Genomic insights into sex determination evolution in yam, an important staple food crop

Benjamen Henry White

This thesis is submitted for the degree of

Doctor of Philosophy

University of East Anglia

School of Biological Sciences

September 2018

This thesis is dedicated to the loving memory of:

Tony Lee

Grandad, you were a hero to many and truly supported me right up until the end of this process, even through you didn't understand most of what I waffled on about.

You are missed but never forgotten.

Norman White

Pops, without you none of this would of been possible. I'm sorry you never got to see where I'd end up, but I hope I did you proud.

You are missed but never forgotten.

Jane Hedgeland

You were only in my life a short while, but were a wonderful person, especially to my Father.

You are missed but never forgotten.

"*I yam what I yam, and that's all what I yam.*"

— Popeye the Sailor

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other University. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text. This dissertation contains fewer than 100,000 words including, footnotes, tables and bibliography.

Benjamen Henry White

Acknowledgements

I am exceptionally thankful to my secondary supervisor, Sophien Kamoun, for putting me in touch with Ryohei Terauchi, who was been fundamental in the progress of my PhD and has made it possible for me to write this thesis. I'd also like to thank the rest of the Terauchi group for their assistance and for accommodating me in the collaborations we share so well. I'd especially like to thank my supervisor, Wilfried Haerty, for adopting myself and my project, and being so supportive and going beyond what would be expected to guide me through my PhD. I'd also like to additionally thank the Haerty and Di Palma labs for letting me their help and advice throughout. Jose De Vega should be mentioned as he really guided me through my initial teething into the process of annotation and assembly early on. I'd also like to thank Ranjana Bhattacharjee for introducing me to yam and for access to materials. My industrial partner, Eagle Genomics Ltd, in particular William Spooner and Daniel Barrell, should be thanked for their support of my BBSRC iCASE PhD and placement with them. I'm also grateful to those who initially gave me the opportunity to partake this PhD. And of course my friends, family, colleagues, and everyone lost and found along the way for their support and advice that helped me reach the completion of these past four years.

Abstract

Reductions in the cost of next generation sequencing and expertise required for whole genome assembly and annotation permits improvement of existing assemblies of industrially important models (Chinese hamster ovary cell line - CHO) and sequencing neglected agronomically important species, such as yam. Applying these new technologies, we have produced an improved reference for the CHO lineage, CHO-K1, and generated draft assemblies and annotations for three yam species. Yam is an important staple crop of great cultural and socioeconomic significance to Africa, the Americas, the Caribbean, South Pacific and Asia. I explored the evolutionary history of sex determination in dioecious *Dioscorea* species, a rare trait found in only 5-6% of angiosperms. We identified the most socio-economically important species, guinea yam (*D. rotundata*) to be female heterogametic (ZW), and confirmed the related basal species, oni-dokoro (*D. tokoro*), to be male heterogametic (XY). It is exciting to observe both ZW and XY sex determination systems in *Dioscorea*, as this indicates turnover of sex determination systems. There has been little study to date comparing plant species in the same genus with different sex determination systems, making *Dioscorea* a unique opportunity to investigate the turnover of sex determination. Through comparison of these two species, and generation of a draft reference for *D. alata*, I have begun to elucidate the ancestral state of sex within the genus. Generation of these genomic resources in yam and study of the evolution of sex determination, will assist with breeding programmes that will improve this important staple food crop. Finally, these findings will assist with future studies that aim to improve our fundamental understanding of the mechanisms of recombination and speciation in plants.

Contents

List of Figures

List of Tables

Chapter 1

General Introduction

An impressive array of different inbred and outbred mating systems have evolved in plants. These mating systems have a major impact not only on the ecology and distribution of plants, but also their genomes. Primarily these mating systems can be thought of in terms of being asexual, cosexual or unisexual. Of these, hermaphroditic or cosexual individuals, that produce both male and female gametes, may or may not have mechanisms that will influence the rate self-fertilisation vs outcrossing[1]. Outcrossing refers to the process of exchanging genetic information between non-related individuals that likely results in increased variation and combinations of alleles that may otherwise not have occurred[1]. This can have an impact on fitness, which is defined as an individual's ability to compete for resources and successfully reproduce, passing genetic information onto its progeny. This process of exchanging genetic information between individuals ultimately results in increased variation and combinations of alleles that may otherwise not have occurred, resulting in a change in productivity of the population.

Focusing on seed plants, specifically those that flower (angiosperms), the mating system of the species and associated sex function are often defined phenotypically based on the arrangement of male (stamen) pollen producing, and female (carpels) ovule producing, reproductive organs during inflorescence (Figure 1.1). Pollen is borne within the male gametophyte and leaves the structure it is produced in, while ovules produced by the female gametophyte remain stationary. The purpose of flowers is therefore ultimately to facilitate reproduction. Delivery of pollen to the carpel is dependant on the mechanism of pollination, such as anemophily (wind-pollinated) or by a pollinator. In the case of dioecy and unisexual populations, androecious (male) individuals only have flowers with fully developed stamen and gynoecious (female) individuals only have flowers with fully developed carpels. Conversely, in asexual and cosexual individuals,

such as hermaphrodites, both reproductive organs are represent on the same "perfect" flower or, in the case of monoecious individuals, both organs present on different flowers of the same plant. Some other scenarios of cosexual involve having a combination of hermaphroditic flowers and either female (gynomonoecious) or male (andromonoecious) on the same individual.

Mating system are key life history traits that distinguish sexual from cosexuals, and outcrossing from inbreeding, and ultimately play a role in determining the fitness of a population. All mating systems aim to promote fitness and reproductive success within the constraints of their initial environment and selective pressures. The mating system of a population can be determined by the sexual system of individuals within the population, e.g. hermaphroditism or dioecy (separate sexes). These systems are evolutionarily liable to respond to natural selection and other selective pressures. As such, there are many different mating systems in angiosperms[2]. Differences in the ability a mating system grants to respond to natural selection and inbreeding depression will likely impact the length of time it persists within a lineage[3].

In the following sections I will briefly explore the fundamentals of different mating systems and sexes in angiosperms, and then move onto how unisex and sex determination can evolve.

1.1 Asexuality and self-fertilisation

Asexual populations can avoid the cost of sex, that is the cost of males and females or meiotic reproduction in general, as only around half the genome is transmitted in sexual populations[3]. Similarly in hermaphroditic or cosexual populations, inbreeding through selfing can occur as a form of clonal propagation. Both asexual reproduction and selfing provide a transmission advantage, in which there is an increase in the transmission of gametes to the next generation and an initial advantage given to alleles that promote self-fertilisation. This advantage is especially prevalent when selfing is occurring in an outcrossing population, as outcrossing individuals will often act as a parent only half the time, whereas selfing individuals transmit the entirety of their genome onto their progeny[1]. Additionally, outcrossing bears an additional cost compared to selfing in that outcrossers can only act as pollen parents to outcrossing individuals, but cannot serve as pollen parents to individuals that decide to self-fertilise, causing the rate of outcrossing to decline[3]. As such, mutations that promote selfing will provide a transmission advantage over those promoting outcrossing, making them likely to spread within a population to

fixation, unless there is a sufficiently strong selection pressure against the benefits of selfing[3]. Fixation refers to the point at which a single mutation becomes dominant and all other variants at the same locus are lost from the gene pool (all genetic information in a population).

Selfing also provides reproductive assurance, as there is no need for other mates that may be too distant or unreachable, due to sparsity of pollinators, as an example[3]. When there is a lack of pollinators or inefficient transfer of pollen, selfing can provide a viable means of reproductive assurance. It's therefore plausible that complete selfing can evolve and reach fixation in the population, unless there is a strong selective force against it.

Selfing and asexual mating systems have evolved repeatedly across the majority of angiosperm lineages. Hermaphroditic organisms more often transition to self-fertilisation from outcrossing, as one of the most common changes observed in mating systems. Selfing species also have a high degree of variation between populations, but not within, which is the opposite of what could be expected in outcrossing populations. Larger species, such as trees, are less likely to be selfing as floral displays are normally large and it would mean a lot of opportunity for interference, self-incompatibility and dioecy can avoid this. Smaller and shorter lives species tend to favour selfing as a means of reproductive assurance, as well as those with ephemeral habitats, which is most common in annual plants[2].

Three factors are thought to limit the spread of selfing alleles in a population, these are inbreeding depression, pollen discounting (reduction in the amount of pollen available for outcrossing; male fitness is reduced), and seed discounting (reduction in seed production through outcrossing, also reduces transmission of selfing allele)[2].

1.1.1 Inbreeding depression, hitchhiking, background selection and the Red Queen

The main disadvantage of inbreeding is inbreeding depression, which increases the likelihood of recessive traits manifesting. Recessive alleles can be deleterious, leading to reduced fitness, and overdominance caused by the fitness advantage of heterozygotes[1, 3]. Selfing and asexuality reduces the effective population size and diversity of genotypes on which natural selection can act, limiting the ability of a population to adapt to it. As a results, deleterious mutations are more likely to spread to fixation, reducing the fitness of the population and size. This will further increase the chance of additional deleterious alleles to reach fixation, as part of a 'mutational meltdown'[3].

When natural selection cannot efficiently remove all recessive alleles, purging becomes one of the only options left to remove them from the population, such as through lethal phenotypes[3]. Over an individual's life time, gametic mutations will occur, adding to the variance seen in their progeny. These mutations can be advantageous, increasing an individuals likelihood to reproduce and for the allele to become fixed within the population. But, mutations can also be deleterious or not offering much advantage in terms of fitness, perhaps being silent or neutral. Several evolutionary theories have been proposed for how mutations are selected, all of which are fundamentally impacted by the process of recombination.

Deleterious mutations are able to hitchhike with positive or neutral mutations, and equally go to fixation. This random (stochastic) force hypothesised, by Smith and Haigh, expands upon neutral mutation theory by proposing that advantageous alleles undergoing selective sweep (reduction in variation near a selected mutation) can inadvertently increase the frequency of neighbouring neutral polymorphisms that are not directly under selection[4]. This phenomenon causes linked alleles, regardless of their fitness advantage, to essentially 'hitchhike', thereby reducing heterozygosity within the loci and sweeping towards fixation. The frequency of hitchhiking is dependent on population size under this model. Recombination is thought to put breaks on the drive of hitchhiking polymorphisms towards fixation, as hitchhiking has little effect on the frequency of linked alleles that are distant or if the linkage is broken[5]. The processes of genetic recombination is thought to avoid Muller's ratchet, by alleviating genetic load in subsequence generations through allowing selection to act of deleterious mutations[6, 7]. However, not all variation is advantageous and loss of recombination in particular can lead to a build up of deleterious mutations. A number of models in addition to genetic drift, which explains the changes in allele frequency of a population over generations due to random sampling, have attempted to explain the evolutionary forces acting on the frequency of these mutation. Hartfield and Glémin explored the potential for deleterious alleles to hitchhike in asexual and cosexual species, and proposed that dominant alleles are more likely to hitchhike than recessive alleles[8]. As they will go to fixation quicker, increasing the frequency of their haplotype and making recombination more likely to break linkage with the hitchhiking allele. As such, inbreeding, which lacks recombination, will increase the chance of deleterious alleles hitchhiking to fixation due to reduced recombination, compared to sexual species. However, beneficial mutations are more readily fixed in selfing populations. From this, the authors proposed that breeding strategies should therefore aim for an intermediate selfing rate, as a trade off between both potential consequences of hitchhiking[8]. Innan

and Stephen found, while hitchhiking was most powerful in outcrossing species, the alternative theory of background selection was better suited to explain this correlation in individuals prone to selfing[9]. As both hitchhiking and background selection are two major stochastic forces for describing the positive correlation between variation and recombination, the later seen as the most important in deciding the amount of neutral variation.

Proposed by Charlesworth, *et al*, background selection is where selection in a population actively takes place against deleterious alleles and subsequently sweeps neighbouring (background) regions of neutral variation, reducing heterozygosity[10]. Background selection assumes that deleterious polymorphisms occur frequently, whereas fitness increasing alleles are rarer. In a population of asexual individuals, that lack genetic recombination, over multiple generations a build up of deleterious mutations will lead to inbreeding depression as per Muller's ratchet[6, 7]. In this scenario, just like a ratchet that can only turn in one direction, every subsequent generation is doomed to accumulate more deleterious mutations until selection can remove them or eventual extinction. An alternative scenario could be where both asexual progeny gain a different, but equally, advantageous mutation resulting in competition between the two; clonal interference[11, 12]. Outcrossing, sex between two different individuals, is thought to act as a means of purging the genome of potentially deleterious mutations that would otherwise remain in asexual populations. In this case, recombination, the shuffling of genes between chromosomes, is thought to expose mutations to the forces of natural selection and consequently result in deleterious mutations being unlikely to become fixed and thereby avoiding clonal interference. Hill and Robinson, 1966, proposed that recombination provides an evolutionary advantage as it can lead to a gain of multiple advantageous alleles in an individual that may otherwise not occur together, increasing fitness[13]. As such, sex increases the fixation of beneficial mutations and accelerates adaption of a population[11].

