Comparative genomics of the Pooid grasses *Brachypodium distachyon* and *Triticum aestivum*

Jonathan Wright BSc. MSc.

A thesis submitted to the University of East Anglia in fulfilment of the Degree of Doctor of Philosophy

John Innes Centre, Norwich, Norfolk

September 2010

Abstract

Comparative genomics is a technique whereby whole genomes are compared allowing the use of species with well-characterised genomes to understand the genomes of species with larger, more complex genomes. The development of new varieties of bread wheat (*Triticum aestivum*) to meet the future challenge of feeding a growing population in a time of environmental change is hampered by difficulties sequencing the large, complex wheat genome. The material presented in this thesis describes the development of genomic resources for *Brachypodium distachyon*, a new experimental grass system which due to its close evolutionary relationship to wheat and compact genome provides a template for comparative structural genomic studies in larger crop genomes.

This thesis describes my contributions to the development of a genetic linkage map, an integrated physical map and an annotated genome sequence for *B. distachyon*, the first pooid grass to be sequenced. Evolutionary relationships between representatives from three major grass subfamilies are assessed and a proposed mechanism for chromosome reduction in the grasses is confirmed. More than 15,000 putative regulatory regions within the *B. distachyon* genome were identified using phylogenetic footprinting of which a subset was shown to contain functional motifs. Analysis of T-DNA insertions in 741 mutagenised *B. distachyon* lines identified 364 genes containing an insertion providing a resource for functional annotation of novel grass genes.

Chromosome-based physical mapping is one approach towards sequencing the wheat genome and the novel method of anchoring physical map contigs from a single wheat chromosome arm using synteny to *B. distachyon* is described. One-third of physical map contigs were anchored and additional markers identified to anchor further contigs. The transcriptome of wheat was sampled using Illumina and 454 platforms resulting in the identification of new transcripts. *B. distachyon* gene models were used to accurately define wheat gene models from this transcript sequence.

List of contents

List of tables

List of figures

List of supplementary material

All publications, data tables and scripts referenced in the main text are included on the accompanying CD.

Supplementary information 1: “Genome sequencing and analysis of the model grass *Brachypodium distachyon*” manuscript published in Nature.

Supplementary information 2: “An SSR-based genetic linkage map of the model grass *Brachypodium distachyon*” manuscript published in Genome.

Supplementary information 3: “An integrated physical, genetic and cytogenetic map of *Brachypodium distachyon*, a model system for grass research” manuscript published in PLoS ONE.

Supplementary information 4: “Distribution and characterization of more than 1000 T-DNA tags in the genome of *Brachypodium distachyon* community standard line Bd21” manuscript published in Plant Biotechnology.

Supplementary information 5: Perl script used to annotate the Brachypodium mid-point genome assembly.

Supplementary information 6: List of 34 SSR markers designed from Brachypodium mid-point super-contigs for genetic mapping.

Supplementary information 7: Perl script to estimate the repetitive DNA content of a BAC using the SyMAP database.

Supplementary information 8: List of 29 SSR markers designed from BES to anchor the Brachypodium physical map contigs to the Brachypodium genetic map.

Supplementary information 9: List of BAC clones used to paint chromosome arms Bd2S (26) and Bd2L (29).

Supplementary information 10: Perl scripts and modules used for CNS identification.

Supplementary information 11: Perl CGI scripts used to display HSPs and CNSs surrounding orthologous genes.

Supplementary information 12: List of 15,991 CNSs identified within the Brachypodium genome.

Supplementary information 13: Perl script to identify the location of a T-DNA insertion with respect to predicted genes in the Brachypodium v1.0 annotation.

Supplementary information 14: List of 175 genes with a predicted function that contain a T-DNA insertion.

Supplementary information 15: Perl script to identify candidate BACs in the wheat 3DL MTP pool from screening results.

Supplementary information 16: List of 1,557 markers designed from wheat ESTs aligned to Brachypodium genes for syntenic anchoring of wheat 3DL physical map contigs.

Supplementary information 17: List of 40 wheat 3DL physical map contigs anchored in test region of Brachypodium.

Supplementary information 18: List of 94 markers designed from brachy genes to anchor wheat 3DL physical map contigs.

Supplementary information 19: List of 42 ISBP markers identified from anchored wheat 3DL contigs in test region of Brachypodium.

Supplementary information 20: Shell script, R script and Perl scripts used to process Illumina sequence files.

Acknowledgements

First I would like to thank my supervisors Professor Michael Bevan and Dr. Martin Trick for providing an exciting and absorbing project on which to work as well as offering support, guidance and sage advice throughout the length of my studies. I would also like to thank Dr. Melanie Febrer, Dr. Aidong Yang and Neil McKenzie for the laboratory work undertaken as part of this project.

I would like to thank everyone in the department of Computational and Systems Biology at the John Innes Centre for providing a positive environment in which to work and to the John Innes Centre in general for providing a stimulating atmosphere in which to learn about and contribute to the exciting endeavours of plant scientists.

Most importantly, I would like to express my sincerest thanks to my family for supporting me throughout my studies, especially to Jamie who provided a much needed distraction at difficult times and to Emilia who timed her arrival perfectly to allow me to complete this thesis. I would also like to express my gratitude to my wife Anna for her patience and advice over the last four years as without her this thesis would not exist.

And finally, thanks to the late George Duncan, Professor of Biomedical Sciences and friend who first stimulated my interest in the area of bioinformatics and whose encouragement and advice started me on this path.

1 Introduction

1.1 Overview

Genomics is the study of the structure, content and evolution of genomes. The development and refinement of sequencing chemistry, based on the Sanger method using fluorescently labelled dideoxy terminators (Sanger *et al.* 1977), and the development of novel high throughput cloning systems, led to a continued increase in the scale of genome sequencing projects between 1980 and 1990. The advent of whole genome shotgun sequencing revolutionised genomics (Fleischmann *et al.* 1995) as it provided a direct and cost-effective means of sequencing simpler genomes without the expense and time-consuming stage of cloning. The first complete genomic sequence of a eukaryotic organism, *Saccharomyces cereviseae* was published in 1996 (Goffeau *et al.* 1996) and was followed by the genomes of other eukaryotes, sequenced using a variety of methods including whole-genome shotgun sequencing. In 2000, the sequencing of the *Drosophila melanogaster* genome was completed (Adams *et al.* 2000) validating the controversial whole-genome shotgun approach and was the largest genome sequenced at that time. Later that year the genome sequence of *Arabidopsis thaliana* was published (The Arabidopsis Genome Initiative 2000), the first plant to have its genome sequenced. February 2001 saw the completion of the first draft of the human genome sequence (Lander *et al.*, 2001; Venter *et al.*, 2001) by two rival groups. The availability of many genome sequences spawned major fields of functional and comparative genomics to establish the function of genes and proteins in a genomic context and to determine the evolutionary processes that act upon them. Computational approaches have been developed to manage data, perform analyses, display and disseminate genomic data.

The field of genomics is currently undergoing another revolution with the advent of massively parallel sequencing technologies which have significantly reduced the cost of genome sequencing. These new technologies have ushered in the era of personal genomics, where an individual's genome can be sequenced for a fraction of the cost and time of the Human Genome Project (Bentley *et al.* 2008; Wheeler *et al.* 2008) with the eventual goal being a human genome sequence for less than $1000. Genome projects are becoming more wide-ranging, for example, the 1000 genomes project (www.1000genomes.org) aims to sequence the genomes of more than 1000 people from a number of different ethnic groups to provide a detailed catalogue of human genetic variation. More recently, an extensive project was announced to assemble a collection of DNA sequences representing the genomes of 10,000 vertebrate species in an attempt to capture the genetic diversity across the vertebrate genera (www.genome10k.org). These new technologies provide additional challenges both in terms of storing and analysing the vast amounts of genomic data currently being produced as well as learning how to apply these technologies to tackle genomes previously considered impossible to sequence, such as the large crop genomes.

This thesis applies a combination of software and visualisation tools to map and understand the genome of the wild grass *Brachypodium distachyon* and to evaluate its use for comparative structural genomic studies in *Triticum aestivum* using novel genomics approaches. Chapter 2 describes my contributions towards the development of *B. distachyon* as a model grass genome focussing on the genome sequencing project, the construction of a genetic and physical map and comparison of the genome with other grasses. Chapter 3 details the use of phylogenetic shadowing using *B. distachyon* and the genomes of two more distantly related grasses to identify potential functional regions within the *B. distachyon* genome. Chapter 4 describes my contribution to early functional genomics studies in *B. distachyon* by bioinformatic analysis of T-DNA insertion lines. In Chapter 5, physical mapping of a single chromosome arm of *T. aestivum* is undertaken using the *B. distachyon* genome to anchor physical map contigs. In addition, different methods for anchoring physical map contigs are evaluated. Chapter 6 evaluates the use of new sequencing technologies to sample the transcribed portion of hexaploid wheat and *B. distachyon* gene models are used to more precisely define gene models in wheat. The conclusions from this thesis are drawn together in Chapter 7. To put the following chapters into context, this introduction reviews the current state of genomics and comparative genomics, and then details the genomic approaches to date that have been applied to the grasses.

1.2 Genomics

The primary goal of a genome sequencing project is to establish a complete, accurate and durable record of the sequence of genomic DNA. Several other types of data are required to assemble genome sequences into their correct chromosomal context. Among these are physical and genetic maps of the genome. A physical map displays the location of genes or markers in a genome according to their physical distance, as measured by the size of large insert clones such as Bacterial Artificial Chromosomes or BACs (Shizuya *et al.* 1992). Genetic maps identify the relative position of markers determined by meiotic recombination events and are essential for mapping important loci within the genome and for comparison with other genomes. Genome projects can also generate expressed sequence tag (EST) data to facilitate the analysis of the gene-coding regions of the genome. EST sequences are obtained by sequencing libraries of clones containing reverse transcribed copies of mRNA sequences extracted from a variety of tissues. After determination of the complete genome sequence, annotation of the genome is performed using gene finding programs that use a variety of methods to define the protein-coding regions and other functional areas of the genome. Using similarities between the predicted sequence of proteins in the genome of interest and related proteins with known functions in other organisms, the functions of genes can be inferred and related to other data sets, such as gene expression data.

In addition to sequencing the genome of a single individual, the reduction in cost and increase in sequencing throughput have facilitated resequencing projects where whole genomes or selected genomic regions of interest from many individuals in a population can be sequenced. Detecting and understanding sequence differences or polymorphisms between individuals in a population is important to explicate the biological consequences of genetic variation.

A single genome project generates a vast amount of diverse data that needs to be provided to the global research community in user-friendly formats. In the past 20 years the field of bioinformatics has emerged and has developed a plethora of software tools, data repositories and visualisation interfaces that provide an essential bridge between biologists and these large scale data sets. Many software tools and repositories are available to display and store genomics data for the huge variety of genomic projects either completed or ongoing.

1.2.1 Genetic mapping of genomes

Mapping is central to genomic studies. For many years genetic linkage maps have provided researchers with the tools to navigate genomes with very high precision without knowing the underlying sequences. Linkage maps represent the relative position of genetic elements within linkage groups and if the marker density is sufficiently high, the order and grouping of markers into linkage groups directly corresponds to the organisation of chromosomes. Inherited morphological features were first used for linkage mapping but have largely been superseded by DNA-based molecular markers. These markers have facilitated the development of comprehensive genetic linkage maps for many species. A molecular marker is a unique DNA sequence that can be easily identified by a specific molecular technique. Types of commonly used molecular markers include Restriction Fragment Length Polymorphisms (RFLPs), Random Amplified Polymorphic DNAs (RAPDs), Sequence Tagged Sites (STS), Amplified Fragment Length Polymorphisms (AFLPs), Simple Sequence Repeats (SSRs) and Single Nucleotide Polymorphisms (SNPs).

Linkage maps are constructed from observations of the frequencies of recombination between markers during crossing over of homologous chromosomes. Crossing over is the process that occurs at meiosis when genetic material is exchanged between paired homologous chromosomes. The observation that if a crossover occurs between markers, they will not be inherited together led to the hypothesis that the probability of a crossover occurring between two markers would increase as the distance between the markers increases (Morgan 1911). Based on these ideas, the first genetic map was developed in *D. melanogaster* (Sturtevant 1913) using the centiMorgan (cM) as a measure of genetic distance. This unit was defined as ‘a portion of the chromosome of such length that, on the average, one crossover will occur in it out of every 100 gametes formed’. Mapping functions such as the Kosambi map function (Kosambi 1944) are used to infer the genetic distance between two loci from the observable recombination rate between them.

In modern genetic studies involving large numbers of markers, computer software such as MapMaker (Lander *et al.* 1987) or JoinMap (Stam 1993) is used to construct genetic maps. Various algorithmic approaches are used to calculate the most likely genetic map from the frequency of recombination between each marker. Genetic maps are widely used in eukaryotic organisms to determine the location of a gene controlling a particular trait and the inheritance pattern can be determined using markers linked to the trait.

1.2.2 Physical mapping of genomes

A physical map consists of a set of ordered clones of large DNA fragments, whose relative positions in the genome are known. Physical maps describe the location of chromosomal features according to relative positions on genomic DNA and the distance between features is measured as a physical distance in kilobases (kb) or megabases (Mb). Construction of a physical map is commonly the first stage in genome sequencing projects aimed at tackling large and complex genomes as it provides multiple anchor points for aligning and assembling sequences and breaks down the genome into smaller (generally 100-200 kb fragments in BAC cloning vectors) units for sequencing and assembly.

Physical maps are constructed by partially digesting genomic DNA into segments of 50-200 kb using restriction enzymes, then cloning these segments into large insert vectors such those based on the *Escherichia coli* F factor, a low-copy plasmid. A library of these vectors or BACs, is created such that any region of the genome will be represented 5 to 10 times in the clone library. Clones are selected and analysed to determine regions of overlap which are then aligned to form larger contiguous segments or contigs. Contigs are extended to cover large tracts of DNA, ultimately producing a series of contigs that correspond to the chromosome structure of the organism. Once optimal coverage has been achieved, a set of clones is selected from the library called the minimal tiling path (MTP) which gives the most complete coverage of the genome with a minimal number of overlapping clones.

Three methodologies have been developed to determine overlaps between clones; restriction enzyme fingerprinting, end-sequencing of clone inserts and hybridisation assays. Restriction enzyme fingerprinting was originally developed for *Caenorhabditis elegans* (Coulson *et al.* 1986) and *S. cerevisiae* (Olson *et al.* 1986). Each clone in a genomic library is digested with restriction enzymes, the resulting fragments separated on a high resolution gel and detected by autoradiography. A fingerprint is produced for each clone which is used to align clones based on overlap of fingerprints. Sets of clones with overlapping fingerprints are assembled into contiguous structures or contigs.

A rapid and more accurate method called high-information content fingerprinting or HICF (Marra *et al.* 1997) is now more commonly used. This method uses large-insert clones (P1 or Bacteria Artificial Chromosomes – abbreviated to PACs or BACs) which yield more fragments when digested with restriction enzymes. In addition, the different sized fragments can be resolved accurately on capillary arrays using fluorescent dyes and sensitive fluorescent imagers rather than by radioactive methods. At the same time as the advent of HICF, a software called FingerPrinted Contigs or FPC (Soderlund *et al.* 1997) was developed to compare the fragment profiles to determine whether clones overlap. The underlying algorithm calculates a probability that the number of shared bands between two clones could occur by chance and uses this to cluster the clones into contigs. FPC also has editing facilities to modify alignments and to remove poorly fingerprinted clones.

Additional information on potential overlap of clones can be gathered by end-sequencing the insert of each clone. This essentially gives a ~500 bp read from each end of the genomic sequence contained within the clone and is called a BAC-end sequence or BES. Overlapping clones can then be identified by overlapping sequence regions. This method was originally proposed to assist in the sequencing of the human genome (Venter *et al.* 1996) and is used in conjunction with restriction enzyme fingerprinting.

A hybridisation assay uses short, radioactively labelled probes called overgos created from overlapping oligonucleotides in the end sequences of BACs in contigs. These probes are used to select BACs from the library that match the ends of contigs. If the probe sequence is present in low copy number within the genome, only a few clones are selected from the library representing potential overlapping clones. A probe is selected from the end of the new clone and used to select another BAC from the library. This allows contigs to be extended forming larger assemblies, a technique known as chromosome walking (Bender *et al.* 1983).

Physical and genetic maps can be aligned by use of common markers. Although both types of maps show the order of features along the genome, the distances between features may vary in each map. This is due to variation in recombination rates over different parts of the chromosome, for example, at the centromeres and telomeres of the chromosomes which usually show less recombination than the gene-rich euchromatin. In chromosomal regions showing low recombination one cM will represent a much larger physical distance than in regions showing high recombination. The advantage of a physical map is that regions of the genome are mapped independently of recombination frequencies, therefore most areas of the genome that can be cloned can be placed in physical maps. In contrast, large tracts of genomes such as pericentromeric heterochromatin and other repetitive regions have suppressed recombination and cannot be accurately genetically mapped.

1.2.3 Genome sequencing strategies

Traditional genome sequencing utilises the chain terminator DNA sequencing technique developed by Frederick Sanger (Sanger *et al.* 1977) which uses dideoxynucleotide triphosphates as DNA chain terminators. A single-stranded DNA template is incubated with a DNA primer, a DNA polymerase, a mixture of the four deoxynucleotides (dATP, dGTP, dCTP, and dTTP) one or more of which is radioactively labelled, and a dideoxy terminator (either ddATP, ddGTP, ddCTP or ddTTP). This reaction results in a mixture of DNA fragments of differing lengths complementary to the template with labelled nucleotides incorporated into the chain and each fragment terminated by a dideoxy terminator. The dideoxynucleotides are added at a lower concentration than the standard deoxynucleotides such that only partial incorporation of the terminator occurs. Analogous reactions are performed using the other three dideoxy terminators and the four samples are fractionated in parallel using electrophoresis on a denaturing acrylamide gel which separates the fragments by size. This results in a pattern of bands on the gel which are visualised by autoradiography and from which the DNA sequence can be read directly. Early DNA sequencing was a manual, labour intensive process requiring four reactions for each sequence - one for each dideoxy terminator. The four reactions were run on a gel for a specified time, then stopped so that the results could be visualised by exposing the gel to X-ray film, which was read manually. Modern genome sequencing operations are based on this technique but are heavily automated and can accurately read sequences of up to 1000 bases. Improvements to the original technique include the use of fluorescent dyes to label the four dideoxy terminators. This means that only one reaction is needed as the four terminators can be distinguished by colour and the sequencing reaction no longer needs to be stopped as the products are scanned for laser–induced fluorescence as they run through the electrophoresis medium. The base sequence is collected as a set of trace files which indicates the intensity of each of the four colours as a set of peaks which are read automatically by base-calling software. Programs such as phred (Ewing and Green 1998; Ewing *et al.* 1998) enable sequence reads to be taken directly from an automated sequencer and converted into data files along with quality scores assigned to each base-call. Complementary software tools phrap and consed (Gordon *et al.* 1998) allow the alignment and editing of the final DNA sequence. These tools use similarity matching algorithms that also take into account the quality scores assigned in the first step. Any ambiguities in the sequence or regions of poor coverage are highlighted at this stage and can be targeted individually. Multiple high-quality reads per base means that the final consensus sequence has high accuracy.

Traditional strategies for genome sequencing are based on physical maps where a minimum tiling path of clones has already been identified. To determine the genome sequence, the clones from the MTP (usually 100-200 kb in size) are randomly fragmented into fractions of 3-5 kb and 10 kb and cloned into plasmid vectors for sequencing. The sequence of each BAC clone is reconstructed from the plasmid clone sequences and the genomic sequence determined from the known order of the BAC clones in the physical map. A finishing step is normally required to check for errors in the sequence and to identify physical and sequence gaps. The hierarchical clone-based approach was used in early sequencing projects including *E. coli* (Blattner *et al.* 1997) and *C. elegans* (The *C. elegans* Sequencing Consortium 1998).

A more recent approach to genome sequencing is the whole-genome shotgun (WGS) method originally proposed in 1992 by Eugene Myers (Kececioglu and Myers 1992). This approach takes DNA from a whole genome and shears it randomly into separate preparations containing fragments of different sizes (usually 2 kb, 15-20 kb, 50 kb and 150 kb) which are ligated into cloning vectors of the appropriate size. The vector inserts are sequenced from both ends yielding two reads for each clone called mate-pairs. The orientation of each read in the mate-pair and their approximate distance apart is known and this information can be used when reconstructing the original sequence. Overlapping reads are first used to assemble regions of contiguous sequence (contigs), then mate-pair information is used to link contigs together to produce longer scaffolds. In this way the longer sequences provide a larger-scale framework for assembly of the smaller stretches of DNA. A useful statistic to compare the quality of genome assemblies is the N50 length. An assembly described as having an N50 length of 100 kb means that 50% of the bases within that assembly reside in contigs longer than 100 kb. More complete assemblies will have a higher N50 length. Assembly of WGS sequence is computationally expensive requiring specialist assembly software such as ARACHNE (Batzoglou *et al.* 2002). Myers suggested using the WGS approach in the Human Genome Project (Weber and Myers 1997) as a low cost alternative to the hierarchical clone-based approach chosen by the public consortium (Lander *et al.* 2001). WGS was criticised at the time (Green 1997) for downplaying the expense of the finishing stage and producing sequences of low quality, but a private company Celera opted to use this method (Venter *et al.* 2001). Both groups published draft sequences in 2001, instigating arguments as to which sequence was more accurate. The first organism to have its genome sequenced by the whole-genome shotgun method was *D. melanogaster* (Adams *et al.* 2000), and several others such as mouse (Waterston *et al.* 2002) and chimpanzee (Chimpanzee Sequencing and Analysis Consortium 2005) have followed. The hierarchical approach has the advantage of an easy to assemble final sequence, as the position of each BAC clone in the genome is known. However, it is more expensive and time consuming to perform the extra cloning step. The WGS approach is faster and less expensive but more prone to errors due to the complexity of the assembly process.

Over the last few years, new techniques for high-throughput genome sequencing have emerged, driven by the need to reduce the cost of the process. These next-generation sequencing (NGS) technologies have lessened the reliance on traditional Sanger sequencing and competition between different companies has reduced the cost still further. The main differences between these technologies and traditional sequencing is the length of the sequence reads and the amount of data produced. In general, the new methods produce much shorter reads although this field is advancing at a remarkable rate generating increases in read length and reductions in cost. The first NGS method to become commercially available was from 454 Life Sciences (now owned by Roche) and makes use of a technology called 'pyrosequencing' (Margulies *et al.* 2005). Genomic DNA is fragmented into small sections of a few hundred base pairs. Short adapters are added to each fragment and the single-stranded DNA is ligated onto a bead. Each bead is captured in a droplet of an emulsion mixture creating a microreactor where emulsion PCR amplification takes place. The fragments on each bead are amplified in parallel resulting in several million copies per bead. Each bead is deposited into a well on a specially designed plate containing hundreds of thousands of wells, and then individual nucleotides in a fixed order are flowed over the wells. When a nucleotide complementary to the template strand is incorporated, a chemiluminescent signal is emitted and recorded by a camera. Images containing signals from all the wells on the plate are processed to determine the sequence of the fragments from each well in parallel. This technology can produce more than one million reads per run with each read around 400 bp in length. Run time is around 10 hours. The limitation of this technology is that it is difficult to accurately determine the length of stretches of sequence where a single base is repeated, called homopolymeric runs. This is due to problems in determining the intensity of the signal when multiple nucleotides are incorporated.

The second NGS technology to appear on the market was developed by Solexa and is now owned by Illumina. This technique uses a reversible terminator-based method called ‘sequencing-by-synthesis’ (Bennett *et al.* 2005) where genomic DNA is randomly fragmented and adapters are ligated to each end of the fragments. These adapters allow single-stranded fragments to be attached to the planar, optically transparent surface of a flow cell. Each fragment is subjected to a solid-phase bridge amplification reaction that creates hundreds of millions of dense clusters, each consisting of about 1000 copies of the original fragment. The sequencing cycle proceeds by flowing a mixture of four labelled reversible terminators, primers and polymerase over the flow cell so each cluster will have a base incorporated that is complementary to the first base of the fragment. After washing, a laser is used to excite the labelled terminators and the emitted fluorescence from each cluster is captured as an image. The fluorophore is removed and a 3’ hydroxyl group is regenerated on the terminator ready for the next cycle of terminator addition. Subsequent cycles identify the sequence of bases in each fragment and as before, all the clusters are sequenced in parallel. The read length obtained using this technology is currently around 100 bp with 150-200 Gb of sequence being produced in a single 8 day run.

The ‘Sequencing by Oligonucleotide Ligation and Detection’ (or SOLiD) system from Life Technologies (formerly called Applied Biosystems) was the third NGS technology to become available (Valouev *et al.* 2008). The first stages of this method are similar to the pyrosequencing method where genomic DNA is fragmented and each fragment ligated onto a bead. Each bead is placed in a microreactor where amplification occurs such that each bead contains many copies of the fragment. The 3' ends of the fragments are modified so the beads can be attached to a glass slide. Sequencing by ligation proceeds by hybridising a primer of length n to the adapter sequence that attaches the fragment to the bead, then a set of labelled di-base probes compete for ligation to the primer. In total, sixteen di-base probes are used corresponding to every combination of the four nucleotides. Each probe consists of a di-base followed by three degenerate bases and a fluorophore. Once a probe is hybridised, it is interrogated to determine which di-base probe is attached before removal of the labelling fluorophore. This leaves the di-base and three degenerate bases attached to the template sequence. The identities of the two bases immediately following the primer sequence have been determined. Another round of ligation is initiated and a second probe attaches to the template sequence adjacent to the degenerate bases of the previous round. Multiple cycles of ligation, detection and cleavage proceed until the end of the fragment is reached. The next stage is a primer reset where the original primer and extended sequence is melted off and a primer of length n-1 is annealed. Again, multiple rounds of ligation, detection and cleavage proceed until the end of the fragment is reached. After this round, half of the bases interrogated in the previous round will have been interrogated again. Once again, a primer reset is performed and the whole cycle is executed three more times with primers of length n-2, n-3 and n-4. Upon completion, each base in the fragment will have been interrogated twice meaning that sequencing errors are reduced. The output of the sequencing reaction is a sequence of fluorophores corresponding to the positions of the di-base probes on the sequence. This is referred to as 'colour space' and is translated into 'sequence space' using a conversion table. As only four fluorophores are used to label sixteen di-base probes, translating a specific colour into sequence is dependant on the colour both before and after the position being translated. The SOLiD system can currently produce read lengths of 50 bp and up to 100 Gb of sequence in one 12 day run.

In order to increase the usefulness of the short reads produced by NGS technologies, the ability to produce reads separated by a known distance was quickly introduced into all of the methods described above. This addition goes some way to ameliorating the disadvantages of short reads by providing additional long range positional information which can be used for subsequent alignment or assembly of reads. The reads are called 'mate-pairs' or 'paired-end reads' and although these terms are sometimes used interchangeably, the difference between them refers to the protocol used in the library preparation and thus the distance between the reads. Mate-pair reads are separated by 2-5 kb whereas paired-end reads are rarely more than 500 bp apart and can even be overlapping if the insert size is small.

A less popular sequencing platform, called the Polonator, has been developed as an open source solution with software and protocols freely available in contrast to the other commercial options. This technology is also based on sequencing-by-ligation and was originally proven by sequencing the *E. coli* MG1655 genome (Shendure *et al.* 2005). The first stage of the process requires the construction of a mate-pair library from size-selected genomic fragments. The fragments are attached to beads and amplified using emulsion PCR before being attached to a flow cell. The sequencing reaction proceeds using one of four anchor primers (named A1 to A4). The primers are designed to hybridise to each side of the two mate-pair reads contained within each fragment, called the proximal and the distal tag. Fluorescently labelled probes are used to interrogate the sequence in each primer cycle; seven probes are used with primers A1 and A3 and six probes with primers A2 to A4. The probes are degenerate in all positions except the query position. After hybridisation of the first primer (A1) followed by seven rounds of interrogation, a 7-base sequence has been read from one side of the proximal tag. This is removed, then the second primer (A2) is hybridised to the other end of the proximal tag and the sequence is interrogated by six probes. This process continues using the other two primers resulting in one six and one seven base sequence from each tag. This platform produces 26 base pair reads, with 4-5 Gb of sequence per run.

The NGS platforms discussed so far involve an amplification step, either bridge amplification or clonal amplification using emulsion PCR. Direct sequencing of a single DNA molecule that doesn't require a cloning step does not produce PCR artefacts, and is therefore potentially more accurate. Helicos Biosciences was the first company to offer this service in their HeliScope Single Molecule Sequencer. Genomic DNA fragments (200-250 bp) are modified by the addition of a common adapter allowing them to be directly bound to a solid support (Harris *et al.* 2008). A sequencing-by-synthesis method similar to that used by Illumina is used to interrogate the fragment. As a labelled reversible terminator is incorporated into the synthesised strand complementary to the template, a fluorescent signal is recorded. The terminator is unblocked and the reaction proceeds in a cycle manner until the read length is reached. This method produces reads of 25 to 35 nucleotides per template and 21-35 Gb of sequence per run.

Another generation of sequencing technologies are already becoming available, called third generation technologies, these methods are based around the ability to directly sequence long single-stranded DNA molecules in real-time. Leading the way with this approach is Pacific BioSciences who have developed a Single Molecule Real Time (SMRT) sequencing technology to capture the incorporation of labelled nucleotides during DNA synthesis (Eid *et al.* 2009). A DNA polymerase molecule is attached to the bottom of a nano-scale chamber. Single stranded DNA is pulled through the polymerase and sequence information is captured as phospholinked nucleotides are incorporated into the strand. The nucleotides have a fluorescent label attached to the terminal phosphate rather than the base meaning that as a nucleotide is incorporated, the fluorescent label can be detected. The nano-scale chamber is called a zero-mode waveguide and is designed to provide a minimal observation volume, thus reducing the background fluorescence and allowing the fluorescent label of the incorporated nucleotide to be detected before it is released and diffuses away (Levene *et al.* 2003). The polymerase incorporates bases at a rate of around 10 per second and this data is captured in real time. In addition, thousands of detection chambers can be imaged in parallel giving an exceptionally high throughput. This sequencing platform was unveiled at the Advances in Genome Biology and Technology conference held in Marco Island, Florida in February 2010 (Munroe and Harris 2010) and was reported to have produced reads of up to 10 kb. Furthermore, a 'strobe sequencing' method was discussed where images are taken for short time periods resulting in blocks of sequence separated by unknown regions distributed over very long insert lengths which should provide useful scaffolds for *de novo* assembly of complex genomes (Pacific Biosciences 2009).

At the same conference, Life Technologies outlined their third-generation approach which uses a quantum dot (Qdot) nanocrystal attached to a DNA polymerase. The Qdot absorbs laser light and as labelled nucleotides get incorporated into the DNA strand by the polymerase, light is transferred from the Qdot to the labelled nucleotide by a process called fluorescence resonance energy transfer, or FRET. This light is emitted by the fluorescent dye and detected. FRET only happens in close proximity to the Qdot so the emitted signal is spatially confined and additional labelled nucleotides in solution are not detected. The advantage of this method is that the excitation laser requires less energy than other methods which reduces damage to the polymerase and therefore should allow longer read lengths.

The third main competitor in real-time single molecule DNA sequencing is Oxford Nanopore whose technology combines nanopores and electrical detection. A DNA strand is guided towards a pore in a membrane surrounded by exonuclease enzymes which sequentially cleave each base from the strand and direct them through the nanopore. A cyclodextrin molecule within the nanopore transiently binds each base as it passes through and the disturbance in electrical current is measured to determine the identity of the base (Astier *et al.* 2006). The advantage of this method is potentially reduced running costs compared to other methods as it is not dependant on fluorescent labelling and advanced optimal imaging.

These new sequencing technologies promise to vastly reduce the cost of DNA sequencing and push the frontiers of what is currently possible in genomic and genetic research. Shorter run-times and less complex assembly will allow a multitude of genomes (or genomic regions) to be sequenced. One of the challenges for the future is to develop new methods of data storage and analysis to deal with this deluge of genomic data.

1.2.4 Alignment and assembly of NGS sequence data

NGS sequencing technologies have spawned a multitude of computational tools to deal with the huge amount of sequence data being produced. These tools fall into two classes, those designed to align short reads to reference genomes and those designed to assemble short reads *de novo*, required when no reference genome is available. When used for resequencing projects, short genomic reads from the species of interest are aligned to a reference genome to determine differences between the genomes, for example, to identify SNPs and insertions/deletions (called indels). In general, the first stage in aligning short reads is to use a heuristic algorithm to identify a small subset of positions in a genome where a read might align, before applying more sensitive alignment algorithms to this smaller dataset. Two approaches have been used in the heuristic stage of the alignment; hash-based alignment and Burrow-Wheeler transform methods. Hash-based methods use a hash table data structure to index either the set of input reads or the reference genome. The hash table index allows for rapid searching of large datasets. This method was implemented in the first generation of short read alignment software including MAQ (Li *et al.* 2008a) and SOAP (Li *et al.* 2008b). Alignment methods based on the Burrows-Wheeler transform or BWT (Burrows and Wheeler 1994) first apply the BWT to the genome sequence which involves transforming the sequence such that repeated characters are grouped together to allow more efficient compression. An index is created from the transformed sequence which is used to place short reads onto the genome. Using the BWT on a genome before the creation of an index often produces a much smaller index than that produced from a non-transformed genome. This approach is implemented in BWA (Li and Durbin 2009), BOWTIE (Langmead *et al.* 2009) and SOAP2 (Li *et al.* 2009b) which generally run much faster than hash-based algorithms. Standard file formats are beginning to emerge for storing aligned reads produced by the different tools which simplifies the downstream analysis of datasets (Li *et al.* 2009a).

Traditional assembly algorithms such as those used for WGS assembly, are based on identifying overlapping reads to assemble into contigs but this approach is not feasible for the huge numbers of short reads produced by NGS platforms. Assembly algorithms for NGS data are based on de Bruijn graph data structures containing nodes joined with edges, an approach first applied to sequence assembly in 1995 (Idury and Waterman 1995). Each node represents a short fixed-length subsequence, called a *k*-mer, where *k* is the length of the subsequence. A set of *k*-mers is constructed from the sequence reads with one node for each *k*-mer found in the reads and an edge between nodes if the *k*-mers they represent are found together in a read. In addition, the coverage of each *k*-mer is recorded. In the graph, continuous stretches of nodes joined with edges represent error-free sequence and sequencing errors appear as bubbles or tips that end abruptly. At this point the graph is simplified by amalgamating the continuous stretches and correcting errors by bubble removal and tip clipping. Repetitive regions in the sequence will be represented as different paths through the graph, called cycles. Similar to traditional assemblers based on sequence overlap, paired-end reads are essential for traversing repeats in a genome and most NGS assembly algorithms incorporate paired-end read information when building assemblies. The first commercial NGS assembler to be released was Newbler, designed to deal with sequence reads produced by pyrosequencing and supplied with every 454 sequencing machine. Subsequently, assembly algorithms such as Velvet (Zerbino and Birney 2008) and EULER-SR (Chaisson and Pevzner 2008) were released, designed for Illumina reads and used to assemble small bacterial genomes. Most recently, assemblers such as ABySS (Simpson *et al.* 2009), AllPaths (Maccallum *et al.* 2009), and SOAPdenovo (Li *et al.* 2009c) have been released, implementing algorithms with reduced memory requirements to allow the assembly of larger genomes. For example, SOAPdenovo was used to assemble a draft sequence of the 2.4 Gb genome of the giant panda (Li *et al.* 2010).

1.2.5 Genome annotation

Knowledge of the physical DNA sequence of a genome does not provide much useful information. Identification of the gene-rich regions in each chromosome, characterisation of the intron-exon structure of the genes within these regions and determination of the protein products derived from these genes provide the basis for functional genomics and proteomic studies. This helps to define the components of the system and determine how they work together. A gene comprises one or more exons, possibly separated by introns, transcribed to form an RNA transcript in the first stage of protein production. A single gene can give rise to multiple transcripts, and thus multiple distinct proteins with multiple functions, by alternative splicing and alternative transcription initiation and termination sites. Identification of the DNA sequences where transcription is initiated and regions that determine how exons are spliced together helps to describe the structure of each gene and transcriptional unit and is the first step in characterising the genome.

Gene annotation by experimental means is a costly and time consuming process. Therefore, driven by the large number of completed and ongoing sequencing projects, automated computational or *ab initio* gene prediction has become the norm. The software recognises structural features of the genome common to genes such as initiation sites and splicing signals at intron-exon boundaries. This method of gene identification is used as a ‘first pass’ annotation and any predictions need to be verified by additional evidence before being accepted as likely to be an actual gene. In order to identify genes, algorithms must identify all the exons within a gene by detection of promoter regions, 3’ polyadenylation sites and sequences that control the excision of introns. In small bacterial genomes this is a relatively simple task due to the lack of introns but in the larger genomes of higher eukaryotes it is more difficult due to complex gene structures and processes such as alternative splicing. The first algorithms to attempt this used positional weight matrices to search DNA sequences for defined motifs.

A more rigorous approach makes use of Markov models which allow the local characteristics of sequence motifs to be modelled within a statistical framework as well as information from previous studies to be incorporated into the analysis. A model is defined as a series of ‘states’ and the transition between each state is assigned a probability score. Probability scores are obtained by training the model using a dataset of DNA sequences that contain experimentally confirmed gene structures, usually from a related species. A model will have a number of paths though it and each path will have an associated probability score. Each path gives a different combination of sub-sequences and the path with the highest probability score is the one that fits the model most closely and thus is the most probable prediction. The first Hidden Markov

Model (HMM) for gene prediction was published in the mid-1990s (Stormo and Haussler 1994) and was followed by software implementations such as Genie (Kulp *et al.* 1996) and GENSCAN (Burge and Karlin 1997). An alternative approach used linear discriminant analysis to identify splice sites, exons and polyadenylation sites (Solovyev *et al.* 1994). A variant of this algorithm called FGENESH (Salamov and Solovyev 2000) is based on a HMM and is used in a variety of genome sequencing projects today. *Ab initio* gene prediction methods make predictions from the information present in the DNA sequence only. These algorithms give high sensitivity but low specificity, meaning that a high proportion of actual coding regions are identified correctly but many of the predictions are incorrect. Gene boundaries, small exons and large introns are not predicted consistently and the accuracy of these methods is largely dependant on the quality and scope of the training dataset. In addition, regulatory elements are difficult to identify using sequence information alone.

More recent improvements in computational gene-prediction software have been made by programs that use comparisons between two genomes as an indicator of which regions are conserved between species and therefore may be functional. These comparative methods supplement traditional methods to give improved specificity and sensitivity. Examples include SLAM (Alexandersson *et al.* 2003) and DoubleScan (Meyer and Durbin 2002), which use a variation of HMM called pair-HMM. Other methods are comparative extensions of previous gene-prediction algorithms such as TwinScan (Korf *et al.* 2001), SGP-2 (Parra *et al.* 2003) and GenomeScan (Yeh *et al.* 2001). Exofish is an approach developed for cross-species gene prediction which compares exon structures to identify conserved regions (Roest Crollius *et al.* 2000). TwinScan has been used to identify conserved genomic regions between *A. thaliana* and *Brassica oleracea* (Ayele *et al.* 2005; Katari *et al.* 2005) and Exofish has been optimised for *Oryza sativa* and *A. thaliana* comparisons (Castelli *et al.* 2004). Both studies resulted in many new gene predictions.

An efficient method used to identify genes is to study the transcribed part of the genome, the transcriptome, by sequencing reverse transcribed copies of mRNA sequences, called complementary DNA (cDNA). A partial sequence from a cDNA forms an expressed sequence tag (EST). A cDNA library is built up by sampling a wide array of tissues from the organism under a variety of conditions. EST sequences obtained from the library are aligned to genomic DNA to identify the positions of the open reading frames (ORFs) which code for the proteins. This requires 'spliced alignment' where the transcript sequence is aligned to genomic DNA allowing for large gaps within the alignment representing the introns. Many alignment tools have been developed to perform this type of alignment such as BLAT (Kent 2002), GMAP (Wu and Watanabe 2005), Spidey (Wheelan *et al.* 2001), and GenomeThreader (Gremme *et al.* 2005).

BLAT and GMAP have been incorporated into a comprehensive genome annotation pipeline called the Program to Assemble Spliced Alignments or PASA (Haas *et al.* 2003). PASA takes a collection of transcript sequences, quality checks them and cleans them for vector contamination, aligns them to the genome sequence, clusters high quality alignments into gene structures, separates alternatively spliced isoforms, then uses the resulting gene models to annotate the genome. PASA was developed to improve the *A. thaliana* genome annotation and resulted in the modelling of several novel genes, more than 1,000 alternative splicing variations and updates to just under half the annotated protein coding genes.

Genome annotation using transcript data relies on the availability of an extensive collection of EST sequences, which are both time consuming and expensive to collect, especially for more complex organisms. In addition, the identification of sequences with low-abundance transcripts is sometimes problematic. In cases where little or no transcriptome data is available for the species being studied, existing EST sequence from closely related organisms can also be used. Transcript sequences from many different organisms are available in NCBI's RefSeq database (Pruitt *et al.* 2007). Novel transcribed sequences in *A. thaliana* have been identified using a combination of full-length cDNA data from *A. thaliana* and additional EST data from *Brassica*, rice and wheat, many of these falling in hitherto unannotated regions (Yamada *et al.* 2003).

In practice, genome annotation projects use a combination of *ab initio* and experimental methods to accurately identify protein-coding genes. An automated pipeline using *ab initio* predictions and alignment of transcript assemblies provides a first pass annotation. In addition, a training set of accurately curated genes is constructed by manual annotation. This training set is used with a statistical combiner such as JIGSAW (Allen and Salzberg 2005) to assess the accuracy of each line of evidence for the predicted genes. Consensus gene models are constructed from statistically accurate predictions derived from the training set.

Genes producing functional RNA rather than proteins are difficult to identify because the transcripts are not polyadenylated and therefore do not exist in cDNA libraries. In addition, orthologous regions display little similarity as the function of these sequences relies on the conservation of secondary structure rather than coding sequence. Transfer RNAs (tRNAs) fold into a cloverleaf structure by base pairing between short sequences where local complementarily is preserved rather than more long range sequence conservation. Software such as tRNAscan-SE (Lowe and Eddy 1997) has been developed to identify such structures using an advanced search grammar based on sequence similarity and base-pairing potential.

In addition to the identification of protein-coding genes, comparative methods are used to identify potential regulatory regions in genomic sequences. It is assumed that functional regions will be under selective pressure and therefore will remain more conserved between species than the surrounding non-functional regions. This technique is called ‘phylogenetic footprinting’ and has been used to identify conserved sequences from multiple alignments of orthologous regulatory regions from several species (Blanchette and Tompa 2002; Lenhard *et al.* 2003). Regulatory regions are usually located upstream of the transcription start site but can also be within introns and downstream of protein-coding genes. A variant of this technique called ‘phylogenetic shadowing’ compares sequences from closely related species and takes into account the phylogenetic relationship of the species being compared (Boffelli *et al.* 2003). Phylogenetic shadowing is implemented in a program called eShadow, developed for the identification of elements under selective pressure by the alignment of multiple sequences from closely related genomes (Ovcharenko *et al.* 2004). Comparative methods are also used to identify sequences that code for micro-RNAs (miRNAs), a type of functional RNA that plays an important role in the regulation of gene expression. A study exploited conservation between the genomes of *A. thaliana* and *O. sativa* to identify miRNA encoding genes in the *Arabidopsis* genome (Bonnet *et al.* 2004).

Repetitive elements are generally annotated by similarity-based detection as implemented in RepeatMasker (Smit *et al.* 1999) which uses a database of known repetitive elements. For genomes whose repeat content is not known, *de novo* detection methods are used. The RECON tool (Bao and Eddy 2002) uses pairwise similarity, clustering and boundary calling to build a set of repeat families and more recent tools like RepeatScout (Price *et al.* 2005) employ *k*-mer frequency analysis to identify high copy reads to use as seeds for finding repeats. A genome sequence is represented as a set of unique *k*-mers alongside a count of the number of times each *k*-mer occurs within the genome. *K*-mers occurring at high frequency are likely to represent repetitive sequence.

A simple method of annotating an uncharacterised genome is to transfer the annotation from a well characterised model genome to the new genome based on sequence similarity. Software such as Projector (Meyer and Durbin 2004), GeneMapper (Chatterji and Pachter 2006) and GeneWise (Birney *et al.* 2004) have been developed for this purpose and use a variety of gene mapping methods.

Once a gene has been identified, its function needs to be determined. Genome projects generally ascribe functions to genes using sequence similarity (with varying levels of similarity between projects) to relate genes to those with known function from other organisms. Whilst efficient and useful at a general level, inferring gene function between very different organisms such as plants and animals can sometimes be misleading. This method also means that existing annotation errors can be propagated onto new annotations. Even in well characterised genomes, many genes still fall into the category of 'unknown function'. As the functions of many genes have not been experimentally determined (only ca. 20% of *Arabidopsis* genes have been subject to experimentation), more robust yet high-throughput informatics methods have been developed that aim to classify a gene through a Gene Ontology (GO). The Gene Ontology consortium (Gene Ontology Consortium 2006) has established a standard vocabulary for the classification of genes that can be used across species and databases. The consortium also coordinates an ongoing effort to completely annotate a set of twelve reference genomes for eventual use in the annotation of other genomes (Reference Genome Group of the Gene Ontology Consortium 2009). This ontology is now widely used and a variety of software tools incorporate GO classifications. Other ontologies have been produced such as MapMan (Thimm *et al.* 2004; Usadel *et al.* 2005) and the MIPS Functional Catalogue (Ruepp *et al.* 2004).

1.2.6 Genomic repositories and visualisation of genomics data

One of the main challenges in genomics is the storage and visualisation of data. A wealth of genomics information is being generated by groups across the world from a variety of organisms. This information needs to be available in standard data formats and easily accessible to be of use. A number of data repositories exist including EMBL (Kulikova *et al.* 2004), Genbank (Benson *et al.* 2007), UniProt (Wu *et al.* 2006) and DDBJ (Miyazaki *et al.* 2004) which in addition to storing and providing raw biological data also provide a wealth of analytical tools. Genomic data is provided in a variety of formats, depending on the source of the data with common data formats including EMBL, Genbank and FASTA. Many specialised resources also exist such as Gramene (Jaiswal *et al.* 2006) which provides tools for comparative genome analysis in the grasses.

Tools to visualise genomics data can be divided into two main areas; visualising sequence data, and browsing genome annotations. Automated Sanger sequencing produces trace files which are converted into base calls and assembled using programs such as phred and phrap (Ewing and Green 1998; Ewing *et al.* 1998). Viewers such as consed are used to inspect and edit the final assembly and allow the trace files upon which the consensus sequence is based to be viewed (Gordon *et al.* 1998). Over the last few years, a new generation of visualisation tools have been developed to deal with the large amounts of data produced by high-throughput sequencing projects. These tools are generally designed to visualise the alignment of short reads to a reference genome. Maqview is an example of such a tool, designed specifically to view the output from MAQ (Li *et al.* 2008a). More generic tools also exist including Tablet

(Milne *et al.* 2010) and Eagleview (Huang and Marth 2008) which are designed to display data from different analysis pipelines.

Once a genome has been assembled and annotated the data is usually made available to the scientific community via web-based genome browsers. Several implementations of this concept exist including the University of California Santa Cruz (UCSC) Genome Browser (Kuhn *et al.* 2007), the ENSEMBL genome browser (Hubbard *et al.* 2007) and the NCBI MapViewer (Wheeler *et al.* 2007). These browsers allow the user to navigate through the genomes of multiple species at varying levels of detail. Gene annotations and other data (including additional genomes to allow the comparison of multiple genomes) are displayed as tracks aligned to the reference genome and can be customised depending on the data required. GBrowse, the Generic Genome Browser (Stein *et al.* 2002) has been developed by the Generic Model Organism System Database (GMOD) Project and is used for a wide range of organism specific sites including The *Arabidopsis* Information Resource (Swarbreck *et al.* 2008), WormBase (Harris *et al.* 2010) and FlyBase (Tweedie *et al.* 2009).

Following developments in web technologies such as Asynchronous JavaScript and XML (AJAX) which allow asynchronous loading of content (as opposed to traditional synchronous page loading), more interactive genome browsers have been developed. A JavaScript version of GBrowse called JBrowse is available which allows smoother scrolling and a more intuitive selection of genomic features to display (Skinner *et al.* 2009). An unpublished browser called Anno-J (Tonti-Filippini 2010) utilises these new technologies to the full providing an interactive browser interface with the ability integrate data from multiple web services as different tracks. It is used to display maps of the epigenome of *Arabidopsis* (Lister *et al.* 2008).

1.2.7 Interoperability and analysis workflows

A key requirement of all the tools used in genome informatics is interoperability. Genome annotation and sequence analysis is often performed in stages by different components integrated into workflows. Components can be locally installed software tools or services on other internet servers that allow remote access to web services. Each component performs a specific analysis and passes data to the next stage of the workflow using standard data formats. The FASTA format is used for DNA and protein sequences and the General Feature Format or GFF (Durbin and Haussler 2010) is used for features associated with biological sequences such as genes. The Perl programming language is often used to link the components of a workflow together and this has led to the development of BioPerl, a suite of open source software libraries to facilitate the creation of customised pipelines to perform a wide range of bioinformatics tasks (Stajich *et al.* 2002). These libraries include modules for common DNA and protein sequence manipulation and the import and export of data between different file formats or data sources. In addition, BioPerl also enables various molecular biology programs to be integrated into a single workflow. Many of these libraries are also available in the Java, Ruby and Python programming languages.

Graphical tools have also been developed to enable the development of workflows without requiring any programming knowledge. The Taverna tool allows different web services to be integrated into bioinformatic workflows using a graphical interface which can be run, saved, modified and reused for subsequent analysis (Hull *et al.* 2006). This approach is taken a stage further with the development of ^{my}Experiment, a web-based 'Virtual Research Environment' for the discussion and sharing of workflows (De Roure *et al.* 2007).

The completion and publication of genome sequences from multiple species allows researchers to compare genomes to understand how evolutionary processes have shaped them. This field is called comparative genomics.

1.3 Comparative genomics

Biological investigations often involve comparisons, from Darwin's observations of beak shape in the Galapagos Archipelago finch population to comparison of phenotypic markers in fruitfly. Comparative genetics focussed on investigations into the similarities and differences between genes of related organisms and comparative genomics takes this a step further to compare the whole genomes of related organisms. The development of genetic and physical maps for a range of species coupled with the elucidation of genome sequences for many model organisms has provided the data for these studies. Comparative genomics provides a foundation for understanding evolutionary relationships between organisms, for understanding the processes of domestication, and for exploiting natural variation in genes to understand complex genetic traits.

A genome sequence captures a complete, accurate and durable record of the changes experienced by an organism over evolutionary timescales. Therefore comparing genomes can help us understand changes that have occurred within each genome since divergence from a common ancestor. Genomes of closely related species exhibit conservation of gene position and order which is gradually eroded as evolutionary distance increases due to the many processes contributing to genome change. At very large phylogenetic distances (>1 billion years, such as that separating the last common ancestor of plants and animals) the order of genes and other regulatory regions are not generally conserved but comparisons at this distance are used to make general observations about the encoded protein sets of different eukaryotes (Rubin *et al.* 2000). Comparison of genomes separated by intermediate phylogenetic distances (70-100 million years) can be used to identify conserved regions that contain functional and non-functional DNA. The assumption when making these type of comparisons is that the evolution of functional areas of the genome (coding exons, noncoding RNAs and regulatory regions) will be constrained by negative selection and these regions will have changed less than non-functional regions (Jukes and Kimura 1984). Comparison of genomic sequences from closely related species such as human and chimpanzee uncovers the sequence differences that may account for differences between the species. These sequences will be under positive selection, i.e. nucleotide substitutions that impart something of benefit to the organism and will thus be retained.

The concept of homology is central to comparative genomics. Two genes are homologous if they are related through descent from a common ancestor and are distinct from analogous genes which are unrelated genes that have evolved similar functions due to selective pressures. Homologues can be further classified by their biological relationship as orthologues or paralogues. Orthologues are true homologues; they have diverged from a common ancestor via a speciation event. Paralogues are genes that have arisen by gene duplication since the speciation event that separates the two species being compared. The terms orthologue and paralogue first appeared in the literature in a discussion by Fitch on the importance of distinguishing between homologous and analogous proteins (Fitch 1970).

The analysis of orthologous and paralogous gene-pairs can suggest a hypothetical evolutionary tree or phylogeny that explains the hierarchical ancestral relationship of the genes (Bowers *et al.* 2003b). It can also enable a gene to be classified into an appropriate subfamily. Phylogenetic analysis involves constructing a multiple alignment, determining the substitution model, building the tree and evaluating the result.

1.3.1 Model species

Model species are a selected set of organisms chosen to represent a wider group of organisms. Models tend to be the primary focus of basic research due to relative ease of experimental manipulation, with the assumption and expectation that information gained in the model can be transferred to other closely related organisms. Nearly all model organisms have a completely sequenced genome that is highly accurate and fully annotated. The usual approach is to define a model species, fully characterise its genome and use it to investigate species with more complex genomes that are less well defined. This approach relies on the high degree of conservation observed between related genomes. Using a model, one can infer structure and function to a more complex or intractable genome, investigate genomic rearrangements that have taken place and define evolutionary events and relationships between the species. If the evolutionary distance between the model and the genome of interest is relatively small the model can be used to define functional coding regions within the genome, aid in defining protein-coding genes, in recognising conserved non-coding regions, and in finding regulatory sequences and other functional elements of that genome. An important model in the Human Genome Project was the mouse, whose genome was sequenced in order to assist the annotation of the human genome (Waterston *et al.* 2002). Additional model organisms such as *E. coli*, *S. cerevisiae*, *C. elegans* and *D. melanogaster* were used in this annotation. On a genetic level, mice and humans are very similar. Mice also have the advantages of rapid reproduction, short life spans and can be genetically manipulated by the insertion of foreign genetic material into mouse DNA to assess the function of genes.

The original plant model system, *A. thaliana* was widely adopted in the 1980s initially due to its small size, short generation time, fecundity and small genome and later due to experimentally derived properties such as ease of mutagenesis and transformation. The widespread use of *Arabidopsis* as a plant genetic model lead to its genome being sequencing at the turn of the century (The Arabidopsis Genome Initiative 2000).

1.3.2 Comparative mapping and synteny

Comparative mapping refers to the alignment of chromosomes of related species based on genetic mapping of common DNA markers or genes. These aligned maps enable comparisons to be made between the positions and order of the markers or genes. The precise genetic definition of synteny is the property of being on the same chromosome but in comparative genomics this term has been somewhat generalised to refer to the degree of similarity in the order of markers on genetic, physical or sequence-based maps. A conservation of synteny is observed between closely related species and this conservation becomes less evident as the evolutionary distance between the species increases. A more accurate term to describe the similarity of gene order between chromosomal segments is collinearity. Collinear is derived from the root collimate meaning aligned, in contrast to co-linear meaning in the same straight line. If genes or markers are found in the same order on two chromosome segments they are said to be collinear, or described as showing a high level of synteny. These definitions are also applied to comparisons at the DNA sequence level and referred to as microsynteny or microcollinearity as opposed to macrosynteny or macrocollinearity which is used at the gene or marker level.

1.3.3 Identification of functional regions

The Neutral Theory of molecular evolution is the theory on which the categorisation of functional and non-functional regions of DNA is based. This theory was postulated before the advent of DNA sequencing methods (Kimura 1969) and recognised that negative (or purifying) selection is a potent force in the evolution of genes (Kimura 1983). Negative selection is the removal of deleterious mutations from a population and the theory states that this type of selection is more common than positive selection, in which mutations are retained (Kimura and Ota 1974). It predicted that the rate of nonsynonymous substitution (where a nucleotide substitution would alter the resulting amino acid) would be lower than the rate of synonymous substitution (where a substitution would not affect the amino acid). This is interpreted to mean that synonymous mutations are effectively neutral because they do not change the resulting amino acid and that most nonsynonymous mutations are eliminated over time by natural selection. The Neutral Theory also predicted that functionally important gene regions should evolve more slowly than non-functional regions and that duplicated genes should differ in their evolutionary rate.

The ratio of nonsynonymous (k_A) to synonymous (k_S) substitutions provides a measure of a region of the genome being under selection. Pairwise comparisons between most homologous protein coding sequences give k_A/k_S of less than 1. This is explained in Neutral Theory by the fact that most nonsynonymous substitutions are deleterious so are not often retained relative to synonymous substitutions and indicates that these sequences are functional. If $k_A/k_S = 1$ then replacement of an amino acid has no effect on the protein and both types of substitution are neutral. If nonsynonymous substitutions are retained, it must be evolutionarily advantageous for the gene to do so, meaning the protein is under diversifying selection ($k_A/k_S > 1$). A comparative study of the closely related soil nermatodes *C. briggsae* and *C. elegans* used this technique to explore the extent of conservation existing between these two species (Stein *et al.* 2003). Differing rates of synonymous substitution are often observed between different lineages such the twofold increase in mouse compared to human (Waterston *et al.* 2002). It is postulated that this could be due to the shorter life span of mice resulting in many more generations along that lineage.

Neutral Theory also introduced the idea of using amino acid replacements as a ‘molecular clock’ to estimate divergence times between different taxa. This idea assumes that as synonymous substitutions are largely free from selection they will accumulate changes in a neutral manner. Therefore, the number of synonymous substitutions (k_S) between two homologous sequences will increase linearly with divergence time and can be used to temporally order the sequences.

The timings of duplications within the *Arabidopsis* genome have been characterised by the molecular clock technique (Blanc *et al.* 2003) indicating that duplicated blocks could be grouped into two main families, one arising from an ancient duplication event and one that occurred more recently.

1.3.4 Understanding genome evolution

Comparing genomes by nucleotide similarity results in a description of matching regions and areas of mismatch. Comparative studies between the mouse and the human genome have shown that a large proportion of the gene order has been conserved in the 75-80 million years since divergence (Waterston *et al.* 2002). It is estimated that about 5% of the mammalian genome is under negative selection, which is about three times larger than the protein-coding portion of the genome. These additional regions include noncoding RNA genes, regulatory sequences and other functional regions which are not protein-coding. In addition, extensive conservation of protein-coding regions is observed including conservation of intron-exon structures, with nearly all the genes in the human genome aligning with homologues in mouse. This conservation extends to the nucleotide level where about 40% of the human genome aligns with the mouse genome. The remainder is composed of insertions, deletions and other genomic rearrangements that have occurred in both lineages since divergence of the two species. Many of the orthologous regions that do not align may have diversified from the ancestral sequence to such an extent that similarity is no longer detectable. As more genome sequences become available from species at differing phylogenetic distances, lineage-specific changes can be characterised across a wide taxonomic spectrum and alignment methods improved for more accurate detection of homologues.

Plant genome evolution is a vibrant field of research. The diverse group of plants classified as angiosperms or flowering plants are divided into monocotyledons and dicotyledons (Cronquist 1988) which are distinguished by various morphological differences. Monocots and dicots are estimated to have diverged approximately 150 million years ago (Chaw *et al.* 2004). Comparative studies of the genomes of *A. thaliana* (the dicot model) and rice (the monocot model) have revealed variable conservation of collinearity (Devos *et al.* 1999; Liu *et al.* 2001; Mayer *et al.* 2001). A possible explanation for this is the occurrence of genome duplication events in *Arabidopsis* (Vision *et al.* 2000) and the subsequent loss of many duplicated genes (The Arabidopsis Genome Initiative 2000), complicating patterns of conservation. Genome duplication events are a major force in the evolution of plant genomes providing redundant gene copies free from selective constraints that can undergo mutation, resulting either in loss of function or the evolution of new and diverse functions. Plant genomes display high levels of collinearity between closely related taxa (Moore *et al.* 1995). Rice is a member of the grass family and has been used to investigate closely related species such as maize and wheat (Chantret *et al.* 2004; La Rota and Sorrells 2004; Lai *et al.* 2004a). The conservation of gene order between grass species has lead to its adoption as a key system for plant comparative genomics studies.

1.3.5 Sequence alignment

The application of computational techniques to align nucleotide and protein sequences is at the core of comparative genomics. The basis of most comparative genomic studies is pairwise alignment, mapping of the nucleotides in one sequence onto the nucleotides in another sequence with gaps introduced into one or other of the sequences to increase the number of matches. The global pairwise alignment algorithm (Needleman and Wunsch 1970) takes two sequences and aligns them optimally over their entire length using a dynamic programming algorithm. This method is effective if the two sequences are closely related and of similar length. Local pairwise alignment (Smith and Waterman 1981) aligns similar regions of two sequences which is often more useful than the global approach since areas of local similarity may have biological significance. Both these methods are guaranteed to find the optimum alignment between two sequences and are useful when relatively short sequences are being compared.

The alignment of long DNA sequences using the optimal global and local alignment algorithms is computationally intensive. Heuristic algorithms, i.e. algorithms that use trial and error to approximate a solution for computationally difficult problems, have been developed to improve the speed of similarity matching. These algorithms do not compromise the sensitivity of the matching, although the optimal solution is not guaranteed. The most popular of these is the Basic Local Alignment Search Tool or BLAST (Altschul *et al.* 1990) which detects all possible regions of local alignment between two sequences and scores each of the matches for biological significance. The BLAST algorithm is available on many web servers and is used to search DNA and protein sequence repositories for sequences that have significant matches to a query sequence. Different versions of BLAST are available depending on the type of search required; BLASTN to search a nucleotide database with a nucleotide query sequence and BLASTP to search a protein database using a protein query sequence. The FASTA algorithm (Lipman and Pearson 1985) was the first heuristic method designed to perform database similarity searching in addition to assessing the significance of each alignment. In order to score similarity matches for statistical significance, both BLAST and FASTA use scoring matrices. These matrices give scores for residue matches and are constructed in such a way as to reflect actual biological observations of conservation of residues, frequency of occurrence and evolutionary patterns. Modification of parameters such as gap and gap extension penalties can also be used to fine-tune these alignment algorithms.

For aligning long genomic sequences, faster and more efficient methods of alignment have been developed. A product of the human genome project, BLASTZ (Schwartz *et al.* 2003b) is an algorithm that is particularly useful for aligning similar regions of two genomic sequences and was used to find homology between portions of the human and mouse genomes. This algorithm uses a 19 position match-mismatch pattern to assess similarity and performs local alignments. Other genomic-scale global alignment algorithms include LAGAN (Brudno *et al.* 2003) and AVID (Bray *et al.* 2003). These methods define anchors of best alignment within the sequences before performing a more detailed alignment on small regions. A key stage in many of these algorithms is the conversion of sequence data into a data structure that can be efficiently searched. MUMmer (Kurtz *et al.* 2004) uses a suffix tree to achieve high-speed alignment and SSAHA (Ning *et al.* 2001) uses a hash table data structure. Bounded sparse dynamic programming has also been applied to the sequence alignment problem to further increase speed and accuracy of alignments and is implemented in Exonerate (Slater and Birney 2005).

In additional to pairwise alignments, many comparative genomic studies require alignment of multiple sequences in order to determine features that are common to distantly related species. A software package called ClustalW allows this by first computing a guide tree for the sequences by comparing each of them with the other sequences (Thompson *et al.* 1994). The guide tree groups similar sequences together and this grouping is used to build up the alignment starting with the pair of sequences that are most similar. This algorithm also allows sequences that were used early on in the alignment process to be realigned against the multiple alignments in an iterative fashion. The T-Coffee algorithm builds a library of pairwise alignments between all the sequences then uses the library to find a multiple alignment that tries to preserve the pairwise alignments (Notredame *et al.* 2000). T-Coffee produces very good alignments but is only suitable for short sequences as it is much slower than other methods. A similarly accurate but much faster algorithm for multiple sequence alignment is implemented in MUSCLE (Edgar 2004). In addition to tools designed specifically for multiple sequence alignment, many of the pairwise genomic alignment algorithms described previously have been extended to align multiple sequences, for example Multi-LAGAN (Brudno *et al.* 2003) and M-AVID (Bray and Pachter 2004).

1.3.6 Visualisation in comparative genomics

The ability to visualise aligned genomes is a key requirement in many aspects of comparative genomics and many tools are available which provide global and local views of genomic alignments. The dot-plot is a powerful way to visualise pairwise whole-genome alignments using tools such as DAGChainer (Haas *et al.* 2004) and MUMmer (Kurtz *et al.* 2004). The two genomes being compared are arranged on each axis of a rectangular array and a dot is placed in the grid at all the points where the sequences are identical. A more sophisticated approach uses a sliding window of a particular size where a number of bases in each sequence are compared to generate one dot. A 'mismatch limit' defines how many mismatches are allowed before the sequences within the window are not considered similar. Using a dot-plot, areas of local alignment, repeats, inversions and deletions can be easily identified. An alternative representation of a whole-genome comparison consists of one genome shown as an ideogram, with blocks of colour defining synteny with another genome (Figure 1.1). This is the approach used by software such as Cinteny (Sinha and Meller 2007).

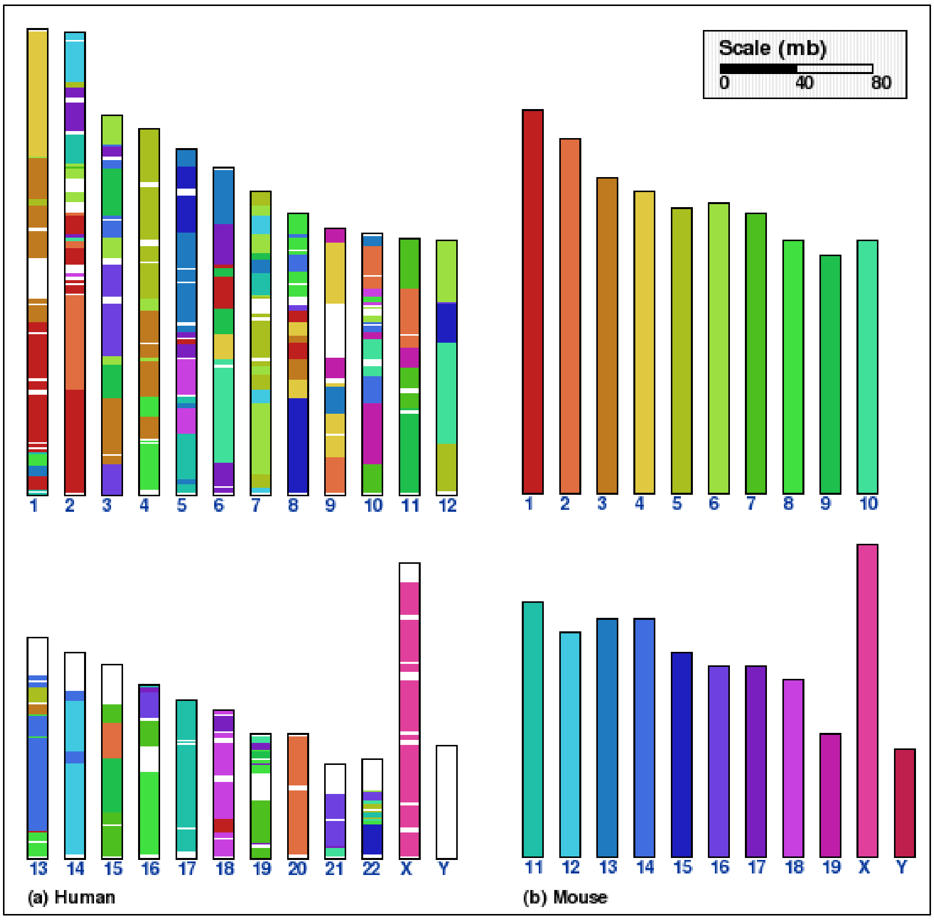


Figure 1.1: Whole genome comparison between human and mouse.

Generated by the Cinteny software and obtained from http://cinteny.cchmc.org

A recent alternative to the dot-plot and ideogram views are the highly customisable images produced by the Circos package (Krzywinski *et al.* 2009). Chromosomes from single or multiple genomes are represented as arcs forming a circle. These arcs can be coloured to indicate properties of each chromosome and annotation tracks added within or without the circle. Syntenic regions between chromosomes are represented by lines that cross the centre of the circle thus reducing the clutter usually seen when representing synteny between multiple genomes as lines joining a stack of linear representations.

In order to compare aligned genomic regions in more detail, different visualisation tools are required. Most genome browsers allow the representation of sequence conservation between a reference genome and a related genome as an annotation track in the form of a histogram. An example is the Java-based VISTA browser which displays pre-computed alignments between the genomes of many species (Frazer *et al.* 2004). These histograms are used to identify functionally conserved regions between the genomes. Where no pre-computed alignments are available to assess conservation between genomic sequences, tools such as PipMaker (Elnitski et al., 2002) or mVISTA (Dubchak and Ryaboy 2006) are used. These tools use the genomic alignment algorithms BLASTZ and AVID respectively to produce graphical representations of percent identity throughout the sequences being compared. mVISTA allows multiple sequences to be aligned as does a variant of PipMaker, called MultiPipMaker (Schwartz *et al.* 2003a).

To visualise the alignment of multiple pairwise comparisons, tools such as Artemis Comparison Tool or ACT (Carver *et al.* 2005) and CMap (Youens-Clark *et al.* 2009) are available. On the simplest level, ACT takes as input two sequence files and a comparison file obtained from BLAST. The software displays the high-scoring pairs (HSPs) identified between the sequences as a series of lines linking the two sequences and the user can move over the length of the sequences and alter the level of detail being displayed. Additional sequences and comparison files can be loaded with annotations displayed alongside the comparisons if required. CMap is used to display genomic maps (either genetic, physical or sequence-based) as a linear array of annotated features and multiple maps can be added to the display with common features in each pairwise comparison linked by lines.

1.4 Composition and dynamics of plant genomes

1.4.1 Introduction

Plant genomes are among the largest and most complex of the eukaryotes. Even within quite closely related groups, a huge variability in DNA content is observed (Bennett and Leitch 1995). The amount of DNA contained within a nucleus of a gamete is called the c-value, measured in picograms (pg) or megabase pairs (Mbp). DNA content is estimated using the Feulgen densitometry technique (Rasch 1985) which involves staining tissue preparations with a dye that specifically binds DNA. The amount of stain bound to the nucleus is proportional to the amount of DNA present and this is determined by the amount of light absorbed, measured using computer-based image analysis software (Vilhar *et al.* 2001). The absolute genome size is calculated by comparison with species of known DNA content, for example sequenced genomes. These sequence-based values may not be entirely accurate in plant studies because a fully sequenced genome usually does not contain sequence data for the highly repetitive regions. C-values within the plant kingdom range from 10 Mbp for the algae *Ostreococcus tauri* to 124,852 Mbp for the fritillary *Fritillaria assyriaca* (Bennett and Leitch 2005). The observation that the amount of DNA contained within a cell nucleus seems to bear no correlation to the complexity of the organism was termed the c-value paradox (Thomas 1971). For example, many plant species contain more DNA per cell nucleus than is found in human cells. It is now known that that the complexity of an organism is not dependent on the size of its genome but on the number of genes that are encoded by the genome as well as the regulatory pathways that control the expression of these genes. In addition to genes, a high percentage of eukaryotic DNA is non-coding and the amount of coding to non-coding DNA varies greatly between species, accounting for the genomic size difference. The chromosome complement of species within the *Planteae* kingdom is also highly variable as well as the size of chromosomes. Furthermore, many plants exhibit polyploidy where the cell nucleus contains more than two paired sets of chromosomes. Polyploidy and the subsequent genome dynamics bought about by this change contributes to the tremendous heterogeneity in the size and complexity of plant genomes (Feldman and Levy 2009).

The different regions of plant genomic DNA can be classified by their frequency of occurrence within the genome, as determined by reassociation kinetics (Britten *et al.* 1974). This technique involves shearing genomic DNA into short lengths followed by heating to denature the double-stranded DNA. As the solution cools, the DNA renatures and the amount of double stranded DNA is measured. Three components are observed that reassociate at different speeds: these are highly repetitive, intermediately repetitive and low-copy regions. Regions that are unique or that exist in low-copy numbers are the actively transcribed genic regions called euchromatin. Intermediately repetitive DNA is found throughout the genome and can represent tandemly repeated genes such as ribosomal RNA (rRNA) and transfer RNA (tRNA) genes or genes present in multiple copies, such as those that encode the storage protein zein in maize (Song *et al.* 2001). Highly repetitive regions are low in gene content and called heterochromatin. Highly repetitive DNA is found in intergenic regions, at the termini of chromosomes (the telomeres), and at the central pairing site involved in mitosis and meiosis (the centromere). Centromeric and telomeric chromosomal regions are not generally sequenced thoroughly due to their highly repetitive nature although these regions are beginning to be characterised in plants (Wu *et al.* 2009).

The GC-content of a DNA sequence is the amount of guanine-cytosine base pairs in the sequence and a wide variation of GC content is observed in different regions of genomic DNA. This pairing forms an extra hydrogen bond compared to the two that are observed between adenine-thymine pairs and provides a measure of the stability of the double-stranded DNA. In general, plant genomes exhibit a higher GC-content in gene-rich regions than gene-poor areas, with exons having a particularly high GC-content compared to introns (Montero *et al.* 1990).

1.4.2 Transposable elements

The complete genome sequence of the model dicot *Arabidopsis thaliana* (The Arabidopsis Genome Initiative 2000) provided the first opportunity to characterise the repetitive sequence component of a complete plant genome (Le *et al.* 2000); however the repetitive content of this small genome is relatively low. Historically, repetitive regions have been studied more extensively in larger cereal genomes such as maize (SanMiguel *et al.* 1998; Meyers *et al.* 2001; Messing *et al.* 2004), wheat (Wicker *et al.* 2003a; Li *et al.* 2004; Sabot and Schulman 2006) and barley (Kalendar *et al.* 2004; Schulman and Kalendar 2005; Wicker *et al.* 2005) where repetitive regions account for a much larger percentage of the genome. Plants with small genomes such as *Arabidopsis* contain about 20 % repetitive DNA, found mainly in heterochromatic regions and plants with larger genomes such as maize and wheat contain much greater proportions.

The majority of repetitive regions in plant genomes are due to transposable elements first identified in maize in the 1940s and since found throughout the genomes of all prokaryotes and eukaryotes. Transposable elements, or transposons, are able to move from one location to another in a genome and are classified depending on their mode of transmission (Finnegan 1989; Flavell *et al.* 1994). Class I transposons (retrotransposons) replicate throughout the genome by means of an mRNA intermediate and therefore increase their copy number after each replication. Retrotransposons are further classified into those that contain long terminal repeats (LTRs) such as Ty1-*copia* and Ty3-*gypsy* and non-LTR retrotransposons such as long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs). LTR-retrotransposons encode a reverse transcriptase to convert the mRNA intermediate to DNA, and an integrase to incorporate the DNA into the chromosome. The sub-classes *copia* and *gypsy* differ in the order of these elements within the encoded protein. LINEs are several kilobases long and encode proteins for retrotransposition without using an integrase. SINEs are shorter structures and do not encode their own reverse transcriptase, instead relying on proteins encoded by other transposable elements for transposition. LTR-retrotransposons are the single largest component of most plant genomes and are located largely in intergenic regions (Kumar and Bennetzen 1999).

Class II transposons (DNA transposons) consist of a simple DNA sequence which transposes throughout the genome without increasing in number using a mechanism by which the transposon is excised from one place in the genome and inserted elsewhere. An abundant repeat of this type found in plants is the miniature inverted-repeat transposable element (MITE) which comprises of almost identical sequences of about 400 base pairs flanked by characteristic inverted repeats of around 15 base pairs. MITEs were first discovered in non-coding regions of grass genes (Bureau and Wessler 1992) and have since been found in the gene-rich regions of sequenced plant genomes (The Arabidopsis Genome Initiative 2000; International Rice Genome Sequencing Project 2005). MITEs are classified as *Tourist*-like MITEs or *Stowaway*-like MITEs by the similarity of their inverted repeats. Two particularly important DNA transposon families in maize are the *Activator* (*Ac*) and *Mutator* (*Mu*) elements and these are utilised in large-scale mutagenesis experiments. The *Activator* element and its deletion derivative *Dissociation* (*Ds*) are used to isolate genes in plants, by a technique known as transposon tagging (Fedoroff *et al.* 1984). The first step in this procedure is to identify a mutant for a specific trait which is caused by the insertion of a transposable element into a gene and its subsequent inactivation. A genomic library of this mutant is created and screened for the transposable element, resulting in a number of clones that contain the element. The sequences flanking the inserted element will be from the gene of interest and probes made from these sequences are then used to screen a genomic library containing DNA from a normal plant. The clones selected from this library will contain a normal copy of the gene. The *Mutator* element is the most active plant transposon discovered to date and has been used in cloning, sequencing and mutating many genes in maize (Walbot 2000). Similarly useful transposons have been found in the sorghum genome (Chopra *et al.* 1999).

In recent years, a new type of class II transposon has been described and computationally identified in *A. thaliana*, *O. sativa* and *C. elegans* (Kapitonov and Jurka 2001). Called a *helitron*, they are believed to transpose via rolling circle replication, similar to some known prokaryotic rolling circle transposons. *Helitrons* are able to acquire fragments of genes from different locations in the genome and these fragments are moved around the genome within the transposon. *Helitrons* carrying gene fragments have been described in the maize genome and investigations have revealed that some of these gene fragments are transcribed, generating transcripts containing portions of different genes (Brunner *et al.* 2005; Lai *et al.* 2005; Morgante *et al.* 2005). Other types of DNA transposon have been shown to acquire multiple fragments of genes, for example *mutator*-like transposable elements (MULEs) called Pack-MULEs have been described in rice (Jiang *et al.* 2004; Juretic *et al.* 2005) and a small proportion of those identified appear to be transcribed. In addition, a member of the CACTA superfamily that carries multiple gene fragments has been identified in soybean (Zabala and Vodkin 2005). In these cases, the mechanism of duplication preserves the intron-exon structure of the gene fragments. A unified classification system for eukaryotic transposable elements has recently been proposed (Wicker *et al.* 2007).

1.4.3 Gene content

The genome of *A. thaliana* was initially annotated with 25,498 genes (The Arabidopsis Genome Initiative 2000) and predictions of gene function suggested these genes were grouped into 11,601 gene families with differing numbers of gene copies in each family. Initial annotations of the draft rice genome predicted 40,000 to 60,000 genes (Feng *et al.* 2002; Goff *et al.* 2002; Sasaki *et al.* 2002; Yu *et al.* 2002; Rice Chromosome 10 Sequencing Consortium 2003). These early studies also indicated that half the genes in rice had no homologues in *Arabidopsis* (Goff *et al.* 2002; Yu *et al.* 2002) and that gene content varied even between subspecies (Feng *et al.* 2002). It is now known that the predictions in rice were excessive due to the misidentification of genes and a failure to account for gene duplication. Annotation of the finished rice genome indicated the presence of approximately 37,000 genes but only 61% of these predictions were supported by EST or cDNA data (International Rice Genome Sequencing Project 2005). Of the supported predictions, 88% have homologues in *Arabidopsis*.

Misidentification of genes in DNA sequences has been ascribed to the presence of gene fragments and pseudogenes. Gene fragments are carried around the genome by transposable elements such as *Helitrons*, Pack-MULEs and CACTA elements, as previously described. In 2004, Bennetzen and colleagues reassessed the estimate of gene content in rice and suggested the figure was less than 40,000 (Bennetzen *et al.* 2004). The authors identified what they termed "the five fantasies of gene prediction" in an attempt to characterise the problems associated with gene discovery and to outline ways in which these problems could be minimised. Pseudogenes were originally identified in *Xenopus laevis* (Jacq *et al.* 1977) and have since been identified in plants (Loguercio and Wilkins 1998). Pseudogenes were believed to be non-functional sequences of genomic DNA originally derived from functional genes where an accumulation of mutations such as premature stop codons and frameshift mutations prevents their expression. They can arise either from reverse transcription of mRNA and reintegration into the genomic DNA (Vanin 1985) or by gene duplication. Once present in genomic DNA the pseudogene accumulates mutations causing loss of function (Mighell *et al.* 2000). Further investigation into pseudogenes has shown that the boundary between actual genes and pseudogenes is less well defined. In eukaryotes, the observation that pseudogenes exhibit a higher degree of conservation to related functional genes than would be expected by a non-functional region indicates a possible functional role for these sequences (Harrison and Gerstein 2002). More recent evidence indicates that some pseudogenes identified in the genomes of *Arabidopsis* and rice are expressed at low levels (Thibaud-Nissen *et al.* 2009; Zou *et al.* 2009).

Since the initial analysis of the *Arabidopsis* and rice genomes, the genomes of many more plants have been sequenced and annotated (Table 1.1) indicating a plant gene complement of between 20,000 and 60,000 genes.

Name	**Common name**	**Number of genes**	**Release information**
Arabidopsis thaliana	Mustard cress	33,410	TAIR 9
Populus trichocarpa	Black cottonwood	41,377	JGI v2.0 annotation of assembly v2
Vitis vinifera	Wine grape	30,434	September 2007 release
Carica papaya	Papaya	28,629	Ming *et al* (2008)
Cucumis sativus	Cucumber	21,491	v1.0
Medicago truncatula	Barrel medic	50,962	Mt3.0
Sorghum bicolor	Sorghum	34,496	v1.0
Zea mays	Maize	32,540	Filtered gene set from annotation of the 4a.53 assembly
Oryza sativa	Rice	31,232	RAP-DB v5

Table 1.1: Predicted gene complements of sequenced plant genomes.

Early genetic studies of wheat suggested that genes are clustered into islands and interspersed with long sequences of repeated regions (Gill *et al.* 1996) with larger genomes having longer repeated regions (Chen *et al.* 1997). Similar observations were made in maize (SanMiguel *et al.* 1996). More detailed studies have since been conducted on sequenced genomic data. A semi-automated annotation of 100 random maize BAC clones representing 0.6% of the maize genome, identified 330 genes with an average length of 4 kb (Haberer *et al.* 2005) compared to 2.6 kb and 2 kb in rice and *Arabidopsis* respectively (The Arabidopsis Genome Initiative 2000; International Rice Genome Sequencing Project 2005). The average exon lengths are similar in rice and maize, both slightly longer than experimentally confirmed exons in *Arabidopsis*. However, the introns in maize are longer compared to rice which accounts for the larger gene length in maize. This study also found that genes in the maize BAC clones were not grouped in islands but that single genes are interspersed with repetitive elements. Gene density in the maize BAC clones is calculated to be one gene in 43.5 kb, compared to one gene in 9.9 kb and one gene in 4 kb for rice and *Arabidopsis* respectively which correlates with the increasing genome size between these species.

BAC sequencing in wheat initially concentrated on gene-rich regions (Wicker *et al.* 2001; Brooks *et al.* 2002; SanMiguel *et al.* 2002; Faris *et al.* 2003; Yan *et al.* 2003; Gu *et al.* 2004). These studies produced gene density estimates of one gene per 5-20 kb but a small investigation of four random BAC clones produced an estimation of one gene per 75 kb which suggested genes were much more widely distributed in wheat (Devos *et al.* 2005). Fluctuations in gene density over different regions of plant chromosomes are also observed, for example in rice, gene-rich regions are found near the telomeres with the centromeres consisting of gene-poor, repeat-rich regions (International Rice Genome Sequencing Project 2005).

1.4.4 Organelle genomes

Plant cells contain mitochondria, nearly ubiquitous in eukaryotes, and chloroplasts (or plastids), unique to plants. Both these organelles are endosymbionts descended from prokaryotic origins and both contain genomes separate from the nuclear genome of the plant. Chloroplast genomes consist of a circular, double-stranded chromosome and due to their small size they have been characterised in many plant species. Chloroplast genomes are densely packed and contain approximately 110 genes which are highly conserved in sequence and organisation across all plant species (Sugiura 1995). Mitochondrial genomes are also circular but unlike chloroplast genomes vary greatly in size between plant species (Levings and Brown 1989). The mitochondrial genomes of plants are generally much larger than other eukaryotes and the presence of large repeated sequences makes these genomes very dynamic and difficult to analyse (Palmer and Shields 1984). In *Arabidopsis*, 57 genes were identified in the 366,924 nucleotides of the mitochondrial genome covering only 10% of the genome (Unseld *et al.* 1997).

Genetic material is often transferred from the organelles to the nucleus resulting in plastid and mitochondrial DNA insertions into nuclear DNA and a corresponding reduction in the size of the organellar genome (International Rice Genome Sequencing Project 2005). Nuclear DNA insertions originating from the plastid genome are called NUPTs (<u>nu</u>clear <u>pt</u>DNA) and insertions originating from the mitochondrial genome are called NUMTs (<u>nu</u>clear <u>mt</u>DNA), both types being common in flowering plants (Richly and Leister 2004b; Richly and Leister 2004a). NUMTs and NUPTs are found in clusters which retain their homology to the original sequence to varying degrees indicating that initial insertions can be large but decay over time by processes such as transposon insertion (Noutsos *et al.* 2005). These initial insertions are thought to arise by the incorporation of DNA fragments that have “escaped” from the mitochondria or plastid by processes such as nonhomologous recombination or the repair of double stranded breaks in nuclear DNA (Timmis *et al.* 2004). Investigations into tobacco plants have also shown that genetic transfer from the organelles to the nuclear genome is an ongoing process (Huang *et al.* 2003). In *Arabidopsis* and rice, 25% of NUPT and NUMT insertions occur preferentially in gene-poor regions (Richly and Leister 2004b).

1.4.5 Genome dynamics

The smallest modification a genome can be subjected to is a point-mutation of a single nucleotide, a transition (where a purine is changed to a different purine or a pyrimidine is changed to a different pyrimidine) or less commonly, a transversion (where a purine is substituted for a pyrimidine or a pyrimidine is substituted for a purine). In addition, modification can occur by the insertion or deletion of a single nucleotide or longer sequence into a genome (an indel). Larger mutational events where chromosomal rearrangement is observed can be divided into those which modify the size of the genome and those that move genetic material around the genome with no overall change in size. Examples of the latter type of rearrangement are transposition, translocation and inversion. Transposition occurs when a segment of a chromosome moves to another location in the genome, called a shift when the segment is inserted into the same chromosome and an insertion when incorporated into a different chromosome. Translocation is the swapping over of two segments of different chromosomes and inversion is when a segment of a chromosome is removed, inverted and reinserted into the same place.

Genetic recombination is a general term which refers to a DNA molecule being broken and joined to a different one. Genetic recombination can occur between similar molecules (homologous recombination) or dissimilar molecules (nonhomologous recombination). Homologous recombination occurs during the first phase of meiosis where homologous chromosomes pair to form bivalents by a process called synapsis. Each bivalent consists of four chromatids; two sister chromatids from each chromosome. The close proximity of homologous chromosomes allows genetic material to be exchanged between the chromatids. Homologous recombination, or crossing over, was first postulated by Morgan (Morgan 1911) and provided the basis for constructing the first genetic linkage map, but it was not until 1931 that crossing over was experimentally demonstrated in maize (Creighton and McClintock 1931). In 1964, a model was proposed to explain homologous recombination initiated by a single-strand break in the DNA followed by the formation and cleavage of Holliday junctions (Holliday 1964). More recent models have been proposed to explain observations not accounted for by the Holliday model such as the double-strand break repair model (Szostak *et al.* 1983). In most cases, homologous recombination involves the reciprocal exchange of near-identical chromatid segments and after meiosis is complete, four haploid gametes are produced containing chromosomes that contain a combination of alleles from the parental cells. Homologous recombination can also involve non-reciprocal exchange where homologous regions misalign by virtue of multiple regions of high sequence identity, for example in tandem arrays. In these cases, paired chromosomes break at slightly different loci during crossing over resulting in a duplication of genes on one chromosome and a deletion of genes on the other. Homologous recombination can also give rise to gene conversion where a homologous region is used to repair a double stranded break in DNA. If the homologous region and the region being repaired contain a base mismatch, the repair mechanism may convert one of the two mismatched bases. If gene conversion occurs the resulting haploid cells will contain alleles in non-mendelian ratios. Both recombination and gene conversion contribute to an increase in genetic diversity by giving rise to offspring with a different combination of alleles to its parents.

There are three distinct mechanisms of genome amplification; polyploidy, tandem amplification and segmental duplication. Polyploidy arises by the doubling of the chromosome set of a single individual (autopolyploidy) or the combination of the genomes from two closely related species (allopolyploidy) and results in multiple sets of chromosomes in a cell nucleus. Bread wheat is an example of an allopolyploid, being formed by the hybridisation of three diploid progenitors. Polyploidy is a prominent force in the evolution of plants (Soltis *et al.* 2004) and it is estimated that most plants have undergone multiple polyploidisation events, even those with relatively small genomes such as *Arabidopsis* (Blanc *et al.* 2003). A change in ploidy level initiates a range of transformations within the genome including gene silencing, gene loss (Kashkush *et al.* 2002), and diploidisation (Tian *et al.* 2005).

Tandem amplification is the repeated duplication of a portion of the genome, generally a single gene, which results in long stretches of repeating sequences called tandem arrays. In *Arabidopsis*, around 17% of genes are arranged in tandem arrays (The Arabidopsis Genome Initiative 2000) and analysis of the rice genome provided a similar estimate (International Rice Genome Sequencing Project 2005). Ribosomal RNA genes are found in long tandem repeats with copy numbers that are highly polymorphic between species.

Segmental duplication is a process causing a large portion of one chromosome to be duplicated and moved to a new area in the genome, either on the same or on a different chromosome. These duplicated regions can arise as a result of whole genome duplication events followed by subsequent gene loss as well as many cycles of chromosome breakage and fusion rearrangements (Simillion *et al.* 2002; Blanc *et al.* 2003; Bowers *et al.* 2003b; Ermolaeva *et al.*

2003). The complete genome sequence of *Arabidopsis* indicated that tandem gene duplication and segmental duplication is much more prevalent in plants than in other model organisms, possibly indicating more relaxed constraints on genome size in plants (The Arabidopsis Genome Initiative 2000). A phylogenetic analysis of 50 large *Arabidopsis* gene families enabled the segmental and tandem duplication history of the genome to be characterised (Cannon *et al.* 2004).