Pollen discounting can further increase inbreeding depression. Pollen used in selfing is consequently not made available for competitive outcrossing; pollen discounting. While ovules similarly used in selfing are also discounted from outcrossing, seeding discounting, that is also costly. Highly selfing individuals may not sire any outcrossing seeds on other plants, potentially being the result of a complete loss of male fitness that has subsequently abolished or reduced the advantages of selfing[1]. This situation can arise from a mutation that affects flower morphology, e.g. making the smaller, such that there is little separation of anther and stigma. As a result, pollen is made less available to pollinators for outcrossing.

Finally, one other pressure on the evolution and maintenance of mating system and sex of a species comes in the form of the Red Queen hypothesis. This hypothesis considers that individuals are in a constant evolutionary arms race to compete with one another and their environment[14, 15]. Empirical evidence for this has been demonstrated in the form of parasitic infections, where parasitism increases the frequency of recombination and survival of sexual, compared to asexual, individuals[16, 17]. This is also true of plants where a constant evolutionary arms race against pathogens takes place, in which both plants and pathogens could capitalise on sex to increase variation to adapt new specialised defence mechanisms. Effector triggered immunity (ETI) is a potential example of this, as plants have to constantly evolve new resistance genes to recognise pathogen effectors (proteins that aid colonisation of the host), while pathogens must also constantly evolve new efforts to outcompete the plant host's immunity[18]. Recombination could therefore result in shuffling of mutations associated with resistance genes, producing allelic variation that could confer resistance to specific pathogen effectors. However, specialisation of a pathogen to a particular plant genotype does not also mean losing the ability to successfully attack other genotypes[19]. Furthermore, while sex has been shown to increase defence against specialist herbivores, it can also increase susceptibility to generalist herbivores compared to asexuals[20]. Finally, adaptation to the environment may be the potential overall deciding factor in the eventual extinction of a species, regardless of an individuals ability to generate new genotypes and adapt[21].

1.1.2 Mechanisms to prevent self-fertilisation

In response to the effects of selfing on fitness, plants have evolved multiple systems to limit self-fertilisation. In self-incompatibility an S-locus is responsible for protein-protein interactions between the haplotypes on male and female gametophytes, that can in term inhibit selfing through halting pollen tube growth or embryogenesis[1]. However, self-incompatibility is not a self-recognition and requires two coadapted alleles at the incompatibility loci; one for the receptor and another for the ligand[1]. Self-incompatibility is also not to be confused with herkogamy, a form of heterostyly, or dichogamy. These can reduce the rate of selfing in hermaphrodites through separation of pistil and stigma within and/or between flowers spatially (herkogamy) or temporal separation (dichogamy).

In herkogamy, a population may have two or three different types of flower morphology, that can be distinguished by the length of stamen and pistil, and can also be linked to genes associated with self-incompatibility[1, 22]. In which, the stamens can be longer,

short or intermediate, in comparison to the pistils. This acts as an adaptation to pollination by different pollinators. As an example, in some species the stigma can be longer than anthers in an attempt to attract pollinators to the stigma first, before taking pollen from the anther; aiding pollen export[22]. Additionally, pollen from one floral morphology often cannot fertilise another individual with the same flower morphology, due to self-incompatibility[22].

This strategy employed in herkogamy is different to that of dichogamy. In dichogamy there is a temporal, rather than spacial, separation of anther and stigma. Whereby there is a minimal overlap in the presence of stigma and anther on each inflorescence. As a result, selfing cannot occur due to the absence of either sex organ at any one time, and pollen discounting and geitonogamy can be thereby be reduced[22]. Both herkogamy and dichogamy systems encourage outcrossing and allogamy, helping the avoid inbreeding depression as a result.

In principle, the evolution of these mechanism to avoid self-fertilisation is similar to that of full unisexuality, dioecy, in that two mutations are required; one affecting the expression of femaleness and another of maleness[1].

1.2 Evolution of dioecy

1.2.1 Two genes model of sex determination

One possible scenario leading to dioecy is a mutation in a single locus that is responsible for the allocation of reproductive resources, causing the resource allocation to become unbalanced and entirely/mostly favour either sex[23]. This leads to individuals with alleles dedicated to either male or female sex function. Alternatively, Charlesworth and Charlesworth hypothesised a two genes model that requires mutations in two or more separate alleles that can lead to dioecy[24]. In this model, at least two mutations with complementary dominance are required, one to make a female and another to make a male, often through genes that regulate gynoecium and androecium development[25]. Regulation of maleness and femaleness through sterility factors is essential to achieving unisexual male or female heterogametic individuals. In a population of cosexual individuals, gynodioecy, through establishment of a male sterility mutation, is more likely than one in which gynodioecy evolves first[23]. As selfing in gynodioecious individuals reduces the availability of ovules for outcrossing with cosexuals, and therefore, androdioecious individuals in a population of partially self-fertilising cosexuals cannot greatly gain

outcrossing opportunities[23]. Gynodioecious individuals are also self-incompatible and cannot produce seeds through self-fertilisation, avoiding potential inbreeding depression. As such, androdioecy is rare and is through to appear most often to evolve via mutations through which females in dioecious populations gain some male function, and full dioecy breaks down[23].

It is therefore most likely that first a loss-of-function mutation in the allele responsible for maleness results in a male-sterility factor will create females[23]. A second mutation will then occur in another allele on a different haplotype, creating a female suppressor or male enhancer that can overcome the male-sterility factor, leading to a predominantly male individual.

As such, many studies have looked at the expression of ABC model genes in the developmental pathways of unisex flowers on dioecious and monoecious species[26]. Some empirical evidence to directly support the two gene model has been observed in dioecious persimmon (*Diospyros lotus*), as sex determination is controlled through a small RNA encoding transcription factor (*OGI*; Oppressor of *MeGI*) on the MSY that targets the autosomal gene Male Growth Inhibitor (*MeGI*), which thought to promote androecia sterility[27]. Although this model may not completely suffice the hypothesis of both male and female sterility genes being present on the same chromosome, other species with less well defined sex determination, such as those in the dioecious *Phoenix* genus, have shown this[28].

1.2.2 Genetic and environmental sex determination

There are many diverse factors that can impact or decide the sex of an individual, such as temperature dependency, infections, lifestyle, social cues and other environment environmental effects[29]. As such, sex may not always been decided by genotypic sex determination (GSD), where an individual is predetermined by their genotype to develop as male or female, and instead can be determined by the environment in the case of environmental sex determination (ESD). African oil palm (*Elaeis guineensis*) is a monoecious angiosperm that is been shown to cycle between unisexes, with this change impacted by environment factors, such as water stress, that have a knock-on effect on downstream genetic factors that regulate male and female inflorescence[30]. The line between GSD and ESD if often blurred, where ESD is thought to occur more often in environments that offer benefit to one sex more than another. GSD is thought to occur more often in environments that are not varied enough for a fitness difference between sexes or where the environment is unpredictable, as this would unbalance ESD

or potentially lead to intersex and neuter individuals[29]. Selection on sex ratio can trigger transition between ESD and GSD[29].

1.2.3 Molecular evolution of sex chromosomes

Sex in dioecious plants can be thought of as being XY or ZW sex, where males (XY/ZZ) and females (XX/ZW), respectively, are the limiting (maternal) sex (Figure 1.1a, b). While not reviewed here, as no UV sex determination has been reported in angiosperms, an example of this can be found in Bryophytes, such as *Marchantia polymorpha*[31]. This species is dioecious, with X and Y chromosomes present, but the sex of an individual is dependent on the life cycle stage of the parents (either a haploid gametophyte or diploid sporophyte). Gametophytes can reproduce asexually, but in the presence of water motile sperm are able to migrate from the male to female sex organs and fertilise the egg cell (Figure 1.1c). After fertilisation the zygote will develop into a sporophyte that is still attached to the parent plant, and will produce spores for germination. The sex phenotype in dioecious species is often associated with sex chromosomes, harbouring the underlying genetic factors that determine the sex of an individual. Sex chromosomes often arise from an ancestral pair of non-sex autosomes, that will often follow a set evolutionary trajectory towards becoming strikingly heterochromatic sex chromosomes, as those seen in mammals[32]. Establishment of dioecy is preceded by the emergence of sex chromosomes. A review by Ming, *et al*, concluded that we can now begin to recognise distinct stages of sex chromosome evolution in angiosperms, after divergence from the ancestral autosomes[33, 34]. Here we will visit the potential evolutionary trajectory of sex chromosomes, as five potential stages (Figure 1.3). In the first stage, sex chromosomes start out as a normal pair of ancestral autosomes belonging to a cosexual individual. For simplicity we will follow the potential evolution of an XY pair of sex chromosomes, although the same model can be applied to female heterogametic sex determination, but with Z-linked female sterility and W linked male-sterility genes present.

The presence of the male-sterility factor and female suppressor/male enhancer act as sexually antagonistic alleles present on separate haplotypes, forming the proto-X and proto-Y chromosomes (Stage 2). Neuters are also possible in this scenario, whereby recombination between these two chromosomes could cause inheritance of both a male-sterility factor and female suppressor in a single individual[23].

Through recombination or mutation, both dominant sexually antagonistic alleles become present on the same proto-Y chromosome, that stabilises in the population. Linkage of these alleles form an evolutionary strata, the male-specific male-determining

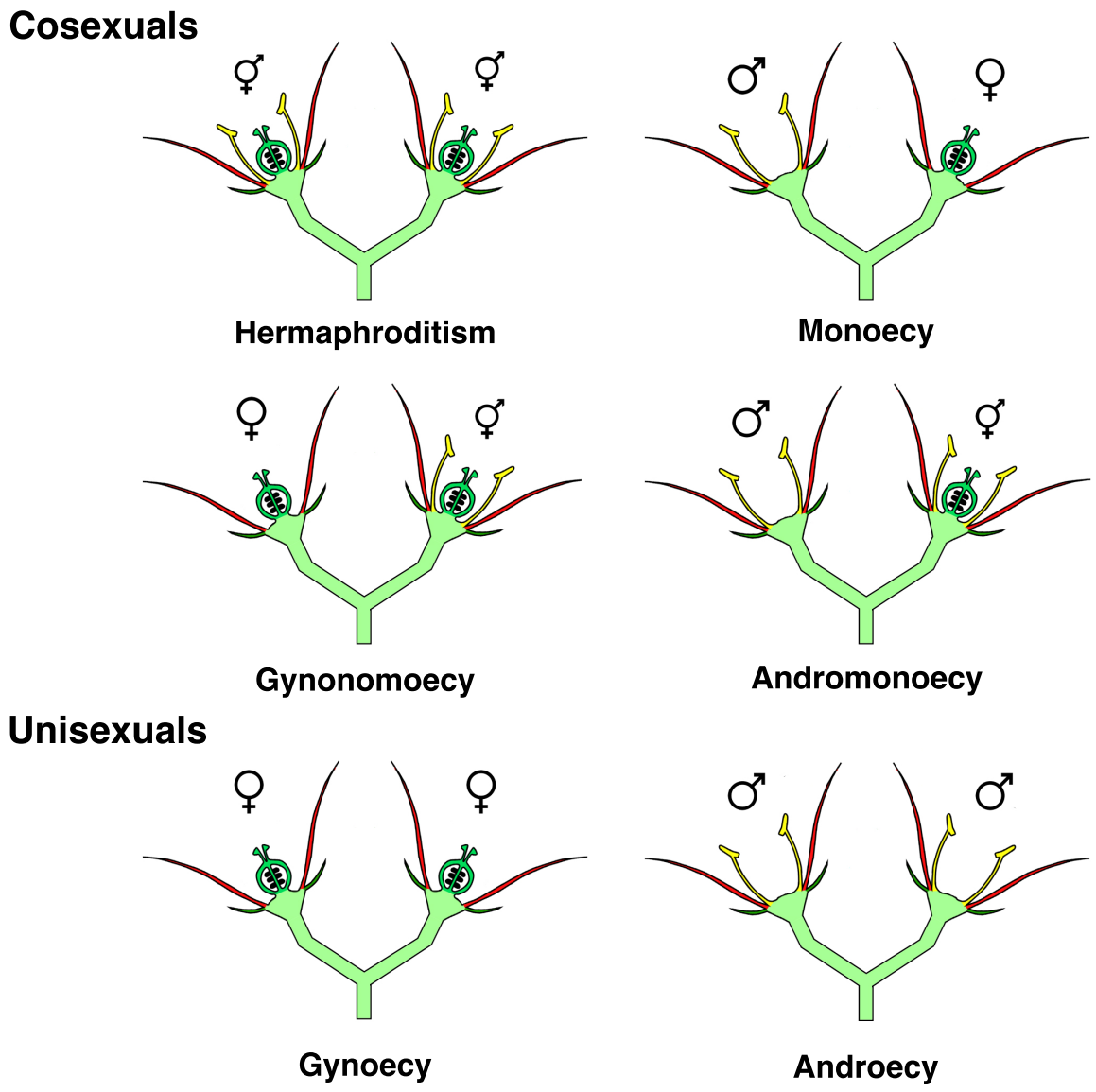

Figure 1.1: Various types models of sex in angiosperms. Hermaphrodites have flowers with both sets of reproductive organs on the same flower, while in monoecy sexual phenotypes develop later in life, forming inflorescence of male and/or female flowers on the same plant. Gynomonoecious and andromonoecious populations, have either female or male, and hermaphrodite flowers, respectively, present. Unisexual individuals are either gynoecious (female only flowers) or androecious (male flowers). In dioecious populations, both gynoecious and androecious phenotypes are seen, as the male and females of the species. Figure does not show models of sub-dioecy, where male, female and cosexuals are present in a population, or trimonoecy (male, female and hermaphrodite flowers on same pant).

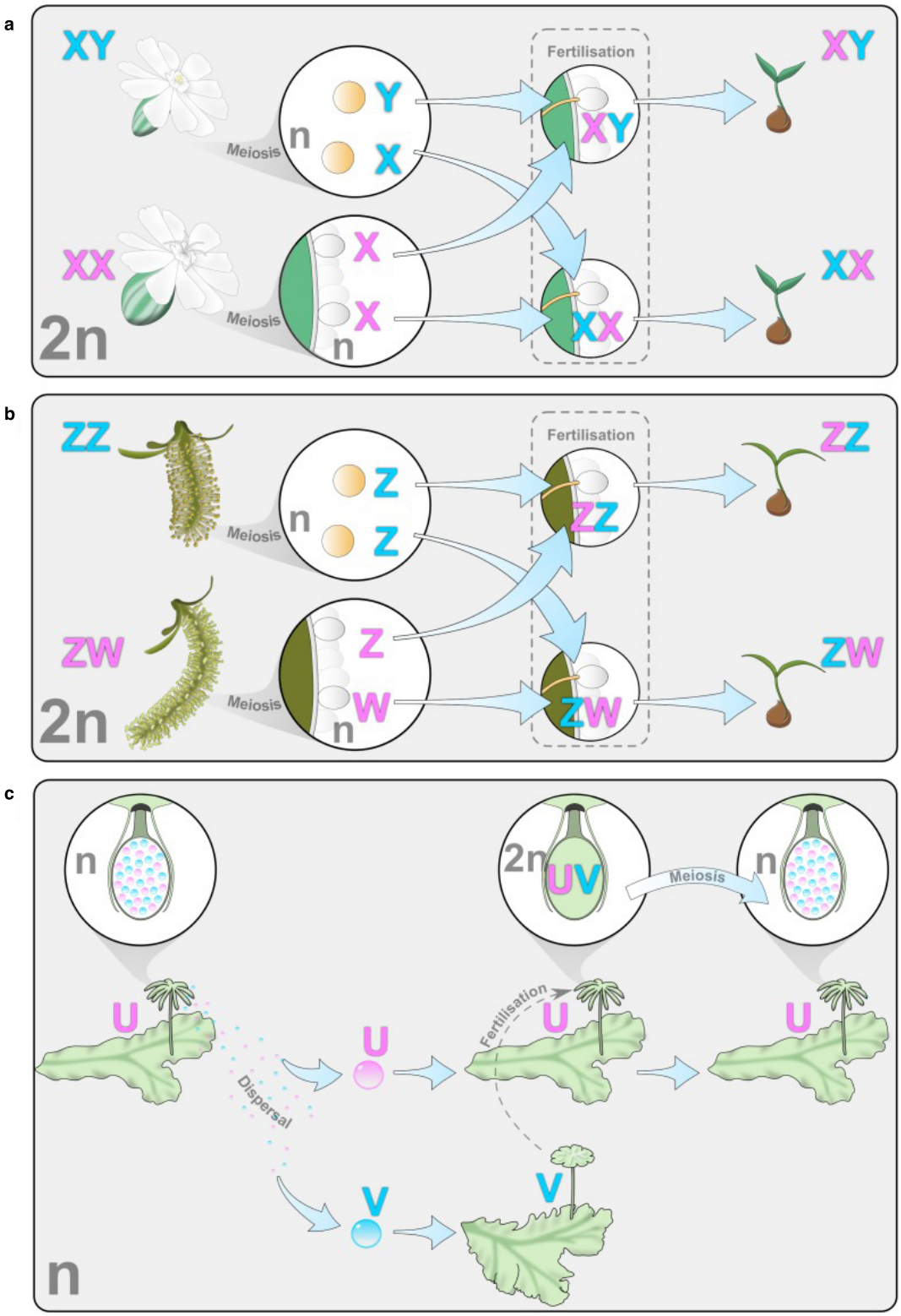

Figure 1.2: Examples of sex chromosome systems in land plants. **a.** XY: male heterogametry (*Silene latifolia*), **b.** ZW: female heterogametry (*Salix suchowensis*) and **c.** UV: haplo-diploid system (*Marchantia polymorpha*), showing maternal (pink) and paternal (blue) sex chromosomes. Not shown are potential X0 sex chromosomes, where the Y/W chromosome has been lost through degradation. **This figure has been reproduced from from A. Muyle, *et al*, 2017, with permission from Oxford University Press.**

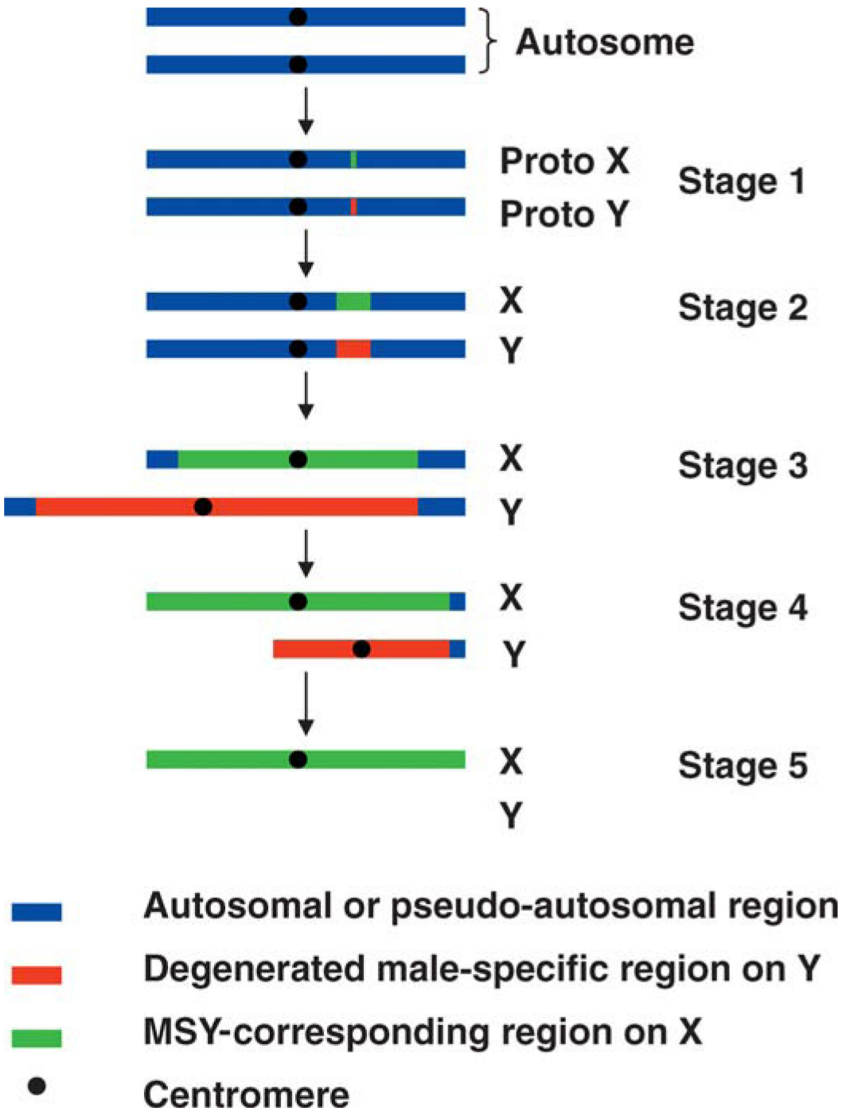

Figure 1.3: Potential model of sex chromosome evolution, using XY system as example for later stages. Stage 1. Sterility mutation causes emergence of sexually antagonist allele on the autosomes of two cosexual individuals in regions for regulating development of male/femaleness, such as an advantageous male sterility. Stage 3. Formation of male-specific region (red; corresponding region on X in green) as mutation suppressing femaleness or promoting maleness occurs, and evolutionary strata, the male-specific male-determining region. Recombination begins to become suppressed in pseudoautosomal regions and new strata form, extending the male-specific region. Stage 4. Transposable elements begin to accumulate along the chromosome, causing an increase in length of the chromosome. Stage 5. Degradation of chromosome through pseudogensiation of genes and selection against nonfunctional DNA. Non-recombined region spreads throughout much of the chromosome. Stage 6. Recombination is suppressed throughout the entire chromosome, causing it to be lost and result in an XO sex determination system. **This figure has been reproduced from Ming, *et al*, 2007, with permission from John Wiley and Sons**.

region (MSY), in which recombination with the female-determining region on the X chromosome has been selected against in response to the presence of linkage between both sexual polymorphisms, that will lead to the inheritance of sex (Stage 3)[35]. Halting recombination is essential for the divergence of X and Y chromosomes. Suppression of recombination can occur through structural changes, such as inversion and duplication events or indels that change the sequences around the MSY. This also stops selection from activating on the MSY and other strata that can lead to accumulation of deleterious elements, repeats and transposable elements. The degree of heterochiasmy, the sex specific rate of recombination between sexes, varies between species, across the sex chromosomes and in other autosomes. Recombination may not occur at all in the sex of one species in a situation known as achiasmy that can occur prior to divergence of sex chromosomes and lead to instance suppression across the full length of the Y chromosome[36]. The MSY region will become flanked by pseudoautosomal region(s) (PAR) that still recombine with the proto-X chromosomes[34]. At this point, X and Y chromosomes are still homomorphic (cytologically indistinguishable). The MSY region will still contain a number of non-sex determining genes, that may overtime gain male-specific variants, that are beneficial to male fitness and detrimental to females, will join the linkage group with the sex determination genes[34]. It is however difficult to tell if these genes evolve to gain sex specific functions before or after divergence of sex chromosomes and is therefore necessary to explore these ancestral stages to determine this[36]. Comparing the differences of paralogs between the two sex chromosomes, can allow us to date and investigate the ancestral autosomal state[37]. When the existence of a non-recombining region is known, the region is called a genetic sex-determination locus; implying that at least one sex determining gene is present in the locus. Given the scattered taxonomic distribution of dioecious plants, suggesting that genetic sex determination often evolved recently, many species may have not yet evolved extensive sex-linked regions[23].

Overtime, sexually antagonistic alleles begin to appear throughout the PARs and with them recombination suppression spreads, expanding the MSY and possibly creating new strata to cover more of the chromosome[36]. The MSY further increases in size through accumulation of mostly randomly distributed retrotransposable elements, translocations of DNA from organelles, and changes in epigenetic regulation of the chromosome[38]. However, it should also be noted that elements are also found to be enriched on the X chromosome, although the families of repeats will likely vary between chromosomes[37]. Eventually the MSY will expand to encompass the majority of the chromosome, with loss of recombination spreading throughout. Both X and Y chromosomes are heteromorphic

during this stage and the Y chromosome can become far larger than the X chromosome, as seen in (*Silene latifolia*)[39].

Degeneration of the Y chromosome through pseudogensiation of genes and loss of non-functional sequences through ectopic recombination causes shrinkage of the Y chromosome[34]. Recombination still occurs through disjunction in the shortened PAR at the ends of chromosome. Not all Y chromosomes are fated to undergo this phase, with some continuing to grow until lost or stabilising at a larger size[29].