Processes such as polyploidy, tandem amplification and segmental duplication increase the size of the genome. These expansion processes are counteracted by deletion mechanisms that can remove large parts of a chromosome and it is the dynamic balance between these two forces over evolutionary time that determines the size of the genome. Similar mechanisms of expansion and contraction act on the genomes of different lineages and the observed differences in size and structure are a result of differing levels of action of each process (Vitte and Bennetzen 2006). Polyploidy is by far the biggest factor in the expansion of genome size but the amplification of LTR-retrotransposons also causes genome expansion. A study of the rice lineages *indica* and *japonica* indicated an increase in the genome size of both species since their divergence from a common ancestor, largely due to LTR-retrotransposon amplification (Ma *et al.* 2004). Moreover, the *japonica* lineage had undergone a greater expansion than its sister lineage. A more extensive study of the ten genome types within the genus *Oryza* also found that the number of LTR-retrotransposons increased linearly with genome size (Ammiraju *et al.* 2006). This supported previous observations from the comparison of an orthologous region in maize and sorghum which showed that a large percentage of the maize region consisted of retrotransposons that were not present in sorghum (Tikhonov *et al.* 1999). Characterisation of retrotransposon structures in maize and wheat (SanMiguel *et al.* 1996; Wicker *et al.* 2001) have found multiple levels of nested elements corresponding to different waves of invasion.

Reductions in genome size are predominantly due to deletion of large segments of DNA by unequal homologous recombination involving LTR-retrotransposon sequences. An excess of LTR flanking regions has been observed in wheat and barley relative to internal retrotransposon sequences suggesting that recombination can occur between the LTRs resulting in loss of the internal region (Shirasu *et al.* 2000; Wicker *et al.* 2001). An *Arabidopsis* study showed that this mechanism can occur within the same retroelement resulting in loss of the element, or between two elements resulting in the loss of the DNA sequence between the elements (Devos *et al.* 2002). Illegitimate or non-homologous recombination is also an important mechanism for removal of DNA from plant genomes (Devos *et al.* 2002; Wicker *et al.* 2003b) which is thought to arise during the repair of double-stranded breaks in genomic DNA (Puchta 2005). Studies on insect species have found that illegitimate recombination seems to remove DNA faster from species with small genomes compared to the rate of removal from larger genomes (Petrov *et al.* 2000), an observation that seems to extend to plants (Kirik *et al.* 2000).

Many of the processes of genome expansion create duplicate genes and this duplication is an important evolutionary mechanism in plants. The redundancy of duplicated genes is thought to be important for the development of new functions and processes and has been proposed to be a major contributor to the divergence of species (Ohno 1970). Once duplicated, paralogous gene copies are commonly silenced or deleted but occasionally, once freed from selective constraints, some of these genes may undergo a mutation causing them to diverge from their sister paralogue and develop new function (called neofunctionalisation) or further functional specificity (called subfunctionalisation). Functional divergence of duplicated genes in *Arabidopsis* has been studied in depth (Blanc and Wolfe 2004a), providing evidence that a high proportion of duplicated genes have undergone functional diversification after polyploidy and that some duplicated gene pairs have diverged in parallel, forming two new networks that are expressed in different cell types or under different environmental conditions.

1.5 The Grass family

1.5.1 Introduction

The family *Poaceae* is also known as *Gramineae* and comprises approximately 10,000 species classified into 600 to 700 genera (Clayton and Renvoize 1986; Watson and Dallwitz 1992). These species form part of the monocotyledonous group which includes all flowering plants presenting one cotyledon (or seed leaf) as seedlings. Species within the *Poaceae* are commonly referred to as grasses and are believed to have descended from a common ancestor that lived between 50 and 80 million years ago based on phylogenetic and fossil evidence (Crepet and Feldman 1991; Paterson *et al.* 2004b; Prasad *et al.* 2005). The family is broadly divided into two clades, named using acronyms of the subfamilies they contain. The BEP clade contains the subfamilies Bambusoideae, Ehrhartoideae and Pooideae and the PACCMAD clade contains the subfamilies Panicoideae, Arundinoideae, Chloridoideae, Centothecoideae, Micrairoideae, Aristidoideae, and Danthonioideae. The major genera in each clade are shown in Figure 1.2.

The grass family contains many important species. Agriculturally important crop species such as rice (*Oryza sativa*), wheat (*Triticum aestivum*), barley (*Hordeum vulgare*), maize (*Zea mays*) and sorghum (*Sorghum bicolor*) are members of this family and together, these crops provide a substantial proportion of calories consumed globally by humans. Forage grasses such as *Lolium* and *Festuca* are a major source of nutrition to grazing animals, including domesticated species, meaning that much of humankind's animal protein diet is also directly dependent on grasses. In addition, an emerging novel application of some grass species is that of feedstock for biomass energy production (DOE 2006), an example being switchgrass (*Panicum virgatum*); a large, fast growing species native to North America. Species within the grass family are ecologically dominant and are estimated to cover 20 % of the earth's land surface (Shantz 1954).

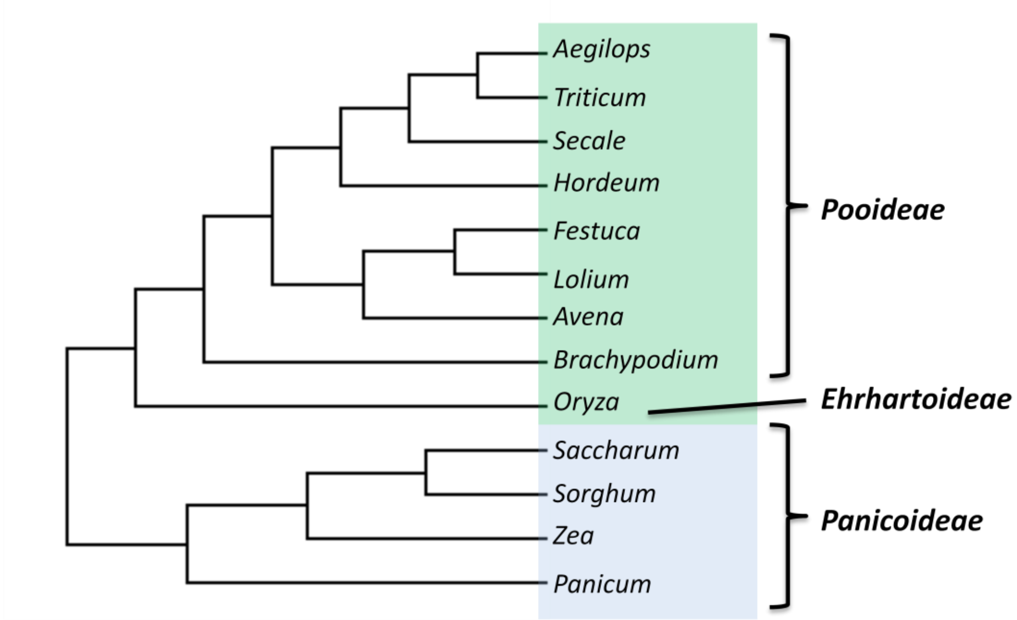


Figure 1.2: Cladogram showing the major genera in the grass family.
Genera in the BEP clade are shaded green, PACCMAD clade genera are shaded blue. Subfamilies are indicated on the right.

The crop species within the grass family have been the subject of much biological study, driven by the need to feed a global population that has more than doubled in size in the last fifty years and is predicted to grow to over 9 billion people by 2050 (Royal Society of London 2009). A major improvement in crop yields occurred in the years following the "green revolution" of the 1960s, when the development of high-yielding varieties of wheat, rice and maize, coupled to the use of fertilizers, resulted in a large increase in cereal production across the world. The population growth predicted to occur over the next thirty years will require an estimated 50-100 % increase in food production from todays levels (Royal Society of London 2009). Thus, research into the biological mechanisms underlying important agronomic traits such as disease and pest resistance, drought tolerance, and grain size, is more relevant than ever.

1.5.2 Grass genomics

Despite the relatively recent origins of the *Poaceae*, member species exhibit a huge diversity in chromosome number and genome size as well as a range of ploidy levels (Kellogg and

Bennetzen 2004). Rice is a diploid species with a relatively small genome of 466 Mbp and a haploid chromosome complement of 12. This is usually expressed as $2n = 2x = 24$ where n is the haploid number (the number of chromosomes in a gamete) and x is the monoploid number (the number of chromosomes in a single set). Bread wheat is hexaploid and has twenty-one chromosomes in each gamete, three sets of seven ($2n = 6x = 42$). This genome is estimated to be 17,000 Mbp in size (Bennett and Leitch 1995).

Its small genome size and agronomic importance prompted rice to be the first member of the grass family to have its genome sequenced. Cultivars of the two major subspecies *japonica* and *indica* were sequenced by separate research groups using different approaches. The International Rice Genome Sequencing Project (IRGSP) used a hierarchical clone-by-clone approach to produce a high-quality draft sequence of the *Nipponbare* cultivar of *japonica* in 2002 (Feng *et al.* 2002; Sasaki *et al.* 2002; Rice Chromosome 10 Sequencing Consortium 2003). The same cultivar was also sequenced using a whole-genome shotgun approach by the private company Syngenta resulting in a draft sequence (Goff *et al.* 2002). In 2005, the IRGSP published a finished map-based sequence of the *Nipponbare* cultivar (International Rice Genome Sequencing Project 2005). This finished sequence was derived from 3,401 BAC and PAC clones and was based on a genetic linkage map, EST sequences, two physical maps, BAC-end sequence data and incorporated the Syngenta WGS sequence as well as a draft sequence from Monsanto, another private company. The finished assembly contained 370.7 Mbp in 12 pseudo-molecules with an accuracy of one error per 10 kb. An initial draft of the *indica* cultivar was obtained using a whole-genome shotgun approach (Yu *et al.* 2002) and an improved sequence was published in 2005 alongside a reassembly of the WGS *japonica* sequence from Syngenta (Yu *et al.* 2005).

Initial analysis of the rice genome produced varying estimates of gene content, due to the different gene prediction methods used, although analyses of specific chromosomes had already revealed a proportion of these predictions were transposable elements (Sasaki *et al.* 2002; Rice Chromosome 10 Sequencing Consortium 2003). Assembly and annotation of the WGS *japonica* and *indica* sequences using the same method revealed a difference in the gene content between the subspecies and massive intergenic differences (Yu *et al.* 2005). A recent hand-curated annotation of the *japonica* sequence estimated a gene number of approximately 32,000 (Itoh *et al.* 2007). The two leading rice genome annotation projects, the Rice Annotation Project Database (RAP-DB) and the Michigan State University Rice Genome Annotation (MSU) estimate vastly different numbers of protein coding loci (31,439 compared to 56,797) but this is due to differing annotation methods. The RAP-DB annotation is largely manually curated and the automated annotation performed by MSU probably identifies many transposable elements or pseudogenes as real genes.

The annotated sequence of the rice genome provides a detailed view of gene content and distribution, areas of repetitive sequences, transposable elements and gene transfer from organelles (International Rice Genome Sequencing Project 2005). In addition, the rice genome has been used to develop markers for investigations into quantative trait loci (QTLs) underlying important agronomic traits such as disease resistance, plant architecture and flowering time (Yamamoto *et al.* 2009). Other genomic resources for rice include 1,249,110 rice ESTs in public databases (dbEST release 050710) and a collection of full-length rice cDNA sequences (Kikuchi *et al.* 2003). Responsibility for the functional characterisation of the genome using EST and cDNA libraries lies with the International Rice Functional Genomics Consortium (Hirochika *et al.* 2004). Efforts have also been made to categorise the SNPs that underlie genetic variation in both sub-species of rice (Feltus *et al.* 2004; Katagiri *et al.* 2004; Zhao *et al.* 2004) although different methods have generated differing estimates.

The last two years have seen the sequencing and publication of two additional grass genomes, those of sorghum and maize (Paterson *et al.* 2009; Schnable *et al.* 2009). Sorghum is an important grain crop in Africa, Central America and South Asia due to its tolerance of arid conditions. It is a tropical grass with a genome size of approximately 736 Mbp and a haploid chromosome number of 10. Sorghum was the first genome to be analysed using C_0t-based techniques to filter the genome into highly repetitive, moderately repetitive, and low-copy components (Peterson *et al.* 2002). Methylation filtration has also been used to capture the functional parts of the genome (Bedell *et al.* 2005). Both C_0t filtration (CF) and methylation filtration (MF) allow the selective characterisation of different parts of the genome and were proposed as methods to investigate plant species with more complex genomes (Rabinowicz and Bennetzen 2006). CF exploits the reassociation kinetics of DNA to separate repetitive sequences from low-copy DNA using single-stranded DNA affinity column chromatography (Peterson *et al.* 2002). MF utilises bacterial strains that destroy methylated DNA inserts in clone libraries to preferentially clone hypomethylated (gene-rich) regions (Rabinowicz 2003). Alternatively, hypomethylated DNA can be been cloned by partial restriction with methylation-sensitive enzymes (Emberton *et al.* 2005).

Sorghum is closely related to grasses in the *Saccharum* genus (sugarcane) and more distantly related to maize. The sorghum and maize lineages diverged around 12 million years ago (mya) (Swigonova *et al.* 2004) before the maize genome underwent a whole-genome duplication and numerous retrotransposon expansions. Since the divergence of sorghum and sugarcane ~5 mya, sugarcane has undergone two whole-genome duplications (Ming *et al.* 1998) therefore sorghum is an important intermediate species which can be used to identify the mechanisms and processes of genomic evolution in maize and sugarcane. The sorghum genome was sequenced by the United States Department of Energy's Joint Genome Initiative Community Sequencing Program using a whole-genome shotgun approach (Paterson *et al.* 2009). More than 60% of the sorghum genome was estimated to be repetitive and this was the first time a WGS strategy had been used to sequence such a complex grass genome. In order to deal with repeats, the sequencing strategy incorporated 8.5x coverage of paired-end reads from genomic libraries with insert sizes ranging from 2 kb to 108 kb. The read length obtained was of particularly high quality (average 723 bp) which aided the assembly. The assembly was validated by comparison to 27 finished BACs and was estimated to be 98.46 % complete with less than 1 error per 10 kb. The assembly was also improved by comparison to a high-density genetic map (Bowers *et al.* 2003a), a physical map and the completed rice genome sequence. The final assembly (v1) is 697.6 Mbp in size comprising 10 pseudo-molecules.

Maize is a domesticated crop descended from wild teosinte (Doebley *et al.* 2006) with a haploid chromosome number of 10 and an estimated genome size of 2300-2500 Mbp (Rayburn *et al.* 1993). This large size is due to a proliferation of LTR-retrotransposons occurring in the last 3 million years (SanMiguel *et al.* 1998) which constitute a large portion of the genome. The genome has also undergone whole-genome duplication (WGD) between 5 and 12 mya (Blanc and Wolfe 2004b; Swigonova *et al.* 2004) but has since lost most of its duplicated centromeric regions and paired gene sets resulting in a near-diploid genome with extensive fragmentation of ancestral gene orders (Ilic *et al.* 2003; Lai *et al.* 2004b). The maize genome was also shown to be particularly dynamic with the discovery of differences in gene content between inbred lines (Fu and Dooner 2002). The size, complexity and high-repeat content of the maize genome presented a challenge for traditional genome sequencing strategies. Initially, the maize genome was tackled using C_0t-based (Yuan *et al.* 2003) and methylation filtration techniques to remove methylated and highly repetitive regions before shotgun sequencing (Whitelaw *et al.* 2003). Analysis of 100 random BACs representing 0.6 % of the genome indicated a repetitive element content of 66% and between 42,000 and 56,000 genes were estimated to be present in the genome (Haberer *et al.* 2005).

Despite the potential difficulties, a project to sequence the whole maize genome was initiated in 2005 using a minimum tiling path (MTP) of clones from an integrated physical and genetic map (Schnable *et al.* 2009). A key resource was the creation of an optical map to span repetitive regions and to serve as independent verification of the sequence assembly (Zhou *et al.* 2009). Optical mapping involves shearing genomic DNA into fragments and imaging each DNA molecule whilst it is digested by restriction enzymes (Samad *et al.* 1995). This produces a restriction map for each molecule which are combined and assembled iteratively using computer software to form larger optical contigs in a largely automated process (Zhou *et al.*

2007b). The contigs from the physical map are then anchored to the optical map contigs *in silico*. The MTP clones were shotgun sequenced to 4-6x coverage, assembled, and the unique regions finished by automated and manual methods. The final release contains 2,048 Mbp of sequence in 61,161 super-contigs and is estimated to contain 89.2% of the genome. The remaining sequence is either not present in the physical map or contains tandem repeats that cannot be assembled.

Although the maize genome is more representative of the larger crop genomes, it is diminutive in comparison to the genomes of some wheat species, for example, the genome of *Triticum aestivum* (bread wheat or common wheat) has an estimated size of 17,000 Mbp (Bennett and Leitch 1995). This species is an allohexaploid, containing three homoeologous nuclear genomes from related progenitor species. The three constituent genomes are very similar in gene content and order, each consisting of seven chromosome pairs. Wheat forms part of the *Triticeae* tribe alongside barley and rye (*Secale cereale*) which are crops more suited to temperate climates. The *Triticeae* tribe also contains a number of wild grasses. The evolutionary history of wheat is complex and has been shaped by both domestication and hybridisation events between species in the *Triticum* and *Aegilops* genera. Extant wheat species are found in diploid, tetraploid and hexaploid accessions; diploid and tetraploid accessions are found both in the wild and as domesticated crops and hexaploid species only exist as domesticated crops. The earliest cultivated wheat species were *Triticum monococcum* (einkorn wheat) and *T. turgidum* subspecies (ssp.) *dicoccum* (emmer wheat), domesticated from wild grasses in south-eastern Turkey approximately 10,000 years ago (Heun *et al.* 1997). Einkorn wheat is a diploid species and emmer wheat is a subspecies of the main tetraploid wheat species *T. turgidum*, which also contains the subspecies *durum* (durum wheat). *T. turgidum* was formed by the hybridisation of the wild diploid grass *T. urartu* and an unknown diploid species within the *Sitopsis* section of the *Aegilops* genera closely related to *Ae. speltoides* less than half a million years ago but long before domestication (Dvorak *et al.* 1993; Huang *et al.* 2002; Salse *et al.* 2008b). The genome of tetraploid wheat contains two diploid genomes, A and B, with the A genome contributed by *T. urartu* and the B genome contributed by the unknown diploid species. Soon after the domestication of *T. turgidum*, spontaneous hybridisation occurred between tetraploid wheat and *Ae. tauschii* (Huang *et al.* 2002). This gave rise to hexaploid wheat *T. aestivum* containing three constituent genomes, A, B and D. At meiosis, pairing is only observed between homologous chromosomes as the *Ph1* locus (a region on chromosome 5B) prevents the pairing of homoeologous chromosomes (Sears 1976).

In general, polyploid plant species are more aggressive, competitive and adaptable to different environments than their diploid progenitors and this is the case with wheat species. Species such as emmer and einkorn are not widely grown in the modern world and hexaploid wheat accounts for the majority of wheat produced globally (FAOStat 2010). Compared to diploid or tetraploid wheat, hexaploid wheat is adaptable to a wider range of environmental conditions due to the genetic diversity captured within its constituent genomes (Dubcovsky and Dvorak 2007). In addition, the duplicated genetic material in the genome is tolerant to genomic changes that would be fatal in a diploid species but which contribute to accelerated evolution in the polyploid species by neo- and subfunctionalisation (Feldman and Levy 2009).

The huge size and the polyploid nature of the wheat genome mean that the application of traditional genome sequences strategies, such as the hierarchical or WGS approaches, is challenging. An additional complication is the high repeat content of the genome, estimated to be more than 80 % (Smith and Flavell 1975). However, several approaches have been used to gain insight into the structure and content of this large genome. In order to reveal information about gene content, a large EST collection has been generated for wheat. There are currently 1,071,050 ESTs sequences available (dbEST release 051410) which have been used to predict around 30,000 unique genes (Ogihara *et al.* 2003; Lazo *et al.* 2004; Zhang *et al.* 2004). Transcript assemblies built from existing wheat sequence data are available from the J. Craig Venter Institute (formally The Institute of Genomic Research, TIGR), the Gene Index Project at Harvard School of Public Health, and NCBI UniGene. These groups use different assembly methods which produce different datasets but currently 40,870 unique genes (UniGene Build #56) and 221,925 unique transcript assemblies (TaGI release 12.0) are predicted.

An important resource to identify the approximate location of genes in the wheat genome is the wheat deletion bin map (Qi *et al.* 2004). This map was constructed by the U.S. National Science Foundation-funded wheat EST project and uses wheat deletion stocks where a known portion of a single wheat chromosome has been removed (Endo and Gill 1996). In wheat, the deletion of a whole chromosome, a chromosome-arm or a large segment of a chromosome can be tolerated due to the redundancy provided by the polyploid genome. ESTs are used as probes against each stock until the absence of a signal allows the allocation of the EST into a deletion bin. Each chromosome arm in the wheat genome contains 3 or 4 deletion bins and a total of 7,107 ESTs have been cytogenetically mapped, defining 16,099 loci.

The availability of the rice genome sequence (International Rice Genome Sequencing Project 2005) has facilitated gene isolation in wheat producing a collection of BAC clones covering many agronomically important genes (Chantret *et al.* 2005; Isidore *et al.* 2005; Griffiths *et al.* 2006; Gu *et al.* 2006). In some cases, BAC clones have been fully sequenced revealing groups of genes extending over 10-30 kb interspersed by long stretches (> 100 kb) of repetitive DNA (Wicker *et* *al.* 2001; SanMiguel *et al.* 2002). Repetitive elements have also been characterised and are available in the Triticeae Repeat Database (Wicker *et al.* 2002). In some cases, the same locus has been sequenced from different species (Chantret *et al.* 2005; Isidore *et al.* 2005) revealing major differences in the intergenic repetitive DNA between these species. From an analysis of 4 BAC clones, gene density in wheat was estimated to be one gene every 75 kb, lower than previous estimates (Devos *et al.* 2005). CF has been used in wheat resulting in over 13-fold enrichment of genic sequences and a 3-fold reduction in repetitive DNA suggesting that CF could be used to sample the gene-space of wheat (Lamoureux *et al.* 2005). Analysis of BAC-end sequence data representing approximately 1% of wheat chromosome 3B predicted a gene content for the B-genome of wheat of ~36,000, comparable to that of rice and slightly lower than maize (Paux *et al.* 2006). The authors suggest that the size difference between the wheat and rice genomes is not due to gene amplification but to the amplification of transposable elements. A more recent analysis of 18.2 Mb of sequence from wheat chromosome 3B confirmed that genes are clustered into small islands containing an average of 3 genes separated by up to 800 kb of intergenic sequence (Choulet *et al.* 2010). Gene density was higher towards the distal ends of chromosomes compared to pericentromeric regions and it was estimated that between 36,000 to 50,000 genes are present on the seven chromosomes of the B genome. Around 800 new transposable element families were also identified in this analysis.

Much of the complexity within the wheat genome arises from its hexaploid nature with each gene being present as three homoeologous copies. Useful knowledge can be gleaned from analysis of other species within the *Triticeae* with less complex genomes such as barley and *Ae. tauschii*, the D-genome donor of hexaploid wheat. Barley is genetically very similar to wheat having diverged between 10 and 14 MYA (Wolfe *et al.* 1989) but is a diploid species (2n = 2x = 14). Barley has an estimated genome size of 5,100 Mb (Dolezel *et al.* 1998). The International Barley Sequencing Consortium was set up in 2006 with the aim of constructing a physical map of the barley genome using fingerprinted BAC clones representing 14 genome equivalents (Schulte *et al.* 2009). This map will be anchored to a high-density barley genetic map (Close *et al.* 2009) to facilitate the eventual goal of whole-genome sequencing. A genetic map was recently constructed for the seven chromosomes of *Ae. tauschii* and compared to the sequenced genomes of rice and sorghum (Luo *et al.* 2009). These comparisons identified the genome rearrangements that have taken place within the three lineages and postulated a mechanism of chromosome reduction in the *Triticeae* whereby an entire chromosome is inserted into the centromere of another chromosome. Physical mapping in *Ae. tauschii* is also underway and a physical map covering a large portion of the short arm of chromosome 3 has been constructed (Fleury *et al.* 2010). Although genomic studies in these larger genomes will help to pave the way for sequencing the wheat genome, they are still large and complex and bring with them many of the problems associated with wheat genomics.

1.5.3 Composition of grass genomes

Investigations into the genomes of grass species have revealed high levels of repetitive DNA, with larger repetitive proportions within larger genomes. For example, over 80 % of the wheat genome is repetitive, estimated using reassociation kinetics (Smith and Flavell 1975), compared to 50 % in rice (Deshpande and Ranjekar 1980). Further characterisation of genome sequences has shown that the majority of repetitive sequences are due to transposable elements. Thirty-five percent of the rice genome comprises transposable elements; 19 % retrotransposons and 13 % DNA transposons (Table 1.2). However, the smaller size of DNA transposons means that the actual number of these elements is more than double the number of retrotransposons in the genome. The larger genomes of sorghum and maize contain much larger proportions of retrotransposons but similar proportions of DNA transposons suggesting that expansion of retrotransposons accounts for much of the size differences observed between grass genomes.

Species	Genome size (Mb)	Retrotransposon content (%)	DNA transposon content (%)
Rice	400	19.4	13.0
Sorghum	800	54.5	7.5
Maize	2500	75.9	8.6

Table 1.2: Genome size and repeat content of three sequenced grass genomes.

(Data obtained from International Rice Genome Sequencing Project 2005; Paterson *et al.* 2009; Schnable *et al.* 2009)

The latest version of the Rice Annotation Project Database predicts 31,439 genes in rice and very similar gene complements are predicted for sorghum and maize of 27,640 and 32,540 respectively, suggesting a relatively consistent haploid gene content within the genomes of different grass species (Tanaka *et al.* 2008; Paterson *et al.* 2009; Schnable *et al.* 2009). These genes are not distributed evenly over chromosomes as observed in the genome of *Arabidopsis* but tend to be found in gene-rich euchromatic regions towards the distal ends of chromosomes with retrotransposon density highest in pericentromeric regions. In contrast, most DNA transposons tend to be found in gene-rich regions. In the larger genome of maize, individual genes or islands of up to four genes are separated by long stretches of repetitive sequence and also show differential gene distribution along chromosomes (Liu *et al.* 2007; Schnable *et al.* 2009).

1.5.4 Comparative genomics in the grasses

Comparative genomics in the grasses relies on the remarkable levels of conservation that exist between members of the grass family and this has been used to facilitate gene discovery and to characterise the evolutionary mechanisms that determine the structure and organisation of grass genomes. The conservation of both marker position and order (macrocollinearity) between chromosomes of different members of the grass family was revealed by early comparative mapping studies (Ahn and Tanksley 1993; Devos *et al.* 1994; Grivet *et al.* 1994; Kurata *et al.* 1994; Van Deynze *et al.* 1995; Kuhn *et al.* 2007) and the first consensus map to align the genetic maps of six grass species to the genetic reference map of rice was published in 1995 (Moore *et al.* 1995). In this study it was shown that the rice chromosomes could be dissected into linkage blocks and that it was possible to describe the genomes of wheat, maize, foxtail millet, sugar cane and sorghum by rearranging these blocks. The consensus map consisted of aligned linkage segments from the various genomes displayed in concentric circles and was called the 'crop circle'. This map has been updated since its initial publication to include other important members of the grass family (Devos and Gale 1997; Gale and Devos 1998; Devos 2005).

Comparative mapping at the marker level provided an initial evaluation of macrocollinearity between genomes. These observations led to more detailed investigations focusing on whether this conservation was also reflected at the DNA sequence level (microcollinearity). Microcollinearity is important for the isolation of genes from a genomic library as conservation must exist at the level of the large-insert clones that comprise the library. A comparison of gene organisation at the maize locus *sh2/a1* and the homologous regions in rice and sorghum revealed a conservation of gene order and composition (Chen *et al.* 1997). Another study revealed collinearity between orthologous loci in wheat and barley (Feuillet and Keller 1999). However, investigations also revealed that significant gene rearrangements had occurred in otherwise collinear regions of related grasses (Tikhonov *et al.* 1999; Tarchini *et al.* 2000). In contrast to the large scale rearrangements detailed in early mapping studies, these were small rearrangements that had occurred at the DNA sequence level and had not been detected at the genetic map level. Comparisons between several grass species (Bennetzen and Ma 2003) revealed many small rearrangements such as gene duplications and deletions, inversions and translocations in genic regions, together with high variability in intergenic regions that consisted mostly of transposable elements.

The sequencing of *Arabidopsis thaliana* provided the first analysis of a complete plant genome (The Arabidopsis Genome Initiative 2000), however this dicot species has limited use as a model for the monocot grasses as these two groups diverged approximately 150 MYA (Chaw *et al.* 2004). Comparative mapping between *Arabidopsis* and rice concluded that although some conservation exists, the level of conservation between the genomes is not great enough to facilitate the isolation of important genes (Devos *et al.* 1999). As a result of earlier comparative studies, rice was chosen as the “Rosetta Stone” of the grass genomes due to its relatively small genome compared with many other grass species (Messing and Llaca 1998). The genome sequences of rice subspecies *japonica* and *indica* (Yu *et al.* 2002; International Rice Genome Sequencing Project 2005; Yu *et al.* 2005) allowed detailed comparison between the genome of rice and the genetic and physical maps of other grasses increasing the resolution of previous studies. The first genome-scale comparison between rice and wheat utilised ESTs mapped into wheat deletion bins to align the wheat genome to that of rice (Sorrells *et al.* 2003; La Rota and Sorrells 2004). These alignments revealed an overall conservation of collinearity with numerous rearrangements, insertions, deletions and duplications that have eroded the synteny between the species.

Early studies using molecular markers observed duplicated loci in many cereal genomes indicating genome duplications and polyploidisation events within seemingly diploid genomes. Indications of duplicated segments within the chromosomes of maize were observed using molecular markers (Ahn and Tanksley 1993). These duplications were subsequently identified as arising from a WGD by allotetraploidisation from analysis of duplicated genes and comparison of orthologous loci in rice, maize and sorghum (Gaut and Doebley 1997; Swigonova *et al.* 2004). Duplications have also been identified in the rice genome (Yu *et al.* 2005), providing evidence of a WGD that occurred before the divergence of the cereal genomes 55-70 Mya (Kellogg 2001), and thus predicted to be shared by all the grasses. A segmental duplication between chromosomes 11 and 12 was also identified as well as many other gene duplications. One of the difficulties in performing this type of analysis is distinguishing paralogous and orthologous relationships between genes. To overcome this, Jerome Salse and colleagues applied modified alignment criteria to identify inter- and intra-specific duplications in the grass genomes and performed statistical validation on the results to determine whether the same gene order within two chromosomal segments truly reflects collinearity between the genomes or simply occurs by chance (Salse *et al.* 2008a). This method confirmed previously identified duplications in the rice genome as well as three new duplications. In addition, 12 duplicated regions in the wheat genome were identified using bin-mapped ESTs. A similar analysis identified orthologous regions between the two genomes which allowed the duplications shared between rice and wheat to be determined. Combining this analysis with previous comparative analyses between rice and maize (Salse *et al.* 2004; Wei *et al.* 2007) and rice and sorghum (Paterson *et al.* 2004a) allowed the authors to propose a common cereal ancestor containing five chromosomes (2n=10). Subsequently, this ancestor underwent a whole genome duplication to give 10 chromosomes (2n=4x=20) followed by further rearrangements and diploidisation to give an ancestral intermediate with 12 chromosomes (2n=2x=12). The authors suggest evolutionary mechanisms for the formation of the genomes of rice, wheat, sorghum and maize from this intermediate. The recent comparison of an *Ae. tauschii* genetic map to the pseudo-molecules of rice and sorghum supported the hypothesis that the common ancestor of *Triticeae*, rice and sorghum had 12 chromosomes (Luo *et al.* 2009). An alternate evolutionary model has recently been proposed by Katrien Devos (Devos 2010) whereby the ancestral genome consisted of 7 chromosomes that underwent allotetraploidisation to 14 chromosomes (2n=4x=28) before reduction by chromosomal rearrangements and diploidisation to a 12 chromosome intermediate (2n=2x=24). She argued that evolutionary processes tend to lead to a reduction in chromosome number rather than an increase and that the scenario of two chromosomes breaking and recombining to form three (as proposed in the Salse *et al.* model) is less likely than the insertion of a chromosome into the centromere of another chromosome forming a single chromosome.

Comparative investigations into grass genomes using the complete rice genome sequence as a reference have relied on the high degree of collinearity between grass genomes. Although rice has been an exceptionally useful model, in some cases, especially for investigations into wheat, its use as a model has been questioned due to the many rearrangements that have occurred since the two species diverged. An attempt to use the collinearity between rice and wheat to characterise the *Ph1* pairing locus in wheat used markers from rice to locate the orthologous region in wheat (Griffiths *et al.* 2006). The authors found that many of the markers failed to hybridise, indicating significant differences between these regions. An investigation into the evolutionary events observed at the *hardness* locus in wheat also uncovered a breakdown in rice-wheat collinearity, with genes missing from the orthologous region in rice (Chantret *et al.* 2005). Investigations such as these have revealed that synteny between wheat and rice is complex with many rearrangements disrupting the collinearity between important loci. In addition, the use of rice as a model for wheat elicits other problems in that rice is a specialised, semi-aquatic tropical grass that does not exhibit many of the important traits that exist in the temperate grasses such as resistance to specific pathogens, vernalisation and freezing tolerance. In practical terms, rice is a large plant with a long life cycle and is not particularly easy to cultivate in Europe, which has limited its uptake by the non-specialist research community. It is for these reasons that attention has turned to the identification of a new model grass with a closer relationship to wheat.

2 Brachypodium genomics

2.1 *Relevant publications*

International Brachypodium Initiative (2010). "Genome sequencing and analysis of the model grass *Brachypodium distachyon*." Nature **463**(7282): 763-768.

David F. Garvin, Neil McKenzie, John P. Vogel, Todd C. Mockler, Zachary J. Blankenheim, Jonathan Wright, Jitender J.S. Cheema, Jo Dicks, Naxin Huo, Daniel M. Hayden, Yong Gu, Christian Tobias, Jeff H. Chang, Ashley Chu, Martin Trick, Todd P. Michael, Michael W. Bevan, and John W. Snape (2010). "An SSR-based genetic linkage map of the model grass *Brachypodium distachyon*." Genome **53**(1): 1-13.

Melanie Febrer, Jose Luis Goicoechea, Jonathan Wright, Neil McKenzie, Xiang Song, Jinke Lin, Kristi Collura, Marina Wissotski, Yeisoo Yu, Jetty S.S. Ammiraju, Elzbieta Wolny, Dominika Idziak, Alexander Betekhtin, Dave Kudrna, Robert Hasterok, Rod A. Wing, Michael W. Bevan (2010). "An integrated physical, genetic and cytogenetic map of *Brachypodium distachyon*, a model system for grass research." PLoS ONE **5**(10): e13461

2.2 *Introduction*

The tribe *Brachypodieae* consists of a single genus that contains several temperate wild grass species grouped under the common name of False Bromes. This tribe is a member of the sub-family *Pooideae* within the BEP clade of grasses which diverged from the rice lineage around 50 million years ago (Catalan *et al.* 1995; Kellogg 2001; Paterson *et al.* 2004a). *Pooid* grasses are classified as temperate grasses in contrast to the tropical grasses found in the *Panicoid* and *Ehrhartoid* sub-families such as maize, sorghum and rice (Figure 2.1). In addition to the *Brachypodieae* tribe, the *Pooideae* sub-family contains the tribes *Triticeae, Aveneae* and *Poeae* which include temperate grasses of major economic importance (Kellogg 2001). The *Triticeae* tribe contains the genera *Triticum* and *Hordeum* which include wheat and barley, as well as the *Aegilops* genus which contains species that have been involved in the evolution of polyploid wheat (Dvorak *et al.* 1993; Huang *et al.* 2002). The *Aveneae* tribe contains oat species, used for animal feedstock and the *Poeae* tribe contains the genera *Lolium* and *Festuca* which contain important forage grass species (Clayton *et al.* 2006 onwards).

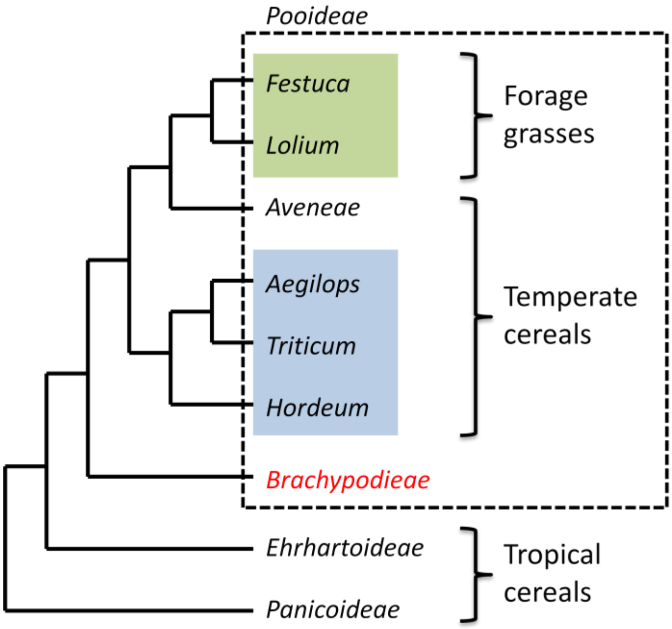


Figure 2.1: Cladogram showing the subfamilies, tribes and genera within the grass family. The tropical cereals such as rice and maize are found in the *Ehrhartoideae* and *Panicoideae* subfamilies and the temperate cereals and forage grasses are found in the *Pooideae*. The *Brachypodieae* tribe is positioned within the *Pooideae* alongside the *Aveneae*, *Poeae* and *Triticeae* tribes. Genera within the *Poeae* tribe are shaded green and genera within the *Triticeae* tribe are shaded blue (adapted from Draper *et al.* 2001).

Many species within the temperate grasses have large genomes but the genomes of several species within the *Brachypodieae* tribe are relatively small (Bennett and Leitch 1995). This, coupled with the close evolutionary relationship between the *Brachypodieae* and the *Triticeae*, prompted the suggestion by Moore and colleagues that the small genome of *Brachypodium sylvaticum* could be used for gene isolation and comparative analysis in the larger genomes of wheat and barley (Moore *et al.* 1993b). In the same year, Moore had used a methylation-sensitive restriction enzyme to assess the distribution of genes over wheat, barley, rice and *B. sylvaticum* chromosomes and identified that gene density appeared to be higher in the distal/subtelomeric regions than in the pericentromeric regions (Moore *et al.* 1993a). Two years later the important ‘crop circle’ publication emerged from the same group showing that the genomes of six major grass species could be aligned by dissecting individual linkage groups into blocks and arranging these blocks into different configurations, thus indicating a general conservation of gene order between the grasses (Moore *et al.* 1995). *B. sylvaticum* was also used to identify archetypal centromere sequences in the grasses (Aragon-Alcaide *et al.* 1996).

In 2004, a BAC library was constructed from *B. sylvaticum* consisting of 320,228 clones providing 6.6x genome coverage (Foote *et al.* 2004). This library was screened with probes from rice, barley and wheat to ascertain the level of synteny between *B. sylvaticum* and the other grass genomes. The study concluded that synteny between these species was largely maintained in the region studied. The same BAC library was used to identify and characterise the *Ph1* chromosome pairing locus in wheat (Griffiths *et al.* 2006). Initially, rice markers were used to screen wheat lines containing deletions within the *Ph1* locus but many failed to give a clear signal, so the same markers were used to screen the *B. sylvaticum* BAC library to identify homologous markers. In addition, new markers were identified from *B. sylvaticum* as a result of the screening. All the markers derived from *B. sylvaticum* gave a greater level of specific hybridisation than was obtained by the original rice markers, indicating a greater degree of useful similarity between wheat and *B. sylvaticum* than wheat and rice. This investigation also showed the potential of Brachypodium species to facilitate gene discovery in the large genomes of the temperate grasses. The *B. sylvaticum* BAC library was also screened using wheat ESTs to identify three BACs that were subsequently sequenced (Bossolini *et al.* 2007). The resulting 371 kb of sequence was compared with orthologous regions in rice and wheat to assess collinearity between the species and a general conservation of gene order and content between wheat and *B. sylvaticum* was observed. This study also estimated that the divergence of *B. sylvaticum* from the temperate grass lineage occurred between 35 and 40 MYA, significantly more recently than the divergence of rice and wheat (Paterson *et al.* 2004a).

B. distachyon was proposed as a model system for functional genomics in temperate grasses in 2001 (Draper *et al.* 2001). *B. distachyon* is an annual species with an estimated genome size of 355 Mb (Bennett and Leitch 1995). It is the preferred choice of *Brachypodieae* as *B. sylvaticum* is a perennial species with a larger tetraploid genome of 470 Mb. In their 2001 publication, Draper and colleagues made a compelling case for the adoption of *B. distachyon* (hereafter referred to as Brachypodium) as a model experimental system based on its physical properties, its genomic properties and its phylogenetic position in the grasses. Brachypodium is short in stature, self fertile, inbreeding, has a life cycle of about four months and undemanding growth requirements (Draper *et al.* 2001; Garvin *et al.* 2008). The small genome of Brachypodium was estimated to contain less than 15 % highly repeated DNA (Shi *et al.* 1993; Catalan *et al.* 1995; Catalan and Olmstead 2000). Draper and colleagues also reported the results of preliminary cytogenetic investigations to characterise the karyotype of Brachypodium showing that diploid ecotypes contained five chromosomes with chromosome 1 being much larger than the other chromosomes (Figure 2.2). Chromosomes 3 and 4 were very similar in size and chromosome 5 was much smaller. The locations of the 5S and 45S rDNA loci were identified as chromosomes 4 and 5 respectively (Draper *et al.* 2001). Using a combination of morphology, relative size and rDNA loci, all five pairs of chromosomes in the Brachypodium karyotype were distinguishable.

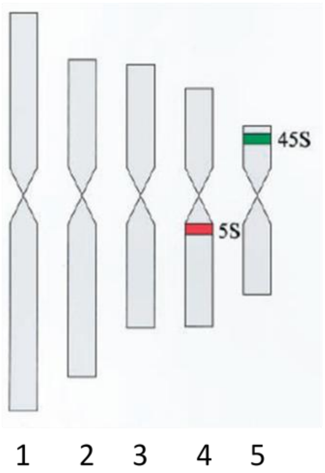


Figure 2.2: Ideogram showing the haploid set of five Brachypodium chromosomes.
Each chromosome has a distinctive and diagnostic shape and length. The 5S and 45S rDNA loci are indicated on chromosomes 4 and 5 (obtained from Draper *et al.* 2001).

The karyotype of Brachypodium was characterised in more detail using fluorescence *in situ* hybridisation or FISH (Hasterok *et al.* 2006). FISH is a technique whereby BACs are labelled with fluorophores and used as probes to decorate chromosome substrates (Pinkel *et al.* 1986). The location on the chromosome where the probe is bound is determined by epifluorescence microscopy. Usually two different fluorophores are used in a single experiment (called dual-colour FISH) allowing the location of two BACs to be determined simultaneously. In the resulting images the probes appear as red or green spots on a blue chromosome which is stained with DAPI, a fluorescent stain that binds strongly to DNA (Figure 2.3). Multi-coloured FISH utilises more than two different fluorophores. In order to determine which chromosome each BAC hybridises to, labelled BACs of known location are generally used in combination with probe BACs. In the Brachypodium study, two BAC libraries were prepared from Brachypodium diploid ecotypes and screened with 13 primer pairs designed to amplify genic sequences from a locus on rice chromosome 6 as well as 9 markers that had been genetically mapped in *Lolium perenne* (a forage grass), *Triticeae* species and rice. This screen identified 15 BACs and an additional 24 were selected randomly from the library. In total, 39 BAC clones were used as probes to interrogate the chromosomes. In addition, a clone containing the 25S rDNA region from *A. thaliana* and a wheat clone containing the 5S rDNA region were used to detect rDNA loci. Of the 39 clones, 32 hybridised to single locus and could be assigned to individual chromosome arms of the karyotype (designated p for short arm and q for long arm).

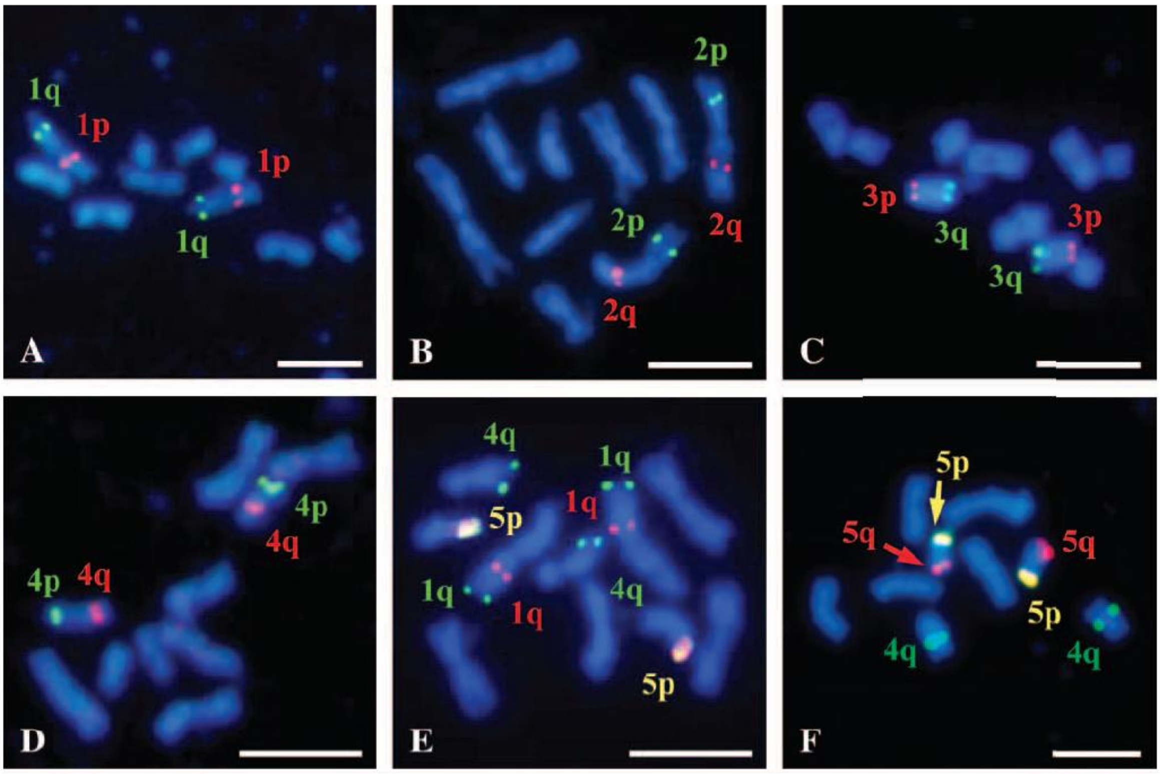


Figure 2.3: Identification of Brachypodium accession ABR1 chromosomes by FISH using BAC clones. Panels A-D use dual-colour FISH to identify the chromosome arms of chromosomes 1 to 4. The chromosomes appear as blue shapes due to DAPI staining and BAC clones as red and green spots. The chromosomal location is shown in the same colour as the corresponding BAC probe. Panels E and F use multi-colour FISH and show (E) the location of the 25S rDNA region in yellow and (F) the location of the 5S rDNA region in green and the 25S rDNA region in yellow. The bar indicates a distance of 5 µm. (Images obtained from Hasterok *et al.* 2006)

The high level of single-locus hybridisation by BAC clones, especially randomly selected ones, indicated that the Brachypodium genome is compact with low levels of repetitive DNA. In addition, BACs selected on the basis of synteny with other grasses mostly hybridised to chromosomal positions predicted by their marker locations in the original species. For example, all but one BAC identified by markers that mapped to different chromosomes in *L. perenne* also hybridised to different chromosomes in Brachypodium and all BACs identified by markers mapping to the same chromosome in *L. perenne* hybridised to the same chromosome in Brachypodium. The markers from the rice chromosome 6 loci also indicated conserved gene order between rice and the orthologous loci in Brachypodium with many markers being physically associated by anchoring to the same BAC.

In order to focus research efforts in Brachypodium, the International Brachypodium Initiative (IBI) was formed in 2005 and the first meeting was held at the 2006 Plant and Animal Genome conference in San Diego with the aim of establishing resources to promote Brachypodium as an experimental system for temperate grass research. The five objectives agreed during this meeting were;

1. Develop a set of community standard lines that are genetically well-defined and distribute these to all researchers free of encumbrances.
2. Promote the development and distribution of genomic and genetic resources such as BAC libraries, genetic markers and mapping populations.
3. Initiate a genome sequencing programme.
4. Establish collaborative links with other researchers e.g. in comparative genomics of crops.
5. Develop a web portal and genome database to link the research community, promote discussion and provide access to data, lines and other information.

Later the same year, a press release by the US Department of Energy's Joint Genome Institute (DOE-JGI) indicated its intention to sequence Brachypodium stating "This choice responds to the urgent need for developing grasses into superior energy crops and improving grain crops and forage grasses for food production" (DOE-JGI 2006). The JGI runs a community sequencing program which provides the scientific community access to high-throughput sequencing facilities. Projects are chosen by scientific merit and judged by peer-review and relevance to the aims of the DOE. In addition to its relevance to improve grain crops and forage grasses by providing a template genome, the potential of using Brachypodium to develop prospective biofuel grasses such as *Miscanthus giganteus* and switchgrass (*Panicum virgatum*) was a key factor in this decision (DOE 2006). Furthermore, a Brachypodium genome sequence would provide the first representative genome from the *Pooideae*, the third major grass sub-family alongside the sequenced genomes of rice (International Rice Genome Sequencing Project 2005) and sorghum (Paterson *et al.* 2009) providing representatives from the *Ehrhartoid* and the *Panicoid* respectively. This would afford the opportunity to examine genome evolution in all the major grass lineages to generate a complete picture of grass chromosome evolution and to validate the hypothesised evolutionary model of the grasses from a common ancestor containing five chromosomes (Salse *et al.* 2008a).

When defining a potential model experimental system, care should be taken to define the genetic stock on which to base the research. This stock should be genetically and phenotypically well characterised with a clear provenance and widely available to the research community. Two main collections of Brachypodium ecotypes were available: one at the United States Department of Agriculture (USDA) National Plant Germplasm System (USDA ARS 2010) and one at Brachyomics, a company set up by the University of Aberystwyth in Wales to promote and distribute Brachypodium Germplasm (Draper *et al.* 2001). USDA lines were designated with the prefix 'PI' and Brachyomics lines with the prefix 'ABR'. Twenty-eight inbred lines were developed from the USDA collection by single-seed descent and assigned the 'Bd' prefix (Vogel *et al.* 2006a; Garvin 2007). As Brachypodium ecotypes exist in different ploidy levels, the Bd accessions were assessed by flow cytometry and five were identified as diploid (Vogel *et al.* 2006a). The diploid inbred line designated Bd21 was eventually chosen and distributed widely to the research community (Garvin *et al.* 2008). One of the factors for choosing this line was the lack of vernalisation required for flower development (Vogel *et al.* 2006a). If plants are grown under conditions simulating 20 hour day lengths, seed-to-seed development can be as little as two months meaning that high-throughput genetics is possible.

Much of the previous cytogenetic investigations of the Brachypodium karyotype used diploid ecotypes (ABR1 and ABR5) from the Brachyomics collection (Draper *et al.* 2001; Hasterok *et al.* 2006). However, investigations into the karyotype of Bd21 showed that it was identical to the ABR accessions and that information gleaned in these studies could readily be transferred to Bd21 (Garvin *et al.* 2008). For example, marker BACs used to label specific chromosome arms in previous experiments (Hasterok *et al.* 2006) could be used in new FISH experiments.

Bd21 was used to develop BAC libraries (Huo *et al.* 2006; Febrer *et al.* 2010) and an EST collection from sequencing cDNA (Vogel *et al.* 2006b). The BAC library produced by Huo and colleagues was end-sequenced generating 1.3 Mb of random sequence from the Brachypodium nuclear genome (Huo *et al.* 2006). Only 4 % of BES showed similarity to repetitive sequences compared to 40.0 % showing similarity to ESTs indicating that a large proportion of the genome was likely to be transcribed and supporting previous evidence for low repetitive content in the genome (Catalan *et al.* 1995). In addition, when compared to species-specific EST collections the BES showed more similarity to wheat ESTs than to maize or rice, reflecting the close evolutionary relationship between Brachypodium and wheat. A subset of the 20,440 ESTs generated from five cDNA libraries made from leaves, stems plus leaf sheaths, roots, callus and developing seed heads was used to construct a phylogenetic tree that supported the close evolutionary relationship between Brachypodium and the temperate grasses (Vogel *et al.* 2006b). These ESTs were deposited in Genbank which substantially increased the publically available Brachypodium ESTs from 9 to 20,449.

This chapter describes my contributions to the development of Brachypodium genomics resources which involved a number of international collaborations. A Brachypodium genetic linkage map was constructed in collaboration with David Garvin at USDA. I contributed molecular markers and comparative genomics expertise to this project. Melanie Febrer (MF) from the John Innes Centre (JIC) and I constructed a BAC-based physical map of the Brachypodium genome using two BAC libraries which we anchored to the genetic map. I annotated the initial Brachypodium genome assembly and provided this information to the wider research community via a genome browser at the website I developed for the dissemination of Brachypodium genomics resources (www.modelcrop.org). This web resource was also used to display the genetic and physical maps. As part of the assembly process I provided data and analysed results from fluorescence *in situ* hybridisation experiments performed by Robert Hasterok (RH) at the University of Silesia in Poland. These were designed to validate the sequence contigs emerging from the assembly process. I also compared sequence contigs from the Brachypodium genome to genomic data from rice, wheat and sorghum to understand how Brachypodium relates to these other important grasses. This chapter concludes with a discussion of the results and conclusions from the Brachypodium genomics activities.

2.3 Overview of sequencing, assembly and annotation

Nuclear DNA from the diploid single seed descent line Bd21 (Garvin *et al.* 2008) was used to prepare clone libraries of varying insert size (Table 2.1).

Library	Insert size (bases)	Reads	Coverage
3kb (1)	3,215	277,248	0.65
3kb (2)	3,237	1,519,924	3.17
8kb (1)	6,381	855,422	2.04
8kb (2)	6,392	1,448,347	2.46
fosmid (1)	32,823	60,767	0.06
fosmid (2)	35,691	325,536	0.52
BAC BRA (BAC DH)	94,073	110,592	0.22
BAC BRB (BAC DB)	101,562	36,864	0.08
BAC DH (*Hind*III)[1]	103,216	30,704	0.05
BAC DB (*Bam*HI)[1]	108,177	36,388	0.04
BAC BD_CBa (*Eco*RI)[2]	124,935	25,948	0.05
BAC BD_ABa (*Hind*III)[2]	149,112	34,177	0.07
TOTAL		**4,761,917**	**9.43**

Table 2.1: Clone libraries end-sequenced in the Brachypodium genome sequencing project.

Libraries of varying sizes were prepared and end-sequenced. The total number of reads and final genome coverage is shown. [1] BAC libraries described in Huo *et al.*, 2006, [2] BAC libraries described in Febrer *et al.* 2010.

Cosmid libraries contained an insert size of 3 to 6 kb, fosmid libraries 30 to 40 kb and BAC libraries 100 to 150 kb. The clones were end-sequenced in a whole-genome shotgun approach using standard Sanger protocols on ABI 3730 instruments. A total of 4,761,917 reads were obtained providing an estimated genome coverage of 9.4x. These reads were assembled at the JGI using the ARACHNE assembler (Batzoglou *et al.* 2002). ARACHNE first constructs contigs based on overlapping reads then uses mate-pair reads from the larger clones to join contigs together into larger super-contigs.

The timeline of the sequencing programme is shown in Figure 2.4. BAC-end sequencing began in April 2007 and the initial output to the IBI was a mid-point assembly in August 2007, resulting from sequencing the genome to 4x coverage and assembling the reads. The mid-point assembly consisted of 281.2 Mb of sequence in 1,015 super-contigs, including 21.4 Mb of sequence gaps. 17 super-contigs (super_0 to super_16) were larger than 1 Mb and contained 97.8% of all sequenced nucleotides.

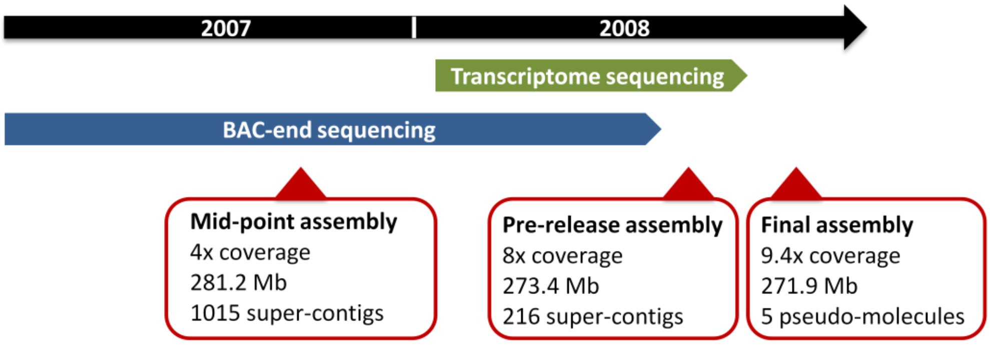


Figure 2.4: Timeline of the Brachypodium sequencing program at the JGI.
BAC-end sequencing started in 2007 and continued into 2008. Three assemblies were released, the mid-point assembly, the pre-release assembly and the final assembly. Towards the end of the sequencing, cDNA libraries were sequenced to generate ESTs for genome annotation.

Once sequencing was completed, all the reads were assembled which resulted in set of 216 super-contigs consisting of 273.4 Mb of sequence, including 1.1 Mb of sequence gaps. 10 super-contigs (super_0 to super_9) were larger than 1 Mb and these contained 99.3% of the sequenced nucleotides. This assembly was called the pre-release assembly and was released to the research community in August 2007.

The final assembly resulted from the alignment of the 10 largest super-contigs from the pre-release assembly (super_0 to super_9) to a high-density genetic map constructed from 562 SNP markers (Figure 2.5). This alignment indicated two false joins within super-contigs 2 and 4 which were also identified by cytogenetically anchoring the super-contigs. These super-contigs were broken before all super-contigs were joined as indicated by alignment to the genetic map (Figure 2.6). This produced an assembly consisting of five pseudo-molecules totalling 271.9 Mb of sequence. An additional 774 kb of sequence was left unassembled in additional scaffolds. The final assembly was provided by the JGI to the research community for analysis and annotation in September 2008.

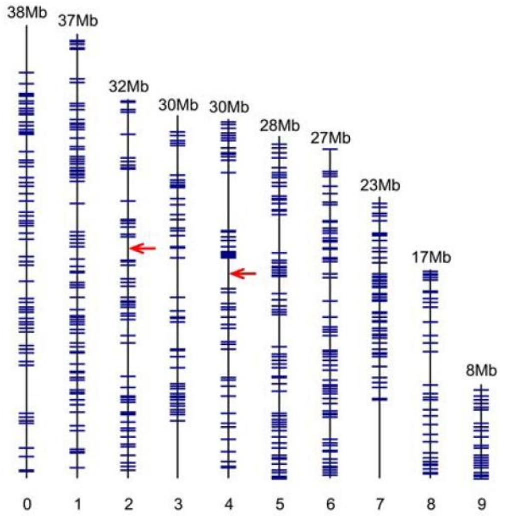


Figure 2.5: Identification of false joins in the pre-release assembly.
The 10 largest super-contigs from the pre-release assembly (numbered 0 to 9) were compared to a high-density genetic map constructed from 562 SNP markers. Blue lines indicate the positions of genetic markers on the super-contigs. Two false joins were detected in super-contigs 2 and 4 (shown by red arrows) which were broken in the final assembly. The length of each super-contig is shown at the top of the figure.

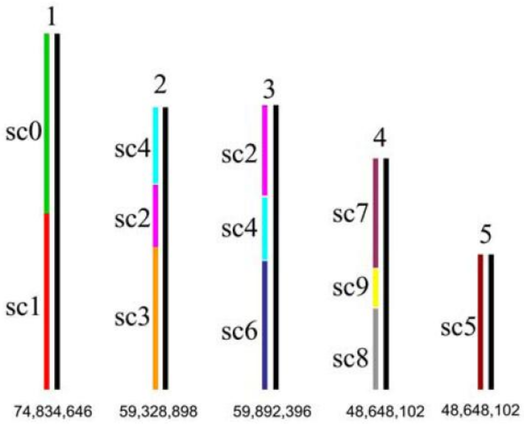


Figure 2.6: Building the pseudo-molecules of the final assembly.
Assignment of the pre-release super-contigs (labelled sc0 to sc9) to the five Brachypodium pseudo-molecules (labelled 1 to 5). Super-contigs 2 and 4 have been broken after false joins were indicated by comparison to the genetic map. The length of each pseudo-molecule is shown below each one.

In order to obtain experimental evidence for gene expression for the annotation phase of the genome project, the JGI also sequenced a number of cDNA libraries. Four normalised cDNA libraries were Sanger sequenced as well as 11 cDNA libraries that were sequenced using the 454 platform (Table 2.2). The cDNA libraries were made from a diverse collection of tissues, different environmental conditions and different developmental stages including diurnal sampling and other treatments designed to maximise transcript diversity.

Library	Number of ESTs	Platform	Sequenced by	Tissue/Stage /Treatment	Normalisation
CCXF	25494	Sanger	JGI	abiotic and biotic stress	DSN
CCXG	28229	Sanger	JGI	superpool	DSN
CCYO	26237	Sanger	JGI	flower, flower drought	DSN
CCYP	27821	Sanger	JGI	leaf, leaf drought	DSN
CCXU	49540	454	JGI	callus	N/A
CFAA	948	454	JGI	roots	DSN
CFAB	234	454	JGI	developing seeds	DSN
CFAC	1851	454	JGI	diurnally sampled seedlings	DSN
CFCF	405974	454	JGI	diurnally sampled roots	DSN
CFCG	317095	454	JGI	diurnally sampled leaves, stems	DSN
CFCH	362432	454	JGI	diurnally sampled flowers	DSN
CFCI	253491	454	JGI	callus	DSN
CFFH	129769	454	JGI	diurnally sampled leaves, stems, callus	DSN
CFFI	139968	454	JGI	diurnally sampled leaves, stems, callus	DSN
CFFN	93222	454	JGI	diurnally sampled leaves, stems, callus	DSN
AC60	170521	454	Schnable	root tips	N/A
AC61	89277	454	Schnable	root tips	N/A
AC63	157349	454	Schnable	root tips	N/A
AC64	122320	454	Schnable	root tips	N/A
callus	4196	Sanger	Vogel	callus	N/A
leaf	3780	Sanger	Vogel	leaf	N/A
root	3869	Sanger	Vogel	root	N/A
seed	4688	Sanger	Vogel	seed	N/A
stem	3907	Sanger	Vogel	Stem	N/A

Table 2.2: ESTs generated and used in the Brachypodium genome annotation.

JGI sequenced 4 cDNA libraries using traditional Sanger-sequencing and 11 libraries using 454. These were combined with ESTs from the Vogel group and ESTs generated from 4 cDNA libraries from the Schnable group. All libraries were made from Bd21 except for the 4 libraries generated by the Schnable group which were made from different USDA Brachypodium accessions. DSN = Double-strand nuclease, N/A = none

All of these cDNA libraries were enriched for full length transcripts and all but one were normalised to remove abundant transcripts using a duplex-specific nuclease (DSN) from the Kamchatka crab (Zhulidov *et al.* 2004). This normalisation method proceeds by heating double stranded cDNA to denature it, then slowly cooling it to allow the strands to anneal. At this stage the more abundant transcripts will anneal faster than the less abundant transcripts as they will more quickly find their complementary strand. The sample is then treated with the DSN which degrades the double stranded molecules representing the more abundant transcripts and leaving the single stranded rare transcripts unaffected.

For genome annotation, these ESTs were combined with existing ESTs from the Vogel group (Vogel *et al.* 2006b) and 454 ESTs generated from 4 additional cDNA libraries made from different USDA Brachypodium accessions. A total of 128,221 Sanger ESTs and 2,293,991 454 ESTs were used in the genome annotation and 126,072 Sanger ESTs (98.3 %) aligned to the genome sequence. This indicates a near complete assembly with the remaining ESTs probably originating from unassembled regions.

A superpool of cDNA was sequenced using the Illumina platform by the Mockler group and 280 million 32-base reads were generated. These reads were aligned to the annotated genome to assess the accuracy of the assembly and 92.7 % of predicted coding sequences were supported by these reads.

2.4 Annotation of the mid-point assembly

In August 2007, the mid-point assembly was released and consisted of 1015 super-contigs totalling more than 280 Mb of sequence, the largest genome-wide sequence sample from Brachypodium to date. Analysis and annotation of these super-contigs was crucial to more fully understand the structure of the Brachypodium genome including the number of genes it contains and its repetitive DNA content. I implemented an annotation pipeline to perform this task and loaded the assembly and annotation into a genome browser.

2.4.1 Methods

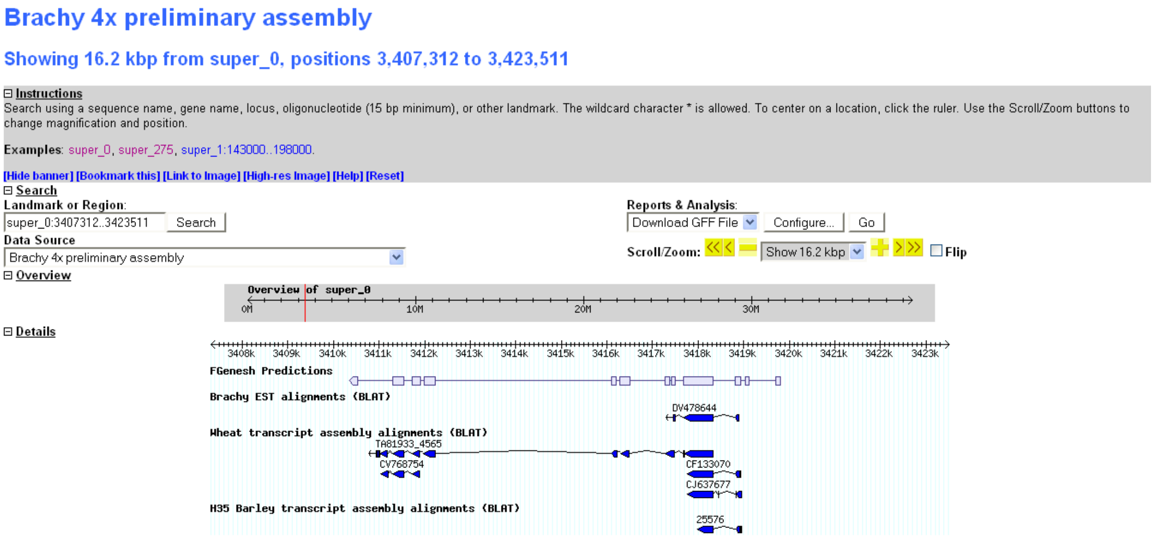
To annotate the super-contigs I adopted and modified an annotation pipeline originally developed by Martin Trick from the JIC for annotating *Brassica* BACs (http://brassica.bbsrc.ac.uk/annotate.html). The pipeline is written in Perl and utilises various BioPerl libraries to perform the annotation (Stajich *et al.* 2002). A FASTA file containing one or more genomic sequences is required as input and various analyses are performed producing output files in GFF (Durbin and Haussler 2010).

The script is provided as Supplementary information 5 and the annotation steps are as follows;

- Create a reference file to define the length of each of the super-contigs.
- Identify any gaps within the super-contigs.

- Annotate SSRs and identify primers for each SSR using Msatfinder (Thurston and Field 2005).
- Annotate repetitive sequence using RepeatMasker (Smit *et al.* 1999) with the built-in *O. sativa* library of repetitive elements.
- Identify gene models using FGENESH (Salamov and Solovyev 2000) using the monocot parameters file on the assembly after it had been masked for repeats. BLASTX (Gish 1996-2006) was used to identify the best hit from the rice genome annotation (TIGR version 5) for each predicted exon to give some indication of the function of predicted genes.
- Convert the BAC assembly coordinates file to GFF.
- Align ESTs and transcript assemblies using BLASTN (Gish 1996-2006) to find initial homologies, followed by BLAT (Kent 2002) to perform more accurate alignments. A threshold score (E-value for BLAST, percent identity for BLAT) is used to restrict alignments depending on the evolutionary distance of the species providing the ESTs. Brachypodium ESTs, deletion bin-mapped wheat ESTs, TIGR wheat transcript assemblies, TIGR barley transcript assemblies and barley transcript assemblies (harvest35 from http://harvest.ucr.edu/) were aligned to the assembly.

The annotated genomic sequence assemblies were displayed at www.modelcrop.org (Figure 2.7) using the genome browser GBrowse (Stein *et al.* 2002). GBrowse uses a MySQL relational database to store the genomic data displayed in the browser and the BioPerl module Bio::DB::GFF was used to query and analyse the genomic data programmatically using Perl scripts. The website also provides a BLAST server implementing WU-BLAST (Gish 1996-2006) to allow users to search the sequence assemblies, predicted genes and proteins. In addition, I provided Perl scripts to allow users to download Brachypodium sequence to enable them to perform their own analyses on the data. The site runs on a Dell Dual Core Xeon 3070 running the Linux-based operating system Fedora 6 and uses Apache HTTP Server v.2.2.3 as the web server.



Brachy 4x preliminary assembly

Showing 16.2 kbp from super_0, positions 3,407,312 to 3,423,511

⊟ Instructions

Search using a sequence name, gene name, locus, oligonucleotide (15 bp minimum), or other landmark. The wildcard character * is allowed. To center on a location, click the ruler. Use the Scroll/Zoom buttons to change magnification and position.

Examples: super_0, super_275, super_1:143000..198000.

[Hide banner] [Bookmark this] [Link to Image] [High-res Image] [Help] [Reset]

⊟ Search

Landmark or Region: super_0:3407312..3423511 Search

Data Source: Brachy 4x preliminary assembly

Reports & Analysis: Download GFF File Configure... Go

Scroll/Zoom: Show 16.2 kbp □ Flip

Figure 2.7: Screenshot from the Brachypodium genome browser at www.modelcrop.org.

The region displayed shows a predicted gene on the reverse strand of super-contig 0 consisting of 12 exons (shown as white boxes). The alignments of Brachypodium, wheat and barley ESTs (shown in blue) support 9 of the predicted exons.

2.4.2 Results

The gene prediction algorithm FGENESH (Salamov and Solovyev 2000) predicted 39,228 genes in the 281.2 Mb repeat-masked mid-point assembly. This gives a gene density of 13.9 genes per 100 kb. Approximately 34 % of the predicted genes were supported by aligned wheat ESTs, 28 % were supported by barley ESTs and only 15 % were supported by Brachypodium ESTs. At this stage, only 20,449 Brachypodium ESTs where available, hence the low percentage of predicted genes supported by Brachypodium ESTs.

EST dataset	Total # ESTs	# genes supported	% genes supported
Wheat	319,949	13,420	34
Barley	50,938	11,065	28
Brachypodium	20,449	6,084	15

Table 2.3: Support for predicted genes in the preliminary Brachypodium assembly

Brachypodium, barley and wheat ESTs were used to support gene predictions.

The average GC content in genic regions is 57.1 %, higher than the background GC content of the assembly as a whole (46.2 %). The annotation pipeline also identified 21,026 simple sequence repeats in the assembly. RepeatMasker identified that 20.1 Mb (7.2 %) of the assembly consisted of repetitive elements (Table 2.4), 17.6 Mb (6.3 %) of these are classified as retrotransposons (class I) and 2.4 Mb (0.1 %) are classified as DNA transposons (class II). Retrotransposons from the *Gypsy* superfamily are the predominant LTR retrotransposons comprising 3.44 % of the assembly and 55.0 % of the total retrotransposon complement in the super-contigs.

	Number	Length (bp)	Percentage of assembly
Retroelements			
SINEs	1,610	261,496	0.09
LINEs	3,274	2,225,494	0.79
Ty1/copia	7,712	5,222,336	1.87
Ty3/gypsy	16,085	9,663,425	3.44
Other LTR elements	3,118	225,837	0.06
Total class I	**31,799**	**17,598,588**	**6.25**
DNA transposons			
hobo-Activator	854	231,314	0.08
Tc1/Mariner	3,242	511,120	0.18
En-Spm	1,535	892,479	0.32
Mutator	1,867	623,679	0.22
Tourist/Harbinger	1,268	178,934	0.06
Total class II	**8,766**	**2,437,526**	**0.86**
Unclassified	**165**	**62,924**	**0.02**
Total	**40,730**	**20,099,038**	**7.13**

Table 2.4: Categorisation of transposable elements in the mid-point assembly. Obtained using Repeatmasker with the *Oryza sativa* library of repeats.

2.5 *Genetic linkage mapping*

2.5.1 Introduction

To fulfil one of the initial objectives of the IBI a Brachypodium genetic linkage map was constructed in a collaboration between groups at the JIC, the United States Department of Agriculture – Agricultural Research Service (USDA-ARS), Oregon State University, Rutgers University and University of California, Davis, under the leadership of David Garvin at USDA-ARS. This was the first genetic linkage map developed for Brachypodium and was seen as an essential resource to assess the quality of WGS sequence assemblies emerging from the JGI, to relate physical and genetic map distance in the Brachypodium genome and to evaluate collinearity to other grass genomes. A mapping population of 183 F_2 plants derived from a cross between inbred diploid Brachypodium lines Bd21 and Bd3-1 was used (Vogel *et al.* 2006a; Garvin *et al.* 2008). The linkage map used simple sequence repeat (SSR) markers designed from Brachypodium EST and BAC sequences, conserved orthologous sequence (COS) markers from other grasses and additional Brachypodium-derived molecular markers. During its construction the map went through three iterations. Builds 1 and 2 were performed by Neil McKenzie at the JIC using Joinmap (Stam 1993) and Build 3 by David Garvin at USDA-ARS using Map Manager

QTX (Manly *et al.* 2001) which is the published version of the genetic map (Garvin *et al.* 2010). Build 1 contained 151 markers distributed over 16 linkage groups with a total length of 1010 cM. Build 2 of the map contained 165 markers distributed over 13 linkage groups with a total length of 1007 cM. Build 3 contains 139 markers distributed over 20 linkage groups (a-t) with a total length of 1386 cM.

I contributed thirty-two markers to the genetic map, six of which were included in the final map. I also performed *in silico* alignments between the linkage groups and Brachypodium super-contigs as well as rice chromosomes. In addition, a number of linkage groups were anchored to the Brachypodium karyotype using FISH.

2.5.2 Methods

2.5.2.1 Analysis of Build 1 linkage groups using the mid-point assembly

I aligned the 16 linkage groups from Build 1 of the genetic map to the super-contigs in the mid-point assembly in order to assess their quality and to determine if any joins between linkage groups could be made. These alignments were performed by either locating the BAC or EST from which the marker was derived onto the Brachypodium super-contigs using BLAT, or identifying the position of the primer pairs in the Brachypodium assembly using Primersearch from the EMBOSS package (Rice *et al.* 2000). This data was used to build a formatted text file containing the position of each marker on the linkage groups as well as a file containing the identified positions of the markers on the super-contigs. These files were loaded into CMap (Youens-Clark *et al.* 2009) to visualise and compare the marker positions on the linkage groups and super-contigs.

Ideally, a genetic map will have a single linkage group representing each physical chromosome in the genome. Brachypodium has five chromosomes so in order to coalesce linkage groups I designed more markers, targeting specific regions on the mid-point assembly super-contigs where alignments indicated that additional markers would be useful. For this purpose, SSR markers were extracted from the GBrowse GFF database having been annotated by the annotation script described in Section 2.4. These additional markers were genetically mapped.

2.5.2.2 Analysis of Build 2 linkage groups using the mid-point assembly and FISH

Aligning the linkage groups from Build 2 of the genetic map to the super-contigs allowed me to again assess the quality of the linkage groups and, in addition, to predict possible joins between linkage groups based on linking super-contigs. In order to test one of these predicted joins I selected 10 BACs from a super-contig that defined the region of alignment. These BACs were sent to Poland for FISH to determine whether the linkage groups that were predicted to be joined could be anchored to the same physical Brachypodium chromosome. The BACs are labelled with fluorophores and used as probes against chromosome spreads made from root tip meristematic cells of Brachypodium. Epifluorescence microscopy was used to identify where each probe hybridises to the chromosome. In this case, two-colour FISH was used where two BACs are labelled with different coloured fluorophores (red or green) and applied to the chromosome spreads simultaneously. Subsequent FISH analyses were performed to attempt to locate all linkage groups onto Brachypodium chromosomes.

2.5.2.3 Comparison of Build 3 linkage groups with Brachypodium and rice

The 20 linkage groups in Build 3 were first aligned to the five Brachypodium pseudo-molecules by aligning the BAC and EST sequences from which the markers were derived to the sequence assembly using BLAST. The BLAST hit with the lowest E-value was used as the most likely position of the marker on the pseudo-molecules. The comparisons were visualised in CMap.

To assess collinearity between the Brachypodium and rice genomes, the linkage groups were compared to the rice genome. The BAC and EST sequences from which the markers were derived were aligned to the TIGR rice pseudo-molecules (release 5) using BLASTN with an E-value cut-off of 1 e^{-15}. For the sequences that returned hits less than the cut-off value, those with the lowest E-value were considered to be rice orthologues for purposes of collinearity comparisons. An exception was marker INTR5-11, where a secondary BLASTN hit with a very low E-value was used since the position of this secondary hit was in a collinear block of genes defined by other Brachypodium markers aligned to rice. Once again, these alignments were visualised using CMap.

2.5.3 Results

2.5.3.1 Analysis of Build 1 linkage groups using the mid-point assembly

Aligning the 16 linkage groups in Build 1 of the genetic map to super-contigs showed that all but four linkage groups showed good synteny with super-contigs indicating robust linkage groups had been constructed that reflect the assembled genome sequence. An example of a syntenic alignment is shown in Figure 2.8 where linkage group 1 is aligned to super-contigs 3, 5 and 8 by multiple anchor-points and shows three distinct blocks of collinearity within the linkage group.

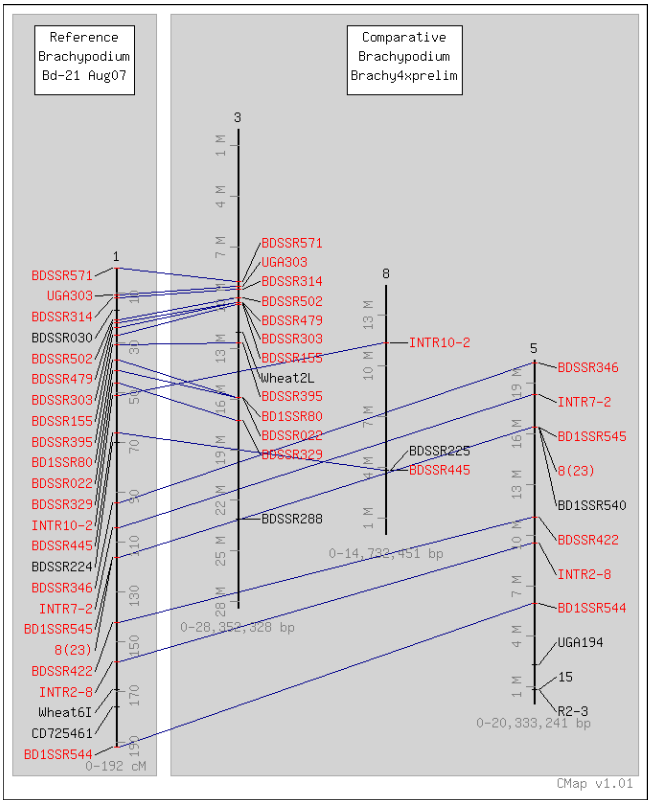


Figure 2.8: Alignment of linkage group 1 to super-contigs.

Linkage group 1 from Build 1 of the genetic map (on the left) is aligned to super-contigs 3, 5 and 8 from the mid-point assembly (on the right) by genetic markers anchored to the super-contigs (shown as blue lines). Marker names are indicated on the linkage group and the super-contigs.

Alignments between super-contigs and linkage groups indicated eight regions on the super-contigs where additional markers may help to join smaller linkage groups (Table 2.5).

Super-contig	Start position (Mb)	End position (Mb)	To join linkage groups
0	15.8	26.8	2, 7
0	34.6	36.9	7, 8
1	24.6	35.6	2, 4
1	6.6	9.0	6, 2
2	16.0	23.8	11, 13
3	17.3	23.1	1, 5
5	2.3	5.9	1, 15
7	1.9	5.9	12, 5, 6

Table 2.5: Regions on the super-contigs to target for new markers.

Regions were identified based on alignments between mid-point super-contigs and Build 1 linkage groups.

An example of such a region is shown in Figure 2.9. Simple sequence repeats were identified from these regions and flanking primer pairs designed resulting in 34 additional markers (Supplementary information 6). Six of these markers proved to be polymorphic between the two parental lines and were included in the next build of the genetic map. It is difficult to assess whether these markers alone allowed the joining of linkage groups, as additional markers from other sources were included in the next map build.

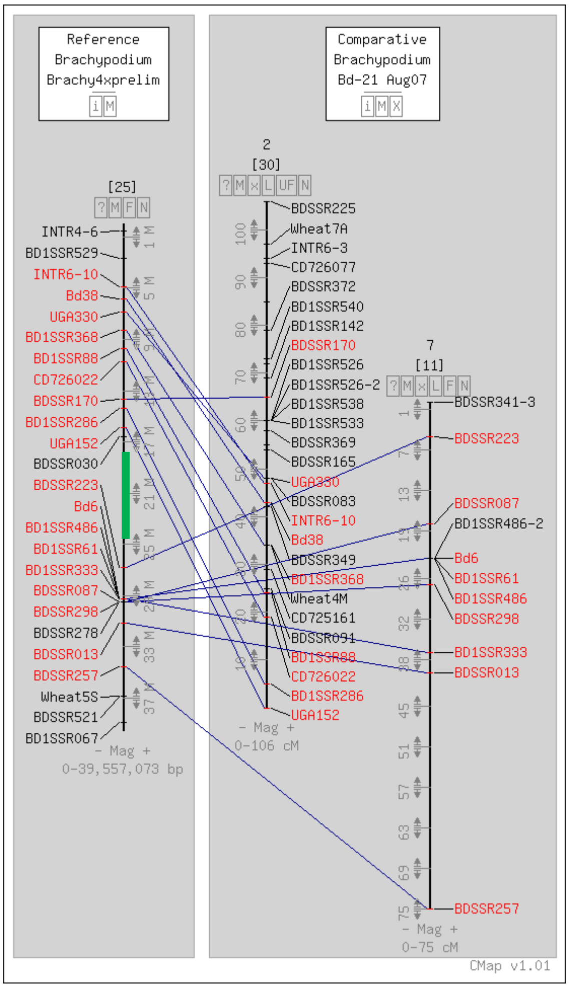


Figure 2.9: Alignment of super-contig 0 to linkage groups.

Super-contig 0 (shown on the left) aligned to linkage groups 2 and 7.The green area highlighted in the super-contig is an example of a region where additional markers may help to join these two linkage groups.

2.5.3.2 Analysis of Build 2 linkage groups using the mid-point assembly

Four of the thirteen linkage groups in Build 2 of the genetic map comprised merged linkage groups from Build 1 and one of these merges was predicted from the Build 1 alignments to super-contigs. The Build 2 linkage groups were aligned to the super-contigs and all but one linkage group (Lg9) showed good synteny with one or more super-contigs indicating robust linkage between markers was being achieved (Table 2.6). Some larger linkage groups (1, 2, 4 and 5) showed alignments to multiple super-contigs indicating possible false linkage within these groups but this could equally be due to false joins in the super-contigs at this early stage of the genome assembly.

Linkage group	Length (cM)	Aligns to super-contig
1	237	0, 1, 5, 3
2	96	7, 10, 9
3	89	0
4	174	7, 1, 12
5	158	4, 2
6	40	1
7	80	6
8	42	13
9	11	-
10	21	7
11	22	5
12	35	2
13	16	1

Table 2.6: Aligning the Build 2 linkage groups to mid-point super-contigs.

Each linkage group is shown alongside the length of the linkage group and the aligning super-contigs.

Some linkage groups align to the same super-contig, for example, linkage groups 1 and 3 both align to super-contig 0 (Table 2.6). These alignments were visualised to determine if the alignment indicated that linkage groups could be joined. To indicate a potential join, a super-contig should align to one end of two linkage groups. These alignments predicted possible joins between linkage groups 1 and 3, 4 and 10, and 5 and 12 to form three larger linkage groups (Figure 2.10).

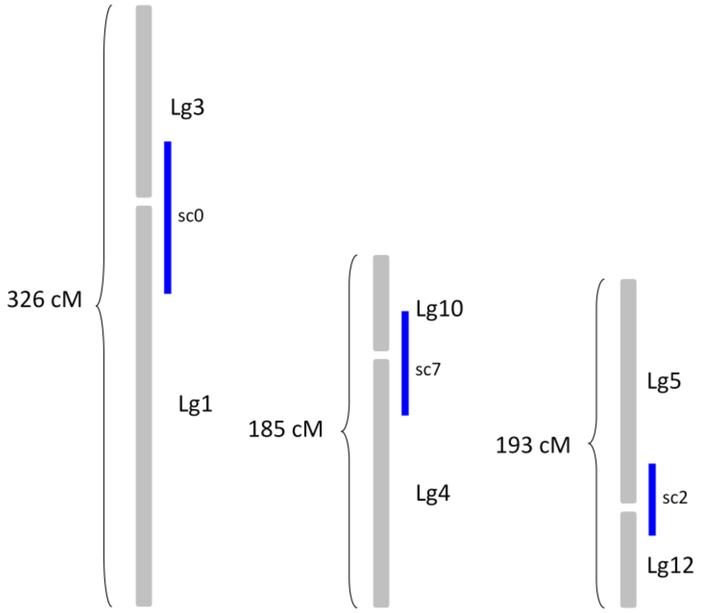


Figure 2.10: Three possible linkage group joins predicted by alignment to super-contigs.

Linkage groups are labelled as Lg1, Lg3 etc. (shown in grey) and super-contigs are labelled sc0, sc7 (shown in blue).

The length of the combined linkage group is shown to the left in centiMorgans (cM).

2.5.3.3 Validation of predicted join between linkage groups 1 and 3 of Build 2 using FISH

To validate the predicted join between linkage groups 1 and 3 using FISH, ten BACs were identified from super-contig 0 at evenly spaced intervals in aligning regions (Figure 2.11). Linkage group 1 aligns to super-contig 0 between approximately 5 and 18 Mb and linkage group 3 aligns to a region between 27 and 35 Mb. These BACs were used as probes for FISH.

Approx. Position (Mb)	BAC name
6	DB154I09
8	DH045C22
10	DB159K03
12	DB150J17
14	DH005B15

Approx. position (Mb)	BAC name
27	DH019G11
28	DB164C13
29	DB019H23
30	DH017H07
31	DH024M24

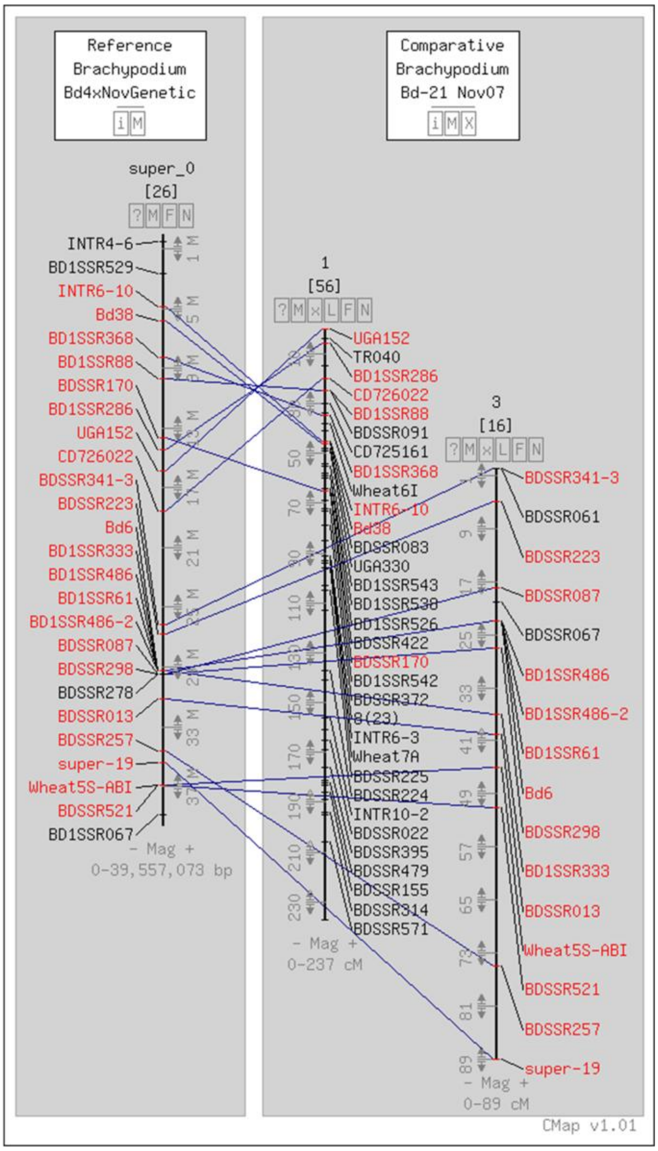


Figure 2.11: BAC clones selected for FISH from super-contig 0.

BACs were selected from super-contig 0 to validate the predicted join between linkage groups 1 and 3. The two tables on the left show the BAC clones selected from evenly spaced intervals from super-contig 0 (shown on the left of the comparison). Linkage groups 1 and 3 (shown on the right of the comparison) are aligned to super-contig 0 by genetic markers (shown as blue lines).

The images obtained (Figure 2.12 & Figure 2.13) show the location of each BAC probe as a red or green spot on the chromosomes that are stained light blue with DAPI. BAC probes that hybridise in close proximity to one another appear as yellow spots in the images. Figure 2.12 shows BAC probes taken from the region that aligns to linkage group 1 (image A) and BAC probes taken from the region that aligns with linkage group 3 (image B). Image A indicates that both probes hybridise to a single centromeric locus as the probes are displayed as red/green/yellow spots. Image B indicates that both probes hybridise to a single locus on the distal end of a chromosome.