Eventually, one potential final outcome is that the entire Y chromosome will succumb to recombination suppression and become lost, resulting in a new X-to-autosome based sex determination system. From this, a new Y chromosome could potentially form, but it would not instantly play a role in sex determination[34]. Potentially leading to transitions or evolution of new entirely new sex determination systems[29].

1.2.4 Diversity of sex determination in dioecious angiosperms

Despite being rare, only observed in 5-6% of angiosperms, dioecy is diverse and present throughout much of the monocots and eudicots[40]. One of the most studied models of sex in angiosperms is that of the eudicot species, white campion (*S. latifolia*). This species is male heterogametic, with sex chromosomes that thought to have begun recombination suppression and begun to diverge 5-10 million years ago[39, 41]. A more recent study has further investigated the age of the sex chromosomes through long read whole-genome sequencing and estimation of mutation rates, and predicted the two evolutionary strata to be 6 and 11 million years old[42]. A slow down in the degeneration of the Y chromosome is observed, compared to what would be expected based on animal sex chromosomes. Study of region of the PAR in *S. latifolia* using bacteria artificial chromosome (BAC) sequencing, also show evidence of increased repetitive elements and pseudogensiation of genes compared to the autosome of related species *S. vulgaris*[43]. Interestingly, no elevation in GC content was observed as maybe expected based on animal sex chromosomes[43]. Physical mapping of the Y chromosome's non-recombining region shows evidence for a large pericentric inversion that likely occurred after evolution of the first strata, also showing evidence for increases in repetitive elements and loss of genes[41]. These combined observations are characteristic of mature sex chromosomes, despite their young age compared to mammals whose sex chromosomes diverged about 180 million years ago[44].

Another emerging model of dioecy in this order, spinach (*Spinacia oleracea* L.), is also male heterogametic, with monoecious and gynomonoecious individuals also observed[45].

The full length of the MSY of spinach has been sequenced, revealing large stretches of repeat sequences, primarily Ty1-copia, and subsequently a relatively low gene density of only 45 genes in the 504 kb of the MSY sequence[46]. Further work using an YY individual provided has provided additional support for the Charlesworth and Charlesworth two gene model of sex determination in this species[47, 48].

Older examples of dioecy can be observed in willows *Salix*, belonging to the eudicot family *Salicaceae*, that are thought to have evolved dioecy more than 45 million years ago (Mya)[49]. This family also contains the *Populus* genus, consisting of more than 25 species of trees that have been observed to be either male or female heterogametic, suggesting transition in sex determination[49]. Both families consist of species that are widely cultivated in the UK and are of industrial importance in the production of biofuels[50]. Genomes of shrub willow *Salix suchowensis* and black cottonwood *Populus trichocarpa* have both been published and have assisted in identifying sex determination loci of chromosome 15, in *S. suchowensis*, and chromosome 19, in *P. trichocarpa* and all other studied *Populus* to date[49, 51–54]. Chromosome 15 of *S. suchowensis* and *S. viminalis* have high degrees of synteny, that is also shared with chromosome 19 of *P. trichocarpa*[49]. Although, it is not thought the sex determination loci are conserved between these *Salix* species and *P. trichocarpa*, despite all these species thought to be female heterogametic[49]. This suggests turnover of sex determination in the genus.

Similarly, Cucurbitaceae are a eudicot tribe of the Cucurbitoideae subfamily, of which the majority are monoecious or dioecious, with shifts between both sex determination systems observed within genera and species[55]. Ivy gourd (*Coccinia grandis*), belonging to the entirely dioecious genus ~25 species, has been observed to be male heterogametic and to have heteromorphic sex chromosomes[56]. Repeat sequences associated with the Y chromosome of *C.grandis* have been observed on other single autosomal pairs in several related species, through Fluorescent *in situ* hybridisation, indicating potential turnover of sex chromosomes in the genus[56].

Kiwifruit (*Actinidia*) make up a genus of likely dioecious ancestry, in which most species are male heterogametic, but some hermaphroditism has also been reported[57]. Whole genome sequencing and assembly of a female *Actinidia chinensis* individual, the most commercially important kiwifruit, and later RAD-seq studies of male and female populations from a crosses between *A. chinensis* and *A. rufa*, to create a map of the homomorphic Y chromosome, led to the characterisation of a suppressor of female function, cytokinin signalling gene (*Shy girl*), in the MSY region of *A. chinensis*[57–59]. Phylogenetic analysis of *Shy girl* by Akagi, *et al*, suggests this gene originated from

an ancestral duplication even within the genus or family, predating the divergence of *Actinidia* species studied as the gene is conserved across these species and expressed developing carpels of male flowers in all of the studied species[57]. Causing suppression of carpel development, which was also observed in transgenic *A. thaliana* and *Nicotiana tabacum*[57].

Conversely to these mostly dioecious geniuses, asparagus (*Asparagus officinalis* L.) shows only one or two occurrences of dioecy evolution in a single clade of the monoecious genus, in the form of male heterogametry, that likely occurred recently due to the lack of cytologically heteromorphic sex chromosomes and viability of double haploid YY individuals[60]. This genotype is of particularly interest agriculturally, as males have higher yields and improved longevity over females[61]. Harkes, *et al*, have carried out whole genome sequencing of the YY genotype, using a combination of short and long reads, and optical mapping, to the assembly and characterisation of 13 homozygous genes within a MSY region[60]. Two genes, Defective In Tapetum Development And Function 1 *Arabidopsis* homolog (AspTDF1) and a gene of unknown function, have been identified as potential male sterility and suppressor of female function genes, respectively[60]. Deletion of the MSY region has been shown to cause conversion of males to hermaphrodites, while deletion of the suppressor of female function gene has been reported to cause male to female conversion; consistent with two gene model of sex determination[24, 60]. A recent follow on study by Mitoma, *et al*, showed that previously used markers of sex were not useable in some cultivars of *A. officinalis* and has led to the development of a DNA marker for AspTDF1, that can be used for early sexing in multiple cultivars and related dioecious *Asparagus* species[61].

Finally, yams (*Dioscorea*) are an attractive model for studying the evolution of sex determination in angiosperms. This monocotyledonous genus, belonging to the family Dioscoreaceae, consist of over 600 mostly dioecious species[62]. Of which cytologically heteromorphic and homomorphic chromosomes have been observed[63]. The species are mainly anemophilous, but some species also propagate clonally through their tubers. Studies in one species, *D. tokoro*, have shown male heterogametic sex determination and have developed AFLP markers linked to sex[64]. However, the genus lacks genomics resources and the majority of species remain understudied.

Studies of sex determination and evolution in the majority of species discussed here have overwhelmingly benefited from DNA sequencing efforts, in particular next generation sequencing, leading to identification of putative sex determination genes and development of markers that can be used in translational research. In particular, the *Dioscorea* genus

is of interest, given the unexplored diversity of sex and impact a better understanding of sex determination would have.

1.3 Developing genomic resources in *Dioscorea*

Yams are most well known for their tubers, that are mostly used to store water, but other storage organs exists within the genus in the form of rhizomes and bulbils[65]. As the majority of *Dioscorea* are dioecious, they require cross-pollination for fertilisation and have sexually dimorphic flowers, such as those seen in *D. tokoro* (Figure 1.4). In the majority of species, flower morphology is thought to play a role in attracting pollinators, including tissue puncture insects, such as thrips, ants and also flies[65, 66]. It's therefore possible this sexual dimorphism plays a role in attracting pollinators to maximise reproduction. However, species that rely on anemophily and avoid the cost of nectar or similar reward to pollinators are known[65]. Biases in sex ratio have also been observed that may be an adaptation to maximise reproduction, with male to female ratios as high as 60:1 have observed in *D. communis*, and other species, such as *D. tokoro* showing a more males present in the population[65, 67]. Lastly, propagation has been shown to occur clonally through tubers or bulbils, or through seed propagation via winged and unwinged seeds, and fruit[65, 68].

While *Dioscorea* are distributed throughout much of the tropical and sub-tropical regions of the world[62]. Only about 10 species have been independently domesticated in West Africa, Southeast Asia, and the Pacific and Caribbean islands, becoming the third most important root crop in these continents[69–71]. In 2016, approximately 94% of the 65 million tons of yam produced globally came from the West-African countries[72]. Here, yam tuber and bulbils are mostly consumed for carbohydrates, but varieties have been shown to have comparable total dietary fibre to wheat flour, with high levels of amylose and potassium, and low levels of sodium[70, 73]. Making yam and excellent staple food in addition to commonly consumed grains and other tubers.

Aside from yams role as a staple crop, this geographical region is often referred to as the "Civilisation of the yam", which elegantly captures the West African societies that are tightly linked to yam cultivation[74, 75]. Where yams are used as symbols of social status in ethnic groups, have festivals and traditions surrounding their cultivation, and are ultimately considered the 'King of crops', as elegantly captured by Chinua Achebe's novel, 'Things Fall Apart'[76, 77]. Interestingly, traditions also surround yam in East Asia. In Japan, both tuberous and rhizomatous species are consumed, and these have

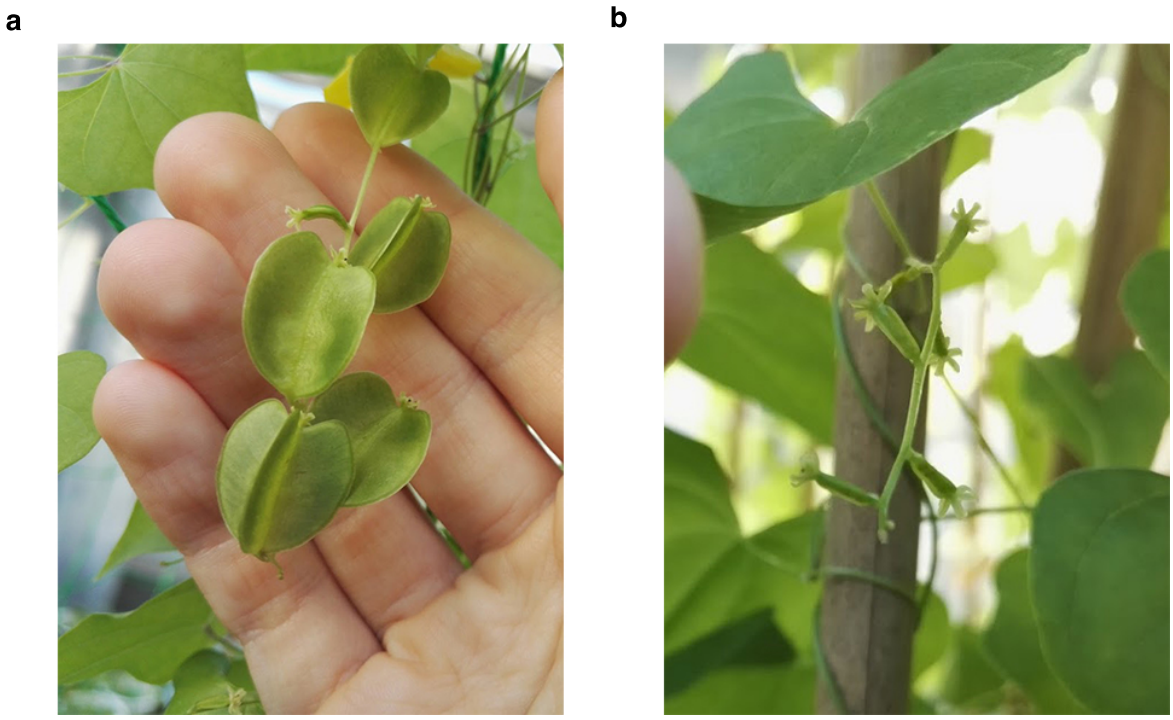

Figure 1.4: Flowers of *D. tokoro*, showing sexual dimorphism. Inflorescence of **a.** female flowers and **b.** male flowers.

an historic role as famine food in times of natural disasters (personal communication, S. Natsume, IBRC).

While demand for yam in sub-Saharan Africa is particularly high, there is a decline in production due to pests, declines in soil fertility and anthracnose disease caused by *Colletotrichum gloeosporioides*[78]. Further to this, environmental policy integrated climate combined modelling of climate change has predicted a 18-33% yield loss between 2041-2050 due to a lack of nitrogen and soil mineralisation through reduced precipitation[70]. When combined with the expected population boom predicted to occur in sub-Saharan Africa by 2030, this makes yam an important stable crop for further research.

Despite their considerable importance, the majority of yam have been considered as "orphan" crops and little is known about their genomes, evolutionary history or sex determination. Through application of sequencing techniques to *Dioscorea*, we would be able to generate genomic resources that could be used to generate new hypotheses outside of sex evolution, such as resistance to disease, as well as improving our fundamental understanding of sex evolution and having direct impact towards food security.

1.4 Next Generation Sequencing

There has been an ever growing demand for DNA sequencing, the process of deciphering individual nucleotides from a molecule of DNA, since the introduction of the chain-termination (Sanger sequencing) and subsequent commercialisation and innovation of new technologies have driven by demand from academia and industry, in many fields biological (agriculture) and medicinal fields (personal genomics). These new advances in sequencing technologies bring new opportunities that were previously out of reach, and equally, many novel challenges. With the advent of Massive Parallel or Next Generation Sequencing (NGS), it has become possible to generate unprecedented and continually increasing, volumes of short read (36 - 600 bp) genomic data from second generation sequencing platforms such as those offered by Illumina Inc and Thermo Fisher Scientific. The two most remarkable changes to occur surround the volume of data and associated costs. NGS technologies offered by Illumina's sequencing by synthesis, promise to produce terabases of data in a matter of days, compared to the near kilobases previously possible only a decade ago. Significant decreases in the cost of per basepair of DNA sequenced make it possible to now achieve the milestone of a $1,000 human genome, at sufficient coverage (theoretical number of times the whole length of a genome is sequenced) for most downstream informatics purposes[79]. Moreover, the accuracy and length of reads (the sequence calls from each DNA molecule) now rivals and/or exceeds Sanger sequencing. Furthermore, third generation sequencing platforms developed by Pacific Biosciences and Oxford Nanopore are able to generate reads spanning kilobases (Kb) to megabases (Mb) in length, albeit at currently lower data yields per sequencing run, compared to Illumina sequencing.

NGS can be applied to a range of techniques and biological questions, with or without an assembled reference genome for the species of interest. In the presence of a pre-established reference genome, sequencing and alignment of the data to the reference genome can be used to identify variants, such as single nucleotide polymorphism (SNPs), copy number variants (CNVs) and short sequence repeats (SSRs), for instance. These can yield information for genotyping and development of markers for a range of biological applications, such as breeding programs[80, 81]. As seen in the previous section, NGS and associated analyses have been fundamental in the recent classification of sex determination in angiosperm species[59, 60, 82]. Furthermore, in the absence of a reference genome, sequencing costs are now becoming low enough that smaller groups are able sequence and assemble their species of interest to a sufficient standard for tackling fundamental

biological questions; this is especially becoming in the case for orphan crop species[83].

In the next sections, I aim to review some of the current state of the art NGS technologies, focusing mainly on Illumina sequencing and how this can be applied to developing resources for non-model organisms. Due to the focus of my PhD project and hypothesis, I will omit to review the majority of background related to wet lab techniques, instead I will focus on the computational side of NGS and downstream analyses.

1.4.1 Short read sequencing

1.4.1.1 Illumina - sequencing by synthesis

Currently considered the 'workhorse' of NGS, Illumina's sequencing by synthesis platforms offers the most used and diverse number of sequencing applications available. The technology works on the basis of bridge amplification, whereby fragments of DNA molecules are ligated to adaptors containing sequencing primers, these include a barcode sequence (index) for each sample of DNA being sequenced (Figure 1.5a.)[84, 85]. These adaptors bind the fragments of DNA to primers on the surface of a flowcell, a channeled slide where the amplification of DNA and chemical aspect of sequencing takes place (Figure 1.5b.). This process of ligating sequencing primers and immobilising fragments of DNA is prevalent throughout most NGS technologies. The use of multiple indices allow for multiple samples to be pooled into the same solution (multiplexed) and sequenced together at the same time. Each separate index can then be 'demultiplexed' computationally after sequencing, correctly assigning reads to each respective sample. Prior to sequencing, libraries are amplified to increase the strength of their signal for sequencing, using solid-phase bridge amplification (Figure 1.5c.). This occurs through bending both ends of the DNA fragment to neighbouring random primers, forming a bridge, on the surface of the flowcell. The DNA fragments are then replicated using unlabelled nucleotides and and the bridge broken to leave two neighbouring copies of the library (Figure1.5d.). With the process repeated multiple times to generate clusters of each library (Figure 1.5e.). From this, sequencing reactions can then take place in parallel. In essence, the sequencing reactions involve the addition of fluorescently labeled single nucleotides (dNTPs), with polymerase, to the flowcell in alternating cycles (Figure 1.5f.)[86]. The incorporation of labeled nucleotides causes a distinct emission that is captured and converted into a base call depending on the intensity and wavelength of the emission. This process is then repeated with a new set of nucleotides, beginning the next cycle. The same process of base incorporation and image capture of an associated signal is also used by the third

generation single molecule real time sequencing technology, albeit with some differences to the chemistry and immobilisation method explored later in this chapter[87]. One drawback of this technology is an inherent GC bias, whereby low or high GC regions cause issues with the chemistry used in Illumina's sequencing methods and will result in low read quality or low/no coverage of extreme GC/AT rich regions. Controls using libraries containing known sequences, e.g. PhIX, can be 'spiked' into a sequencing run to try and even out GC/AC content or increase the diversity of sequences if the libraries being sequenced are very homogenous and repetitive, e.g. in the case of amplicons obtained from 16S ribosomal RNA[88, 89].

The sequencing platforms offered by Illumina come in two separate tiers, bench-top sequencers aimed at labs and production-scale sequencers marketed towards sequencing providers and/or institutes (www.emea.illumina.com/systems/sequencing-platforms.html). The bench-top sequencers consist of the iSeq100, MiSeq series and NextSeq, with the MiSeq's being the most prevalent, offering outputs of 15 gigabases (Gb) and 300-600 base pair (bp) reads. With the NextSeq sitting somewhere between the MiSeq and the production-scale HiSeq series, that offer some of the highest throughput from Illumina. While these generally have shorter reads, between 50-250 bp, the potential output is far higher, with the HiSeq X series able to produce up to 1.8 Tb of sequence in less than three days, making use of improved chemistry, optics, and patterned flowcells with a defined layout of primers (previously randomly distributed). Similar advances are seen in the latest NovaSeq platforms, with an increased output compared to earlier HiSeq sequencers, such as the 3000/4000, that still remain as the workhorses of the field.

1.4.1.2 Ion Torrent

Offering similar read lengths and outputs to the Illumina bench-top sequencers, Thermo Fisher's Ion Torrent NGS series offers a different NGS technology based around hydrogen ion emissions that can be measured by a semiconductor, with no optical imaging involved (www.thermofisher.com/uk/en/home/brands/ion-torrent.html). The workflow of this method begins with ligation of sequencing adaptors to fragments of DNA that are immobilised by complementary adapters on beads. These are then amplified with emulsion PCR, within the wells of a semiconductor chip[90, 91]. As free flowing single nucleotides are added to the wells, these bind to the DNA and cause the release of a hydrogen ion that results in a change in pH of the well. This change in pH is measured as change in voltage across the semiconductor. This produces not only single base calls for each cycle of single nucleotide addition, as per Illumina sequencing, but can also detect

the incorporation of homopolymers in each cycle. However, as an increase in pH from a sequence of 5 identical bases to 6 is only a 1.2 fold it is difficult to call multiple similar stretches of homopolymers, resulting in a characteristic decrease the overall read quality of repeat dense regions[92]. Despite this, as the method does not require the use of optics it is faster at producing reads than Illumina and can offer a similar Gb/hr output to the Illumina bench-top sequencers. This makes Ion Torrent well suited to applications that require rapid turnaround, of well characterised samples that don't require large amounts of coverage or extensive accuracy.

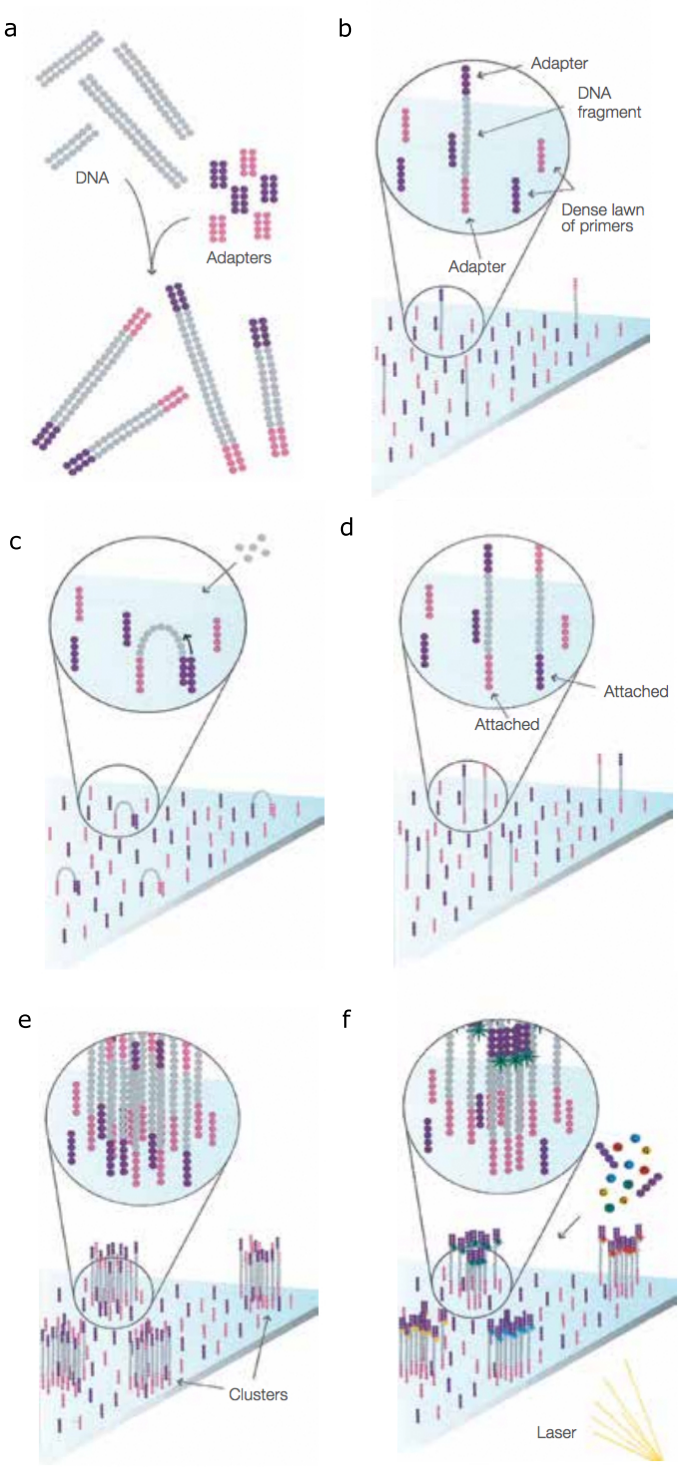

Figure 1.5: Overview of Illumina library preparation, clustering and sequencing. **a.** Library adaptors are ligated to the insert DNA fragments. **b.** Libraries bind primers on surface of flowcell. **c.** Bridge amplification using unlabelled random nucleotides. **d.** Copies of amplified library. **e.** After multiple rounds of amplification, clusters of identical library are formed. **f.** Sequencing occurs through cyclic addition of fluorescently labeled nucleotides that are incorporated into the library insert by a polymerase. Emissions from a laser allow excitation and capture of fluorescence from released florophore that can be converted into a base call. **This figure has been adapted from K. R. Mitchelson, *et al*, 2011, with permission from Elsevier**.