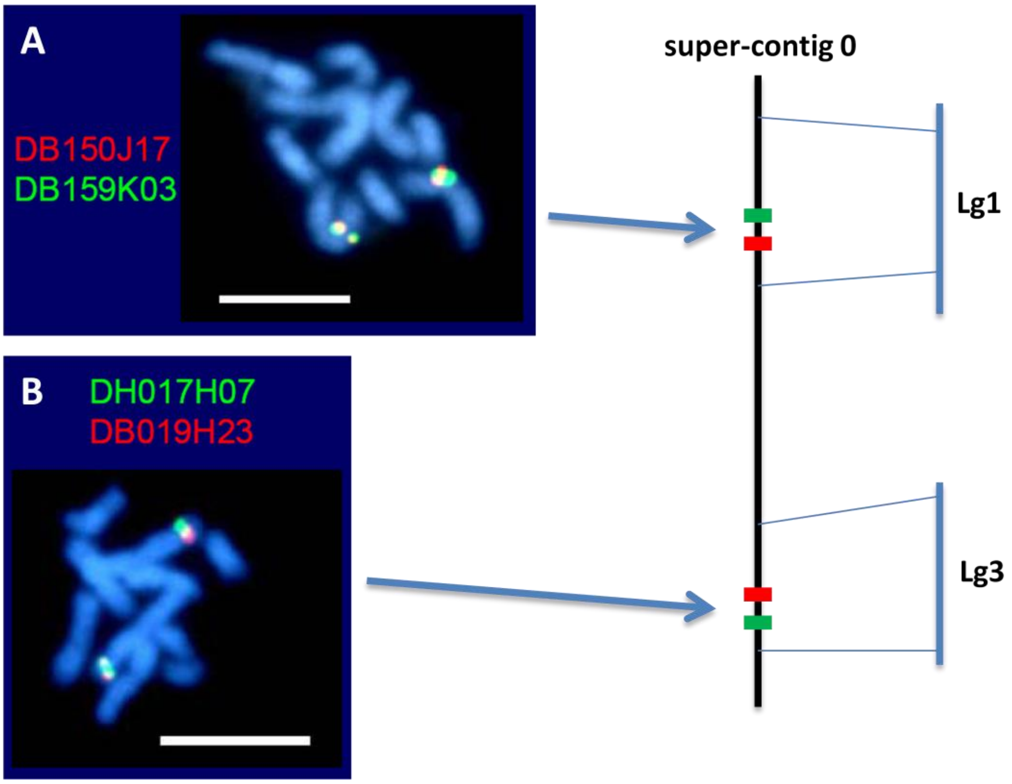


Figure 2.12: Locating linkage groups 1 and 3 onto chromosomes.
Epifluorescence microscopy images (A and B) showing the locations of labelled BAC probes on Brachypodium chromosome (light blue). The BAC probes are taken from super-contig 0 where linkage groups 1 and 3 align and are labelled with red and green fluorophores. The BAC names are shown on each image in the colour of the probe. The diagram on the right shows the relative locations of the probes on super-contig 0 as green and red blocks coloured relative to the fluorophore colour. (scale bar = 5μm)

Figure 2.13 shows two BAC probes, one from each aligned region. Each probe hybridises to distinct loci within a single chromosome arm indicating that the nucleotide sequence in super-contig 0 represents a large portion of a single Brachypodium chromosome arm and that this arm is represented genetically, at least in part, by linkage groups 1 and 3.

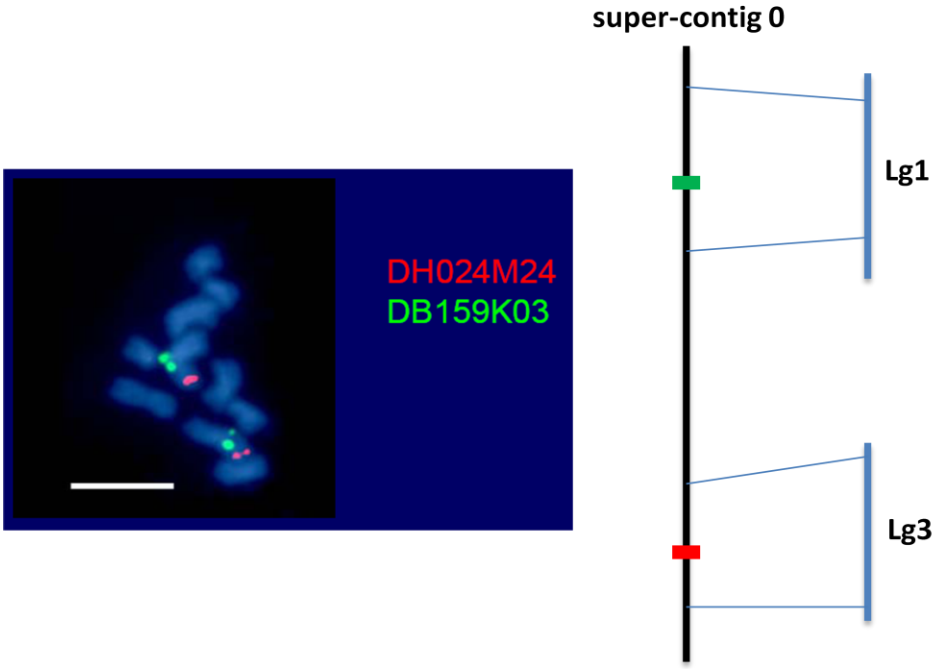


Figure 2.13: Locating linkage groups 1 and 3 onto chromosomes.

Epifluorescence microscopy image showing the locations of two labelled BAC probes on Brachypodium chromosomes (blue). One BAC probe is taken from super-contig 0 where linkage group 1 aligns and one from where linkage group 3 aligns. The probes are labelled with red and green fluorophores and the BAC names are shown on the image in the colour of the probe. The diagram on the right shows the relative location of each probe on super-contig 0 as a green or red block coloured relative to the fluorophore colour. (scale bar = 5µm)

2.5.3.4 Additional anchoring of Build 2 linkage groups to chromosomes using FISH

In order to cytogenetically anchor additional linkage groups from Build 2, I selected a further 16 BACs from regions within the mid-point super-contigs where linkage groups aligned (Table 2.7).

Linkage group	Super-contig	start (bp)	end (bp)	start BAC	end BAC	Brachypodium chromosome
1	0	18,562,705	4,844,784	DB096E17	DB156I06	1S
1	1	24,612,535	10,759,094	DB091E04	DB086H24	1L
1	5	9,545,763	20,161,615	DB041E07	DB019K13	3L
1	3	17,287,869	2,259,858	DB060H11	DB024K15	3S
2	7	7,092,044	13,115,900	DB087J23	DB006P15	4L
2	9	4,408,046	4,971,939	DH009L18	DH040M22	5S/5L
3	0	26,156,786	36,924,347	DH033N14	DB069C07	1S
4	7	5,908,251	17,257,560	DH002F14	DH009H14	4L

Table 2.7: Aligning further linkage groups to chromosomes.

BAC clones selected for FISH based on alignments between the Build 2 linkage groups and mid-point super-contigs. A BAC was selected from the start and end of each alignment. The last column shows the resulting Brachypodium chromosome assignment for the BACs (S=Short arm, L=Long arm)

In regions where a linkage group aligned to a super-contig, a BAC was selected from each end of the aligning region to delineate it resulting in a two BACs per alignment. These BAC probes were sent for FISH, alongside anchor clones of known location (Hasterok *et al.* 2006) so that the chromosome to which the BAC probe hybridised could be identified.

The results from this experiment (Table 2.7) indicate that unique chromosomal locations can be assigned to super-contigs 0, 1, 3, 5, 7 and 9 (1S, 1L, 3S, 3L, 4L, and 5 respectively). Super-contigs 0 and 1 represent different arms of chromosome 1 and super-contigs 3 and 5 represent different arms of chromosome 3. However, assigning chromosomal locations to linkage groups is not so straightforward. Linkage group 1 appears to align to chromosomes 1 and 3 indicating false linkage within this group. A similar observation can be made for linkage group 2 which appears to align to chromosomes 4 and 5. Linkage groups 3 and 4 can be assigned to unique chromosome arms (1S and 4L).

2.5.3.5 Comparison of Build 3 linkage groups to the pseudo-molecules of the final assembly

Ninety-five of the 139 genetic markers showed good BLAST hits to the Brachypodium pseudo-molecules (E-value < $1e^{-50}$). Fourteen out of the 20 linkage groups can be anchored by multiple markers and all but one linkage group (r) can be anchored by a single marker (Figure 2.14).

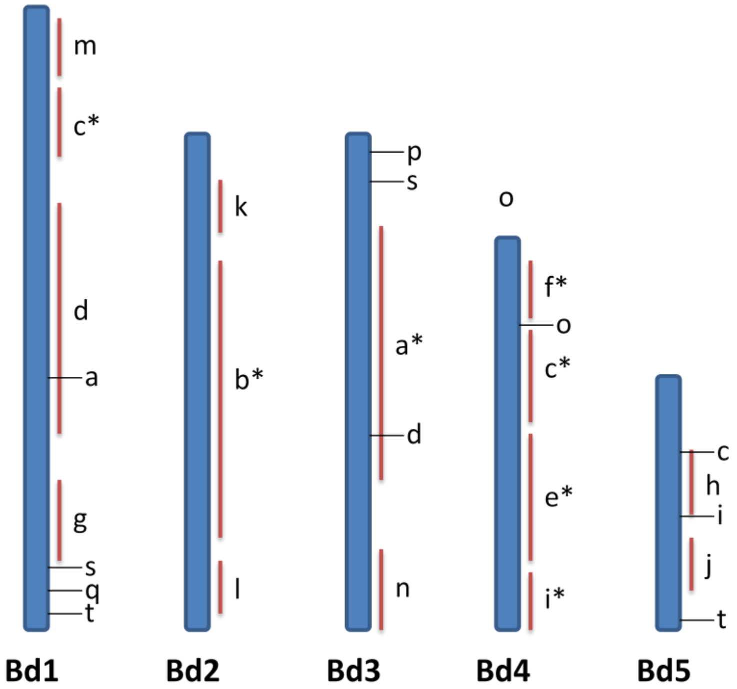


Figure 2.14: Genetic linkage groups from Build 3 (a-t) aligned to the psuedo-molecules of the final assembly. Linkage groups are labelled a-t and psuedo-molecules are shown in blue (Bd1-5). Linkage groups anchored by multiple markers are shown as red lines indicating the region of alignment. Linkage groups labelled with an asterisk show some small local rearrangement in the order of the markers between the maps. Linkage group positions shown as black lines indicate the linkage group is anchored by a single marker.

Linkage groups anchored with multiple markers show a high level of synteny to the sequence assemblies with a few small rearrangements in marker order. Markers on the upper portion of linkage group ‘c’ show synteny to Bd1 and markers on the lower portion show synteny to Bd4 indicating that this linkage group probably contains a false linkage. In addition, the presence of single markers from linkage groups that are anchored elsewhere by multiple markers (for example linkage group ‘a’ is anchored to Bd3 by multiple markers but a single marker is located on Bd1) indicates other possible discontinuities in the linkage groups.

2.5.3.6 Comparison of Build 3 linkage groups to rice chromosomes

Seventy out of 139 markers showed significant (<1e^{-15}) hits to the rice genome and were used to assess macrosynteny between the rice genome and the Brachypodium linkage groups (Figure 2.15). Eight of the 20 linkage groups show synteny to a single rice chromosome (b, e, i, j, k, l, n and p) whilst others have more complex relationships where markers from a single linkage group hit multiple rice chromosomes, for example, linkage group ‘a’. Linkage group b shows the largest region of collinearity to a single rice chromosome with 9 markers collinear to rice chromosome 1. Markers TR024 and Wheat1S at the bottom of this linkage group also have putative orthologues on rice chromosome 1 but an apparent rearrangement has disrupted the order of these markers relative to the others. Many local rearrangements in overall syntenic regions are also observed which may indicate inversions that have occurred between the species (on linkage groups b, d, f, I, j, k and l). In addition, four cases are found where adjacent loci on the linkage groups hit the same rice gene (on linkage groups b, c, g and n).

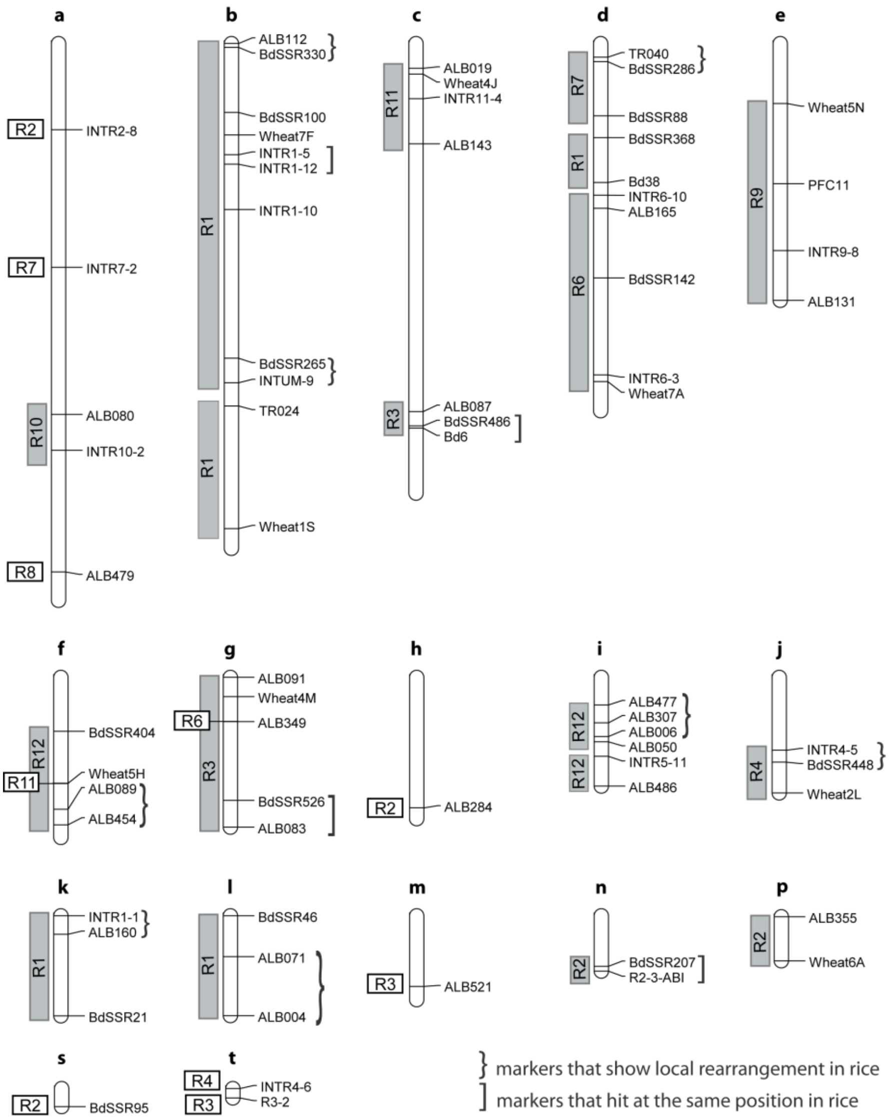


Figure 2.15: Inferred macrosynteny between Brachypodium linkage groups and the rice genome.
Linkage groups are labelled a-t. The open and shaded boxes to the left of each linkage group identify the chromosome locations of the best BLASTN hits in rice for the listed Brachypodium molecular markers (R1 = rice chromosome 1; R2 = rice chromosome 2, etc.) The open boxes next to linkage groups f and g identify the rice chromosome for individual loci associated with different rice chromosomes. Curly brackets indicate local rearrangement of markers relative to rice and square brackets identify marker pairs that BLASTN analysis revealed had highest hits to the same rice gene. Only Brachypodium markers with BLASTN hits giving E-values of less than 1e^{-15} against the rice genome sequence are displayed.

2.6 Physical mapping and validation of WGS sequence assemblies

2.6.1 Introduction

A physical map consists of a linear array of genome fragments that have been cloned into vectors. The order of cloned fragments is determined by detecting overlaps between the fragments. Restriction enzyme fingerprinting was developed to construct physical maps in *C.*

elegans and *S. cerevisiae* (Coulson *et al.* 1986; Olson *et al.* 1986). In this technique each clone is digested with restriction enzymes to yield a number of size fragments which are detected by autoradiography producing a 'fingerprint'. Overlapping fingerprints are used to cluster the clones and build up contigs. Modern physical mapping methods use the high-information content fingerprinting (HICF) method developed for the human genome sequencing project (Marra *et al.* 1997). HICF uses large-insert clones (PACs or BACs) which when digested, yield more fragments compared to the smaller cosmid libraries used in previous physical mapping projects. In addition, fluorescent dyes are used to resolve the different sized fragments using an automated DNA analyser rather than the radioactive labelling and polyacrylamide gel electrophoresis previously used (Gregory *et al.* 1997). The FingerPrinted Contigs software (FPC) is used in conjunction with HICF to compare clone fingerprints and assemble clones into contigs based on a probability of coincidence calculated from clones containing similar fragment sizes (Soderlund *et al.* 1997).

In order to increase the accuracy of BAC fingerprinting, a multiplexing strategy was developed where each BAC clone in a library is digested by three pairs of restriction enzymes and visualised by three fluorescent dyes (Ding *et al.* 1999). The rational behind this method is that the probability of two clones overlapping is increased if shared fragment sizes are identified by all three restriction enzymes. This led to the development of a multiplexing technique where each BAC is simultaneously digested with five restriction endonucleases, one that cuts at a 4 bp site and four that cut at a 6 bp site. The restriction sites cut by the 6 bp endonucleases are labelled using four different fluorescent dyes and the fragments are sized by capillary electrophoresis (Luo *et al.* 2003). This technique is marketed by Applied Biosystems and called the SNaPshot Multiplex System. Subsequently, the GeneMapper software is used to size the fragments. As FPC was originally developed to assemble contigs from fingerprints on polyacrylamide gel the output file from GeneMapper containing the fluorescently labelled and sized peaks from SNaPshot fingerprinted data is converted to one that resembles a gel file using GenoProfiler (You *et al.* 2007).

Genome-wide physical maps provide a foundation for gene isolation by positional cloning and are the basis for a hierarchical genome sequencing strategy as the minimum tiling path of clones provides the template for genome sequencing and subsequent reassembly. A WGS approach to genome sequencing can also benefit from a physical map, especially for complex genomes as the map can be used to assess the consistency of the sequence contigs emerging from the assembly process and to align and orientate them (Meyers *et al.* 2004). Physical maps can also be used to facilitate the analysis of genome structure and to compare the genome to other genomes. Physical maps have been developed for many plant species including

Arabidopsis, the original model plant (Mozo *et al.* 1999), and rice, the model grass (Chen *et al.* 2002) as well as other grass genomes, for example, sorghum (Klein *et al.* 2000) and maize (Wei *et al.* 2007).

A genome-wide physical map of Brachypodium is a key genomic resource to study the structure of the genome and to enable comparison with other grass genomes and is the first pooid grass to be physically mapped. A Brachypodium BAC-based physical map containing 671 contigs covering 410 Mb was recently published (Gu *et al.* 2009). The BACs from this physical map were end-sequenced and aligned to the rice genome and to wheat deletion bins using gene tags identified from the BES.

A second Brachypodium physical map was constructed by MF and myself using two BAC libraries made from the Brachypodium community standard line Bd21 (Garvin *et al.* 2008). Physical map contigs were built using FPC (Soderlund *et al.* 1997) and aligned to the super-contigs of the mid-point assembly to assess the emerging WGS sequence assemblies. Although the two BAC libraries used to construct the physical map were also used in the WGS sequencing of the Brachypodium genome, no physical mapping data was included in the WGS assembly so these comparisons provided an independent method to assess the sequence assemblies. Upon release of the final genome assembly, the physical map contigs were aligned to the pseudo-molecules to assess the final assembly. The contigs of our physical map were integrated to the genetic map to assess the consistency between these two genomic resources (Garvin *et al.* 2010).

Anchoring the physical map contigs to Brachypodium genome sequence also facilitated the development of a novel *in silico* BAC selection and screening process used to identify BACs containing low proportions of repetitive DNA for FISH experiments. At each stage of the sequencing project, the sequence assemblies released by the JGI were validated by anchoring them to the Brachypodium cytogenetic map. This anchoring provided a method to determine whether the sequence assemblies constructed *in silico* accurately reflected the underlying structure of the Brachypodium chromosomes. The integrated physical and genetic map was anchored to the Brachypodium cytogenetic map by FISH experiments that were performed in collaboration with RH. The integrated physical, genetic and cytogenetic map has been published (Febrer *et al.* 2010).

2.6.2 Methods

2.6.2.1 BAC library construction, fingerprinting and end-sequencing

Two BAC libraries were produced by MF from the Brachypodium community standard line Bd21 (Garvin *et al.* 2008). These libraries were also sequenced by the JGI as part of the genome sequencing project. Nuclear DNA was partially digested with restriction enzymes *Hin*dII and *Eco*RI and ligated into cloning vectors pAGIBAC1 and pIndigoBAC536 *Swa*I respectively. Each library consisted of 36,864 clones, the average insert size of the *Hin*dIII library (BD_ABa) was 128 kb and the average insert size of the *Eco*RI library (BD_CBa) was 124 Kb (Table 2.8). In total, the libraries represented 9.7x genome coverage based on a genome size of 272 Mb (International Brachypodium Initiative 2010).

Library name	Restriction enzyme	Average insert size	Number of BACs	Number of BACs fingerprinted	Number of BES	Genome coverage
BD_ABa	*Hin*dIII	128 kb	36,864	15,565	34,001	4.5X
BD_CBa	*Eco*RI	124 kb	36,864	14,947	24,893	4.5X
Total		126 kb	73,728	30,512	58,894	9.7X

Table 2.8: Details of the two BAC libraries used to construct the Brachypodium physical map.

Half of the BACs in each library (a total of 36,864 clones) were fingerprinted at the Arizona Genome Institute using the High-Information Content Fingerprinting (HICF) method followed by SNaPshot reagent labelling (Luo *et al.* 2003). This resulted in 30,512 (82.7 %) clones with fingerprints suitable for contig assembly. The BACs in each library were also end-sequenced to allow anchoring of the contigs to Brachypodium genomic sequence assemblies using SyMAP (Soderlund *et al.* 2006). This resulted in 58,894 BAC-end sequences with a combined length of 41.1 Mb.

2.6.2.2 Initial build of the physical map

Fingerprinted BACs were assembled into contigs using the FingerPrinted Contigs (FPC) software version 8.9 (Soderlund *et al.* 1997). These contigs were merged manually by identifying fingerprint overlaps at a lower stringency than was initially used to build contigs. The initial build was created by aligning these contigs to the largest 17 super-contigs of the mid-point assembly in SyMAP using the BAC end-sequences (BES) to validate the assembled physical map contigs and to orient them according to the mid-point assembly. At this point, any contigs that did not align to the mid-point assembly super-contigs were broken and reassembled.

2.6.2.3 Final build of the physical map

Upon release of the final Brachypodium assembly the FPC-derived contigs were aligned to the five pseudo-molecules using SyMAP. The BioPerl FPC module was used to export the physical mapping data from the FPC database to GFF files which were then loaded into the genome browser GBrowse for display on www.modelcrop.org (Pampanwar *et al.* 2005). GBrowse provides a more user-friendly interface for browsing physical mapping data and in addition allows this data to be made available to the wider community via a website.

2.6.2.4 Anchoring the physical map to the genetic map

Two complimentary approaches were used to anchor the physical map contigs to the genetic map (Garvin *et al.* 2010); first I used BLASTN to locate marker sequences from the genetic map within the BES of the BACs used in the physical map. An E-value cut-off of $1e^{-50}$ was used to select good hits. Identified BES were located on a BAC clone which was either assembled into a contig by FPC or was not assembled into a contig but could be located on the Brachypodium sequence assembly via BES BLAST hit from SyMAP. I also designed additional molecular markers by identifying simple sequence repeats or SSRs (Thurston and Field 2005) from the BES. Markers were selected from the beginning, middle and end of each physical map contig and tested for amplification by PCR in the mapping parents used to generate the original map (Bd21 and Bd3-1). The PCR products from markers that amplified in the mapping parents were tested for polymorphism by direct sequencing before screening against a subset of the original mapping population. MF performed this phase of the anchoring.

2.6.2.5 Validation of pre-release super-contigs using FISH

Once the physical map contigs were aligned to sequence assemblies, BACs could be selected from specific positions in the sequence contigs and used as FISH probes to identify the likely position of the super-contig on the physical Brachypodium chromosomes (Figure 2.16). This also provided an additional method to validate the sequence assemblies emerging from the JGI. BACs were selected from each end of the super-contigs and from positions along their lengths where joins between mid-point super-contigs had been made. These BACs were sent for FISH and, in most cases, reference BACs of known location (Hasterok *et al.* 2006) were used alongside the probes to facilitate the chromosomal assignment of each super-contig.

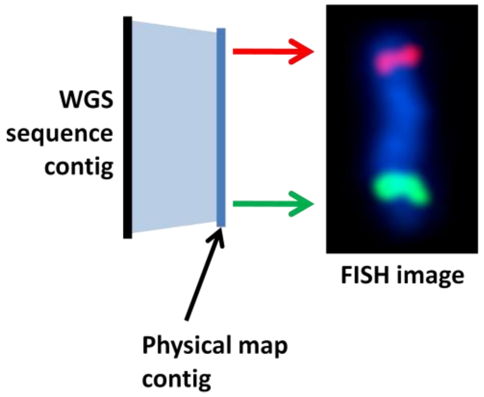


Figure 2.16: Locating sequence contigs onto chromosomes using FISH.
The principle of using the alignment between a WGS sequence contig (black line) and a physical map contig (blue line) to select BACs (green and red arrows) from the physical map to be used as FISH probes. The resulting epifluorescence images indicate the position of the physical map contig (and hence the sequence contig) on the chromosome and indicate whether the sequence contig has been robustly constructed.

2.6.2.6 Screening BACs for Repetitive DNA

It was observed that a small number of BACs selected as probes for FISH gave multi-locus signals. This occurs when a BAC contains a high proportion of repetitive DNA such as centromeric sequence or transposable elements as the BAC will hybridise to multiple places on the chromosomes. Hybridisation to a single locus is critical if one is to determine the position of the BAC on a single chromosome and due to time and financial constraints we needed to maximize the probability of selecting good BAC probes. Having the physical map contigs anchored to the genome sequence assembly facilitated the development of a Perl script to assess the repeat content of a BAC based on its predicted location in the assembly (Supplementary information 7). My script first extracts the genomic coordinates of the BAC-end reads for the BAC from the SyMAP MySQL database using the Perl DBI library. These coordinates are used to retrieve the predicted sequence of the BAC from a FASTA sequence file of the assembly, then RepeatMasker (Smit *et al.* 1999) is run on this sequence using a *Pooideae*-specific repeat element database provided by Klaus Mayer at the Munich Information Centre for Protein Sequence (MIPS). RepeatMasker masks any repetitive elements in the sequence with Ns so by counting the number of Ns and comparing this to the total length of the sequence, the percentage of masked bases can be calculated and used to estimate the repeat content of the BAC. The script was tested using BACs previously used as probes for FISH.

2.6.2.7 Validation of the final assembly using FISH

Three methods were used to validate the final assembly. Firstly, the FISH hybridisation results from previous experiments were applied to the pseudo-molecules and more BACs identified from physical map contigs aligned to pseudo-molecules that had not yet been anchored to provide a complete picture of alignments between the physical map contigs, the final sequence assembly and the Brachypodium cytogenetic map. Secondly, BACs with a low repetitive DNA content were identified from physical map contigs aligning to each pseudo-molecule and combined to create five chromosome-specific pools. Each pool of BACs was hybridised simultaneously, alongside reference BACs (Hasterok *et al.* 2006), to confirm that each group of physical map contigs, and hence each pseudo-molecule, aligned to its respective chromosome. Finally, in order to give a higher resolution picture of continuity between the physical map contigs, the sequence assemblies and Brachypodium chromosomes, larger pools of BACs were selected from physical map contigs aligning to each arm of the pseudo-molecules and applied to meiotic chromosome spreads (Jenkins and Hasterok 2007) to highlight or 'paint' large chromosomal regions. Plant genomes exhibit an abundance of highly dispersed DNA repeats which normally precludes this type of experiment but this technique has been used successfully to paint chromosome 4 of *Arabidopsis* (Lysak *et al.* 2001). For accurate hybridisations it is essential to select probe BACs containing low proportions of repetitive DNA. Therefore, BACs were selected to provide even coverage along chromosome arms and BACs estimated to consist of more than 33% repetitive DNA were excluded from the pools using the repeat estimator script. This stringent criterion allowed for the selection of pools large enough to highlight large regions whilst at the same time excluding BACs that may hybridise to multiple loci when applied to the chromosomes. The clones were applied to chromosome spreads either as whole pools to paint the entire chromosome arm, or as sub-pools to highlight small areas within a single chromosome arm.

2.6.3 Results

2.6.3.1 Initial build of the physical map

Assembling the fingerprinted clones using FPC produced 208 contigs assembled using 30,195 (99%) clones, with 317 (1%) clones remaining as singletons (Table 2.9).

	Initial build		**Final build**
	Automatic contig assembly	**After manual editing using the mid-point assembly**	**After manual editing using the final assembly**
Number of clones fingerprinted	30,512	30,512	30,512
Number of clones used for map assembly	30,195	26,800	26,800
Number of singletons	317	472	475
Number of contigs	208	35	26
Contigs containing			
> 1000 clones	1	9	9
999-800 clones	1	6	7
799-600 clones	5	6	3
599-400 clones	9	2	2
399-200clones	24	6	1
199-100 clones	45	4	3
< 99 clones	123	2	1
Unique bands of the contigs	270,216	253,114	252,810
Physical length of the contigs (Mb)	324.26	303.74	303.37

Table 2.9: Features of the Brachypodium physical map.

Details are shown at the three stages of assembly, the automatic contig assembly stage, the comparison with the mid-point assembly and the comparison with the final assembly.

Aligning the physical map contigs from the automated build to the 17 largest super-contigs of the mid-point assembly using BES indicated good congruence between the assemblies. An example of such an alignment is shown in Figure 2.17 where BACs in physical map contig 9 are aligned to a region of super-contig 1 (called chrB in the SyMAP software). A high level of synteny is shown between the two contigs indicating a high level of accuracy in both the assembly of the physical map and the sequence assembly. FPC constructs contigs in a random manner so alignment to the sequence assemblies enabled the physical map contigs to be oriented and ordered. Figure 2.18 shows a high-level view where six physical map contigs (6-11) are aligned end-to-end with super-contig 1 of the mid-point assembly.

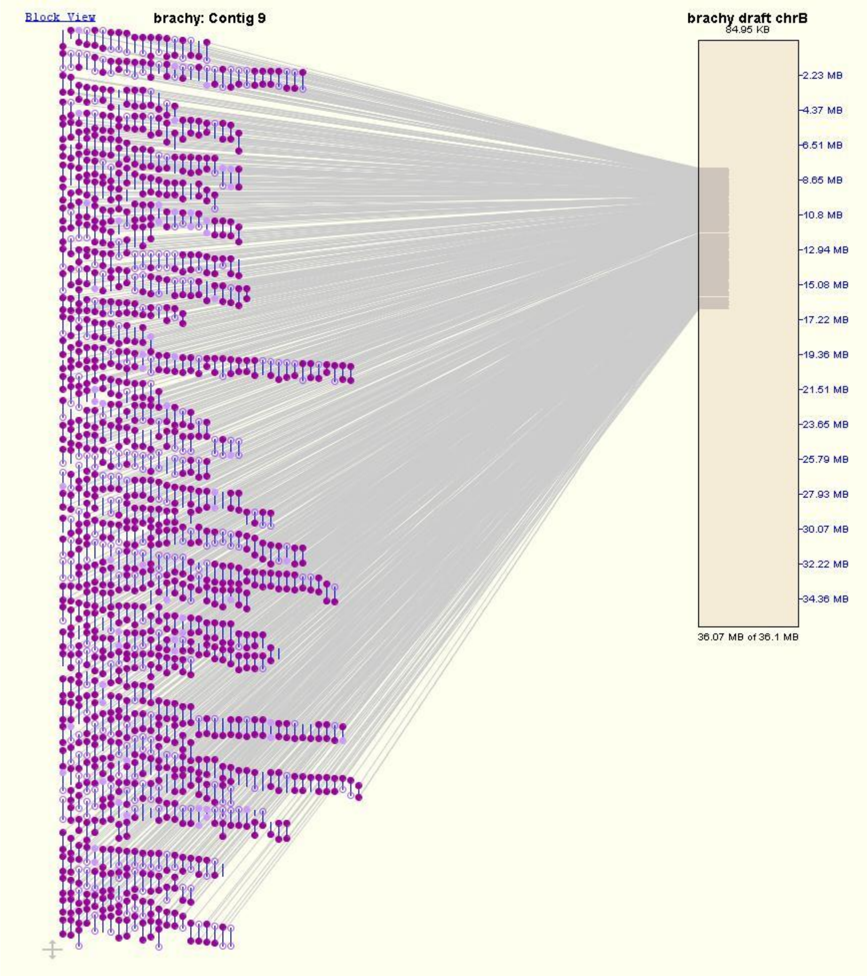


Figure 2.17: Alignment between physical map contig 9 and super-contig 1 using SyMAP.

The physical map contig is comprised of BACs shown as blue lines with BES shown as purple circles, super-contig 1 (chrB) of the mid-point assembly is shown as a beige rectangle. Grey lines indicate BLAST hits between the BES in the physical map contig and the super-contig. The physical map contig aligns to the central region of the super-contig.

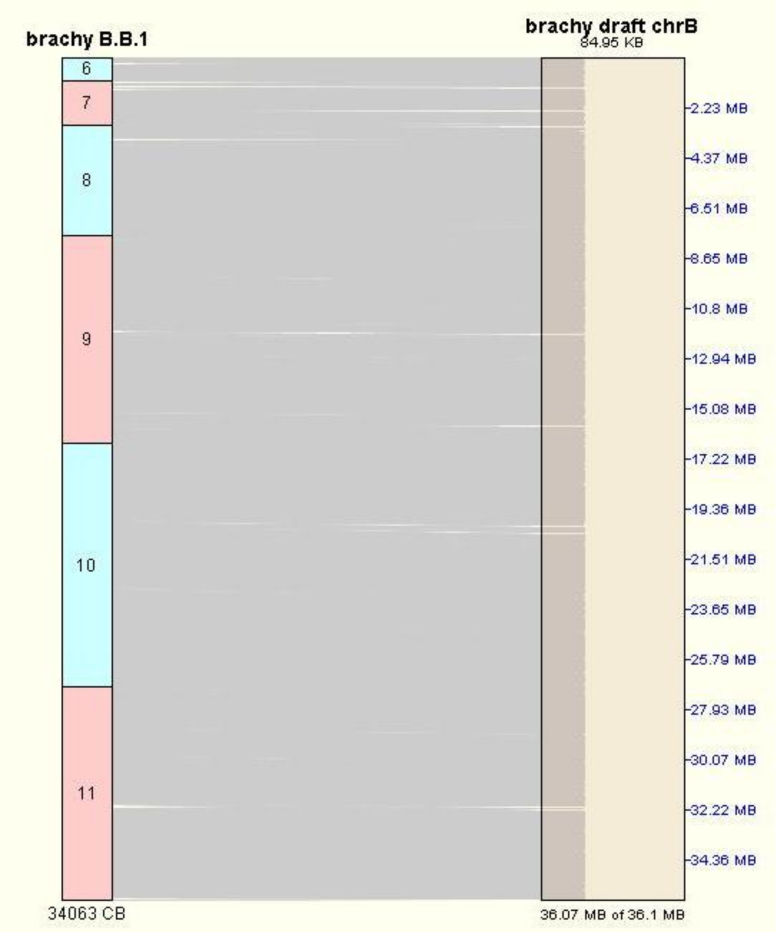


Figure 2.18: Alignment of six FPC-derived contigs to super-contig 1 using SyMAP.
The FPC-derived contigs are shown as pink and blue boxes, to super-contig 1 (chrB) of the mid-point assembly is shown as a beige rectangle. The grey area shows the high concentration of individual BLAST hits between the BES in the FPC contigs and the super-contig from the mid-point assembly.

Aligning the physical map contigs to the sequence assemblies also suggested joins between the super-contigs of the mid-point assembly where a physical map contig aligned to two sequence contigs. Three cases were identified where the clones from one end of a physical map contig had strong sequence alignments with one super-contig and clones from the opposite end of the same contig showed good alignments with another super-contig. Figure 2.19A shows contigs 18, 19 and the top portion of contig 20 aligning to super-contig 4 (SC4) and the bottom half of contig 20 aligning to super-contig 13 (SC13). This indicates that super-contigs 4 and 13 can be merged. A similar situation is shown in Figure 2.19B where contig 30 aligns to both super-contig 10 and super-contig 9. Figure 2.19C indicates that super-contig 11 can be merged with super-contig 14 by virtue of good alignments to physical map contig 34. In this way, constructing the physical map helped to improve the existing sequence assembly by creating larger super-contigs.

At the end of this process, the physical map comprised 35 contigs assembled from 26,800 (87.8 %) clones with 472 (1.5 %) clones remaining as singletons (Table 2.9). These contigs were aligned to 11 original super-contigs from the mid-point assembly as well as three larger super-contigs created by merging super-contigs.

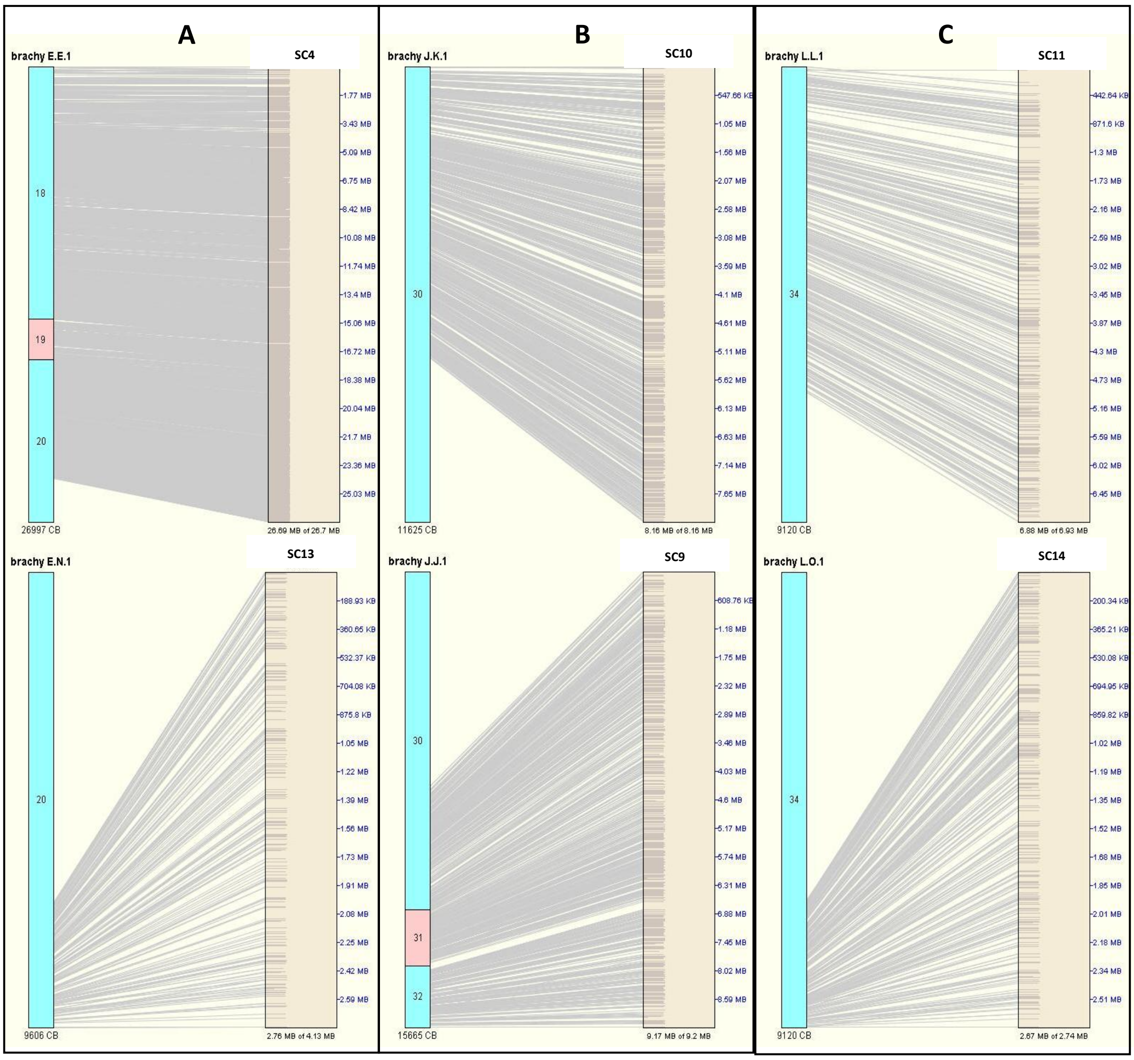


Figure 2.19: Merging super-contigs of the mid-point assembly.

The super-contigs are merged by alignment to physical map contigs using BAC end sequences. Physical map contigs are shown as blue and pink boxes to the left of each panel, and sequence contigs are shown as beige boxes. The grey lines show individual BES hits to the sequence assemblies. Panel A shows BACs from the upper portion of physical map contig 20 align to super-contig 4 (SC4) and BACs from the lower portion of the contig align to super-contig 13 (SC13) and indicates that these two contigs can be merged. Panel B shows BACs from physical map contig 30 aligning to super-contigs 10 and 9 (SC10 and SC9) indicating these super-contigs can be merged. Panel C shows BACs from physical map contig 34 aligning to super-contigs 11 and 14 (SC11 and SC14) indicating these super-contigs can be merged.

2.6.3.2 Final build of the physical map

Aligning the FPC-derived contigs to the pseudo-molecules of the final WGS assembly again showed good congruence between the physical map and the sequence assemblies and also provided independent validation of all the WGS sequence assemblies. These alignments resulted in 26 physical map contigs anchored and oriented according to the five pseudo-molecules of the final assembly (Figure 2.20). The contigs contained 26,800 (87.8%) clones, with 475 (1.55%) remaining as singletons (Table 2.9). The final size of the physical map of BACs was 303.4 Mb based on an approximation of the band size of fingerprints.

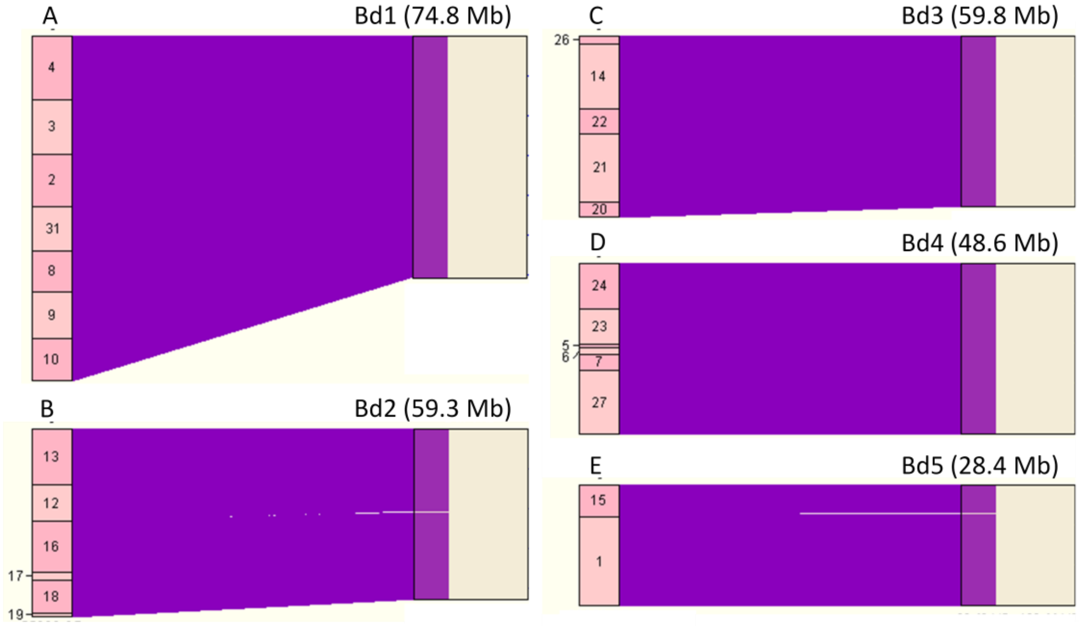


Figure 2.20: Physical map contigs aligned to the final assembly.
Physical map contigs are shown in light and dark pink and are aligned and oriented to the pseudo-molecules of the final assembly (Bd1-5), shown as beige rectangles. The contigs are grouped together to form five assemblies of physical map contigs (A-E).

2.6.3.3 Anchoring the physical map to the genetic map

BLASTN identified 29 markers from the genetic map that could be located on the BES from the physical map BAC clones (E-value < $1e^{-50}$) and 23 of these where from clones assembled into physical map contigs so could be located onto the sequence assembly (Table 2.10). SyMAP alignment of the BES to the sequence assembly was used to place the remaining 6 BACs.

Fourteen linkage groups could be anchored to 12 unique physical map contigs using this method. All but two linkage groups are anchored to physical map contigs aligning to the same pseudo-molecule. Linkage group 'a' is anchored to physical map contigs 21 and 22 (aligned to Bd3) as well as contig 31 (aligned to Bd1). Linkage group 'c' is anchored to physical map contig 4 (aligned to Bd1) as well as contigs 23 and 24 (aligned to Bd4).

				Predicted chromosome	
Marker	**Linkage group**	**BES hit**	**Contig**	**from contig alignment**	**from BES blast**
ALB329	a	a0017B20.r	31	Bd1	Bd1:46674554..46832512
ALB346	a	a0004A14.r	21	Bd3	Bd3:39998994..40158254
ALB502	a	a0003B19.r	22	Bd3	Bd3:14467921..14508862
ALB112	b	a0028A07.r			Bd2:52341444..52542965
ALB376	b	a0047D21.f			Bd2:34360649..34544003
INTR1-12	b	a0002F05.f	18	Bd2	Bd2:47264595..47444884
INTR1-5	b	a0002F05.f	18	Bd2	Bd2:47264595..47444884
INTR1-10	b	a0044K02.f	16	Bd2	Bd2:44170168..44300832
BdSSR330	b	a0011E13.r	18	Bd2	Bd2:52343100..52542965
ALB181	b	a0047D21.f			Bd2:34360649..34544003
ALB087	c	a0033O05.r	4	Bd1	Bd1:9863517..10021969
ALB143	c	a0010G09.r	23	Bd4	Bd4:15503867..15684692
ALB266	c	a0029E21.r	24	Bd4	Bd4:9176853..9359162
INTR6-10	d	a0004B16.r	2	Bd1	Bd1:33880013..34040483
BdSSR142	d	b0011N13.f	31	Bd1	Bd1:46605642..46750089
ALB184	e	a0013J03.f			Bd4:36679202..36865740
Wheat5H	f	a0007O24.f	24	Bd4	Bd4:5235135..5299208
ALB349	g	a0021G20.f			Bd1:62436810..62593244
ALB311	h	a0015J02.r	1	Bd5	Bd5:18580275..18736126
ALB006	i	b0016N17.f	27	Bd4	Bd4:46592449..46762416
ALB307	i	a0045G06.r	27	Bd4	Bd4:46843082..46994022
ALB477	i	a0045G06.r	27	Bd4	Bd4:46843082..46994022
ALB486	i	a0022E14.f	27	Bd4	Bd4:46773621..46913354
ALB160	k	a0028F07.r			Bd2:8154796..8305485
ALB071	l	b0039I21.f	18	Bd2	Bd2:57118446..57289793
BdSSR46	l	b0015G03.r	19	Bd2	Bd2:58185841..58333092
ALB257	m	a0040G20.r	4	Bd1	Bd1:4839832..4990165
ALB163	s	a0017N13.r	14	Bd3	Bd3:3383484..3526031
INTR4-6	t	a0021B08.r	1	Bd5	Bd5:27958933..28153690

Table 2.10: Anchoring the genetic map to the physical map using genetic marker sequences. Marker sequences that give good BLAST hits to BES are used. The marker name is shown alongside the linkage group it comes from, the BES it hits and the chromosome location either predicted by the contig containing the BES or the BES hit to the assembly from SyMAP.

In order to increase the number of potential anchor points, an additional 29 SSR markers were identified from the beginning, middle and end of each physical map contig (Supplementary information 8). Twenty-four of these were successfully genetically mapped (Table 2.11).

Marker name	BAC	Contig	Linkage group	Pseudo-molecule
BDBES_SSR0108_1378	b0047H14_f	21	a	3
BDBES_SSR0108_1031	a0019A19_r	16	b	2
BDBES_SSR0108_1048	b0041L12_f	17	b	2
BDBES_SSR0108_0223	a0029E10_f	4	c	1
BDBES_SSR0108_0332	a0018K11_r	6	c	4
BDBES_SSR0108_1970	b0040E12_f	24	c	4
BDBES_SSR0108_1917	a0003G24_f	31	d	1
BDBES_SSR0108_0371	a0036L13_f	7	e	4
BDBES_SSR0108_1673	a0037L08_f	27	e	4
BDBES_SSR0108_1455	a0002I07_f	24	f	4
BDBES_SSR0108_2004	b0033O01_f	24	f	4
BDBES_SSR0108_1819	b0003P21_f	1	g	5
BDBES_SSR0108_0014	b0033D02_f	1	h	5
BDBES_SSR0108_0034	a0002N04_f	1	h	5
BDBES_SSR0108_0570	b0003N09_f	1	h	5
BDBES_SSR0108_0934	a0013L16_r	15	h	5
BDBES_SSR0108_1786	b0029E05_r	1	h	5
BDBES_SSR0108_1779	b0041P08_r	1	j	5
BDBES_SSR0108_0712	a0044B16_r	13	k	2
BDBES_SSR0108_0719	b0005P05_r	13	k	2
BDBES_SSR0108_1164	a0039D05_f	19	l	2
BDBES_SSR0108_1169	a0039O07_f	20	n	3
BDBES_SSR0108_1189	a0010O04_f	20	n	3
BDBES_SSR0108_0523	b0003G22_r	10	q	1

Table 2.11: Anchoring the genetic map to the physical map using SSR markers.

Table shows markers successfully genetically mapped, the clone each was identified from, the contig where the BAC is found and the linkage group in which the marker resides. From the contig position we can predict the pseudo-molecule to which this linkage group aligns.

These markers anchored 15 unique physical map contigs to 13 linkage groups. Six linkage groups (a, d, g, j, l and k) were anchored to a physical map contig by a single anchor-point and the remaining seven were anchored by multiple anchor-points. Three linkage groups (f, k and n) are anchored by multiple anchor-points to the same contig and three (b, e and h) are anchored by multiple anchor-points to different physical map contigs that align to the same pseudo-molecule indicating good congruence between the physical and genetic maps. A single linkage group (c) was ambiguously anchored to physical map contigs 6 and 24 (aligned to Bd4) as well as contig 4 (aligned to Bd1).

Combining the results from the two anchoring methods allows 17 linkage groups to be anchored to physical map contigs (Table 2.12). Seven linkage groups (a, b, c, d, f, h and l) are anchored to physical map contigs by both methods and in all but two cases (linkage groups 'a' and ‘c’) contigs aligning to the same pseudo-molecule are anchored to. Linkage group ‘a’ cannot be unambiguously anchored although it is more likely to be anchored to contigs aligning to Bd3 than Bd1 as both methods support this anchoring. In addition, the anchoring of linkage groups to pseudo-molecules by aligning the marker sequences directly (Figure 2.14) places linkage group ‘a’ on Bd3. Both anchoring methods indicate that linkage group ‘c’ is anchored to physical map contigs aligning to Bd1 and Bd4 so this linkage group cannot be unambiguously anchored. Two linkage groups (e and k) are anchored to BACs not assembled into contigs but their BES aligns them to pseudo-molecules. Both of these of these linkage groups are also genetically anchored to contigs aligning to the same pseudo-molecules. Linkage group ‘g’ is anchored to Bd1 by virtue of a BES BLAST hit and to Bd5 by a single genetic marker so it cannot be uniquely anchored. The seven remaining linkage groups are anchored either *in silico* or genetically to a physical map contig.

	***In silico* anchoring**		**Genetic anchoring**	
Linkage group	**Contigs**	**Pseudo-molecules**	**Contigs**	**Pseudo-molecules**
a	31, 21, 22	1 or 3	21	3
b	16, 18	2	16,17	2
c	4, 23, 24	1 or 4	4, 6, 24	1 or 4
d	2, 31	1	31	1
e		4	7, 27	4
f	24	4	24	4
g		1	1	5
h	1	5	1, 15	5
i	27	4		
j			1	5
k		2	13	2
l	18, 19	2	19	2
m	4	1		
n			20	3
s	14	3		
t	1	5		
q			10	1

Table 2.12: Comparison between *in silico* and genetic methods of anchoring linkage groups.

Seventeen linkage groups were anchored to physical map contigs and pseudo-molecules. Seven linkage groups are anchored by both methods and all but two cases (linkage groups ‘a’ and ‘c’) contigs aligning to the same pseudo-molecule are anchored to.

2.6.3.4 Validation of the pre-release super-contigs using FISH

Seventeen BAC clones were selected to be used as FISH probes (Table 2.13). These BACs were taken either from the ends of the pre-release super-contigs or from points along their length where joins were predicted between mid-point super-contigs (Figure 2.21). For example, super-contig 0 from the pre-release assembly comprises super-contigs 0 and 16 from the previous mid-point assembly so probes were chosen from each end of the 38 Mb super-contig as well as a single probe from the locus at 2 Mb where the join between the mid-point super-contigs occurs.

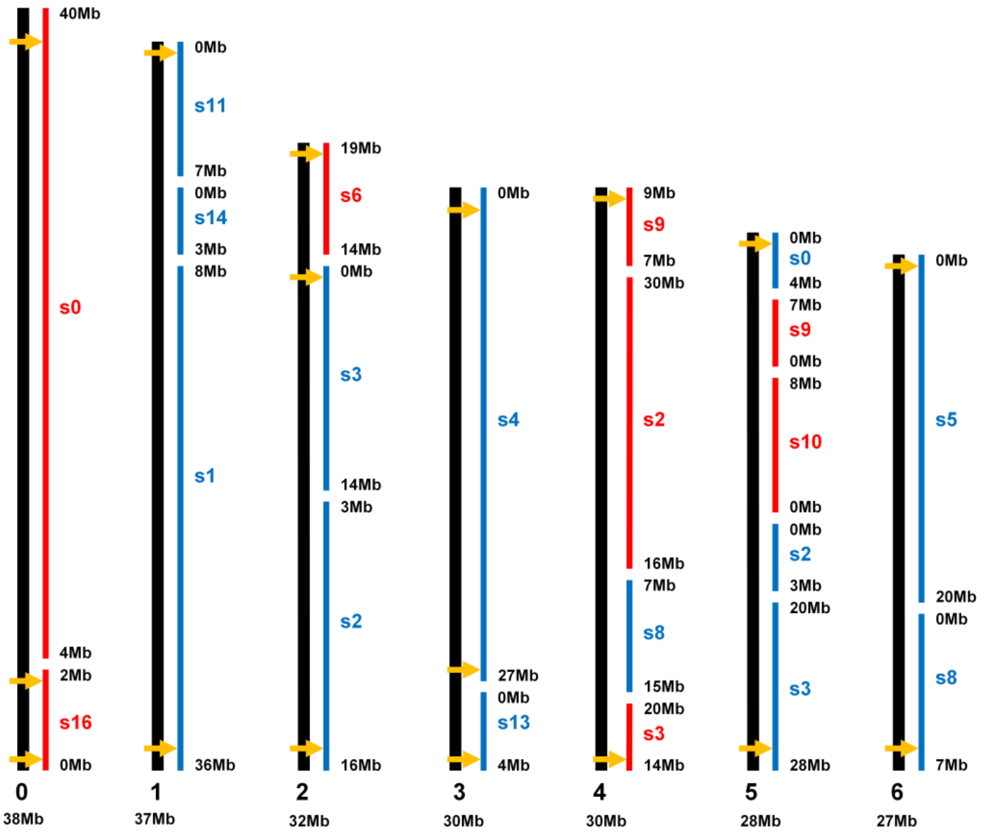


Figure 2.21: Position of BAC clones selected as FISH probes on the pre-release super-contigs.
The 7 largest super-contigs are shown as black lines (labelled 0-6) with BAC clones as yellow arrows. Probes were selected from the ends of super-contigs and from positions within them next to joins between mid-point super-contigs (shown as red and blue lines; labelled s0-s14).

Twelve probes could be assigned unambiguously to a Brachypodium chromosome arm (Table 2.13). One probe (a0015M04) was assigned to the short arm of either chromosome 3 or chromosome 5 and two probes (a0039J03 and b0001B16) hybridised to multiple loci and could not be assigned. One FISH experiment was performed without a reference probe so a chromosome location cannot be assigned to probe a0039L14 and a0017C21 but the results indicate that they hybridise to the same chromosome meaning that the super-contig 6 appears to be assembled correctly.

The largest super-contigs (0, 1) can be assigned to each arm of the largest Brachypodium chromosome. In addition, super-contig 3 can be assigned to chromosome 2S and super-contig 5 can be assigned to chromosome 5 in its entirety. Possible mis-assemblies are indicated in super-contig 2 where the clone from one end of the super-contig hybridises to chromosome 2L and two probes from the other end hybridise to chromosome 3S. Similarity, probes from each end of super-contig 4 hybridise to three different chromosomes (3S, 4S and 5S).

BAC clone	Super-contig	Approximate position (Mb)	Chromosome assignment from FISH	Est. repetitive content (%)
a0044D23	0	0	1S	59.4
a0039J03	0	2	Multi-locus hybridisation	86.4
b0010E24	0	38	1S	7.4
a0019H09	1	0	1L	41.6
b0033J12	1	37	1L	58.3
b0016C15	2	0	2L	12.9
a0044P16	2	27	3S	19.6
b0019G01	2	32	3S	16.8
b0023M21	3	0	2S	14.6
b0017M03	3	5	2S	24.3
b0001B16	3	30	Multi-locus hybridisation	76.4
a0015M04	4	0	3S, 5S	41.2
b0016F18	4	30	4S	7.4
a0040K04	5	0	5S	18.2
b0038G13	5	28	5L	10.0
a0039L14	6	0	Same chromosome as a0017C21*	38.2
a0017C21	6	27	Same chromosome as a0039L14*	27.6

Table 2.13: BAC clones from pre-release super-contigs used as probes in FISH.

BAC name, approximate position on the pre-release super-contigs and chromosome assignment from FISH hybridisation is shown. (S=short arm, L=Long arm), * indicates that no reference BAC was used so chromosomal location could not be determined. The repetitive content for each BAC is estimated using the Perl script described in the methods section.

2.6.3.5 Screening BACs for repetitive DNA

The BACs used as probes in the previous section were retrospectively screened for repetitive DNA using my Perl script described in the Methods section of this chapter to determine whether the BAC selection procedure could be optimised. The two BACs that hybridised to multiple locations on the chromosomes (a0039J03 and b0001B16) were estimated to contain particularly high proportions of repetitive DNA (86.4% and 76.4% respectively). The probes hybridising to a single locus had a mean estimated repetitive content of 26.5% (σ = 17.4).

To increase the productivity of our FISH experiments, the Perl script was used to estimate repetitive content in all BACs before using them as FISH probes. BACs were selected with the lowest possible repetitive content to maximize the chances that they would hybridise to a single locus on the Brachypodium chromosomes.

2.6.3.6 Validation of the final assembly using two-colour FISH

An additional 10 BACs with low repetitive DNA content were identified from the final assembly, targeting regions that had not yet been anchored and regions that contained predicted joins between pre-release super-contigs (Figure 2.22).

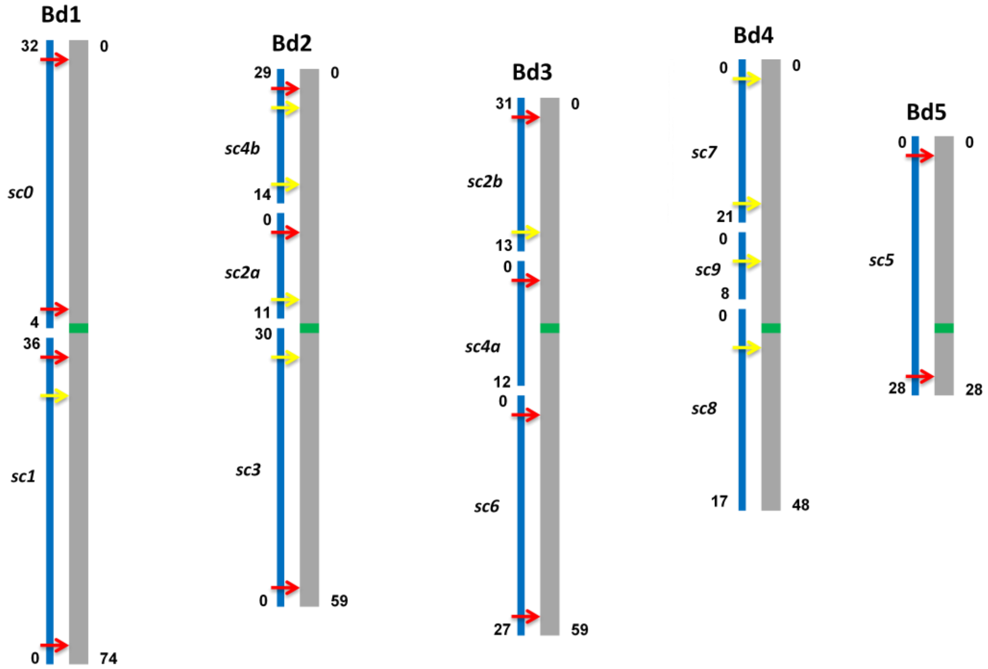


Figure 2.22: Positions of additional BAC probes on the pseudo-molecules of the final assembly.
The pseudo-molecules are shown as grey blocks with centromeres in green. Red arrows indicate the positions of BAC probes identified from previous experiments, yellow arrows indicate new BAC probes identified in this experiment. New probes were chosen to anchor the unanchored ends of pre-release super-contigs (shown as blue lines aligned to the pseudo-molecules and labelled sc0-sc8). Approximate lengths (in Mb) of super-contigs are indicated to the left of each super-contig and approximate lengths of pseudo-molecules are shown on the right.

Table 2.14 shows the positions of these probes on the pre-release super-contigs and pseudo-molecules alongside the results from FISH. With the exception of a0043I11 and a0021O21, the BAC probes hybridise to the expected chromosome. For example, Bd4 comprises of super-contigs 7, 9 and 8 according to the alignment of the super-contigs to the high-density genetic map. These results indicate that probes a0021B17, a0011N01 and b0026I15 identified from these super-contigs all hybridise to chromosome 4. In addition, the chromosome assignments of super-contigs from previous FISH experiments are confirmed. Pseudo-molecule Bd5 comprises super-contig 5 from the pre-release assembly which has previously been anchored to chromosome 5. BAC probes a0043I11 and a0021O21 do not hybridise to the expected chromosomes (Bd2 and Bd4 respectively) but to other chromosomes highlighting anomalies in the 0 to 12 Mb region of Bd2 and the 0 to 20 Mb region of Bd4. To investigate these anomalies, two BAC pools were identified using clones taken from evenly spaced intervals spanning these regions. The probe BACs used and the chromosomal locations from FISH are shown in Table 2.15.

BAC clone	Super-contig / position (Mb)	Pseudo-molecule / position (Mb)	Chromosome assignment from FISH	Est. repetitive content
a0029E05	1: 33	1: 41	1S	18.9
a0043I11	4: 21	2: 1	Not 2L	13.5
a0001I12	4: 15	2: 15	2L	8.1
a0031O03	2: 10	2: 26	2L	21.5
a0040A20	3: 28	2: 32	2L	20.0
b0002A16	2: 14	3: 17	3L	13.2
a0021O21	7: 0	4: 1	Not 4L	16.0
a0021B17	7: 18	4: 17	4L	28.2
a0011N01	9: 3	4: 26	4L	20.2
b0026I15	8: 0	4: 32	4L	4.7

Table 2.14: BACs probes chosen for the second phase of FISH experiments.

The approximate position of each BAC on the pre-release super-contigs and the pseudo-molecules of the final assembly are shown alongside the chromosomal location obtained from FISH. The repetitive content for each BAC is estimated using the Perl script described in the Methods section.

BAC clone	Pseudo-molecule	Distance from end of pseudo-molecule (Mb)	FISH position
a0036M12	2	0	2L sub-terminal
b0003O11	2	3	2L sub-terminal
b0015L19	2	6	2L sub-terminal
b0027B07	2	9	2L interstitial
a0017F11	2	12	2L interstitial
b0022C23	4	0	4S sub-terminal
b0035I12	4	4	4S sub-terminal
a0035E03	4	12	4S interstitial
b0006E18	4	16	2
a0036J07	4	20	1-5 centromeric

Table 2.15: BAC probes used to investigate the anomalies in Bd2 and Bd4.

Chromosomal positions from FISH hybridisation are also shown.

As all the probes identified from Bd2L hybridise to chromosome 2L we can conclude that the anomaly previously highlighted in Bd2 was probably due to a single BAC probe (a0043I11) from the previous experiment mapping incorrectly, as in this more detailed experiment all probes from Bd2L landed on the long arm of chromosome 2. However, one of the BAC probes from Bd4S region (b0006E18) still indicates an anomaly approximately 16 Mb from the end of the pseudo-molecule. This probe hybridizes to chromosome 2 whereas all other probes (except the centromeric probe a0036J07) hybridise to chromosome 4 as expected.

To investigate this further, a more detailed FISH experiment was performed using 4 BAC probes spanning the 6 Mb region surrounding the predicted breakpoint on Bd4 at evenly spaced intervals. BAC probe details are shown in Table 2.16.

BAC clone	**Pseudo-molecule**	**Distance from end of pseudo-molecule (Mb)**	**FISH position**
a0002I10	4	13.2	4S pericentromeric
a0003P20	4	15.2	4S interstitial
a0002J19	4	17.3	2L interstitial
a0009D02	4	17.7	4S sub-terminal

Table 2.16: BAC probes used to investigate in more detail the anomaly in Bd4.

Chromosomal positions from FISH hybridisation are also shown.

Once again, a single probe BAC (a0002J19) landed on Bd2, confirming the presence of an anomaly between 16 and 17.5 Mb from the distal end of Bd4S. This could be due to a mis-assembly in the pseudo-molecule, a problem with the physical map or could be due to the incorrect BAC being selected for FISH hybridisation. In the physical map, this region is in the centre of contig 23 which has been assembled from overlapping clone fingerprints. Both probe BACs (b0006E18 and a0002J19) are assembled into this contig and the BES from these BACs all align to Bd4. The alignment between contig 23 and the 12 to 23 Mb region of Bd4 indicates that this contig has been correctly constructed and that it accurately reflects the sequence in the pseudo-molecule. Conversely, the fact that one BAC from two separate experiments hybridises to Bd2 indicates that this result is probably not due to the wrong BAC being selected for FISH hybridisation. Further hybridisation experiments are required to accurately define this potentially problematic region and to identify the cause of the incorrectly mapping BACs, something that should be elucidated when Bd4S is 'painted' using a larger BAC pool (described in section 2.6.3.8).

A summary of the FISH chromosome assignments from multiple BAC landing experiments is shown in Figure 2.23 and shows that each pseudo-molecule can be assigned to an individual chromosome in the Brachypodium karyotype using reference BACs. In addition, BAC probes taken at intervals along the length of each pseudo-molecule hybridise to relative positions on the chromosome as is expected by their position on the pseudo-molecule. For example, Bd1 is assigned to chromosome 1 by virtue of the reference BAC ABR1-26H1 which is known to hybridise to the distal end of the long arm of chromosome 1. From the epifluorescence images, it is evident that probe BACs b0010E24, a0029E05 and a0019H09 all hybridise to this chromosome with probe b0010E24 hybridising to the opposite end of the chromosome to the reference BAC, probe a0029E05 hybridising in a centromeric position and probe a0019H09 hybridising in close proximity to the reference BAC.

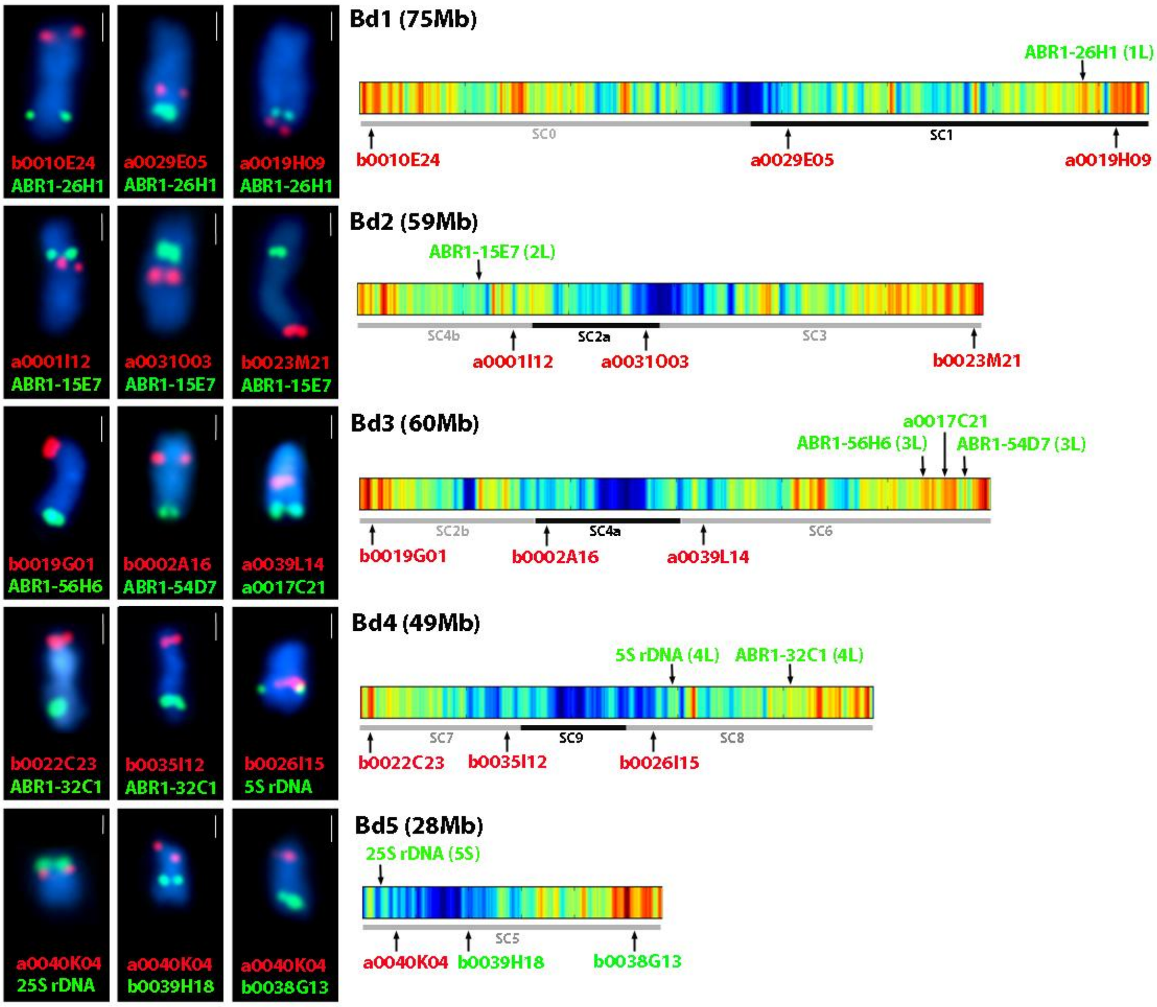


Figure 2.23: Brachypodium pseudo-molecules aligned to the karyotype using fluorescently labelled BACs. The pre-release super-contigs comprising each pseudo-molecule are shown below each pseudo-molecule. Super-contigs are labelled SC0-9 and pseudo-molecules labelled Bd1-5.Reference BACs (ABR1 clones, 5S rDNA and 25S rDNA) are used to identify the target chromosome. On the right hand side of the images, each pseudo-molecule is shown as a heat map indicating gene density. The position of BAC probes (both physical map BACs and reference BACs) are shown on the on the right. Each pseudo-molecule has three related epifluorescence images obtained from FISH showing the relative position of hybridisation for each BAC probe on the Brachypodium karyotype. Probe BACs are labelled on each image and on the pseudo-molecules in the in the same colour as their fluorescent label (scale bar = 1μm).

2.6.3.7 Validation of the final assembly using chromosome-specific BAC pools

The five chromosome-specific BACs pools are shown in Table 2.17 alongside the predicted positions of each BAC from alignment of the physical map contigs to the sequence assembly. Each pool contained between 4 and 8 BACs and was hybridised alongside chromosome-specific markers. The epifluorescence images obtained from FISH are shown in Figure 2.24.

Pool	BAC name	Pseudo-molecule	Position (Mb)
1	a0040G20	Bd1	48.4
	a0029E10	Bd1	9.7
	a0030F11	Bd1	28.0
	b0003G22	Bd1	68.5
2	a0044B16	Bd2	8.2
	b0005P05	Bd2	7.7
	a0044B16	Bd2	8.3
	b0023P23	Bd2	13.3
	a0019A19	Bd2	42.1
	a0039D05	Bd2	59.1
3	a0004A14	Bd3	40.0
	a0010O04	Bd3	57.0
	a0041J04	Bd3	57.5
	a0039O07	Bd3	59.5
4	b0033O01	Bd4	5.4
	a0002I07	Bd4	10.8
	a0018K11	Bd4	25.4
	a0036L13	Bd4	30.0
	a0037L08	Bd4	31.8
	a0045G06	Bd4	46.8
	b0016N17	Bd4	46.6
	a0022E14	Bd4	46.8
5	a0013L16	Bd5	0.7
	b0033D02	Bd5	28.0
	b0003N09	Bd5	9.8
	a0015J02	Bd5	18.6
	b0013P03	Bd5	28.0
	a0002N04	Bd5	26.8
	b0029E05	Bd5	19.8
	b0041P08	Bd5	18.7

Table 2.17: BACs comprising the five chromosome-specific BAC pools.

These BACs were used as probes to validate the final sequence assembly and physical map contigs by hybridisation to Brachypodium chromosomes.

BACs from pools 2, 3, 4 and 5 hybridise to a single chromosome which is the expected chromosome as defined by the reference anchor probes (Panels B, C, D and E: Figure 2.24). Three of the four BACs identified from physical map contigs aligning to Bd1 (shown in Panel A) hybridise to chromosome 1, but one probe BAC hybridises to the distal end of another chromosome. The most likely explanation for this is that an error occurred when selecting the BACs for the pool but the experiment would need to be repeated to ascertain whether this was the case. Internal duplications have since been identified between Brachypodium chromosomes 1 and 3 which may be the cause of this anomaly. With the exception of one BAC, these results indicate that the WGS sequence assemblies and the physical map contigs aligned to them are accurate representations of the underlying chromosomes.

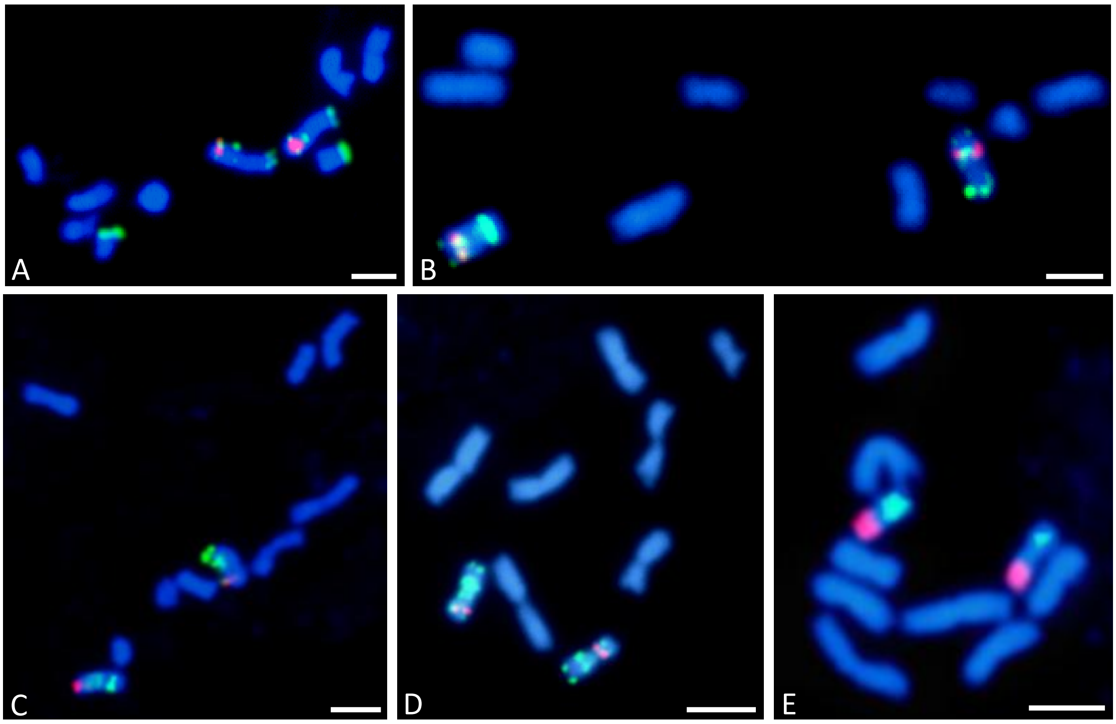


Figure 2.24: Epifluorescence images from hybridisation of chromosome specific BAC pools.
A. BAC pool of Bd1 (green) with ABR1-26-H1 – anchor for 1L (red). B. BAC pool of Bd2 (green) with ABR1-41-E10 – anchor for 2S (red). C. BAC pool of Bd3 (green) with ABR5-33-F12– anchor for 3S (red). D. BAC pool of Bd4 (green) with ABR5-33-F2 – anchor for 4S (red). E. BAC pool of Bd5 (green) with 25S rDNA – anchor for 5S (red), scale bars = 5 µm.

2.6.3.8 Validation of the final assembly using chromosome painting

This part of the analysis is ongoing and although BACs have been selected from physical map contigs aligning to each chromosome arm and screened for repeats only a small region from chromosome 1 and a more extensive region of chromosome 2 has been painted. Table 2.18 shows the number of clones selected from each chromosome arm of Brachypodium. Each chromosome arm is evenly covered by clones with an estimated low amount of repetitive DNA but only 4 suitable clones were selected from the short arm of Bd5 due to its small size (~7 Mb) and high levels of retrotransposons. This is the region containing the 45S rDNA locus so probes specific to this locus can be used in addition to identified clones.

Pool	Number of clones
Bd1S	70
Bd1L	72
Bd2S	35
Bd2L	55
Bd3S	42
Bd3L	54
Bd4S	25
Bd4L	39
Bd5S	4
Bd5L	19

Table 2.18: Number of BACs in each chromosome arm pool used to 'paint' chromosome arms

The BAC clones identified by alignment to Bd1S were applied in smaller sub-pools along each chromosome arm. The epifluorescence microscopy image in Figure 2.25 shows hybridization of a clone sub-pool spanning approximately 4 Mb on Bd1S. All clones appear to hybridise to a distinct region on a single chromosome indicating good consistency between the physical map, the sequence assembly and the cytogenetic map in this region.

Clone name	Start (Mb)	End (Mb)	Repeat content (%)
b0002O16	20.0	20.2	0.00
a0024N14	20.5	20.6	19.18
a0027K03	21.2	21.3	0.00
a0010K04	21.5	21.6	19.83
a0011I01	21.9	22.0	0.00
a0018B03	22.4	22.6	32.71
b0022H13	23.1	23.2	0.00
a0023E14	23.2	23.4	21.52
a0043B06	24.0	24.2	22.76
a0042C21	24.2	24.4	16.48

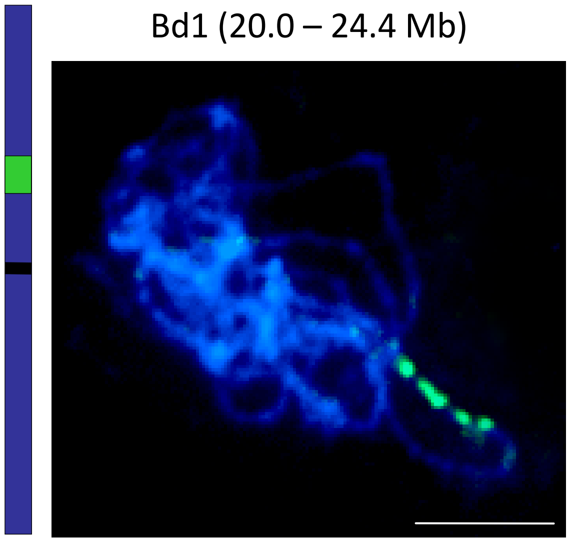


Figure 2.25: Chromosome painting using a sub-pool of BACs from Bd1S.

The clone name, positions on the pseudo-molecules and estimated repeat content are given in the table to the left. The ideogram is a representation of the Brachypodium chromosome 1 showing the position of the BACs in the sub-pool. The hybridised BACs appear as green spots on the chromosome spreads and all appear to hybridise to a distinct region on one chromosome (scale bar = 5µm).

BAC clones from physical map contigs aligning to both arms of Bd2 were hybridised simultaneously to Brachypodium chromosome spreads using two-colour FISH to distinguish the arms. The full list of 26 BAC clones from Bd2S and the 29 BAC clones from Bd2L are given in Supplementary information 9. The epifluorescence microscopy image of hybridised probes from these pools is shown in Figure 2.26 and clearly shows each chromosome arm (Bd2L in red, Bd2S in green) with the pericentromeric region shown faintly in blue where no probes hybridise. BACs in this region tend to contain a high percentage of repeats so they were excluded from the BAC pools. Due to the orientation of the chromosome in the chromosome preparation, the region of Bd2S represented by BAC probes is much larger than the region of Bd2L. The ability to paint an entire chromosome using BAC probes identified from physical map contigs indicates that the WGS sequence assemblies and the physical map contigs are accurate representations of the underlying chromosomes.

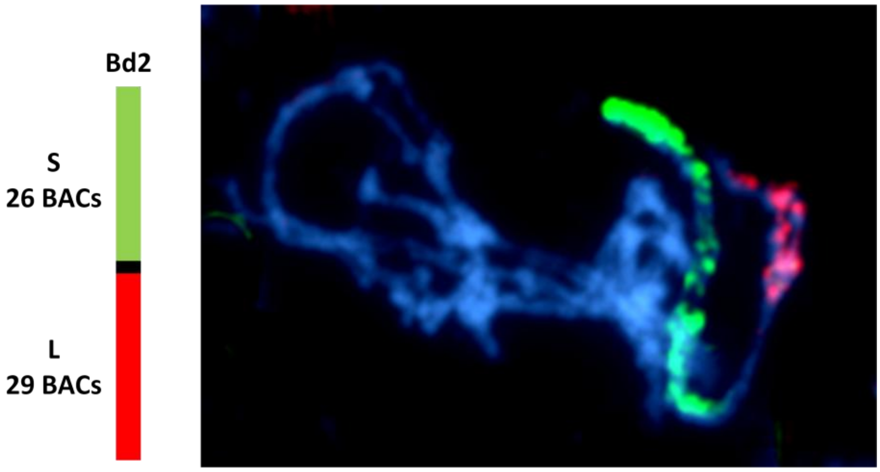


Figure 2.26: Epifluorescence microscopy image from chromosomal FISH.
A Brachypodium chromosome spread was decorated using a pool of 26 BAC probes from Bd2S (labelled in green) and 29 BAC probes from Bd2L (labelled in red). Each arm of Bd2 is clearly visible with an absense of probes in the pericentromeric region.

2.7 Comparative genomics between Brachypodium and other grasses

The first grass consensus map was built based upon the observation that the order of markers in the genetic maps from multiple grass species was largely conserved (Moore *et al.* 1995). This consensus map defined 25 rice linkage blocks that could be rearranged to represent the genomes of the other grasses. The genetic maps used in these comparisons provided relatively low resolution, thus only large rearrangements could be detected. Subsequently, the availability of genome sequence from various grasses shifted the focus of research to identify whether this conservation extended to the DNA sequence level. Comparative studies of small homologous regions revealed a conservation of gene order (Chen *et al.* 1997; Feuillet and Keller 1999) but also significant rearrangements that were not detected at the genetic map level (Tikhonov *et al.* 1999; Tarchini *et al.* 2000). The availability of the rice genome sequence in 2005 (International Rice Genome Sequencing Project 2005) as well as more than 1 million wheat ESTs, of which 7,107 have been mapped into deletion bins (Qi *et al.* 2004) allowed large scale comparisons between rice and wheat (Sorrells *et al.* 2003; La Rota and Sorrells 2004; Singh *et al.* 2007). These comparative studies revealed more rearrangements than had previous been observed at the genetic map level. A more statistically rigorous analysis used comparisons between rice and wheat combined with the results from previous comparative analysis between rice, sorghum and maize, to propose an evolutionary model for the grasses based on a five chromosome ancestor common to all the grasses (Salse *et al.* 2008a). Rice is a representative of the Ehrhartoid grass sub-family and most recently, the completed genome sequences of sorghum (Paterson *et al.* 2009) and maize (Schnable *et al.* 2009) have provided representative genome sequences from the Panicoid sub-family. The genome sequence of Brachypodium provides the first complete genome sequence from a representative of the Pooid grass sub-family so direct comparisons between this genome sequence and the other grasses will allow the relationship between these three grasses to be examined in more detail than was previously possible.

I performed genome-wide comparisons between Brachypodium, rice, wheat, barley and sorghum at each stage of the sequencing project to identify syntenic regions. For sequenced grasses such as rice and sorghum, sequence level comparisons were performed. To compare Brachypodium to other Triticeae species for which no sequenced genome exists, deletion bin-mapped or genetically mapped ESTs were used.

2.7.1 Methods

2.7.1.1 Sequence-level comparisons

For comparisons between genomic DNA, for example, Brachypodium to rice and Brachypodium to sorghum, I used the NUCmer alignment script with default settings from the MUMmer package (Delcher *et al.* 2002). The output of this script is a text file of alignments that is converted to a dot-plot using MUMmerplot script. The dot-plots show a visual representation of the chromosome scale sequence similarity between the sequences. For comparison to rice I used the TIGR version 5 sequence assembly and for comparisons to sorghum I used the version 1.0 genome release.

I used the Artemis Comparison Tool to obtain more detailed comparisons between genomic DNA (Carver *et al.* 2005). These comparisons were guided by the dot-plot comparisons and used DNA sequences that had been masked for repeats. BLASTN with the –m8 parameter was used to obtain high-scoring pairs (HSPs) in tabular format from the sequences being compared.

2.7.1.2 Comparison to wheat and barley

Where a direct sequence comparison was not possible, ESTs were used to obtain chromosome-level alignments. First, ESTs of known position on the wheat or barley genomes were aligned to the Brachypodium assembly using the annotation pipeline described in Section 2.4, then loaded into the genome browser. A Perl script was written using the BioPerl Bio::DB::GFF interface to extract the location of each EST from the Brachypodium super-contigs. These coordinates were used to create map and correspondence files to load into CMap (Youens-Clark *et al.* 2009) and good alignments were visually identified.

Wheat ESTs mapped into deletion bins (Qi *et al.* 2004) were used to compare Brachypodium to wheat. Each wheat chromosome arm has between 3 and 4 deletion bins defined along its length. Although ESTs from deletion bins on all three homoeologous chromosome sets (A, B and D) of the wheat genome were aligned to Brachypodium, each group of three homoeologous wheat chromosomes were treated as a single entity to create a consensus wheat genome. This approach is justified for initial genome-wide comparisons as the Brachypodium genome is estimated to have diverged from the wheat lineage between 35 and 40 MYA (Bossolini *et al.* 2007) whereas the diploid *Triticum* / *Aegilops* progenitors of the A, B and D genomes diverged from a common ancestor much more recently, between 2.5 and 4.5 MYA (Huang *et al.* 2002). Thus the A, B and D genomes of wheat will be more conserved with respect to each other than they will to Brachypodium. For comparisons between Brachypodium and barley, I aligned a set of barley transcript sequences from the HarvEST-21 assembly (http://harvest.ucr.edu/) to the Brachypodium sequence assemblies. The HarvEST database provides accurate transcript assemblies for various crop plants generated from trimming, cleaning and assembling ESTs using CAP3 (Huang and Madan 1999). The HarvEST-21 assembly is assembled using stringent parameters to define accurate barley transcripts. The transcript sequences within this assembly had been assigned into bins on a barley genetic map (D. Marshall, pers. comm.). Each barley chromosome was represented by a single linkage group and each contained 40-60 bins which allowed fine resolution alignments compared to the lower resolution that was possible using the larger bins of the wheat deletion-bin map.

2.7.2 Results

2.7.2.1 Comparative genomics using the mid-point assembly

2.7.2.1.1 Comparison to rice

The super-contigs from the mid-point assembly were aligned to the rice chromosomes in order to assess the coverage of the rice genome by these super-contigs and to assess the collinearity between the sequences at this early stage. The super-contigs were aligned to each rice chromosome in turn, resulting in twelve dot-plots each identifying the super-contigs showing sequence similarity to that chromosome. Significant sequence similarity was detected between rice chromosomes and the largest 16 super-contigs and is summarized in Figure 2.29. The individual dot-plot for rice chromosome 1 is shown in Figure 2.27 and rice chromosome 3 in Figure 2.28. The dot-plots have the super-contigs (0-15) placed end-to-end on the y-axis and the rice chromosome on the x-axis. Within the plot a red or blue dot is placed where significant sequence similarity is seen between the sequences and coloured to indicate the orientation of the match. Long stretches of similar sequence appear as red or blue lines and can be easily identified from the background matches using this visualisation technique. Figure 2.27 shows that rice chromosome 1 has significant sequence similarity to super-contigs 2 and 9 on one arm and super-contig 4 and 13 on the other arm with super-contig 4 matching in reverse orientation. An apparent duplication can also be seen where a region of rice chromosome 1 matches super-contig 4 as well as exhibiting a weaker match to super-contig 2.