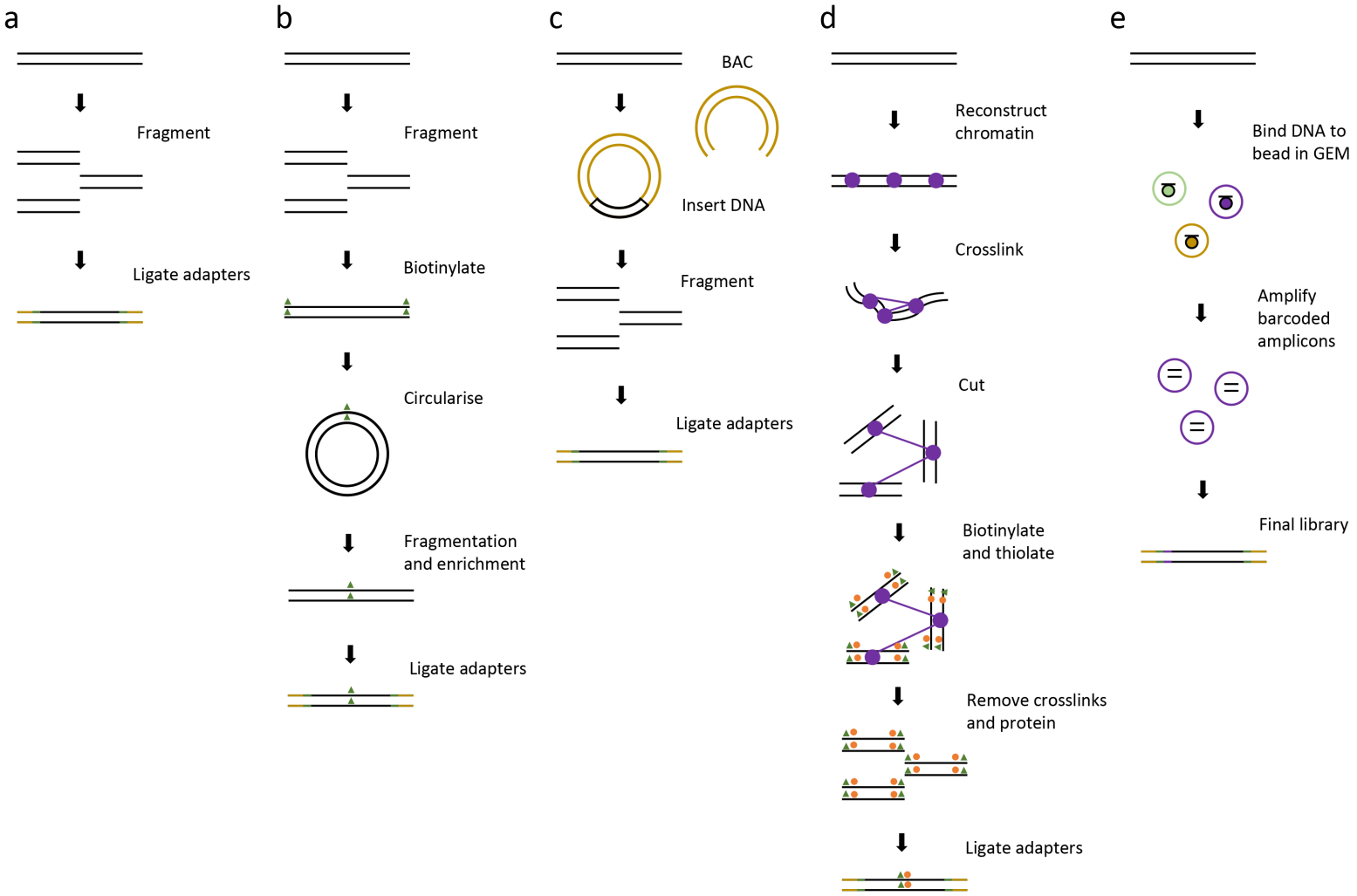
a
Fragment
Ligate adapters
b
Fragment
Biotinylate
Circularise
Fragmentation
and enrichment
Ligate adapters
c
BAC
Insert DNA
Fragment
Ligate adapters
d
Reconstruct
chromatin
Crosslink
Cut
Biotinylate
and thiolate
Remove crosslinks
and protein
Ligate adapters
e
Bind DNA to
bead in GEM
Amplify
barcoded
amplicons
Final library

Figure 1.5: Summary and comparison of workflows for five different library types used in Illumina sequencing. **a.** Paired-end library generated through fragmentation (200-800 bp) and size selection of genomic DNA, and ligation of library adaptors (orange line) that can include indicies (green line). **b.** Long mate-pair of jumping library, generated through biotinylation and circularisation of DNA fragments that are several Kb in length, that is then fragmented (200-800 bp) and enriched for biotinylated fragments to which library adaptors are ligated. **c.** Dovetail Chicargo libraries are generated from HMW DNA using *in vitro* reconstruction of chromatin (purple circle), that is then fixed, crosslinking the DNA. This is cut to produce sticky and blunt ends. The sticky ends are biotinylated (green triangle) and thiolated (orange circle), while the blunt ends are ligated. Crosslinking is then reversed, leaving fragments for library adapator ligation. **d.** BAC libraries used in BAC-by-BAC sequencing are generated through digestion of HMW DNA and insertion into vectors, that are then isolated and cloned in bacteria, such as *E. coli*, that are then fragmented (200-800 bp) and have adaptors ligated. **e.** Lastly, 10X Genomics libraries are generated using their Chromium platform with HMW DNA. The DNA is seperated into paritions that are captured by a bead in a gel bead emlusion (GEM) with library adpators and unique GEM indices, these fragments are then amplified in the GEM to produce the final library for sequencing.

1.4.1.3 Diversity of library types for whole genome sequencing

The short read libraries explained in the previous sections on Illumina and Ion Torrent sequencing are generally constructed from fragmented DNA of generally up to 1 Kb (Figure 1.5a). These DNA fragments are most often produced from physical fragmentation using sonication and size selected with agarose gel electrophoresis or magnetic beads. Obtaining the desired fragment size for insertion into the library is critical, as amplification if often biased towards smaller fragments and in the case of paired reads, Illumina sequencing from both ends of the same library, it's necessary to obtain or avoid reads that overlap each other, depending on the desired application[93].

While long read sequencing technologies are now available, it is still possible and often desirable to obtain long range sequence information using short read technologies. Mate pair libraries are one way to achieve this, as these make it possible to sequence two segments of a DNA that are separated by up to tens of Kb, with short reads[93]. These libraries are generated by fragmenting DNA to the desired overall mate pair length, e.g. 5 Kb, biotinylating (ligating a molecule of biotin) both ends of the fragment to act as tags for either end (Figure 1.5b). Then circularising the DNA so that the tag's are now adjacent to each other. A further fragmentation step to the desired insert size for amplification and selection for only biotinylated DNA, results in a linear DNA fragment with both ends, the mate pairs, of the original larger DNA molecule now next to each other. As before, sequencing adaptors are then ligated to this to generate the final library. Through sequencing of the library and downstream computational separation of each mate pair we can gain long range information, as we know the distance between the two mates and can infer this when aligning to other sequencings to act as a bridge or scaffold between two sequences. This technique can then be extended to using mate pair libraries of incrementally increasing size to give us multiple levels of complimentary long range information.

Another novel approach to achieving long range information with short read sequencing, are the Cell-free Hi-C for Assembly and Genome Organisation (Chicago)[94] libraries and subsequent sequencing offered by Dovetail Genomics (Figure 1.5c; www.dovetailgenomics.com). Using high molecular weight DNA, e.g. 150 kb, chromatin is reconstructed *in vitro* though packaging of the DNA on histones[95]. The chromatin is then fixed with formaldehyde, cross-linking the DNA and cut with restriction enzymes. Producing several pieces of DNA with sticky and blunt ends. The sticky ends are biotinylated and thiolated (similarly filled with thiol), and blunt ends ligated. Cross-linking is finally then reversed and the histones removed, leaving DNA fragments for ligation of sequencing adaptors, as with

mate pair libraries. Sequencing a single Chicago library allows for a similar result to using different mate pair libraries. This procedure can be used to determine multiple intervals of distance, between the sequenced fragments, along the full length of the original DNA molecule. This allows for linkage information beyond that of mate pairs. Phasing information on the haplotype being sequenced is also to a degree possible with this method, providing insight into the variants occurring between parental alleles.

An alternative approach to using mate-pairs and Dovetail, can be found in the use of bacteria artificial chromosomes (BAC) clones, to potentially sequence across whole chromosomes (Figure 1.5d). BAC-by-BAC is method where a physical map is generated from BAC clones that are packaged with overlapping sequences and combined cover the entire region of interest. Each BAC is fragmented and sequenced, and the reads from these aligned based on the order of the originating BACs[79]. In principle this method is a solution to sequencing and assembling difficult regions as BAC isolates are easier to sequence and assemble than a large repeat dense region or region of high GC, which whole end to end sequencing of chromosomes would fall under. However, many BACs clones would need to be prepared and sequenced to cover an entire eukaryote genome of mega or gigabases in size, making BAC-by-BAC sequencing of a genome prohibitively expensive and time consuming.

Lastly, an other approach that can be used to gain an insight into phasing and additional linked long read information is offered by 10X Genomics (Figure 1.5e; www.10xgenomics.com). Here a microfludic device, in the form of 10X Genomic's 'Chromium', is used to carry out a similar reaction to emulsion PCR used in the Ion Torrent sequencing workflow[96, 97]. High molecular weight DNA is separated into partitions of ~10 molecules, with each partition encapsulated in a gel bead emulsion (GEM) with primers and barcoded Illumina sequencing adaptors added to this. The DNA is fragmented and adaptors are ligated, with a separate barcode indexing each individual GEM. The library is then amplified through thermocycling, and the emulsion is broken and the barcoded libraries are retrieved for sequencing. As it is unlikely for two molecules from the same loci and different haplotypes to end up in the same GEM, due to the partitioning phase at the start, the accuracy for demultiplexing the sequences back to their respective haplotypes is high. However, there is a loss of sensitivity with this method that can make calling variants of small heterozygous sites difficult[96, 97].

1.4.2 Long read sequencing

1.4.2.1 PacBio - single molecule real time sequencing

Pacific Biosciences were the first to break into the third generation of sequencing at the commercial level in 2011 with their PacBio RS sequencer

(www.pacificbiosciences.com), using single molecule, real-time sequencing (SMRT). This method of sequencing uses SMRT cells, that contain tens of thousands of wells known as zero-mode waveguides (ZMWs) and is explained briefly here[98]. Pre-processing of samples involves ligation of templates called SMRTbells, close circular DNA hairpins, to both ends of the DNA. This is then bound by an immobilised template-polymerase at the bottom of the ZMW. As with Illumina's sequencing technologies, single fluorescence labeled dNTPs are added to the ZMW and incorporated across the DNA by the template-polymerase. The dNTPs used in SMRT have their fluorescence labels phospholinked to the nucleotide itself. As each dNTP is incorporated into the chain by the polymerase, the phosolinked label is released and the florescence emission recorded continuously as a movie, rather than in cycles as with the previously mentioned short read technologies. The bottom of the ZMW is illuminated just enough for light to pass into the bottom of the waveguide for excitation of the florophore. Allowing capture of fluorescence emission for each individual nucleotide incorporation, that can be converted into base calls.

This method of sequencing offers reads of tens of Kb, far exceeding second generation sequencing technologies. Furthermore, this technology allows for detection of epigenetic changes to bases without the chemical conversion steps required by Illumina. Though the accuracy of the reads is lower than that of Illumina, with an error rate at around 11-15%, compared to Illumina which can have 99.99% accuracy[98]. However, due to the circular nature of the SMRTbells, it's possible to sequence around the hairpins and back again across the same molecule of DNA multiple times in what is known as 'reads of insert'. This can produce a consensus read of all the sub-read passes along the DNA molecule, that can be >99% accurate, like Illumina, while still maintaining reads kilobases in length. Furthermore, reads can be complemented with existing Illumina reads to error correct any incorrect base calls in a read, using downstream processing after a successful run. It is also important to note that errors in PacBio reads are randomly distributed, unlike Illumina where they more often occur in regions of extreme GC, repeats or towards the ends of reads as phasing begins to desynchronise and errors accumulate as a result of the chemistry used[86]. There is also a lower output generated per run due to limitations of fitting ZMWs on a SMRT cell.

One final drawback for PacBio and other third generation technologies is the amount of high molecular weight DNA required, which can be a challenge to extract for some tissues or organisms. Whereby 10 kb or longer reads are unachievable if the molecular weight of the DNA does not meet or exceed this length[99].

1.4.2.2 Nanopore

One of the latest advances in third generation technology is the introduction of nanopore sequencing, from Oxford Nanopore (www.nanoporetech.com). Offering even longer reads than PacBio, though with similar accuracy, for a fraction of the cost associated with purchasing a PacBio sequencer[100]. The technology relies on a protein nanopore embedded in an electrically resistant membrane. The sequencing process begins with the ligation of sequencing adaptors, that have hairpins similar to SMRTbells, to generate duplex libraries. Enzymes are then captured at the 5' end, allowing unidirectional loading of libraries into the nanopore. The enzyme adheres to the nanopore and acts as a motor, pulling the DNA through the nanopore. The rate at which the enzyme feeds library into the nanopore is also controllable. The nucleotides passing through the nanopore obstruct the flow of the ionic current and produce a signature associated with of them, based on the duration of the disruption. Base calls are computationally calculated as sequence fragments, instead of single bases as with previously mentioned sequencing technologies, in the form of 3-6 bp k-mers. Once a nanopore has sequenced a predetermined length (or matched sequence), it'll start a new read. The hairpin at the end of the adaptor used, allows the same molecule to be pulled back through the nanopore and sequenced twice to produce a higher accuracy consensus read, similar to PacBio. Nanopore reads, whilst improvising with updates to chemistry and software, are still less accurate (up to 92% correct base calls) than what is possible with Illumina[100]. Like PacBio also, sequencing is carried out in real time and not in cycles as with the short-read technologies previously discussed. As base calls are determined by the overall physical structure of each nucleotide and how this interacts with the nanopore, it's possible to detect an array of different base modifications. Both Nanopore and PacBio are able to sequence full length RNA, allowing accurate characterisation of RNA splice isoforms, that would otherwise be difficult to reconstruct from short-read sequences alone.