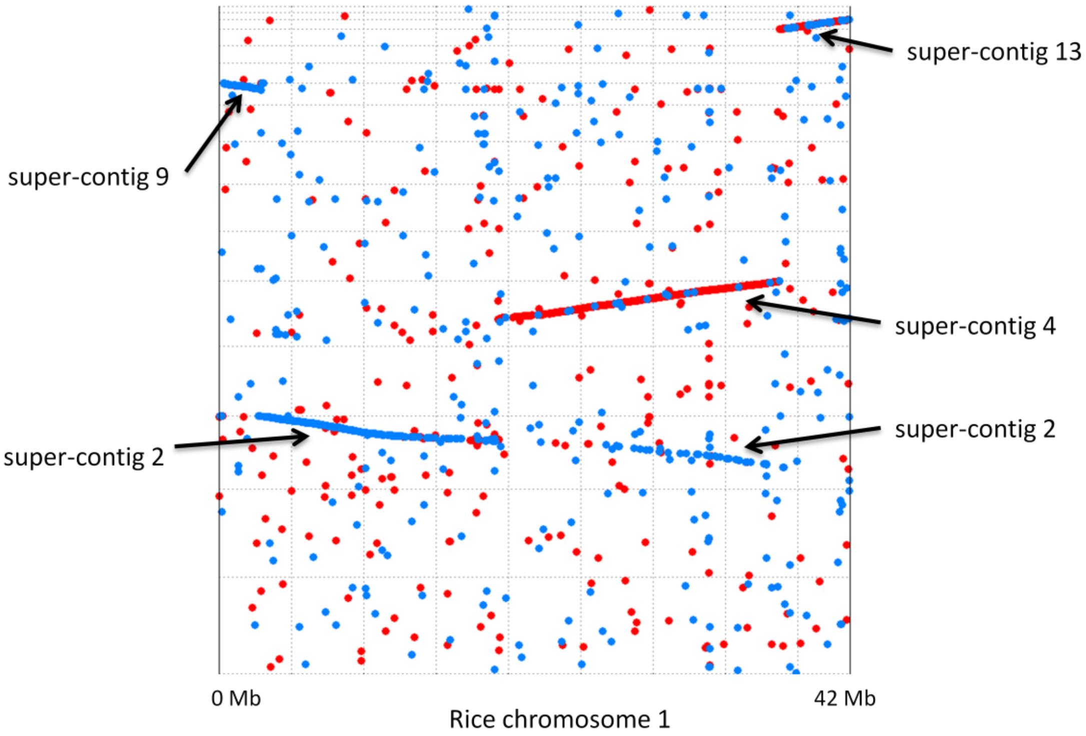


Figure 2.27: MUMmer dot-plot comparison of rice chromosome 1 to the largest 16 super-contigs. Super-contigs are placed end-to-end on the y-axis and the rice chromosome on the x-axis. Within the plot a red or blue dot indicates significant sequence similarity between the sequences and coloured to indicate the orientation of the match. Significant sequence similarity is observed between rice chromosome 1 and super-contigs 2, 4, 9 and 13.

A simpler pattern of sequence similarity is observed between rice chromosome 3 and the super-contigs (Figure 2.28). One arm of the rice chromosome shows significant sequence similarity to super-contig 0 and the other arm of the chromosome matches super-contig 1.

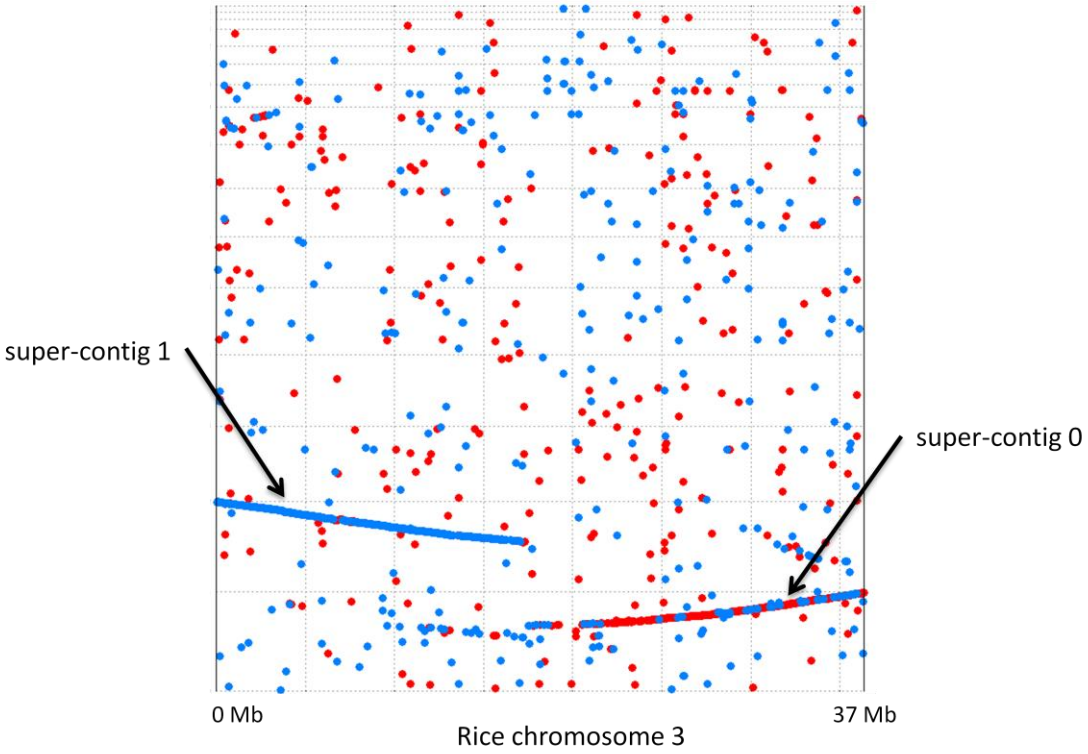


Figure 2.28: MUMmer dot-plot comparison of rice chromosome 3 to the largest 16 super-contigs. Significant sequence similarity is observed between rice chromosome 3 and super-contigs 0 and 1.

The alignment of all 16 super-contigs to rice chromosomes (Figure 2.29) shows that most of the rice genome is covered by super-contig alignments. Duplications are also evident where a region of weak similarity overlaps with a super-contig alignment, shown as a dotted line on the figure. Previous analysis of duplication within the rice genome has identified duplications that exist between chromosomes 1 and 5 (Salse *et al.* 2004) and chromosomes 11 and 12 (Yu *et al.* 2005). These duplications are both evident in Figure 2.29 where the same super-contigs align to different rice chromosomes. For example, rice chromosomes 1 and 5 are predominantly covered by super-contigs 2 and 4 and the end of chromosomes 11 and 12 show similarity to super-contigs 7 and 8. None of the additional duplications identified by Salse and colleagues were reflected in these alignment most probably due to the less rigorous alignment method used in this analysis (Salse *et al.* 2008a).

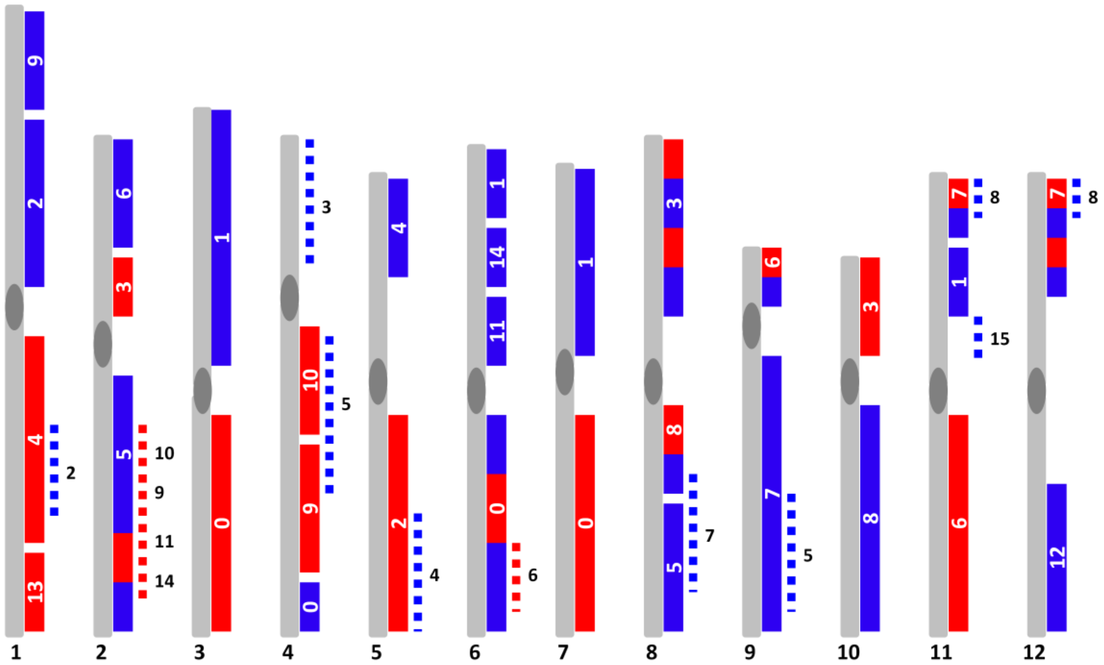


Figure 2.29: Sequence similarity identified between rice chromosomes and Brachypodium super-contigs. The rice chromosomes are shown in light grey with centromeres in dark grey and Brachypodium super-contigs from the mid-point assembly and shown as blue or red blocks indicating orientation. Dotted lines show weaker areas of sequence similarity.

Guided by the dot-plot comparisons in Figure 2.29 I produced more detailed comparisons using the Artemis Comparison Tool (ACT). Figure 2.29 shows that both rice chromosomes 3 and 7 align to super-contigs 0 and 1. If these comparisons are visualised within ACT, the precise locations of synteny breakpoints between rice chromosomes and contigs can be identified. Figure 2.30 shows a high-level pair-wise comparison between super contigs 0 and 1 and rice chromosomes 3 and 7 and shows that different regions of a single super-contig show similarity to different rice chromosomes. For example, the top of super-contig 0 shows similarity to the top of rice chromosome 3 and the middle of the same super-contig shows similarity to the bottom of rice chromosome 7. The remainder of rice chromosomes 3 and 7 show similarity to super-contig 1. In addition to showing collinearity to rice 3, the lower part of super-contig 1 also shows a weaker similarity to rice 7 indicating the presence of a segmental duplication within the sequences.

These alignments show good collinearity between the Brachypodium super-contigs and the rice chromosomes suggesting that the WGS super-contigs are being accurately assembled. Specific regions of alignment can be identified between the sequences which are separated by syntenic breakpoints where synteny is disrupted. These breakpoints could reflect chromosomal rearrangements that have taken place between Brachypodium and rice but at this stage of the genome assembly they could also reflect mis-assemblies in the Brachypodium super-contigs.

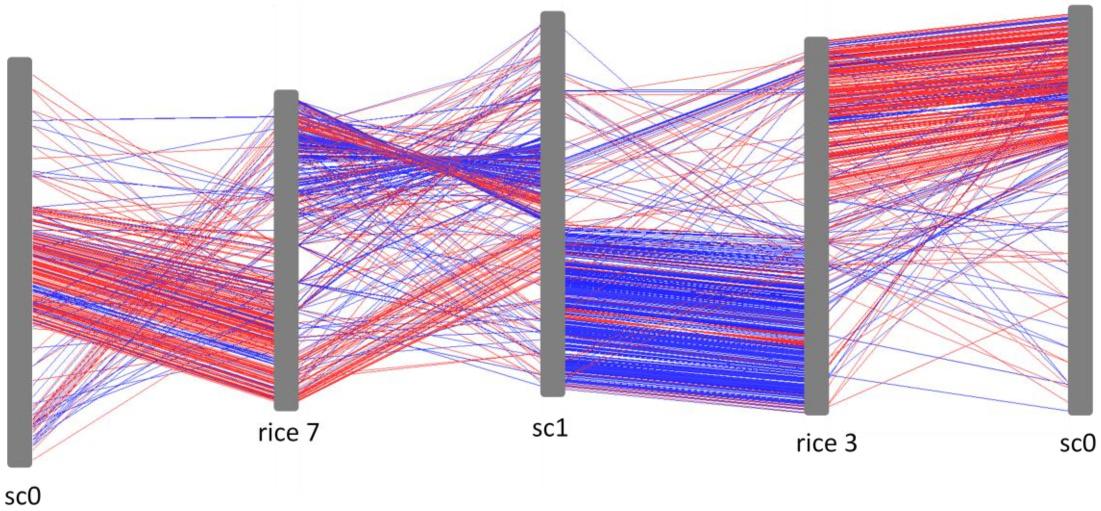


Figure 2.30: Fine-scale sequence alignment between rice chromosomes 3 and 7 and super-contigs 0 and 1. Super-contigs 0 and 1 are labelled sc0 and sc1. Regions of sequence similarity in each pair-wise comparison are joined by red or blue lines to indicate the orientation of the comparison. Specific regions of collinearity are identified separated by syntenic breakpoints.

2.7.2.1.2 Comparison to wheat

The alignments between super-contigs and the wheat chromosomes identified via deletion bin-mapped ESTs are summarised in Figure 2.31. At this resolution, all of the wheat genome is covered by super-contigs with some super contigs aligning to more than one wheat chromosome. Although not shown in the figure, it is interesting to note that the majority of ESTs aligning to Brachypodium come from deletion bins on the D-genome of wheat (59.9%). This genome is the most recent addition to the hexaploid wheat genome, incorporated within the last 10,000 years, and originates from the wild grass species *Aegilops tauschii* (Huang *et al.* 2002). The fact that the majority of aligned ESTs come from D-genome deletion bins suggests that Brachypodium genes are more closely related to wheat genes in the D-genome than the A or B genomes as there is no overrepresentation of ESTs from the D-genome in the bin-mapped EST dataset. However, this observation is most likely artifactual as the Brachypodium lineage is estimated to have diverged from the *Triticeae* lineage more than 35 MYA (Bossolini *et al.* 2007) and the diploid *Triticum* and *Aegilops* progenitors diverged between 2.5 and 4.5 MYA (Huang *et al.* 2002). There is no reason to assume that Brachypodium genes are more closely related to genes on any one of the homoeologous chromosomes of wheat.

There is evidence of segmental duplication in these alignments where the same super-contig aligns to two wheat chromosomes. Many of these duplications are the same as those previously identified in wheat, for example the duplication between wheat 1 and 3, shown by alignment to super-contig 2, and the duplication between wheat 5 and 7, shown by alignment to super-contig 0 (Salse *et al.* 2008a).

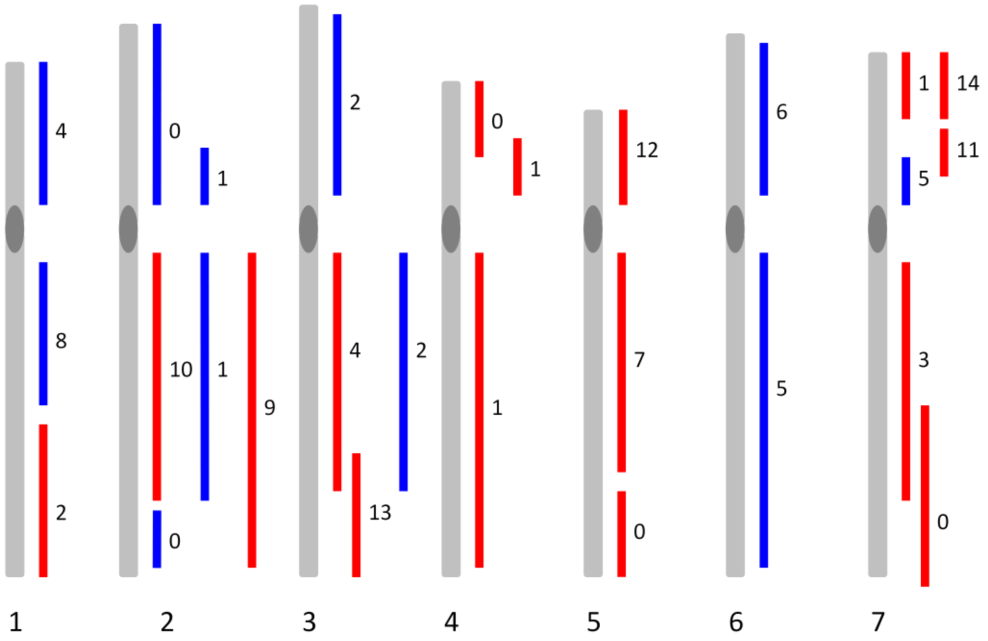


Figure 2.31: Summary of alignments between wheat and Brachypodium super-contigs.
Each wheat consensus chromosome group is shown in light grey with approximate centromeres marked in dark grey. Brachypodium super-contigs are shown in blue or red to indicate orientation.

2.7.2.1.3 Comparison to barley

The alignment of the super-contigs to barley linkage groups identified via bin-mapped ESTs is summarised in Figure 2.32. The super-contigs aligning to each chromosome are almost identical to the alignments between wheat and Brachypodium (Figure 2.31). These alignments appear to be more accurately delineated, reflecting the higher resolution of the barley genetic map compared to the wheat deletion-bin map. The increased resolution provided by the barley alignments also indicate new alignments such as a short segment of super-contig 9 aligning to the distal end of barley linkage group 3. This alignment is not identified from the alignments to wheat. In addition, super-contigs 1, 5, 11 and 14 align to the short arm of wheat consensus chromosome 7, however the alignment to barley indicates that only super-contigs 1 and 14 are aligned. Some of these differences will be due to the types of map being compared; the wheat deletion-bin map is based on physical distances since bins are defined by chromosomal deletions whereas the barley map is based on recombination rates. This means that at the distal ends of chromosomes where the rate of recombination is generally higher, one cM may correspond to a smaller physical distance than in pericentromeric regions where the rate of recombination is lower.

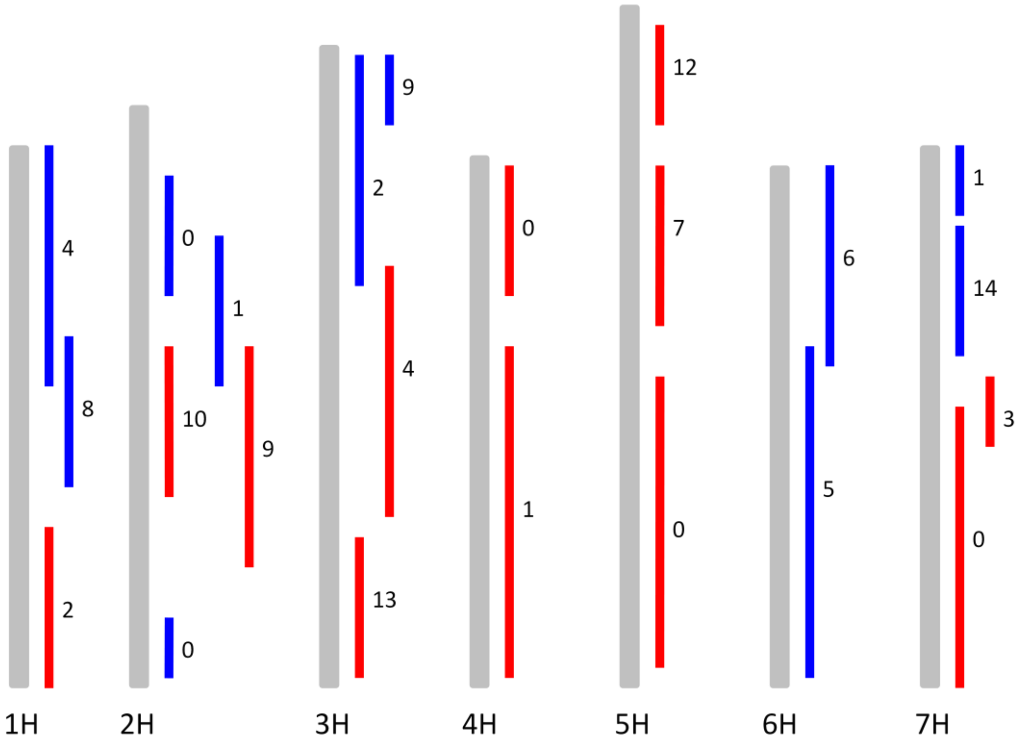


Figure 2.32: Summary of alignments between barley and Brachypodium super-contigs.
Each barley linkage group is shown in light grey and the Brachypodium super-contigs are shown in blue or red to indicate orientation.

2.7.2.1.4 Validation of alignments using existing comparisons between rice and wheat

Since elucidation of its genome sequence, rice has been used for map-based cloning of genes in wheat and as the reference grass genome sequence (International Rice Genome Sequencing Project 2005). Detailed comparisons have been performed between the wheat deletion-bin maps and the rice genome (La Rota and Sorrells 2004). I used these comparisons to validate the alignments of the Brachypodium super-contigs to rice, wheat and barley. For example, it is known that the chromosomes of wheat group 6 are predominantly syntenic with rice chromosome 2. The alignments obtained herein show that Brachypodium super-contigs 5 and 6 align to both wheat chromosome group 6 and rice chromosome 2. This agreement is observed for each syntenic wheat-rice relationship although the Brachypodium-rice alignments provide more detail due to the more thorough alignment made possible by comparing complete genome sequences (Table 2.19). These are the first genome-wide alignments between Brachypodium, rice, wheat and barley.

In order to further validate the Brachypodium-rice and Brachypodium-wheat alignments, a set of 31 conserved orthologous set (COS) markers were obtained, designed from wheat ESTs aligning to rice chromosome 1 and therefore assumed to be from wheat chromosome group 3 (Simon Griffiths, JIC, pers. comm.). Twenty-four of these marker sequences could be aligned to the Brachypodium super-contigs and the majority of these (75%) aligned to super-contigs 2, 4 and 13 which are the super-contigs predicted to align to wheat chromosome group 3 by the alignments shown here.

Barley chromosome	**Wheat chromosome**	**Rice chromosomes**	**Brachypodium super-contigs**
1H	1	5, 10	4, 8, 2, 3
2H	2	4, 7	0, 1, 9, 10
3H	3	1	2, 4, 13, 9
4H	4	3	0, 1
5H	5	3, 9, 12	12, 7, 0, 1, 6
6H	6	2	5, 6, 3
7H	7	6, 8	1, 14, 11, 5, 3, 0, 8

Table 2.19: Summary of alignments between Brachypodium super-contigs related grasses.
The super-contigs aligning to each *Triticeae* chromosome are shown alongside the corresponding rice chromosome. The underlined super-contig numbers indicate alignments only observed in the Brachypodium-rice comparisons due to more detailed alignments made possible by genome sequence comparisons.

There is striking similarity between the Brachypodium-wheat alignments and the Brachypodium-barley alignments as one would expect given that the barley and wheat lineages diverged only 10 to 14 MYA (Wolfe *et al.* 1989) and that the barley genome is essentially equivalent to a diploid wheat genome. These alignments were obtained by aligning Brachypodium super-contigs to two separate species in the *Triticeae* using two independent datasets and indicate that the alignments shown are both accurate and robust.

2.7.2.2 Comparative genomics using the final assembly

2.7.2.2.1 Chromosome level alignments between Brachypodium, rice, sorghum and wheat

Upon release of the Brachypodium pseudo-molecules (Bd1-5) comprising the final assembly I again compared each rice chromosome to each of the pseudo-molecules and analysed the resulting dot-plots to obtain an overview of synteny between the species. In addition, I compared each Brachypodium pseudo-molecule to the ten sorghum chromosomes to add this newly sequenced grass into the comparison (Paterson *et al.* 2009). In order to obtain a wheat-Brachypodium genome comparison, I aligned bin-mapped wheat ESTs (Qi *et al.* 2004) to the Brachypodium pseudo-molecules. These genome-wide alignments are summarised in Table 2.20 and are the first between grass species representing the three major grass sub-families, the Pooideae, the Ehrhartoideae and the Panicoideae.

Brachypodium pseudo-molecule	**Rice chromosomes**	**Sorghum chromosomes**	**Wheat chromosomes**
Bd1	Os3, Os6, Os7	Sb1, Sb2, Sb10	Ta2 ,Ta4, Ta5, Ta7
Bd2	Os1, Os5	Sb3, Sb9	Ta1, Ta3
Bd3	Os2, Os8, Os10	Sb1, Sb4, Sb7	Ta1, Ta6, Ta7
Bd4	Os9, Os11, Os12	Sb2, Sb5, Sb8	Ta4, Ta5
Bd5	Os4	Sb6	Ta2

Table 2.20: Synteny between Brachypodium, rice, sorghum and wheat.

The five Brachypodium pseudo-molecules (Bd1-5) are shown alongside aligning chromosomes from rice, sorghum and wheat.

These results are in complete agreement with published work of shared syntenic blocks between grass genomes as displayed in the most recent ‘crop circle’ alignment of grass genomes (Devos 2005) and allow the Brachypodium chromosomes to be added to these alignments (Figure 2.33). For example, the alignments reported here show that Bd2 is syntenic to Os1 / Os5, Sb3 / Sb9, and Ta1 / Ta3 and the existing circular alignment of grass genomes indicates shared synteny between Os1, Sb3 and Ta3, and between Os5, Sb9 and Ta1. The agreement between genomic alignments from multiple species indicates that the syntenic relationships indicated are correct. Only Brachypodium chromosomes 2 and 5 remain unbroken in the alignments, the other three chromosomes appear to comprise of syntenic blocks that align to individual rice chromosomes. For example, Bd1 comprises 3 syntenic blocks that align to Os3, Os6 and Os7. This pattern of syntenic blocks arranged in different combinations is characteristic of the grass genomes.

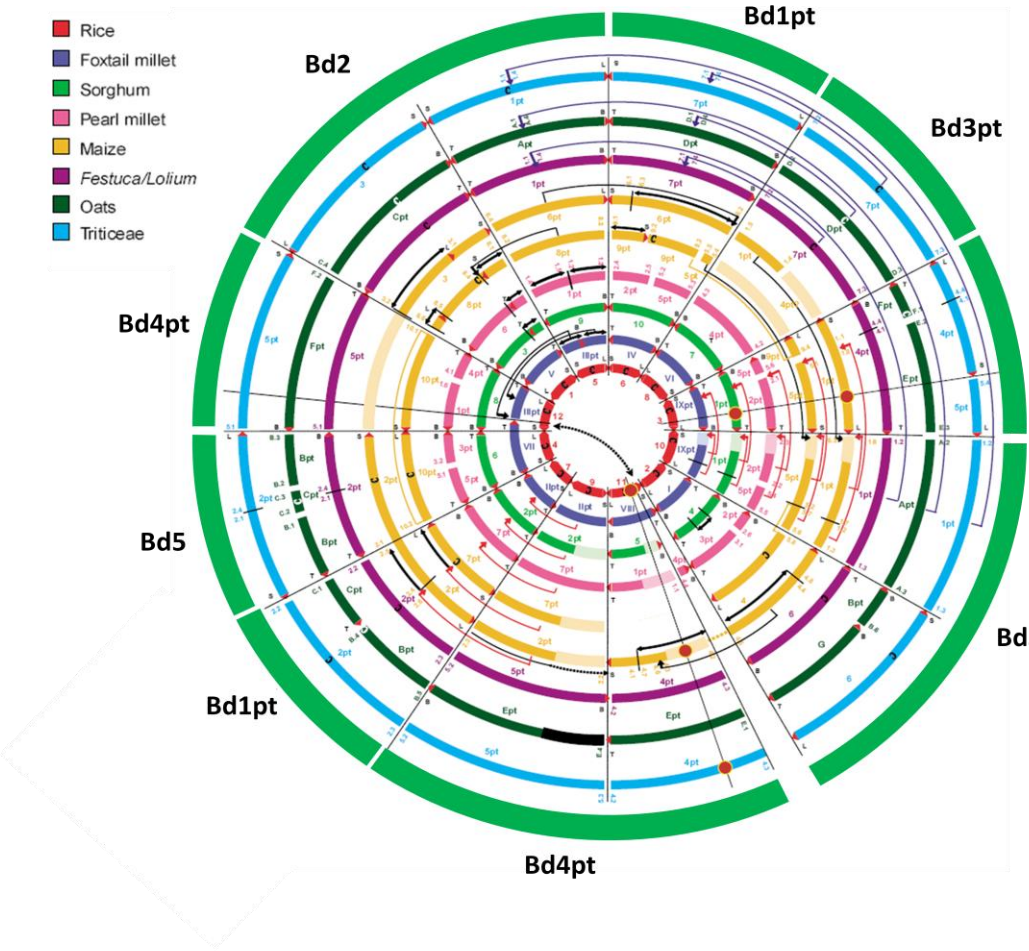


Figure 2.33: Brachypodium integrated into the crop circle alignment of grass genomes.

The Brachypodium pseudo-molecules are represented as the outermost circle, shown in green. Bd1 refers to pseudo-molecule 1 and 'pt' indicates that only part of the psuedo-molecule aligns to this syntenic block. (Original image obtained from Devos 2005)

2.7.2.2.2 Detailed sequence comparison between Brachypodium and rice

The dot-plot comparisons between Brachypodium pseudo-molecule assemblies and rice genome sequence were used again to guide more detailed sequence comparisons. Figure 2.34 shows a summary of these alignments with each Brachypodium chromosome represented as a block and colour-coded by sequence similarity to rice chromosomes. It is striking that three of the five Brachypodium chromosomes consist of bands of homology to rice chromosomes centred on the centromere. For example, Bd1 shows homology to Os6 around the centromere, to Os7 in the middle of both the long and short chromosome arms and to Os3 distally in both arms. If the ancestral grass genome was similar in structure to the rice genome (Salse *et al.* 2008a), this indicates that Brachypodium chromosomes have been formed by the insertion of one rice chromosome into the centromere of another and also provides a possible explanation for the reduction in chromosome number from rice (12) to Brachypodium (5). Ten precise syntenic breakpoints can be identified in the three largest Brachypodium chromosomes defining specific regions of homology to rice chromosomes. In Bd4 and Bd5 the pattern is less clear.

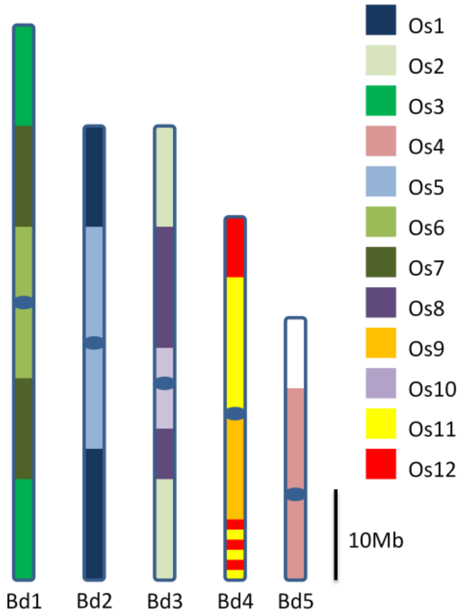


Figure 2.34: Brachypodium chromosomes aligned to rice.
Each Brachypodium chromosome is colour-coded by homology to rice chromosomes. All chromosomes show a pattern of nested chromosome insertion centred on the centromere.

A similar picture of nested homology was obtained for some wheat chromosomes when compared to rice using bin-mapped ESTs (Sorrells *et al.* 2003; La Rota and Sorrells 2004; Salse *et al.* 2008a). For example, wheat chromosome 1 shows similarity to rice chromosome 10 in the pericentromeric region and rice chromosome 5 at the distal ends of the chromosome. Wheat chromosome 7 shows similarity to rice chromosome 8 in the pericentromeric region and rice chromosome 6 at the distal ends of the chromosome. If this also indicates a mechanism of chromosome number reduction from rice (12) to wheat (7) it means that the reduction from rice to wheat occurred independently to the reduction from rice to Brachypodium as in Brachypodium, no single chromosome shows similarity to rice chromosomes 5 and 10 (as seen in wheat chromosome 1) or to rice chromosomes 6 and 8 (as seen in wheat chromosome 7). This pattern of nested homology to rice is also observed in *Ae. tauschii* and similarly, no rice chromosome insertions are shared between Brachypodium and *Ae. tauschii* supporting the hypothesis that reduction in chromosome number occurred independently in Brachypodium and the *Triticeae* (Luo *et al.* 2009).

Similar blocks of synteny are seen when comparing Brachypodium and wheat at a chromosome level, albeit slightly less clear due to the lower resolution of the comparison provided by the size of deletion bins on the wheat chromosomes (Figure 2.35). Four Brachypodium chromosomes (Bd1-4) show homology to a single wheat chromosome at the distal end of each chromosome arm. For example, Bd2 shows homology to Ta1 around the centromere and to Ta3 distally in each chromosome arm.

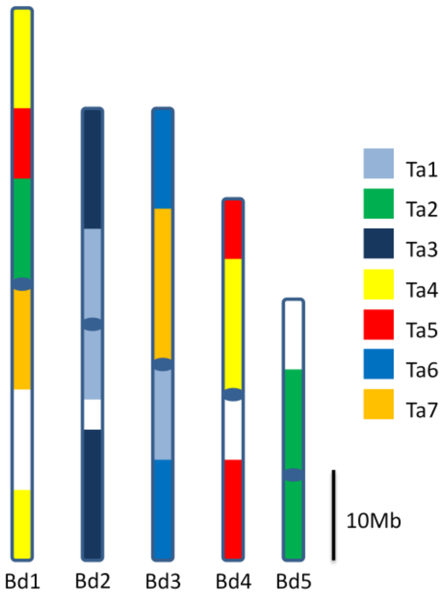


Figure 2.35: Brachypodium chromosomes aligned to wheat.

Each Brachypodium chromosome is colour-coded by homology to wheat chromosomes.

In addition to regions of strong sequence similarity, the sequence comparisons between Brachypodium and rice show regions of weaker similarity between chromosomes in the two species. For example, although the distal ends of Bd2s and Bd2l align to Os1 and the central part of each arm aligns to Os5, there is reduced similarity observed between these regions (Figure 2.36). This is evidence of the duplication previously identified between rice chromosomes 1 and 5 (Salse *et al.* 2004) which also exists in the Brachypodium genome. There are many other examples of weak homology in the comparisons between Brachypodium and rice and similar regions of duplication can be seen in the comparisons between Brachypodium and sorghum. However the picture is less clear due to the increased evolutionary distance between the temperate and the tropical grasses.

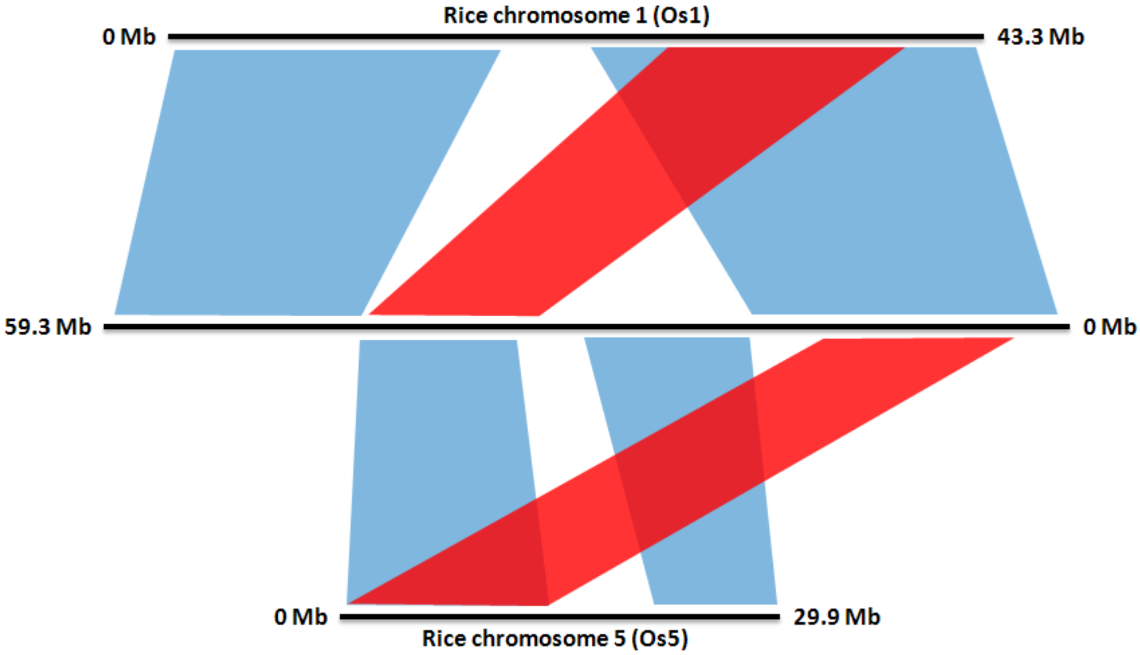


Figure 2.36: Secondary regions of sequence similarity between Brachpodium and rice.
Brachypodium chromosome 2 is shown in the center and secondary regions of similarity to rice chromsomes 1 and 5 are shown on red, in addition to primary regions of similarity shown in blue.

2.8 Discussion

The *Brachypodium* genus in the *Brachypodieae* tribe contains several wild grasses that are phylogenetically well placed in the grass family at the base of the temperate grasses (Catalan *et al.* 1995). The temperate grasses include several major crop species such as wheat and barley as well as important forage grasses (Kellogg 2001). *Brachypodium* species are estimated to have diverged from the other temperate grasses between 35 to 40 MYA (Bossolini *et al.* 2007), at least 10 MY after the divergence of the temperate grasses from rice (Paterson *et al.* 2004a). In contrast to the large genomes of many other temperate grasses, species within the *Brachypodium* genus have relatively small genomes (Bennett and Leitch 1995) and it was this property combined with its phylogenetic position in the grasses that first prompted the suggestion that *Brachypodium* species could be used for comparative analysis of the larger genome of wheat (Moore *et al.* 1993b). More recently, the same group used a *B. sylvaticum* BAC library to characterise the *Ph1* pairing locus in wheat (Foote *et al.* 2004; Griffiths *et al.* 2006) and found that a greater level of specific hybridisation was obtained using *B. sylvaticum* markers then was obtained using markers derived from rice, indicating that *Brachypodium* species would be more useful than rice for structural genomic studies in the *Triticeae*.

Brachypodium distachyon was first proposed as a model system for functional genomics in temperate grasses in 2001 (Draper *et al.* 2001). Brachypodium is a small grass, easily cultivated and with a short life cycle (Garvin *et al.* 2008). It is superior to rice, the existing monocot model, as rice is a large semi-aquatic tropical grass which is difficult to cultivate in Europe. In addition, rice does not exhibit many of the traits that exist in the temperate grasses such as resistance to specific pathogens, vernalisation and freezing tolerance. As well as being of use to the crop research community, Brachypodium possesses many traits that are relevant to research into novel uses of grasses such as that of a feedstock for biofuels (DOE 2006).

The International Brachypodium Initiative was formed in 2005 to establish resources to promote Brachypodium as an experimental system for grass research. These include a set of community standard lines (Vogel *et al.* 2006a), BAC libraries (Huo *et al.* 2006; Febrer *et al.* 2010), genetic markers and mapping populations (Garvin *et al.* 2008). The diploid inbred line Bd21 developed from the USDA germplasm collection was chosen as the primary community standard line (Garvin *et al.* 2008). In addition, a genome sequencing programme was initiated to sequence the genome of Bd21 using a WGS approach. This chapter describes three key resources that have been developed from Bd21 for Brachypodium genomics: a genetic linkage map, an integrated physical map and an annotated genome sequence. In addition, comparative genomics methods have been used to understand the relationship between Brachypodium and other grasses of major economic importance such as rice, sorghum, wheat and barley.

2.8.1 Genetic linkage mapping

A genetic linkage map represents the relative positions of genetic markers within linkage groups and is constructed from observations of the frequencies of recombination between these markers during crossing over of homologous chromosomes during meiosis. Genetic maps allow genomes to be navigated without knowledge of the underlying sequence. The genetic map described herein was the first Brachypodium genetic linkage map and used a mapping population of 183 F_2 plants derived from a cross between inbred diploid Brachypodium lines Bd21 and Bd3-1 (Vogel *et al.* 2006a; Garvin *et al.* 2008). A combination of simple sequence repeat (SSR) markers designed from Brachypodium EST and BAC sequences, conserved orthologous sequence (COS) markers from other grasses and additional Brachypodium-derived molecular markers was used. Genetic linkage mapping was performed at the same time as WGS sequencing meaning that during the construction and analysis of the genetic map we had access to a substantial amount of genome sequence. This allowed me to analyse the three builds of the genetic map by aligning the linkage groups of each build to the WGS sequence assemblies. In Build 1 of the genetic map, the linkage groups showed good consistency with the super-contigs of the mid-point assembly although larger linkage groups showed multiple distinct regions of synteny indicating possible false linkages. Based on these alignments, specific areas of the super-contigs were identified to design new markers which were screened against the mapping population and included in the next build. Aligning the Build 2 linkage groups to the super-contigs again showed that the larger linkage groups aligned to multiple super-contigs. These alignments were also used to determine if linkage groups could be joined and joins were predicted between three pairs of linkage groups. One of the predicted joins was tested by selecting BACs from super-contig 0 that defined the regions of alignment with linkage groups 1 and 3. Using these BACs as FISH probes indicated that super-contig 0 represented a large portion of a single Brachypodium chromosome arm (Figure 2.13) and that based on the alignment of this super-contig with linkage groups, linkage groups 1 and 3 genetically represented this chromosome arm. Linkage group 1 also showed good alignment to super-contigs 1, 3 and 5 (Table 2.6) indicating that these super-contigs also formed part of the same chromosome. This experiment located a portion of assembled genomic sequence into its correct chromosomal context although at this point it was not known to which chromosome the super-contig was anchored as no reference BAC were used.

Additional BACs were selected using the linkage group alignments in order to cytogenetically anchor all the linkage groups. These BACs were used as probes alongside reference BACs so that the actual chromosome location could be determined. These alignments supported the previous hypothesis that super-contig 0 represented a single chromosome arm by virtue of its assignment to 1S, the short arm of the largest Brachypodium chromosome. In addition, the cytogenetic anchoring indicated that super-contig 1 represented the other arm of the same chromosome. My initial alignments obtained by locating marker sequences on the super-contigs indicated that linkage group 1 showed good alignment to super-contigs 0, 1, 3, and 5 (Table 2.6) and therefore might represent these super-contigs. However, the results from the cytogenetic anchoring showed that BACs selected from super-contigs 3 and 5 hybridised to both arms of chromosome 3 therefore highlighting inconsistencies within this linkage group and indicating that it should be broken up into two smaller linkage groups, one representing chromosome 1 and another representing chromosome 3. This was also observed with linkage group 2 which showed good alignment to super-contigs 7, 9 and 10 (Table 2.6). Cytogenetic analysis also assigned super-contigs 7 and 9 to different chromosomes (4L and 5 respectively) meaning that linkage group 2 also probably contains a false join.

The conclusion from analysis of Build 2 of the genetic map was that the linkage groups in this build appeared to contain many false linkages and for this reason Build 3 was created using more stringent criteria for accepting genetic linkage between markers. Build 3 contained 20 linkage groups with the 5 largest linkage groups encompassing 64 % of the genetic map distance and potentially representing the five chromosomes of Brachypodium. When aligned to the

Brachypodium pseudo-molecules (Figure 2.14), these linkage groups showed good synteny with occasional breaks where individual markers landed on different chromosomes. In addition, one linkage group (c) contained a probable false linkage, aligning to both Bd1 and Bd4. The other 4 largest linkage groups aligned to individual chromosomes. In general, the alignments indicated that the linkage groups in Build 3 are accurate representations of small portions of the underlying chromosomes but the marker density on the map is not sufficiently high to coalesce these linkage groups further. Only 11 % of the SSR markers used in the map identified polymorphisms between the mapping parents (Bd21 and Bd3-1) meaning that many potential markers could not be genetically mapped. This success rate indicates a low level of EST-SSR marker polymorphism between the mapping parents. A more recent study used SSR markers to evaluate genetic diversity in a large Turkish diploid Brachypodium collection and also included the original inbred lines used for genetic mapping (Vogel *et al.* 2009). The Brachypodium accessions clustered into two major groups and interestingly, Bd21 and Bd3-1 were phylogenetically distant within one of the groups, indicating they are genetically quite diverse.

Aligning the Brachypodium linkage groups to the rice genome sequence showed a complex relationship between the genomes with markers from a single linkage group hitting many rice chromosomes. In syntenic regions, many local rearrangements were observed as well as adjacent markers hitting the same rice gene. No real understanding of the relationship between Brachypodium and rice could be gained from these comparisons due the fragmentary nature and low resolution of this first Brachypodium genetic map. Increasing the marker density by using a wider range of markers should decrease the number of linkage groups to better resemble the five Brachypodium chromosomes and will allow a more complete comparison to genetic, physical and sequence maps of other grasses.

The FISH analysis performed on linkage groups from Builds 1 and 2 of the genetic linkage map was not performed on the Build 3 linkage groups but the chromosomal assignment of super-contigs provided a basis for subsequent FISH analysis on the sequence contigs which proved an essential method to validate the WGS sequence assembly.

2.8.2 Validation of the WGS genome assembly

The diploid inbred line Bd21 (Garvin *et al.* 2008) was sequenced using a WGS approach by the JGI. Twelve clone libraries were prepared from Bd21 with insert sizes ranging from 3 kb to 150 kb. These were end-sequenced and assembled using the ARACHNE assembler (Batzoglou *et al.* 2002). The genome was initially sampled to a depth of 4x at which point a mid-point assembly was released consisting of 1,015 super-contigs comprising 281.2 Mb of sequence. The pre-release assembly, representing 9.4x genome coverage, was released at the end of the sequencing phase of the project and this consisted of 216 super-contigs comprising 273.4 Mb of sequence. Two of the clone libraries sequenced by the JGI were also fingerprinted and end-sequenced which allowed a physical map to be constructed using FPC (Soderlund *et al.* 1997). FPC-derived contigs were aligned to the WGS sequence assemblies to assess their accuracy and to improve both the physical map and the sequence assembly. Validation methods such as comparison to physical maps are essential to assess sequence assemblies produced by WGS sequencing, especially for complex genomes (Meyers *et al.* 2004). In addition, aligning the physical map contigs to the sequence assemblies allowed extensive validation of both the physical map contigs and the sequence assemblies using FISH.

The two BAC libraries were made from Bd21 nuclear DNA using the restriction enzymes *Eco*RI and *Hin*dIII. An average insert size of 126 kb was used and a total of 36,864 clones were fingerprinted generating 58,894 BES comprising 41.1 Mb of sequence. FPC assembled 208 contigs from 30,195 fingerprinted clones. The physical map generated by Gu *et al.* (2009) was made from 67,151 *Hin*dIII- and *Bam*HI-derived clones with an average insert size of 100 kb which assembled into 671 contigs. Both maps used a similar Sulston score to assess significant fingerprint overlaps to build the contigs. This suggests that the longer insert size used to construct the BAC libraries for this physical map allowed more contiguous regions to be assembled. In addition, the use of the *Eco*RI restriction enzyme may have allowed cloning of a more representative set of genomic regions leading to a more accurate assembly in this physical map.

The 208 FPC-derived contigs were aligned to the 17 largest mid-point sequence assemblies using the BES by SyMAP (Soderlund *et al.* 2006). A high level of synteny was observed between the physical map contigs and the sequence assemblies indicating a high level of accuracy in both the sequence assemblies and the FPC contigs. Aligning FPC contigs to the sequence assemblies allowed the FPC contigs to be ordered and oriented according to the WGS sequence assemblies. These alignments also indicated joins between physical map contigs where two contigs aligned to a single super-contig. In addition, these alignments indicated possible merges between WGS sequence contigs where a FPC contig aligned to the ends of two sequence contigs. This showed the benefit of having two independently constructed representations of the genome, each approach has different strengths and combining the approaches improved both the sequence assembly and the physical map. The alignment of FPC contigs to sequence assemblies resulted in a physical map consisting of 35 contigs assembled from 26,800 clones aligned to 14 sequence contigs (11 original and 3 merged) with an estimated physical length of 303.7 Mb.

An important and novel aspect of assessing the WGS sequence contigs was the analysis facilitated by the physical map anchored to the sequence assembly. I developed a Perl script to estimate the repeat content of BACs based on the predicted position of the BAC on the sequence assembly. BACs containing low amounts of repetitive DNA were selected from the physical map from regions defining the ends of each sequence contig. These BACs were used as FISH probes to determine whether the physical map contigs and WGS sequence contigs accurately reflected the underlying chromosomes. In addition, it enabled me to identify the chromosomal locations of super-contigs prior to having access to chromosome-scale pseudo-molecules. The FISH analysis of the genetic linkage groups had resulted in the assignment of 6 super-contigs from the mid-point assembly to chromosomes (Table 2.7). Additional BAC probes were identified from the physical map to define the pre-release assembly super-contigs and these were anchored to the Brachypodium cytogenetic map confirming the previous anchoring results and indicating that the WGS sequence assemblies were accurate. This experiment also located two additional two super-contigs onto chromosomes. Mis-assemblies in pre-release super-contigs 2 and 4 were indicated by this experiment.

The five pseudo-molecules of the final Brachypodium genome assembly were constructed by aligning the pre-release super-contigs to a high-density genetic map. These alignments highlighted the mis-assembles in super-contigs 2 and 4 previously identified by the cytogenetic analysis. These super-contigs were broken in the final assembly (Figure 2.5). The final assembly consisted of five pseudo-molecules comprising a total of 271.9 Mb.

The FPC-derived physical map contigs were aligned to the five pseudo-molecules in the final sequence assembly by BES using SyMAP and again good congruence was observed between the physical map contigs and the pseudo-molecules. Twenty-six physical map contigs containing 26,800 clones aligned to the pseudo-molecules. The total length of the physical map contigs was 303.4 Mb based on an approximation of the band size of fingerprints. This is slightly larger than the total length of the pseudo-molecules (272 Mb) but is consistent with a tendency for the size of physical map contigs to be overestimated when calculated from the band size of fingerprints. In contrast, the existing Brachypodium physical map comprises contigs totalling 410 Mb which suggests that many contigs in this map overlap (Gu *et al.* 2009). The integration of the WGS sequence assembly and the physical map contigs allowed a more accurate physical map to be assembled with more contiguous sequences.

The alignment between physical map contigs and pseudo-molecules was used to select probe BACs for FISH to validate joins between pre-release super-contigs and to locate the ends of pseudo-molecules that had not already been validated. This analysis enabled all the pseudo-molecules to be validated cytogenetically (Figure 2.23). A 1.5 Mb region was identified at the distal end of Bd4S where two probe BACs hybridised to Bd2L and this could not be fully explained, however this issue will be resolved when high-resolution chromosome painting is performed on this chromosome. To provide additional evidence of the agreement between the WGS sequence assemblies, the physical map contigs and the cytogenetic map, chromosome-specific BAC pools were identified and hybridised alongside reference BACs to identify each chromosome. With one exception, BACs from each pool hybridised to the expected chromosome and provided further evidence that the pseudo-molecules accurately reflect the physical chromosomes of Brachypodium. The Brachypodium chromosome painting experiment is ongoing but my script to screen BACs for repetitive DNA was particularly useful to select BAC clones containing low amounts of repetitive DNA as accurate hybridisation is required to obtain clear images. At the present time, BAC pools have been identified for all 10 chromosome arms and pools from Bd1 and Bd2 have resulted in high resolution images. These results not only indicate good agreement between the physical map, the sequence assembly and the cytogenetic map at a much higher level of resolution than was previously obtained using single BACs but also open the door for further chromosome studies in Brachypodium and related grasses.

In order to provide an integrated genomic resource and to assess synteny between the two maps, the genetic linkage groups from the recently developed genetic map (Garvin *et al.* 2010) were anchored to the physical map contigs using two complimentary approaches. Fourteen of the 20 linkage groups could be anchored unambiguously, three linkage groups were anchored in two possible positions and three small linkage groups could not be anchored. These alignments agreed with previous alignments of linkage groups to pseudo-molecules made by locating the marker sequences directly onto the pseudo-molecules (Figure 2.14). In addition, combining these results with the previous alignments allowed ambiguously anchored linkage groups to be clarified. For example, the previous results aligned linkage group ‘g’ to Bd1 whereas this anchoring indicated Bd1 or Bd5, so Bd1 is the more likely candidate, being supported by two independent methods. The same is true for linkage group ‘a’ which was previously aligned to Bd3 but in this case was anchored to either Bd1 or Bd3. Bd3 is the more likely candidate as it is supported by three independent methods. Interestingly, aligning linkage group ‘c’ to the physical map contigs once again highlighted the false join that was previously identified and indicates that this linkage group should be split into two as markers at one end of the linkage group align to Bd1 and markers at the other end to Bd4. The long range collinearity between the physical and the genetic map supports the accuracy of the physical map contigs and provides a resource for map-based gene isolation.

The WGS sequence assemblies produced by the JGI were of high quality and were validated by integration with the physical map as well as additional cytogenetic validation. This high quality sequence assembly will provide an important resource for understanding the genomes of other pooid grasses with large and complex genomes such as wheat and barley.

2.8.3 Genome annotation

The release of the mid-point assembly in August 2007 provided the first genome-scale view of Brachypodium and my annotation of the 281 Mb of sequence was essential to gain an understanding of the structure and properties of this first pooid grass genome. The gene prediction program FGENESH identified 39,228 genes in the repeat-masked assembly, giving a gene density of 13.9 genes per 100 kb which falls between that of the smaller genome of *Arabidopsis* (TAIR8: 22.64 genes per 100 kb) and the larger genome of rice (RAP2: 7.39 genes per 100 kb). However, this estimate is higher than a previous estimate of gene density in Brachypodium (7.14 genes per 100 kb) obtained from BES representing 10.9 % of the Brachypodium genome (Huo *et al.* 2008). Less then 40 % of the gene predictions are supported by EST alignments suggesting that this estimate is high. FGENESH is an a*b initio* gene prediction program which bases predictions purely on DNA sequence. The incorporation of additional evidence for genes such as EST alignments would probably reduce the number of predicted genes. The GC-content of the assembly (46.2 %) is similar that obtained from Brachypodium BAC-end sequences (45.2 %) and to other monocot genomes (SanMiguel *et al.* 2002; Yu *et al.* 2002; Huo *et al.* 2006). The GC content is higher in genic regions (57.1 %) than in the assembly as a whole which is a general trend in plant genomes (Montero *et al.* 1990). The simple sequence repeats and flanking primer sequences identified in the super-contigs provided a source of potential molecular markers for genetic mapping.

The transposable element content of the mid-point assembly was estimated as 7.1 %, with more than 80 % of these classified as class I retrotransposons. This trend is observed in other sequenced grass genomes, for example, around 30 % the rice genome consists of transposable elements, two thirds of which are retrotransposons (International Rice Genome Sequencing Project 2005). Analysis of Brachypodium BES estimated that 7.87 % of the genome consisted of retrotransposons (mainly LTR-retrotransposons) and 1.28 % was due to DNA transposons (Huo *et al.* 2008). My estimation of transposable element content within the super-contigs is slightly lower than this but the proportions of class I to class II elements is similar. I did not attempt to identify repeats specific to Brachypodium, estimated to comprise 7.4 % of the genome from BES analysis (Huo *et al.* 2008). A *de novo* repeat finding strategy would be required to fully annotate the repetitive portion of the Brachypodium super-contigs.

The five pseudo-molecules of the final assembly were annotated by teams at MIPS and JGI. In total, the pseudo-molecules comprised 272 Mb of sequence and 25,532 protein-coding loci were predicted containing 32,532 protein-coding transcripts. These predictions were performed using *ab initio* methods as well as evidence-based prediction incorporating the ESTs generated by the JGI as part of the sequencing project. This gene complement is within the same range as those of rice (Tanaka *et al.* 2008) and sorghum (Paterson *et al.* 2009) having 28,236 and 27,640 genes respectively. This gives an average gene density in Brachypodium of approximately 10 genes per 100 kb. Annotation of the final genome assembly also quantified the fluctuations in gene density over the chromosomes showing that gene density was highest at the distal ends of chromosomes with very few genes found near the centromeres (Figure 2.37). This distribution of genes is characteristic of other sequenced grass genomes (International Rice Genome Sequencing Project 2005; Paterson *et al.* 2009; Choulet *et al.* 2010).

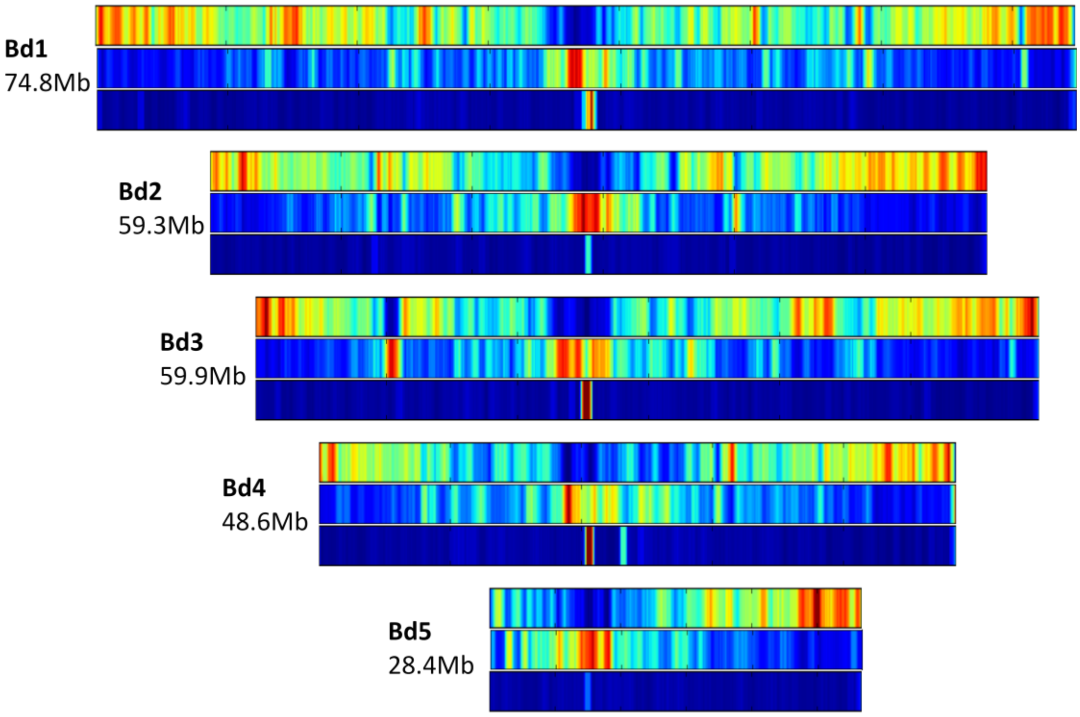


Figure 2.37: Structural characteristics of the Brachypodium genome.
The five pseudo-molecules (Bd1-5) are shown as three heat maps with the size each pseudo-molecule to the left. The heat maps show gene density on the top, LTR-retrotransposon density in the middle and centromeric repeats on the bottom with red indicating high density and blue indicating low density. Gene density is highest at the distal end of chromosome arms and LTR-retrotransposon density is highest around the centromere (heat maps courtesy of Heidrun Gundlach, MIPS).

The transposable element content of the Brachypodium genome was estimated to be 28.1 %; comprising 23.3 % class I retrotransposons and 4.8 % class II DNA transposons. The majority of class I elements were LTR-retrotransposons (21.39 %) and the examples from all the major class

II families were found with *gypsy* elements being the predominant superfamily. Representatives from the DNA transposon superfamilies *CACTA*, *hAT*, *Mutator*, *Tc1*/*Mariner* and *Harbinger* were identified as well as both *stowaway*- and *tourist-like* MITEs and *helitrons*. The transposable element content of Brachypodium is consistent with that observed for other sequenced grasses where a similar proportion of each genome comprises DNA transposons and the proportion of retrotransposons increases with genome size (International Rice Genome Sequencing Project 2005; Paterson *et al.* 2009; Schnable *et al.* 2009). Brachypodium has the lowest transposable element content of any grass genome sequenced to date as well as the lowest proportion of DNA transposons. Its retrotransposon content is slightly higher than that reported for rice. A transposable element from the *Harbinger* superfamily was identified containing a portion of a gene belonging to the nucleotide-binding site–leucine-rich repeat (NBS-LRR) family which implicates this class of transposable element in gene mobility, a phenomenon already associated with Pack-MULEs and *helitrons* (Kapitonov and Jurka 2001; Jiang *et al.* 2004).

By determining the number of synonymous substitutions that have occurred in the long terminal repeats of inserted retrotransposons, the approximate date of retrotransposon insertion can be estimated (SanMiguel *et al.* 1998). In the Brachypodium genome, many retrotransposons appear to have been inserted within the last 100,000 years which is evidence of much more recent activity than has been observed in the rice (Wicker and Keller 2007) and wheat genomes (Choulet *et al.* 2010). The small size of the Brachypodium genome indicates that this retrotransposon activity is somehow mitigated to prevent genome expansion. An estimate of the amount of DNA lost due to LTR : LTR recombination can be calculated from the number of solo LTRs identified in a genome as a solo LTR is generally left behind when an LTR-retrotransposon is excised by unequal homologous recombination (Vicient *et al.* 1999). In Brachypodium, 1,814 solo LTRs were identified equating to 17.4 Mb of sequence lost due to excision of LTR-retrotransposons. This indicates that in Brachypodium, genome expansion by retrotransposons is countered by their removal by recombination. In the Triticeae lineages, retrotransposons are not removed to such an extent leading to genome expansion (Wicker and Keller 2007).

Particularly low gene density was identified on the short arm of chromosome 5 compared to the rest of the genome (Figure 2.37). This chromosome arm also contains a high number of LTR-retrotransposons, many of them young and few solo LTRs are found. These observations indicate that in this region of the genome, LTR retrotransposons are accumulating and not being removed by recombination. This genomic region is syntenic to the short arms of rice chromosome 4 and sorghum chromosome 6 which also show low gene density and high retrotransposon density (International Rice Genome Sequencing Project 2005; Paterson *et al.* 2009).

2.8.4 Comparative genomics

At each stage of the genome assembly, the Brachypodium WGS sequence assemblies were compared with rice, wheat, barley and sorghum to provide the first genome-wide alignments between the three major grass sub-families. At the mid-point stage of genome assembly, I compared the super-contigs with rice, wheat and barley. The super-contigs showed good collinearity at the sequence level to rice with small rearrangements breaking the pattern of conservation. The comparison obtained between wheat and Brachypodium was at a lower resolution as deletion bin-mapped ESTs (Qi *et al.* 2004) were used, however good alignments were obtained at the chromosome-level that were supported by alignments to barley obtained from genetically mapped barley ESTs aligned to Brachypodium. In addition, the alignments of Brachypodium super-contigs to rice and wheat agreed with previous alignments between rice and wheat (La Rota and Sorrells 2004). Wheat ESTs aligning to rice chromosome 1 also aligned to super-contigs 2, 4 and 13 which are the super-contigs that align to wheat chromosome group 3 predicted by my alignments providing an additional method of validation. Aligning each pseudo-molecule of the final assembly to rice, sorghum and wheat allowed the Brachypodium chromosomes to be added into the classic crop circle alignment of grass genomes (Devos 2005) alongside their syntenic regions in rice, millet, sorghum, maize, festuca, oat and wheat (Figure 2.33).

At a more detailed level, comparisons between the Brachypodium pseudo-molecules and rice showed a pattern of nested homology to rice chromosomes centred on the centromere (Figure 2.34). This indicated a possible method of reduction in chromosome number from rice (12) to Brachypodium (5) as the insertion of multiple rice chromosomes into the centromeres of other chromosomes. For example, by this mechanism Bd1 would be formed by Os6 inserting into the centromere of Os7, followed by insertion of this new chromosome into Os3 (Figure 2.38). Multiple syntenic breakpoints were identified in each Brachypodium chromosome delineating blocks of homology to each rice chromosome.

This pattern of nested chromosome insertion has previously been observed when comparing the linkage groups of rice to the panicoid grasses, for example, maize, sorghum and sugarcane (Kellogg 2001) and a comparative map of finger millet (a grain crop cultivated in East Africa and Southern India) to the rice genome (Srinivasachary *et al.* 2007). In addition, comparison of the recent *Ae. tauschii* genetic map to rice and sorghum identified nested chromosome insertions as the likely mechanism by which the reduction in chromosome number from rice (x = 12) to

Triticeae species (x = 7) and sorghum (x = 10) had occurred (Luo *et al.* 2009). The insertion of one chromosome into the centromere of another with no overall loss in gene content is thought to be the dominant mechanism of chromosome number reduction in grasses and these results confirm that this is also the case in Brachypodium (x = 5). This provides confirmation that the nested insertion mechanism is a unifying mechanism that describes the evolution of grass chromosomes and chromosome synteny. In addition, comparing these alignments to previous rice-wheat alignments (Salse *et al.* 2008a) indicated that the reduction in chromosome number had occurred independently in the Brachypodium and wheat lineages.

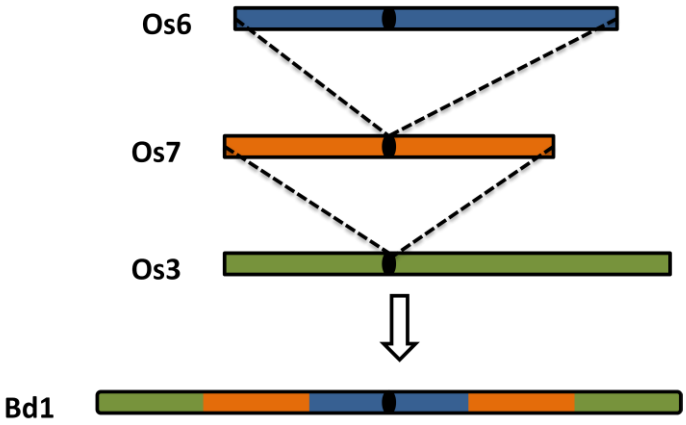


Figure 2.38: Possible mechanism of chromosome reduction from rice (12) to Brachypodium (5).
This explains the pattern of nested homology observed between Brachypodium and rice. Multiple rice chromosomes are inserted into the centromeres of other chromosomes to form a new chromosome.

The final Brachypodium genome analysis extended the comparisons presented here by using the stringent alignment criteria and statistical validation that had already been applied to the genomes of rice, wheat, sorghum and maize (Salse *et al.* 2008a) to identify duplications within the Brachypodium genome and duplications shared between Brachypodium and other grasses. This more sophisticated analysis distinguishes paralogous and orthologous relationships and identified six major chromosomal duplications in the Brachypodium genome. In addition, identification of orthologous genes-pairs between Brachypodium, rice, sorghum and the *Triticeae* defined 21,045 orthologous relationships which were consistent with the current evolutionary model of grasses originating from an ancestral grass genome containing five chromosomes via a 12 chromosome intermediate (Salse *et al.* 2008a). The Brachypodium genome sequence provided the first opportunity to add a complete pooid grass genome sequence into this model and thus was key to confirm the existing model. This analysis identified the nested insertion of a whole rice chromosome into the centromere of another chrosmome as a mechanism of chromosome reduction. This observation was also extended to sorghum and barley. Detailed analysis also identified ‘footprints’ of chromosome insertions at the syntenic breakpoints where centromeric repeats from the inserted chromosome still existed in the target chromosome. In addition, retrotransposon density was high around these insertion sites as would be expected at a centromere and high gene density could still be observed at the former distal end of the inserted chromosome. This more detailed analysis confirmed that the reduction in chromosome number had occurred independently in the Brachypodium and *Triticeae* lineages as no chromosome fusions were shared between these genomes.

In the final genome analysis, Brachypodium genes were clustered with sorghum and rice genes in addition to wheat ESTs to represent wheat genes. This identified orthologues between the four species and allowed the time of divergence of these four grasses to be estimated by measuring the mean synonymous substitutions rates between syntenic gene-pairs. It was estimated that sorghum (representative of the panicoid grasses) diverged from the rice lineage 45-60 MYA, the pooid grasses (wheat and Brachypodium) diverged from the rice lineage 40-54 MYA and Brachypodium diverged from wheat 32-39 MYA. These estimates agree with previous results obtained from small gene samples (Gaut 2002; Charles *et al.* 2009). In addition, ancestral whole-genome duplication was identified in the Brachypodium genome, and estimated to have occurred 56-73 MYA, consistent with the ancient whole-genome duplication that predated the divergence of the grasses previously identified in rice and sorghum (Yu *et al.* 2005; Paterson *et al.* 2009).

Brachypodium is the first member of the Pooideae grass subfamily to be sequenced. Its small genome and a phylogenetic position in the grass family make it an ideal model for agriculturally important grasses such as wheat and barley as well as potential biofuel crops. The genome resources described in this chapter such as the first genetic map, a physical map, an annotated genome sequence and comparison of this genome to other grasses comprise a major contribution towards the adoption of Brachypodium as a model organism for a range of applications in plant genomics. In particular, having a well annotated pooid grass genome establishes a template for analysis of the large genomes of other pooid grasses such as wheat.

3 Identification of conserved non-coding sequences (CNS) in the Brachypodium genome

3.1 Relevant publications

International Brachypodium Initiative (2010). "Genome sequencing and analysis of the model grass *Brachypodium distachyon*." Nature **463**(7282): 763-768.

3.2 Introduction

The protein-coding regions and intron-exon structures of genes can generally be accurately predicted *ab initio* from DNA sequence. However, there is a wealth of other information encoded by DNA sequences that have thus far proved to be more difficult to analyse. For example, the identification of regulatory elements controlling the spatiotemporal expression of genes is an essential step towards understanding the complex regulatory networks that operate within a cell. Regulatory elements are short DNA sequences located in noncoding DNA that control gene expression by binding transcription factors at specific sites within the DNA sequence (Kadonaga 2004). These sites are called transcription factor binding sites (TFBS) or *cis*-regulatory elements, meaning regulatory elements that are located on the same strand and in close proximity as the regulated gene.

Experimental identification of regulatory elements is time consuming and in most cases can only be applied to individual genes. DNA footprinting was developed to study the interaction of DNA binding proteins with their target sites (Galas and Schmitz 1978). First, PCR is used to amplify the DNA region to be studied and the DNA is radioactively or fluorescently labelled. The sample is divided into two, and one sample has the protein of interest added. A cleavage agent such as deoxyribonuclease (or DNase) is added to both samples at a low concentration to ensure that each DNA strand will be cut once at a random position along its length. Both samples are separated by electrophoresis and the pattern of fragments compared. If the protein has successfully bound to the DNA it will protect the DNA strand from enzymatic cleavage and thus a ‘footprint’ of missing bands will be observed in one sample. A technique called a mobility shift assay also uses electrophoresis to determine whether a protein has bound to a DNA sequence (Garner and Revzin 1981). Again, two samples are required, one containing a labelled DNA fragment with no protein present and the other containing the same labelled DNA fragment and the protein expected to bind to it. If the protein binds to the DNA fragment, the resulting complex will be larger and will this move more slowly in the polyacrylamide gel and a ‘shift’ in the resulting gel band will be observed.

In 1988, the chromatin immunoprecipitation (ChIP) assay was developed to determine whether specific proteins such as transcription factors bind to sites within a genomic region *in vivo* (Solomon *et al.* 1988). Cells are first treated with formaldehyde to generate stable cross-links between bound proteins and DNA sites. The chromatin is isolated from the cells and the DNA is sheared into small fragments. An antibody specific to the DNA-binding protein being investigated is used to isolate the protein-DNA complex by precipitation, then the protein is unbound and PCR is used to amplify specific DNA sequences to determine if they were precipitated with the protein. More recently, ChIP has been combined with high-throughput sequencing (ChIP-Seq) to provide a genome-wide approach that can be used to identify regulatory elements (Johnson *et al.* 2007). DNA sites with bound transcription factors are isolated *in vivo*, sequenced and aligned back to the target genome to determine their binding locations.

Phylogenetic footprinting (Tagle *et al.* 1988) is an *in silico* technique used to identify regions containing regulatory elements and exploits the fact that functional sequences tend to evolve more slowly than non-functional sequences as they are subject to selective pressure. Comparison of orthologous sequences from different species enables these potentially functional regions to be identified within the non-functional surrounding sequence. The success of phylogenetic footprinting was pivotal in the decision to sequence the mouse genome in order to identify regulatory regions in the human genome sequence (Hardison *et al.* 1997). These regions are called conserved noncoding sequences (CNSs). Phylogenetic footprinting was used to identify ninety CNSs from a comparison of 1 Mb orthologous regions in human and mouse (Loots *et al.* 2000). The largest sequence identified (401 bp) was analysed for function *in vitro* using transgenic mice and appeared to function as a coordinate regulator of three genes, *interleukin-4*, *interleukin-13*, and *interleukin-5* by modification of chromatin structure. Nobrega and colleagues used phylogenetic footprinting between human and mouse to identify regulatory regions surrounding the human *DACH* gene, a gene involved in the development of brain, limbs and sensory organs (Nobrega *et al.* 2003). A subset of these elements were tested *in vivo* using a reporter assay in transgenic mice and were found to drive expression of the reporter gene. Although phylogenetic footprinting will only identify regulatory regions under selective constraint, *in silico* methods are significantly cheaper and faster than experimental methods and can be used to inform experimental methods.

Following its use in mammalian systems, phylogenetic footprinting was used to identify CNSs in plants from comparisons between orthologous regions of rice and maize (Kaplinsky *et al.* 2002; Inada *et al.* 2003) and a more extensive study using maize, rice, barley, wheat and sorghum (Guo and Moose 2003) . These studies indicated that plant CNSs are characteristically different to mammalian CNSs; plant CNSs are much shorter and plant genes have fewer CNSs within their gene-space with no CNSs found around some genes. In mammals, CNSs are detected throughout the genome and can regulate genes several kilobases away (Loots *et al.* 2000). Conversely, plant CNSs tend to be found much closer to the genes they regulate. In general, plant CNSs appear to be less complex than mammalian CNSs possibly indicating less complex regulatory mechanisms in the *Plantae* kingdom (Kaplinsky *et al.* 2002; Guo and Moose 2003; Inada *et al.* 2003). It has been hypothesised that due to the increased redundancy of genetic material in plants arising from various gene duplication processes, genes do not require such complex regulatory mechanisms as a duplicated gene can develop a new or more specialised function after duplication (Lockton and Gaut 2005).

The completion of genome sequencing projects in *Arabidopsis* (The Arabidopsis Genome Initiative 2000) and rice (International Rice Genome Sequencing Project 2005) enabled genome-wide identification of CNSs from these genomes by comparison of paralogous gene-pairs resulting from intra-genomic duplication (Thomas *et al.* 2007; Li *et al.* 2009d). These were called α-CNSs as they are identified from paralogous rather than orthologous comparisons. In *Arabidopsis*, nearly 15,000 α-CNSs were identified, located upstream, downstream and within the introns of protein-coding genes, most commonly in the 5' promoter region. This investigation also showed that the typical *Arabidopsis* gene contains approximately 1.7 α-CNSs and that genes with a high number of α-CNSs tend to be associated with GO terms relating to transcription factor activity and environmental, metabolic or pathogenic stress. Genes with no α-CNSs are associated with GO terms inferring housekeeping and basal metabolic processes (Thomas *et al.* 2007).

In silico identification of CNSs requires pairwise local alignment of orthologous DNA sequences. These sequences should be from species within a “window of useful divergence” (Freeling and Subramaniam 2009). If the evolutionary distance between the DNA sequences being compared is too small, mutations will not have had time to accumulate sufficiently to distinguish functional from non-functional DNA and if the evolutionary distance is too large, all but the most conserved regions will have diverged to such an extent as to not be detectable. Plant CNSs are less conserved than vertebrate CNSs so more sensitive detection methods are required. The method used in the first plant CNS study has been adopted in most subsequent studies (Kaplinsky *et al.* 2002). Sequences are aligned using BLASTN and syntenic HSPs with E-values less than the E-value resulting from a 15/15 exact base pair match are considered to be CNSs (Thomas *et al.* 2007). Local alignment tools such as BLASTN have been shown to perform better than global alignment methods for the detection of plant CNSs (Lyons and Freeling 2008).

In addition to phylogenetic footprinting, potential regulatory motifs can be identified by searching the promoter regions of functionally related genes. Large databases of gene expression data such as Genevestigator (Zimmermann *et al.* 2004) are used to identify genes that are co-expressed at specific developmental stages or in response to a specific environmental condition and this is used to infer related function. Motifs that are statistically overrepresented in the promoter regions of functionally related genes compared to a background model may be potential regulatory elements. Combining evolutionary conservation and co-expression analysis to identify potential functional motifs increases the predictive power in comparison to relying on a single source of information, as motifs that are present in the promoters of functionally related genes and that are conserved between orthologues are likely candidates for function. Software such as PhyloCon is used to detect motifs by comparing the promoter sequences of co-regulated genes in one species to the promoter regions of orthologous genes in related species (Wang and Stormo 2003). PhyloCon has been used to identify functional motifs in the promoter regions of rice and sorghum genes (Wang *et al.* 2009b) but the lack of large gene expression datasets in Brachypodium precludes this type of analysis at the present time.

Very few CNSs in plants have been experimentally validated. One study investigated a gene found in grasses called *Knotted1* which has two 5' CNSs containing regulatory elements that control the expression of the gene (Uchida *et al.* 2007). Another study showed that insertion of a MITE transposable element into a 5' CNS of a gene involved in flowering time in maize lowered the expression levels of the gene, resulting in late flowering (Salvi *et al.* 2007).

This chapter describes the *in silico* identification of conserved noncoding sequences in the Brachypodium genome by pairwise alignment of putative orthologues from Brachypodium, rice (International Rice Genome Sequencing Project 2005) and sorghum (Paterson *et al.* 2009). These grasses represent three major subfamilies of economically important grasses: the Pooideae (Brachypodium), the Ehrhartoideae (rice) and Panicoideae (sorghum). The level of evolutionary divergence between these three grasses ensures that non-functional sequence will have been randomised by background mutation and will allow the identification of CNSs that have been conserved over the 50-70 million years since these grasses diverged from a common ancestor (International Brachypodium Initiative 2010). Part of this analysis involved the design of web pages to visualise the alignments and identified CNSs. Gene expression data from rice was used to identify Brachypodium genes potentially involved in drought response and CNSs surrounding these genes were searched for regulatory elements known to be involved in drought response.

3.3 Methods

3.3.1 Identification of putative orthologues

FASTA files containing all predicted protein sequences from Brachypodium (v1.0), sorghum (v1.4) and rice (TIGR v5) were used as input into OrthoMCL v1.4 (Li *et al.* 2003) to determine putative rice and sorghum orthologues to each Brachypodium gene. OrthoMCL uses BLASTp to identify reciprocal best hits within a genome as potential paralogues and reciprocal best hits across two genomes as potential orthologues. Related proteins are linked in a graph structure which is decomposed into discrete clusters representing orthologous and paralogous groups of genes. The orthologous relationships were loaded into a MySQL database alongside the genomic coordinates of genes for later processing.

3.3.2 Identification of conserved sequences

Each Brachypodium gene with identified orthologues in rice and sorghum was processed in turn using Perl scripts (Supplementary information 10) to extract the genomic sequence containing the gene and surrounding region (gene-space) from FASTA files containing the three genome assemblies. The gene-space of each rice and sorghum orthologue was also extracted. The gene-space for a gene was defined to include 10 kb upstream of the first exon and 10 kb downstream of the last exon to capture conserved sequences that may occur far from the proximal promoter (Freeling *et al.* 2007). In cases where a neighbouring gene was found less than 20 kb away, the intervening distance was divided equally between the neighbouring genes.

Once the gene-space for orthologous genes were extracted, the exons of each gene were masked and bl2seq v2.2.18 (Tatusova and Madden 1999) was used to run pair-wise comparisons between the Brachypodium sequence and each of the rice and sorghum orthologues using settings designed to identify short conserved sequences as previously described in the literature (Kaplinsky *et al.* 2002). A spike sequence of 15 bp was added to the end of each sequence prior to the comparison and any high-scoring pairs (HSPs) giving E-values greater than the E-value obtained from the HSP between the spike sequences were discarded (Lyons and Freeling 2008).

The resulting HSPs were post-processed to identify regions on the Brachypodium sequence that were covered by both a Brachypodium - rice HSP and a Brachypodium - sorghum HSP, i.e. regions that are conserved between the three genomes. Only HSPs having a percentage identity of 85% or higher were included in this step and overlapping regions of less than 15 bp were excluded as this is the minimum length for consideration as a plant CNS (Lyons and

Freeling 2008). CNSs were classified as class1 where sequence and position in relation to the gene was conserved in orthologous gene-space and class2 where sequence only was conserved.

3.3.3 Analysis of CNS-rich genes using Gene Ontology (GO) terms

Previous studies have shown that genes with a high number of CNSs are associated with GO terms related to transcription factor activity and environmental, metabolic or pathogenic stress (Thomas *et al.* 2007; Li *et al.* 2009d). To test this I extracted GO terms for the 100 most CNS-rich genes and compared it to the GO terms for 100 random genes with no defined CNSs.

3.3.4 Visualisation of a large dataset

Dealing with such a large dataset required tools to visualise the HSPs and resulting CNSs with respect to genes. In order to do this, the HSPs between all pair-wise comparisons were loaded into the MySQL database together with the identified CNS. Web pages were written to display the orthologous relationships between Brachypodium, rice and sorghum genes and the HSPs obtained from BLAST for each pair-wise comparison. In addition, the Bio::Graphics library was used to implement web pages to visualise the HSPs resulting from each pair-wise comparison and the CNSs surrounding each Brachypodium gene (Supplementary information 11). Identified CNSs were loaded into the Brachypodium genome browser at www.modelcrop.org.

3.3.5 Identification of CNS regulatory motifs using rice gene expression data

In order to determine whether particular functional motifs were overrepresented in CNSs surrounding functionally related genes, a set of rice genes were obtained that were shown experimentally to be up-regulated in drought conditions (Zhou *et al.* 2007a). This data was provided as a list of oligonucleotides so first the corresponding rice gene for each oligonucleotide was identified by comparing each sequence against the TIGR rice v6 pseudo-molecules using BLASTn. Then BLASTp was used to find putative Brachypodium orthologues for each rice gene using an E-value cut-off of 1e-10.

The CNSs surrounding these genes were searched for the core motif ([GA]CCGAC) of the DRE/CRT (dehydration-responsive element/C-repeat) cis-acting element found in many genes in *Arabidopsis* and rice (Yamaguchi-Shinozaki and Shinozaki 1994). A Pearson chi-square test was used to test for the null hypothesis that no overrepresentation of the drought motif was observed in the CNSs surrounding the putative drought response genes, compared with the occurrence of drought motifs in the CNSs surrounding the genes not involved in drought response.

3.4 Results

3.4.1 Identification of putative orthologues

The 25,532 predicted protein coding loci in the Brachypodium genome (v1.0 annotation) encode 32,255 predicted proteins and these were clustered into 22,012 orthologous groups by OrthoMCL together with the 66,710 and 36,338 predicted proteins of rice and sorghum respectively.

Eight-hundred and seventy-one clusters (4 %) contained a single gene and the remaining clusters contained between 2 and 1,831 genes. Brachypodium genes were present in 16,892 clusters (78 %) and 21,480 Brachypodium genes (84 %) were clustered. In total, 14,635 clusters (66 %) contained genes from all three species and could be used to identify putative CNSs.

3.4.2 Identification of conserved sequences

After running the CNS detection pipeline, 15,997 putative conserved non-coding sequences were identified in the Brachypodium genome with lengths ranging from 15 to 2,255 nucleotides. Manual inspection of the longest CNSs revealed that six were clustered around a single gene on the distal end of the short arm of Brachypodium chromosome 5 (Figure 3.1). BLAST analysis showed these sequences had significant similarity to 25S ribosomal RNA genes which were not included in the Brachypodium v1.0 genome annotation and had been identified as highly conserved between the three species. These CNSs were excluded from further processing.

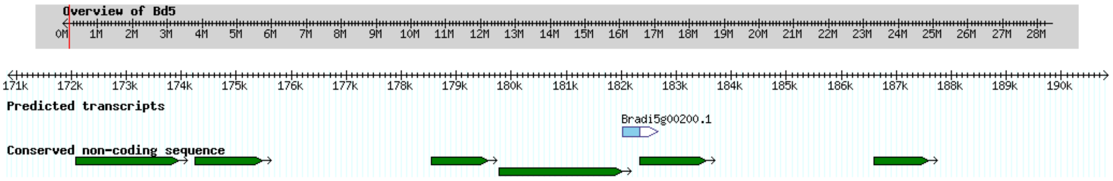


Figure 3.1: The six longest CNSs clustered around Brachypodium gene *Bradi5g00200*.
These CNSs were identified as 25S ribosomal RNA genes and excluded from further processing.

After these exclusions, 15,991 CNSs remained and were classified as 10,194 class 1 and 5,797 class 2 (Table 3.1). The complete list of CNSs is presented in Supplementary information 12. The lengths of identified CNSs ranged from 15 to 581 bp with a mean length of 30.1 bp. The median length was 23 bp meaning that half the CNSs were longer than 23 bp and half were shorter showing that the distribution was skewed towards shorter CNSs within this range. Only 30% of CNSs were longer than 40 bp. In total, 480.7 kb of genomic sequence was represented by CNSs.

	Number identified	Mean length (bp)	Median length (bp)	Length range (bp)
Class 1	10,194	32.0	24	15 - 476
Class 2	5,797	26.8	21	15 – 581
All	15,991	30.1	23	15 – 581

Table 3.1: Properties of putative CNSs identified in the Brachypodium genome

Low complexity sequence was not masked before identification of putative CNSs as many functional motifs contain low complexity sequence, for example the GAGA regulatory element which is a di-nucleotide repeat and has been shown to regulate gene expression in plants (Sangwan and O'Brian 2002). Because low complexity sequences were not masked, these sequences may be identified as potential CNSs if they occur in the gene-space of orthologous genes. However, they will be classified as class 2 CNSs as they do not satisfy the requirement of showing conservation of both sequence and location. All identified CNSs were searched for low complexity sequence using the DUST program (Altschul *et al.* 1990) to determine if class 2 CNSs (which are not conserved in location) were higher in low complexity sequence than class1 CNSs. CNSs containing 50 % or more low complexity sequence were identified (321 class1 CNSs and 1270 class2 CNSs).

Out of the 21,480 Brachypodium genes clustered, 7,486 (34.9 %) had identified CNSs, leaving 13,994 genes with no detectable CNSs using this method. More than half (52.4 %) of genes with CNSs had a single CNS with the maximum number of CNSs identified for one gene being 21 (Figure 3.2).

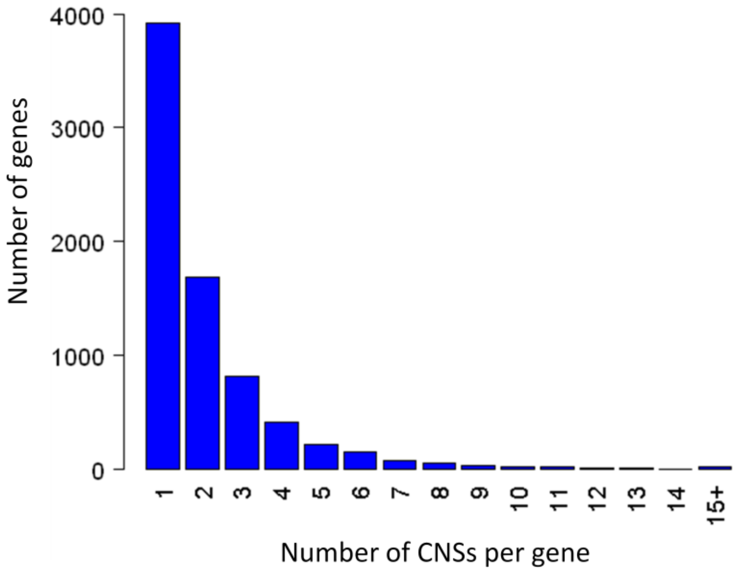


Figure 3.2: Distribution of the number of CNSs per gene considering only genes with CNSs.

Most genes included in have one identifiable CNS and the most CNSs identified for one gene is 21.

The location of CNSs in relation to their corresponding genes is important as, although they are usually found in the promoter region of genes, regulatory elements can also be found at the 3' end of genes and within introns (Thomas *et al.* 2007). Analysis of the locations of the CNSs identified in the Brachypodium genome (Table 3.2) showed that the majority of CNSs (45.2 %) occur in the 5' region of genes including the 5' untranslated region (UTR). The regions containing the second highest number of CNSs (29.7 %) were 3' regions and the remaining CNSs (25.1 %) were found in intronic regions. CNSs found in 5' and 3' UTR regions were counted separately in this analysis but as not all genes in the Brachypodium v1.0 annotation have defined UTR regions, these figures could be misleading if used on their own. For this reason, CNSs falling in defined 5' and 3' UTRs were counted as occurring within 5' upstream or 3' downstream regions.

	All CNSs		**Class 1 CNSs**	
Region	**Observed occurrences**	**Percentage of total**	**Observed occurrences**	**Percentage of total**
Intron	4,020	25.1	2,864	28.1
5' UTR	705	4.4	344	3.4
3' UTR	2,024	12.7	1,382	13.6
5' region	6,525	40.8	4,121	40.4
3' region	2,717	17.0	1,483	14.5
Total	15,991		10,194	

Table 3.2: Locations of identified CNSs considering all CNSs and class 1 CNSs only.

CNSs are found mainly in the 5' regions of genes, and also in 3' and intronic regions, but at a lower frequency.

Class 1 CNSs are conserved both in location and sequence and if the same analysis is performed on these CNSs only, a similar trend is observed. The majority of CNSs (42.8 %) occurred upstream of genes in the classic promoter region and occurred in 3' and intronic regions at a similar frequency (28.1 %).

Figure 3.3 shows how far away from genes 5' and 3' CNSs were found. The majority of both 5' and 3' CNS were close to genes, generally less than 2 kb. However some CNSs were as far as 10 kb away from the gene they were associated with. The mean distance of 5' and 3' CNSs from genes was 1.2 kb and if one considers the 'footprint' of a gene as extending 2 kb upstream from the first exon and 2 kb downstream from the last exon then 86% of detected CNSs fell within this region.

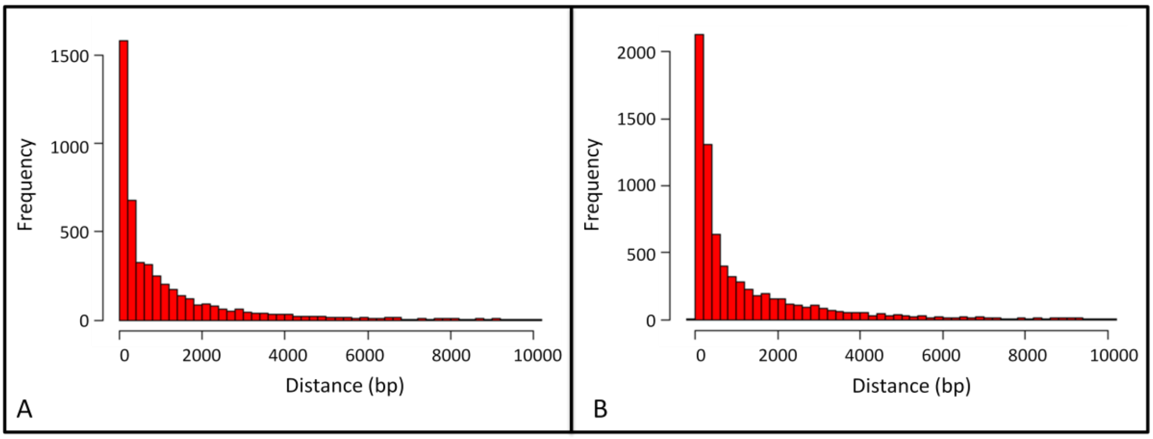


Figure 3.3: Distributions of the distance from genes where CNSs are found.

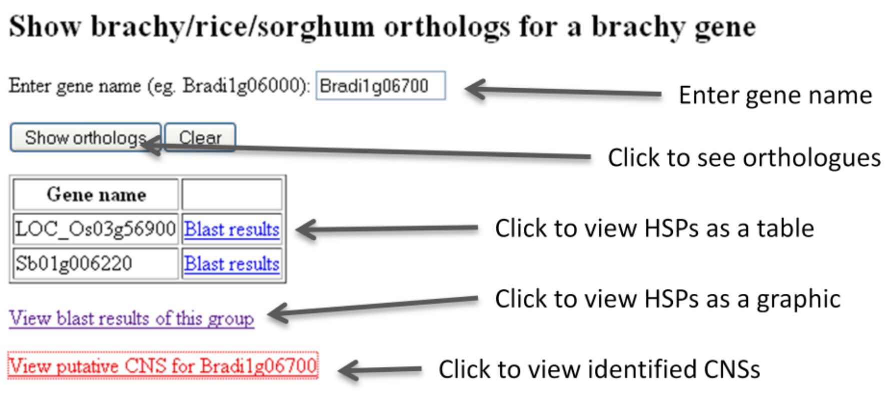
(A) CNSs in the 3' region (B) CNSs in the 5' region. In both cases the x-axis measures the distance from the gene. Most 5' and 3' CNSs are found close to genes (within 2kb) but some are found up to 10 kb away.

3.4.3 Analysis of CNS-rich genes using Gene Ontology (GO) terms

The 100 most CNS-rich genes had CNS numbers ranging from 21 to 6. The list of GO terms extracted from these genes contained 41 terms associated with transcription factor activity, either through transcriptional regulation or binding of a transcription factor. The list of GO terms extracted from 100 genes containing no identified CNSs contained only 3 terms associated with transcription factor activity.

3.4.4 Visualisation of a large dataset

The webpage I designed as the point of entry to the CNS database requires a Brachypodium gene name and displays all the orthologues identified for that gene (Figure 3.4). Clicking on one of the orthologue links in the resulting table shows the individual HSPs from each pairwise comparison between the Brachypodium gene and the orthologue, displayed as a table (Figure 3.5).



Show brachy/rice/sorghum orthologs for a brachy gene

Enter gene name (eg. Bradi1g06000): Bradi1g06700

Gene name	
LOC_Os03g56900	Blast results
Sb01g006220	Blast results

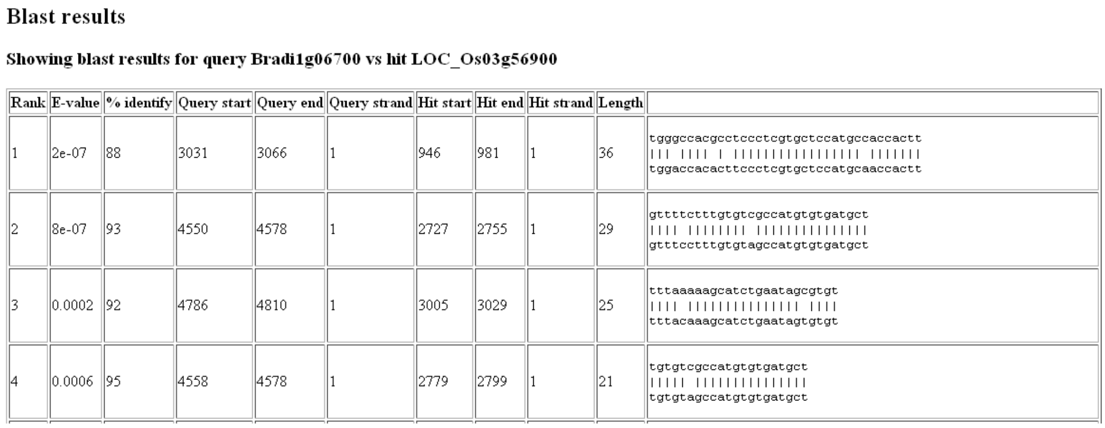
View blast results of this group

View putative CNS for Bradi1g06700

Enter gene name
Click to see orthologues
Click to view HSPs as a table
Click to view HSPs as a graphic
Click to view identified CNSs

Figure 3.4: Screenshot of the entry page for the CNS viewer.

From this page a user can view predicted orthologues, blast results and visualise BLAST results and CNSs with respect to genes.



Blast results

Showing blast results for query Bradi1g06700 vs hit LOC_Os03g56900

<table>
<tr><th>Rank</th><th>E-value</th><th>% identify</th><th>Query start</th><th>Query end</th><th>Query strand</th><th>Hit start</th><th>Hit end</th><th>Hit strand</th><th>Length</th><th></th></tr>
<tr><td>1</td><td>2e-07</td><td>88</td><td>3031</td><td>3066</td><td>1</td><td>946</td><td>981</td><td>1</td><td>36</td><td>tgggccacgcctccctcgtgctccatgccaccactt
||| |||| | ||||||||||||||||| |||||||
tggaccacacttccctcgtgctccatgcaaccactt</td></tr>
<tr><td>2</td><td>8e-07</td><td>93</td><td>4550</td><td>4578</td><td>1</td><td>2727</td><td>2755</td><td>1</td><td>29</td><td>gttttctttgtgtcgccatgtgtgatgct
|||| |||||||| |||||||||||||||
gtttcctttgtgtagccatgtgtgatgct</td></tr>
<tr><td>3</td><td>0.0002</td><td>92</td><td>4786</td><td>4810</td><td>1</td><td>3005</td><td>3029</td><td>1</td><td>25</td><td>tttaaaaagcatctgaatagcgtgt
|||| ||||||||||||||| ||||
tttacaaagcatctgaatagtgtgt</td></tr>
<tr><td>4</td><td>0.0006</td><td>95</td><td>4558</td><td>4578</td><td>1</td><td>2779</td><td>2799</td><td>1</td><td>21</td><td>tgtgtcgccatgtgtgatgct
||||| |||||||||||||||
tgtgtagccatgtgtgatgct</td></tr>
</table>

Figure 3.5: Screenshot showing individual HSPs from a pairwise comparison using BLAST.

In addition to the tabular view, HSPs can be viewed graphically positioned around a gene. Figure 3.6 shows a screenshot of the HSPs resulting from two pair-wise comparisons (Brachypodium – rice and Brachypodium – sorghum). From this image it is easy to see that the HSPs labelled ‘hsp1’ (linked with green arrows) in both comparisons are located upstream of the 5’ end of each gene model and identify a CNS. The resulting CNS can be viewed graphically in the CNS viewer or the Brachypodium genome browser.

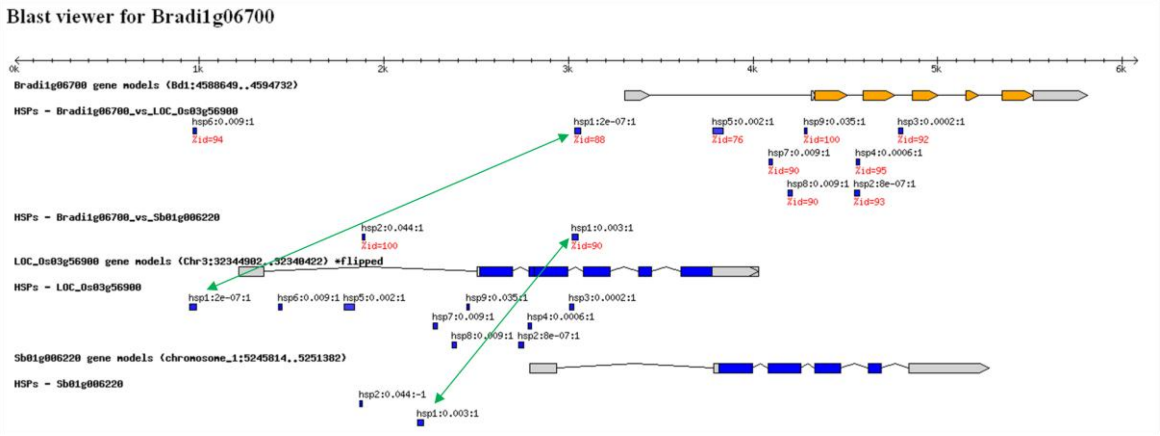


Figure 3.6: Screenshot showing a graphical view of HSPs.

Two pairwise comparisons of orthologous genes are shown with HSPs shown as small blue blocks labelled 'hsp'. The Brachypodium gene model is shown at the top in orange and the rice and sorghum gene models in blue. Directly below the Brachypodium gene model are the HSPs resulting from the comparison with the rice gene and below this, the HSPs resulting from the comparison with the sorghum gene. Each hit in an HSP is labelled the same as its matching pair. For example, hsp1 in the Brachypodium – rice comparison matches hsp1 under rice gene model and hsp1 in the Brachypodium – sorghum comparison matches hsp1 under the sorghum model. These matches are indicated by green arrows in the figure.

3.4.5 Identification of CNS regulatory motifs using rice gene expression data

The 1,629 *Brachypodium* genes orthologous to rice genes shown to be involved in drought response had 1,567 CNSs associated with them and 23 of these contained DRE/CRT motifs. The 19,851 genes not predicted to be involved in drought response had 14,424 CNSs surrounding them and 99 of these contained the DRE/CRT motif. A Pearson chi-square test showed that there was a statistically significant difference (χ^2 (1, N=15,991) = 11.4, p < 0.001) in enrichment of the motif within CNSs surrounding genes predicted to be involved in drought response compared to CNSs surrounding genes not predicted to be involved in drought response. There is a less than 0.1 % probability that this difference would occur by chance.

An example of a CNS containing a DRE/CRT motif is shown in Figure 3.7. The CNS was located in the 5' promoter region of a gene and was highly conserved between orthologues of Brachypodium, rice and sorghum differing in only 4 positions along its 27 bp length. The gene (Bradi1g55560) was identified as being involved in drought response by orthology to a rice gene shown to be involved in drought response. The protein sequence of Bradi1g55560 showed sequence similarity to a light inducible protein from barley belonging to the ELIP/psbS family. Genes from this family have been shown to increase gene expression in response to a variety of stresses, including drought (Zeng *et al.* 2002).

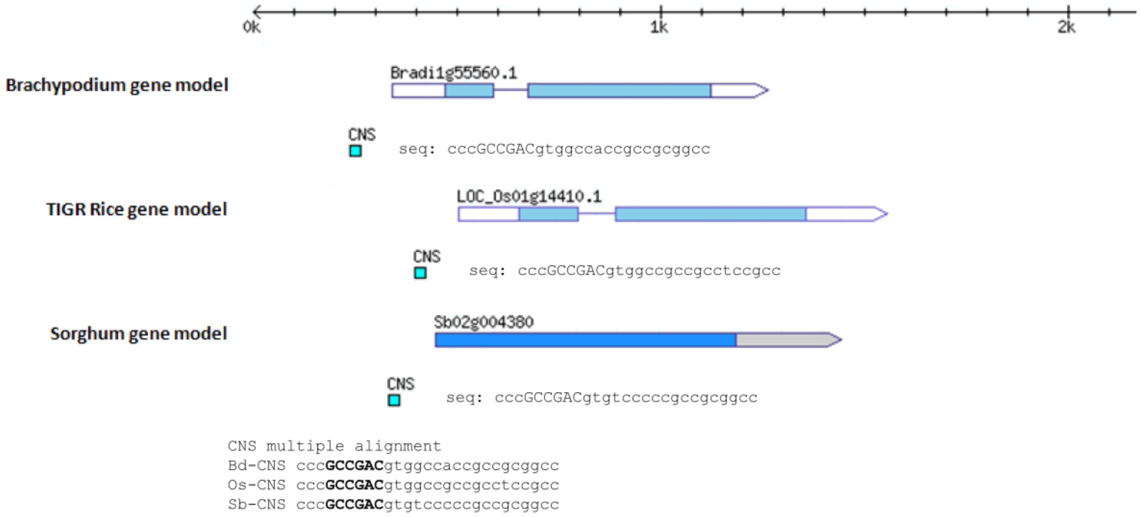

0k
1k
2k
Brachypodium gene model
Bradi1g55560.1
CNS
seq: cccGCCGACgtggccaccgccgcggcc
TIGR Rice gene model
LOC_Os01g14410.1
CNS
seq: cccGCCGACgtggccgccgcctccgcc
Sorghum gene model
Sb02g004380
CNS
seq: cccGCCGACgtgtcccccgccgcggcc
CNS multiple alignment
Bd-CNS ccc**GCCGAC**gtggccaccgccgcggcc
Os-CNS ccc**GCCGAC**gtggccgccgcctccgcc
Sb-CNS ccc**GCCGAC**gtgtcccccgccgcggcc

Figure 3.7: A conserved non-coding sequence in Brachypodium, rice and sorghum.

The CNS is located in the 5' upstream region of orthologous genes shown to be involved in drought-response. The multiple sequence alignment of the three CNSs shows the core DRE/CRT (dehydration-responsive element/C-repeat) cis-acting element in bold.

3.5 Discussion

Phylogenetic footprinting is a method used to identify regions in genomes containing regulatory elements by comparing homologous sequences from related species. These comparisons highlight regions that have been conserved over evolutionary timescales due to selective pressure and that may have functional significance, for example as transcription factor binding sites. Non-functional sequence is not evolutionarily constrained and will accumulate mutations. Under this premise I compared orthologous sequences from Brachypodium, rice and sorghum to identify highly conserved regions in Brachypodium that could be candidate regulatory regions. The genomic regions identified using this method are conserved across three diverse grass subfamilies and therefore will probably be conserved in other grasses such as wheat and barley for which a complete genome sequence has yet to be obtained. The method used to identify conserved sequences has previously been used to identify conserved sequences in plants (Kaplinsky *et al.* 2002; Inada *et al.* 2003; Thomas *et al.* 2007; Lyons and Freeling 2008; Li *et al.* 2009d) which appear to have less complex gene regulatory mechanisms than mammalian systems.

Clustering protein sequences from Brachypodium, rice and sorghum identified 22,012 orthologous groups; 21,480 Brachypodium genes were clustered and 14,635 clusters contained genes from all three species that could be used from CNS detection. Similar relationships were identified in the Brachypodium genome analysis where 20,562 Brachypodium/wheat/barley genes were clustered with sorghum and rice genes into 13,580 gene families (International

Brachypodium Initiative 2010). The reason for the difference may be my choice of rice genome annotation; I used the TIGR v5 annotation whereas the RAP2 annotation was used in the clustering performed for the Brachypodium genome analysis and the number of genes predicted by these methods differs significantly (RAP2 - 28,236, TIGRv5 – 56,328). The clustering performed for the Brachypodium genome analysis also collapsed paralogous gene sets before running OrthoMCL, something that was not performed in this analysis.

Previous analyses of regulatory regions in plants identified CNSs clustered around protein-coding genes, either upstream, downstream or within introns (Freeling *et al.* 2007; Thomas *et al.* 2007). In addition, CNSs were identified up to 4 kb away from an exon. For these reasons I chose to define a gene-space for each gene of up to 10 kb, limited by neighbouring genes.

The orthologous comparisons performed in this analysis identified 15,991 CNSs with lengths ranging from 15 to 581 bp with a median length of 23 bp. Two previous efforts to identify CNSs in *Arabidopsis* and rice using paralogous comparisons identified an average CNS as being of similar size (Table 3.3). According to the figures in Table 3.3, the CNSs identified in Brachypodium are most similar to *Arabidopsis* CNSs, which is surprising if one considers the data obtained from previous studies to be generally representative of CNS differences between monocots and dicots. Of the total number of CNSs identified, 63.7 % are classified as ‘true CNSs’ where both sequence and location is conserved. Very little difference is observed in the length distributions of class1 and class2 Brachypodium CNSs.

Species	**Number identified**	**Mean length (bp)**	**Median length (bp)**	**Length range (bp)**
Brachypodium	15,991	30.1	23	15-581
Arabidopsis	14,944	30.7	24	15-285
Rice	31,114	29	17	15-1,684

Table 3.3: Comparison of CNSs identified in Brachypodium, *Arabidopsis* and rice.

(Data obtained from Thomas *et al.* 2007; Li *et al.* 2009d)

When identifying CNSs from paralogous comparisons (Thomas *et al.* 2007; Li *et al.* 2009d), one is limited by the number of genes contained within intra-genomic duplicated regions so the subset of genes used to identify these CNSs is usually small (6,042 in rice and 3,179 in *Arabidopsis*). In this analysis, nearly 15,000 clusters of orthologous genes were used so in theory more CNSs should be identified. However, in this analysis the CNSs being identified are conserved between three species, hence only a small number of CNSs are identified. CNSs conserved between Brachypodium and rice but not sorghum or conserved between

Brachypodium and sorghum but not rice were not included in this dataset but may still be relevant as potential regulatory regions conserved be more closely related grasses.

The identification of six ribosomal RNA genes as potential CNSs (Figure 3.1) highlights the fact that this method of CNS detection is very dependent upon an accurate genome annotation. The regions selected for comparison are determined by the annotation and coding regions defined by the annotation are masked. This means that any annotation errors will be reflected in the CNS dataset, including exons that have not been annotated and incorrect gene models. The CNSs identification performed using version 5 of the *Arabidopsis* genome annotation highlighted many annotation issues (Thomas *et al.* 2007). The Brachypodium CNS dataset was not explicitly screened to identify annotation errors in this analysis due to time restrictions.

Very few identified CNSs contained low complexity sequence or simple sequence repeats although class 2 CNSs contained higher levels than class 1 CNSs. Low complexity sequence in class 1 CNSs is more likely to be functional than similar sequence in class 2 CNSs as class 1 CNSs are conserved in syntenic positions across the three grass genomes. Low complexity sequence in class 2 CNSs probably reflects non-functional repetitive sequence in orthologous gene-space, as location is not conserved in class 2 CNSs. Very few CNSs identified in *Arabidopsis* were simple sequence repeats (Thomas *et al.* 2007).

Most Brachypodium genes have no detectable CNSs using BLASTN comparison, in agreement with previous studies (Thomas *et al.* 2007; Li *et al.* 2009d). This is because the detection of potential regulatory regions using phylogenetic footprinting only identifies regions under selective pressure; it does not identify regulatory regions that are not evolutionarily conserved. Genes without CNSs may be regulated by other less conserved DNA elements. One of the findings from the pilot phase of the ENCODE project was that although there is overlap between genomic regions identified as functional by experimental means and regions under evolutionary constraint, not all experimentally determined functional regions are under evolutionary constraint (Birney *et al.* 2007).

The CNS-rich genes identified in this study appear to be more associated with transcription factor activity compared to genes with no CNSs. Interestingly, no CNS-rich genes were annotated with GO terms associated with environmental, metabolic or pathogenic stress, which have been previously demonstrated in *Arabidopsis* and rice (Thomas *et al.* 2007; Li *et al.* 2009d). This could again be due to the annotation of the Brachypodium genome being at a preliminary stage. A more thorough overrepresentation analysis of GO terms is required before any further conclusions can be drawn.

In this analysis, CNSs were found in 5’ and 3’ regions of genes as well as introns. The majority of CNSs are found upstream of genes, within the ‘classical’ promoter region. Although most CNSs are found close to genes, some are found much further away indicating that the potential regulatory footprint of a gene is much larger than simply the proximal promoter region. In *Arabidopsis*, genes with CNSs identified more than 4 kb from an exon have been termed bigfoot genes (Thomas *et al.* 2007).

Visualisation of such a large dataset is difficult and automated processing is no substitute for manual inspection in the case of CNSs. The visualisation methods designed in this analysis were essential used to manually inspect HSPs and their locations with respect to genes. The Bio::Graphics library is used to display genes and associated features but is of limited use as the library is not designed to display relationships between features, something that is required for the identification of syntenic HSPs as potential CNSs. An excellent visualisation tool for manual CNS identification is GEvo (Lyons and Freeling 2008), part of the CoGe comparative genomics suite. GEvo allows pairwise comparison of orthologous genes using a variety of alignment tools and matching HSPs can be linked to determine whether they are syntenic in orthologous genes. The limitation of GEvo is that batch processing of genes is not possible.

It is hypothesised that CNSs contain functional motifs which regulate gene expression by binding transcription factors, so CNSs should contain functional elements. To identify functional elements in CNSs one can either look for overrepresented motifs in all identified CNSs or identify a subset of CNSs surrounding co-expressed genes and search these for specific motifs. In this analysis a set of orthologous Brachypodium genes was identified from rice genes that have been experimentally identified as being involved in drought response. The DRE/CRT drought response motif was found to be significantly overrepresented in these CNSs compared to CNSs surrounding genes not predicted to be involved in drought response. This indicates that the CNSs identified in this analysis are likely to have functional significance and could be used as a basis for future experimental work to identify gene regulatory mechanisms in Brachypodium. Further analysis is required to identify motifs in CNSs surrounding different groups of co-regulated genes but as more expression data becomes available in Brachypodium, groups of co-regulated genes can be directly identified. An Affymetrix microarray for transcriptome analysis has been developed as part of the genome project and this should be useful for determining gene expression patterns on a genome scale. These studies will provide more accurate co-expression data than inferring co-expression from studies in related grasses.

In conclusion, the analysis presented in this chapter describes the first attempt to identify putative regulatory regions in the Brachypodium genome by phylogenetic footprinting on a genome-wide scale. I have identified a large dataset of potential regulatory regions that can be targeted in future experiments. In addition, the fact that the identified CNSs are conserved between three diverse members of the grass family indicates that these regions should also be conserved in other grasses such as wheat and barley and may used to identify regulatory mechanisms in these species.

4 Analysis of T-DNA insertions in the Brachypodium genome

4.1 Relevant publications

Vera Thole, Barbara Worland, Jonathan Wright, Michael W. Bevan and Philippe Vain (2010). "Distribution and characterization of more than 1000 T-DNA tags in the genome of *Brachypodium distachyon* community standard line Bd21." Plant Biotechnol J. **8**(6): 734-47.

4.2 Introduction

After sequencing and annotating the genome of an organism, the focus of research is directed towards the high-throughput analysis of gene function. This field is called functional genomics. Comparison of newly identified gene sequences with genes of known function *in silico* is one method used to infer gene function but experimental evidence is also required. Even in a well-studied model organism such as *Arabidopsis*, the majority of genes have not been subject to experimental validation and many are annotated as unknown or hypothetical function. One of the most powerful ways to determine the function of a gene is to mutate it and study the resulting phenotype. The traditional approach is called 'forward genetics' and involves the random mutagenesis of a number of individuals followed by screening to identify mutants with a particular phenotype of interest. The gene responsible for the phenotype is then identified by mapping and cloning. The opposite is a 'reverse genetics' approach, first proposed in 1989 (Ballinger and Benzer 1989) where a gene of unknown function is mutated and its resulting phenotype is investigated. For reverse genetics studies, the ideal resource is a population of mutagenised individuals, each having a mutation in a single gene that disrupts the function of that gene. The population should be large enough to ensure that every gene in the organism is disrupted. There are two ways to generate such a population. The first is to systematically mutate each gene in the genome by targeted mutagenesis and the second is to randomly mutate genes in the genome. Targeted mutagenesis is an expensive and time-consuming process but does ensure that each gene in the genome is mutated. Random mutagenesis is fast and relatively inexpensive but due to the random nature of the mutations, one cannot ensure that every gene in the genome has been mutated.

Targeted mutagenesis was pioneered in mouse (Capecchi 1989) and requires a vector containing a disrupted gene to be incorporated into a culture of embryonic stem cells by homologous recombination. Embryos at an early stage of development are injected with these stem cells and if successful, some of the resulting offspring will be heterozygous for the mutated gene. The procedure is complex and dependent on the efficiency of homologous recombination which is highly variable between different species. In bacteria and yeast, homologous recombination is highly efficient and targeted mutagenesis has been used to produce a nearly complete set of gene-deletion mutants in *S. cerevisiae* (Giaever *et al.* 2002). It can also be used in microbial genomes as these contain relatively few genes. The efficiency of targeted mutagenesis in higher eukaryotes is lower, however gene targeting strategies have been developed for some vertebrates including *D. melanogaster* (Maggert *et al.* 2008). Targeted mutagenesis strategies have also been developed in plants (Terada *et al.* 2002; Zhang *et al.* 2010)

Random mutagenesis can be achieved by exposure to radiation, chemical treatment, or the insertion of a short piece of DNA into the genome. Radiation was the first mutagen to be used for forward genetic studies as exposure to radiation in the form of X-rays, fast neutrons, or gamma rays induces genomic deletions. The deletions can be small, resulting in the loss of parts of single genes or can be much larger, encompassing many genes. More recently fast neutron mutagenesis has been used in reverse genetic studies in *Arabidopsis* and rice to generate large mutagenised populations (Li *et al.* 2002). Mutant lines with deletions in a target gene are identified by screening DNA pools generated from all the mutant lines using PCR primers flanking the target gene. Alkylating agents such as ethyl methanesulfonate (EMS) or N-nitroso-N-methylurea (NMU) are chemical mutagens that typically induce single nucleotide polymorphisms (SNPs) in DNA. Chemical mutagenesis by EMS is the basis for TILLING or Targeting Induced Local Lesions IN Genomes (Colbert *et al.* 2001). Treatment with EMS creates point mutations in a population of individuals which are detected using a plant endonuclease (CEL-I) that cleaves at the mismatched sites in heteroduplexes formed between mutant DNA a wild type. TILLING is a high-throughput reverse-genetics technique originally developed for *Arabidopsis* which has since been applied to other plants and other organisms (Till *et al.* 2004; Gilchrist *et al.* 2006; Moens *et al.* 2008).

The most popular strategy to generate a mutagenised population is by insertional mutagenesis where a piece of DNA is inserted randomly into the genome of an organism. The DNA can be an endogenous or heterologous transposon or a foreign DNA sequence. If the DNA inserts into a gene it can cause loss of function of that gene, called a null allele. The advantage of insertional mutagenesis over chemical- or radiation-based approaches is that as well as causing a mutation, the junction between the inserted DNA and the genome sequence can be used to identify and tag the mutation. This junction region is called a Flanking Sequence Tag or FST. Mutagenised populations of *S. cerevisiae* have been generated using the retrotransposon *Ty* as an insertional mutagen (Smith *et al.* 1996). Insertions are located by PCR using a primer that anneals to the inserted element as well as a gene-specific primer.

In plants, transposons such as *Activator* (*Ac*), *Suppressor-mutator* (*Spm*) and *Mutator* (*Mu*) have been used to generate large mutagenised populations. In general, transposons generate single-copy insertions but tend to insert preferentially into certain regions of the genome meaning that it is difficult to achieve whole-genome saturation. In addition, transposons have the intrinsic ability to move by virtue of encoding their own transposase, therefore additional steps are usually required to stabilise the inserted transposons by removing the transposase. The *Activator* transposon is autonomous as it encodes a transposase gene but shorter derivatives called *Dissociation* (*Ds*) do not encode this gene, therefore can only move if *Ac* is present in the genome. If *Ac* is removed by crossing, progeny with stable *Ds* insertions can be recovered. The *Ac*/*Ds* system exhibits a phenomenon called local transposition, meaning it jumps preferentially to linked sites. This is an advantage if mutations are required in a small genomic location (Ito *et al.* 1999) but it can be difficult to achieve genome-wide saturation using this system. However, this problem can be overcome by selecting against closely linked transpositions (Parinov *et al.* 1999). The *Spm* system from maize is an efficient transposon tagging system and has been used to generate mutant *Arabidopsis* lines (Speulman et al. 1999). These elements transpose to unlinked sites without showing a bias to specific genomic locations. The *Mu* transposon exhibits preferential insertion into genes (Cresse *et al.* 1995; Raizada *et al.* 2001) making it a good tool for mutagenesis studies in maize.

An alternative method of insertional mutagenesis applicable to plants is T-DNA mutagenesis. This method exploits the ability of the soil bacterium *Agrobacterium tumefaciens* to integrate a section of DNA (called the T-DNA) from a freely replicating plasmid within the bacterium to the plant genome. In its native form, the bacterium is responsible for the formation of tumours in plants or crown gall (Gelvin 2003) but when used for mutagenesis, the T-DNA is replaced with new sequence. In the plasmid, the T-DNA is flanked by two 25 bp border repeats that allow it to integrate into the plant genome. The advantage of this method over transposon-meditated mutagenesis is that T-DNA has no ability to transpose once inserted thus the insertions generated are more stable. However, the integration of T-DNA into the plant genome can generate complex insertion patterns and make subsequent analysis difficult. For example, in some cases the border repeats are not recognised correctly resulting in the integration of short sequences of plasmid backbone into the genome (Martineau *et al.* 1994). In addition, the T-DNA can integrate at more than one location in the genome (Afolabi *et al.* 2004). A detailed study of T-DNA insertions in rice showed that only 20 % of plants contained a single T-DNA copy at one locus, a further 30 % contained a single copy at multiple loci and 50 % contained multiple T-DNA copied at multiple loci (Vain *et al.* 2003).

Using T-DNA, insertion mutants have been produced for *Arabidopsis* and rice (Alonso *et al.* 2003; Krishnan *et al.* 2009) and plant lines are available to the scientific community for use in biological investigations. In *Arabidopsis*, over 360,000 insertions have been mapped in the genome covering more than 90 % of the genes (Alonso and Ecker 2006).

Brachypodium is an emerging experimental system for the temperate grasses (Garvin *et al.* 2008). As discussed in Chapter 2, Brachypodium is closely related to important crop species, forage grasses of economic importance and has physical properties similar to grasses such as *Miscanthus*, a potential biofuel crop (Gomez *et al.* 2008). A key step towards establishing Brachypodium as an experimental grass system is the generation of a large mutant collection which will provide a resource for gene discovery and analysis of gene function in Brachypodium. In addition to the identification of genes controlling traits specific to cereal crops, these studies have the potential to elucidate the genetic mechanisms controlling traits relevant to the development of grasses as a biofuel feedstock (DOE 2006). For example, identifying genes that control the development and structure of the plant cell wall will facilitate the development of efficient methods to extract cellulose from plant cells.

An efficient transformation system is a key foundation for functional genomics studies. The first report of *Agrobacterium*-mediated transformation of diploid Brachypodium lines achieved an efficiency of 2.5 %, too low for high-throughput applications such as T-DNA tagging (Vogel *et al.* 2006a). However, further research has increased the efficiency of transformation in the community standard line Bd21, to more than 40 % (Vain *et al.* 2008; Alves *et al.* 2009). This efficiency is similar to that of rice meaning that T-DNA tagging is possible. Philippe Vain's (PV) group at the JIC are the first to produce and characterise Bd21 T-DNA insertion lines and this chapter describes the bioinformatic analysis of these insertions in the Brachypodium genome that I have performed as part of this work.

4.3 Methods

A collection of 4,117 fertile plant lines containing T-DNA insertions has been developed using an *Agrobacterium*-mediated transformation system (Vain *et al.* 2008). A subset of these lines (741) was analysed at the molecular level to retrieve the regions flanking the right and left borders of the T-DNA insertion. These regions were sequenced and compared to vector sequence (pVec8-GFP) to remove vector contamination. A total of 1,005 FSTs were obtained and compared to the Brachypodium genomic sequence to locate the position of insertion on the Brachypodium genome. This manual process of locating T-DNA insertions was performed by PV with bioinformatics assistance from myself.

To determine locations of insertions with respect to genes, I wrote a Perl script utilising the Bio::DB::GFF BioPerl module to query the relational database storing the Brachypodium genome assembly and version 1.0 of the annotation (Supplementary information 13). The script searches both strands of the genomic sequence at each T-DNA insertion point to determine whether the insertion falls within a defined coding sequence (CDS), then whether it is intronic or exonic. If the insertion falls outside of a coding region, upstream (5') or downstream (3') regions flanking the CDS are calculated using 1500 bp for the upstream region and 500 bp for the downstream region (Figure 4.1, A). If two genes exist in close proximity, flanking regions are calculated by dividing the distance between the genes proportionally (Figure 4.1, B). This ensures that 3' and 5' regions do not overlap when calculating the position of insertion. However, overlap may occur between genes in close proximity but on opposite strands. In these cases a T-DNA insertion can occur within the upstream (or downstream) region of both genes (Figure 4.1, C). A T-DNA insertion is classified by the region in which it inserts as exonic, intronic, 5' upstream, 3' downstream or intergenic. The script was also run with the size of the 5' upstream region set at 500 bp to determine whether insertions into this region occurred at higher frequency close to the start codon.

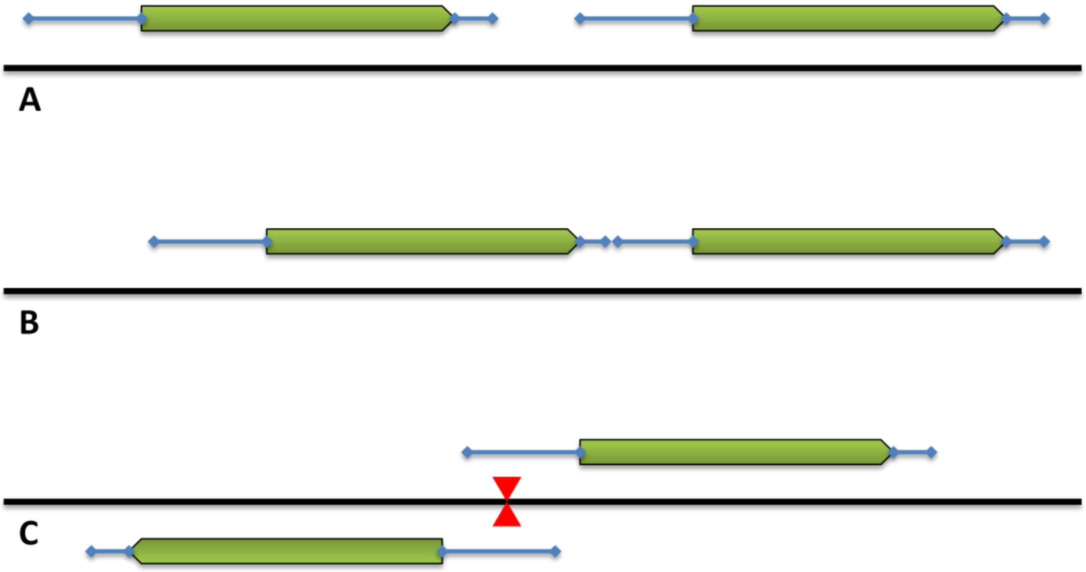


Figure 4.1: Defining the position of a T-DNA insertion.

Genes are shown in green on the genomic sequence (black line), the strand is indicated by the direction of the arrow. Upstream and downstream regions are indicated by blue lines. (A) The normal case where upstream and downstream regions for two genes are defined as 1500 bp and 500 bp respectively. (B) When the proximity of two genes on the same strand would give overlapping upstream/downstream regions the distance between the genes is divided proportionally. (C) Showing a T-DNA insertion (red triangles) potentially affecting genes on both DNA strands. The upstream regions of both strands are overlapping.

The distribution of insertions within the exons of genes was also analysed. The coding region of a gene is defined as the distance from the start codon in the first exon to the stop codon in the last exon and the position of the insertion between these points is calculated as a percentage along the length of the coding region.

Genes containing a T-DNA insertion were analysed to identify if they were supported by evidence of expression in Brachypodium, wheat and barley. Evidence of expression was defined as showing homology to an expressed sequence tag. The predicted function of disrupted genes was extracted from the v1.0 annotation.

All FST insertion coordinates were converted to GFF, loaded into the Brachypodium genome browser at www.modelcrop.org and displayed as an annotation track. Each FST is linked to a page showing the FST sequence and the plant line from which it originates.

4.4 Results

A total of 991 T-DNA FSTs were located on Bd21 genomic sequence representing 660 individual T-DNA insertions. Two of these insertions (with 4 corresponding FSTs) were located on unanchored scaffolds. Fourteen FSTs could not be anchored to a unique location on the chromosomes. The distributions of T-DNA insertions along each Brachypodium chromosome using a window size of 2 Mb are shown in Figure 4.2. The distributions along each chromosome are non-uniform and in general shown a low frequency of insertion at each centromere with more insertions found towards the end of each chromosome arm. In most cases, this insertion pattern correlates with the gene density along each chromosome as shown by the heat maps aligned to each chromosome where red indicates a high gene density and blue a low gene density (Figure 4.2). A particularly low frequency of insertion is observed on the short arm of chromosome 5. This is a region containing very few genes and a high number of retroelements (International Brachypodium Initiative 2010).

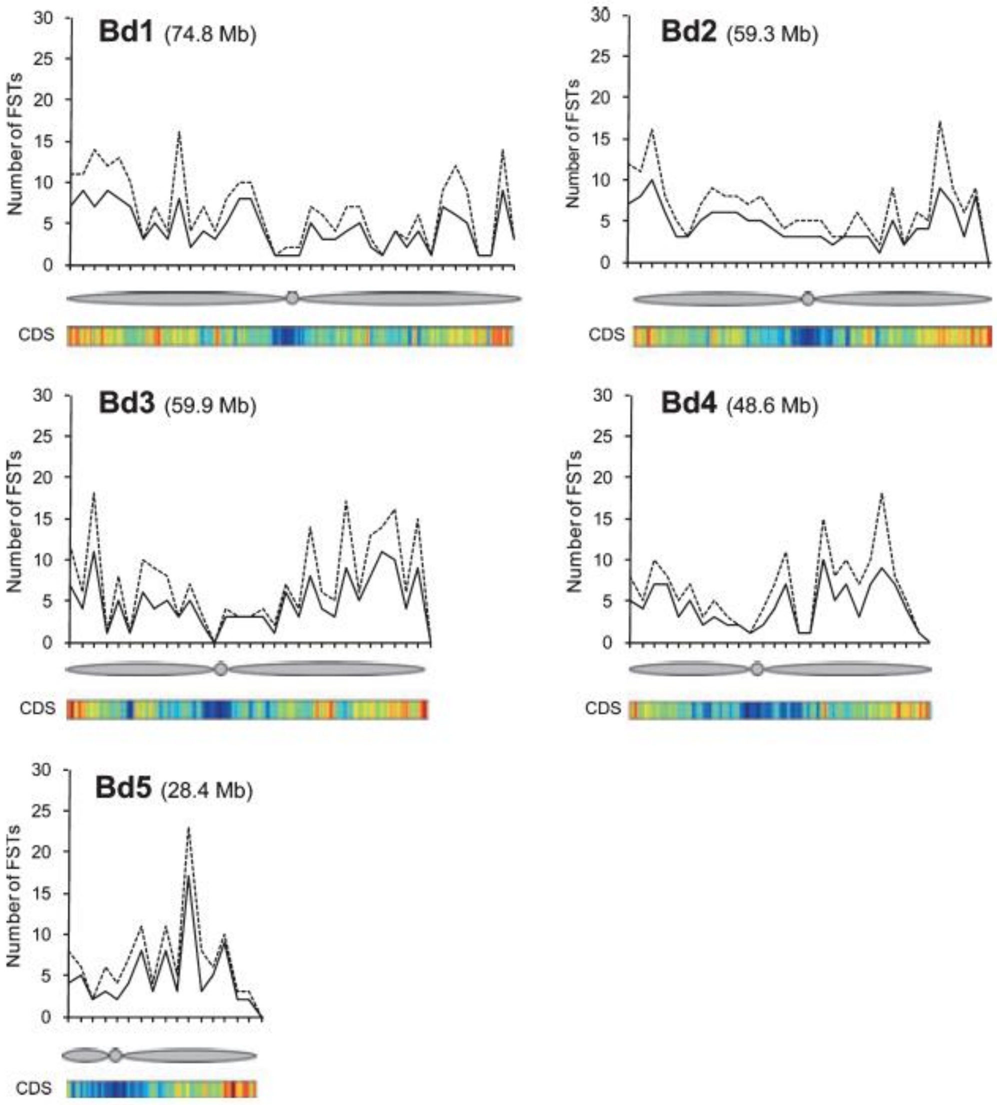


Figure 4.2: Chromosomal distribution of T-DNA insertions in the genome of Brachypodium.
The number of insertions is represented at 2 Mb intervals. Solid lines indicate unique T-DNA insertion loci and dotted lines represent flanking sequence tags (FSTs). The heat maps represent the estimated density of Brachypodium coding sequence (CDS) in the annotated v1.0 genome sequence (International Brachypodium Initiative, 2010). CDS density ranges from blue (low) to red (high). Mb: megabase. (Figure from Thole *et al.* 2010)

The locations of the 660 T-DNA insertions with respect to genes are presented in Table 4.1. Insertions are classified as either intergenic or genic, with genic insertions further divided into exon, intron, 5’ upstream (first 500 bp), 5’ upstream (500 to 1500 bp), and 3’ downstream regions (500 bp). For the purposes of this analysis a genic region consists of a 1500 bp 5’ region upstream of the ATG start codon, the exons and introns of the gene and a 500 bp 3’ region downstream of the stop codon. The expected insertion frequency for each class of insertion is calculated assuming random integration of T-DNA across the genome based on published Brachypodium genome statistics (International Brachypodium Initiative 2010). For example, the

Brachypodium genome contains 25,532 protein coding genes with a mean size (not including UTRs) of 2,956 bp. Therefore, an average gene is 4,956 bp in length including 1500 bp for the 5' flanking region and 500 bp for the 3' flanking region. The total length of genomic sequence estimated to be genic is 126.5 Mb, or 46.7% of the total genome size (271.1 Mb).

Classification	Number of T-DNA insertions	Observed insertion frequency	Expected insertion frequency
Total intergenic	311	47.1%	53.3%
Total genic	349	52.9%	46.7%
Exon	66	10.0%	14.0%
Intron	93	14.1%	13.9%
5' upstream (-500 bp)	83	12.6%	4.7%
5' upstream (-500 to -1500 bp)	66	10.0%	9.4%
3' downstream (+500 bp)	41	6.2%	4.7%
Total	660		

Table 4.1: Distribution of T-DNA insertions with respect to predicted genes in the Brachypodium genome. The expected insertion frequency assuming random integration of T-DNA across the genome is also shown for comparison.

These results show that T-DNA insertion is higher in genic regions (52.9 %) than would be expected by random insertion (46.7 %). This is mainly due to much higher than expected levels of insertion into the 500 bp region immediately upstream of the ATG start codon of genes (12.6 % compared to 4.7 %) but also due to slightly higher than expected insertion into 3' downstream regions (6.2 % compared to 4.7 %). T-DNA insertions into exons are lower than expected (10.0 % compared to 14.0 %) and insertion into introns and 500 bp to 1500 bp 5' upstream regions are close to expected values.

T-DNA insertions within exons are not uniformly distributed throughout the gene but a strong preference for insertion close to the 5' and 3' end of the gene, more so in 3' exons. Very few exonic insertions occur in the central region of a gene (Figure 4.3).

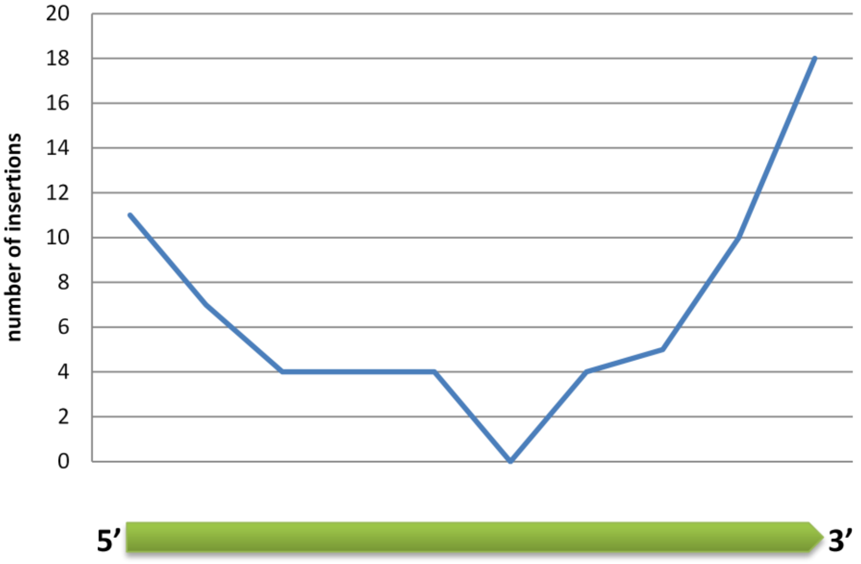


Figure 4.3: Distribution of exonic T-DNA insertions within genes.

The frequency of insertion is highest at the 3' end of genes and slightly less at the 5' end of genes. Very few T-DNAs appear to insert in the middle of genes in Brachypodium.

Of the 349 insertions occurring within predicted genes, 15 disrupt genes on both strands of the genomic sequence therefore 364 genes are potentially disrupted. Genic insertions occur at similar frequencies on each chromosome taking into account the chromosome size; between 1.2 and 1.5 disrupted genes per Megabase (Table 4.2). One T-DNA inserts into a predicted gene on an unassembled scaffold. Evidence of expression was observed for more than 60% of tagged genes by homology to Brachypodium and wheat/barley ESTs.

Chromosome number	**Predicted genes with T-DNA insertions**	**Expressed genes (Brachypodium)**	**Expressed genes (wheat/barley)**
1	96	65	56
2	86	55	57
3	82	50	51
4	60	35	33
5	39	21	27
Scaffolds	1	1	1
Total	364	227	225

Table 4.2: Genes interrupted by a T-DNA insertion on each Brachypodium chromosome.

If a gene showed homology to an EST it was considered to be expressed.

Predicted functions could be found for 175 of the 364 genes containing a T-DNA insertion and these are listed in Supplementary information 14. Many of these genes have been ascribed function using the Blast2Go software (Conesa *et al.* 2005) and some are from manual annotation.

4.5 Discussion

Generating a large population of systematically mutagenised individuals is a key step towards understanding the function of genes within the genome of an organism. Insertional mutagenesis using T-DNA is a high-throughput approach to generate this type of population and combined with a protocol to identify the potentially mutagenised gene in each individual line provides researchers with a functional genomics resource for the analysis of gene function. Brachypodium is an emerging experimental system for the temperate grasses (Garvin *et al.* 2008). It is closely related to important crop species, forage grasses of economic importance and has physical properties similar to grasses such as *Miscanthus*, a potential biofuel crop (Gomez *et al.* 2008). Functional genomics studies in Brachypodium will enable the identification of genes controlling traits specific to cereal crops as well as traits relevant to the development of grasses as a biofuel feedstock (DOE 2006).

A population of Brachypodium Bd21 plant lines containing T-DNA insertions has been generated and a subset of these were analysed to locate the position of the T-DNA insertion using FSTs. In total, 1,005 FSTs were obtained corresponding to 660 T-DNA insertions and of these 991 could be placed on the Brachypodium pseudo-molecules or unanchored scaffolds. The average frequency of T-DNA insertion across the genome was 2.4 insertions per Mb which agrees with similar analyses in *Arabidopsis* and rice (Rosso *et al.* 2003; Jeong *et al.* 2006). The distribution of T-DNA insertions over the five chromosomes of Brachypodium is non-uniform and appears to be correlated with gene density, with more insertions occurring at the distal ends of chromosomes than surrounding the centromere indicating a preference for insertion close to genes. In addition, very few T-DNA insertions occur in the short arm of chromosome 5, a region with a low gene-density and a high concentration of retrotransposons. T-DNA insertions in rice occur more frequently within gene-rich regions and at a much lower frequency within areas rich in transposon-related sequences (Zhang *et al.* 2007).

In general, the distribution of T-DNA insertions over each chromosome of rice and Brachypodium was similar and showed a low frequency of insertion at the centromere, gradually increasing as the distance from the centromere increases (Zhang *et al.* 2007). In *Arabidopsis*, the insertion frequency is also low at the centromere but increases more steeply as the distance from the centromere increases before leveling off at the end of the chromosome arms (Alonso *et al.* 2003). This difference presumably reflects the different genome organizations of monocots and dicots; genes in rice and Brachypodium are concentrated into a small gene space towards the distal ends of the chromosome arms whereas genes in *Arabidopsis* are distributed more evenly over the chromosomes.

A more detailed analysis of the locations of T-DNA insertion with respect to predicted genes showed a higher insertion frequency in genic regions than would be expected if T-DNAs inserted into the genome at random. This is mainly due to higher than expected levels of insertion into the 500 bp region upstream of an ATG start codon and to a lesser extent the 500 bp region downstream of a stop codon. Preferential T-DNA insertion into upstream and downstream regions flanking transcribed genes has previously been reported in rice (An *et al.* 2003; Zhang *et al.* 2007). In *Arabidopsis*, T-DNA insertion has also been shown to occur preferentially in promoter regions compared to the transcribed region of genes (Sessions *et al.* 2002).

Schneeberger and colleagues showed that the position of T-DNA insertion is positively correlated with an overrepresentation of cytosines compared to guanines at around 100 bp each side of an ATG start codon (Schneeberger *et al.* 2005). They also found that the DNA strand at the point of insertion showed a high level of flexibility indicating that both sequence composition and flexibility of the DNA strand may influence where a T-DNA will insert. T-DNA insertion may also be influenced by the methylation state of the DNA sequence as regions immediately upstream and downstream of transcribed sequences tend to exhibit lower methylation levels than the exons and introns of genes (Cokus *et al.* 2008). Such analyses were not performed for the T-DNA insertions described herein so no definite conclusions can be drawn in this regard but the similarity of T-DNA insertion patterns in the genomes of different plant species points towards a common mechanism for T-DNA insertion influenced by factors such as sequence composition and accessibility of the insertion site to the T-DNA.

This analysis has also shown that the frequency of T-DNA insertion into Brachypodium genes is highest at the initial and terminal exons with a low frequency of insertion within the central region of genes. There may be some bias in these results introduced by the method of calculation as there will always be an exon at the start and end of a gene but exon density may be lower in the central region due to introns. The insertion of T-DNA into rice genes also shows a preference for the 3’ and 5’ ends of genes (An *et al.* 2003).

The T-DNA insertions analysed here disrupt a total of 364 Brachypodium genes and more than 60 % of these genes show evidence of expression in Brachypodium and wheat/barley. Overall, 53 % of predicted genes in Brachypodium are supported by evidence of expression in Brachypodium and only 16 % by evidence of expression in wheat and barley. This indicates that T-DNAs insert preferentially into expressed genes, a finding that agrees with previous observations in *Arabidopsis* (Schneeberger *et al.* 2005).

The T-DNA insertion lines produced in this project are the first to be analysed in Brachypodium and although they comprise a relatively small dataset they provide a useful resource for

Brachypodium functional genomics. For high-throughput functional genomics, automated analysis is essential and due to the sometimes complex nature of T-DNA insertion, this analysis is non-trivial. A high-throughput and automated pipeline implementing some of the methods manually performed in this analysis is essential in order to scale-up the analysis of mutagenised Brachypodium lines. A number of groups within the IBI are currently producing and analysing additional lines as it is estimated that around 100,000 T-DNA lines will be required to tag a large proportion of Brachypodium genes (Thole *et al.* 2010). These resources will enable further biological investigations into this wild grass species and facilitate the elucidation of gene function in grasses, particularly in important crop species such as wheat and barley to complement the mutagenised populations that have already been developed in these species (Caldwell *et al.* 2004; Slade *et al.* 2005).

5 Physical mapping of wheat chromosome 3DL

5.1 Introduction

The importance of wheat to human welfare and civilisation cannot be overstated. It was one of the first grasses to be domesticated and today is the most widely grown crop in the world (Heun *et al.* 1997; USDA Foreign Agricultural Service 2010). The predicted increase in human population over the next 30 years will require a substantial increase in food production and increasing wheat yields will play an important part in addressing this need (Royal Society of London 2009). Despite this urgency, genomic resources for wheat are lagging behind those of other crops such as rice, sorghum and maize which have been subject of genome sequencing projects over the last five years (International Rice Genome Sequencing Project 2005; Paterson *et al.* 2009; Schnable *et al.* 2009). A sequenced genome provides a complete catalogue of genes in the genome and is the basis for understanding phenotypic variation. This information can be used to develop new varieties of crop plants with increased resistance to environmental stresses such as drought and salinity and increased resistance to disease. Bread wheat is an allohexaploid, meaning its genome consists of three homoeologous genomes (A, B and D) from closely related progenitors. In addition, the genome is estimated to be 17 Gb in size, over 40 times larger than the genome of rice (Bennett and Leitch 1995) and contains more than 80 % repetitive DNA (Smith and Flavell 1975). These factors mean that sequencing the genome of wheat probably represents one of the most challenging genome sequencing projects undertaken to date.

The various technologies used for DNA sequencing generate read lengths of between 80 bp and 1,000 bp, depending on the sequencing technology used. Standard Sanger sequencing, although low-throughput, still produces the longest read lengths. Even the relatively small genome of *S. cerevisiae* is several orders of magnitude larger than the longest reads (Goffeau *et al.* 1996) so the challenge of genome assembly is to reconstruct a genome sequence from these short reads. Genome assembly can be performed in two ways; the slower and more methodical method is to break the genome into fragments, order the fragments by physical mapping, then sequence each fragment (a hierarchical clone-by-clone approach), the other is to fragment the genome, sequence all the fragments, then reconstruct the sequence based on overlaps between the sequences (a whole-genome shotgun approach or WGS). Shorter reads, such as those produced using the Illumina platform, are more complex to assemble due to the number of reads required to obtain sufficient genomic coverage. Repeats such as tandem repeats, segmental duplications and transposable elements provide additional complexity to the process of genome assembly and repetitive regions occur with particularly high frequency in eukaryotic genomes. Reads obtained from one of these regions will be similar to the reads obtained from the duplicated region and the two regions cannot be resolved during the assembly stage. The accurate assembly of repeats requires reads to span the repeat either by virtue of the length of the read or by the use of a range of clone insert sizes (mate-pair or paired-end). Although faster due to the lack of an additional cloning step, the WGS method usually requires a more complex assembly stage and sequence contigs assembled from WGS reads generally need to be assessed by comparison to physical or genetic maps to validate their accuracy. The clone-by-clone approach is less dependant on repeats as genome fragments are ordered by physical mapping before sequencing.

There is no obvious fast, accurate and cost effective strategy to sequence the large, complex genome of wheat. A WGS approach would require deep genome coverage using clones of sufficient length to span repeats. Furthermore, the presence of homoeologous chromosomes would mean that assembling these reads using overlapping sequence and long-range scaffold information would be a formidable task. A clone-by-clone approach would first require the construction of a physical map from a set of fingerprinted clones. To achieve 10x coverage of the 17 Gb wheat genome would require nearly 1.2 million 150 kb insert clones which would need to be fingerprinted and assembled into contigs, a feasible but technically complex procedure. Contig assembly would be complicated by the presence of BACs from the three homoeologous chromosome sets as similar fingerprints may be generated from BACs in homoeologous regions which would be assembled into common contigs rather than homoeolog-specific contigs. A recent study simulated a physical map build using a merged set of fingerprinted BAC libraries made from homoeologous wheat chromosomes and found that the vast majority of contigs contained clones from a single library meaning that homoeolog-specific chromosomes had been constructed (Luo *et al.* 2010). More investigation is required to determine whether this approach can be scaled up to generate homoeolog-specific contigs from a genome-wide BAC library from wheat.

Physical map construction using HICF fingerprinting followed by SNaPshot fragment labelling (Luo *et al.* 2003) and assembly of contigs using FPC (Soderlund *et al.* 1997) is explained in Chapter 2. After initial contig assembly, a physical map generally consists of short kilobase-scale contigs comprised of overlapping BACs that need to be ordered and oriented according to the underlying sequence. This is done by integrating a physical map with a genetic map or a whole genome shotgun sequence assembly. Integrating a physical map and a genetic map requires the placement of genetic markers from the genetic map onto the physical map. This can be performed by hybridisation- or PCR-based approaches. Hybridisation-based approaches using overgo probes are limited, especially in genomes containing duplications or a high proportion of repetitive DNA as probes may hybridise to a number of clones. In contrast, PCR-based screening is more accurate provided that specific primers are designed. Screening a large library of BAC clones is very time-consuming but can be made more efficient by adopting a pooling strategy. The theoretical basis behind N-dimensional pooling was demonstrated in the early 1990s (Barillot *et al.* 1991) and involves arranging a BAC library in such a way as to reduce the number of PCR reactions needed to screen the library. Screening a pooled library results in a set of coordinates, one for each dimension of the pool, which are deconvoluted to identify the correct BAC clone. A three-dimensional pooling strategy was adopted to allow screening of a BAC library made from DNA extracted from the Chinese Spring wheat cultivar (Febrer *et al.* 2009). The library consisted of 715,776 clones arranged in 1,817 384-plates and represented 5 haploid genome equivalents. Pooling the clones meant that 17 PCR reactions using a multi-channel pipette are required to screen the library compared to 1,817 PCR reactions if the plates were not pooled. I wrote a Perl script to identify plates containing candidate BAC clones from a list of coordinates obtained from screening. The BAC pool and screening script was used to identify clones containing the reduced height (Rht-1a) gene from the BAC library. Thirteen clones were identified and sequence analysis of the amplified products showed that all three Rht homoeologues were represented.

In wheat, anchoring physical map contigs to genetic maps is difficult due to the differing rates of recombination in different parts of the chromosomes. Rates of recombination are higher at the distal ends of chromosomes compared to the pericentromeric regions (Saintenac *et al.* 2009). This means that in centromeric regions a much larger physical distance is represented by one unit of genetic distance than in the distal regions and genetic markers cluster in peri-centromeric regions. In addition, the low level of polymorphism exhibited between different varieties of wheat means that is it difficult to identify suitable polymorphic markers with which to genetically anchor contigs.

NGS sequencing platforms are able to generate huge quantities of sequence data at reduced cost when compared to traditional Sanger sequencing. Ultimately these platforms or their successors will provide the means to sequence and assemble large and complex eukaryotic genomes such as wheat. Recently, a draft sequence of the 2.4 Gb giant panda genome was assembled using SOAPdenovo from reads generated using the Illumina platform, proving that large genomes can be sequenced in this way (Li *et al.* 2010). A group of UK scientists, including members from my group, recently announced that 5x coverage of the wheat genome had been achieved using the 454 platform (JIC Press Office 2010). New assembly methods such as Curtain (http://code.google.com/p/curtain) and Cortex (M. Caccamo, pers. comm.) are being developed to assemble datasets such as this although due to the quantity and complexity of the data, this is a challenging and computationally intensive task. On the horizon are so-called third-generation platforms such as the SMRT technology being developed by Pacific Biosciences (Eid *et al.* 2009) and the Nanopore technology being developed by Oxford Nanopore (Astier *et al.* 2006). These technologies will provide the means to sequence single DNA molecules and will generate much longer read lengths than is currently possible. In the case of wheat this should overcome the problems associated with assembling sequence reads from highly repetitive genomic regions. The strobe sequencing protocol available with the SMRT technology (Pacific Biosciences 2009) will allow short blocks of sequence to be read over very long insert lengths, rather than the traditional mate-pair reads from each end of an insert. For the highly repetitive wheat genome, this will produce reads that span repetitive regions and provide long-range scaffolds for the assembly of shorter reads that provide deep coverage.

An approach that has recently been applied to the large genomes of various *Triticeae* species including wheat, is flow cytometry (Dolezel *et al.* 2007). This method enables the separation of individual chromosomes within a genome and is especially useful for species with large, complex genomes. Once separated, individual chromosomes from these large genomes can be tackled as separate genomic entities using a hierarchal clone-by-clone approach, a WGS approach or by NGS sequencing methods.

For flow cytometry, a chromosome suspension is prepared by the mechanical homogenisation of formaldehyde-fixed root tips and the suspension is flowed through a light beam. Each chromosome causes the light to be scattered and the amount of light scattered is correlated with the size of the chromosome. The chromosome suspension is usually first stained with a DNA-binding fluorochrome which magnifies the scattering effect. Measuring the scattered light generates a series of peaks for each chromosome in the suspension, called a ‘flow karyotype’ which is used to differentiate each chromosome. In reality, the heterogeneity observed in the size of chromosomes present in a given species is not large enough to discriminate between all chromosomes but combining this technique with the various aneuploid stocks available in the *Triticeae* has enabled the separation of individual chromosomes and chromosome arms in most major *Triticeae* species.

TriticeaeGenome (www.triticeaegenome.eu) is a collaborative project involving a number of European partners and was instigated to coordinate research into wheat genomics. One of the main objectives of this collaboration is to produce physical maps of wheat chromosome groups 1 and 3 using flow-sorted chromosomes. From these physical maps, a MTP of clones can be selected for sequencing. The largest chromosome in wheat, 3B (993 Mb), was readily separated from the others in the wheat genome by flow cytometry and for this reason chromosome 3B was the first to be analysed (Paux *et al.* 2008). Physical mapping was initiated by producing a BAC library from purified 3B DNA and 11 Mb of BAC-end sequence was generated from clones within the library providing the most comprehensive picture of randomly generated wheat sequence up to that point (Paux *et al.* 2006). In total, 67,968 clones were fingerprinted and assembled using FPC (Soderlund *et al.* 1997) resulting in 1036 contigs. A minimum tiling path of 7,440 BAC clones was identified. The contigs were validated using a new type of marker designed for use in wheat called insertion site-based polymorphism (ISBP) markers (Paux *et al.* 2010). ISBP markers are based on junctions between transposable elements (TEs) and flanking sequence and consist of one primer within the TE and one in the flanking DNA. In addition to ISBP markers, EST and SSR markers were used to anchor contigs to both a genetic and a deletion bin map of chromosome 3B. The genetic map was constructed from a F_2 population of 376 individuals resulting from a cross between Chinese Spring and Renan cultivars and 213 contigs were anchored to this map. Additional anchoring using a deletion bin map was required due to the nonhomologous distribution of recombination events along the chromosome and the low levels of polymorphism detected between the mapping parents. The deletion bin map consisted of 16 intervals spanning the chromosome and anchored 599 contigs. Construction of this physical map provided a proof of concept for the chromosome-based approach in wheat genomics. Thirteen contigs from the physical map have now been sequenced using a combination of Sanger and 454 sequencing methods resulting in 18.2 Mb of sequence taken from different regions of chromosome 3B (Choulet *et al.* 2010).

I am part of a group at the JIC tasked with creating a physical map of the long arm of wheat chromosome 3D (3DL). Our approach closely follows the approach used to construct the wheat 3B physical map (Paux *et al.* 2008). A BAC library is constructed from flow-sorted 3DL genomic DNA, fingerprinted and assembled into contigs using FPC. These contigs are currently being anchored to a deletion bin map and a genetic map. In addition to this we have applied the novel approach of anchoring and ordering the physical map contigs using a region in the Brachypodium genome syntenic to 3DL to create a ‘syntenic build’. Physical map contigs are anchored to Brachypodium genes by designing markers from wheat ESTs aligning to the Brachypodium genes and in addition, BES from the MTP BACs are used to anchor contigs *in silico*. Other approaches to increase the number of anchored contigs were evaluated.

My contributions to this project centred mainly on the construction of the syntenic build. I designed markers from wheat ESTs to anchor physical map contigs to the Brachypodium syntenic region. These markers were complemented by BES anchoring. I also designed a database to record the results from screening markers against the MTP BACs which provided a basis from which to analyse and visualise the anchored contigs. In addition, I performed *in silico* analysis of BES data and designed ISBP markers to screen the MTP BAC pool.

5.2 Methods

5.2.1 Construction of 3DL BAC library and clone fingerprinting

Flow-sorting of chromosomes was performed by Jaroslav Dolezel's group in The Institute of Experimental Botany in the Czech Republic. A BAC library was constructed from purified 3DL DNA using the restriction enzyme *Hin*dIII. BAC library fingerprinting was performed at the Istituto de Genomica Applicata in Italy. DNA was extracted from the BACs, purified, then digested with multiple restriction endonucleases (*Eco*RI, *Xba*I, *Bam*HI, *Xho*I and *Hae*III). The resulting fragments were labelled using fluorescent SNaPshot chemistry and separated by electrophoresis (Luo *et al.* 2003). Fragment peaks were detected using an ABI3730 automated capillary sequencer and sized using GeneMapper. The GeneMapper output files were downloaded and I used GenoProfiler to extract the peak data, remove background contamination and vector-related bands and to detect cross contamination of clones in 384- and 96- well plates. The files produced by GenoProfiler were used in the FPC contig assembly.

5.2.2 Building physical map contigs using FPC

MF used Fingerprinted Contigs (Soderlund *et al.* 1997) to build contigs from the fingerprinted BACs according to TriticeaeGenome general guidelines (Paux *et al.* 2009). The first contig build used a stringent Sulston cut-off of $1e^{-75}$ designed to avoid the assembly of contigs from randomly overlapping clones. During the automated finishing stage the cut-off was reduced in a step-wise fashion to $1e^{-45}$. The contigs were manually finished using a cut-off of $1e^{-25}$ and the FPC 'Pick MTP clones' function was used to identify a minimum tiling path of clones.

5.2.3 Pooling and screening the MTP BAC library

The MTP clones were sent to the French Plant Genomic Resource Centre to be arranged into 16 384-well plates and pooled using a three-dimension pooling strategy to facilitate efficient screening. The library consisted of 24 column pools, 16 row pools and 16 plate pools as well as two wells containing superpools of all MTP clones to use as a positive control during screening. This pooling design allowed the library to be screened using 58 PCR reactions. The library was screened by MF and Aidong Yang using melting curve analysis on a Roche LightCycler 480 and I wrote a Perl script to deconvolute the pool coordinates to identity a specific BAC (Supplementary information 15).

5.2.4 Analysis of BES and identification of markers

The 5,826 BACs in the MTP were end-sequenced at the Istituto de Genomica Applicata using the vector pIndigoBAC-5 and the restriction enzyme *Hin*dIII. The raw ABI chromatogram and phred base call files were downloaded and converted to FASTA files using the phd2fasta script (Ewing *et al.* 1998). Quality trimming and vector screening was achieved using lucy (Chou and Holmes 2001). BES passing the quality requirements were masked for repeats using the TREP database (Wicker *et al.* 2002) and compared to wheat genes (NCBI unigenes build 56) using BLASTn (e-value < 1e-100). High and medium confidence insertion site-based polymorphism (ISBP) markers were identified from the wheat BES using a Perl script (isbpfinder.pl) provided by Etienne Paux.

5.2.5 Identification and assessment of syntenic 3DL region in Brachypodium

Identification of orthologous gene-pairs in the genomes of Brachypodium and wheat showed that the long arm of Brachypodium chromosome 2 is syntenic to the long arm of wheat group 3 chromosomes (Figure 3e; International Brachypodium Initiative 2010). To confirm this synteny, 11,014 bin-mapped wheat ESTs were aligned to the Brachypodium genome using BLASTn and the results visualised using CMap (Youens-Clark *et al.* 2009). Wheat chromosome 3D was compared to each Brachypodium chromosome in turn to determine which had the highest number of 3D bin-mapped wheat ESTs aligning to it.

A high resolution, EST-based genetic map was recently published for *Aegilops tauschii*, the diploid source of the D-genome of wheat (Luo *et al.* 2009). In this study a mapping population of 572 F_2 *Ae. tauschii* plants were genotyped to produce a genetic map containing 878 loci distributed over seven linkage groups. From this dataset, a set of 101 wheat ESTs was identified which had been mapped to loci covering the long arm of chromosome 3, syntenic to the region in wheat being physically mapped. These ESTs were aligned to Brachypodium genes using BLASTx (evalue < $1e^{-10}$) to assess collinearity between Brachypodium 2L and wheat 3DL to determine whether the syntenic approach would be useful.

5.2.6 Anchoring contigs to the *Ae. tauschii* genetic map

I designed markers from the 101 wheat ESTs genetically mapped in *Ae. tauschii* 3DL using Primer3 to identify PCR primer pairs (Rozen and Skaletsky 2000). These markers were first screened against 3DL specific genomic DNA, then screened against the MTP BAC pools to identify anchor points between the genetic map and physical map contigs.

5.2.7 Anchoring contigs to the syntenic region in Brachypodium

5.2.7.1 Anchoring contigs using PCR-based markers

Wheat ESTs were aligned to the 3,061 genes in Brachypodium 2L, the region syntenic to wheat 3DL, using BLASTn and PCR primer-pairs were designed from these ESTs to be used as markers (Rozen and Skaletsky 2000). The ESTs were aligned in two stages, first ESTs assigned to deletion bins on wheat 3DL were used, then all other wheat ESTs. In addition, 17 PCR-based Landmark Unique Gene or PLUG markers (Ishikawa *et al.* 2007) and 187 Conserved Orthologous Sequence (COS) markers (Quraishi *et al.* 2009) were included in the experiment. The sequences from which the PLUG and COS markers were derived from were anchored to genes in the Brachypodium syntenic region using BLASTn. All markers were first screened against 3DL genomic DNA, then screened against the MTP BAC pools if a 3DL-specific product was observed.

5.2.7.2 Anchoring contigs using BES

First the BES were masked for repeats using the TREP database of repeats (Wicker *et al.* 2002). Non-repetitive BES were aligned to genes in the Brachypodium syntenic region using BLASTx and considered to be anchored if they showed greater than 70 % sequence identity over greater than 50 bp in length.

5.2.7.3 Visualisation of syntenically anchored contigs

I designed a MySQL database to record the position of each MTP BAC clone in the physical map contigs, and to store the results from the marker screening. Perl scripts were used to query the database to link MTP BACs (and therefore physical map contigs) to genes in the Brachypodium syntenic region. In this way, the physical map contigs could be oriented and ordered according to the syntenic region in Brachypodium. CMap was used to visualise the contigs anchored to the scaffold.

5.2.8 Detailed analysis of a small Brachypodium syntenic region

A small region of Brachypodium chromosome 2 was identified and contigs predicted to be anchored in this region were visually examined using CMap to assess anchoring methods. Contigs that were not anchored robustly were discarded at this point. Brachypodium genes and anchored contigs in this region were used to assess additional anchoring methods.

5.2.8.1 Designing markers from Brachypodium genes

In order to increase the marker density, markers were designed directly from Brachypodium genes in the test region which had no aligning wheat EST. These markers were screened against the MTP BACs to determine if they could be used to confirm already anchored contigs, or to anchor new contigs to the syntenic region.

5.2.8.2 Using ISBP markers identified from anchored contigs

ISBP markers were identified from the BES of BACs at both ends of each contig anchored within the test region. Ideally, each contig should have four markers, two from each terminal BAC (Figure 5.1). These markers were screened against the MTP BAC pools to test the robustness of the contigs constructed by FPC and also to identify BACs in new contigs which may overlap, thus extending contigs. Markers designed from the internal BES on a terminal BAC (ISBP-2 and ISBP-3 in Figure 5.1) should detect an overlap with the neighbouring BAC in the contig and markers designed from terminal BES (ISBP-1 and ISBP-4 in Figure 5.1) may detect an overlap with a new contig. If successful, this approach can be used on all contigs both to validate the contigs and to extend them by joining two contigs.

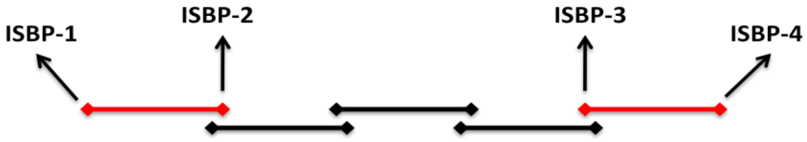


Figure 5.1: Designing ISBP markers from terminal BAC BES from each anchored contig.
Terminal BACs on a hypothetical contig are shown in red and central BACs in black. BES are indicated at the end of each BAC. Four ISBP markers should be identified from each contig, two from internal BES (ISBP-2 and ISBP-3) and two from terminal BES (ISBP-1 and ISBP-4). Markers from internal BES should identify an overlap with an adjacent BAC in the same contig. Markers from terminal BES may identify overlaps with new contigs.

5.3 Results

5.3.1 Construction of 3DL BAC library and clone fingerprinting

Based on estimates from flow-sorting the wheat chromosomes, the size of wheat chromosome arm 3DL was estimated to be 449 Mb this could be sorted to a purity of 86 % from a chromosome preparation of hexaploid wheat (Dolezel 2008). The BAC library made from 3DL-specific DNA contained 55,780 clones with an average insert size of 105 kb. The estimated coverage of 3DL was 11.2x.

5.3.2 Building physical map contigs using FPC

A total of 5,478 contigs resulted from the initial FPC build which was reduced to 2,848 contigs after the automated finishing stage. The manual finishing stage produced a final set of 1,000 contigs. The contigs contained 35,689 clones (64.0 % of total) and 20,091 clones remained as singletons. The average band size was estimated to be 1.1 kb based on 152 sequenced and fingerprinted BAC clones and the contigs contained 367,848 bands in total meaning that the physical map contigs represented approximately 404.6 Mb of sequence or 90.1 % of the 449 Mb chromosome arm. The distribution of contig lengths is shown in Figure 5.2. The mean contig length is 404.6 kb and the longest contig is contig 181 which contains 1,859 bands equating to an approximate length of 2,044 kb. A minimum tiling path of 5,826 clones was identified from these contigs.

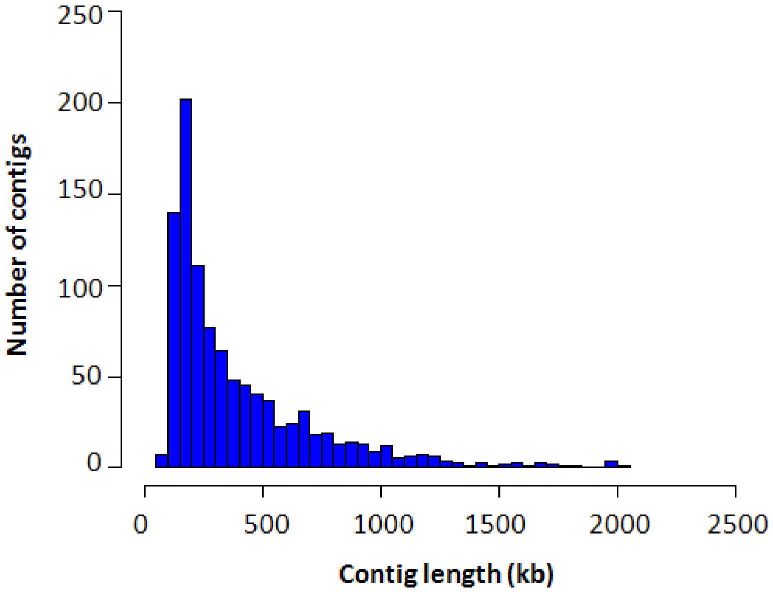


Figure 5.2: Distribution of contig lengths from wheat 3DL BACs after the FPC assembly stage.

5.3.3 Analysis of BES and identification of markers

After quality trimming and vector removal, 10,900 BES remained. This represented 94 % of expected sequence considering 5,826 MTP clones were end-sequenced. The total length of sequence reads was 7,062,723 bp, approximately 1.6 % of the estimated length of 3DL. The reads obtained were between 101 and 909 bp in length. After masking repeats, 3,828 (35.1 %) BES remained and 94 (0.86 %) of these showed good homology to wheat unigenes. Using the isbpfinder.pl script, 1,083 high and 4,390 medium confidence ISBP markers were designed from the MTP BES. This provided markers for 2,967 (50.9 %) of the MTP clones which have the potential to anchor 877 (87.7 %) of the 3DL physical map contigs.

5.3.4 Identification and assessment of syntenic 3DL region in Brachypodium

Figure 5.2 shows the comparison between wheat chromosome 3D and each Brachypodium chromosome obtained by aligning wheat ESTs assigned to deletion bins to the Brachypodium chromosomes. A higher density of ESTs align to Brachypodium chromosome 2 (Bd2) compared to the other chromosomes confirming that the long arm of Bd2 is syntenic to wheat group 3 chromosomes.

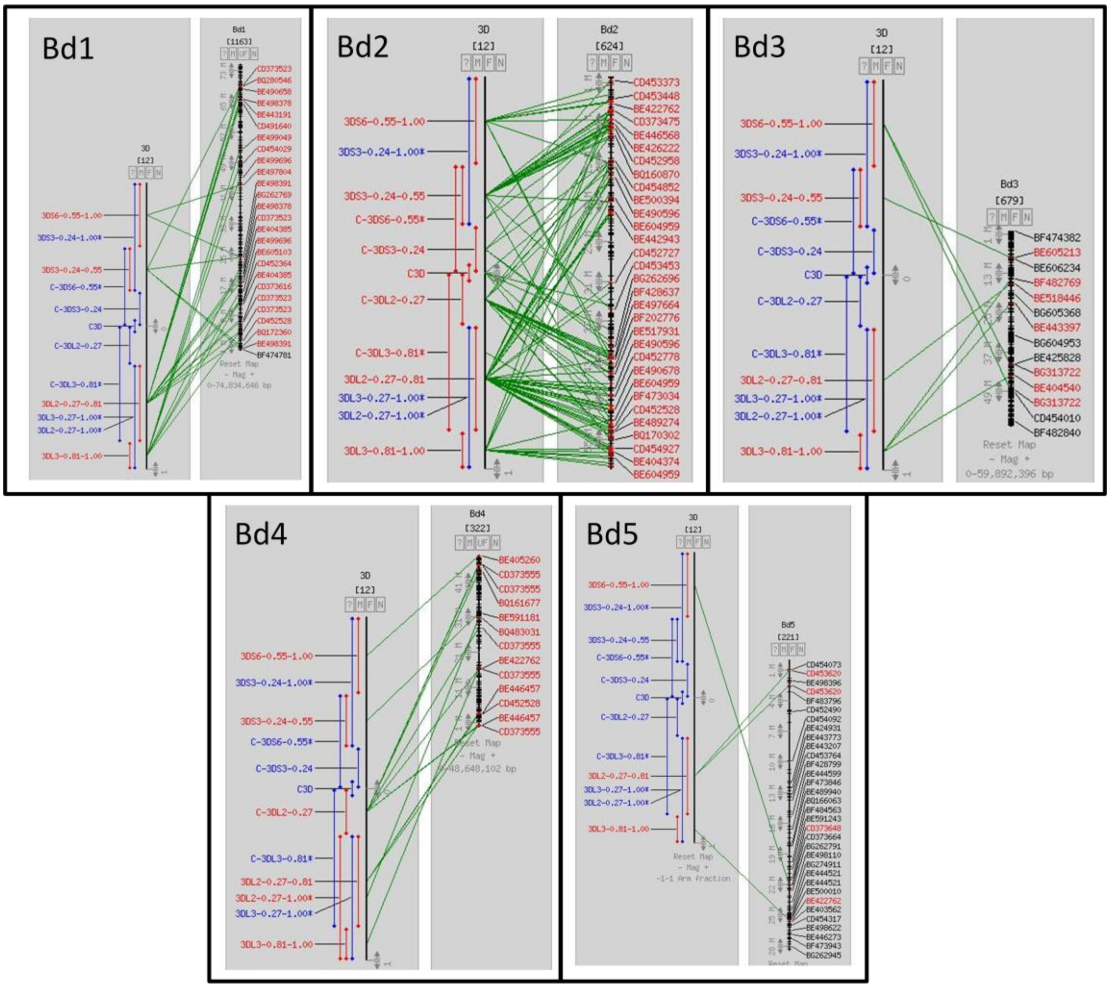


Figure 5.3: Brachypodium chromosomes (Bd1-5) aligned to wheat chromosome 3D by deletion bin mapped ESTs. In each panel, the deletion bins on the wheat chromosome are shown on the left and the position of the ESTs on the Brachypodium chromosomes are shown on the right. The green lines connect each EST to the deletion bin to which they are assigned. A high density of ESTs from wheat 3D is observed aligning to Bd2 compared to the other chromosomes.

The region of Brachypodium chromosome 2 syntenic to wheat 3DL is shown at a larger scale in Figure 5.4. There are three main deletion bins covering this portion of the wheat genome, 3DL3-0.81-1.00 at the distal end, 3DL2-0.27-081 in the central region and C-3DL2-0.27 near the centromere. ESTs from all three bins align to Bd2. The long arm of Brachypodium chromosome 2 is approximately 30 Mb in length and contains 3061 genes. Synteny to wheat appears to be highest in the 40 to 60 Mb region of this chromosome where most of the ESTs align. From 29 to 40 Mb there are very few aligning ESTs.

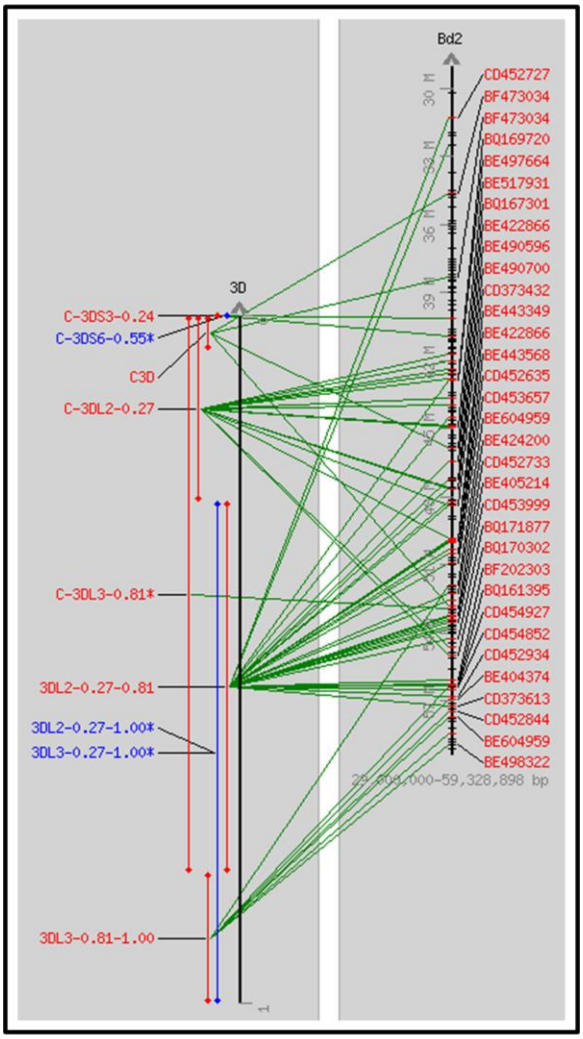


Figure 5.4: Comparison between Bd2 and wheat 3DL shown at a larger scale.
Wheat 3DL deletion bins are shown on the left and the 29-59 Mb region of Bd2 is shown on the right. The position of wheat ESTs is shown on Bd2 and green lines connect each EST to its originating deletion bin. Synteny appears to be highest in the 40 to 60 Mb region of Bd2 where most of the ESTs align. From 29 to 40 Mb there are very few aligning ESTs.

Twenty-eight of the ESTs genetically mapped in *Ae. tauschii* showed good homology to genes on the long arm of Brachypodium chromosome 2, the region syntenic to wheat 3DL, and a comparison between the maps is shown in Figure 5.5.

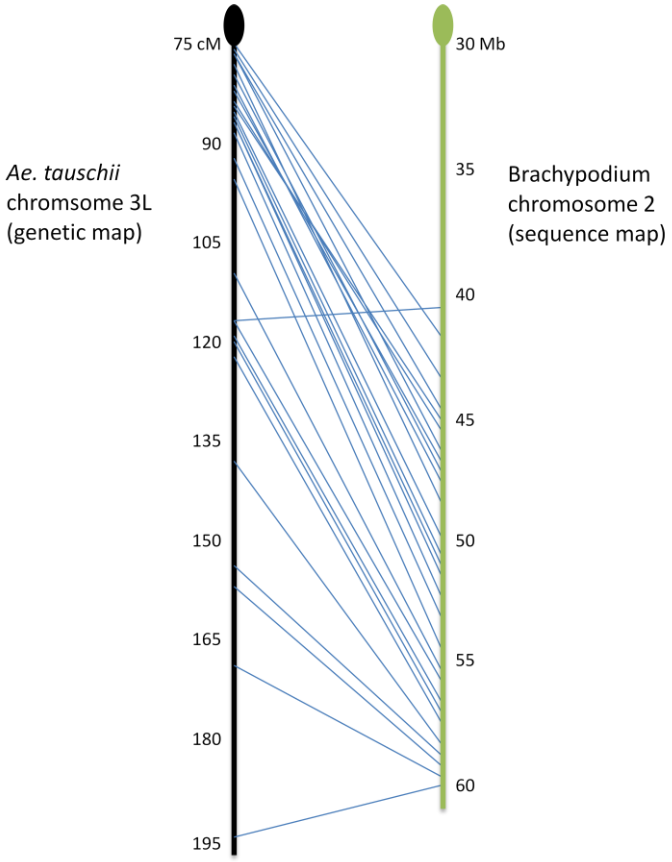


Figure 5.5: Comparison between *Ae. tauschii* 3L and Brachypodium 2L.

The region of the *Ae. tauschii* genetic map is shown as a black line with distances in centiMorgans and Brachypodium chromosome 2L is shown as a green line with distances shown in megabases. Blue lines link the position of each EST on the two maps. The centromere of each chromosome is shown at the top of each map as an oval.

This comparison shows that with the exception of a small inversion between 80 and 90 cM and a single EST genetically mapped to 115 cM aligning to a non-collinear gene, there is very good collinearity between this region of *Ae. tauschii* and Brachypodium sequence map. However, the distribution of the ESTs on each map is strikingly different with the ESTs from the genetic map covering the whole chromosome arm and the same ESTs in Brachypodium covering the 40 to 60 Mb region of the chromosome. No ESTs align to genes found in the 10 Mb region proximal to the centromere indicating a break in synteny within the overall alignment. It should be noted that the genetic and physical distances shown in these maps cannot be directly compared and the unequal interval sizes between markers on each map is due to differing rates of recombination along the chromosome. On the genetic map, markers appear to cluster in the peri-centromeric region but are evenly distributed on the physical map.

5.3.5 Anchoring contigs to the *Ae. tauschii* genetic map

Screening the 101 markers designed from genetically mapped ESTs against 3DL genomic DNA indicated that 71 would potentially be useful for anchoring contigs. These were screened against the MTP BAC pools and 32 could be anchored to BACs. These anchor points enabled 30 unique contigs to be anchored to the *Ae. tauschii* genetic map (Table 5.1).

EST name	Genetic map position (cM)	Marker name	BAC name	Contig	Position in contig
BE446403	73.65	Ta-EST-0026	3DL048_B12	5551	7
BQ162314	73.72	At-GEN-0010	3DL107_B19	24	9
			3DL108_D12	24	5
BG262775	76.02	Ta-EST-0109	3DL140_G16	5593	1
BE403201	76.73	Ta-EST-0014	3DL014_A11	461	6
			3DL123_O08	461	5
BG607570	77.76	Ta-EST-0117	3DL058_I07	411	13
			3DL136_H24	411	12
BE500330	77.88	At-GEN-0001	3DL071_I24	4089	2
			3DL119_D14	4089	1
BE585797	77.95	At-GEN-0038	3DL122_F18	5593	5
BE494888	78.38	Ta-EST-0078	3DL011_H08	916	1
			3DL048_K07	916	2
BG262775	78.38	Ta-EST-0109	3DL140_G16	5593	1
BE444252	80.22	Ta-EST-0133	3DL071_L22	104	18
			3DL082_M04	104	19
BF429272	80.89	At-GEN-0008	3DL043_F03	202	17
			3DL144_J24	202	18
BE607113	81.28	At-GEN-0007	3DL069_M14	1225	2
			3DL165_F03	1225	3
BE494474	85.76	Ta-EST-0030	3DL150_A14	617	2
			3DL153_K23	617	1
BE490651	86.86	Ta-EST-0073	3DL077_I16	60	21
BF483884	86.86	At-GEN-0002	3DL102_H20	1057	4
			3DL164_L18	1057	3
BG313636	87.05	Ta-EST-0180	3DL063_N02	2785	3
			3DL075_J22	2785	2
			3DL151_F21	2785	1
BF478406	92.91	Ta-EST-0149	3DL061_A23	99	10
			3DL095_H05	99	9
BE443349	97.60	Ta-EST-0130	3DL073_L08	14	5
			3DL165_K06	14	4

Table 5.1: Contigs anchored to the *Ae. tauschii* genetic map using markers designed from genetically mapped ESTs. Each EST is shown with its position on the genetic map, the marker designed from it and the BAC (or BACs) identified by screening the MTP pool. The contig containing the BAC and the position of the BAC within the contig is also shown.