One final consideration for use of Nanopore is the rapid development of new chemistries and software, both of which are moving at a staggering pace and could result in the redundancy of earlier work using the technology[101, 102].

1.4.3 Optical mapping

Optical mapping is a means by which to get structural information for the karyotype being sequenced, as this is exceptionally difficult to achieve with sequencing reads alone. With optical mapping acting as an effective means of high throughput karyotyping of multiple individuals, investigating CNVs, structural rearrangements and repeat regions, compared to offer long ranged approaches.

Two predominant technologies, OpGen (www.opgen.com) and BioNano (www.bionanogenomics.com), offer potential approaches to address this. Both of these technologies require a preexisting reference genome to be generated, in order to make the most of the results. In principle, they both work by using restriction enzymes to cut molecules of high molecular weight DNA (150-1000 Kb or greater)[103]. The DNA is then immobilised on a microfluidic device, stained with a fluorescent dye and elongated so that DNA molecules are stretched out unidirectionally for imaging. With a reference genome it's then possible to infer which parts of the sequence relate to each cut site and from this determine the actual distance between them. Allowing the genomic reference to be reordered and orientated based on the cytological evidence; optical mapping. With the reference or a preexisting restriction map to determine the best combinations of endonucleases to use. There are however differences between the platforms and technology offered by OpGen and Bionano, and their overall output.

The Opgen Argus platform works by immobilising the molecules of DNA on the surface of a microfluidic 'MapCard', carrying out digestion with the chosen endonuclease and staining, and then elongating the DNA through the capillary action and positive charge of the microchannels of the MapCard[103]. Bionano's Irys or Saphyr platforms similarly aim to pull the DNA into a linear strand, but there are a number of differences. Double stranded DNA is nicked on a single strand by an endonuclease of choice, the fluorescently labeled nucleotides that can be associated with the motif target are incorporated into the DNA through repair by a polymerase, allowing recognition of multiple sequence motifs on a single molecule. The mapping takes place on 'Nanochannel' chip, with a fluidics system that acts much like a pachinko machine. The DNA runs through multiple positively charged tumbler like structures that uncoil the DNA and deposit elongated molecules into individual Nanochannels[104].

Automated computational processing of images taken by the platforms can then be converted into optical maps based on the distance between cut site motifs/molecule ends. With a run of BioNano's Saphy platform able to produce much more data at 640 Gb per run (www.bionanogenomics.com/products/saphyr), as opposed to 15-20 Gb in OpGen

(www.genomics.cn/en/navigation/show_navigation?nid=2646), and both generating optical maps of potentially megabases in length.

1.4.4 Potential applications

NGS can be applied to a range of techniques and biological questions, with or without an assembled reference genome for the specie(s) of interest. Whole genome sequencing aims to cover all genic and intergenic regions of the genome, often requiring a far greater amount of sequencing than other applications, depending on the genome size and availability of a reference genome to which reads can be aligned. From aligned reads, SNPs, inserts and deletions (indels) and CNVs can be called, to provide insight into the variation of an individual or whole population from a wild type control (or whatever the reference genome was). Furthermore, comparative genomics can also be used to investigate and compare the orthology of gene evolution across related species, as well as exploring conserved syntenic regions across genomes, providing new evolutionary insight to the biological question at hand.

Studying population level genotypic and phenotypic variation can also be achieved through use of genome-wide association studies (GWAS) to associate SNPs to traits of interest in populations of genotypically and phenotypically diverse individuals, to determine the genomic loci where regulation of these phenotypes are regulated. This can be achieved through linkage disequilibrium, association of alleles at different loci, mapping of quantitative trait locus (QTL), representing a genomic region that contains one or more genes responsible for a polygenic trait of interest in a population. Likewise, the identification of expression quantitative trait locus (eQTLs) acts as a means of measuring the association between gene expression and SNPs that are within close proximity of a gene[105]. This workflow of GWAS and eQTL has been implemented to compare healthy and diseased tissues to identify SNPs and changes in gene expression, that are associated with the disease[105]. Additionally, linkage maps can be used show the relation of the QTLs between individuals in terms of recombination frequency. Analysis of the results from these applications can be used for marker-assisted selection and/or genomic selection, e.g. in breeding crops with improved agronomically important traits[106].

Other methods to investigate variation at the population level include, restriction site associated DNA sequencing (RAD-seq), which acts as a minimal bias method of sequencing restriction site associated tags across whole genomes, allowing the identification and analysis of SNPs associated with genotypes of interest at the population level, without the need for whole genome sequencing. Reducing the amount of sequencing required and

subsequent costs, that would otherwise may make the sequencing of tens or hundreds of individuals impractical.

RNA-seq aims at capturing transcript sequences from extracted RNA and measure the expression of genes of interest or to assemble and annotate a full transcriptome. RNA-seq provides information on gene expression and, with or without a reference genome, can be used to reveal new genes, transcripts, alternative splicing, fused sequences and novel RNAs[107]. While methods are being developed for direct sequencing of full length RNAs, most workflows require extracted RNA to first be translated to cDNA in order to generate libraries for sequencing[108]. This process however can incorporate errors due to biases associated with the polymerases used, transcripts captured and over or under amplification of specific transcripts[93].

Aside from use of long read sequencing to directly call a number of base modifications, chemical conversion steps can be used to study epigenetics with short read sequencing. One such example of this can be found in chromatin immunoprecipitation followed by sequencing (ChIP-seq), that can be used to study DNA-protein interactions, such as histone modifications and transcription factors. In ChIP-seq, DNA-protein complex are crosslinked by formaldehyde. The chromatin is then sheared by nucleases/sonication and immunoprecipitated using antibodies specific to the target protein or histone modification as part of pull down using magnetic beads or centrifugal force. Alignment of sequenced ChIP-seq reads to the genome shows peaks of coverage at the point of DNA-protein interaction[93]. Another method of investigating epigenetic regulation is through Methylseq, to study genome-wide or region specific methylation. With one approach to this being to chemically convert unmethylated cytosine nucleotides to uracil, while retaining intact methylated cytosines, using bisulphite. The bisulphate treated DNA can then be used as standard input for sequencing libraries, generating a map of DNA methylation[93].

The majority of applications described here require a reference genome. Where a reference is unavailable, it is necessary to undergo the process of *de novo* sequencing the whole genome, often using multiple sequencing technologies, library types and bioinformatics techniques in order to build a reliable genomic reference for answering biological questions. The next sections will therefore discuss the principals and fall backs of generating a reference genome and annotating it.

1.5 Assembly and Annotation

The process of assembling and annotating a genome can take many weeks just to complete computationally, without accounting for iterations, revisions and/or time for biological inference of the subsequent models produced. Not only this, but the computational resources required may not be readily or financially available for many groups considering to carry out an assembly and annotation project by themselves[99]. The complexity and cost also generally increases with the size of the genome being sequenced. As E. Birney, Director of The European Bioinformatics Institute (2018), elegantly puts it, "sequencing, analysing and interpreting genomes is 'routine' in the same way the US Navy 'routinely' lands planes on aircraft carriers. It might happen regularly by well trained crew with the right equipment but it is not an easy thing to do."[109].

However, while having a draft reference genome may not always be necessary to answer specific biological questions, it is becoming more affordable to generate the data required. Improvements to the ease of use of the computational pipelines involved and reductions in sequencing costs are making it easier for smaller groups to generate a first pass assembly and annotation of their organism of interest. In the next sections I will present some of the fundamentals of assembly and annotation, and address some of the considerations that should be made when approaching this in the context of plant genomes.

1.5.1 Initial quality control of data and considerations

Before NGS data can be used in downstream analyses, it must be carefully quality controlled as the data are almost never perfect and some sequencing platforms are prone to specific errors (as discussed) that need to be accounted for.

One of the first steps in quality control of NGS data is to check the quality of the reads, that is the confidence that the correct nucleotide as been called at each basepair. One metric for this, mostly used by Illumina, is the Q30 standard, where reads matching or exceeding the Q30 standard have a maximum one error in every 1,000 bases called[110]. Various tools are available for this, with one of the most used being FASTQC[111] that offers a range of visualisations and sequence content metrics, e.g. N's and overrepresented sequences, to assess the per base quality of reads in FASTQ[112] or BAM[113] formats. Of which, FASTQ is the main file format used to handle short read data, as it provides the quality score for each base and details about the run and library used.

If the reads are of too low quality for their intended purpose or there are errors, these

can be filtered or trimmed. Filtering involves binning reads that are either below a certain quality threshold, e.g. Q30, or that contain potential contamination from sequencing primers or other organisms, e.g. *Escherichia coli*. For low quality reads, tools such as the FASTX-Toolkit[114] can be used to remove these prior to any downstream analysis based on their quality score. In order to remove potential contaminants, Kontaminant[115] can be used to filter reads that have high identity to libraries of k-mers (see Overlap-layout-consensus vs De bruijn graphs section) for known contaminant sequences. This approach can also be used towards adaptor removal and reads from organelles, ensuring that only nuclear DNA sequence is included in the assembly process. In terms of trimming, this is often required if a read contains sequence from the adaptor, perhaps due to the insert being shorter than the read length. Tools such as Trimmomatic[116] offer a means of trimming reads of certain length, quality, or those with known adaptor sequences.

A final important consideration for NGS data, while not necessarily an issue with data quality, is the problem of 'Big Data'. Where NGS is able to produce and increasing amount of data, at an increasingly fast pace, with platforms such as the Illumina's HiSeq X and NovaSeq, able to produce terabytes of data in less than a few days (www.emea.illumina.com/systems/sequencing-platforms.html). While the boom in the volumes of data that can be produced, at an ever reducing cost, can provide exceptionally valuable insight and resolution to the biological question at hand. There are often not sufficient means to store or compute such large volumes of data. Leading to a constant expensive battle to ensure there are sufficient computational resources available, to handle an exponentially growing amount of data[117]. As technology continues to advance, even once we reach the point of one perfectly accurate read per chromosome, NGS projects will continue to become more ambitious in scope and breadth.

1.5.2 Overlap-layout-consensus vs De bruijn graphs

A major starting point for studying the genome of any organism, where a reference does not already exist, is to go about the daunting task of assembling one. The majority of modern non-greedy assemblers rely on various flavours and combinations of overlapping consensus and De bruijn graphs in order to piece together the many millions of reads used to cover entire genomes.

The most basic method, the overlapping consensus graph, involves taking all reads and overlapping them to create a directed graph based on consensus overlaps. Where each node is a read and the directed edge between pairs of nodes, or reads, is drawn when the suffix and prefix of the first and second read overlap. Then the layout of nodes and edges

within the graph is determined, and the consensus sequence generated. The required lengths of overlapping sequences to form nodes and the number of mismatches allowed within each overlap are considering in constructing the graph, to provide the desired level of confidence in the full consensus, or continuous sequence (contig), generated. With a number of assemblers, that use and expand upon this method, e.g. Celera[118] and more recently Canu[119]. While this method can be simple and effective at resolving the complete sequence of all reads, it is very computationally demanding as the method has to consider an all-vs-all pair-wise comparison of millions of reads[120]. However for smaller sets of long error prone reads, such as those from third generation sequencing technologies, this can produce an effective means of assembling a genome without the need to generate a more complex graph with limited k-mer size; unique sequences of length k from the complete set of all reads or sequence[100].

A more computationally efficient and widely adopted alternative, that does not rely on alignments of reads, is to use De bruijn graphs[121]. This approach to the assembly problem uses k-mers to act as nodes within the graph. These are linked by subtracting the lengths of k-1 from the start and end of each k-mer, and the edges then drawn between pairs of k-mers with matching k-1-mers. This produces a graph with one edge per k-mer and one node per k-1-mer. Assemblers then conduct a Eulerian walk (aiming to visit each edge only once) through the graph following the edges from each node, using each k-kmer once, to reconstruct contigs based on the complete unbroken path through the graph. Examples of this can be found in ALLPATHS[122] and the more recent, DISCOVAR *de novo*[123].