EST name	Genetic map position (cM)	Marker name	BAC name	Contig	Position in contig
BE444392	110.07	At-GEN-0043	3DL131_B12	139	5
			3DL143_B07	139	6
BG262785	114.55	At-GEN-0037	3DL058_I07	411	13
			3DL162_P11	411	15
BE497664	114.91	Ta-wEST-0049	3DL096_K02	376	2
			3DL077_K24	376	1
BF200942	114.91	At-GEN-0046	3DL094_E14	316	8
			3DL145_P13	316	7
BF200549	116.51	Ta-EST-0145	3DL095_F05	234	4
BE444736	117.77	Ta-EST-0134	3DL045_E15	195	4
BF474720	122.08	At-GEN-0014	3DL132_P02	1698	7
			3DL153_E16	1698	6
BE443397	125.18	At-GEN-0034	3DL062_K23	4	3
			3DL131_E01	4	4
BF483498	137.13	At-GEN-0019	3DL150_J13	217	12
BE488620	155.13	At-GEN-0053	3DL023_L18	66	12
			3DL108_I08	66	11
BE426763	157.53	Ta-EST-0063	3DL003_O12	38	10
			3DL079_F01	38	9
			3DL130_G16	38	6
BE444864	169.32	At-GEN-0039	3DL086_F20	38	3
			3DL130_G16	38	6
			3DL125_B13	3863	4
			3DL167_G09	3863	3
			3DL027_H05	4665	1
			3DL086_K13	4665	2
BE443092	175.40	At-GEN-0020	3DL007_K23	1185	6
BM137927	189.32	At-GEN-0042	3DL041_B20	224	2
			3DL114_J09	224	1

Table 5.1 (contd.)

It is evident from these results that although some markers anchor single BACs (eg. At-GEN-0020), many markers anchor more than one BAC. However, these BACs are usually from the same contig indicating that there is some degree of overlap between BACs in a contig. For example, marker At-GEN-0042 gives a positive screening result for two BACs in contig 224 and these BACs are located at positions 1 and 2 of this contig. A single marker (At-GEN-0039) anchors 3 different contigs, suggesting a duplicated gene in the BACs that comprise these contigs.

Thirty contigs are anchored in total, 27 of these are anchored to a single position. Contigs 5593 and 411 are anchored to multiple distinct positions indicating that BACs within these contigs contain a duplicated gene. The markers that anchor contig 5593 are genetically very close with only 2.38 cM separating them so this contig could be anchored uniquely. Supporting this hypothesis, two of the three markers anchoring this contig anchor different BACs within the contig at positions 1 and 5. Contig 38 is anchored by two markers in adjacent positions on the genetic map indicating that this contig may span these two positions. Additional evidence for this is shown by the positions of the anchoring BACs within the contig; the first marker anchors at positions 10, 9 and 6 with the second marker anchoring to BACs at positions 6 and 3. It is possible that contig 38 extends from genetic map position 157.53 cM to 169.32 cM.

The 30 anchored contigs are not distributed evenly over the region of the genetic map with 15 contigs anchored to the first 20 cM of the genetic map. The remaining 15 contigs are anchored to a region of more than 90 cM. This reflects the general distribution of the genetic markers in the original *Ae. tauschii* map which are more highly concentrated in the region closest to the centromere. This is likely due to the rate of recombination being lower in this region meaning that a larger physical distance is represented by this part of the genetic map and more markers are concentrated in this region.

5.3.6 Anchoring contigs to the syntenic region in Brachypodium

5.3.6.1 Anchoring contigs using markers

A total of 1,557 markers (Supplementary information 16) were designed from wheat ESTs aligning to genes within the syntenic region of Brachypodium 2L. These markers included 181 designed from ESTs assigned to a deletion bin on wheat 3DL. These markers were screened against 3DL genomic DNA and 726 (46.6%) amplified a PCR product. The markers designed from bin-mapped ESTs were more successful than the markers designed from non bin-mapped ESTs with 62.4 % amplifying a PCR product from 3DL genomic DNA compared to 44.5 % of non bin-mapped markers. The markers amplifying a 3DL-specific product were screened against the MTP BAC pools and 547 (35.1%) were anchored to BACs. Including the PLUG and COS markers, this experiment anchored 320 unique contigs (32 %) to the Brachypodium syntenic region (Table 5.2).

Marker type	Number of markers	Amplified in 3DL	Anchored to MTP BAC	Anchored vs. Amplified	Number of contigs anchored
Bin-mapped EST	181	113	82	73%	79
Non bin-mapped EST	1376	613	465	76%	283
PLUG	17	14	10	71%	10
COS	187	102	62	61%	55
Total	**1761**	**842**	**619**	**70%**	**320**

Table 5.2: Marker screening results from anchoring contigs using markers.

The last column shows the number of unique contigs anchored by each group and the total for this column shows the number of unique contigs anchored by all markers.

The markers designed from bin-mapped ESTs, the PLUG markers and the COS markers were generally more successful than the markers designed from non bin-mapped ESTs. However, the percentage of 3DL-specific markers that successfully anchored an MTP BAC was less variable between different marker types; between 61 and 76 % of markers were useful.

Nearly 100 contigs were anchored multiple times to the scaffold region. This explains why the total number of unique contigs anchored (320) is less than the total number of contigs anchored by all the markers (427).

5.3.6.2 Anchoring contigs using BES

After masking the BES for repeats, BLASTx aligned 46 (0.4%) BES to 57 genes in the Brachypodium syntenic region, anchoring 47 unique contigs.

The 1,000 contigs of the physical map were classified by how they are anchored to the Brachypodium scaffold (Table 5.3). Thirty-five contigs were anchored by both BES and PCR-based anchoring methods. The majority of anchored contigs (287) were anchored by PCR-based methods only and 8 were anchored by BES only. As of July 2010, approximately two-thirds of contigs (670) were not anchored to the Brachypodium syntenic region.

	Number of contigs
Contig anchored by BES AND marker(s)	35
Contig anchored by marker(s) only	287
Contig anchored by BES only	8
Contig not anchored	670
Total	**1000**

Table 5.3: Classification of physical map contigs based on their anchoring to the Brachypodium syntenic region.

5.3.6.3 Visualisation of syntenically anchored contigs

Anchor points obtained from both marker anchoring and BES anchoring were loaded into CMap for visualisation. Due to the complexity of the data, visual inspection proved to be the only reliable way of determining where a contig was anchored. Many anchoring scenarios were observed including contigs anchored by multiple syntenic anchor points (Figure 5.6), contigs anchored by a single anchor point (Figure 5.7) and contigs anchored to multiple positions in the scaffold (Figure 5.8). The various different anchoring scenarios can be used to classify anchored contigs, for example a contig anchored by multiple syntenic anchor point is considered to be well anchored but further investigation is required to determine the correct position of a contig anchored to multiple places in the Brachypodium syntenic region.

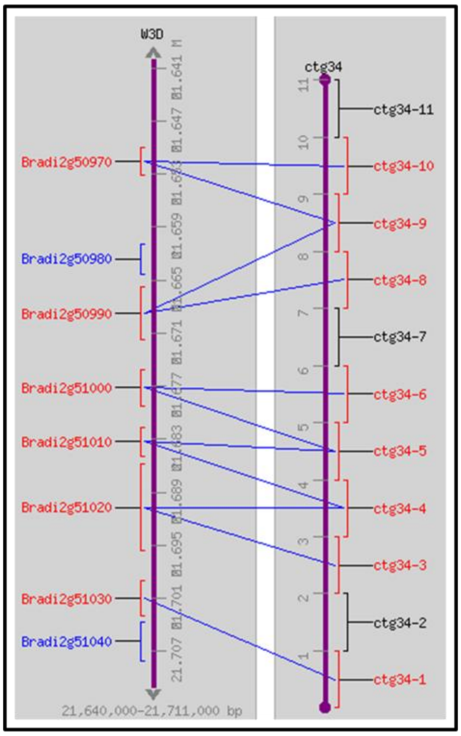


Figure 5.6: Physical map contig 34 anchored by multiple syntenic anchor points.

The Brachypodium genes are displayed in the left panel and the 11 MTP BACs comprising the contig are shown in the right panel (labelled ctg34-1 to ctg34-11). Blue lines indicate BACs that are anchored to Brachypodium genes. Eight of the BACs in the contigs are anchored to neighbouring genes in Brachypodium (Bradi2g50970 to Bradi2g51030) indicating good collinearity between these regions of Brachypodium and wheat.

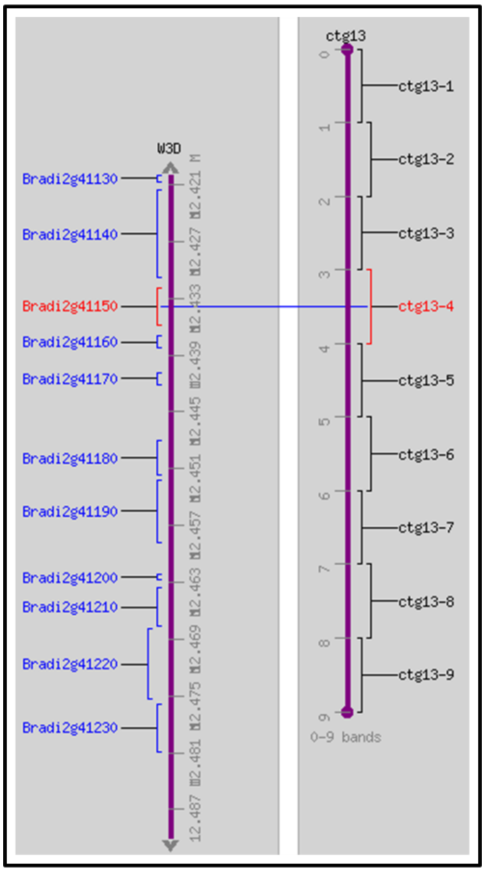


Figure 5.7: Physical map contig 13 anchored by a single anchor point.

The Brachypodium genes are displayed in the left panel and the 9 MTP BACs comprising the contig are shown in the right panel (labelled ctg13-1 to ctg13-9). The blue line indicates that the fourth BAC in the contig is anchored to Brachypodium gene Bradi2g41150.

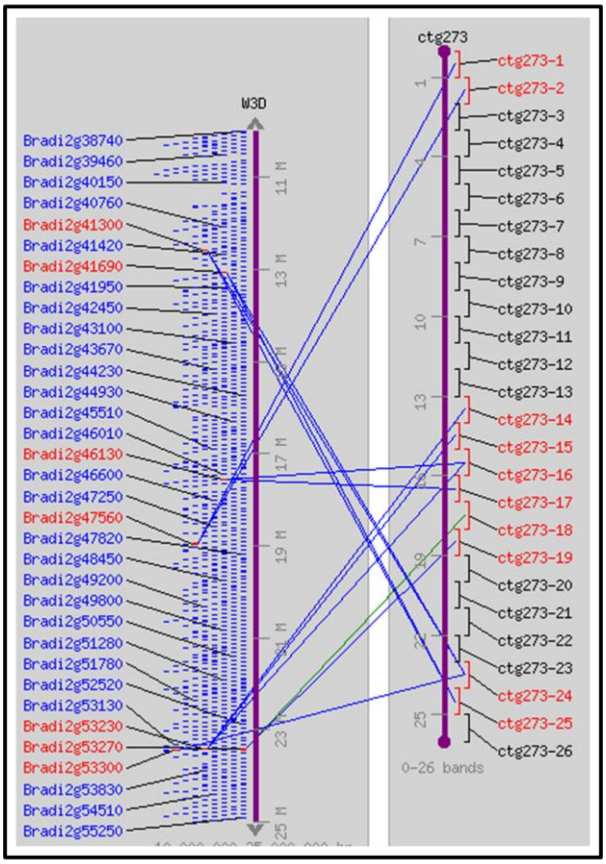


Figure 5.8: Physical map contig 273 anchored to multiple places in the Brachypodium syntenic region.
A 20 Mb region of Brachypodium is displayed on in the left panel and the 26 MTP BACs comprising the contig are shown in the right panel (labelled ctg273-1 to ctg273-26). The blue lines indicate that BACs from different parts of the contig are anchored to different genes within the syntenic region.

A number of well anchored contigs show evidence of synteny breakdown between wheat and Brachypodium. An example is shown in Figure 5.9 where a contig anchored by three anchor points suggests that the relative order of two neighbouring genes is different between Brachypodium and wheat.

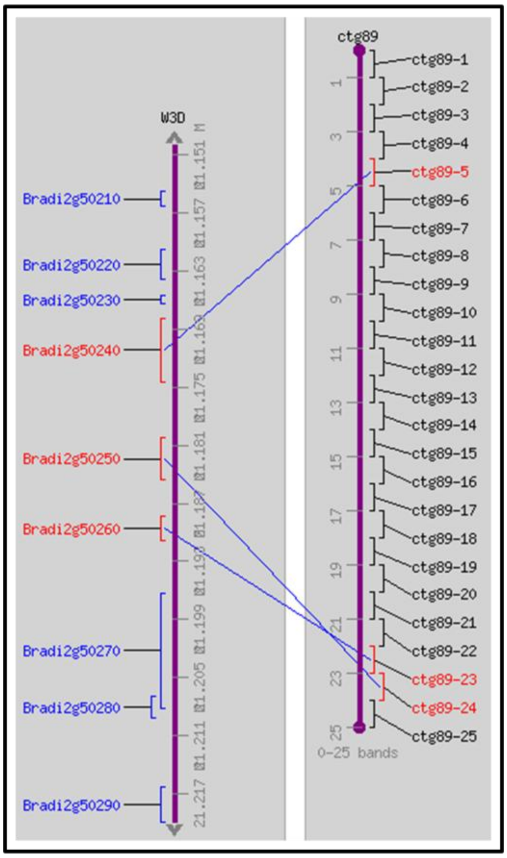


Figure 5.9: Physical map contig 89 anchored to the Brachypodium syntenic region.

The Brachypodium genes are displayed in the left panel and the 25 MTP BACs comprising the contig are shown in the right panel (labelled ctg89-1 to ctg89-25). The blue line indicates that the three BACs are anchored to three Brachypodium genes and that the order of two of these genes is different in wheat compared to Brachypodium.

5.3.6.4 *Comparing the genetic anchoring to the syntenic anchoring*

To assess the value of the syntenic build approach I compared the location of each contig anchored to the *Ae. tauschii* genetic map to its location in the syntenic build. There are 15 contigs anchored to both maps (Table 5.4). All but three of these contigs are anchored in the same relative order in each map indicated by cM position or incrementing Brachypodium gene names. Contigs 411, 104 and 376 are the exceptions to this pattern and are anchored to non-syntenic positions.

Contig	Position on genetic map (cM)	Anchored to gene in syntenic build
5551	73.65	Bradi2g43190
461	76.73	Bradi2g45580
411	77.76	Bradi2g48260
916	78.38	Bradi2g47600
5593	78.38	Bradi2g47760
104	80.22	Bradi2g46790
617	85.76	Bradi2g50290
60	86.86	Bradi2g51470
2785	87.05	Bradi2g51660
99	92.91	Bradi2g53620
14	97.60	Bradi2g54950
376	114.91	Bradi2g42360
234	116.51	Bradi2g58680
195	117.77	Bradi2g59050
38	157.53	Bradi2g60870

Table 5.4: Comparison between physical map contigs that are both genetically and syntenically anchored.

5.3.7 Detailed analysis of small syntenic region

A 1.7 Mb region of the Brachypodium chromosome 2 was identified spanning from 48.1 Mb to 49.8 Mb. This region contains 204 genes, Bradi2g47730 to Bradi2g49770. Forty contigs were anchored in this region by PCR-based markers or BES (Supplementary information 17).

5.3.7.1 Designing markers from Brachypodium genes

Ninety-four markers were designed directly from Brachypodium genes in the test region which had no aligning wheat EST (Supplementary information 18). Eighteen (19.1 %) of these markers amplified in 3DL genomic DNA and of these, 9 could be anchored to MTP BACs (Table 5.5). All but one marker anchored BACs from a single contig. Marker BD-CDS-0076 anchored BACs from two contigs (481 and 1072). In total, 10 contigs could be anchored using this method.

Marker	Anchored to BAC	Contig	BAC position in contig
BD-CDS-0012	3DL148_H24	1044	5
BD-CDS-0020	3DL006_A07	26	2
	3DL012_P09	26	1
BD-CDS-0039	3DL149_E07	154	1
	3DL159_O03	154	2
BD-CDS-0045	3DL055_C02	453	6
BD-CDS-0046	3DL035_C21	14	11
	3DL087_G20	14	8
BD-CDS-0064	3DL091_A14	341	2
	3DL134_D07	341	1
BD-CDS-0076	3DL052_N13	481	3
	3DL065_D22	481	4
	3DL057_A18	1072	8
	3DL062_A17	1072	7
	3DL074_M16	1072	6
BD-CDS-0088	3DL069_M14	1225	2
	3DL165_F03	1225	3
BD-CDS-0092	3DL026_E05	427	7
	3DL095_I24	427	8

Table 5.5: Physical map contigs anchored by markers designed from Brachypodium genes.

When these 10 contigs were integrated with contigs already anchored in the test region, 7 of them were already anchored to neighbouring genes (Table 5.6). The remaining 3 contigs (14, 453 and 481) were new contigs, not already anchored in the test region.

Gene	Anchored contig	Position in contig	Anchoring method
Bradi2g47990	1044	9	Marker
Bradi2g48010	1044	5	Marker
	4070	3	Marker
Bradi2g48020	1044	5	Bd marker
Bradi2g48160	26	4	Marker
Bradi2g48190	26	1	Bd marker
Bradi2g48470	154	14	BES
Bradi2g48480	154	14	BES
Bradi2g48490	154	14	BES
Bradi2g48550	154	4	Marker
Bradi2g48590	154	1	Bd marker
Bradi2g48690	350	20	Marker
Bradi2g48710	453	6	Bd marker
Bradi2g48720	805	6	Marker
Bradi2g48730	14	11	Bd marker
Bradi2g49100	341	5	Marker
Bradi2g49110	341	4	Marker
	341	1	Bd marker
Bradi2g49330	481	3	Bd marker
	1072	6	Bd marker
Bradi2g49340	1072	7	Marker
Bradi2g49380	1072	6	Marker
Bradi2g49530	1225	8	Marker
Bradi2g49540	1225	8	Marker
Bradi2g49570	1225	3	Marker
	1225	2	BES
Bradi2g49580	1225	2	Bd marker
Bradi2g49630	427	7	Marker
Bradi2g49640	427	7	Bd marker

Table 5.6: Integration of contigs anchored by Brachypodium markers with previously anchored contigs. All 10 contigs anchored by Brachypodium markers (labelled as Bd marker) are shown alongside contigs already anchored (labelled as Marker or BES) in the surrounding region. Seven contigs anchored by Brachypodium markers are already anchored by existing methods and three contigs (14, 453 and 481) are new contigs.

An example of agreement between the different anchoring strategies is shown by contig 154 which consists of 14 BACs (Figure 5.10). Two BACs are already anchored to Brachypodium genes (1 by a marker and 1 by 3 BES hits) and a marker designed from a Brachypodium gene provides an additional anchor point.

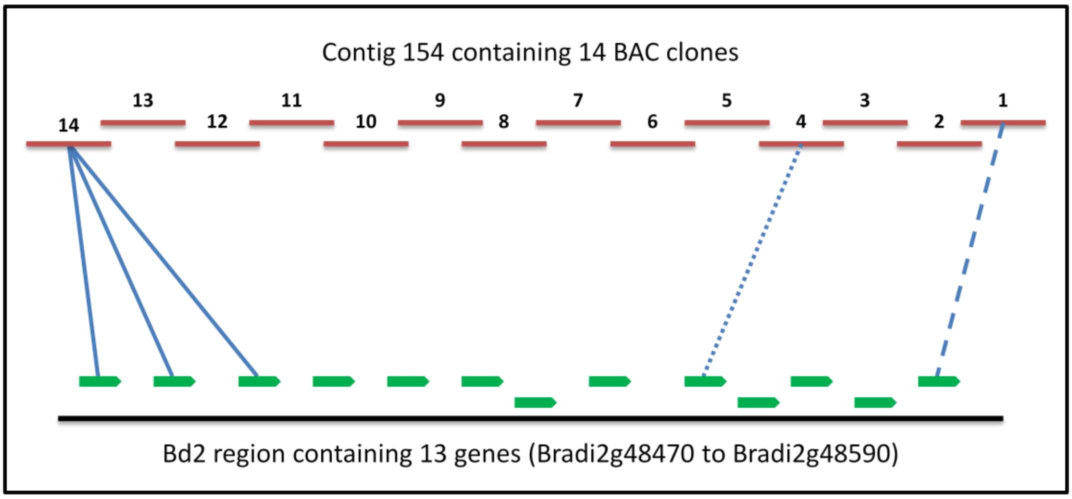

Contig 154 containing 14 BAC clones
14
13
12
11
10
9
8
7
6
5
4
3
2
1
Bd2 region containing 13 genes (Bradi2g48470 to Bradi2g48590)

Figure 5.10: Agreement between different anchoring strategies.

Contig 154 consists of 14 MTP BACs (red lines), one of which is anchored by three BES hits (solid blue lines) and one by a marker (dotted blue line). An additional BAC is anchored by a marker designed from a Brachypodium gene (dashed line).

5.3.7.2 Using ISBP markers identified from anchored contigs

ISBP markers were identified from 30 of the 40 contigs anchored in the test region but four markers could not be identified from each contig. In total, 42 markers were identified for analysis (Supplementary information 19), 13 of these were high confidence markers and 29 medium confidence. Twenty-one (50 %) of these markers successfully identified one or more MTP BACs (Table 5.7).

Marker name	Designed from BAC	Contig	Position	Hits BAC	Contig	Position
TA-ISBP-0001	3DL096_F19 (f)	14	1	3DL168_F05	14	2
				3DL096_F19	14	1
TA-ISBP-0003	3DL117_P08 (r)	26	17	3DL015_B06	342	3
				3DL151_H08	342	4
TA-ISBP-0007	3DL010_I17 (r)	73	1	3DL161_I16	76	1
TA-ISBP-0008	3DL161_I16 (f)	76	1	3DL161_I16	76	1
				3DL093_A08	76	2
TA-ISBP-0010	3DL057_K09 (r)	76	15	3DL165_L12	5390	2
				3DL165_H16	5390	1
TA-ISBP-0011	3DL056_L04 (r)	104	32	3DL056_L04	104	32
				3DL142_K08	104	31
TA-ISBP-0013	3DL047_N21 (r)	120	8	3DL047_N21	120	8
				3DL139_M13	120	7
TA-ISBP-0020	3DL153_A06 (f)	266	8	3DL136_H14	151	8
				3DL099_D04	151	9
TA-ISBP-0021	3DL147_K02 (f)	327	12	3DL128_K19	327	11
				3DL147_K02	327	12
TA-ISBP-0022	3DL164_O18 (r)	350	1	3DL062_C08	350	2
				3DL164_O18	350	1
TA-ISBP-0023	3DL083_A04 (f)	350	21	3DL083_A04	350	21
				3DL130_O19	350	20
TA-ISBP-0024	3DL104_I15 (f)	395	1	3DL073_K06	395	2
				3DL104_I15	395	1
TA-ISBP-0028	3DL097_O20 (r)	744	4	3DL097_O20	744	4
				3DL134_F05	744	3
TA-ISBP-0029	3DL006_J18 (r)	747	1	3DL006_J18	747	1
				3DL103_L23	747	2
TA-ISBP-0030	3DL137_C05 (r)	747	5	3DL023_A21	747	4
				3DL137_C05	747	5
TA-ISBP-0032	3DL162_G22 (r)	805	19	3DL162_G22	805	19
TA-ISBP-0033	3DL167_O05 (f)	1044	1	3DL167_O05	1044	1
TA-ISBP-0034	3DL010_D14 (f)	1072	1	3DL010_D14	1072	1
TA-ISBP-0037	3DL157_L19 (f)	1410	1	3DL048_G18	1410	2
				3DL157_L19	1410	1
TA-ISBP-0038	3DL007_I21 (f)	4070	3	3DL007_I21	4070	3
				3DL117_P14	4070	2
TA-ISBP-0039	3DL007_I21 (r)	4070	3	3DL007_I21	4070	3
				3DL117_P14	4070	2

Table 5.7: Contigs identified by screening the ISBP markers against the MTP BAC pool.
Each marker is shown alongside the BAC it was designed from, the position of the BES (f = forward, r = reverse) and the contig and position in the contig of the BAC. The last three columns show the result of the screening, the BAC(s) identified the contig containing the BAC and the position of the BAC in the contig.

Seventeen of the successful markers identified BACs from within the contig they were designed from. For example, TA-ISBP-0037 was designed from a BAC (3DL157_L19) in the first position of contig 1410. Screening this BAC against the MTP BAC pool identified itself and the BAC in the second position of the contig (3DL048_G18). Four markers (TA-ISBP-0003, TA-ISBP-0007, TA-ISBP-0010 and TA-ISBP-0020) identify BACs on different contigs indicating several places where contigs might be joined. For example, TA-ISBP-0003 is designed from a BAC at the end of contig 26 and this marker identified two BACs at the end of contig 342. Contig 76 is predicted to overlap with contig 73 by marker TA-ISBP-0007. In addition, the other end of contig 76 is predicted to overlap with contig 5390 by marker TA-ISBP-0010. This potentially means that contigs 73, 76 and 5390 can all be joined to form a long contig.

Contigs 76, 350 and 747 have markers identified from both ends but only one marker from contig 76 (TA-ISBP-0010) identifies a new contig. This means that markers designed from the ends of contigs 350 and 747 are probably designed from internal BES as they only identify BACs in the same contig or that there is no overlap with other contigs. Although BAC 3DL007_I21 has markers designed from both BES (TA-ISBP-0038 and TA-ISBP-0038), both markers identify BACs within the same contig indicating that there is a large overlap between these two BAC clones in contig 4070.

5.4 Discussion

Sequencing the wheat genome is an essential foundation to understand phenotypic variation in this important crop species and will provide a basis from which to develop of new varieties of wheat with increased resistance to environmental stresses such as drought and salinity and increased resistance to disease. These new varieties will help to address the future demand for food from a growing human population (Royal Society of London 2009). However, sequencing the genome of wheat represents one of the most challenging genome sequencing projects to date due to its huge size, high repeat content, and hexaploid nature (Smith and Flavell 1975; Bennett and Leitch 1995).

Recently, the technique of chromosome sorting has been applied to wheat aneuploid stocks which allows individual chromosomes or chromosome-arms in the genome to be separated by flow cytometry (Dolezel *et al.* 2007). In this way, each of the 21 chromosomes or 42 chromosome arms can be tackled as a separate genomic entity. With a highly-repetitive genome such as wheat, a physical map provides a foundation for genome sequencing as a minimum tiling path of BAC clones can be identified from such a map and then sequenced. A physical map has recently been constructed for wheat chromosome 3B (Paux *et al.* 2008) which was readily separated from the other chromosomes by flow cytometry. A similar approach is being used for the long arm of chromosome 3D. As part of the TriticeaeGenome project we have built a BAC library of over 55,000 clones from flow-sorted 3DL genomic DNA, fingerprinted the clones using HICF, labelled the fragments using SNaPshot technology (Marra *et al.* 1997; Luo *et al.* 2003) and used FPC to build contigs based on overlapping clone fingerprints (Soderlund *et al.* 1997). The final physical map consists of 1,000 contigs with a minimum tiling path (MTP) of 5,826 clones. The MTP clones were pooled using a three dimensional pooling strategy to facilitate subsequent screening. The wheat 3B physical map at the same stage consisted of 1,036 contigs and a minimum tiling path of 7,440 clones (Paux *et al.* 2008). Wheat chromosome 3B has an estimated size of 900 Mb and 3DL is approximately 450 Mb so the clones provide roughly the same potential coverage. The MTP clones were end-sequenced to provide markers for anchoring. In addition, the BES were compared to wheat unigenes and 0.86 % showed good homology to unigenes. This is similar to the previous estimate of 1.2 % from chromosome 3B BES (Paux *et al.* 2006).

In order to anchor physical map contigs I developed a novel approach of anchoring contigs to the syntenic region of Brachypodium, a recently sequenced grass genome closely related to wheat (International Brachypodium Initiative 2010). The long arm of Brachypodium chromosome 2 (~30 Mb) was identified as syntenic to wheat 3DL and further comparisons using deletion bin-mapped wheat ESTs confirmed this. The recently published *Ae. tauschii* genetic map (Luo *et al.* 2009) was also compared to Bd2L to assess synteny as *Ae. tauschii* is the donor of the D-genome in wheat. This provided a more detailed comparison and identified a 20 Mb region at the distal end of Bd2L as syntenic to wheat 3DL. Very few 3DL wheat ESTs aligned to genes in the 10 Mb proximal to the centromere and markers from the genetic map did not cover this region. Interestingly, this is the position of one of the syntenic breakpoints identified in the Brachypodium genome by comparison to rice, sorghum and wheat marked by a high density of retroelements and centromeric repeats (International Brachypodium Initiative 2010). In general, these comparisons indicated that the 40 to 60 Mb region of Bd2L shows a high level of synteny to the entire wheat 3DL chromosome arm and that the syntenic anchoring strategy should provide a reasonable guide to the relative positions of contigs. These syntenically anchored positions can then be validated and refined by additional anchoring methods using a genetic map and a deletion-bin map, leading to the creation of a physical map.

Markers designed from ESTs assigned to genetic loci on the *Ae. tauschii* map were screened against the MTP clones and 30 contigs were anchored. In most cases, single markers identified multiple BACs from the same contig indicated that the contigs built by FPC were robust, insofar as they reflect the underlying DNA sequence. Some evidence of gene duplication was observed where single markers anchored BACs on multiple contigs. This might also reflect contamination in the flow-sorted library where BACs have been made from other chromosomes in the wheat genome and assembled into small contigs. The flow-sorted DNA is estimated to be 86 % pure meaning 14 % of the DNA used to construct the BAC library is potentially from other wheat chromosomes. A recent study used three BAC libraries constructed from homoeologous wheat chromosomes to build physical map contigs and showed that the resulting contigs mostly contained BACs from a single library (Luo *et al.* 2010). However, contigs representing non-3DL DNA may have been built from contaminant clones in the 3DL BAC library and would provide additional primer-pair binding sites for markers.

A total of 619 markers were used to anchor physical map contigs to Brachypodium genes in the Bd2L syntenic region. The majority of these (75.1 %) were markers designed from non bin-mapped wheat ESTs aligned to Brachypodium genes. In total, the markers resulted in 320 anchored contigs. The BES were also used to anchor contigs to Brachypodium genes but this strategy proved to be less reliable as a large number of BES showed sequence similarity to four transposable element-related genes in Brachypodium. This is to be expected in a repeat-rich genome such as wheat and again indicates the necessity of accurate masking of repetitive elements when working with wheat genome sequence. Combining the marker and BES anchoring resulted in the physical map 'syntenic build' and with nearly one third of contigs anchored to Brachypodium. The anchored contigs were visualised in CMap and many different anchoring scenarios were observed. In addition, evidence of breaks in collinearity between wheat 3DL and Brachypodium 2L were evident by BACs in the same physical map contig anchoring to multiple places in the Brachypodium syntenic region as well as BACs in the same contig showing more localised rearrangements when anchored to Brachypodium 2L. Due to the number of contigs anchored to Brachypodium, the CMap visualisation tool did not provide the ideal platform for visualising the anchored contigs, however, nothing more suitable could be found. CMap is a web-based tool requiring a page refresh to update the display and it is designed to display comparisons between a small number of maps. More recent visualisation tools such as CMap3D (Duran *et al.* 2010) may provide the functionality required to fully explore the data presented here.

A comparison between contigs anchored genetically and syntenically showed that in general, contigs were anchored in the same relative order by each method. This indicates that the syntenic approach is useful, however this analysis was based on only 15 contigs so the resolution of the comparison was not very high. To more accurately assess the value of the syntenic build, contigs anchored in a small region of Bd2L were manually inspected. In general it was found that the results obtained from anchoring contigs using PCR-based markers and *in silico* anchoring using BES agreed.

Markers designed directly from Brachypodium genes in the test region were screened against the MTP BACs to test whether these markers could be used on a larger scale to increase the marker density over the Brachypodium syntenic region. It was found that although a low percentage of the markers designed from Brachypodium genes were useful for anchoring contigs (~10 %), the majority of the contigs anchored by these markers showed good agreement with contigs anchored by previous methods. The low success rate of markers designed from Brachypodium genes reflects the evolutionary divergence of wheat and Brachypodium as although most genes are shared between the pooid grasses, only genes that are sufficiently conserved in sequence would be useful in this approach.

ISBP markers (Paux *et al.* 2010) were identified from contigs anchored in the test region in order to test the robustness of the contigs and to identify new contigs that may overlap with those already anchored. The original strategy was to design four markers from each contig, two from BES on each terminal BAC but the density of identified ISBP markers was not high enough to achieve this. Although ISBP markers were obtained from more than half of the MTP clones enabling the potential anchoring of more than 87 % of contigs, these markers are from central as well as terminal BACs so the probability of obtaining two markers from each terminal BAC is low. ISBP markers are estimated to occur in wheat at a frequency of one every 3.8 kb (Paux *et al.* 2010) so if the average length of a BES is 500 bp, one would expect to find one ISBP marker in every eight BES. This analysis identified an average of one high confidence ISBP marker in every 10 BES indicating that all possible ISBP markers are being identified but that shorter contigs are less likely to contain a useful ISBP.

ISBP markers are identified by a Perl script that uses junctions between repeat elements to design primer-pairs and allocates a confidence level to the makers depending on how accurately the repeat elements are identified. I did not include low confidence markers when ISBP markers were designed from the MTP BES as the isbpfinder documentation indicates that these markers may not be particularly accurate when used experimentally. However, it would be useful to experimentally test a number of low confidence markers to determine whether they could be used to screen the MTP BAC pools as this would provide greater marker coverage. ISBP markers that successfully identified MTP BACs most often identified BACs from the contigs from which they were designed, confirming that the majority of contigs tested have been robustly constructed by FPC. Four markers also identified overlaps between contigs and interestingly, these markers all identify BACs at the end of the new contig rather than BACs in the middle. This is encouraging if one considers that the ends of contigs would overlap if such overlaps occurred. These results indicate that screening ISBP markers from all contigs against the MTP BAC pools would be a useful approach to identify joins between contigs and to reduce their number. One anomaly common to all the markers that identify new contigs is that none of them identify the BAC from which they are designed. One would expect that in addition to identifying new BACs, the BAC from which the marker was designed would also be identified. However, this was not the case in the four examples observed in this analysis.

Another encouraging result which supports the value of the syntenic build is that ISBP marker (TA-ISBP-0007) designed from contig 73, identifies a potential overlap with contig 76. Contigs 73 and 76 are anchored in close proximity in the syntenic build which indicates that these contigs are indeed found adjacent to each other in wheat 3DL (Figure 5.11).

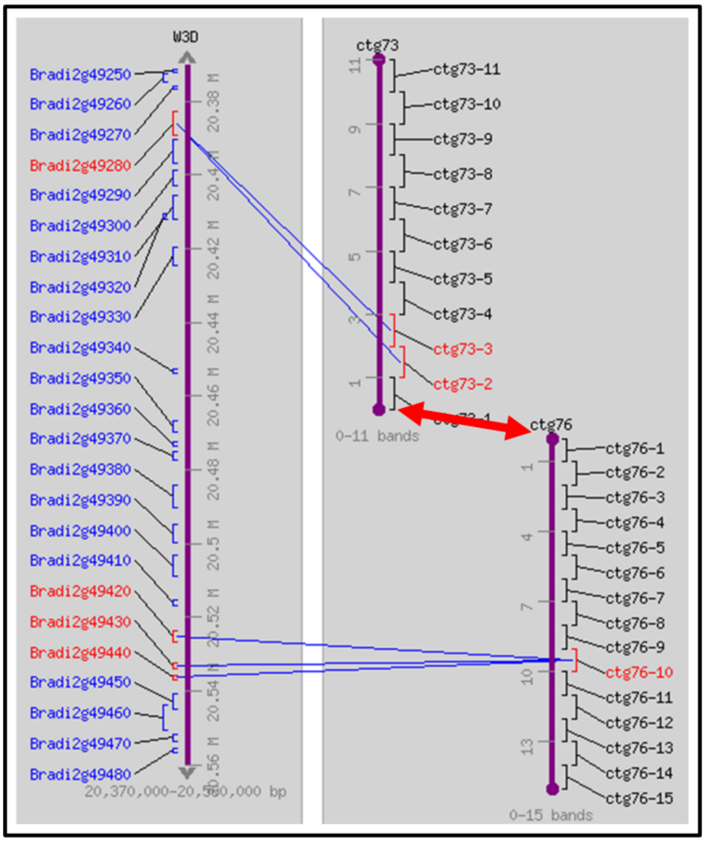


Figure 5.11: Agreement between syntenic anchoring and anchoring using ISBP markers.

Contigs 73 and 76 are anchored in close proximity in the syntenic build (indicated by blue lines) and also predicted to be linked by an ISBP designed from a BES on the first BAC of contig 73 which identifies the first BAC on contig 76 from the MTP BAC pool (indicated by a red arrow).

In conclusion, the wheat 3DL physical map is progressing well based on my strategy of incorporating syntenic anchoring that provides a long-range initial structure for subsequent steps. Multiple marker types have been designed and tested and the syntenic approach using Brachypodium appears to be useful although cannot fully be assessed until additional resources are available to anchor contigs. A deletion-bin map of wheat is being constructed using gamma-irradiated Chinese Spring seeds and 22 F_2 lines have been identified as containing deletions in chromosome 3DL (A. Yang, pers. comm.). Work is ongoing to more accurately define the extent of these deletions so that the markers identified in this study can be used to allocate physical map contigs into deletion bins. In addition, Both COS and ISBP markers have been designed from contigs which will be genetically mapped using a mapping population of 200 F_2 lines derived from Chinese Spring and Renan parental lines (N. McKenzie, pers. comm.). These approaches will complement the syntenic build presented here and allow the 1,000 contigs in the wheat 3DL physical map to be anchored. Once this is achieved, the foundation will be in place for sequencing another portion of the huge genome of wheat.

6 Sequencing the transcriptome of *Triticum aestivum*

6.1 Introduction

The complete set of RNA molecules produced from the genome of an organism is called the transcriptome (Velculescu *et al.* 1997) and the study of these molecules is called transcriptomics. Understanding the transcriptome of an organism is essential to elucidate the functional elements of its genome and to identify the molecular components within cells and tissues. Transcriptomics can be used to determine the structure of genes, for example, the 3' and 5' ends and the position of introns and exons, as well as the alternative splicing processes that occur after transcription giving rise to different mature transcripts from the same gene. It can also be used to measure the relative amounts of each transcript to characterise the expression level of genes. Different genes are expressed at different stages of cell development so transcriptomics can be used to identify which genes are expressed in different cell types and how gene expression changes as a cell develops. In addition, the expression of genes can be monitored in response to environmental factors such as biotic or abiotic stress.

As early as the 1970s, techniques were developed to monitor the expression of individual genes in different tissues or developmental stages by detecting RNA (Alwine *et al.* 1977). These techniques use electrophoresis to separate RNA molecules by size which are then immobilised on a membrane. A labelled probe complimentary to the target sequence is used to detect the RNA by hybridisation. In the mid-1990's DNA microarray technology emerged which allowed the expression levels of thousands of genes to be measured simultaneously (Schena *et al.* 1995). DNA microarrays consist of an ordered arrangement of DNA spots (the probes) covalently attached to a solid surface by photolithography or electrochemistry, where each spot represents a gene. The array is interrogated using cDNA from mRNA (the target) that has been purified from cells of a particular type and labelled with a fluorescent dye. The amount of cDNA hybridised to each site on the array is measured by the levels of fluorescence when the array is excited by a laser. The intensity of fluorescence is captured and analysis software is used to determine the gene expression levels. Hybridisation-based approaches to measure gene expression are limited by high background noise levels due to cross-hybridisation (Okoniewski and Miller 2006) and limited dynamic range of detection due to background noise and signal saturation.

Sequence-based approaches to transcriptomics have been used for many years to produce libraries of expressed sequence tags (ESTs) from cDNA clones. RNA is extracted from a range of tissue samples representing cells in different stages of development and cells that have been subjected to various biotic and abiotic stresses. Messenger-RNA is purified from the total RNA and reverse-transcribed into cDNA. Standard Sanger sequencing techniques are used to sequence a portion of each cDNA clone resulting in an EST. Collections of ESTs representing genes from a wide range of organisms are available in public databases (Boguski *et al.* 1993) and can be aligned to newly sequenced genomes to provide a fast and accurate method of gene annotation. Although an EST sequence can provide evidence of gene expression, Sanger sequencing of full-length cDNA clones is relatively low-throughput so quantative measurement of expression levels using this technology is not feasible. Approaches such as serial analysis of gene expression or SAGE (Velculescu *et al.* 1997) and massively parallel signature sequencing (Brenner *et al.* 2000) were developed to address this limitation. SAGE involves capturing a 9 to 14 nucleotide sequence tag from the 3' end of each cDNA molecule. These tags are concatenated and sequenced to identify the individual tags. Because one tag is generated from each cDNA molecule, gene expression levels can be determined from the frequency of occurrence of unique tags. Massively parallel signature sequencing (MPSS) is a procedure that produces longer sequence tags and with higher throughput than SAGE. In MPSS, cDNA molecules are attached to microbeads using a method that ensures each cDNA in a sample is likely to be represented and attached to a single microbead. The microbeads are aligned in a flow cell and a 17 to 20 nucleotide sequence tag is generated from the 3' end of each cDNA in parallel by successive rounds of cleavage and interrogation using fluorescently labelled probes. Multiple copies of the same transcript will result in identical reads from which expression levels can be deduced.

RNA extracted from tissue samples (total RNA) consists of all RNA molecules found in the cells. In eukaryotic cells, only 1 to 5 % of total RNA represents messenger RNA and within this portion, the quantity of different transcripts ranges from thousands to tens of thousands. A small number of highly expressed genes account for a large proportion of the mRNA and rare transcripts will exist at much lower frequencies. If a quantative measurement of gene expression is not required, a more accurate representation of the transcribed portion of the genome can be obtained if the prevalence of highly abundant transcripts is reduced, thus increasing the representation by rarer transcripts. This process is called normalisation and is commonly applied to cDNA libraries before sequencing. The method of cDNA normalisation using a double-stranded nuclease (DSN) is explained in Chapter 2.

The advent of next-generation sequencing (NGS) technologies (Metzker 2010) has revolutionised the field of transcriptomics as cDNA can be directly sequenced in a high-throughput manner and at reduced cost compared to traditional sequencing technologies. The resulting sequence reads can be aligned to a reference genome or assembled *de novo* to determine both gene structure and gene expression levels. The application of NGS technologies to transcriptomics is called RNA-Seq. RNA-Seq has many advantages over previous hybridisation-based approaches. Unlike these approaches, RNA-Seq doesn't rely on the availability of a genome sequence as RNA can be extracted from any organism, sequenced and assembled *de novo* to produce contigs representing transcribed regions of the genome. Where a reference genome is available, sequence reads can be aligned to the genome sequence to reveal gene structure to single base resolution and identify junctions between exons. Furthermore, these alignments can be used to identify SNPs between different varieties for use as molecular markers. RNA-Seq doesn't exhibit the background noise levels that are inherent in hybridisation-based approaches as reads can be mapped uniquely to the genome and in addition, there is no limit to the number of transcripts that can be sequenced, hence the dynamic range of the RNA-Seq is much wider. In *S. cerevisiae*, RNA-Seq has been used to measure the expression levels of genes and identified genes that are expressed at levels more than 8,000 times higher than others, something that would have been impossible to measure using microarrays (Nagalakshmi *et al.* 2008).

RNA-Seq has been used to sequence and characterise the transcriptomes of many species. In 2007, Weber and colleagues sequenced mRNA from *Arabidopsis* seedlings (Weber *et al.* 2007). The resulting ESTs were aligned to the genome and provided evidence of expression for over 17,000 gene loci, some of which were predicted genes with no supporting experimental evidence. 454 sequencing was used to generate more than 600,000 ESTs from two cDNA libraries constructed from the Glanville fritillary butterfly (Vera *et al.* 2008). These ESTs were assembled into contigs and approximately 9,000 genes were identified. Another study investigated the transcriptome of *Artemisia annua*, a plant used to produce the anti-malaria drug artemisinin from its glandular trichomes (Wang *et al.* 2009a). Assembling these reads generated over 42,000 unigenes, more than half of which could be assigned function by comparison with the NCBI protein database. For species with large and complex genomes, RNA-Seq provides a fast and relatively inexpensive method to sample the transcribed portion of the genome. One such study has been performed on Lodgepole pine, a tree species in the genus *Pinus* (Parchman *et al.* 2010). Genomic data from a related pine in combination with *de novo* assembly was used to generate nearly 64,000 contigs from cDNA, 17,000 of which were identified as genes. In addition, a large number of retrotransposons sequences were identified in the contigs indicating that these elements are transcriptionally active and that there may be regions of retroelements inserted into transcribed regions of genes.

It is perhaps counter-intuitive that the problem of transcriptome assembly is more complex than genome assembly as the transcribed portion of the genome is significantly smaller than the complete genome. The goal in genome assembly is to assemble chromosome-scale contiguous regions of sequence representing the entire genome. In contrast, but the transcriptome consists of thousands of short stretches of sequence representing transcribed genes meaning that the contiguity of an assembled transcriptome is low. In addition, transcripts are present in many different copy numbers and include a high proportion of variant transcripts that are the products of alternative splicing from the same gene. As with genomic reads, assembling the longer reads produced by 454 sequencing is computationally more feasible than assembling short Illumina reads. Illumina cDNA reads are generally aligned to a reference genome to identify transcribed regions or to identify SNPs. However, the development of more powerful assembly algorithms and the availability of high-performance compute clusters means that assembling transcriptomes from short reads is realistic. The ABySS assembler has been used to assemble Illumina reads obtained from human tumour tissue into contigs (Birol *et al.* 2009). Nearly 67,000 contigs of 100 bp or longer were generated representing over 30 Mb of transcriptome sequence. Another potentially useful assembler specifically designed for transcriptome data is Oasis, an extension to the Velvet assembler (http://github.com/dzerbino/oases), although results from using this assembler are yet to be published.

Sequencing the transcribed portion of the bread wheat genome is a tractable approach to apply to this large, complex genome. In the past, traditional Sanger sequencing has been used to sequence wheat cDNA clones (Ogihara *et al.* 2004) resulting in an extensive EST collection containing 1,071,199 sequences (dbEST release 081310). In addition, more than 8,000 full-length cDNA sequences are available (Mochida *et al.* 2009) and a handful of sequenced BACs in public databases. Transcript assembles (TAs) have been built from publicly available wheat sequences using automated pipelines resulting in putative sets of wheat unigenes. Several TA datasets exist and are outlined in Table 6.1.

Because these transcript assemblies are constructed from ESTs they are unlikely to provide complete end-to-end coverage of genes. In addition, the datasets vary in size from less than 40 Mb to nearly 200 Mb indicating that the smaller datasets are probably incomplete in terms of representation of all wheat genes and the larger datasets are highly redundant. The wheat genome is complicated by the presence of homoeologous genes; each gene has the potential to be present as three homoeologues, each exhibiting a high sequence similarity to the other two homoeologues. These homoeologues are frequently expressed simultaneously (Mochida *et al.* 2004). If high stringency criteria are used when assembling these ESTs, homoeologous transcripts will be assembled separately. Assembling at lower stringency would cluster homoeologous ESTs into single transcripts.

Name and URL	Date of release	Total sequences	Total length (Mb)
TIGR transcript assembly (http://plantta.jcvi.org)	2007, release 2	319,949	198
DFCI gene index (http://compbio.dfci.harvard.edu/tgi)	2010, release 12	221,925	154
NCBI unigenes (http://www.ncbi.nlm.nih.gov/unigene)	2010, Build 56	40,870	36

Table 6.1: Existing wheat transcript assemblies built from EST data.

The recent sequencing of 18.2 Mb of wheat chromosome 3B contigs identified genes ranging from 309 to 15.8 kb in length and the average coding sequence (CDS) of these genes was 1,382 bp (Choulet *et al.* 2010). This is close to the average size of 1,143 bp predicted from 6,137 full-length wheat cDNAs (Mochida *et al.* 2009). These estimates also agree with the average coding sequence length of genes in the Brachypodium genome of 1,392 bp (Supp. Table 6, International Brachypodium Initiative 2010). Choulet and colleagues estimate that between 36,000 and 50,000 genes are present on the wheat group B chromosomes (Choulet *et al.* 2010) meaning that the group B transcriptome is between 49.7 and 69.1 Mb in size. The additional two homoeologous chromosome sets would contain a similar number of genes so the complete hexaploid wheat transcriptome is estimated to be between 149.1 and 207.3 Mb. The NCBI unigene dataset is 36 Mb indicating that this dataset may represent a large proportion of wheat transcripts where homoeologous sequences are represented as a single sequence. As well as redundant sequence, the TIGR and DFCI datasets are more likely to contain separate homoeologous transcript sequences as these datasets are larger.

A limitation of transcript sequence is that it contains no inherent information on gene structure as cDNA is made from mRNA after excision of introns. In wheat, this limitation can be partially overcome by using homologous genes from closely related species such as rice to estimate gene models. Such a method is implemented in the Wheat Estimated Transcript Server (Mitchell *et al.* 2007) which identifies wheat ESTs similar in sequence to a selected rice gene. The ESTs are assembled and aligned to the rice gene to give an estimation of the intron-exon structure of the wheat gene based on the rice homologue. Gene-based markers that span predicted introns can then be designed from these sequences.

Applying high-throughput sequencing methods to the wheat transcriptome is an exciting new prospect and in theory can be used to obtain a more accurate picture of the transcripts present in hexaploid wheat. Provided a broad range of transcripts can be sampled, this should enable long transcript sequences to be constructed as well as identifying new transcripts. In addition, the availability of the Brachypodium genome sequence (International Brachypodium Initiative

2010) means that gene models from Brachypodium can be used to more accurately define gene models in wheat. Brachypodium is estimated to have diverged from the wheat lineage between 8 and 15 million years after the divergence of the temperate grasses from the rice lineage (International Brachypodium Initiative 2010), so gene models should be more conserved between wheat and Brachypodium than between wheat and rice. More than 16,300 genes in the Brachypodium genome are predicted to have highly related orthologues in wheat and barley, meaning that a significant proportion of the genes in the wheat genome could be defined by alignment with Brachypodium genes (International Brachypodium Initiative 2010).

In this chapter I describe the analysis of sequence data generated by sequencing wheat cDNA using both the pyrosequencing (454) and sequencing-by-synthesis (Illumina) methods and the subsequent analysis of this data using existing wheat transcript sequence and Brachypodium sequence. Sequence reads generated from the Illumina platform were aligned to existing wheat transcriptome sequence to determine what proportion of the reads represented existing wheat sequence and whether any new sequence was generated. Both normalised and unnormalised libraries were prepared and sequenced to determine whether cDNA normalisation could be used to identify a larger number of transcripts. Sequence reads generated by 454 sequencing were assembled into contigs and compared to existing wheat transcriptome sequence. Contigs representing new sequence were aligned to Brachypodium gene models to identify putative gene structures and contigs representing existing wheat transcript sequence were used to extend these transcript assemblies before alignment to Brachypodium gene models. Illumina sequencing was performed in collaboration with The Genome Analysis Centre (TGAC) in Norwich and 454 sequencing was performed in collaboration with Neil Hall’s group at the University of Liverpool.

6.2 Methods

6.2.1 Preparation of cDNA libraries for sequencing

Four samples from Chinese Spring batch 42 were used to prepare cDNA libraries and were designed to maximise the transcript diversity sampled for sequencing. The samples comprised;

- Roots, young leaves, young flowers and immature seeds.
- 6 and 11 day drought induced leaves.
- 3 and 5 days-after-anthesis senescent flag leaf.
- 24 hr circadian time course, young leaves collected at 6 hourly intervals.

Both Illumina and 454 sequencing technologies were used to take advantage of their relative merits. Illumina sequence is relatively cheap to generate and deep coverage can be achieved but at the expense of shorter read lengths. 454 sequencing is more expensive but generates longer read lengths. In addition, using two sequencing methods reduces potential bias in generated sequence compared to using a single method. For Illumina sequencing, normalised and unnormalised cDNA libraries were prepared from each individual sample using the DSN normalisation method (Zhulidov *et al.* 2004). Both normalised and unnormalised cDNA samples were sequenced using the Illumina paired-end protocol to give 80 bp paired end reads with an insert size of 250 bp.

For 454 sequencing, two superpools were produced consisting of cDNA from all three samples. One superpool was sent to Evrogen (www.evrogen.com) for normalisation. The normalised and unnormalised superpools were sequenced separately and the resulting reads were combined and assembled using the MIRA assembler (http://www.chevreux.org/projects_mira.html).

6.2.2 Analysis of Illumina sequence files

The Illumina GA analysis pipeline produces Illumina-FASTQ sequence files (Cock *et al.* 2010). These were first processed using the fastx_quality_stats utility from the FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit) to produce statistics on the reads in the file. Assessing the quality of reads is essential as the quality of base calls tends to deteriorate towards the end of the read due to the chemistry of the Illumina sequencing method. An R script (R Development Core Team 2010) was used to plot the distribution of the read quality profiles for reads in each file. The Illumina-FASTQ file is then converted to Sanger-FASTQ format using the MAQ 'ill2sanger' function (Li *et al.* 2008a).

The FASTA reference sequence files (TIGR transcript assembly or NCBI unigenes) were converted to binary FASTA format using the MAQ 'fast2bfa' function and the standard MAQ pipeline was used to align the reads to the reference sequence. First the 'match' function was used to align the reads in each file to the reference sequences, then the 'assemble' function was used to build the mapping assembly. The Tablet viewer (Milne *et al.* 2010) was used to visualise these alignments by converting the alignment file to text format using the MAQ 'mapview' function and the mapping assembly file to FASTQ using the 'cns2fq' function.

Aligning the reads to a reference sequence using BWA (Li and Durbin 2009) first requires the reference sequences to be indexed using the BWA 'index' function, then the 'aln' function is used to get the suffix array coordinates of the input reads. The alignments are generated using the 'sampe' function. These alignments are created in Sequence Alignment/MAP (SAM) format. The functions 'view', 'sort' and 'index' in the SAMTools utility (Li *et al.* 2009a) are used to convert the SAM file to Binary Alignment/Map (BAM) format and to sort and index it. The shell script and additional files used to process an Illumina run are included as Supplementary information 20. Statistics are generated from the BAM file using the SAMTools 'flagstat' function and a custom Perl script was written utilising the Bio::DB::Sam library to determine the average read depth exhibited by each unigene and the percentage of each unigene covered by reads.

6.2.3 Analysis of contigs assembled from 454 sequence reads

The redundancy of assembled contigs was tested using CD-HIT-EST (Li and Godzik 2006) to identify contigs that were represented identically by a longer sequence. This stage is important if contigs from multiple runs are analysed together as there is no need to include contigs in the analysis that are represented multiple times.

The analysis of the 454 contigs was performed in two stages. The first stage is depicted in Figure 6.1. Contigs were first compared to NCBI unigenes using BLASTn to determine what proportion show high sequence similarity to unigenes. An e-value cut-off of 1e-10 was used. The contigs that show no sequence similarity to unigenes were then compared to Brachypodium proteins using BLASTx to identify putative Brachypodium homologues. These contigs may represent genes that are present in wheat and have homologues in Brachypodium but are not represented in the wheat unigene set. Contigs were aligned to their putative Brachypodium homologues using the Spidey alignment tool which is designed to align transcript sequence to genome sequence (Wheelan *et al.* 2001). From these alignments, the percentage of coding sequence covered by transcript sequence for each Brachypodium gene is calculated. Any contigs that don't show sequence similarity on the protein level to Brachypodium are compared to the RefSeq (release 41) database of plant proteins (Pruitt *et al.* 2007). These contigs may represent wheat genes that are not present in Brachypodium but have been characterised in other plants, a proportion likely to be small.

Contigs that show good sequence similarity to NCBI unigenes were used to extend unigenes in the following way. Each unigene was processed in turn and contigs showing good sequence similarity (e-value < $1e^{-50}$) to that unigene were extracted. Each unigene with matching contigs were assembled using CAP3 (Huang and Madan 1999) and saved as an extended unigene if the resulting unigene is 100 bp or longer than the original unigene. In this way, a new set of unigenes (the extended unigene set) was created containing the extended unigene if one had been created or the original unigene if no extension was possible.

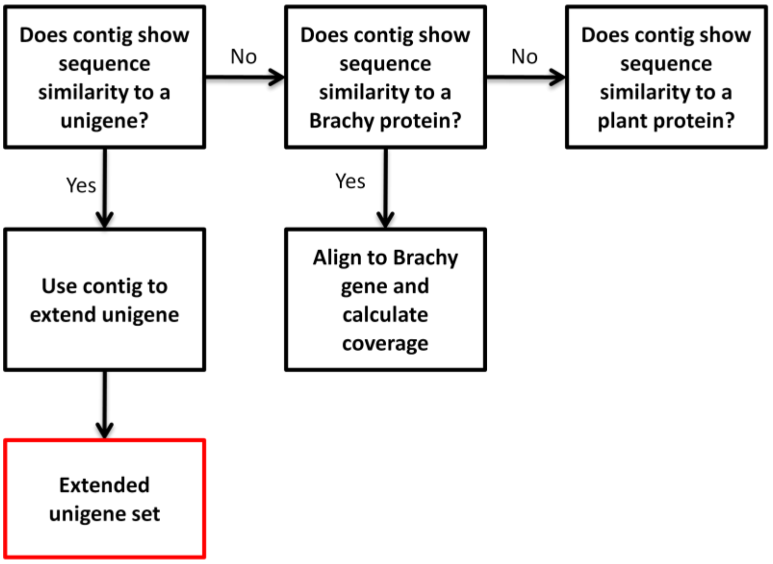

Does contig show sequence similarity to a unigene?
No
Does contig show sequence similarity to a Brachy protein?
No
Does contig show sequence similarity to a plant protein?
Yes
Yes
Use contig to extend unigene
Align to Brachy gene and calculate coverage
Extended unigene set

Figure 6.1: The workflow designed to analyse 454 contigs.

Contigs showing similarity to unigenes are used to extend the unigenes to create a new unigene set (red box) and contigs showing no similarity to unigenes are compared to Brachypodium and other plant proteins.

The second stage is depicted in Figure 6.2. The extended unigene set was compared to Brachypodium proteins using BLASTx to identify putative homologues. Any unigene showing 80 % or more sequence similarity to a Brachypodium gene and covering 80 % or more of its length was aligned to its putative Brachypodium homologue using Spidey and the coverage of each Brachypodium gene was calculated to generate a mean coverage. To provide a baseline with which to compare the extended unigenes set, the original unigenes were analysed in the same way, first compared to Brachypodium proteins using BLASTx to determine putative homologues, then aligning each unigene to the homologous Brachypodium gene and calculating the mean coverage. By comparing the results from the original and extended unigene alignments one can determine whether extending the unigenes has increased the number of gene models that can be characterised by alignment to Brachypodium.

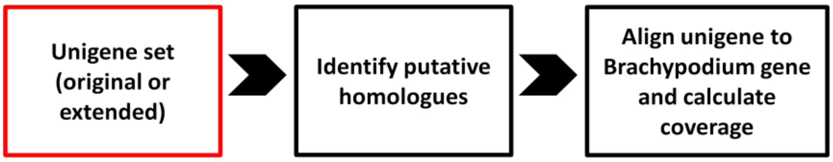

Unigene set (original or extended)
Identify putative homologues
Align unigene to Brachypodium gene and calculate coverage

Figure 6.2: Comparison of the original and extended unigenes by alignment to Brachypodium.

First putative homologues are identified, then from these, the best homologues are aligned and the coverage of the Brachypodium gene calculated.

6.3 Results

6.3.1 Analysis of Illumina reads

Approximately 532 million 80 bp reads were generated from wheat cDNA containing 42.6 Gb of sequence; representing a theoretical 200 x coverage of the wheat transcriptome.

6.3.1.1 Assessment of read quality

An Illumina flow cell provides 8 available lanes, 7 of these are generally loaded with samples and 1 is used as a control lane. The control lane is loaded with phiX, a circular bacteriophage genome consisting of 5,386 bases which is used by the base-calling software provided with the sequencing machine. In addition, the quality of reads from the control lane can be used to assess the quality of the run. In a successful run, the quality of the reads from the 7 sample lanes should be of similar quality as reads from the control lane. The accuracy of an individual base-call is expressed as a phred quality score which is a logarithmic scale ranging from 4 to around 60 with higher values corresponding to higher quality (Ewing and Green 1998). The read profile from the phiX control lane in the unnormalised samples is shown in Figure 6.3, Panel B. It shows an initial phred score of approximately 37, falling to around 30 by position 80 in the read. This indicates that the base-calling accuracy is more than 99.9 % until very late on in the read when it is still above 99 % accurate.

The read quality profile for a lane of unnormalised cDNA is shown in Figure 6.3, Panel A. It shows a quality score at the first base position of approximately 36 indicating the accuracy of the base call is more than 99.9 %. The quality score smoothly decreases as the length of the read increases but remains above 29. The read quality profile for the unnormalised cDNA is very similar to that of the control lane (Panel B) indicating these sequence reads are of high quality.

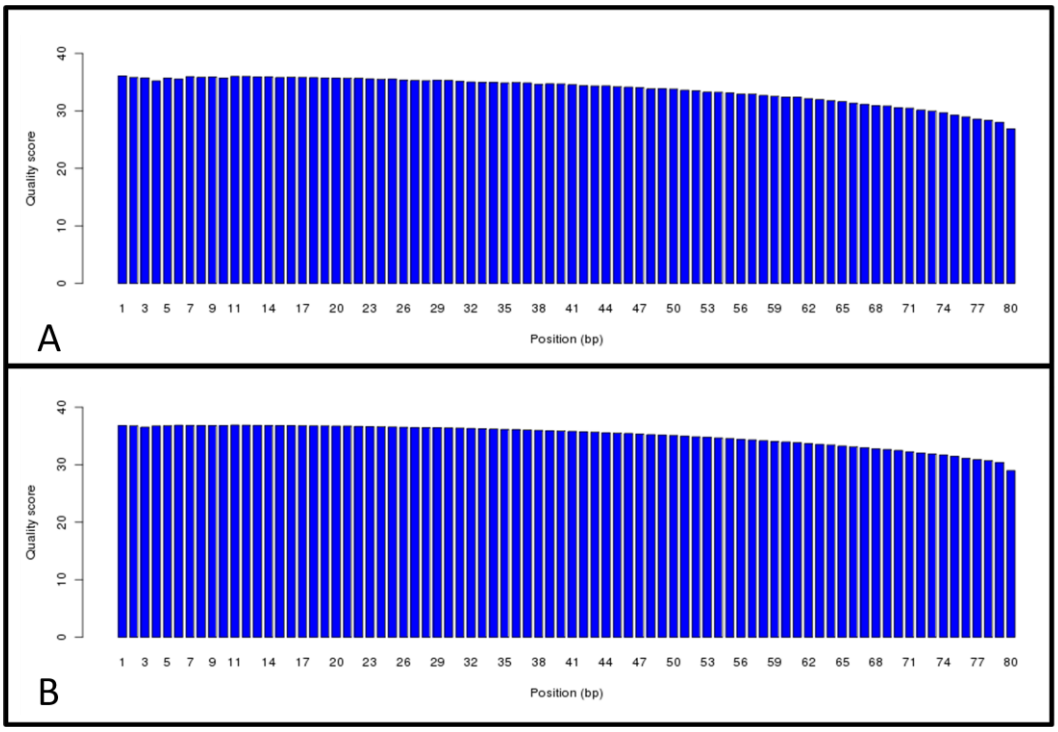


Figure 6.3: Read quality profile from an unnormalised cDNA sample.

The cDNA sample is shown in Panel A and the corresponding phiX control lane is shown in Panel B. The quality of the base call in each position in the read is a mean of the quality scores for the bases at that position in all the reads and is presented as a phred quality score. Both profiles show read quality gradually decreasing as the length of the read increases. The read qualities in the control lane are generally higher than those in the sample lane.

The read quality profile from the normalised cDNA sample alongside the quality profile of the corresponding phiX control lane is shown in Figure 6.4. The control lane indicates that the sequencing run has been successful, however the quality profile from the sample lane is strikingly different and shows a much steeper decline in the quality of base-calling as the length of the read progresses. In addition, the profile from the sample lane is irregular, unlike the smooth profile obtained from the unnormalised sample.

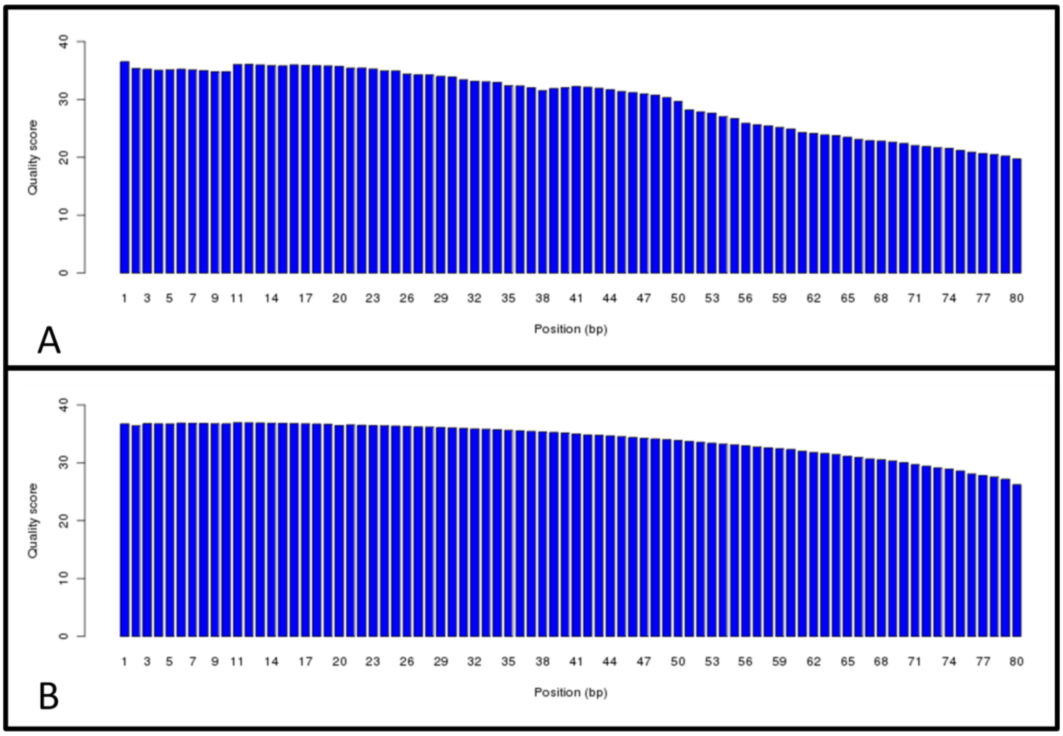


Figure 6.4: Read quality profile from a normalised cDNA sample.

The cDNA sample is shown in Panel A and the corresponding phiX control lane is shown in Panel B. The quality of the base call in each position in the read is a mean of the quality scores for the bases at that position in all the reads and is presented as a phred quality score. The control lane shows a profile similar to that observed for the unnormalised sample but the sample lane shows are steeper decline in quality and also an irregular pattern of decline.

The difference in the read quality profiles shown in Figure 6.3 and Figure 6.4 was observed between all the normalised and unnormalised samples and indicates a problem with the normalised samples. This could be due to a problem in the normalisation procedure. Upon further inspection it was found that the reads from normalised samples also showed a pattern of biased base composition. Figure 6.5 shows the number of thymine residues at each base position from reads taken from both normalised and unnormalised samples. Thymine residues are distributed relatively evenly along the reads in the unnormalised sample (Panel A) but the normalised sample shows very few thymine residues in the first 25 bases of the read (Panel B).

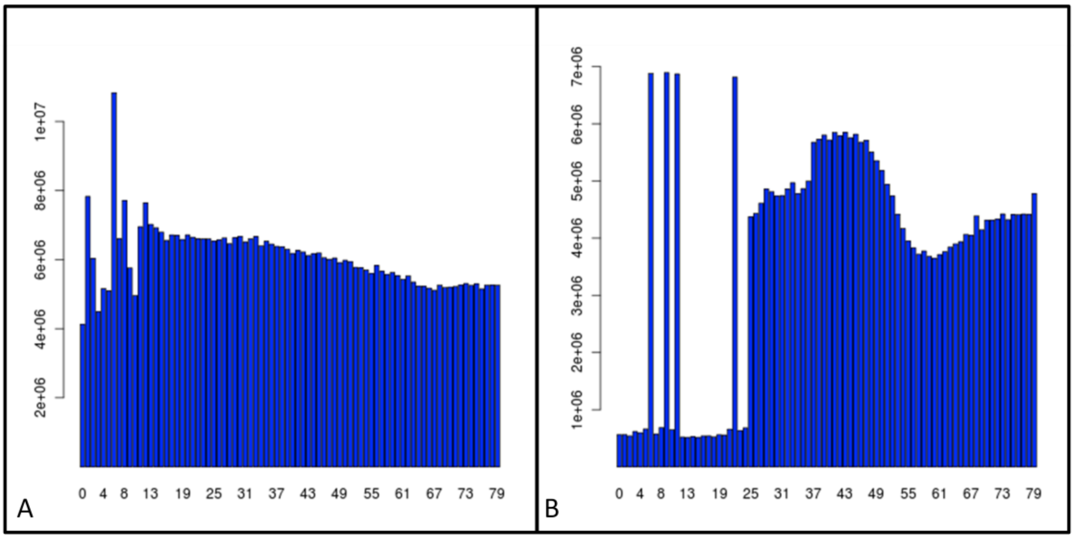


Figure 6.5: Frequency of thymine residues in Illumina reads.
Unnormalised cDNA (Panel A) shows a relatively even distribution and normalised cDNA (Panel B) shows very few thymine residues in the first 25 nucleotides of the reads.

6.3.1.2 Alignment of Illumina reads to TIGR wheat transcript assembles using MAQ

Reads from each sample were mapped to TIGR wheat transcript assemblies. From unnormalised samples, more than 23 million reads were obtained per lane on average. Normalised samples gave over 30 million reads per lane (Table 6.2).

cDNA sample	Average number of reads obtained per lane	Average number of reads mapped	Reads mapped (%)	TAs with reads mapped (%)
Unnormalised	23,228,673	20,819,543	89.6	65.1
Normalised	30,046,978	15,555,617	51.8	61.7

Table 6.2: Analysis of Illumina reads by alignment to TIGR transcript assemblies using MAQ.
Number of reads obtained from the normalised and unnormalised samples and analysis of the alignment of these reads to TIGR wheat transcript assemblies.

When the reads were mapped to the TIGR transcript assemblies a much larger difference was observed between the samples. Nearly 90 % of reads from the unnormalised samples map to the reference sequences compared to 51.8 % of reads from the normalised samples. The percentage of reference sequences with 1 or more aligning reads is between 61 and 65 % with reads from unnormalised samples hitting more reference sequences. This is contrary to the view that a normalised cDNA sample should represent a broader range of genes than an unnormalised sample. These results show that the normalisation procedure is working to some extent as reads from the normalised libraries hit a similar number of reference sequences compared to the reads from the unnormalised samples, yet fewer reads are involved. The normalisation procedure appears to be removing a large proportion of the multiple cDNA copies from highly expressed genes, however there is not a general increase in unigenes represented as a result of this. These results indicate that little has been gained by normalising the cDNA before sequencing, assuming that the normalisation procedure is working properly.

6.3.1.3 Alignment of Illumina reads to TIGR wheat transcript assembles using BWA

BWA was used to align the Illumina reads as it is a faster tool and has the additional benefit of storing the alignments it generates in the SAM/BAM format. Alignments stored in this format can be interrogated programmatically and downstream analysis is easier. For both samples, fewer reads mapped to the transcript assemblies using the BWA alignment tool compared to mapping the reads using MAQ (Table 6.3). Despite this, a similar disparity was observed in the number of reads mapping from the normalised samples compared to the unnormalised samples. A much larger proportion of reads mapped from the unnormalised samples (72.2 % compared to 27.7 %).

cDNA sample	Reads mapped (%)	TAs with reads mapped (%)
Unnormalised	72.2	58.9
Normalised	27.7	57.1

Table 6.3: Aligning Illumina reads to TIGR transcript assemblies using BWA.

Many more reads are mapped from the unnormalised sample than the normalised samples but reads from both libraries hit a similar percentage of unigenes.

If substantially fewer reads align to the reference sequences one would expect fewer reference sequences to be hit by a read but this was not observed. The proportion of reference sequences with aligned reads using BWA (57 to 59 %) is only marginally less than the number of reference sequences with aligned reads when using MAQ (61 to 65 %). As observed from the MAQ alignments, there is no real difference between the number of reference sequences hit by reads from normalised samples compared to unnormalised samples. These results show that BWA aligns fewer reads but the same numbers of reference sequences have reads aligned to them. This indicates that BWA uses more stringent criteria for aligning reads than those used by MAQ. Once again, there is no evidence that the normalisation procedure is increasing the number of reference sequences represented by reads.

6.3.1.4 Alignment of Illumina reads to NCBI unigenes using BWA

BWA was used to align the Illumina reads to the NCBI wheat unigenes as the TIGR transcript assembly is very large and slow to process. The wheat unigene set is much smaller and potentially provides a more accurate representation of unique transcript sequence (Table 6.1). BWA aligned fewer reads to the unigenes than to the TIGR transcript assemblies as would be expected for a smaller reference set (Table 6.4). Once again, more reads are mapped from the unnormalised samples than the normalised samples (48.7 % compared to 20.1 %). However, in this case the reads from normalised samples hit marginally more reference sequences than the reads from unnormalised samples.

cDNA sample	Reads mapped (%)	Unigenes with reads mapped (%)
Unnormalised	48.7	65.3
Normalised	20.1	73.0

Table 6.4: Aligning Illumina reads to NCBI unigenes using BWA.

These alignments show a small difference between the number of unigenes hit by unnormalised reads compared to normalised reads.

In order to obtain a more accurate measure of the number of unigenes covered by reads, the alignments were analysed in more detail using a custom Perl script. Each unigene was inspected to determine two statistics; the average read depth of the unigene (the total number of aligned bases divided by the length of the unigene) and the percentage of the unigene covered by reads. A unigene was counted if it exhibited an average read depth of 5 or greater, an arbitrary cut-off to indicate a unigene with moderately good read depth. For example, a 500 nucleotide unigene with forty 80 bp reads aligning to it will have an average read depth of 6.4. These reads may be distributed over the unigene such that only 10 % of the unigene is covered (the reads stack up in one position providing high read depth at that position) or that 90 % of the unigene is covered (the reads are distributed more evenly over the unigene but at lower read depth). Calculating both statistics is an accurate measure of the quality of the coverage provided by reads to each unigene.

From this data I calculated the number of unigenes with reads covering 10 % or greater of their length and the percentage of mapped reads used to obtain this coverage. This process was repeated at 20 % coverage, 30 % coverage etc. up to 100 % coverage. Figure 6.6 presents these results for one normalised sample (blue lines) and one unnormalised sample (red lines). The solid lines represent the percentage of unigenes that are counted for each coverage cut-off and the dashed lines represent the percentage of mapped reads that provide this coverage. The x-axis measures the percentage of the unigene covered by reads and the y-axis measures the percentage of unigenes counted (solid lines) or percentage of mapped reads aligned to counted unigenes (dashed line). Considering only unigenes with an average read depth of 5 or greater, the data shows that 30 to 40 % of unigenes have reads covering 10 % or more of their length. For these alignments, more than 90 % of mapped reads are used. At the other extreme, less than 5 % of unigenes have reads covering 100 % of their length. These alignments involve between 5 and 20 % of mapped reads.

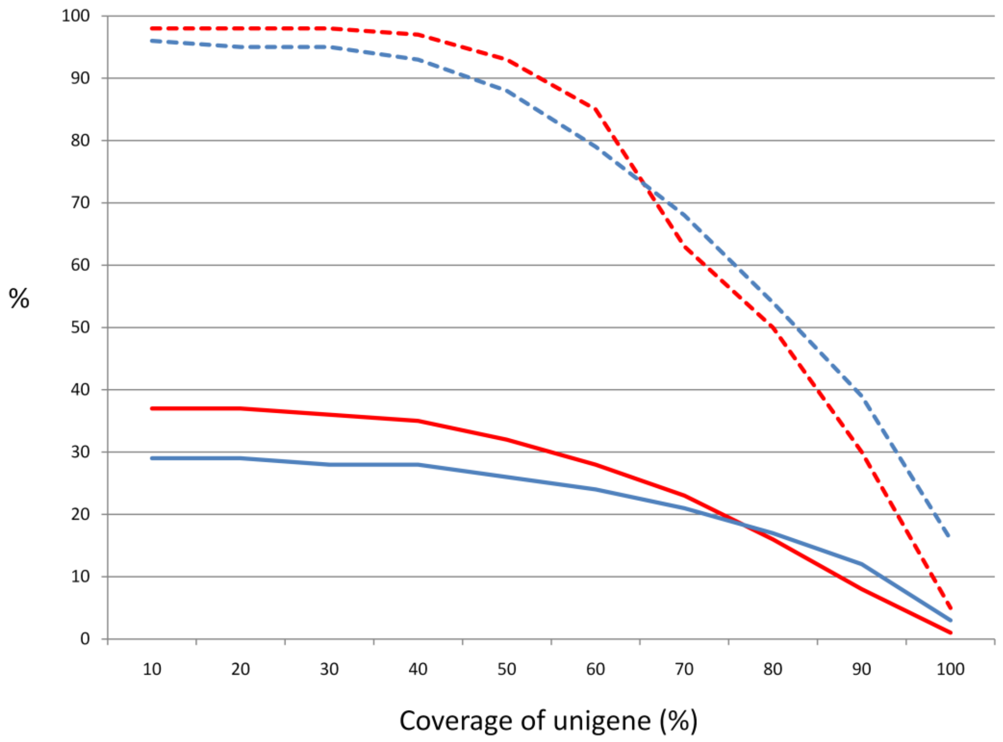


Figure 6.6: Alignment statistics for reads from Illumina samples.

The unnormalised sample is represented by a red line, the normalised sample is represented by a blue line. The method for generating these statistics is explained in the main text. The solid lines indicate the percentage of total unigenes counted at each coverage cut-off (x-axis) and the dotted lines indicate the percentage of mapped reads that generate this coverage.

There are two main points to note from Figure 6.6. First, there is no obvious difference in the number of unigenes counted (solid lines) between reads from different samples (normalised and unnormalised). This shows that reads from the normalised samples are not hitting more unigenes that reads from the unnormalised samples as was indicated by the previous, less detailed analysis (Table 6.4). The second point is that using this measure to determine the number of unigenes hit by reads we can see that a maximum of 30 to 40 % of unigenes have reads aligned to them. This indicates that the coverage being achieved by our sequencing runs is unexpectedly low. The cDNA samples are obtained from a wide range of tissues and environmental conditions so one would expect that a larger proportion of existing unigenes should be represented.

6.3.2 Analysis of 454 sequence data

6.3.2.1 Analysis of 454 raw reads

The details of the reads obtained from sequencing the normalised and unnormalised superpools are shown in Table 6.5. Nearly 900,000 reads were generated with a total length of 296 Mb.

This represents between 1.4 and 2.0 times coverage of the wheat transcriptome. More reads were obtained from the normalised library than the unnormalised library resulting in a larger amount of total sequence from the normalised library. The N50 length was also higher for the normalised reads indicating that the read length is longer.

Library type	Number of reads	Total length (Mb)	N50 (bp)
Normalised	450,943	168	464
Unnormalised	435,085	128	392

Table 6.5: Reads obtained by sequencing the normalised and unnormalised 454 libraries.

Profiles of the read lengths obtained from the two libraries are shown in Figure 6.7 and Figure 6.8. The read lengths obtained from the normalised library are weighted towards longer sequences with the majority of reads between 400 and 550 bp. This profile of read lengths is characteristic of the read length profiles obtained from 454 transcriptome sequencing in other species (Parchman *et al.* 2010). In contrast the reads from the unnormalised library (Figure 6.8) are more evenly distributed with a higher number of shorter reads. This indicates a potential problem in the preparation of this library for sequencing, possibly in the DNA nebulisation stage resulting in highly fragmented DNA.

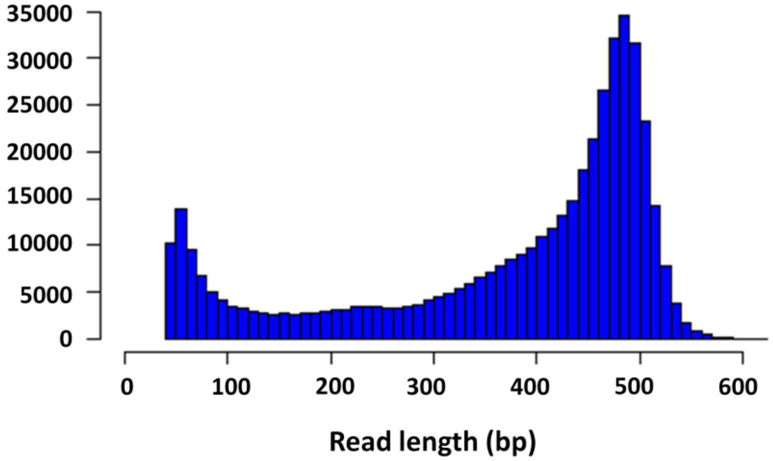


Figure 6.7: Profile of 454 read lengths obtained from the normalised library. Read length is shown on the x-axis and frequency is shown on the y-axis.

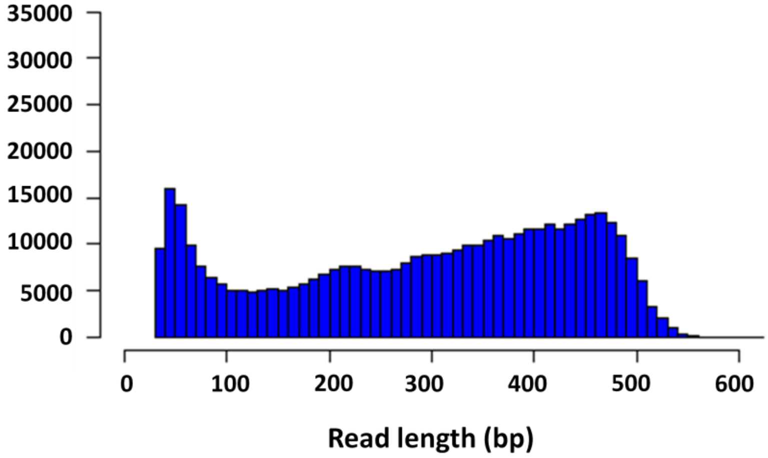


Figure 6.8: Profile of 454 read lengths obtained from the unnormalised library.
Read length is shown on the x-axis and frequency is shown on the y-axis.

The reads from both libraries were assembled together resulting in 68,488 sequence contigs with a total length of 41.3 Mb. The length of the contigs ranged from 40 bp to 4,349 bp with the mean length being 603 bp. The length profile of the assembled contigs is shown in Figure 6.9. The majority of contigs (67.3 %) are between 250 and 750 bp in length.

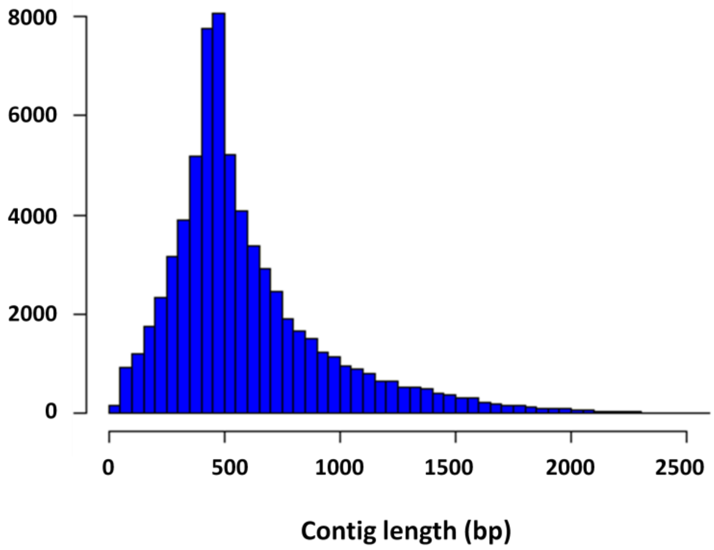


Figure 6.9: Profile of contig lengths obtained from assembling 454 reads.
The normalised and unnormalised reads were assembled together. Contig length is shown on the x-axis and frequency is shown on the y-axis.

6.3.2.2 Analysis of assembled contigs from 454 reads

Testing the assembled contigs (built from both normalised and unnormalised reads) for redundancy identified only 591 contigs (less than 1 %) that were represented by longer contigs. This shows that the majority of assembled contigs represent unique wheat transcript sequence.

Comparing the assembled contigs to NCBI unigenes using BLAST identified 51,396 contigs (75.1 % of all contigs) that showed significant sequence similarity to unigenes and 28,133 unigenes (68.8 % of unigenes) that showed significant sequence similarity to contigs (Figure 6.10). This leaves a portion of NCBI unigenes that are not represented by contigs (light pink area) and a portion of contigs that are not represented in unigenes (light blue area). This means that although the 454 contigs are not representative of the complete unigene set, there are contigs in the set that may represent new sequence. The portion of the contigs showing no similarity to unigenes (17,052 contigs) were analysed further to determine whether they represented new sequence. BLAST analysis showed that 9,946 contigs (58 %) showed significant sequence similarity to 3,295 unique Brachypodium genes and a further 376 contigs (3 %) showed sequence similarity to other plant proteins. This indicates that 9,946 contigs may represent wheat genes that have homologues in Brachypodium but that are not represented in the NCBI unigenes. These contigs have a total length of 4.8 Mb and the longest contig is 2.28 kb. Furthermore, 376 contigs represent wheat genes that are not found in Brachypodium but have homologues identified in other plant species. The total length of these contigs is 179 kb. The remaining 6,730 contigs (39 %) showed no sequence similarity to any plant proteins indicating that these contigs may represent either contaminant or new wheat sequence with no homologues in Brachypodium or other plant species. Using BLAST to compare these contigs to all public nucleotide sequence (evalue cut-off 1e^{-20}) indicated that the majority of these contigs showed no significant similarity to any known sequence (Figure 6.11). However, more than 700 contigs did show significant DNA sequence similarity to wheat sequence. These included mainly *Triticum aestivum* cDNA clones and chromosome 3B BAC clones and also BAC clones, genes and transcript sequences from *T. turgidum*, *T. monococcum* and *T.urartu*. This analysis indicates that there is wheat sequence represented in the 454 contigs that is not present in the NCBI wheat unigenes. Nucleotide sequence from other *Triticeae* species and other grasses are also represented in the contigs in addition to a small amount of contamination in the form of human and non-plant DNA.

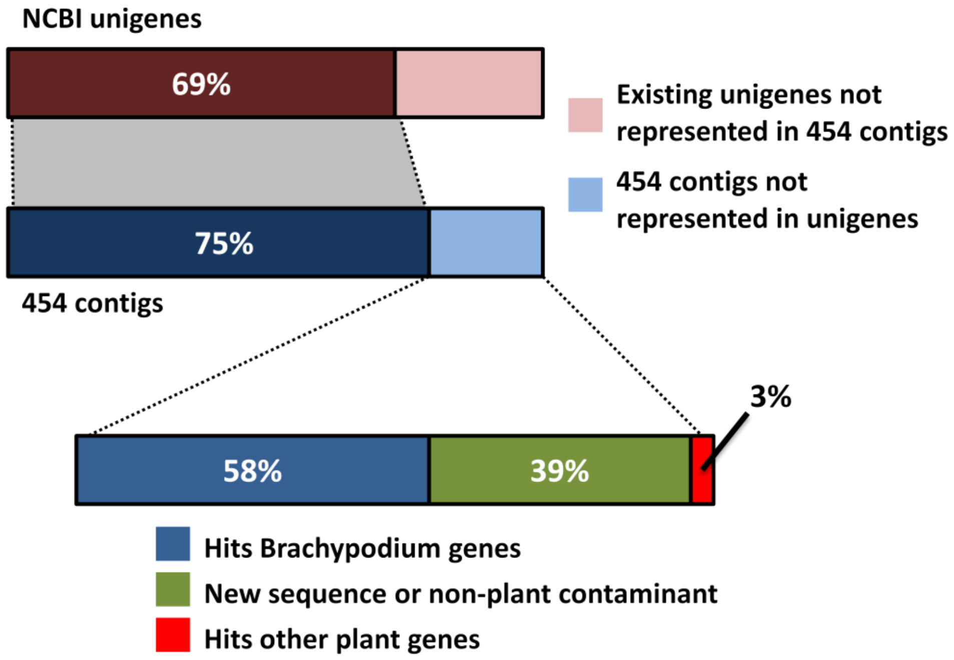


Figure 6.10: Results from analysis of 454 contigs.

First the contigs were compared to NCBI unigenes to determine the proportion of contigs hitting unigenes and the proportion of unigenes hitting contigs. Subsequently, the contigs not represented by unigenes were analysed by comparison to Brachypodium and other plant proteins to determine whether these contigs represented new wheat sequence.

The contigs that hit Brachypodium genes were aligned to these genes to determine what percentage of the coding exons in each gene was covered by contig sequence and 351 of the 9,946 contigs were identified that covered 50 % or more of the coding exons of the homologous Brachypodium gene. On average, 20.5 % of each homologous Brachypodium gene was covered by 454 contig sequence.

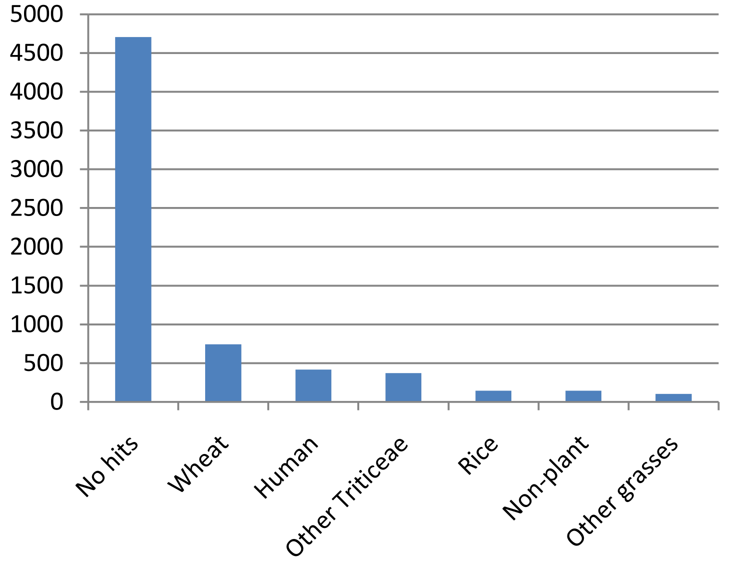


Figure 6.11: BLAST analysis of 454 contigs showing no similarity to existing plant proteins.
These contigs showed no similarity to Brachypodium proteins or to other plant proteins. The majority of contigs could not be identified, a smaller proportion represented wheat sequence, other *Triticeae* species or other grasses and the presence of some human and non-plant contaminant was detected.

Using 454 contigs to extend unigenes resulted in 7,547 unigenes being extended by 100 bp or more and a new unigene set was constructed. This new unigene set included 4.3 Mb of additional sequence, increasing the total length of the unigene set from 36.4 Mb to 40.7 Mb.

Comparing the original NCBI unigene set with Brachypodium genes identified 4,746 unigenes that showed high sequence similarity to a Brachypodium gene and these were classified as putative homologues. Aligning these unigenes to their corresponding Brachypodium homologues gave a mean coverage of 77 %, meaning that on average a unigene sequence covered 77 % of the coding sequence of its homologous Brachypodium gene. The same analysis using the extended unigenes identified 6,008 unigenes that showed high sequence similarity to a Brachypodium gene and aligning these unigenes to their putative Brachypodium homologues gave a mean coverage of 78 %. This shows that extending the unigenes using the 454 contigs has significantly increased the number of Brachypodium genes that can be identified as putative homologues. In addition, extending the unigenes has slightly increased the proportion of Brachypodium coding sequence covered by unigenes.

A few examples of extended unigenes aligned to Brachypodium gene models were visually inspected and two of these are shown in Figure 6.12. Panel A shows unigene S12902301 aligned to Brachypodium gene Bradi5g21900.1. The unigene has been extended from a length of 723 to 1,512 nucleotides by reassembly with similar 454 contigs. This extension allows every exon of the gene to be covered by wheat sequence. Panel B shows unigene S17974984 aligned to Brachypodium gene Bradi1g03170.1. This unigene has been extended from 850 to 1,778 nucleotides and when aligned to Brachypodium, this extended unigene covers all exons of the gene and extends into the UTR region.

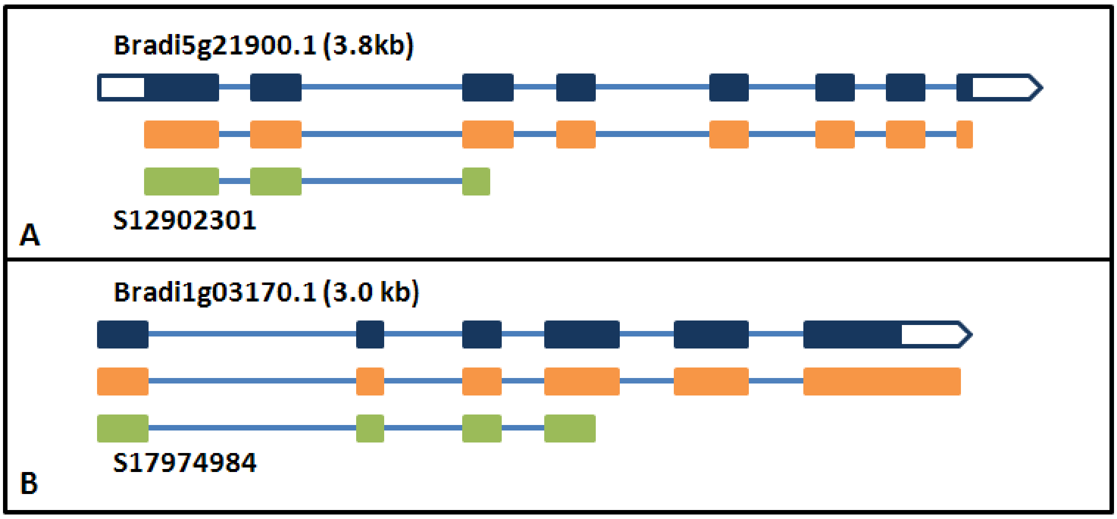


Figure 6.12: Extending unigenes using 454 contigs and aligning to Brachypodium genes to identify gene models. In each panel, the Brachypodium gene is shown at the top with exons in blue, UTRs in white and introns as a blue line. The extended unigene is shown in orange aligning to the gene and the unextended unigene is shown in green.

These examples show that although the mean coverage of Brachypodium genes by unigenes only increased by 1 % after extension, in some cases longer extensions were possible which generate unigenes of a length comparable to a whole gene.

6.4 Discussion

RNA-Seq, the use of next generation sequencing methods to sequence RNA isolated from cells, provides a high-throughput and quantative method to analyse the transcripts present in those cells at specific developmental stages and under different environmental conditions. If a wide range of tissues are sampled, a complete picture of the transcriptome of a species can be generated. RNA-Seq has been used to characterise the transcriptomes of many species which do not yet have sequenced genomes such as the Pine tree (Parchman *et al.* 2010). *T. aestivum* is an important crop species for reasons that have been previously described and due to a large and complex hexaploid genome, has yet to have a well-defined unigene set sequenced, or to be completely sequenced. RNA-Seq can be used to sequence the transcribed portion of this large genome in order to increase the amount of wheat transcriptome sequence that already exists. This existing sequence is largely composed of partial cDNA sequences (ESTs) which have been assembled by automated pipelines and which are incomplete or contain redundant sequence.

The availability of the Brachypodium genome sequence provides wheat researchers with a set of ‘gene templates’ to accurately define wheat gene models by aligning these transcript sequences to Brachypodium using specialised transcript to genomic alignment tools (Wheelan *et al.* 2001). This approach will complement existing wheat gene models derived from rice genes that have been of use in the past. Brachypodium is more closely related to wheat than rice is to wheat, meaning that wheat gene models defined by alignment to Brachypodium will be potentially more accurate than gene models defined by alignment to rice genes.

Wheat cDNA was prepared from a wide range of tissue samples and sequenced using 454 and Illumina platforms. Both normalised and unnormalised samples were sequenced. The Illumina platform generated between 23 and 30 million reads from each lane and when aligned to NCBI unigenes between 20 and 48 % of these reads could be placed. Fewer reads mapped to unigenes from the normalised samples than expected. However, between 65 and 73 % of NCBI unigenes had at least one aligning read, depending on the library used. Only 20 % of reads mapped to unigenes from the normalised library and 73 % of unigenes had reads aligned. A more detailed analysis showed that mapped reads from both normalised and unnormalised libraries generated a similar coverage of unigenes. For example, approximately 30 % of unigenes had reads aligned to an average depth of 5 or more with 50 % or more of their length covered by reads (Figure 6.6).

The read quality profiles and analysis of base frequencies in the reads obtained from the normalised samples indicated a possible problem with the normalisation procedure which may have accounted for the low number of reads mapping from these samples. However, although a low proportion of reads from the normalised samples mapped to unigenes compared to the unnormalised library, a similar number of unigenes are represented. This indicates that the normalisation procedure was working and some abundant transcripts are being removed. A large proportion of reads from the normalised samples did not align to reference sequences which could mean that they were of low quality, or that they represented new wheat sequence. The next stage of investigation would be to assemble these reads with ABySS (Birol *et al.* 2009) or Oases to determine whether the 50 % of unmapped reads from the unnormalised samples or the 79 % of unmapped reads from the normalised samples can be used to generate new sequence contigs that represent wheat transcript sequence not included in the NCBI unigenes set.

Combining all normalised reads together and aligning them to unigenes produced no significant increase in unigene coverage and the same observation was made when all unnormalised reads were combined and aligned to unigenes. However, combining all reads together (normalised and unnormalised) and aligning them to unigenes was not done due to time and computational limitations. It is not expected that this would significantly increase the percentage of unigenes covered as the same transcripts are represented in both libraries although deeper coverage of more abundant transcripts should be obtained from reads in the unnormalised samples.

Normalised and unnormalised wheat cDNA samples were also sequenced using the 454 sequencing platform. This technology produces longer reads meaning that *de novo* assembly is more feasible. The transcriptomes of other plants have been successfully characterised in this way (Novaes *et al.* 2008; Parchman *et al.* 2010). Although the read lengths obtained from the unnormalised library were generally shorter than those obtained from the normalised library (N50 of 392 bp compared to 464 bp) all reads were assembled together resulting in 41.3 Mb of sequence contigs. Seventy-five percent of contigs showed significant similarity to 69 % of unigenes showing that the sequence generated represented a large proportion of existing unigenes as well as potentially new wheat transcript sequence. The contigs matching existing unigenes were used to extend the unigenes and the size of unigene dataset was increased by 4.3 Mb, from 36 to nearly 41 Mb. This is slightly closer to the lowest estimation for the size of the transcriptome of a single wheat chromosome set (49.7 Mb). Interestingly, although in most cases assembling unigenes with matching contigs generated a single extended unigene, in some cases multiple new unigenes were generated which may represent homoeologous genes or the products of alternative splicing from the same gene. Aligning the extended unigenes to Brachypodium gene models showed that in some cases, an extended unigene represented all coding exons of its putative Brachypodium homologue resulting in complete wheat gene models (Figure 6.12). In contrast the unextended unigene covered less than half of coding exons.

From the set of contigs showing no similarity to unigenes, 9,946 contigs showed similarity to Brachypodium protein sequences and 376 contigs showed similarity to other plant proteins. These contigs represent an additional 5 Mb of potential wheat transcript sequence that can be added to the extended unigene set. A small proportion of the contigs showing similarity to Brachypodium genes could be aligned to these genes probably due to synonymous substitutions in the DNA sequences not reflected in the protein sequence. In some cases more than 50 % of the intron-exon structure of the wheat gene could be determined from these alignments. A small proportion of the 6,730 contigs that show no similarity to existing plant proteins were identified as showing similarity to wheat sequence not present in the wheat unigenes. However, these sequences may be represented in the larger transcript assemblies from TIGR or DFCI. The sequences that could not be identified may represent small RNAs and non-coding RNAs which are present in the transcriptome and were not explicitly searched for in this analysis. It is also possible that retroelements in the repeat-rich wheat genome are transcriptionally active and may be represented in the dataset. In the analysis of the Pine transcriptome, 6.2 % of 454 reads that represented transcriptionally active retroelements (Parchman *et al.* 2010). Again, these were not explicitly searched for in this analysis.

This analysis shows that using next-generation sequencing technologies is a useful approach to generate large amounts of transcript sequence from wheat cDNA. However, analysis of this sequence is more challenging. One particular difficulty in this analysis resulted from the low depth of reads from 454 sequencing and difficulties in normalisation of samples for Illumina. Many different transcript assemblies are available meaning that identifying whether a contig resulting from an RNA-Seq experiment is unique is a time-consuming and computationally intensive process. Smaller datasets such as the NCBI unigenes are faster to work with but are incomplete and larger datasets such as the TIGR transcript assemblies are slow to process and contain redundant sequences. The longer reads produced by 454 sequencing appear to be more useful to characterise the transcriptome of wheat as these reads can be assembled *de novo*. However, *de novo* assembly of Illumina reads will undoubtedly provide useful sequence as new transcriptome assembly algorithms are developed. Aligning wheat transcript sequence to Brachypodium genes is an accurate method of defining gene models in wheat and with greater coverage should enable the structure of more than half the genes in the wheat genome to be determined. Work is continuing to increase the variety of tissues from which RNA samples are isolated, to improve normalisation methods, and to increase sequence coverage using both 454 and Illumina platforms. A pipeline has been developed during this analysis to analyse 454 reads and it is envisaged that more unique contigs would provide additional coverage of unigenes as well as new transcript sequence. Although challenging, sequencing the transcriptome of hexaploid wheat is essential to accelerate gene discovery in this important crop plant and to provide an accurate genomic resource in the years before a complete genome sequence is available for wheat.

7 General discussion and conclusions

At the simplest level, comparative genomics describes the comparison between two genomic entities from different taxa, however the range of methodologies covered by comparative genomics is very broad. Comparative genomics provides a foundation for understanding how evolution acts upon genomes and to understand the evolutionary relationship between organisms. The techniques of comparative genomics came of age during the Human Genome Project when the genomes of model species such as *D. melanogaster* (Adams *et al.* 2000) and mouse (Waterston *et al.* 2002) were used to annotate and understand the human genome based on conserved synteny between the species being compared. Comparative genomics approaches are used very widely today due to the rapid increase in sequenced genomes. They have proved to be a powerful component in the toolbox of genomics and a key approach to understanding the evolutionary relationships between diverse organisms. Comparative genomics is also an indispensible research strategy for gene identification and has the potential to transform marker-assisted crop breeding by identifying genes underlying traits and their functions. The material presented in this thesis describes the genome sequencing of *Brachypodium distachyon*, a new experimental grass system and its subsequent use to understand the large and complex genome of bread wheat (*Triticum aestivum*) using a wide range of comparative genomics approaches.

Over the last 10,000 years, wheat has played an important role in providing nutrition to humans and their livestock since it was first domesticated in the fertile crescent of Western Asia (Heun *et al.* 1997). Today, wheat is increasingly important as a crop due to its high yield in relatively unfavourable conditions and high nutritional content, with nearly 700 million metric tonnes produced worldwide in 2008 (FAOStat 2010). With a large increase in the global population as well as changes to the earth's climate predicted to occur over the next 50 years, wheat will also play a central role in the future provision of nutrition, alongside other staple crops such as rice and maize. The development of high yielding varieties of crop species was the basis of the 'green revolution' that occurred in the 1960s and resulted in a large increase in cereal production across the world. In the era of genomics, research into crop plants will once again play a central role in meeting the future challenge of food security (Wollenweber *et al.* 2005).

Once annotated, the complete genome sequence of an organism provides a durable and accurate record of all the genes in the organism's genome. In the case of crop plants, this information can be used to identify the genes underlying important traits and to develop new varieties of crop species with improved yield and increased resistance to environmental factors such as biotic and abiotic stress. Rice was the first crop plant to be sequenced due to its small genome and agricultural importance (International Rice Genome Sequencing Project 2005) and has been followed by the larger genomes of sorghum (Paterson *et al.* 2009) and maize (Schnable *et al.* 2009). The rice and maize genomes were sequenced using a hierarchical clone-by-clone approach and the sorghum genome was sequenced using a whole-genome shotgun approach. Due to its large size, complexity and high repeat content (Smith and Flavell 1975; Bennett and Leitch 1995), sequencing the wheat genome using a clone-by-clone approach or a WGS approach is not practical. To achieve 10x coverage of the 17 Gb wheat genome would require more than one million 150 kb BACs which would require fingerprinting, assembling into contigs and subsequently ordered. Moreover, the high repeat content and presence of homoeologous chromosomes means that reassembling reads from a WGS approach would be very challenging. The availability of the rice genome sequence has facilitated gene isolation in wheat, producing a collection of BAC clones covering many agronomically important genes (Chantret *et al.* 2005; Isidore *et al.* 2005; Griffiths *et al.* 2006; Gu *et al.* 2006). These investigations relied on the high degree of collinearity between genomes in the grass family evident since early comparative analysis using molecular markers (Moore *et al.* 1995). Although useful in this regard, some studies indicated that the evolutionary distance between rice and wheat had resulted in a breakdown of synteny between the species thus limiting the use of rice as a model genome for wheat genomics (Chantret *et al.* 2005; Griffiths *et al.* 2006). The wild grass *B. sylvaticum* had been used to fill this evolutionary gap as early as 1993 (Moore *et al.* 1993b) but it was not until 2001 that *B. distachyon* was proposed as a new experimental system for grasses (Draper *et al.* 2001). Brachypodium was estimated to have diverged from the wheat lineage 35 to 40 MYA (Bossolini *et al.* 2007), significantly more recently than the estimated divergence time (50 MYA) of rice and wheat (Paterson *et al.* 2004a). In 2006, a genome sequencing project was initiated and my contributions to the production and analysis of the Brachypodium genome sequence (International Brachypodium Initiative 2010), the construction of a genetic linkage map (Garvin *et al.* 2010) and an integrated BAC-based physical map (Febrer *et al.* 2010) comprise the topics covered in Chapter 2 of this thesis.

The Brachypodium genome was sequencing using a whole-genome shotgun strategy, then assembled and annotated by an international consortium. I annotated the intermediate release of the genome sequence and provided this information to the research community as an immediate resource using a website developed specifically for this purpose (www.modelcrop.org). Brachypodium was the fourth grass species to be sequenced and the first pooid grass meaning for the first time, genome-wide comparisons between grasses from the three major economically important sub-families within the Poaceae could be performed, the Ehrhartoideae, the Panicoideae and the Pooideae (International Rice Genome Sequencing

Project 2005; Paterson *et al.* 2009). These comparisons allowed Brachypodium to be integrated into the recent detailed model of grass chromosome evolution (Salse *et al.* 2008a). Furthermore, remnants of nested chromosome insertion were identified in the Brachypodium genome. These comprised centromeric repeats from the inserted chromosome that still existed in the target chromosome as well as a high retrotransposon density concomitant with a centromeric region. In addition, a high gene density could still be observed at the former distal end of the inserted chromosome. These observations provided direct evidence for the hypothesis of nested chromosome insertion that had been proposed based on the order of genetic markers in grasses (Kellogg 2001; Srinivasachary *et al.* 2007; Luo *et al.* 2009). The Brachypodium genome analysis confirmed that the insertion of one chromosome into the centromere of another appears to be a unifying mechanism that describes the evolution of grass chromosomes and conservation of chromosome synteny. The BAC-based physical map was used to assess the quality of the WGS sequence assemblies generated by the sequencing project and was essential to provide independent validation of the pseudo-molecules in the final assembly. The physical map was integrated to the genetic map, the karyotype and the WGS assemblies to enhance the accuracy of the genomic sequence and to provide a very high quality whole-genome shotgun assembly for future genomic studies in the grasses.

The development of genomic resources in Brachypodium over the last five years has facilitated its use as an experimental system for temperate grasses and the genome sequence provides a template for the analysis of the genomes of larger crop species. It should be stressed however, that using Brachypodium in comparative studies has the same inherent limitations as observed for all model species in that they remain evolutionarily distant from the species under study. Although Brachypodium is more closely related to wheat than rice is to wheat (International Brachypodium Initiative 2010), comparative studies rely on collinearity between the species being compared which breaks down as evolutionary distance increases. A comparison between wheat sequence from chromosome 3B, rice and Brachypodium identified an ancestral backbone of conserved genes interspersed with non-collinear genes (Choulet *et al.* 2010). Sequence homology to wheat was higher in Brachypodium than in rice but the authors observed that in the regions studied, lineage-specific rearrangements had disrupted synteny in Brachypodium to the same extent as is observed in rice. It is possible that this reduction in synteny is due to accelerated evolution in the *Triticeae* genomes as recently hypothesised by Luo and colleagues from a comparison between *Ae. tauschii*, rice and sorghum (Luo *et al.* 2009). Nevertheless, approaches that exploit synteny to wheat using both the rice and Brachypodium genomes are likely to be more successful than using only one of these species. An approach recently applied to wheat is the use of conserved orthologous set (COS) markers designed from wheat ESTs aligned to orthologous rice genes (Quraishi *et al.* 2009). Primers are designed to span putative introns based on gene structure in rice and have the advantage of a known position in rice and therefore a putative position in wheat. Using orthologous genes from both Brachypodium and rice to design COS markers is likely to provide more useful markers than using a single species.

Phylogenetic footprinting is a powerful comparative genomics technique used to discover potentially functional sequence in genomes based on identifying sequences in related species that have been conserved over evolutionary timescales (Tagle *et al.* 1988). This approach has been used extensively in mammalian genomics (Loots *et al.* 2000; Nobrega *et al.* 2003) to identify putative regulatory elements and more recently has been applied to plants (Thomas *et al.* 2007; Li *et al.* 2009d). In Chapter 3, phylogenetic footprinting was used to compare homologous genes from three diverse members of the grass family (rice, Brachypodium and sorghum) to identify regions with potential gene regulatory function that are conserved in all three species. This analysis produced a set of more than 15,000 sequences with potential regulatory functions. An overrepresentation of the core motif from the DRE/CRT element shown to be involved in drought response (Yamaguchi-Shinozaki and Shinozaki 1994) was found within conserved sequences surrounding genes predicted to be involved in drought response. This suggests that the sequences identified contain functional motifs. The identified sequences are conserved between three diverse grass genomes and therefore are likely to be conserved in grasses with less well defined genomes such as crop species in the *Triticeae* and potential biofuel crops. The availability of the Brachypodium genome sequence provided the first opportunity to perform this type of genome-wide analysis in the grass family and provides a dataset of potential regions to direct future experiments aimed at establishing gene regulatory networks. The validation of these conserved sequences was performed *in silico* and using a dataset extrapolated from rice. Ideally, the next stage of this analysis would be to experimentally determine genes that are differentially expressed in Brachypodium under different conditions and to analyse the conserved sequences surrounding these genes for overrepresented motifs. Experimental validation of these putative regulatory elements could then be undertaken.

Functional genomics aims to assign function to genes and insertional mutagenesis using T-DNA is a fast and cost effective way to develop the large, mutagenised populations which provide a foundation for functional genomics research. T-DNA insertional mutagenesis has been used to develop populations of rice and *Arabidopsis* lines (Alonso et al. 2003; Rosso et al. 2003; Krishnan et al. 2009), which have been used to characterise many genes such as those involved in plant-pathogen interactions (Dellagi et al. 2005; Ramonell et al. 2005). In addition to being closely related to temperate grass crops, Brachypodium is related to grasses with potential use as biofuel crops such as such as *Miscanthus giganteus* and switchgrass (*Panicum virgatum*) and having the experimental resources to identify genes with important function such as those involved in the biosynthesis of lignin polymers, holds great potential (DOE 2006). More than 1,000 mutagenised Brachypodium lines were produced and the analysis of these lines to identify the genomic location of T-DNA insertion is described in Chapter 4. A total of 364 unique genes were identified containing an insertion and 175 of these had already been functionally annotated. This pilot study, to which I contributed the bioinformatics required for insertional analysis, is an important first step to provide a functional genomics resource in this new experimental grass system. Many more lines are required to achieve a mutation in every gene and in order to analyse T-DNA insertions more completely and to compare these with T-DNA insertions in other plants, a larger dataset is required. Mutant lines are being developed by a number of groups internationally which will provide this large dataset. Furthermore, additional bioinformatics expertise is required to develop an automated method to identify and characterise these insertions in a high-throughput manner and with the high level of accuracy required to deal with the complex insertion patterns associated with T-DNA insertion which have hitherto been ignored in large-scale analysis (PV, pers. comm.).

As discussed previously, sequencing the bread wheat genome is an essential but highly challenging step towards the development of improved varieties of wheat. The current approach to generate accurate and contiguous sequence over large distances is to sequence BACs from a minimal tiling path that forms part of an anchored physical map (Paux *et al.* 2008). Although technically feasible, this approach requires physical mapping which is very costly and time consuming. As part of an international effort to sequence the wheat genome, my group at the JIC is physically mapping the long arm of wheat chromosome 3D. A chromosome-based approach is being used which relies on flow cytometry using aneuploid stocks to separate individual chromosomes and chromosome arms (Dolezel *et al.* 2007). Our approach is described in Chapter 5 and closely follows that taken by Paux *et al.* who used a genetic map and a detailed deletion bin map of chromosome 3B to produce a physical map of this chromosome (Paux *et al.* 2008). It also includes my novel method of ordering and orientating the physical map contigs using synteny between the genomes of Brachypodium and wheat. We found that markers designed from wheat ESTs aligned to Brachypodium genes combined with a BES anchoring strategy provided syntenic anchor points for more than one-third of contigs. In addition, more than 5,000 ISBP markers (Paux *et al.* 2010) were designed, providing the potential to anchor 87 % of the 3DL physical map contigs. The novel syntenic anchoring strategy provides a first-pass physical map which will be validated by additional anchoring using a genetic map and a deletion bin map of 3DL which are currently under construction. The syntenic anchoring strategy is based on synteny between Brachypodium and wheat, but as previously described this is sometimes disrupted. Although every effort was made to validate this approach with current data, for example the *Ae. tauschii* genetic map (Luo *et al.* 2009), subsequent comparison to a genetic map and a high-resolution deletion bin map may indicate that this approach is not particularly accurate. Although one-third of physical map contigs have been anchored syntenically to Brachypodium to date, this method was more successful at the distal end of the chromosome where gene density is highest. It may be that this method will ultimately be limited, and the use of markers derived from next generation sequencing and BAC end-sequences could be used to anchor more contigs.

The emergence of the high-throughput sequencing platforms described in Chapter 1 has revolutionised the field of genomics and has enabled broader and more extensive genomic investigations to be performed (Metzker 2010). One application of these technologies is RNA-Seq, where the RNA complement of an organism is sequenced to characterise the transcribed portion of the genome or to quantitatively measure gene expression. RNA-Seq has been used to characterise the transcriptome of plants with large and complex genomes (Parchman *et al.* 2010). In Chapter 6, the Illumina and 454 sequencing platforms were used to characterise the transcriptome of the hexaploid wheat genome. The transcript sequences obtained from assembling 454 reads were used to expand the existing wheat unigene set and were also aligned to Brachypodium genes to accurately define wheat gene models. In some cases, complete wheat gene models were defined. The Illumina sequence data generated provides the opportunity to assemble transcript sequence *de novo* using new assembly algorithms. Further sequencing will result in increased gene coverage, and a more accurate set of wheat gene sequences can be constructed. These can then be aligned to Brachypodium genes to define accurate gene models in wheat. Two particular limitations in this analysis were firstly the relatively low amount of 454 sequence coverage generated during this study and secondly, technical problems in cDNA normalisation when preparing cDNA for Illumina sequencing. These problems have now been overcome, and it is anticipated that in the coming months deep transcriptome coverage will be generated, and assembly and alignment using the strategy I have developed in this thesis will form an important part of this analysis.

A significant proportion of the research in this thesis has already been published (Febrer *et al.* 2009; Febrer *et al.* 2010; Garvin *et al.* 2010; International Brachypodium Initiative 2010; Thole *et al.* 2010). Other research projects are still ongoing, such as chromosome painting in Brachypodium, the physical mapping of wheat 3DL and the wheat transcriptome analysis. In addition to the current effort to physically map and sequence the wheat genome using a chromosome-based approach, new sequencing technologies mean that direct sequencing of the wheat genome is now feasible. The recent announcement that 5x coverage of the wheat genome has been achieved using the 454 sequencing platform is a milestone in wheat genomics (JIC Press Office 2010). This has been achieved by a group of UK scientists with contributions from my own group at the JIC. Although assembly and annotation of this sequence is a monumental task, a new generation of assembly algorithms such as Curtain (http://code.google.com/p/curtain) and Cortex (M. Caccamo, pers. comm.) are being developed that are designed to deal with the assembly of large eukaryotic genomes. The 85 Gb of wheat sequence should provide a challenging dataset for these new algorithms. In addition, the availability of real-time, single molecule DNA sequencing technologies, such as the SMRT technology being developed by Pacific Biosciences (Eid *et al.* 2009) described in Chapter 1, will provide more robust methods to sequence large and complex genomes. The strobe sequencing protocol available with this technology (Pacific Biosciences 2009) will allow short blocks of sequence to be read over very long insert lengths, rather than the traditional mate-pair reads from each end of an insert. For the highly repetitive wheat genome, this will allow reads to span repetitive regions and provide a long-range scaffold for assembly of shorter reads that provide deep coverage.

In conclusion, it is an exciting time for grass genomics. The genome sequencing technologies are emerging which, combined with new methods of genome assembly and analysis will allow the genomes of many key species to be sequenced and analysed over the coming years. One of these will be the large genome of wheat which only a few years ago, was considered impossible to sequence. This wealth of genomic information will provide the means to address one of the most pressing issues of our time - how to feed a growing population. The material presented in this thesis is an important contribution towards this goal.

Abbreviations

BAC: Bacterial artificial chromosome

BES: BAC-end sequence

BLAST: Basic local alignment search tool

cDNA: Complementary DNA

CNS: Conserved non-coding sequence

COS: Conserved Orthologous Set

DNA: Deoxyribonucleic acid

EST: Expressed sequence tag

FISH: Fluorescence *in situ* hybridisation

FPC: Fingerprinted contigs software

FST: Flanking sequence tag

GFF: Generic feature format

HICF: High information content fingerprinting

HSP: High-scoring pair

IBI: International Brachypodium Initiative

ISBP: Insertion site-based polymorphism

JGI: Joint Genome Institute

JIC: John Innes Centre

LTR: Long terminal repeat

MIPS: Munich Information Centre for protein sequences

mRNA: Messenger RNA

MTP: Minimum tiling path

MYA: Million years ago

NGS: Next-generation sequencing

ORF: Open reading frame

PAC: P1 artificial chromosome

PCR: Polymerase chain reaction

RNA: Ribonucleic acid

rRNA: Ribosomal RNA

SNP: Single nucleotide polymorphism

SSR: Simple sequence repeat

tRNA: Transfer RNA

USDA: United States Department of Agriculture

WGD: Whole-genome duplication

WGS: Whole-genome shotgun

References

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., George, R. A., Lewis, S. E., Richards, S., Ashburner, M., Henderson, S. N., Sutton, G. G., Wortman, J. R., Yandell, M. D., Zhang, Q., Chen, L. X., Brandon, R. C., Rogers, Y. H., Blazej, R. G., Champe, M., Pfeiffer, B. D., Wan, K. H., Doyle, C., Baxter, E. G., Helt, G., Nelson, C. R., Gabor, G. L., Abril, J. F., Agbayani, A., An, H. J., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R. M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E. M., Beeson, K. Y., Benos, P. V., Berman, B. P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M. R., Bouck, J., Brokstein, P., Brottier, P., Burtis, K. C., Busam, D. A., Butler, H., Cadieu, E., Center, A., Chandra, I., Cherry, J. M., Cawley, S., Dahlke, C., Davenport, L. B., Davies, P., de Pablos, B., Delcher, A., Deng, Z., Mays, A. D., Dew, I., Dietz, S. M., Dodson, K., Doup, L. E., Downes, M., Dugan-Rocha, S., Dunkov, B. C., Dunn, P., Durbin, K. J., Evangelista, C. C., Ferraz, C., Ferriera, S., Fleischmann, W., Fosler, C., Gabrielian, A. E., Garg, N. S., Gelbart, W. M., Glasser, K., Glodek, A., Gong, F., Gorrell, J. H., Gu, Z., Guan, P., Harris, M., Harris, N. L., Harvey, D., Heiman, T. J., Hernandez, J. R., Houck, J., Hostin, D., Houston, K. A., Howland, T. J., Wei, M. H., Ibegwam, C., Jalali, M., Kalush, F., Karpen, G. H., Ke, Z., Kennison, J. A., Ketchum, K. A., Kimmel, B. E., Kodira, C. D., Kraft, C., Kravitz, S., Kulp, D., Lai, Z., Lasko, P., Lei, Y., Levitsky, A. A., Li, J., Li, Z., Liang, Y., Lin, X., Liu, X., Mattei, B., McIntosh, T. C., McLeod, M. P., McPherson, D., Merkulov, G., Milshina, N. V., Mobarry, C., Morris, J., Moshrefi, A., Mount, S. M., Moy, M., Murphy, B., Murphy, L., Muzny, D. M., Nelson, D. L., Nelson, D. R., Nelson, K. A., Nixon, K., Nusskern, D. R., Pacleb, J. M., Palazzolo, M., Pittman, G. S., Pan, S., Pollard, J., Puri, V., Reese, M. G., Reinert, K., Remington, K., Saunders, R. D., Scheeler, F., Shen, H., Shue, B. C., Siden-Kiamos, I., Simpson, M., Skupski, M. P., Smith, T., Spier, E., Spradling, A. C., Stapleton, M., Strong, R., Sun, E., Svirskas, R., Tector, C., Turner, R., Venter, E., Wang, A. H., Wang, X., Wang, Z. Y., Wassarman, D. A., Weinstock, G. M., Weissenbach, J., Williams, S. M., WoodageT, Worley, K. C., Wu, D., Yang, S., Yao, Q. A., Ye, J., Yeh, R. F., Zaveri, J. S., Zhan, M., Zhang, G., Zhao, Q., Zheng, L., Zheng, X. H., Zhong, F. N., Zhong, W., Zhou, X., Zhu, S., Zhu, X., Smith, H. O., Gibbs, R. A., Myers, E. W., Rubin, G. M. and Venter, J. C. (2000). "The genome sequence of Drosophila melanogaster." Science **287**(5461): 2185-2195.

Afolabi, A. S., Worland, B., Snape, J. W. and Vain, P. (2004). "A large-scale study of rice plants transformed with different T-DNAs provides new insights into locus composition and T-DNA linkage configurations." Theor Appl Genet **109**(4): 815-826.

Ahn, S. and Tanksley, S. D. (1993). "Comparative Linkage Maps of the Rice and Maize Genomes." Proceedings of the National Academy of Sciences of the United States of America **90**(17): 7980-7984.

Alexandersson, M., Cawley, S. and Pachter, L. (2003). "SLAM: cross-species gene finding and alignment with a generalized pair hidden Markov model." Genome Res **13**(3): 496-502.

Allen, J. E. and Salzberg, S. L. (2005). "JIGSAW: integration of multiple sources of evidence for gene prediction." Bioinformatics **21**(18): 3596-3603.

Alonso, J. M. and Ecker, J. R. (2006). "Moving forward in reverse: genetic technologies to enable genome-wide phenomic screens in Arabidopsis." Nat Rev Genet **7**(7): 524-536.

Alonso, J. M., Stepanova, A. N., Leisse, T. J., Kim, C. J., Chen, H., Shinn, P., Stevenson, D. K., Zimmerman, J., Barajas, P., Cheuk, R., Gadrinab, C., Heller, C., Jeske, A., Koesema, E., Meyers, C. C., Parker, H., Prednis, L., Ansari, Y., Choy, N., Deen, H., Geralt, M., Hazari, N., Hom, E., Karnes, M., Mulholland, C., Ndubaku, R., Schmidt, I., Guzman, P., Aguilar-Henonin, L., Schmid, M., Weigel, D., Carter, D. E., Marchand, T., Risseeuw, E., Brogden, D., Zeko, A., Crosby, W. L., Berry, C. C. and Ecker, J. R. (2003). "Genome-wide insertional mutagenesis of Arabidopsis thaliana." Science **301**(5633): 653-657.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-410.

Alves, S. C., Worland, B., Thole, V., Snape, J. W., Bevan, M. W. and Vain, P. (2009). "A protocol for Agrobacterium-mediated transformation of Brachypodium distachyon community standard line Bd21." Nat Protoc **4**(5): 638-649.

Alwine, J. C., Kemp, D. J. and Stark, G. R. (1977). "Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes." Proc Natl Acad Sci U S A **74**(12): 5350-5354.

Ammiraju, J. S., Luo, M., Goicoechea, J. L., Wang, W., Kudrna, D., Mueller, C., Talag, J., Kim, H., Sisneros, N. B., Blackmon, B., Fang, E., Tomkins, J. B., Brar, D., MacKill, D., McCouch, S., Kurata, N., Lambert, G., Galbraith, D. W., Arumuganathan, K., Rao, K., Walling, J. G., Gill, N., Yu, Y., SanMiguel, P., Soderlund, C., Jackson, S. and Wing, R. A. (2006). "The Oryza bacterial artificial chromosome library resource: construction and analysis of 12 deep-coverage large-insert BAC libraries that represent the 10 genome types of the genus Oryza." Genome Res **16**(1): 140-147.

An, S., Park, S., Jeong, D. H., Lee, D. Y., Kang, H. G., Yu, J. H., Hur, J., Kim, S. R., Kim, Y. H., Lee, M., Han, S., Kim, S. J., Yang, J., Kim, E., Wi, S. J., Chung, H. S., Hong, J. P., Choe, V., Lee, H. K., Choi, J. H., Nam, J., Park, P. B., Park, K. Y., Kim, W. T., Choe, S., Lee, C. B. and An, G. (2003). "Generation and analysis of end sequence database for T-DNA tagging lines in rice." Plant Physiol **133**(4): 2040-2047.

Aragon-Alcaide, L., Miller, T., Schwarzacher, T., Reader, S. and Moore, G. (1996). "A cereal centromeric sequence." Chromosoma **105**(5): 261-268.

Astier, Y., Braha, O. and Bayley, H. (2006). "Toward single molecule DNA sequencing: direct identification of ribonucleoside and deoxyribonucleoside 5'-monophosphates by using an engineered protein nanopore equipped with a molecular adapter." J Am Chem Soc **128**(5): 1705-1710.

Ayele, M., Haas, B. J., Kumar, N., Wu, H., Xiao, Y., Van Aken, S., Utterback, T. R., Wortman, J. R., White, O. R. and Town, C. D. (2005). "Whole genome shotgun sequencing of Brassica oleracea and its application to gene discovery and annotation in Arabidopsis." Genome Res **15**(4): 487-495.

Ballinger, D. G. and Benzer, S. (1989). "Targeted gene mutations in Drosophila." Proc Natl Acad Sci U S A **86**(23): 9402-9406.

Bao, Z. and Eddy, S. R. (2002). "Automated de novo identification of repeat sequence families in sequenced genomes." Genome Res **12**(8): 1269-1276.

Barillot, E., Lacroix, B. and Cohen, D. (1991). "Theoretical analysis of library screening using a N-dimensional pooling strategy." Nucleic Acids Res **19**(22): 6241-6247.

Batzoglou, S., Jaffe, D. B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., Berger, B., Mesirov, J. P. and Lander, E. S. (2002). "ARACHNE: a whole-genome shotgun assembler." Genome Res **12**(1): 177-189.

Bedell, J. A., Budiman, M. A., Nunberg, A., Citek, R. W., Robbins, D., Jones, J., Flick, E., Rholfing, T., Fries, J., Bradford, K., McMenamy, J., Smith, M., Holeman, H., Roe, B. A., Wiley, G., Korf, I. F., Rabinowicz, P. D., Lakey, N., McCombie, W. R., Jeddeloh, J. A. and Martienssen, R. A. (2005). "Sorghum genome sequencing by methylation filtration." PLoS Biol **3**(1): e13.

Bender, W., Spierer, P. and Hogness, D. S. (1983). "Chromosomal walking and jumping to isolate DNA from the Ace and rosy loci and the bithorax complex in Drosophila melanogaster." J Mol Biol **168**(1): 17-33.

Bennett, M. D. and Leitch, I. J. (1995). "Nuclear-DNA Amounts in Angiosperms." Annals of Botany **76**(2): 113-176.

Bennett, M. D. and Leitch, I. J. (2005). Plant DNA C-values Database.

Bennett, S. T., Barnes, C., Cox, A., Davies, L. and Brown, C. (2005). "Toward the 1,000 dollars human genome." Pharmacogenomics **6**(4): 373-382.

Bennetzen, J. L., Coleman, C., Liu, R., Ma, J. and Ramakrishna, W. (2004). "Consistent over-estimation of gene number in complex plant genomes." Curr Opin Plant Biol **7**(6): 732-736.

Bennetzen, J. L. and Ma, J. (2003). "The genetic colinearity of rice and other cereals on the basis of genomic sequence analysis." Curr Opin Plant Biol **6**(2): 128-133.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. and Wheeler, D. L. (2007). "GenBank." Nucleic Acids Res **35**(Database issue): D21-25.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M., Reed, M. T., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance, P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X., Zhang, L., Alam, M. D., Anastasi, C., Aniebo, I. C., Bailey, D. M., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. A., Benoit, V. A., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. A., Brown, R. C., Brown, A. A., Buermann, D. H., Bundu, A. A., Burrows, J. C., Carter, N. P., Castillo, N., Chiara, E. C. M., Chang, S., Neil Cooley, R., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fuentes Fajardo, K. V., Scott Furey, W., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. A., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., Hoschler, K., Hurwitz, S., Ivanov, D. V., Johnson, M. Q., James, T., Huw Jones, T. A., Kang, G. D., Kerelska, T. H., Kersey, A. D., Khrebtukova, I., Kindwall, A. P., Kingsbury, Z., Kokko-Gonzales, P. I., Kumar, A., Laurent, M. A., Lawley, C. T., Lee, S. E., Lee, X., Liao, A. K., Loch, J. A., Lok, M., Luo, S., Mammen, R. M., Martin, J. W., McCauley, P. G., McNitt, P., Mehta, P., Moon, K. W., Mullens, J. W., Newington, T., Ning, Z., Ling Ng, B., Novo, S. M., O'Neill, M. J., Osborne, M. A., Osnowski, A., Ostadan, O., Paraschos, L. L., Pickering, L., Pike, A. C., Chris Pinkard, D., Pliskin, D. P., Podhasky, J., Quijano, V. J., Raczy, C., Rae, V. H., Rawlings, S. R., Chiva Rodriguez, A., Roe, P. M., Rogers, J., Rogert Bacigalupo, M. C., Romanov, N., Romieu, A., Roth, R. K., Rourke, N. J., Ruediger, S. T., Rusman, E., Sanches-Kuiper, R. M., Schenker, M. R., Seoane, J. M., Shaw, R. J., Shiver, M. K., Short, S. W., Sizto, N. L., Sluis, J. P., Smith, M. A., Ernest Sohna Sohna, J., Spence, E. J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C. L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S. M., Wakelin, S., Walcott, G. C., Wang, J., Worsley, G. J., Yan, J., Yau, L., Zuerlein, M., Mullikin, J. C., Hurles, M. E., McCooke, N. J., West, J. S., Oaks, F. L., Lundberg, P. L., Klenerman, D., Durbin, R. and Smith, A. J. (2008). "Accurate whole human genome sequencing using reversible terminator chemistry." Nature **456**(7218): 53-59.

Birney, E., Clamp, M. and Durbin, R. (2004). "GeneWise and Genomewise." Genome Res **14**(5): 988-995.

Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigo, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Thurman, R. E., Kuehn, M. S., Taylor, C. M., Neph, S., Koch, C. M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J. A., Andrews, R. M., Flicek, P., Boyle, P. J., Cao, H., Carter, N. P., Clelland, G. K., Davis, S., Day, N., Dhami, P., Dillon, S. C., Dorschner, M. O., Fiegler, H., Giresi, P. G., Goldy, J., Hawrylycz, M., Haydock, A., Humbert, R., James, K. D., Johnson, B. E., Johnson, E. M., Frum, T. T., Rosenzweig, E. R., Karnani, N., Lee, K., Lefebvre, G. C., Navas, P. A., Neri, F., Parker, S. C., Sabo, P. J., Sandstrom, R., Shafer, A., Vetrie, D., Weaver, M., Wilcox, S., Yu, M., Collins, F. S., Dekker, J., Lieb, J. D., Tullius, T. D., Crawford, G. E., Sunyaev, S., Noble, W. S., Dunham, I., Denoeud, F., Reymond, A., Kapranov, P., Rozowsky, J., Zheng, D., Castelo, R., Frankish, A., Harrow, J., Ghosh, S., Sandelin, A., Hofacker, I. L., Baertsch, R., Keefe, D., Dike, S., Cheng, J., Hirsch, H. A., Sekinger, E. A., Lagarde, J., Abril, J. F., Shahab, A., Flamm, C., Fried, C., Hackermuller, J., Hertel, J., Lindemeyer, M., Missal, K., Tanzer, A., Washietl, S., Korbel, J., Emanuelsson, O., Pedersen, J. S., Holroyd, N., Taylor, R., Swarbreck, D., Matthews, N., Dickson, M. C., Thomas, D. J., Weirauch, M. T., Gilbert, J., Drenkow, J., Bell, I., Zhao, X., Srinivasan, K. G., Sung, W. K., Ooi, H. S., Chiu, K. P., Foissac, S., Alioto, T., Brent, M., Pachter, L., Tress, M. L., Valencia, A., Choo, S. W., Choo, C. Y., Ucla, C., Manzano, C., Wyss, C., Cheung, E., Clark, T. G., Brown, J. B., Ganesh, M., Patel, S., Tammana, H., Chrast, J., Henrichsen, C. N., Kai, C., Kawai, J., Nagalakshmi, U., Wu, J., Lian, Z., Lian, J., Newburger, P., Zhang, X., Bickel, P., Mattick, J. S., Carninci, P., Hayashizaki, Y., Weissman, S., Hubbard, T., Myers, R. M., Rogers, J., Stadler, P. F., Lowe, T. M., Wei, C. L., Ruan, Y., Struhl, K., Gerstein, M., Antonarakis, S. E., Fu, Y., Green, E. D., Karaoz, U., Siepel, A., Taylor, J., Liefer, L. A., Wetterstrand, K. A., Good, P. J., Feingold, E. A., Guyer, M. S., Cooper, G. M., Asimenos, G., Dewey, C. N., Hou, M., Nikolaev, S., Montoya-Burgos, J. I., Loytynoja, A., Whelan, S., Pardi, F., Massingham, T., Huang, H., Zhang, N. R., Holmes, I., Mullikin, J. C., Ureta-Vidal, A., Paten, B., Seringhaus, M., Church, D., Rosenbloom, K., Kent, W. J., Stone, E. A., Batzoglou, S., Goldman, N., Hardison, R. C., Haussler, D., Miller, W., Sidow, A., Trinklein, N. D., Zhang, Z. D., Barrera, L., Stuart, R., King, D. C., Ameur, A., Enroth, S., Bieda, M. C., Kim, J., Bhinge, A. A., Jiang, N., Liu, J., Yao, F., Vega, V. B., Lee, C. W., Ng, P., Shahab, A., Yang, A., Moqtaderi, Z., Zhu, Z., Xu, X., Squazzo, S., Oberley, M. J., Inman, D., Singer, M. A., Richmond, T. A., Munn, K. J., Rada-Iglesias, A., Wallerman, O., Komorowski, J., Fowler, J. C., Couttet, P., Bruce, A. W., Dovey, O. M., Ellis, P. D., Langford, C. F., Nix, D. A., Euskirchen, G., Hartman, S., Urban, A. E., Kraus, P., Van Calcar, S., Heintzman, N., Kim, T. H., Wang, K., Qu, C., Hon, G., Luna, R., Glass, C. K., Rosenfeld, M. G., Aldred, S. F., Cooper, S. J., Halees, A., Lin, J. M., Shulha, H. P., Zhang, X., Xu, M., Haidar, J. N., Yu, Y., Ruan, Y., Iyer, V. R., Green, R. D., Wadelius, C., Farnham, P. J., Ren, B., Harte, R. A., Hinrichs, A. S., Trumbower, H., Clawson, H., Hillman-Jackson, J., Zweig, A. S., Smith, K., Thakkapallayil, A., Barber, G., Kuhn, R. M., Karolchik, D., Armengol, L., Bird, C. P., de Bakker, P. I., Kern, A. D., Lopez-Bigas, N., Martin, J. D., Stranger, B. E., Woodroffe, A., Davydov, E., Dimas, A., Eyras, E., Hallgrimsdottir, I. B., Huppert, J., Zody, M. C., Abecasis, G. R., Estivill, X., Bouffard, G. G., Guan, X., Hansen, N. F., Idol, J. R., Maduro, V. V., Maskeri, B., McDowell, J. C., Park, M., Thomas, P. J., Young, A. C., Blakesley, R. W., Muzny, D. M., Sodergren, E., Wheeler, D. A., Worley, K. C., Jiang, H., Weinstock, G. M., Gibbs, R. A., Graves, T., Fulton, R., Mardis, E. R., Wilson, R. K., Clamp, M., Cuff, J., Gnerre, S., Jaffe, D. B., Chang, J. L., Lindblad-Toh, K., Lander, E. S., Koriabine, M., Nefedov, M., Osoegawa, K., Yoshinaga, Y., Zhu, B. and de Jong, P. J. (2007). "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project." Nature **447**(7146): 799-816.

Birol, I., Jackman, S. D., Nielsen, C. B., Qian, J. Q., Varhol, R., Stazyk, G., Morin, R. D., Zhao, Y., Hirst, M., Schein, J. E., Horsman, D. E., Connors, J. M., Gascoyne, R. D., Marra, M. A. and Jones, S. J. (2009). "De novo transcriptome assembly with ABySS." Bioinformatics **25**(21): 2872-2877.

Blanc, G., Hokamp, K. and Wolfe, K. H. (2003). "A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome." Genome Res **13**(2): 137-144.

Blanc, G. and Wolfe, K. H. (2004a). "Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution." Plant Cell **16**(7): 1679-1691.

Blanc, G. and Wolfe, K. H. (2004b). "Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes." Plant Cell **16**(7): 1667-1678.

Blanchette, M. and Tompa, M. (2002). "Discovery of regulatory elements by a computational method for phylogenetic footprinting." Genome Res **12**(5): 739-748.

Blattner, F. R., Plunkett, G., 3rd, Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B. and Shao, Y. (1997). "The complete genome sequence of Escherichia coli K-12." Science **277**(5331): 1453-1474.

Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K. D., Ovcharenko, I., Pachter, L. and Rubin, E. M. (2003). "Phylogenetic shadowing of primate sequences to find functional regions of the human genome." Science **299**(5611): 1391-1394.

Boguski, M. S., Lowe, T. M. and Tolstoshev, C. M. (1993). "dbEST--database for "expressed sequence tags"." Nat Genet **4**(4): 332-333.

Bonnet, E., Wuyts, J., Rouze, P. and Van de Peer, Y. (2004). "Detection of 91 potential conserved plant microRNAs in Arabidopsis thaliana and Oryza sativa identifies important target genes." Proc Natl Acad Sci U S A **101**(31): 11511-11516.

Bossolini, E., Wicker, T., Knobel, P. A. and Keller, B. (2007). "Comparison of orthologous loci from small grass genomes Brachypodium and rice: implications for wheat genomics and grass genome annotation." Plant J.

Bowers, J. E., Abbey, C., Anderson, S., Chang, C., Draye, X., Hoppe, A. H., Jessup, R., Lemke, C., Lennington, J., Li, Z., Lin, Y. R., Liu, S. C., Luo, L., Marler, B. S., Ming, R., Mitchell, S. E., Qiang, D., Reischmann, K., Schulze, S. R., Skinner, D. N., Wang, Y. W., Kresovich, S., Schertz, K. F. and Paterson, A. H. (2003a). "A high-density genetic recombination map of sequence-tagged sites for sorghum, as a framework for comparative structural and evolutionary genomics of tropical grains and grasses." Genetics **165**(1): 367-386.

Bowers, J. E., Chapman, B. A., Rong, J. and Paterson, A. H. (2003b). "Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events." Nature **422**(6930): 433-438.

Bray, N., Dubchak, I. and Pachter, L. (2003). "AVID: A global alignment program." Genome Res **13**(1): 97-102.

Bray, N. and Pachter, L. (2004). "MAVID: constrained ancestral alignment of multiple sequences." Genome Res **14**(4): 693-699.

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D. H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., Roth, R., George, D., Eletr, S., Albrecht, G., Vermaas, E., Williams, S. R., Moon, K., Burcham, T., Pallas, M., DuBridge, R. B., Kirchner, J., Fearon, K., Mao, J. and Corcoran, K. (2000). "Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays." Nat Biotechnol **18**(6): 630-634.

Britten, R. J., Graham, D. E. and Neufeld, B. R. (1974). "Analysis of repeating DNA sequences by reassociation." Methods Enzymol **29**(0): 363-418.

Brooks, S. A., Huang, L., Gill, B. S. and Fellers, J. P. (2002). "Analysis of 106 kb of contiguous DNA sequence from the D genome of wheat reveals high gene density and a complex arrangement of genes related to disease resistance." Genome **45**(5): 963-972.

Brudno, M., Do, C. B., Cooper, G. M., Kim, M. F., Davydov, E., Green, E. D., Sidow, A. and Batzoglou, S. (2003). "LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA." Genome Res **13**(4): 721-731.

Brunner, S., Pea, G. and Rafalski, A. (2005). "Origins, genetic organization and transcription of a family of non-autonomous helitron elements in maize." Plant J **43**(6): 799-810.

Bureau, T. E. and Wessler, S. R. (1992). "Tourist: a large family of small inverted repeat elements frequently associated with maize genes." Plant Cell **4**(10): 1283-1294.

Burge, C. and Karlin, S. (1997). "Prediction of complete gene structures in human genomic DNA." J Mol Biol **268**(1): 78-94.

Burrows, M. and Wheeler, D. J. (1994). A block-sorting lossless data compression algorithm., Digital Equipment Corporation.

Caldwell, D. G., McCallum, N., Shaw, P., Muehlbauer, G. J., Marshall, D. F. and Waugh, R. (2004). "A structured mutant population for forward and reverse genetics in Barley (Hordeum vulgare L.)." Plant J **40**(1): 143-150.

Cannon, S. B., Mitra, A., Baumgarten, A., Young, N. D. and May, G. (2004). "The roles of segmental and tandem gene duplication in the evolution of large gene families in Arabidopsis thaliana." BMC Plant Biol **4**: 10.

Capecchi, M. R. (1989). "Altering the genome by homologous recombination." Science **244**(4910): 1288-1292.

Carver, T. J., Rutherford, K. M., Berriman, M., Rajandream, M. A., Barrell, B. G. and Parkhill, J. (2005). "ACT: the Artemis Comparison Tool." Bioinformatics **21**(16): 3422-3423.

Castelli, V., Aury, J. M., Jaillon, O., Wincker, P., Clepet, C., Menard, M., Cruaud, C., Quetier, F., Scarpelli, C., Schachter, V., Temple, G., Caboche, M., Weissenbach, J. and Salanoubat, M. (2004). "Whole genome sequence comparisons and "full-length" cDNA sequences: a combined approach to evaluate and improve Arabidopsis genome annotation." Genome Res **14**(3): 406-413.

Catalan, P. and Olmstead, R. G. (2000). "Phylogenetic reconstruction of the genus Branchypodium P-Beauv. (Poaceae) from combined sequences of chloroplast ndhF gene and nuclear ITS." Plant Systematics and Evolution **220**(1-2): 1-19.

Catalan, P., Shi, Y., Armstrong, L., Draper, J. and Stace, C. A. (1995). "Molecular Phylogeny of the Grass Genus Brachypodium P-Beauv Based on Rflp and Rapd Analysis." Botanical Journal of the Linnean Society **117**(4): 263-280.

Chaisson, M. J. and Pevzner, P. A. (2008). "Short read fragment assembly of bacterial genomes." Genome Res **18**(2): 324-330.

Chantret, N., Cenci, A., Sabot, F., Anderson, O. and Dubcovsky, J. (2004). "Sequencing of the Triticum monococcum hardness locus reveals good microcolinearity with rice." Mol Genet Genomics **271**(4): 377-386.

Chantret, N., Salse, J., Sabot, F., Rahman, S., Bellec, A., Laubin, B., Dubois, I., Dossat, C., Sourdille, P., Joudrier, P., Gautier, M. F., Cattolico, L., Beckert, M., Aubourg, S., Weissenbach, J., Caboche, M., Bernard, M., Leroy, P. and Chalhoub, B. (2005). "Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploid wheat species (Triticum and Aegilops)." Plant Cell **17**(4): 1033-1045.

Charles, M., Tang, H., Belcram, H., Paterson, A., Gornicki, P. and Chalhoub, B. (2009). "Sixty million years in evolution of soft grain trait in grasses: emergence of the softness locus in the common ancestor of Pooideae and Ehrhartoideae, after their divergence from Panicoideae." Mol Biol Evol.

Chatterji, S. and Pachter, L. (2006). "Reference based annotation with GeneMapper." Genome Biol **7**(4): R29.

Chaw, S. M., Chang, C. C., Chen, H. L. and Li, W. H. (2004). "Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes." J Mol Evol **58**(4): 424-441.

Chen, M., Presting, G., Barbazuk, W. B., Goicoechea, J. L., Blackmon, B., Fang, G., Kim, H., Frisch, D., Yu, Y., Sun, S., Higingbottom, S., Phimphilai, J., Phimphilai, D., Thurmond, S., Gaudette, B., Li, P., Liu, J., Hatfield, J., Main, D., Farrar, K., Henderson, C., Barnett, L., Costa, R., Williams, B., Walser, S., Atkins, M., Hall, C., Budiman, M. A., Tomkins, J. P., Luo, M., Bancroft, I., Salse, J., Regad, F., Mohapatra, T., Singh, N. K., Tyagi, A. K., Soderlund, C., Dean, R. A. and Wing, R. A. (2002). "An integrated physical and genetic map of the rice genome." Plant Cell **14**(3): 537-545.

Chen, M., SanMiguel, P., de Oliveira, A. C., Woo, S. S., Zhang, H., Wing, R. A. and Bennetzen, J. L. (1997). "Microcolinearity in sh2-homologous regions of the maize, rice, and sorghum genomes." Proc Natl Acad Sci U S A **94**(7): 3431-3435.

Chimpanzee Sequencing and Analysis Consortium (2005). "Initial sequence of the chimpanzee genome and comparison with the human genome." Nature **437**(7055): 69-87.

Chopra, S., Brendel, V., Zhang, J., Axtell, J. D. and Peterson, T. (1999). "Molecular characterization of a mutable pigmentation phenotype and isolation of the first active transposable element from Sorghum bicolor." Proc Natl Acad Sci U S A **96**(26): 15330-15335.

Chou, H. H. and Holmes, M. H. (2001). "DNA sequence quality trimming and vector removal." Bioinformatics **17**(12): 1093-1104.

Choulet, F., Wicker, T., Rustenholz, C., Paux, E., Salse, J., Leroy, P., Schlub, S., Le Paslier, M. C., Magdelenat, G., Gonthier, C., Couloux, A., Budak, H., Breen, J., Pumphrey, M., Liu, S., Kong, X., Jia, J., Gut, M., Brunel, D., Anderson, J. A., Gill, B. S., Appels, R., Keller, B. and Feuillet, C. (2010). "Megabase Level Sequencing Reveals Contrasted Organization and Evolution Patterns of the Wheat Gene and Transposable Element Spaces." Plant Cell.

Clayton, W. D., Harman, K. T. and Williamson, H. (2006 onwards). "GrassBase - The Online World Grass Flora." from http://www.kew.org/data/grasses-db.html.

Clayton, W. D. and Renvoize, S. A. (1986). Genera Graminum, Her Majesty's Stationery Office, London.

Close, T. J., Bhat, P. R., Lonardi, S., Wu, Y., Rostoks, N., Ramsay, L., Druka, A., Stein, N., Svensson, J. T., Wanamaker, S., Bozdag, S., Roose, M. L., Moscou, M. J., Chao, S., Varshney, R. K., Szucs, P., Sato, K., Hayes, P. M., Matthews, D. E., Kleinhofs, A., Muehlbauer, G. J., DeYoung, J., Marshall, D. F., Madishetty, K., Fenton, R. D., Condamine, P., Graner, A. and Waugh, R. (2009). "Development and implementation of high-throughput SNP genotyping in barley." BMC Genomics **10**: 582.

Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L. and Rice, P. M. (2010). "The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants." Nucleic Acids Res **38**(6): 1767-1771.

Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C. D., Pradhan, S., Nelson, S. F., Pellegrini, M. and Jacobsen, S. E. (2008). "Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning." Nature **452**(7184): 215-219.

Colbert, T., Till, B. J., Tompa, R., Reynolds, S., Steine, M. N., Yeung, A. T., McCallum, C. M., Comai, L. and Henikoff, S. (2001). "High-throughput screening for induced point mutations." Plant Physiol **126**(2): 480-484.

Conesa, A., Gotz, S., Garcia-Gomez, J. M., Terol, J., Talon, M. and Robles, M. (2005). "Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research." Bioinformatics **21**(18): 3674-3676.

Coulson, A., Sulston, J., Brenner, S. and Karn, J. (1986). "Toward a physical map of the genome of the nematode Caenorhabditis elegans." Proc Natl Acad Sci U S A **83**(20): 7821-7825.

Creighton, H. B. and McClintock, B. (1931). "A Correlation of Cytological and Genetical Crossing-Over in Zea Mays." Proc Natl Acad Sci U S A **17**(8): 492-497.

Crepet, W. L. and Feldman, G. D. (1991). "The Earliest Remains of Grasses in the Fossil Record." American Journal of Botany **78**(7): 1010-1014.

Cresse, A. D., Hulbert, S. H., Brown, W. E., Lucas, J. R. and Bennetzen, J. L. (1995). "Mu1-related transposable elements of maize preferentially insert into low copy number DNA." Genetics **140**(1): 315-324.

Cronquist, A. (1988). The evolution and classification of flowering plants, New York Botanical Garden, Bronx, NY.

De Roure, D., Goble, C. and Stevens, R. (2007). "Designing the (my)Experiment Virtual Research Environment for the social sharing of workflows." E-Science 2007: Third Ieee International Conference on E-Science and Grid Computing, Proceedings: 603-610
636.

Delcher, A. L., Phillippy, A., Carlton, J. and Salzberg, S. L. (2002). "Fast algorithms for large-scale genome alignment and comparison." Nucleic Acids Res **30**(11): 2478-2483.

Dellagi, A., Rigault, M., Segond, D., Roux, C., Kraepiel, Y., Cellier, F., Briat, J. F., Gaymard, F. and Expert, D. (2005). "Siderophore-mediated upregulation of Arabidopsis ferritin expression in response to Erwinia chrysanthemi infection." Plant J **43**(2): 262-272.

Deshpande, V. G. and Ranjekar, P. K. (1980). "Repetitive DNA in three Gramineae species with low DNA content." Hoppe Seylers Z Physiol Chem **361**(8): 1223-1233.

Devos, K. M. (2005). "Updating the 'crop circle'." Curr Opin Plant Biol **8**(2): 155-162.

Devos, K. M. (2010). "Grass genome organization and evolution." Curr Opin Plant Biol **13**(2): 139-145.

Devos, K. M., Beales, J., Nagamura, Y. and Sasaki, T. (1999). "Arabidopsis-rice: will colinearity allow gene prediction across the eudicot-monocot divide?" Genome Res **9**(9): 825-829.

Devos, K. M., Brown, J. K. and Bennetzen, J. L. (2002). "Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis." Genome Res **12**(7): 1075-1079.

Devos, K. M., Chao, S., Li, Q. Y., Simonetti, M. C. and Gale, M. D. (1994). "Relationship between Chromosome-9 of Maize and Wheat Homeologous Group-7 Chromosomes." Genetics **138**(4): 1287-1292.

Devos, K. M. and Gale, M. D. (1997). "Comparative genetics in the grasses." Plant Mol Biol **35**(1-2): 3-15.

Devos, K. M., Ma, J., Pontaroli, A. C., Pratt, L. H. and Bennetzen, J. L. (2005). "Analysis and mapping of randomly chosen bacterial artificial chromosome clones from hexaploid bread wheat." Proc Natl Acad Sci U S A **102**(52): 19243-19248.

Ding, Y., Johnson, M. D., Colayco, R., Chen, Y. J., Melnyk, J., Schmitt, H. and Shizuya, H. (1999). "Contig assembly of bacterial artificial chromosome clones through multiplexed fluorescence-labeled fingerprinting." Genomics **56**(3): 237-246.

DOE-JGI. (2006). "Energy-rich Portfolio of New Genome Sequencing Targets for DOE JGI." from http://www.jgi.doe.gov/News/news_7_11_06.html.

DOE. (2006). "Breaking the Biological Barriers to Cellulosic Ethanol. A Joint Research Agenda.", from http://genomicscience.energy.gov/biofuels/2005workshop/b2blowres63006.pdf.

Doebley, J. F., Gaut, B. S. and Smith, B. D. (2006). "The molecular genetics of crop domestication." Cell **127**(7): 1309-1321.

Dolezel, J. (2008). Flow Cytometric Analysis and Sorting of Mitotic Chromosomes in Wheat. TriticeaeGenome kickoff meeting. INRA, Clermont-Ferrand, France.

Dolezel, J., Greilhuber, J., Lucretti, S., Meister, A., Lysak, M. A., Nardi, L. and Obermayer, R. (1998). "Plant genome size estimation by flow cytometry: Inter-laboratory comparison." Annals of Botany **82**: 17-26.

Dolezel, J., Kubalakova, M., Paux, E., Bartos, J. and Feuillet, C. (2007). "Chromosome-based genomics in the cereals." Chromosome Res **15**(1): 51-66.

Draper, J., Mur, L. A., Jenkins, G., Ghosh-Biswas, G. C., Bablak, P., Hasterok, R. and Routledge, A. P. (2001). "Brachypodium distachyon. A new model system for functional genomics in grasses." Plant Physiol **127**(4): 1539-1555.

Dubchak, I. and Ryaboy, D. V. (2006). "VISTA family of computational tools for comparative analysis of DNA sequences and whole genomes." Methods Mol Biol **338**: 69-89.

Dubcovsky, J. and Dvorak, J. (2007). "Genome plasticity a key factor in the success of polyploid wheat under domestication." Science **316**(5833): 1862-1866.

Duran, C., Boskovic, Z., Imelfort, M., Batley, J., Hamilton, N. A. and Edwards, D. (2010). "CMap3D: a 3D visualization tool for comparative genetic maps." Bioinformatics **26**(2): 273-274.

Durbin, R. and Haussler, D. (2010). "GFF (General Feature Format) specifications document." from http://www.sanger.ac.uk/resources/software/gff/spec.html.

Dvorak, J., Terlizzi, P., Zhang, H. B. and Resta, P. (1993). "The evolution of polyploid wheats: identification of the A genome donor species." Genome **36**(1): 21-31.

Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic Acids Res **32**(5): 1792-1797.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J. and Turner, S. (2009). "Real-time DNA sequencing from single polymerase molecules." Science **323**(5910): 133-138.

Emberton, J., Ma, J., Yuan, Y., SanMiguel, P. and Bennetzen, J. L. (2005). "Gene enrichment in maize with hypomethylated partial restriction (HMPR) libraries." Genome Res **15**(10): 1441-1446.

Endo, T. R. and Gill, B. S. (1996). "The deletion stocks of common wheat." Journal of Heredity **87**(4): 295-307.

Ermolaeva, M. D., Wu, M., Eisen, J. A. and Salzberg, S. L. (2003). "The age of the Arabidopsis thaliana genome duplication." Plant Mol Biol **51**(6): 859-866.

Ewing, B. and Green, P. (1998). "Base-calling of automated sequencer traces using phred. II. Error probabilities." Genome Res **8**(3): 186-194.

Ewing, B., Hillier, L., Wendl, M. C. and Green, P. (1998). "Base-calling of automated sequencer traces using phred. I. Accuracy assessment." Genome Res **8**(3): 175-185.

FAOStat. (2010). "World Wheat Production."   Retrieved 20/08/2010, 2010, from http://faostat.fao.org.

Faris, J. D., Fellers, J. P., Brooks, S. A. and Gill, B. S. (2003). "A bacterial artificial chromosome contig spanning the major domestication locus Q in wheat and identification of a candidate gene." Genetics **164**(1): 311-321.

Febrer, M., Goicoechea, J. L., Wright, J., McKenzie, N., Song, X., Lin, J., Collura, K., Wissotski, M., Yu, Y., Ammiraju, J. S., Wolny, E., Idziak, D., Betekhtin, A., Kudrna, D., Hasterok, R., Wing, R. A. and Bevan, M. W. (2010). "An integrated physical, genetic and cytogenetic map of Brachypodium distachyon, a model system for grass research." PLoS ONE **5**(10): e13461.

Febrer, M., Wilhelm, E., Al-Kaff, N., Wright, J., Powell, W., Bevan, M. W. and Boulton, M. I. (2009). "Rapid identification of the three homoeologues of the wheat dwarfing gene Rht using a novel PCR-based screen of three-dimensional BAC pools." Genome **52**(12): 993-1000.

Fedoroff, N. V., Furtek, D. B. and Nelson, O. E. (1984). "Cloning of the bronze locus in maize by a simple and generalizable procedure using the transposable controlling element Activator (Ac)." Proc Natl Acad Sci U S A **81**(12): 3825-3829.

Feldman, M. and Levy, A. A. (2009). "Genome evolution in allopolyploid wheat--a revolutionary reprogramming followed by gradual changes." J Genet Genomics **36**(9): 511-518.

Feltus, F. A., Wan, J., Schulze, S. R., Estill, J. C., Jiang, N. and Paterson, A. H. (2004). "An SNP resource for rice genetics and breeding based on subspecies indica and japonica genome alignments." Genome Res **14**(9): 1812-1819.

Feng, Q., Zhang, Y., Hao, P., Wang, S., Fu, G., Huang, Y., Li, Y., Zhu, J., Liu, Y., Hu, X., Jia, P., Zhang, Y., Zhao, Q., Ying, K., Yu, S., Tang, Y., Weng, Q., Zhang, L., Lu, Y., Mu, J., Lu, Y., Zhang, L. S., Yu, Z., Fan, D., Liu, X., Lu, T., Li, C., Wu, Y., Sun, T., Lei, H., Li, T., Hu, H., Guan, J., Wu, M., Zhang, R., Zhou, B., Chen, Z., Chen, L., Jin, Z., Wang, R., Yin, H., Cai, Z., Ren, S., Lv, G., Gu, W., Zhu, G., Tu, Y., Jia, J., Zhang, Y., Chen, J., Kang, H., Chen, X., Shao, C., Sun, Y., Hu, Q., Zhang, X., Zhang, W., Wang, L., Ding, C., Sheng, H., Gu, J., Chen, S., Ni, L., Zhu, F., Chen, W., Lan, L., Lai, Y., Cheng, Z., Gu, M., Jiang, J., Li, J., Hong, G., Xue, Y. and Han, B. (2002). "Sequence and analysis of rice chromosome 4." Nature **420**(6913): 316-320.

Feuillet, C. and Keller, B. (1999). "High gene density is conserved at syntenic loci of small and large grass genomes." Proceedings of the National Academy of Sciences of the United States of America **96**(14): 8265-8270.

Finnegan, D. J. (1989). "Eukaryotic transposable elements and genome evolution." Trends Genet **5**(4): 103-107.

Fitch, W. M. (1970). "Distinguishing homologous from analogous proteins." Syst Zool **19**(2): 99-113.

Flavell, A. J., Pearce, S. R. and Kumar, A. (1994). "Plant transposable elements and the genome." Curr Opin Genet Dev **4**(6): 838-844.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M. and et al. (1995). "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd." Science **269**(5223): 496-512.

Fleury, D., Luo, M. C., Dvorak, J., Ramsay, L., Gill, B. S., Anderson, O. D., You, F. M., Shoaei, Z., Deal, K. R. and Langridge, P. (2010). "Physical mapping of a large plant genome using global high-information-content-fingerprinting: the distal region of the wheat ancestor Aegilops tauschii chromosome 3DS." BMC Genomics **11**(1): 382.

Foote, T. N., Griffiths, S., Allouis, S. and Moore, G. (2004). "Construction and analysis of a BAC library in the grass Brachypodium sylvaticum: its use as a tool to bridge the gap between rice and wheat in elucidating gene content." Funct Integr Genomics **4**(1): 26-33.

Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M. and Dubchak, I. (2004). "VISTA: computational tools for comparative genomics." Nucleic Acids Res **32**(Web Server issue): W273-279.

Freeling, M., Rapaka, L., Lyons, E., Pedersen, B. and Thomas, B. C. (2007). "G-boxes, bigfoot genes, and environmental response: characterization of intragenomic conserved noncoding sequences in Arabidopsis." Plant Cell **19**(5): 1441-1457.

Freeling, M. and Subramaniam, S. (2009). "Conserved noncoding sequences (CNSs) in higher plants." Curr Opin Plant Biol **12**(2): 126-132.

Fu, H. and Dooner, H. K. (2002). "Intraspecific violation of genetic colinearity and its implications in maize." Proc Natl Acad Sci U S A **99**(14): 9573-9578.

Galas, D. J. and Schmitz, A. (1978). "DNAse footprinting: a simple method for the detection of protein-DNA binding specificity." Nucleic Acids Res **5**(9): 3157-3170.

Gale, M. D. and Devos, K. M. (1998). "Plant comparative genetics after 10 years." Science **282**(5389): 656-659.

Garner, M. M. and Revzin, A. (1981). "A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system." Nucleic Acids Res **9**(13): 3047-3060.

Garvin, D. F. (2007). "Brachypodium: a new monocot model plant system emerges." Journal of the Science of Food and Agriculture **87**(7): 1177-1179.

Garvin, D. F., Gu, Y. Q., Hasterok, R., Hazen, S. P., Jenkins, G., Mockler, T. C., Mur, L. A. J. and Vogel, J. P. (2008). "Development of genetic and genomic research resources for Brachypodium distachyon, a new model system for grass crop research." Crop Science **48**: S69-S84.

Garvin, D. F., McKenzie, N., Vogel, J. P., Mockler, T. C., Blankenheim, Z. J., Wright, J., Cheema, J. J., Dicks, J., Huo, N., Hayden, D. M., Gu, Y., Tobias, C., Chang, J. H., Chu, A., Trick, M., Michael, T. P., Bevan, M. W. and Snape, J. W. (2010). "An SSR-based genetic linkage map of the model grass Brachypodium distachyon." Genome **53**(1): 1-13.

Gaut, B. S. (2002). "Evolutionary dynamics of grass genomes." New Phytologist **154**(1): 15-28.

Gaut, B. S. and Doebley, J. F. (1997). "DNA sequence evidence for the segmental allotetraploid origin of maize." Proc Natl Acad Sci U S A **94**(13): 6809-6814.

Gelvin, S. B. (2003). "Agrobacterium-mediated plant transformation: the biology behind the "gene-jockeying" tool." Microbiol Mol Biol Rev **67**(1): 16-37, table of contents.

Gene Ontology Consortium (2006). "The Gene Ontology (GO) project in 2006." Nucleic Acids Res **34**(Database issue): D322-326.

Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., Andre, B., Arkin, A. P., Astromoff, A., El-Bakkoury, M., Bangham, R., Benito, R., Brachat, S., Campanaro, S., Curtiss, M., Davis, K., Deutschbauer, A., Entian, K. D., Flaherty, P., Foury, F., Garfinkel, D. J., Gerstein, M., Gotte, D., Guldener, U., Hegemann, J. H., Hempel, S., Herman, Z., Jaramillo, D. F., Kelly, D. E., Kelly, S. L., Kotter, P., LaBonte, D., Lamb, D. C., Lan, N., Liang, H., Liao, H., Liu, L., Luo, C., Lussier, M., Mao, R., Menard, P., Ooi, S. L., Revuelta, J. L., Roberts, C. J., Rose, M., Ross-Macdonald, P., Scherens, B., Schimmack, G., Shafer, B., Shoemaker, D. D., Sookhai-Mahadeo, S., Storms, R. K., Strathern, J. N., Valle, G., Voet, M., Volckaert, G., Wang, C. Y., Ward, T. R., Wilhelmy, J., Winzeler, E. A., Yang, Y., Yen, G., Youngman, E., Yu, K., Bussey, H., Boeke, J. D., Snyder, M., Philippsen, P., Davis, R. W. and Johnston, M. (2002). "Functional profiling of the Saccharomyces cerevisiae genome." Nature **418**(6896): 387-391.

Gilchrist, E. J., O'Neil, N. J., Rose, A. M., Zetka, M. C. and Haughn, G. W. (2006). "TILLING is an effective reverse genetics technique for Caenorhabditis elegans." BMC Genomics **7**: 262.

Gill, K. S., Gill, B. S., Endo, T. R. and Boyko, E. V. (1996). "Identification and high-density mapping of gene-rich regions in chromosome group 5 of wheat." Genetics **143**(2): 1001-1012.

Gish, W. (1996-2006). from http://blast.wustl.edu.

Goff, S. A., Ricke, D., Lan, T. H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., Hadley, D., Hutchison, D., Martin, C., Katagiri, F., Lange, B. M., Moughamer, T., Xia, Y., Budworth, P., Zhong, J., Miguel, T., Paszkowski, U., Zhang, S., Colbert, M., Sun, W. L., Chen, L., Cooper, B., Park, S., Wood, T. C., Mao, L., Quail, P., Wing, R., Dean, R., Yu, Y., Zharkikh, A., Shen, R., Sahasrabudhe, S., Thomas, A., Cannings, R., Gutin, A., Pruss, D., Reid, J., Tavtigian, S., Mitchell, J., Eldredge, G., Scholl, T., Miller, R. M., Bhatnagar, S., Adey, N., Rubano, T., Tusneem, N., Robinson, R., Feldhaus, J., Macalma, T., Oliphant, A. and Briggs, S. (2002). "A draft sequence of the rice genome (Oryza sativa L. ssp. japonica)." Science **296**(5565): 92-100.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H. and Oliver, S. G. (1996). "Life with 6000 genes." Science **274**(5287): 546, 563-547.

Gomez, L. D., Bristow, J. K., Statham, E. R. and McQueen-Mason, S. J. (2008). "Analysis of saccharification in Brachypodium distachyon stems under mild conditions of hydrolysis." Biotechnol Biofuels **1**(1): 15.

Gordon, D., Abajian, C. and Green, P. (1998). "Consed: a graphical tool for sequence finishing." Genome Res **8**(3): 195-202.

Green, P. (1997). "Against a whole-genome shotgun." Genome Res **7**(5): 410-417.

Gregory, S. G., Howell, G. R. and Bentley, D. R. (1997). "Genome mapping by fluorescent fingerprinting." Genome Res **7**(12): 1162-1168.

Gremme, G., Brendel, V., Sparks, M. E. and Kurtz, S. (2005). "Engineering a software tool for gene structure prediction in higher organisms." Information and Software Technology **47**(15): 965-978.

Griffiths, S., Sharp, R., Foote, T. N., Bertin, I., Wanous, M., Reader, S., Colas, I. and Moore, G. (2006). "Molecular characterization of Ph1 as a major chromosome pairing locus in polyploid wheat." Nature **439**(7077): 749-752.

Grivet, L., Dhont, A., Dufour, P., Hamon, P., Roques, D. and Glaszmann, J. C. (1994). "Comparative Genome Mapping of Sugar-Cane with Other Species within the Andropogoneae Tribe." Heredity **73**: 500-508.

Gu, Y. Q., Coleman-Derr, D., Kong, X. and Anderson, O. D. (2004). "Rapid genome evolution revealed by comparative sequence analysis of orthologous regions from four triticeae genomes." Plant Physiol **135**(1): 459-470.

Gu, Y. Q., Ma, Y., Huo, N., Vogel, J. P., You, F. M., Lazo, G. R., Nelson, W. M., Soderlund, C., Dvorak, J., Anderson, O. D. and Luo, M. C. (2009). "A BAC-based physical map of Brachypodium distachyon and its comparative analysis with rice and wheat." BMC Genomics **10**: 496.

Gu, Y. Q., Salse, J., Coleman-Derr, D., Dupin, A., Crossman, C., Lazo, G. R., Huo, N., Belcram, H., Ravel, C., Charmet, G., Charles, M., Anderson, O. D. and Chalhoub, B. (2006). "Types and rates of sequence evolution at the high-molecular-weight glutenin locus in hexaploid wheat and its ancestral genomes." Genetics **174**(3): 1493-1504.

Guo, H. and Moose, S. P. (2003). "Conserved noncoding sequences among cultivated cereal genomes identify candidate regulatory sequence elements and patterns of promoter evolution." Plant Cell **15**(5): 1143-1158.

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith, R. K., Jr., Hannick, L. I., Maiti, R., Ronning, C. M., Rusch, D. B., Town, C. D., Salzberg, S. L. and White, O. (2003). "Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies." Nucleic Acids Res **31**(19): 5654-5666.

Haas, B. J., Delcher, A. L., Wortman, J. R. and Salzberg, S. L. (2004). "DAGchainer: a tool for mining segmental genome duplications and synteny." Bioinformatics **20**(18): 3643-3646.

Haberer, G., Young, S., Bharti, A. K., Gundlach, H., Raymond, C., Fuks, G., Butler, E., Wing, R. A., Rounsley, S., Birren, B., Nusbaum, C., Mayer, K. F. and Messing, J. (2005). "Structure and architecture of the maize genome." Plant Physiol **139**(4): 1612-1624.

Hardison, R. C., Oeltjen, J. and Miller, W. (1997). "Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome." Genome Res **7**(10): 959-966.

Harris, T. D., Buzby, P. R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., Causey, M., Colonell, J., Dimeo, J., Efcavitch, J. W., Giladi, E., Gill, J., Healy, J., Jarosz, M., Lapen, D., Moulton, K., Quake, S. R., Steinmann, K., Thayer, E., Tyurina, A., Ward, R., Weiss, H. and Xie, Z. (2008). "Single-molecule DNA sequencing of a viral genome." Science **320**(5872): 106-109.

Harris, T. W., Antoshechkin, I., Bieri, T., Blasiar, D., Chan, J., Chen, W. J., De La Cruz, N., Davis, P., Duesbury, M., Fang, R., Fernandes, J., Han, M., Kishore, R., Lee, R., Muller, H. M., Nakamura, C., Ozersky, P., Petcherski, A., Rangarajan, A., Rogers, A., Schindelman, G., Schwarz, E. M., Tuli, M. A., Van Auken, K., Wang, D., Wang, X., Williams, G., Yook, K., Durbin, R., Stein, L. D., Spieth, J. and Sternberg, P. W. (2010). "WormBase: a comprehensive resource for nematode research." Nucleic Acids Res **38**(Database issue): D463-467.

Harrison, P. M. and Gerstein, M. (2002). "Studying genomes through the aeons: protein families, pseudogenes and proteome evolution." J Mol Biol **318**(5): 1155-1174.

Hasterok, R., Marasek, A., Donnison, I. S., Armstead, I., Thomas, A., King, I. P., Wolny, E., Idziak, D., Draper, J. and Jenkins, G. (2006). "Alignment of the genomes of Brachypodium distachyon and temperate cereals and grasses using bacterial artificial chromosome landing with fluorescence in situ hybridization." Genetics **173**(1): 349-362.

Heun, M., SchaferPregl, R., Klawan, D., Castagna, R., Accerbi, M., Borghi, B. and Salamini, F. (1997). "Site of einkorn wheat domestication identified by DNA fingerprinting." Science **278**(5341): 1312-1314.

Hirochika, H., Guiderdoni, E., An, G., Hsing, Y. I., Eun, M. Y., Han, C. D., Upadhyaya, N., Ramachandran, S., Zhang, Q., Pereira, A., Sundaresan, V. and Leung, H. (2004). "Rice mutant resources for gene discovery." Plant Mol Biol **54**(3): 325-334.

Holliday, R. (1964). "Mechanism for Gene Conversion in Fungi." Genetical Research **5**(2): 282-&.

Huang, C. Y., Ayliffe, M. A. and Timmis, J. N. (2003). "Direct measurement of the transfer rate of chloroplast DNA into the nucleus." Nature **422**(6927): 72-76.

Huang, S., Sirikhachornkit, A., Su, X., Faris, J., Gill, B., Haselkorn, R. and Gornicki, P. (2002). "Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the Triticum/Aegilops complex and the evolutionary history of polyploid wheat." Proc Natl Acad Sci U S A **99**(12): 8133-8138.

Huang, W. and Marth, G. (2008). "EagleView: a genome assembly viewer for next-generation sequencing technologies." Genome Res **18**(9): 1538-1543.

Huang, X. and Madan, A. (1999). "CAP3: A DNA sequence assembly program." Genome Res **9**(9): 868-877.

Hubbard, T. J., Aken, B. L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T., Dyer, S. C., Fitzgerald, S., Fernandez-Banet, J., Graf, S., Haider, S., Hammond, M., Herrero, J., Holland, R., Howe, K., Howe, K., Johnson, N., Kahari, A., Keefe, D., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Melsopp, C., Megy, K., Meidl, P., Ouverdin, B., Parker, A., Prlic, A., Rice, S., Rios, D., Schuster, M., Sealy, I., Severin, J., Slater, G., Smedley, D., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wood, M., Cox, T., Curwen, V., Durbin, R., Fernandez-Suarez, X. M., Flicek, P., Kasprzyk, A., Proctor, G., Searle, S., Smith, J., Ureta-Vidal, A. and Birney, E. (2007). "Ensembl 2007." Nucleic Acids Res **35**(Database issue): D610-617.

Hull, D., Wolstencroft, K., Stevens, R., Goble, C., Pocock, M. R., Li, P. and Oinn, T. (2006). "Taverna: a tool for building and running workflows of services." Nucleic Acids Res **34**(Web Server issue): W729-732.

Huo, N., Gu, Y. Q., Lazo, G. R., Vogel, J. P., Coleman-Derr, D., Luo, M. C., Thilmony, R., Garvin, D. F. and Anderson, O. D. (2006). "Construction and characterization of two BAC libraries from Brachypodium distachyon, a new model for grass genomics." Genome **49**(9): 1099-1108.

Huo, N., Lazo, G. R., Vogel, J. P., You, F. M., Ma, Y., Hayden, D. M., Coleman-Derr, D., Hill, T. A., Dvorak, J., Anderson, O. D., Luo, M. C. and Gu, Y. Q. (2008). "The nuclear genome of Brachypodium distachyon: analysis of BAC end sequences." Funct Integr Genomics **8**(2): 135-147.

Idury, R. M. and Waterman, M. S. (1995). "A new algorithm for DNA sequence assembly." J Comput Biol **2**(2): 291-306.

Ilic, K., SanMiguel, P. J. and Bennetzen, J. L. (2003). "A complex history of rearrangement in an orthologous region of the maize, sorghum, and rice genomes." Proc Natl Acad Sci U S A **100**(21): 12265-12270.

Inada, D. C., Bashir, A., Lee, C., Thomas, B. C., Ko, C., Goff, S. A. and Freeling, M. (2003). "Conserved noncoding sequences in the grasses." Genome Res **13**(9): 2030-2041.

International Brachypodium Initiative (2010). "Genome sequencing and analysis of the model grass Brachypodium distachyon." Nature **463**(7282): 763-768.

International Rice Genome Sequencing Project (2005). "The map-based sequence of the rice genome." Nature **436**(7052): 793-800.

Ishikawa, G., Yonemaru, J., Saito, M. and Nakamura, T. (2007). "PCR-based landmark unique gene (PLUG) markers effectively assign homoeologous wheat genes to A, B and D genomes." BMC Genomics **8**: 135.

Isidore, E., Scherrer, B., Chalhoub, B., Feuillet, C. and Keller, B. (2005). "Ancient haplotypes resulting from extensive molecular rearrangements in the wheat A genome have been maintained in species of three different ploidy levels." Genome Res **15**(4): 526-536.

Ito, T., Seki, M., Hayashida, N., Shibata, D. and Shinozaki, K. (1999). "Regional insertional mutagenesis of genes on Arabidopsis thaliana chromosome V using the Ac/Ds transposon in combination with a cDNA scanning method." Plant J **17**(4): 433-444.

Itoh, T., Tanaka, T., Barrero, R. A., Yamasaki, C., Fujii, Y., Hilton, P. B., Antonio, B. A., Aono, H., Apweiler, R., Bruskiewich, R., Bureau, T., Burr, F., Costa de Oliveira, A., Fuks, G., Habara, T., Haberer, G., Han, B., Harada, E., Hiraki, A. T., Hirochika, H., Hoen, D., Hokari, H., Hosokawa, S., Hsing, Y. I., Ikawa, H., Ikeo, K., Imanishi, T., Ito, Y., Jaiswal, P., Kanno, M., Kawahara, Y., Kawamura, T., Kawashima, H., Khurana, J. P., Kikuchi, S., Komatsu, S., Koyanagi, K. O., Kubooka, H., Lieberherr, D., Lin, Y. C., Lonsdale, D., Matsumoto, T., Matsuya, A., McCombie, W. R., Messing, J., Miyao, A., Mulder, N., Nagamura, Y., Nam, J., Namiki, N., Numa, H., Nurimoto, S., O'Donovan, C., Ohyanagi, H., Okido, T., Oota, S., Osato, N., Palmer, L. E., Quetier, F., Raghuvanshi, S., Saichi, N., Sakai, H., Sakai, Y., Sakata, K., Sakurai, T., Sato, F., Sato, Y., Schoof, H., Seki, M., Shibata, M., Shimizu, Y., Shinozaki, K., Shinso, Y., Singh, N. K., Smith-White, B., Takeda, J., Tanino, M., Tatusova, T., Thongjuea, S., Todokoro, F., Tsugane, M., Tyagi, A. K., Vanavichit, A., Wang, A., Wing, R. A., Yamaguchi, K., Yamamoto, M., Yamamoto, N., Yu, Y., Zhang, H., Zhao, Q., Higo, K., Burr, B., Gojobori, T. and Sasaki, T. (2007). "Curated genome annotation of Oryza sativa ssp. japonica and comparative genome analysis with Arabidopsis thaliana." Genome Res **17**(2): 175-183.

Jacq, C., Miller, J. R. and Brownlee, G. G. (1977). "A pseudogene structure in 5S DNA of Xenopus laevis." Cell **12**(1): 109-120.

Jaiswal, P., Ni, J., Yap, I., Ware, D., Spooner, W., Youens-Clark, K., Ren, L., Liang, C., Zhao, W., Ratnapu, K., Faga, B., Canaran, P., Fogleman, M., Hebbard, C., Avraham, S., Schmidt, S., Casstevens, T. M., Buckler, E. S., Stein, L. and McCouch, S. (2006). "Gramene: a bird's eye view of cereal genomes." Nucleic Acids Res **34**(Database issue): D717-723.

Jenkins, G. and Hasterok, R. (2007). "BAC 'landing' on chromosomes of Brachypodium distachyon for comparative genome alignment." Nat Protoc **2**(1): 88-98.

Jeong, D. H., An, S., Park, S., Kang, H. G., Park, G. G., Kim, S. R., Sim, J., Kim, Y. O., Kim, M. K., Kim, J., Shin, M., Jung, M. and An, G. (2006). "Generation of a flanking sequence-tag database for activation-tagging lines in japonica rice." Plant J **45**(1): 123-132.

Jiang, N., Bao, Z., Zhang, X., Eddy, S. R. and Wessler, S. R. (2004). "Pack-MULE transposable elements mediate gene evolution in plants." Nature **431**(7008): 569-573.

JIC Press Office. (2010). "UK researchers release draft sequence coverage of wheat genome.", from http://www.jic.ac.uk/corporate/media-and-public/current-releases/100827wheatgenome.htm.

Johnson, D. S., Mortazavi, A., Myers, R. M. and Wold, B. (2007). "Genome-wide mapping of in vivo protein-DNA interactions." Science **316**(5830): 1497-1502.

Jukes, T. H. and Kimura, M. (1984). "Evolutionary constraints and the neutral theory." J Mol Evol **21**(1): 90-92.

Juretic, N., Hoen, D. R., Huynh, M. L., Harrison, P. M. and Bureau, T. E. (2005). "The evolutionary fate of MULE-mediated duplications of host gene fragments in rice." Genome Res **15**(9): 1292-1297.

Kadonaga, J. T. (2004). "Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors." Cell **116**(2): 247-257.

Kalendar, R., Vicient, C. M., Peleg, O., Anamthawat-Jonsson, K., Bolshoy, A. and Schulman, A. H. (2004). "Large retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes." Genetics **166**(3): 1437-1450.

Kapitonov, V. V. and Jurka, J. (2001). "Rolling-circle transposons in eukaryotes." Proc Natl Acad Sci U S A **98**(15): 8714-8719.

Kaplinsky, N. J., Braun, D. M., Penterman, J., Goff, S. A. and Freeling, M. (2002). "Utility and distribution of conserved noncoding sequences in the grasses." Proc Natl Acad Sci U S A **99**(9): 6147-6151.

Kashkush, K., Feldman, M. and Levy, A. A. (2002). "Gene loss, silencing and activation in a newly synthesized wheat allotetraploid." Genetics **160**(4): 1651-1659.

Katagiri, S., Wu, J. Z., Ito, Y. K., Karasawa, W., Shibata, M., Kanamori, H., Katayose, Y., Namiki, N., Matsumoto, T. and Sasaki, T. (2004). "End Sequencing and chromosomal in silico mapping of BAC clones derived from an indica rice cultivar, Kasalath." Breeding Science **54**(3): 273-279.

Katari, M. S., Balija, V., Wilson, R. K., Martienssen, R. A. and McCombie, W. R. (2005). "Comparing low coverage random shotgun sequence data from Brassica oleracea and Oryza sativa genome sequence for their ability to add to the annotation of Arabidopsis thaliana." Genome Res **15**(4): 496-504.

Kececioglu, J. D. and Myers, E. W. (1992). "Combinatorial algorithms for DNA sequence assembly." Algorithmica **13**(1-2): 7-51.

Kellogg, E. A. (2001). "Evolutionary history of the grasses." Plant Physiol **125**(3): 1198-1205.

Kellogg, E. A. and Bennetzen, J. L. (2004). "The evolution of nuclear genome structure in seed plants." American Journal of Botany **91**(10): 1709-1725.

Kent, W. J. (2002). "BLAT--the BLAST-like alignment tool." Genome Res **12**(4): 656-664.

Kikuchi, S., Satoh, K., Nagata, T., Kawagashira, N., Doi, K., Kishimoto, N., Yazaki, J., Ishikawa, M., Yamada, H., Ooka, H., Hotta, I., Kojima, K., Namiki, T., Ohneda, E., Yahagi, W., Suzuki, K., Li, C. J., Ohtsuki, K., Shishiki, T., Otomo, Y., Murakami, K., Iida, Y., Sugano, S., Fujimura, T., Suzuki, Y., Tsunoda, Y., Kurosaki, T., Kodama, T., Masuda, H., Kobayashi, M., Xie, Q., Lu, M., Narikawa, R., Sugiyama, A., Mizuno, K., Yokomizo, S., Niikura, J., Ikeda, R., Ishibiki, J., Kawamata, M., Yoshimura, A., Miura, J., Kusumegi, T., Oka, M., Ryu, R., Ueda, M., Matsubara, K., Kawai, J., Carninci, P., Adachi, J., Aizawa, K., Arakawa, T., Fukuda, S., Hara, A., Hashizume, W., Hayatsu, N., Imotani, K., Ishii, Y., Itoh, M., Kagawa, I., Kondo, S., Konno, H., Miyazaki, A., Osato, N., Ota, Y., Saito, R., Sasaki, D., Sato, K., Shibata, K., Shinagawa, A., Shiraki, T., Yoshino, M., Hayashizaki, Y. and Yasunishi, A. (2003). "Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice." Science **301**(5631): 376-379.

Kimura, M. (1969). "The rate of molecular evolution considered from the standpoint of population genetics." Proc Natl Acad Sci U S A **63**(4): 1181-1188.

Kimura, M. (1983). The Neutral Theory of Molecular Evolution, Cambridge University Press, Cambridge, UK.

Kimura, M. and Ota, T. (1974). "On some principles governing molecular evolution." Proc Natl Acad Sci U S A **71**(7): 2848-2852.

Kirik, A., Salomon, S. and Puchta, H. (2000). "Species-specific double-strand break repair and genome evolution in plants." Embo J **19**(20): 5562-5566.

Klein, P. E., Klein, R. R., Cartinhour, S. W., Ulanch, P. E., Dong, J., Obert, J. A., Morishige, D. T., Schlueter, S. D., Childs, K. L., Ale, M. and Mullet, J. E. (2000). "A high-throughput AFLP-based method for constructing integrated genetic and physical maps: progress toward a sorghum genome map." Genome Res **10**(6): 789-807.

Korf, I., Flicek, P., Duan, D. and Brent, M. R. (2001). "Integrating genomic homology into gene structure prediction." Bioinformatics **17 Suppl 1**: S140-148.

Kosambi, D. (1944). "The estimation of map distance from recombination values." Ann. Eugen. **12**: 172-175.

Krishnan, A., Guiderdoni, E., An, G., Hsing, Y. I., Han, C. D., Lee, M. C., Yu, S. M., Upadhyaya, N., Ramachandran, S., Zhang, Q., Sundaresan, V., Hirochika, H., Leung, H. and Pereira, A. (2009). "Mutant resources in rice for functional genomics of the grasses." Plant Physiol **149**(1): 165-170.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J. and Marra, M. A. (2009). "Circos: an information aesthetic for comparative genomics." Genome Res **19**(9): 1639-1645.

Kuhn, R. M., Karolchik, D., Zweig, A. S., Trumbower, H., Thomas, D. J., Thakkapallayil, A., Sugnet, C. W., Stanke, M., Smith, K. E., Siepel, A., Rosenbloom, K. R., Rhead, B., Raney, B. J., Pohl, A., Pedersen, J. S., Hsu, F., Hinrichs, A. S., Harte, R. A., Diekhans, M., Clawson, H., Bejerano, G., Barber, G. P., Baertsch, R., Haussler, D. and Kent, W. J. (2007). "The UCSC genome browser database: update 2007." Nucleic Acids Res **35**(Database issue): D668-673.

Kulikova, T., Aldebert, P., Althorpe, N., Baker, W., Bates, K., Browne, P., van den Broek, A., Cochrane, G., Duggan, K., Eberhardt, R., Faruque, N., Garcia-Pastor, M., Harte, N., Kanz, C., Leinonen, R., Lin, Q., Lombard, V., Lopez, R., Mancuso, R., McHale, M., Nardone, F., Silventoinen, V., Stoehr, P., Stoesser, G., Tuli, M. A., Tzouvara, K., Vaughan, R., Wu, D., Zhu, W. and Apweiler, R. (2004). "The EMBL Nucleotide Sequence Database." Nucleic Acids Res **32**(Database issue): D27-30.

Kulp, D., Haussler, D., Reese, M. G. and Eeckman, F. H. (1996). "A generalized hidden Markov model for the recognition of human genes in DNA." Proc Int Conf Intell Syst Mol Biol **4**: 134-142.

Kumar, A. and Bennetzen, J. L. (1999). "Plant retrotransposons." Annu Rev Genet **33**: 479-532.

Kurata, N., Moore, G., Nagamura, Y., Foote, T., Yano, M., Minobe, Y. and Gale, M. (1994). "Conservation of Genome Structure between Rice and Wheat." Bio-Technology **12**(3): 276-278.

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C. and Salzberg, S. L. (2004). "Versatile and open software for comparing large genomes." Genome Biol **5**(2): R12.

La Rota, M. and Sorrells, M. E. (2004). "Comparative DNA sequence analysis of mapped wheat ESTs reveals the complexity of genome relationships between rice and wheat." Funct Integr Genomics **4**(1): 34-46.

Lai, J., Dey, N., Kim, C. S., Bharti, A. K., Rudd, S., Mayer, K. F., Larkins, B. A., Becraft, P. and Messing, J. (2004a). "Characterization of the maize endosperm transcriptome and its comparison to the rice genome." Genome Res **14**(10A): 1932-1937.

Lai, J., Li, Y., Messing, J. and Dooner, H. K. (2005). "Gene movement by Helitron transposons contributes to the haplotype variability of maize." Proc Natl Acad Sci U S A **102**(25): 9068-9073.

Lai, J., Ma, J., Swigonova, Z., Ramakrishna, W., Linton, E., Llaca, V., Tanyolac, B., Park, Y. J., Jeong, O. Y., Bennetzen, J. L. and Messing, J. (2004b). "Gene loss and movement in the maize genome." Genome Res **14**(10A): 1924-1931.

Lamoureux, D., Peterson, D. G., Li, W., Fellers, J. P. and Gill, B. S. (2005). "The efficacy of C_0t-based gene enrichment in wheat (Triticum aestivum L.)." Genome **48**(6): 1120-1126.

Lander, E. S., Green, P., Abrahamson, J., Barlow, A., Daly, M. J., Lincoln, S. E. and Newburg, L. (1987). "MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations." Genomics **1**(2): 174-181.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S. and Chen, Y. J. (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.

Langmead, B., Trapnell, C., Pop, M. and Salzberg, S. L. (2009). "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." Genome Biol **10**(3): R25.

Lazo, G. R., Chao, S., Hummel, D. D., Edwards, H., Crossman, C. C., Lui, N., Matthews, D. E., Carollo, V. L., Hane, D. L., You, F. M., Butler, G. E., Miller, R. E., Close, T. J., Peng, J. H., Lapitan, N. L., Gustafson, J. P., Qi, L. L., Echalier, B., Gill, B. S., Dilbirligi, M., Randhawa, H. S., Gill, K. S., Greene, R. A., Sorrells, M. E., Akhunov, E. D., Dvorak, J., Linkiewicz, A. M., Dubcovsky, J., Hossain, K. G., Kalavacharla, V., Kianian, S. F., Mahmoud, A. A., Miftahudin, Ma, X. F., Conley, E. J., Anderson, J. A., Pathan, M. S., Nguyen, H. T., McGuire, P. E., Qualset, C. O. and Anderson, O. D. (2004). "Development of an expressed sequence tag (EST) resource for wheat (Triticum aestivum L.): EST generation, unigene analysis, probe selection and bioinformatics for a 16,000-locus bin-delineated map." Genetics **168**(2): 585-593.

Le, Q. H., Wright, S., Yu, Z. and Bureau, T. (2000). "Transposon diversity in Arabidopsis thaliana." Proc Natl Acad Sci U S A **97**(13): 7376-7381.

Lenhard, B., Sandelin, A., Mendoza, L., Engstrom, P., Jareborg, N. and Wasserman, W. W. (2003). "Identification of conserved regulatory elements by comparative genome analysis." J Biol **2**(2): 13.

Levene, M. J., Korlach, J., Turner, S. W., Foquet, M., Craighead, H. G. and Webb, W. W. (2003). "Zero-mode waveguides for single-molecule analysis at high concentrations." Science **299**(5607): 682-686.

Levings, C. S., 3rd and Brown, G. G. (1989). "Molecular biology of plant mitochondria." Cell **56**(2): 171-179.

Li, H. and Durbin, R. (2009). "Fast and accurate short read alignment with Burrows-Wheeler transform." Bioinformatics **25**(14): 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009a). "The Sequence Alignment/Map format and SAMtools." Bioinformatics **25**(16): 2078-2079.

Li, H., Ruan, J. and Durbin, R. (2008a). "Mapping short DNA sequencing reads and calling variants using mapping quality scores." Genome Res **18**(11): 1851-1858.

Li, L., Stoeckert, C. J., Jr. and Roos, D. S. (2003). "OrthoMCL: identification of ortholog groups for eukaryotic genomes." Genome Res **13**(9): 2178-2189.

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., Huang, Q., Cai, Q., Li, B., Bai, Y., Zhang, Z., Zhang, Y., Wang, W., Li, J., Wei, F., Li, H., Jian, M., Nielsen, R., Li, D., Gu, W., Yang, Z., Xuan, Z., Ryder, O. A., Leung, F. C., Zhou, Y., Cao, J., Sun, X., Fu, Y., Fang, X., Guo, X., Wang, B., Hou, R., Shen, F., Mu, B., Ni, P., Lin, R., Qian, W., Wang, G., Yu, C., Nie, W., Wang, J., Wu, Z., Liang, H., Min, J., Wu, Q., Cheng, S., Ruan, J., Wang, M., Shi, Z., Wen, M., Liu, B., Ren, X., Zheng, H., Dong, D., Cook, K., Shan, G., Zhang, H., Kosiol, C., Xie, X., Lu, Z., Li, Y., Steiner, C. C., Lam, T. T., Lin, S., Zhang, Q., Li, G., Tian, J., Gong, T., Liu, H., Zhang, D., Fang, L., Ye, C., Zhang, J., Hu, W., Xu, A., Ren, Y., Zhang, G., Bruford, M. W., Li, Q., Ma, L., Guo, Y., An, N., Hu, Y., Zheng, Y., Shi, Y., Li, Z., Liu, Q., Chen, Y., Zhao, J., Qu, N., Zhao, S., Tian, F., Wang, X., Wang, H., Xu, L., Liu, X., Vinar, T., Wang, Y., Lam, T. W., Yiu, S. M., Liu, S., Huang, Y., Yang, G., Jiang, Z., Qin, N., Li, L., Bolund, L., Kristiansen, K., Wong, G. K., Olson, M., Zhang, X., Li, S. and Yang, H. (2010). "The sequence and de novo assembly of the giant panda genome." Nature **463**(7279): 311-317.

Li, R., Li, Y., Kristiansen, K. and Wang, J. (2008b). "SOAP: short oligonucleotide alignment program." Bioinformatics **24**(5): 713-714.

Li, R., Yu, C., Li, Y., Lam, T. W., Yiu, S. M., Kristiansen, K. and Wang, J. (2009b). "SOAP2: an improved ultrafast tool for short read alignment." Bioinformatics **25**(15): 1966-1967.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., Yang, H. and Wang, J. (2009c). "De novo assembly of human genomes with massively parallel short read sequencing." Genome Res **20**(2): 265-272.

Li, W. and Godzik, A. (2006). "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences." Bioinformatics **22**(13): 1658-1659.

Li, W., Zhang, P., Fellers, J. P., Friebe, B. and Gill, B. S. (2004). "Sequence composition, organization, and evolution of the core Triticeae genome." Plant J **40**(4): 500-511.

Li, X., Lassner, M. and Zhang, Y. (2002). "Deleteagene: a fast neutron deletion mutagenesis-based gene knockout system for plants." Comp Funct Genomics **3**(2): 158-160.

Li, X., Tan, L., Wang, L., Hu, S. and Sun, C. (2009d). "Isolation and characterization of conserved non-coding sequences among rice (Oryza sativa L.) paralogous regions." Mol Genet Genomics **281**(1): 11-18.

Lipman, D. J. and Pearson, W. R. (1985). "Rapid and sensitive protein similarity searches." Science **227**(4693): 1435-1441.

Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H. and Ecker, J. R. (2008). "Highly integrated single-base resolution maps of the epigenome in Arabidopsis." Cell **133**(3): 523-536.

Liu, H., Sachidanandam, R. and Stein, L. (2001). "Comparative genomics between rice and Arabidopsis shows scant collinearity in gene order." Genome Res **11**(12): 2020-2026.

Liu, R., Vitte, C., Ma, J., Mahama, A. A., Dhliwayo, T., Lee, M. and Bennetzen, J. L. (2007). "A GeneTrek analysis of the maize genome." Proc Natl Acad Sci U S A **104**(28): 11844-11849.

Lockton, S. and Gaut, B. S. (2005). "Plant conserved non-coding sequences and paralogue evolution." Trends Genet **21**(1): 60-65.

Loguercio, L. L. and Wilkins, T. A. (1998). "Structural analysis of a hmg-coA-reductase pseudogene: insights into evolutionary processes affecting the hmgr gene family in allotetraploid cotton (Gossypium hirsutum L.)." Curr Genet **34**(4): 241-249.

Loots, G. G., Locksley, R. M., Blankespoor, C. M., Wang, Z. E., Miller, W., Rubin, E. M. and Frazer, K. A. (2000). "Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons." Science **288**(5463): 136-140.

Lowe, T. M. and Eddy, S. R. (1997). "tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence." Nucleic Acids Res **25**(5): 955-964.

Luo, M. C., Deal, K. R., Akhunov, E. D., Akhunova, A. R., Anderson, O. D., Anderson, J. A., Blake, N., Clegg, M. T., Coleman-Derr, D., Conley, E. J., Crossman, C. C., Dubcovsky, J., Gill, B. S., Gu, Y. Q., Hadam, J., Heo, H. Y., Huo, N., Lazo, G., Ma, Y., Matthews, D. E., McGuire, P. E., Morrell, P. L., Qualset, C. O., Renfro, J., Tabanao, D., Talbert, L. E., Tian, C., Toleno, D. M., Warburton, M. L., You, F. M., Zhang, W. and Dvorak, J. (2009). "Genome comparisons reveal a dominant mechanism of chromosome number reduction in grasses and accelerated genome evolution in Triticeae." Proc Natl Acad Sci U S A **106**(37): 15780-15785.

Luo, M. C., Ma, Y., You, F. M., Anderson, O. D., Kopecky, D., Simkova, H., Safar, J., Dolezel, J., Gill, B., McGuire, P. E. and Dvorak, J. (2010). "Feasibility of physical map construction from fingerprinted bacterial artificial chromosome libraries of polyploid plant species." BMC Genomics **11**: 122.

Luo, M. C., Thomas, C., You, F. M., Hsiao, J., Ouyang, S., Buell, C. R., Malandro, M., McGuire, P. E., Anderson, O. D. and Dvorak, J. (2003). "High-throughput fingerprinting of bacterial artificial chromosomes using the snapshot labeling kit and sizing of restriction fragments by capillary electrophoresis." Genomics **82**(3): 378-389.

Lyons, E. and Freeling, M. (2008). "How to usefully compare homologous plant genes and chromosomes as DNA sequences." Plant J **53**(4): 661-673.

Lysak, M. A., Fransz, P. F., Ali, H. B. and Schubert, I. (2001). "Chromosome painting in Arabidopsis thaliana." Plant J **28**(6): 689-697.

Ma, J., Devos, K. M. and Bennetzen, J. L. (2004). "Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice." Genome Res **14**(5): 860-869.

Maccallum, I., Przybylski, D., Gnerre, S., Burton, J., Shlyakhter, I., Gnirke, A., Malek, J., McKernan, K., Ranade, S., Shea, T. P., Williams, L., Young, S., Nusbaum, C. and Jaffe, D. B. (2009). "ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads." Genome Biol **10**(10): R103.

Maggert, K. A., Gong, W. J. and Golic, K. G. (2008). "Methods for homologous recombination in Drosophila." Methods Mol Biol **420**: 155-174.

Manly, K. F., Cudmore, R. H., Jr. and Meer, J. M. (2001). "Map Manager QTX, cross-platform software for genetic mapping." Mamm Genome **12**(12): 930-932.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L., Jarvie, T. P., Jirage, K. B., Kim, J. B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F. and Rothberg, J. M. (2005). "Genome sequencing in microfabricated high-density picolitre reactors." Nature **437**(7057): 376-380.

Marra, M. A., Kucaba, T. A., Dietrich, N. L., Green, E. D., Brownstein, B., Wilson, R. K., McDonald, K. M., Hillier, L. W., McPherson, J. D. and Waterston, R. H. (1997). "High throughput fingerprint analysis of large-insert clones." Genome Res **7**(11): 1072-1084.

Martineau, B., Voelker, T. A. and Sanders, R. A. (1994). "On Defining T-DNA." Plant Cell **6**(8): 1032-1033.

Mayer, K., Murphy, G., Tarchini, R., Wambutt, R., Volckaert, G., Pohl, T., Dusterhoft, A., Stiekema, W., Entian, K. D., Terryn, N., Lemcke, K., Haase, D., Hall, C. R., van Dodeweerd, A. M., Tingey, S. V., Mewes, H. W., Bevan, M. W. and Bancroft, I. (2001). "Conservation of microstructure between a sequenced region of the genome of rice and multiple segments of the genome of Arabidopsis thaliana." Genome Res **11**(7): 1167-1174.

Messing, J., Bharti, A. K., Karlowski, W. M., Gundlach, H., Kim, H. R., Yu, Y., Wei, F., Fuks, G., Soderlund, C. A., Mayer, K. F. and Wing, R. A. (2004). "Sequence composition and genome organization of maize." Proc Natl Acad Sci U S A **101**(40): 14349-14354.

Messing, J. and Llaca, V. (1998). "Importance of anchor genomes for any plant genome project." Proc Natl Acad Sci U S A **95**(5): 2017-2020.

Metzker, M. L. (2010). "Sequencing technologies - the next generation." Nat Rev Genet **11**(1): 31-46.

Meyer, I. M. and Durbin, R. (2002). "Comparative ab initio prediction of gene structures using pair HMMs." Bioinformatics **18**(10): 1309-1318.

Meyer, I. M. and Durbin, R. (2004). "Gene structure conservation aids similarity based gene prediction." Nucleic Acids Res **32**(2): 776-783.

Meyers, B. C., Scalabrin, S. and Morgante, M. (2004). "Mapping and sequencing complex genomes: let's get physical!" Nat Rev Genet **5**(8): 578-588.

Meyers, B. C., Tingey, S. V. and Morgante, M. (2001). "Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome." Genome Res **11**(10): 1660-1676.

Mighell, A. J., Smith, N. R., Robinson, P. A. and Markham, A. F. (2000). "Vertebrate pseudogenes." FEBS Lett **468**(2-3): 109-114.

Milne, I., Bayer, M., Cardle, L., Shaw, P., Stephen, G., Wright, F. and Marshall, D. (2010). "Tablet--next generation sequence assembly visualization." Bioinformatics **26**(3): 401-402.

Ming, R., Hou, S., Feng, Y., Yu, Q., Dionne-Laporte, A., Saw, J. H., Senin, P., Wang, W., Ly, B. V., Lewis, K. L., Salzberg, S. L., Feng, L., Jones, M. R., Skelton, R. L., Murray, J. E., Chen, C., Qian, W., Shen, J., Du, P., Eustice, M., Tong, E., Tang, H., Lyons, E., Paull, R. E., Michael, T. P., Wall, K., Rice, D. W., Albert, H., Wang, M. L., Zhu, Y. J., Schatz, M., Nagarajan, N., Acob, R. A., Guan, P., Blas, A., Wai, C. M., Ackerman, C. M., Ren, Y., Liu, C., Wang, J., Na, J. K., Shakirov, E. V., Haas, B., Thimmapuram, J., Nelson, D., Wang, X., Bowers, J. E., Gschwend, A. R., Delcher, A. L., Singh, R., Suzuki, J. Y., Tripathi, S., Neupane, K., Wei, H., Irikura, B., Paidi, M., Jiang, N., Zhang, W., Presting, G., Windsor, A., Navajas-Perez, R., Torres, M. J., Feltus, F. A., Porter, B., Li, Y., Burroughs, A. M., Luo, M. C., Liu, L., Christopher, D. A., Mount, S. M., Moore, P. H., Sugimura, T., Jiang, J., Schuler, M. A., Friedman, V., Mitchell-Olds, T., Shippen, D. E., dePamphilis, C. W., Palmer, J. D., Freeling, M., Paterson, A. H., Gonsalves, D., Wang, L. and Alam, M. (2008). "The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus)." Nature **452**(7190): 991-996.

Ming, R., Liu, S. C., Lin, Y. R., da Silva, J., Wilson, W., Braga, D., van Deynze, A., Wenslaff, T. F., Wu, K. K., Moore, P. H., Burnquist, W., Sorrells, M. E., Irvine, J. E. and Paterson, A. H. (1998). "Detailed alignment of saccharum and sorghum chromosomes: comparative organization of closely related diploid and polyploid genomes." Genetics **150**(4): 1663-1682.

Mitchell, R. A., Castells-Brooke, N., Taubert, J., Verrier, P. J., Leader, D. J. and Rawlings, C. J. (2007). "Wheat Estimated Transcript Server (WhETS): a tool to provide best estimate of hexaploid wheat transcript sequence." Nucleic Acids Res **35**(Web Server issue): W148-151.

Miyazaki, S., Sugawara, H., Ikeo, K., Gojobori, T. and Tateno, Y. (2004). "DDBJ in the stream of various biological data." Nucleic Acids Res **32**(Database issue): D31-34.

Mochida, K., Yamazaki, Y. and Ogihara, Y. (2004). "Discrimination of homoeologous gene expression in hexaploid wheat by SNP analysis of contigs grouped from a large number of expressed sequence tags." Molecular Genetics and Genomics **270**(5): 371-377.

Mochida, K., Yoshida, T., Sakurai, T., Ogihara, Y. and Shinozaki, K. (2009). "TriFLDB: a database of clustered full-length coding sequences from Triticeae with applications to comparative grass genomics." Plant Physiol **150**(3): 1135-1146.

Moens, C. B., Donn, T. M., Wolf-Saxon, E. R. and Ma, T. P. (2008). "Reverse genetics in zebrafish by TILLING." Brief Funct Genomic Proteomic **7**(6): 454-459.

Montero, L. M., Salinas, J., Matassi, G. and Bernardi, G. (1990). "Gene distribution and isochore organization in the nuclear genome of plants." Nucleic Acids Res **18**(7): 1859-1867.

Moore, G., Abbo, S., Cheung, W., Foote, T., Gale, M., Koebner, R., Leitch, A., Leitch, I., Money, T., Stancombe, P. and et al. (1993a). "Key features of cereal genome organization as revealed by the use of cytosine methylation-sensitive restriction endonucleases." Genomics **15**(3): 472-482.

Moore, G., Devos, K. M., Wang, Z. and Gale, M. D. (1995). "Cereal genome evolution. Grasses, line up and form a circle." Curr Biol **5**(7): 737-739.

Moore, G., Gale, M. D., Kurata, N. and Flavell, R. B. (1993b). "Molecular Analysis of Small Grain Cereal Genomes - Current Status and Prospects." Bio-Technology **11**(5): 584-589.

Morgan, T. H. (1911). "Random Segregation Versus Coupling in Mendelian Inheritance." Science **34**(873): 384.

Morgante, M., Brunner, S., Pea, G., Fengler, K., Zuccolo, A. and Rafalski, A. (2005). "Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize." Nat Genet **37**(9): 997-1002.

Mozo, T., Dewar, K., Dunn, P., Ecker, J. R., Fischer, S., Kloska, S., Lehrach, H., Marra, M., Martienssen, R., Meier-Ewert, S. and Altmann, T. (1999). "A complete BAC-based physical map of the Arabidopsis thaliana genome." Nat Genet **22**(3): 271-275.

Munroe, D. J. and Harris, T. J. (2010). "Third-generation sequencing fireworks at Marco Island." Nat Biotechnol **28**(5): 426-428.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M. and Snyder, M. (2008). "The transcriptional landscape of the yeast genome defined by RNA sequencing." Science **320**(5881): 1344-1349.

Needleman, S. B. and Wunsch, C. D. (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins." J Mol Biol **48**(3): 443-453.

Ning, Z., Cox, A. J. and Mullikin, J. C. (2001). "SSAHA: a fast search method for large DNA databases." Genome Res **11**(10): 1725-1729.

Nobrega, M. A., Ovcharenko, I., Afzal, V. and Rubin, E. M. (2003). "Scanning human gene deserts for long-range enhancers." Science **302**(5644): 413.

Notredame, C., Higgins, D. G. and Heringa, J. (2000). "T-Coffee: A novel method for fast and accurate multiple sequence alignment." J Mol Biol **302**(1): 205-217.

Noutsos, C., Richly, E. and Leister, D. (2005). "Generation and evolutionary fate of insertions of organelle DNA in the nuclear genomes of flowering plants." Genome Res **15**(5): 616-628.

Novaes, E., Drost, D. R., Farmerie, W. G., Pappas, G. J., Jr., Grattapaglia, D., Sederoff, R. R. and Kirst, M. (2008). "High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome." BMC Genomics **9**: 312.

Ogihara, Y., Mochida, K., Kawaura, K., Murai, K., Seki, M., Kamiya, A., Shinozaki, K., Carninci, P., Hayashizaki, Y., Shin, I. T., Kohara, Y. and Yamazaki, Y. (2004). "Construction of a full-length cDNA library from young spikelets of hexaploid wheat and its characterization by large-scale sequencing of expressed sequence tags." Genes Genet Syst **79**(4): 227-232.

Ogihara, Y., Mochida, K., Nemoto, Y., Murai, K., Yamazaki, Y., Shin, I. T. and Kohara, Y. (2003). "Correlated clustering and virtual display of gene expression patterns in the wheat life cycle by large-scale statistical analyses of expressed sequence tags." Plant J **33**(6): 1001-1011.

Ohno, S. (1970). "Enormous Diversity in Genome Sizes of Fish as a Reflection of Natures Extensive Experiments with Gene Duplication." Transactions of the American Fisheries Society **99**(1): 120-&.

Okoniewski, M. J. and Miller, C. J. (2006). "Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations." BMC Bioinformatics **7**: 276.

Olson, M. V., Dutchik, J. E., Graham, M. Y., Brodeur, G. M., Helms, C., Frank, M., MacCollin, M., Scheinman, R. and Frank, T. (1986). "Random-clone strategy for genomic restriction mapping in yeast." Proc Natl Acad Sci U S A **83**(20): 7826-7830.

Ovcharenko, I., Boffelli, D. and Loots, G. G. (2004). "eShadow: a tool for comparing closely related sequences." Genome Res **14**(6): 1191-1198.

Pacific Biosciences. (2009). "Technology Backgrounder: Single Molecule Real Time (SMRT™) DNA Sequencing.", from http://www.pacificbiosciences.com/assets/files/pacbio_technology_backgrounder.pdf.

Palmer, J. D. and Shields, C. R. (1984). "Tripartite Structure of the Brassica-Campestris Mitochondrial Genome." Nature **307**(5950): 437-440.

Pampanwar, V., Engler, F., Hatfield, J., Blundy, S., Gupta, G. and Soderlund, C. (2005). "FPC Web tools for rice, maize, and distribution." Plant Physiol **138**(1): 116-126.

Parchman, T. L., Geist, K. S., Grahnen, J. A., Benkman, C. W. and Buerkle, C. A. (2010). "Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery." BMC Genomics **11**: 180.

Parinov, S., Sevugan, M., Ye, D., Yang, W. C., Kumaran, M. and Sundaresan, V. (1999). "Analysis of flanking sequences from dissociation insertion lines: a database for reverse genetics in Arabidopsis." Plant Cell **11**(12): 2263-2270.

Parra, G., Agarwal, P., Abril, J. F., Wiehe, T., Fickett, J. W. and Guigo, R. (2003). "Comparative gene prediction in human and mouse." Genome Res **13**(1): 108-117.

Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A., Schmutz, J., Spannagl, M., Tang, H., Wang, X., Wicker, T., Bharti, A. K., Chapman, J., Feltus, F. A., Gowik, U., Grigoriev, I. V., Lyons, E., Maher, C. A., Martis, M., Narechania, A., Otillar, R. P., Penning, B. W., Salamov, A. A., Wang, Y., Zhang, L., Carpita, N. C., Freeling, M., Gingle, A. R., Hash, C. T., Keller, B., Klein, P., Kresovich, S., McCann, M. C., Ming, R., Peterson, D. G., Mehboob ur, R., Ware, D., Westhoff, P., Mayer, K. F., Messing, J. and Rokhsar, D. S. (2009). "The Sorghum bicolor genome and the diversification of grasses." Nature **457**(7229): 551-556.

Paterson, A. H., Bowers, J. E. and Chapman, B. A. (2004a). "Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics." Proc Natl Acad Sci U S A **101**(26): 9903-9908.

Paterson, A. H., Bowers, J. E., Chapman, B. A., Peterson, D. G., Rong, J. and Wicker, T. M. (2004b). "Comparative genome analysis of monocots and dicots, toward characterization of angiosperm diversity." Curr Opin Biotechnol **15**(2): 120-125.

Paux, E., Faure, S., Choulet, F., Roger, D., Gauthier, V., Martinant, J. P., Sourdille, P., Balfourier, F., Le Paslier, M. C., Chauveau, A., Cakir, M., Gandon, B. and Feuillet, C. (2010). "Insertion site-based polymorphism markers open new perspectives for genome saturation and marker-assisted selection in wheat." Plant Biotechnol J **8**(2): 196-210.

Paux, E., Roger, D., Badaeva, E., Gay, G., Bernard, M., Sourdille, P. and Feuillet, C. (2006). "Characterizing the composition and evolution of homoeologous genomes in hexaploid wheat through BAC-end sequencing on chromosome 3B." Plant J **48**(3): 463-474.

Paux, E., Scalabrin, S., Bartos, J. and Febrer, M. (2009). "Guidelines for physical map assembly." from http://www.triticeaegenome.eu.

Paux, E., Sourdille, P., Salse, J., Saintenac, C., Choulet, F., Leroy, P., Korol, A., Michalak, M., Kianian, S., Spielmeyer, W., Lagudah, E., Somers, D., Kilian, A., Alaux, M., Vautrin, S., Berges, H., Eversole, K., Appels, R., Safar, J., Simkova, H., Dolezel, J., Bernard, M. and Feuillet, C. (2008). "A physical map of the 1-gigabase bread wheat chromosome 3B." Science **322**(5898): 101-104.

Peterson, D. G., Schulze, S. R., Sciara, E. B., Lee, S. A., Bowers, J. E., Nagel, A., Jiang, N., Tibbitts, D. C., Wessler, S. R. and Paterson, A. H. (2002). "Integration of C_0t analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery." Genome Res **12**(5): 795-807.

Petrov, D. A., Sangster, T. A., Johnston, J. S., Hartl, D. L. and Shaw, K. L. (2000). "Evidence for DNA loss as a determinant of genome size." Science **287**(5455): 1060-1062.

Pinkel, D., Straume, T. and Gray, J. W. (1986). "Cytogenetic Analysis Using Quantitative, High-Sensitivity, Fluorescence Hybridization." Proceedings of the National Academy of Sciences of the United States of America **83**(9): 2934-2938.

Prasad, V., Stromberg, C. A., Alimohammadian, H. and Sahni, A. (2005). "Dinosaur coprolites and the early evolution of grasses and grazers." Science **310**(5751): 1177-1180.

Price, A. L., Jones, N. C. and Pevzner, P. A. (2005). "De novo identification of repeat families in large genomes." Bioinformatics **21 Suppl 1**: i351-358.

Pruitt, K. D., Tatusova, T. and Maglott, D. R. (2007). "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins." Nucleic Acids Res **35**(Database issue): D61-65.

Puchta, H. (2005). "The repair of double-strand breaks in plants: mechanisms and consequences for genome evolution." J Exp Bot **56**(409): 1-14.

Qi, L. L., Echalier, B., Chao, S., Lazo, G. R., Butler, G. E., Anderson, O. D., Akhunov, E. D., Dvorak, J., Linkiewicz, A. M., Ratnasiri, A., Dubcovsky, J., Bermudez-Kandianis, C. E., Greene, R. A., Kantety, R., La Rota, C. M., Munkvold, J. D., Sorrells, S. F., Sorrells, M. E., Dilbirligi, M., Sidhu, D., Erayman, M., Randhawa, H. S., Sandhu, D., Bondareva, S. N., Gill, K. S., Mahmoud, A. A., Ma, X. F., Miftahudin, Gustafson, J. P., Conley, E. J., Nduati, V., Gonzalez-Hernandez, J. L., Anderson, J. A., Peng, J. H., Lapitan, N. L., Hossain, K. G., Kalavacharla, V., Kianian, S. F., Pathan, M. S., Zhang, D. S., Nguyen, H. T., Choi, D. W., Fenton, R. D., Close, T. J., McGuire, P. E., Qualset, C. O. and Gill, B. S. (2004). "A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat." Genetics **168**(2): 701-712.

Quraishi, U. M., Abrouk, M., Bolot, S., Pont, C., Throude, M., Guilhot, N., Confolent, C., Bortolini, F., Praud, S., Murigneux, A., Charmet, G. and Salse, J. (2009). "Genomics in cereals: from genome-wide conserved orthologous set (COS) sequences to candidate genes for trait dissection." Funct Integr Genomics **9**(4): 473-484.

R Development Core Team (2010). R: A Language and Environment for Statistical Computing.

Rabinowicz, P. D. (2003). "Constructing gene-enriched plant genomic libraries using methylation filtration technology." Methods Mol Biol **236**: 21-36.

Rabinowicz, P. D. and Bennetzen, J. L. (2006). "The maize genome as a model for efficient sequence analysis of large plant genomes." Curr Opin Plant Biol **9**(2): 149-156.

Raizada, M. N., Nan, G. L. and Walbot, V. (2001). "Somatic and germinal mobility of the RescueMu transposon in transgenic maize." Plant Cell **13**(7): 1587-1608.

Ramonell, K., Berrocal-Lobo, M., Koh, S., Wan, J., Edwards, H., Stacey, G. and Somerville, S. (2005). "Loss-of-function mutations in chitin responsive genes show increased susceptibility to the powdery mildew pathogen Erysiphe cichoracearum." Plant Physiol **138**(2): 1027-1036.

Rasch, E. M. (1985). "DNA "standards" and the range of accurate DNA estimates by Feulgen absorption microspectrophotometry." Prog Clin Biol Res **196**: 137-166.

Rayburn, A. L., Biradar, D. P., Bullock, D. G. and Mcmurphy, L. M. (1993). "Nuclear-DNA Content in F1 Hybrids of Maize." Heredity **70**: 294-300.

Reference Genome Group of the Gene Ontology Consortium (2009). "The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species." PLoS Comput Biol **5**(7): e1000431.

Rice Chromosome 10 Sequencing Consortium (2003). "In-depth view of structure, activity, and evolution of rice chromosome 10." Science **300**(5625): 1566-1569.

Rice, P., Longden, I. and Bleasby, A. (2000). "EMBOSS: the European Molecular Biology Open Software Suite." Trends Genet **16**(6): 276-277.

Richly, E. and Leister, D. (2004a). "NUMTs in sequenced eukaryotic genomes." Mol Biol Evol **21**(6): 1081-1084.

Richly, E. and Leister, D. (2004b). "NUPTs in sequenced eukaryotes and their genomic organization in relation to NUMTs." Mol Biol Evol **21**(10): 1972-1980.

Roest Crollius, H., Jaillon, O., Bernot, A., Dasilva, C., Bouneau, L., Fischer, C., Fizames, C., Wincker, P., Brottier, P., Quetier, F., Saurin, W. and Weissenbach, J. (2000). "Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence." Nat Genet **25**(2): 235-238.

Rosso, M. G., Li, Y., Strizhov, N., Reiss, B., Dekker, K. and Weisshaar, B. (2003). "An Arabidopsis thaliana T-DNA mutagenized population (GABI-Kat) for flanking sequence tag-based reverse genetics." Plant Mol Biol **53**(1-2): 247-259.

Royal Society of London (2009). Reaping the Benefits: Science and the Sustainable Intensification of Global Agriculture.

Rozen, S. and Skaletsky, H. (2000). "Primer3 on the WWW for general users and for biologist programmers." Methods Mol Biol **132**: 365-386.

Rubin, G. M., Yandell, M. D., Wortman, J. R., Gabor Miklos, G. L., Nelson, C. R., Hariharan, I. K., Fortini, M. E., Li, P. W., Apweiler, R., Fleischmann, W., Cherry, J. M., Henikoff, S., Skupski, M. P., Misra, S., Ashburner, M., Birney, E., Boguski, M. S., Brody, T., Brokstein, P., Celniker, S. E., Chervitz, S. A., Coates, D., Cravchik, A., Gabrielian, A., Galle, R. F., Gelbart, W. M., George, R. A., Goldstein, L. S., Gong, F., Guan, P., Harris, N. L., Hay, B. A., Hoskins, R. A., Li, J., Li, Z., Hynes, R. O., Jones, S. J., Kuehl, P. M., Lemaitre, B., Littleton, J. T., Morrison, D. K., Mungall, C., O'Farrell, P. H., Pickeral, O. K., Shue, C., Vosshall, L. B., Zhang, J., Zhao, Q., Zheng, X. H. and Lewis, S. (2000). "Comparative genomics of the eukaryotes." Science **287**(5461): 2204-2215.

Ruepp, A., Zollner, A., Maier, D., Albermann, K., Hani, J., Mokrejs, M., Tetko, I., Guldener, U., Mannhaupt, G., Munsterkotter, M. and Mewes, H. W. (2004). "The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes." Nucleic Acids Res **32**(18): 5539-5545.

Sabot, F. and Schulman, A. H. (2006). "Parasitism and the retrotransposon life cycle in plants: a hitchhiker's guide to the genome." Heredity **97**(6): 381-388.

Saintenac, C., Falque, M., Martin, O. C., Paux, E., Feuillet, C. and Sourdille, P. (2009). "Detailed recombination studies along chromosome 3B provide new insights on crossover distribution in wheat (Triticum aestivum L.)." Genetics **181**(2): 393-403.

Salamov, A. A. and Solovyev, V. V. (2000). "Ab initio gene finding in Drosophila genomic DNA." Genome Research **10**(4): 516-522.

Salse, J., Bolot, S., Throude, M., Jouffe, V., Piegu, B., Masood Quraishi, U., Calcagno, T., Cooke, R., Delseny, M. and Feuillet, C. (2008a). "Identification and Characterization of Shared

Duplications between Rice and Wheat Provide New Insight into Grass Genome Evolution." Plant Cell.

Salse, J., Chague, V., Bolot, S., Magdelenat, G., Huneau, C., Pont, C., Belcram, H., Couloux, A., Gardais, S., Evrard, A., Segurens, B., Charles, M., Ravel, C., Samain, S., Charmet, G., Boudet, N. and Chalhoub, B. (2008b). "New insights into the origin of the B genome of hexaploid wheat: evolutionary relationships at the SPA genomic region with the S genome of the diploid relative Aegilops speltoides." BMC Genomics **9**: 555.

Salse, J., Piegu, B., Cooke, R. and Delseny, M. (2004). "New in silico insight into the synteny between rice (Oryza sativa L.) and maize (Zea mays L.) highlights reshuffling and identifies new duplications in the rice genome." Plant J **38**(3): 396-409.

Salvi, S., Sponza, G., Morgante, M., Tomes, D., Niu, X., Fengler, K. A., Meeley, R., Ananiev, E. V., Svitashev, S., Bruggemann, E., Li, B., Hainey, C. F., Radovic, S., Zaina, G., Rafalski, J. A., Tingey, S. V., Miao, G. H., Phillips, R. L. and Tuberosa, R. (2007). "Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize." Proc Natl Acad Sci U S A **104**(27): 11376-11381.

Samad, A., Huff, E. F., Cai, W. and Schwartz, D. C. (1995). "Optical mapping: a novel, single-molecule approach to genomic analysis." Genome Res **5**(1): 1-4.

Sanger, F., Nicklen, S. and Coulson, A. R. (1977). "DNA sequencing with chain-terminating inhibitors." Proc Natl Acad Sci U S A **74**(12): 5463-5467.

Sangwan, I. and O'Brian, M. R. (2002). "Identification of a soybean protein that interacts with GAGA element dinucleotide repeat DNA." Plant Physiol **129**(4): 1788-1794.

SanMiguel, P., Gaut, B. S., Tikhonov, A., Nakajima, Y. and Bennetzen, J. L. (1998). "The paleontology of intergene retrotransposons of maize." Nat Genet **20**(1): 43-45.

SanMiguel, P., Tikhonov, A., Jin, Y. K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P. S., Edwards, K. J., Lee, M., Avramova, Z. and Bennetzen, J. L. (1996). "Nested retrotransposons in the intergenic regions of the maize genome." Science **274**(5288): 765-768.

SanMiguel, P. J., Ramakrishna, W., Bennetzen, J. L., Busso, C. S. and Dubcovsky, J. (2002). "Transposable elements, genes and recombination in a 215-kb contig from wheat chromosome 5A(m)." Funct Integr Genomics **2**(1-2): 70-80.

Sasaki, T., Matsumoto, T., Yamamoto, K., Sakata, K., Baba, T., Katayose, Y., Wu, J., Niimura, Y., Cheng, Z., Nagamura, Y., Antonio, B. A., Kanamori, H., Hosokawa, S., Masukawa, M., Arikawa, K., Chiden, Y., Hayashi, M., Okamoto, M., Ando, T., Aoki, H., Arita, K., Hamada, M., Harada, C., Hijishita, S., Honda, M., Ichikawa, Y., Idonuma, A., Iijima, M., Ikeda, M., Ikeno, M., Ito, S., Ito, T., Ito, Y., Ito, Y., Iwabuchi, A., Kamiya, K., Karasawa, W., Katagiri, S., Kikuta, A., Kobayashi, N., Kono, I., Machita, K., Maehara, T., Mizuno, H., Mizubayashi, T., Mukai, Y., Nagasaki, H., Nakashima, M., Nakama, Y., Nakamichi, Y., Nakamura, M., Namiki, N., Negishi, M., Ohta, I., Ono, N., Saji, S., Sakai, K., Shibata, M., Shimokawa, T., Shomura, A., Song, J., Takazaki, Y., Terasawa, K., Tsuji, K., Waki, K., Yamagata, H., Yamane, H., Yoshiki, S., Yoshihara, R., Yukawa, K., Zhong, H., Iwama, H., Endo, T., Ito, H., Hahn, J. H., Kim, H. I., Eun, M. Y., Yano, M., Jiang, J. and Gojobori, T. (2002). "The genome sequence and structure of rice chromosome 1." Nature **420**(6913): 312-316.

Schena, M., Shalon, D., Davis, R. W. and Brown, P. O. (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." Science **270**(5235): 467-470.

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T. A., Minx, P., Reily, A. D., Courtney, L., Kruchowski, S. S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S. M., Belter, E., Du, F., Kim, K., Abbott, R. M., Cotton, M., Levy, A., Marchetto, P., Ochoa, K., Jackson, S. M., Gillam, B., Chen, W., Yan, L., Higginbotham, J., Cardenas, M., Waligorski, J., Applebaum, E., Phelps, L., Falcone, J., Kanchi, K., Thane, T., Scimone, A., Thane, N., Henke, J., Wang, T., Ruppert, J., Shah, N., Rotter, K., Hodges, J., Ingenthron, E., Cordes, M., Kohlberg, S., Sgro, J., Delgado, B., Mead, K., Chinwalla, A., Leonard, S., Crouse, K., Collura, K., Kudrna, D., Currie, J., He, R., Angelova, A., Rajasekar, S., Mueller, T., Lomeli, R., Scara, G., Ko, A., Delaney, K., Wissotski, M., Lopez, G., Campos, D., Braidotti, M., Ashley, E., Golser, W., Kim, H., Lee, S., Lin, J., Dujmic, Z., Kim, W., Talag, J., Zuccolo, A., Fan, C., Sebastian, A., Kramer, M., Spiegel, L., Nascimento, L., Zutavern, T., Miller, B., Ambroise, C., Muller, S., Spooner, W., Narechania, A., Ren, L., Wei, S., Kumari, S., Faga, B., Levy, M. J., McMahan, L., Van Buren, P., Vaughn, M. W., Ying, K., Yeh, C. T., Emrich, S. J., Jia, Y., Kalyanaraman, A., Hsia, A. P., Barbazuk, W. B., Baucom, R. S., Brutnell, T. P., Carpita, N. C., Chaparro, C., Chia, J. M., Deragon, J. M., Estill, J. C., Fu, Y., Jeddeloh, J. A., Han, Y., Lee, H., Li, P., Lisch, D. R., Liu, S., Liu, Z., Nagel, D. H., McCann, M. C., SanMiguel, P., Myers, A. M., Nettleton, D., Nguyen, J., Penning, B. W., Ponnala, L., Schneider, K. L., Schwartz, D. C., Sharma, A., Soderlund, C., Springer, N. M., Sun, Q., Wang, H., Waterman, M., Westerman, R., Wolfgruber, T. K., Yang, L., Yu, Y., Zhang, L., Zhou, S., Zhu, Q., Bennetzen, J. L., Dawe, R. K., Jiang, J., Jiang, N., Presting, G. G., Wessler, S. R., Aluru, S., Martienssen, R. A., Clifton, S. W., McCombie, W. R., Wing, R. A. and Wilson, R. K. (2009). "The B73 maize genome: complexity, diversity, and dynamics." Science **326**(5956): 1112-1115.

Schneeberger, R. G., Zhang, K., Tatarinova, T., Troukhan, M., Kwok, S. F., Drais, J., Klinger, K., Orejudos, F., Macy, K., Bhakta, A., Burns, J., Subramanian, G., Donson, J., Flavell, R. and Feldmann, K. A. (2005). "Agrobacterium T-DNA integration in Arabidopsis is correlated with DNA sequence compositions that occur frequently in gene promoter regions." Funct Integr Genomics **5**(4): 240-253.

Schulman, A. H. and Kalendar, R. (2005). "A movable feast: diverse retrotransposons and their contribution to barley genome dynamics." Cytogenet Genome Res **110**(1-4): 598-605.

Schulte, D., Close, T. J., Graner, A., Langridge, P., Matsumoto, T., Muehlbauer, G., Sato, K., Schulman, A. H., Waugh, R., Wise, R. P. and Stein, N. (2009). "The international barley sequencing consortium--at the threshold of efficient access to the barley genome." Plant Physiol **149**(1): 142-147.

Schwartz, S., Elnitski, L., Li, M., Weirauch, M., Riemer, C., Smit, A., Green, E. D., Hardison, R. C. and Miller, W. (2003a). "MultiPipMaker and supporting tools: Alignments and analysis of multiple genomic DNA sequences." Nucleic Acids Res **31**(13): 3518-3524.

Schwartz, S., Kent, W. J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R. C., Haussler, D. and Miller, W. (2003b). "Human-mouse alignments with BLASTZ." Genome Res **13**(1): 103-107.

Sears, E. R. (1976). "Genetic control of chromosome pairing in wheat." Annu Rev Genet **10**: 31-51.

Sessions, A., Burke, E., Presting, G., Aux, G., McElver, J., Patton, D., Dietrich, B., Ho, P., Bacwaden, J., Ko, C., Clarke, J. D., Cotton, D., Bullis, D., Snell, J., Miguel, T., Hutchison, D., Kimmerly, B., Mitzel, T., Katagiri, F., Glazebrook, J., Law, M. and Goff, S. A. (2002). "A high-throughput Arabidopsis reverse genetics system." Plant Cell **14**(12): 2985-2994.

Shantz, H. L. (1954). "The Place of Grasslands in the Earths Cover of Vegetation." Ecology **35**(2): 143-145.

Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., Wang, M. D., Zhang, K., Mitra, R. D. and Church, G. M. (2005). "Accurate multiplex polony sequencing of an evolved bacterial genome." Science **309**(5741): 1728-1732.

Shi, Y., Draper, J. and Stace, C. (1993). "Ribosomal DNA Variation and Its Phylogenetic Implication in the Genus Brachypodium (Poaceae)." Plant Systematics and Evolution **188**(3-4): 125-138.

Shirasu, K., Schulman, A. H., Lahaye, T. and Schulze-Lefert, P. (2000). "A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion." Genome Research **10**(7): 908-915.

Shizuya, H., Birren, B., Kim, U. J., Mancino, V., Slepak, T., Tachiiri, Y. and Simon, M. (1992). "Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector." Proc Natl Acad Sci U S A **89**(18): 8794-8797.

Simillion, C., Vandepoele, K., Van Montagu, M. C., Zabeau, M. and Van de Peer, Y. (2002). "The hidden duplication past of Arabidopsis thaliana." Proc Natl Acad Sci U S A **99**(21): 13627-13632.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. and Birol, I. (2009). "ABySS: a parallel assembler for short read sequence data." Genome Res **19**(6): 1117-1123.

Singh, N. K., Dalal, V., Batra, K., Singh, B. K., Chitra, G., Singh, A., Ghazi, I. A., Yadav, M., Pandit, A., Dixit, R., Singh, P. K., Singh, H., Koundal, K. R., Gaikwad, K., Mohapatra, T. and Sharma, T. R. (2007). "Single-copy genes define a conserved order between rice and wheat for understanding differences caused by duplication, deletion, and transposition of genes." Funct Integr Genomics **7**(1): 17-35.

Sinha, A. U. and Meller, J. (2007). "Cinteny: flexible analysis and visualization of synteny and genome rearrangements in multiple organisms." Bmc Bioinformatics **8**: -.

Skinner, M. E., Uzilov, A. V., Stein, L. D., Mungall, C. J. and Holmes, I. H. (2009). "JBrowse: a next-generation genome browser." Genome Res **19**(9): 1630-1638.

Slade, A. J., Fuerstenberg, S. I., Loeffler, D., Steine, M. N. and Facciotti, D. (2005). "A reverse genetic, nontransgenic approach to wheat crop improvement by TILLING." Nat Biotechnol **23**(1): 75-81.

Slater, G. S. and Birney, E. (2005). "Automated generation of heuristics for biological sequence comparison." BMC Bioinformatics **6**: 31.

Smit, A. F. A., Hubley, R. and Green, P. (1999). "RepeatMasker." from http://www.repeatmasker.org.

Smith, D. B. and Flavell, R. B. (1975). "Characterization of Wheat Genome by Renaturation Kinetics." Chromosoma **50**(3): 223-242.

Smith, T. F. and Waterman, M. S. (1981). "Identification of common molecular subsequences." J Mol Biol **147**(1): 195-197.

Smith, V., Chou, K. N., Lashkari, D., Botstein, D. and Brown, P. O. (1996). "Functional analysis of the genes of yeast chromosome V by genetic footprinting." Science **274**(5295): 2069-2074.

Soderlund, C., Longden, I. and Mott, R. (1997). "FPC: a system for building contigs from restriction fingerprinted clones." Comput Appl Biosci **13**(5): 523-535.

Soderlund, C., Nelson, W., Shoemaker, A. and Paterson, A. (2006). "SyMAP: A system for discovering and viewing syntenic regions of FPC maps." Genome Res **16**(9): 1159-1168.

Solomon, M. J., Larsen, P. L. and Varshavsky, A. (1988). "Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene." Cell **53**(6): 937-947.

Solovyev, V. V., Salamov, A. A. and Lawrence, C. B. (1994). "Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames." Nucleic Acids Res **22**(24): 5156-5163.

Soltis, D., Soltis, P. and Tate, J. (2004). "Advances in the study of polyploidy since Plant speciation." New Phytologist **161**(1): 173-191.

Song, R. T., Llaca, V., Linton, E. and Messing, J. (2001). "Sequence, regulation, and evolution of the maize 22-kD alpha zein in gene family." Genome Research **11**(11): 1817-1825.

Sorrells, M. E., La Rota, M., Bermudez-Kandianis, C. E., Greene, R. A., Kantety, R., Munkvold, J. D., Miftahudin, Mahmoud, A., Ma, X., Gustafson, P. J., Qi, L. L., Echalier, B., Gill, B. S., Matthews, D. E., Lazo, G. R., Chao, S., Anderson, O. D., Edwards, H., Linkiewicz, A. M., Dubcovsky, J., Akhunov, E. D., Dvorak, J., Zhang, D., Nguyen, H. T., Peng, J., Lapitan, N. L., Gonzalez-Hernandez, J. L., Anderson, J. A., Hossain, K., Kalavacharla, V., Kianian, S. F., Choi, D. W., Close, T. J., Dilbirligi, M., Gill, K. S., Steber, C., Walker-Simmons, M. K., McGuire, P. E. and Qualset, C. O. (2003). "Comparative DNA sequence analysis of wheat and rice genomes." Genome Res **13**(8): 1818-1827.

Speulman, E., Metz, P. L., van Arkel, G., te Lintel Hekkert, B., Stiekema, W. J. and Pereira, A. (1999). "A two-component enhancer-inhibitor transposon mutagenesis system for functional analysis of the Arabidopsis genome." Plant Cell **11**(10): 1853-1866.

Srinivasachary, Dida, M. M., Gale, M. D. and Devos, K. M. (2007). "Comparative analyses reveal high levels of conserved colinearity between the finger millet and rice genomes." Theor Appl Genet **115**(4): 489-499.

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G., Korf, I., Lapp, H., Lehvaslaiho, H., Matsalla, C., Mungall, C. J., Osborne, B. I., Pocock, M. R., Schattner, P., Senger, M., Stein, L. D., Stupka, E., Wilkinson, M. D. and Birney, E. (2002). "The Bioperl toolkit: Perl modules for the life sciences." Genome Res **12**(10): 1611-1618.

Stam, P. (1993). "Construction of Integrated Genetic-Linkage Maps by Means of a New Computer Package - Joinmap." Plant Journal **3**(5): 739-744.

Stein, L. D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M. R., Chen, N., Chinwalla, A., Clarke, L., Clee, C., Coghlan, A., Coulson, A., D'Eustachio, P., Fitch, D. H., Fulton, L. A., Fulton, R. E., Griffiths-Jones, S., Harris, T. W., Hillier, L. W., Kamath, R., Kuwabara, P. E., Mardis, E. R., Marra, M. A., Miner, T. L., Minx, P., Mullikin, J. C., Plumb, R. W., Rogers, J., Schein, J. E., Sohrmann, M., Spieth, J., Stajich, J. E., Wei, C., Willey, D., Wilson, R. K., Durbin, R. and Waterston, R. H. (2003). "The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics." PLoS Biol **1**(2): E45.

Stein, L. D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J. E., Harris, T. W., Arva, A. and Lewis, S. (2002). "The generic genome browser: a building block for a model organism system database." Genome Res **12**(10): 1599-1610.

Stormo, G. D. and Haussler, D. (1994). "Optimally parsing a sequence into different classes based on multiple types of evidence." Proc Int Conf Intell Syst Mol Biol **2**: 369-375.

Sturtevant, A. H. (1913). "The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association." Journal of Experimental Zoology **14**: 43-59.

Sugiura, M. (1995). "The chloroplast genome." Essays Biochem **30**: 49-57.

Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T. Z., Garcia-Hernandez, M., Foerster, H., Li, D., Meyer, T., Muller, R., Ploetz, L., Radenbaugh, A., Singh, S., Swing, V., Tissier, C., Zhang, P. and Huala, E. (2008). "The Arabidopsis Information Resource (TAIR): gene structure and function annotation." Nucleic Acids Res **36**(Database issue): D1009-1014.

Swigonova, Z., Lai, J. S., Ma, J. X., Ramakrishna, W., Llaca, V., Bennetzen, J. L. and Messing, J. (2004). "Close split of sorghum and maize genome progenitors." Genome Research **14**(10A): 1916-1923.

Szostak, J. W., Orr-Weaver, T. L., Rothstein, R. J. and Stahl, F. W. (1983). "The double-strand-break repair model for recombination." Cell **33**(1): 25-35.

Tagle, D. A., Koop, B. F., Goodman, M., Slightom, J. L., Hess, D. L. and Jones, R. T. (1988). "Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints." J Mol Biol **203**(2): 439-455.

Tanaka, T., Antonio, B. A., Kikuchi, S., Matsumoto, T., Nagamura, Y., Numa, H., Sakai, H., Wu, J., Itoh, T., Sasaki, T., Aono, R., Fujii, Y., Habara, T., Harada, E., Kanno, M., Kawahara, Y., Kawashima, H., Kubooka, H., Matsuya, A., Nakaoka, H., Saichi, N., Sanbonmatsu, R., Sato, Y., Shinso, Y., Suzuki, M., Takeda, J., Tanino, M., Todokoro, F., Yamaguchi, K., Yamamoto, N., Yamasaki, C., Imanishi, T., Okido, T., Tada, M., Ikeo, K., Tateno, Y., Gojobori, T., Lin, Y. C., Wei, F. J., Hsing, Y. I., Zhao, Q., Han, B., Kramer, M. R., McCombie, R. W., Lonsdale, D., O'Donovan, C. C., Whitfield, E. J., Apweiler, R., Koyanagi, K. O., Khurana, J. P., Raghuvanshi, S., Singh, N. K., Tyagi, A. K., Haberer, G., Fujisawa, M., Hosokawa, S., Ito, Y., Ikawa, H., Shibata, M., Yamamoto, M., Bruskiewich, R. M., Hoen, D. R., Bureau, T. E., Namiki, N., Ohyanagi, H., Sakai, Y., Nobushima, S., Sakata, K., Barrero, R. A., Souvorov, A., Smith-White, B., Tatusova, T., An, S., An, G., S, O. O., Fuks, G., Messing, J., Christie, K. R., Lieberherr, D., Kim, H., Zuccolo, A., Wing, R. A., Nobuta, K., Green, P. J., Lu, C., Meyers, B. C., Chaparro, C., Piegu, B., Panaud, O. and Echeverria, M. (2008). "The Rice Annotation Project Database (RAP-DB): 2008 update." Nucleic Acids Res **36**(Database issue): D1028-1033.

Tarchini, R., Biddle, P., Wineland, R., Tingey, S. and Rafalski, A. (2000). "The complete sequence of 340 kb of DNA around the rice Adh1-Adh2 region reveals interrupted colinearity with maize chromosome 4." Plant Cell **12**(3): 381-391.

Tatusova, T. A. and Madden, T. L. (1999). "BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences." FEMS Microbiol Lett **174**(2): 247-250.

Terada, R., Urawa, H., Inagaki, Y., Tsugane, K. and Iida, S. (2002). "Efficient gene targeting by homologous recombination in rice." Nat Biotechnol **20**(10): 1030-1034.

The Arabidopsis Genome Initiative (2000). "Analysis of the genome sequence of the flowering plant Arabidopsis thaliana." Nature **408**(6814): 796-815.

The *C. elegans* Sequencing Consortium (1998). "Genome sequence of the nematode C. elegans: a platform for investigating biology." Science **282**(5396): 2012-2018.

Thibaud-Nissen, F., Ouyang, S. and Buell, C. R. (2009). "Identification and characterization of pseudogenes in the rice gene complement." BMC Genomics **10**: 317.

Thimm, O., Blasing, O., Gibon, Y., Nagel, A., Meyer, S., Kruger, P., Selbig, J., Muller, L. A., Rhee, S. Y. and Stitt, M. (2004). "MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes." Plant J **37**(6): 914-939.

Thole, V., Worland, B., Wright, J., Bevan, M. W. and Vain, P. (2010). "Distribution and characterization of more than 1000 T-DNA tags in the genome of Brachypodium distachyon community standard line Bd21." Plant Biotechnol J **8**(6): 734-747.

Thomas, B. C., Rapaka, L., Lyons, E., Pedersen, B. and Freeling, M. (2007). "Arabidopsis intragenomic conserved noncoding sequence." Proc Natl Acad Sci U S A **104**(9): 3348-3353.

Thomas, C. A., Jr. (1971). "The genetic organization of chromosomes." Annu Rev Genet **5**: 237-256.

Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994). "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." Nucleic Acids Res **22**(22): 4673-4680.

Thurston, M. I. and Field, D. (2005). Msatfinder: detection and characterisation of microsatellites.

Tian, C. G., Xiong, Y. Q., Liu, T. Y., Sun, S. H., Chen, L. B. and Chen, M. S. (2005). "Evidence for an ancient whole-genome duplication event in rice and other cereals." Yi Chuan Xue Bao **32**(5): 519-527.

Tikhonov, A. P., SanMiguel, P. J., Nakajima, Y., Gorenstein, N. M., Bennetzen, J. L. and Avramova, Z. (1999). "Colinearity and its exceptions in orthologous adh regions of maize and sorghum." Proc Natl Acad Sci U S A **96**(13): 7409-7414.

Till, B. J., Reynolds, S. H., Weil, C., Springer, N., Burtner, C., Young, K., Bowers, E., Codomo, C. A., Enns, L. C., Odden, A. R., Greene, E. A., Comai, L. and Henikoff, S. (2004). "Discovery of induced point mutations in maize genes by TILLING." BMC Plant Biol **4**: 12.

Timmis, J. N., Ayliffe, M. A., Huang, C. Y. and Martin, W. (2004). "Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes." Nat Rev Genet **5**(2): 123-135.

Tonti-Filippini, J. (2010). "Anno-J: Annotation Browsing 2.0." from http://www.annoj.org/.

Tweedie, S., Ashburner, M., Falls, K., Leyland, P., McQuilton, P., Marygold, S., Millburn, G., Osumi-Sutherland, D., Schroeder, A., Seal, R. and Zhang, H. (2009). "FlyBase: enhancing Drosophila Gene Ontology annotations." Nucleic Acids Res **37**(Database issue): D555-559.

Uchida, N., Townsley, B., Chung, K. H. and Sinha, N. (2007). "Regulation of SHOOT MERISTEMLESS genes via an upstream-conserved noncoding sequence coordinates leaf development." Proc Natl Acad Sci U S A **104**(40): 15953-15958.

Unseld, M., Marienfeld, J. R., Brandt, P. and Brennicke, A. (1997). "The mitochondrial genome of Arabidopsis thaliana contains 57 genes in 366,924 nucleotides." Nat Genet **15**(1): 57-61.

Usadel, B., Nagel, A., Thimm, O., Redestig, H., Blaesing, O. E., Palacios-Rojas, N., Selbig, J., Hannemann, J., Piques, M. C., Steinhauser, D., Scheible, W. R., Gibon, Y., Morcuende, R., Weicht, D., Meyer, S. and Stitt, M. (2005). "Extension of the visualization tool MapMan to allow statistical analysis of arrays, display of corresponding genes, and comparison with known responses." Plant Physiol **138**(3): 1195-1204.

USDA ARS. (2010). "National Plant Germplasm System.", from http://www.ars-grin.gov/npgs/.

USDA Foreign Agricultural Service. (2010). "World Agricultural Production." Retrieved 20/082010, 2010, from http://www.fas.usda.gov/wap/current/toc.asp.

Vain, P., Afolabi, A. S., Worland, B. and Snape, J. W. (2003). "Transgene behaviour in populations of rice plants transformed using a new dual binary vector system: pGreen/pSoup." Theor Appl Genet **107**(2): 210-217.

Vain, P., Worland, B., Thole, V., McKenzie, N., Alves, S. C., Opanowicz, M., Fish, L. J., Bevan, M. W. and Snape, J. W. (2008). "Agrobacterium-mediated transformation of the temperate grass Brachypodium distachyon (genotype Bd21) for T-DNA insertional mutagenesis." Plant Biotechnol J **6**(3): 236-245.

Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K., Malek, J. A., Costa, G., McKernan, K., Sidow, A., Fire, A. and Johnson, S. M. (2008). "A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning." Genome Res **18**(7): 1051-1063.

Van Deynze, A. E., Nelson, J. C., Odonoughue, L. S., Ahn, S. N., Siripoonwiwat, W., Harrington, S. E., Yglesias, E. S., Braga, D. P., Mccouch, S. R. and Sorrells, M. E. (1995). "Comparative Mapping in Grasses - Oat Relationships." Molecular & General Genetics **249**(3): 349-356.

Vanin, E. F. (1985). "Processed pseudogenes: characteristics and evolution." Annu Rev Genet **19**: 253-272.

Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett, D. E., Jr., Hieter, P., Vogelstein, B. and Kinzler, K. W. (1997). "Characterization of the yeast transcriptome." Cell **88**(2): 243-251.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigo, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A. and Zhu, X. (2001). "The sequence of the human genome." Science **291**(5507): 1304-1351.

Venter, J. C., Smith, H. O. and Hood, L. (1996). "A new strategy for genome sequencing." Nature **381**(6581): 364-366.

Vera, J. C., Wheat, C. W., Fescemyer, H. W., Frilander, M. J., Crawford, D. L., Hanski, I. and Marden, J. H. (2008). "Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing." Mol Ecol **17**(7): 1636-1647.

Vicient, C. M., Suoniemi, A., Anamthawat-Jonsson, K., Tanskanen, J., Beharav, A., Nevo, E. and Schulman, A. H. (1999). "Retrotransposon BARE-1 and Its Role in Genome Evolution in the Genus Hordeum." Plant Cell **11**(9): 1769-1784.

Vilhar, B., Greilhuber, J., Koce, J. D., Temsch, E. M. and Dermastia, M. (2001). "Plant genome size measurement with DNA image cytometry." Annals of Botany **87**(6): 719-728.

Vision, T. J., Brown, D. G. and Tanksley, S. D. (2000). "The origins of genomic duplications in Arabidopsis." Science **290**(5499): 2114-2117.

Vitte, C. and Bennetzen, J. L. (2006). "Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution." Proc Natl Acad Sci U S A **103**(47): 17638-17643.

Vogel, J. P., Garvin, D. F., Leong, O. M. and Hayden, D. M. (2006a). "Agrobacterium-mediated transformation and inbred line development in the model grass Brachypodium distachyon." Plant Cell Tissue and Organ Culture **84**(2): 199-211.

Vogel, J. P., Gu, Y. Q., Twigg, P., Lazo, G. R., Laudencia-Chingcuanco, D., Hayden, D. M., Donze, T. J., Vivian, L. A., Stamova, B. and Coleman-Derr, D. (2006b). "EST sequencing and phylogenetic analysis of the model grass Brachypodium distachyon." Theor Appl Genet **113**(2): 186-195.

Vogel, J. P., Tuna, M., Budak, H., Huo, N., Gu, Y. Q. and Steinwand, M. A. (2009). "Development of SSR markers and analysis of diversity in Turkish populations of Brachypodium distachyon." BMC Plant Biol **9**: 88.

Walbot, V. (2000). "Saturation mutagenesis using maize transposons." Curr Opin Plant Biol **3**(2): 103-107.

Wang, T. and Stormo, G. D. (2003). "Combining phylogenetic data with co-regulated genes to identify regulatory motifs." Bioinformatics **19**(18): 2369-2380.

Wang, W., Wang, Y., Zhang, Q., Qi, Y. and Guo, D. (2009a). "Global characterization of Artemisia annua glandular trichome transcriptome using 454 pyrosequencing." BMC Genomics **10**: 465.

Wang, X., Haberer, G. and Mayer, K. F. (2009b). "Discovery of cis-elements between sorghum and rice using co-expression and evolutionary conservation." BMC Genomics **10**: 284.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S. E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M. R., Brown, D. G., Brown, S. D., Bult, C., Burton, J., Butler, J., Campbell, R. D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A. T., Church, D. M., Clamp, M., Clee, C., Collins, F. S., Cook, L. L., Copley, R. R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K. D., Deri, J., Dermitzakis, E. T., Dewey, C., Dickens, N. J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D. M., Eddy, S. R., Elnitski, L., Emes, R. D., Eswara, P., Eyras, E., Felsenfeld, A., Fewell, G. A., Flicek, P., Foley, K., Frankel, W. N., Fulton, L. A., Fulton, R. S., Furey, T. S., Gage, D., Gibbs, R. A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T. A., Green, E. D., Gregory, S., Guigo, R., Guyer, M., Hardison, R. C., Haussler, D., Hayashizaki, Y., Hillier, L. W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D. B., Johnson, L. S., Jones, M., Jones, T. A., Joy, A., Kamal, M., Karlsson, E. K., Karolchik, D., Kasprzyk, A., Kawai, J., Keibler, E., Kells, C., Kent, W. J., Kirby, A., Kolbe, D. L., Korf, I., Kucherlapati, R. S., Kulbokas, E. J., Kulp, D., Landers, T., Leger, J. P., Leonard, S., Letunic, I., Levine, R., Li, J., Li, M., Lloyd, C., Lucas, S., Ma, B., Maglott, D. R., Mardis, E. R., Matthews, L., Mauceli, E., Mayer, J. H., McCarthy, M., McCombie, W. R., McLaren, S., McLay, K., McPherson, J. D., Meldrim, J., Meredith, B., Mesirov, J. P., Miller, W., Miner, T. L., Mongin, E., Montgomery, K. T., Morgan, M., Mott, R., Mullikin, J. C., Muzny, D. M., Nash, W. E., Nelson, J. O., Nhan, M. N., Nicol, R., Ning, Z., Nusbaum, C., O'Connor, M. J., Okazaki, Y., Oliver, K., Overton-Larty, E., Pachter, L., Parra, G., Pepin, K. H., Peterson, J., Pevzner, P., Plumb, R., Pohl, C. S., Poliakov, A., Ponce, T. C., Ponting, C. P., Potter, S., Quail, M., Reymond, A., Roe, B. A., Roskin, K. M., Rubin, E. M., Rust, A. G., Santos, R., Sapojnikov, V., Schultz, B., Schultz, J., Schwartz, M. S., Schwartz, S., Scott, C., Seaman, S., Searle, S., Sharpe, T., Sheridan, A., Shownkeen, R., Sims, S., Singer, J. B., Slater, G., Smit, A., Smith, D. R., Spencer, B., Stabenau, A., Stange-Thomann, N., Sugnet, C., Suyama, M., Tesler, G., Thompson, J., Torrents, D., Trevaskis, E., Tromp, J., Ucla, C., Ureta-Vidal, A., Vinson, J. P., Von Niederhausern, A. C., Wade, C. M., Wall, M., Weber, R. J., Weiss, R. B., Wendl, M. C., West, A. P., Wetterstrand, K., Wheeler, R., Whelan, S., Wierzbowski, J., Willey, D., Williams, S., Wilson, R. K., Winter, E., Worley, K. C., Wyman, D., Yang, S., Yang, S. P., Zdobnov, E. M., Zody, M. C. and Lander, E. S. (2002). "Initial sequencing and comparative analysis of the mouse genome." Nature **420**(6915): 520-562.

Watson, L. and Dallwitz, M. J. (1992). "The grass genera of the world: descriptions, illustrations, identification, and information retrieval; including synonyms, morphology, anatomy, physiology, phytochemistry, cytology, classification, pathogens, world and local distribution, and references." Version: 28th November 2005. http://delta-intkey.com.

Weber, A. P., Weber, K. L., Carr, K., Wilkerson, C. and Ohlrogge, J. B. (2007). "Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing." Plant Physiol **144**(1): 32-42.

Weber, J. L. and Myers, E. W. (1997). "Human whole-genome shotgun sequencing." Genome Res **7**(5): 401-409.

Wei, F., Coe, E., Nelson, W., Bharti, A. K., Engler, F., Butler, E., Kim, H., Goicoechea, J. L., Chen, M., Lee, S., Fuks, G., Sanchez-Villeda, H., Schroeder, S., Fang, Z., McMullen, M., Davis, G., Bowers, J. E., Paterson, A. H., Schaeffer, M., Gardiner, J., Cone, K., Messing, J., Soderlund, C. and Wing, R. A. (2007). "Physical and Genetic Structure of the Maize Genome Reflects Its Complex Evolutionary History." PLoS Genet **3**(7): e123.

Wheelan, S. J., Church, D. M. and Ostell, J. M. (2001). "Spidey: a tool for mRNA-to-genomic alignments." Genome Res **11**(11): 1952-1957.

Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y. J., Makhijani, V., Roth, G. T., Gomes, X., Tartaro, K., Niazi, F., Turcotte, C. L., Irzyk, G. P., Lupski, J. R., Chinault, C., Song, X. Z., Liu, Y., Yuan, Y., Nazareth, L., Qin, X., Muzny, D. M., Margulies, M., Weinstock, G. M., Gibbs, R. A. and Rothberg, J. M. (2008). "The complete genome of an individual by massively parallel DNA sequencing." Nature **452**(7189): 872-876.

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Geer, L. Y., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Ostell, J., Miller, V., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R. L., Tatusova, T. A., Wagner, L. and Yaschenko, E. (2007). "Database resources of the National Center for Biotechnology Information." Nucleic Acids Res **35**(Database issue): D5-12.

Whitelaw, C. A., Barbazuk, W. B., Pertea, G., Chan, A. P., Cheung, F., Lee, Y., Zheng, L., van Heeringen, S., Karamycheva, S., Bennetzen, J. L., SanMiguel, P., Lakey, N., Bedell, J., Yuan, Y., Budiman, M. A., Resnick, A., Van Aken, S., Utterback, T., Riedmuller, S., Williams, M., Feldblyum, T., Schubert, K., Beachy, R., Fraser, C. M. and Quackenbush, J. (2003). "Enrichment of gene-coding sequences in maize by genome filtration." Science **302**(5653): 2118-2120.

Wicker, T., Guyot, R., Yahiaoui, N. and Keller, B. (2003a). "CACTA transposons in Triticeae. A diverse family of high-copy repetitive elements." Plant Physiol **132**(1): 52-63.

Wicker, T. and Keller, B. (2007). "Genome-wide comparative analysis of copia retrotransposons in Triticeae, rice, and Arabidopsis reveals conserved ancient evolutionary lineages and distinct dynamics of individual copia families." Genome Res **17**(7): 1072-1081.

Wicker, T., Matthews, D. E. and Keller, B. (2002). "TREP: a database for Triticeae repetitive elements." Trends in Plant Science **7**(12): 561-562.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P. and Schulman, A. H. (2007). "A unified classification system for eukaryotic transposable elements." Nat Rev Genet **8**(12): 973-982.

Wicker, T., Stein, N., Albar, L., Feuillet, C., Schlagenhauf, E. and Keller, B. (2001). "Analysis of a contiguous 211 kb sequence in diploid wheat (Triticum monococcum L.) reveals multiple mechanisms of genome evolution." Plant J **26**(3): 307-316.

Wicker, T., Yahiaoui, N., Guyot, R., Schlagenhauf, E., Liu, Z. D., Dubcovsky, J. and Keller, B. (2003b). "Rapid genome divergence at orthologous low molecular weight glutenin loci of the A and A(m) genomes of wheat." Plant Cell **15**(5): 1186-1197.

Wicker, T., Zimmermann, W., Perovic, D., Paterson, A. H., Ganal, M., Graner, A. and Stein, N. (2005). "A detailed look at 7 million years of genome evolution in a 439 kb contiguous sequence at the barley Hv-eIF4E locus: recombination, rearrangements and repeats." Plant J **41**(2): 184-194.

Wolfe, K. H., Gouy, M., Yang, Y. W., Sharp, P. M. and Li, W. H. (1989). "Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data." Proc Natl Acad Sci U S A **86**(16): 6201-6205.

Wollenweber, B., Porter, J. R. and Lubberstedt, T. (2005). "Need for multidisciplinary research towards a second green revolution." Curr Opin Plant Biol **8**(3): 337-341.

Wu, C. H., Apweiler, R., Bairoch, A., Natale, D. A., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Mazumder, R., O'Donovan, C., Redaschi, N. and Suzek, B. (2006). "The Universal Protein Resource (UniProt): an expanding universe of protein information." Nucleic Acids Res **34**(Database issue): D187-191.

Wu, J., Fujisawa, M., Tian, Z., Yamagata, H., Kamiya, K., Shibata, M., Hosokawa, S., Ito, Y., Hamada, M., Katagiri, S., Kurita, K., Yamamoto, M., Kikuta, A., Machita, K., Karasawa, W., Kanamori, H., Namiki, N., Mizuno, H., Ma, J., Sasaki, T. and Matsumoto, T. (2009). "Comparative analysis of complete orthologous centromeres from two subspecies of rice reveals rapid variation of centromere organization and structure." Plant J **60**(5): 805-819.

Wu, T. D. and Watanabe, C. K. (2005). "GMAP: a genomic mapping and alignment program for mRNA and EST sequences." Bioinformatics **21**(9): 1859-1875.

Yamada, K., Lim, J., Dale, J. M., Chen, H., Shinn, P., Palm, C. J., Southwick, A. M., Wu, H. C., Kim, C., Nguyen, M., Pham, P., Cheuk, R., Karlin-Newmann, G., Liu, S. X., Lam, B., Sakano, H., Wu, T., Yu, G., Miranda, M., Quach, H. L., Tripp, M., Chang, C. H., Lee, J. M., Toriumi, M., Chan, M. M., Tang, C. C., Onodera, C. S., Deng, J. M., Akiyama, K., Ansari, Y., Arakawa, T., Banh, J., Banno, F., Bowser, L., Brooks, S., Carninci, P., Chao, Q., Choy, N., Enju, A., Goldsmith, A. D., Gurjal, M., Hansen, N. F., Hayashizaki, Y., Johnson-Hopson, C., Hsuan, V. W., Iida, K., Karnes, M., Khan, S., Koesema, E., Ishida, J., Jiang, P. X., Jones, T., Kawai, J., Kamiya, A., Meyers, C., Nakajima, M., Narusaka, M., Seki, M., Sakurai, T., Satou, M., Tamse, R., Vaysberg, M., Wallender, E. K., Wong, C., Yamamura, Y., Yuan, S., Shinozaki, K., Davis, R. W., Theologis, A. and Ecker, J. R. (2003). "Empirical analysis of transcriptional activity in the Arabidopsis genome." Science **302**(5646): 842-846.

Yamaguchi-Shinozaki, K. and Shinozaki, K. (1994). "A novel cis-acting element in an Arabidopsis gene is involved in responsiveness to drought, low-temperature, or high-salt stress." Plant Cell **6**(2): 251-264.

Yamamoto, T., Yonemaru, J. and Yano, M. (2009). "Towards the understanding of complex traits in rice: substantially or superficially?" DNA Res **16**(3): 141-154.

Yan, L., Loukoianov, A., Tranquilli, G., Helguera, M., Fahima, T. and Dubcovsky, J. (2003). "Positional cloning of the wheat vernalization gene VRN1." Proc Natl Acad Sci U S A **100**(10): 6263-6268.

Yeh, R. F., Lim, L. P. and Burge, C. B. (2001). "Computational inference of homologous gene structures in the human genome." Genome Res **11**(5): 803-816.

You, F. M., Luo, M. C., Gu, Y. Q., Lazo, G. R., Deal, K., Dvorak, J. and Anderson, O. D. (2007). "GenoProfiler: batch processing of high-throughput capillary fingerprinting data." Bioinformatics **23**(2): 240-242.

Youens-Clark, K., Faga, B., Yap, I. V., Stein, L. and Ware, D. (2009). "CMap 1.01: a comparative mapping application for the Internet." Bioinformatics **25**(22): 3040-3042.

Yu, J., Hu, S., Wang, J., Wong, G. K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., Cao, M., Liu, J., Sun, J., Tang, J., Chen, Y., Huang, X., Lin, W., Ye, C., Tong, W., Cong, L., Geng, J., Han, Y., Li, L., Li, W., Hu, G., Huang, X., Li, W., Li, J., Liu, Z., Li, L., Liu, J., Qi, Q., Liu, J., Li, L., Li, T., Wang, X., Lu, H., Wu, T., Zhu, M., Ni, P., Han, H., Dong, W., Ren, X., Feng, X., Cui, P., Li, X., Wang, H., Xu, X., Zhai, W., Xu, Z., Zhang, J., He, S., Zhang, J., Xu, J., Zhang, K., Zheng, X., Dong, J., Zeng, W., Tao, L., Ye, J., Tan, J., Ren, X., Chen, X., He, J., Liu, D., Tian, W., Tian, C., Xia, H., Bao, Q., Li, G., Gao, H., Cao, T., Wang, J., Zhao, W., Li, P., Chen, W., Wang, X., Zhang, Y., Hu, J., Wang, J., Liu, S., Yang, J., Zhang, G., Xiong, Y., Li, Z., Mao, L., Zhou, C., Zhu, Z., Chen, R., Hao, B., Zheng, W., Chen, S., Guo, W., Li, G., Liu, S., Tao, M., Wang, J., Zhu, L., Yuan, L. and Yang, H. (2002). "A draft sequence of the rice genome (Oryza sativa L. ssp. indica)." Science **296**(5565): 79-92.

Yu, J., Wang, J., Lin, W., Li, S., Li, H., Zhou, J., Ni, P., Dong, W., Hu, S., Zeng, C., Zhang, J., Zhang, Y., Li, R., Xu, Z., Li, S., Li, X., Zheng, H., Cong, L., Lin, L., Yin, J., Geng, J., Li, G., Shi, J., Liu, J., Lv, H., Li, J., Wang, J., Deng, Y., Ran, L., Shi, X., Wang, X., Wu, Q., Li, C., Ren, X., Wang, J., Wang, X., Li, D., Liu, D., Zhang, X., Ji, Z., Zhao, W., Sun, Y., Zhang, Z., Bao, J., Han, Y., Dong, L., Ji, J., Chen, P., Wu, S., Liu, J., Xiao, Y., Bu, D., Tan, J., Yang, L., Ye, C., Zhang, J., Xu, J., Zhou, Y., Yu, Y., Zhang, B., Zhuang, S., Wei, H., Liu, B., Lei, M., Yu, H., Li, Y., Xu, H., Wei, S., He, X., Fang, L., Zhang, Z., Zhang, Y., Huang, X., Su, Z., Tong, W., Li, J., Tong, Z., Li, S., Ye, J., Wang, L., Fang, L., Lei, T., Chen, C., Chen, H., Xu, Z., Li, H., Huang, H., Zhang, F., Xu, H., Li, N., Zhao, C., Li, S., Dong, L., Huang, Y., Li, L., Xi, Y., Qi, Q., Li, W., Zhang, B., Hu, W., Zhang, Y., Tian, X., Jiao, Y., Liang, X., Jin, J., Gao, L., Zheng, W., Hao, B., Liu, S., Wang, W., Yuan, L., Cao, M., McDermott, J., Samudrala, R., Wang, J., Wong G. K. and Yang, H. (2005). "The Genomes of Oryza sativa: a history of duplications." PLoS Biol **3**(2): e38.

Yuan, Y., SanMiguel, P. J. and Bennetzen, J. L. (2003). "High-C_0t sequence analysis of the maize genome." Plant J **34**(2): 249-255.

Zabala, G. and Vodkin, L. O. (2005). "The wp mutation of Glycine max carries a gene-fragment-rich transposon of the CACTA superfamily." Plant Cell **17**(10): 2619-2632.

Zeng, Q., Chen, X. and Wood, A. J. (2002). "Two early light-inducible protein (ELIP) cDNAs from the resurrection plant Tortula ruralis are differentially expressed in response to desiccation, rehydration, salinity, and high light." J Exp Bot **53**(371): 1197-1205.

Zerbino, D. R. and Birney, E. (2008). "Velvet: algorithms for de novo short read assembly using de Bruijn graphs." Genome Res **18**(5): 821-829.

Zhang, D., Choi, D. W., Wanamaker, S., Fenton, R. D., Chin, A., Malatrasi, M., Turuspekov, Y., Walia, H., Akhunov, E. D., Kianian, P., Otto, C., Simons, K., Deal, K. R., Echenique, V., Stamova, B., Ross, K., Butler, G. E., Strader, L., Verhey, S. D., Johnson, R., Altenbach, S., Kothari, K., Tanaka, C., Shah, M. M., Laudencia-Chingcuanco, D., Han, P., Miller, R. E., Crossman, C. C., Chao, S., Lazo, G. R., Klueva, N., Gustafson, J. P., Kianian, S. F., Dubcovsky, J., Walker-Simmons, M. K., Gill, K. S., Dvorak, J., Anderson, O. D., Sorrells, M. E., McGuire, P. E., Qualset, C. O., Nguyen, H. T. and Close, T. J. (2004). "Construction and evaluation of cDNA libraries for large-scale expressed sequence tag sequencing in wheat (Triticum aestivum L.)." Genetics **168**(2): 595-608.

Zhang, F., Maeder, M. L., Unger-Wallace, E., Hoshaw, J. P., Reyon, D., Christian, M., Li, X., Pierick, C. J., Dobbs, D., Peterson, T., Joung, J. K. and Voytas, D. F. (2010). "High frequency targeted mutagenesis in Arabidopsis thaliana using zinc finger nucleases." Proc Natl Acad Sci U S A **107**(26): 12028-12033.

Zhang, J., Guo, D., Chang, Y., You, C., Li, X., Dai, X., Weng, Q., Chen, G., Liu, H., Han, B., Zhang, Q. and Wu, C. (2007). "Non-random distribution of T-DNA insertions at various levels of the genome hierarchy as revealed by analyzing 13 804 T-DNA flanking sequences from an enhancer-trap mutant library." Plant J **49**(5): 947-959.

Zhao, W., Wang, J., He, X., Huang, X., Jiao, Y., Dai, M., Wei, S., Fu, J., Chen, Y., Ren, X., Zhang, Y., Ni, P., Zhang, J., Li, S., Wang, J., Wong, G. K., Zhao, H., Yu, J., Yang, H. and Wang, J. (2004). "BGI-RIS: an integrated information resource and comparative analysis workbench for rice genomics." Nucleic Acids Res **32**(Database issue): D377-382.

Zhou, J., Wang, X., Jiao, Y., Qin, Y., Liu, X., He, K., Chen, C., Ma, L., Wang, J., Xiong, L., Zhang, Q., Fan, L. and Deng, X. W. (2007a). "Global genome expression analysis of rice in response to drought and high-salinity stresses in shoot, flag leaf, and panicle." Plant Mol Biol **63**(5): 591-608.

Zhou, S., Herschleb, J. and Schwartz, D. C. (2007b). A Single Molecule System for Whole Genome Analysis. Perspectives in Bioanalysis, Elsevier. **Volume 2:** 265-300.

Zhou, S., Wei, F., Nguyen, J., Bechner, M., Potamousis, K., Goldstein, S., Pape, L., Mehan, M. R., Churas, C., Pasternak, S., Forrest, D. K., Wise, R., Ware, D., Wing, R. A., Waterman, M. S., Livny, M. and Schwartz, D. C. (2009). "A single molecule scaffold for the maize genome." PLoS Genet **5**(11): e1000711.

Zhulidov, P. A., Bogdanova, E. A., Shcheglov, A. S., Vagner, L. L., Khaspekov, G. L., Kozhemyako, V. B., Matz, M. V., Meleshkevitch, E., Moroz, L. L., Lukyanov, S. A. and Shagin, D. A. (2004). "Simple cDNA normalization using kamchatka crab duplex-specific nuclease." Nucleic Acids Res **32**(3): e37.

Zimmermann, P., Hirsch-Hoffmann, M., Hennig, L. and Gruissem, W. (2004). "GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox." Plant Physiol **136**(1): 2621-2632.

Zou, C., Lehti-Shiu, M. D., Thibaud-Nissen, F., Prakash, T., Buell, C. R. and Shiu, S. H. (2009). "Evolutionary and expression signatures of pseudogenes in Arabidopsis and rice." Plant Physiol **151**(1): 3-15.