1.5.3 Scaffolding and gap filling

In essence, scaffolders work by using long range information, often from mate pairs, to order and orientate contigs. Where two contigs with either one of the paired reads present are assumed to be related, and a stretch of N's (identifier for non-redundant sequence) added between these to link them together into a scaffold and specify a gap of unknown size. Contigs of repetitive regions will generally have multiple alignments from all reads of the complete repetitive element, the coverage depth of k-mers or number of reads aligning, can be used to infer the correct order of contigs. Accurate insert size of the mate pairs is fundamentally important as scaffolders will rely on this information to estimate the distance between mates and correctly assign them. Larger insert sizes generally also have more variability in their insert size, than smaller insert, mate pair libraries[124]. Potential contamination in the mate pair reads from unknown fractions that have not

undergone circularisation to form the mate pairs (essentially just paired-end) can cause incorrect scaffolding, as the insert size and direction of the paired-end contamination is unknown and is required for correct ordering and orientation of the contigs[124]. As such, mate pair reads with high paired-end contamination should be used cautiously, as it will likely introduce errors into the finally assembly.

A range of different scaffolders are openly available, the ALLPATHS-LG[125] assembly workflow has a scaffolding step that estimates the distribution of insert sizes through alignment of reads and then looks for inconsistencies in separation statistics of mate pair libraries, that can be used to merge and fix scaffolds. Stand alone tools such as SSPACE[126], provide efficient means to scaffold preexisting assembled contigs, using paired-end and/or mate pair reads. Filtering for non-ACTG bases, mapping of reads and extending contigs with unmapped reads, prior to scaffolding.

Other approaches can combine long and short read technologies, such as with the FALCON[127] assembler, using the short reads to error correct long reads, while using these to produce and join contigs into an assembly with minimal gaps.

After scaffolding the next step in the assembly process is often the closure of gaps, by replacing the N's with actual sequence. This can be one of the more challenging steps in an assembly. If the gap is in a repeat dense region, there is a high chance of creating a new misassembly due to low read coverage in these difficult to sequence regions and the likelihood of contigs either side of the gap ending early due to the complexity of assembling the region[128]. Gap fillers, such as Sealer[129], focus on intra-scaffold gaps, with Sealer attempting to use short paired-end reads in a graph based approach to connect reads flanking a gap. Alternately, the tool PBJelly[130] uses PacBio reads that span or overlap, but can be extended through, gaps to produce an overlap-layout-consensus assembly of the gap filling sequences that can close the gaps.

Finally, optical maps and linked reads can be used to map scaffolds into pseudo-molecules/chromosomes. With one example of this being the use of *in silico* maps generated through the use of BioNano optical maps and scaffolds to order, orientate and align the scaffolds into super scaffolds that can be whole chromosomes in length, essentially covering the entire consensus sequence for that chromosome[131]. However, in order to obtain chromosome level assemblies through mapping of scaffolds, a very contiguous assembly is require, at least 50-500 Kb in the case of BioNano super-scaffolding[128].

After sufficient scaffolding and gap filling, or only contigs in the case of a first pass assembly, the assembly is 'frozen' and the process of annotation can begin.

1.5.4 The assembly problem

While the end goal of any assembly is to create a single unitig representing each chromosome of the genome, or often more realistically, the longest set of unitigs possible with the data available. Some of the main difficulties that make up the assembly problem, especially for plant genomes, are the presence of heterozygosity, large repetitive elements and ploidy. With the lengths of reads and read accuracy making it difficult to correctly determine the order of nodes within the graph of repetitive regions, where these are longer than the reads used as they cannot cross the paths between all nodes within the region. Resulting in the repetitive region of the graph being collapsed in a single unit of the repeat. Another issue is the presence of heterozygosity, regardless of read length, as it results in the presence of two possible paths through the assembly graph, that are difficult to resolve into the desired representation of a haploid genome, and is further confounded by ambiguities errors that can be present within reads[120]. Use of inbred individuals for sequencing is therefore advisable, but not always possible. These issues can be somewhat resolved through the use of scaffolding to order and orientate contigs into longer unitigs referred to as scaffolds. Finally, erroneous reads if unfiltered, and simple repeats, can be somewhat alleviated through an error correction step used at the start of several modern assemblers, but this is far from perfect[121]. Much time and effort has been devoted by the community to studying the assembly problem, comparing and contrasting different assembly methods to address this and how best to validate the the assembly, of which the Assemblathon contests are some of the most prominent studies[132, 133].

1.5.5 Validation of assembly

There are three main areas to consider when assessing the quality of a genome assembly, the contiguity, correctness and completeness. Neglecting any one of these will have a subsequent impact on the downstream annotation and overall usefulness of an assembly. However, achieving a model that perfectly addresses all three may not be feasible and therefore it may be necessary to consider what is most important, e.g. good quality gene models, but a highly contiguous assembly. Below I address each of these three quality metrics and some methods of assessing their performance.

Contiguity has classically been measured in terms of N50, whereby all contigs/scaffolds are ordered from smallest to largest and N50 is the sum of all contig/scaffold lengths that cover 50% of the total size of all contigs/scaffolds in the assembly. While longer

contigs/scaffolds are desirable, as they would represent a more complete assembly, they are not a measure of accuracy. Aggressive assemblers can concatenate together sequences incorrectly in an effort to generate as long a contig as possible, leading to a misassembly. An N50 that is at least as long as the estimated median gene length is a good target for an assembly prior to any gene annotation, as at least 50% of genes can be complete on a single scaffold[134]. The tool, QUAST[135], can calculate and visualise the total number, size distribution, and N50 of scaffold and contigs, with some capacity to investigate misassembles and functional elements of the assembly.

Correctness essentially equates to, "do the contigs/scaffolds in the assembly accuracy represent the genome?". One first consideration is to ensure that no contamination, that should have been filtered from the reads, has made it into the assembly. As it is common for contamination of microbial origin to be present in the sample and subsequent reads. It is therefore essential that a final assembly is screened for potential contaminants. As with quality control of NGS reads, a potential approach to this is addressed by Kontaminant, that uses a k-mer library associated with potential contaminants, e.g. several bacterial species, to screen the reads for matching k-mers and remove potentially contaminated reads [115]. Beyond potential contamination, another consideration of genome correctness is through misassembly and invention of false genomic content by the assembly processes. Misassembles can be identified by mapping the short reads to the assembly to identify small errors, such as SNPs and InDels, and long reads for larger structural errors, such as translocations and inversions[133]. Recognition of Errors in Assemblies using Paired Reads (REAPR)[136] is a tool that can provide a per base error report, through uniquely mapping of paired reads and calculation of fragment coverage distribution at each base of the assembly, to score each base and where needed introduce breaks in contigs/scaffolds that have been misassembled. In addition to this k-mers can be used to check the correctness, in additional to completeness, of an assembly by ensuring that k-mers present in the assembly and not in the reads are removed as these have been 'invented' by the assembly process. KAT[137] is one such tool that can be used to align reads to the assembly and investigate discrepancies in k-mer content between the reads and assembly.

Completeness is another measure of assembly quality, essentially asking how many of the expected genes and what amount of total genomic content is present? Estimation of genome size using flow cytometry, nuclear weight or k-mers can give a good indication of the expected size of the assembly, whereby a perfect haploid assembly should correlate with these. Generally, due to repetitive sequences, it's often difficult to fully assemble

all regions of the genome and therefore genome coverage above 90% can be considered good and likely to be including the majority of the gene space[134]. The gene space of the assembly, can be assessed by mapping on transcripts, a perfect genome assembly should have all complete transcripts mapping. Another approach can be through the use of the Core Eukaryotic Genes Mapping Approach (CEGMA) tool, that screens an assembly against a database of universal eukaryote single-copy genes and determines the percentage of each gene lying on a single scaffold[138]. While CEGMA has been replaced by Benchmarking Universal Single Copy Orthologus (BUSCO), that similarly scans across the assembly for BUSCOs, that are assumed to be prevalent across eukaryotes and/or prokaryotes and can be used as a benchmark of how complete the gene space of an assembly is[139]. BUSCO can also be applied directly to gene models to look for discrepancies between the genes annotation and gene space of the assembly. Both CEGMA and BUSCO are useful in assess completeness, as while CEGMA has a smaller database of orthologus (248), these are more likely to be present in novel clades than the complete larger clade specific set offered by BUSCO (952; from six major clades in the plantae dataset). However, both fail to account for the diversity and completeness of other genes outside of those that are primarily housekeeping, required for comparative and evolutionary studies, that only account for a small fraction of the gene space.

1.5.6 Genome annotation

The role of the annotation is to infer the structure and subsequent function of sequences in the assembly. While having as correct and complete an assembly as possible is vital, the sequences by themselves do not provide information about the functional genome. Annotation of a new assembly is therefore a crucial step in being able to address our biological questions. The annotation itself essentially is attached biologically meaningful labels to the assembly, based on the structure and composition of the sequences compared to related species and other sources of information from the species being studied, e.g. transcripts. With the majority of effort spent on identifying and correctly producing models for genes.

In addition to gene models, repetitive elements are another important component of the genome for annotation. Repeats can be low-complexity sequences, e.g. homopolymers, transposable elements (TEs), long interspersed nuclear elements (LINE)s and short interspersed nuclear elements (SINE)s. Low complexity repeats and TEs are abundant in plant genomes, as repeats can represent nearly 90% of the genome, as in wheat[140]. These repeats can have roles in affecting the structure of the genome and expression

on genes, and it is therefore important to ensure these are as accurately assembled and classified as possible. Handling repetitive sequences has to be carried out prior to gene prediction, to remove potential false positives that could be caused by alignment seeding within these sequences[134].

The first step in gene annotation is, therefore, repeat identification and masking. There are multiple methods of identifying and classifying repetitive sequences, based on either homology to a preexisting library of repeats or *de novo* generation of this. One common solution is offered by RepeatModeler[141], encompassing *de novo* repeat identification programs (RECON[142] and Repeatscout[143]), to search the assembly for repeat sequences and create a classified library of repeat families. The repeat library from the assembly, or another from a related species/database, can then be used with RepeatMasker[144] to screen the assembly for matching repeats and low complexity sequences. These sequences are then masked, converting every nucleotide identified as a repeat to an'N' (hard masked) or to lowercase (soft masked).

With repeats identified, the remainder of the gene annotation pipeline is generally broken into two iterative phases: 1) intrinsic evidence that can come from alignment and assembly of transcripts from the same species, and 2) alignment of transcripts and proteins from related species that can act as extrinsic evidence, and. The intrinsic approaches involves producing *ab initio* models for genes, while the extrinsic relies on outside evidence from related species, with combiners used at the end to compare and assess both sets of predictions to produce the most likely gene models possible.

Intrinsic evidence can come from *ab initio* gene predictors, such as

AUGUSTUS[145] and GeneMark[146–148], that use Hidden Markov and/or other probabilistic models to find signals for modelling intron-exon and intergenic structures of genes. However, unless the assembly being annotated is from a species related to the parameter files used for the *ab initio* gene prediction, that contains details of expected GC% content, codon usage, min. and max. intron and exon size for that species, the gene predictions will not be accurate and potential genes will also be missed[134]. While *ab initio* gene predictors can to some extend self train, such as with GeneMark-ES[146], external evidence from EST or RNA-seq alignments from the same species being annotated can be used to train the *ab initio* gene predictor and produce a new custom parameter file for gene prediction.

RNA-seq data from the individual/species being annotated, ideally from multiple tissues and developmental stages to capture more novel transcripts and their isoforms, can be used to support *ab initio* gene models by assisting with intron-exon boundaries

and other gene features, such as untranslated regions (UTRs). Transcriptome assemblers, such as Trinity[149], can carry out *de novo* or reference guided assembly (using reads aligned to the assembly) of RNA-seq data. In order to assembly full-length transcripts through clustering of the transcripts and then carrying out multiple De Bruijn graph assemblies for each individual transcript cluster. However, these transcriptome assembles can be limited by missing gene space in the assembly or by transcripts that are not well represented in the reads.

Program to Assemble Spliced Alignments (PASA)[150] uses splice aware alignments to assemble gene models, with intron-exon corrections, UTRs, novel gene prediction and alternate splice variants. Transcriptome assemblies (*de novo* and reference guided) from Trinity and additional transcript alignments using Cufflinks[151] or similar can be used with PASA to merge and filer poorly mapping transcripts. In order to generate a comprehensive transcriptome database for use in validation of gene predictions and any downstream gene expression studies.

Extrinsic data in the form of protein, expression sequence tags (ESTs; short cDNA sequences) and RNA-seq data from related referenced organisms can be used as external evidence for gene prediction, using the homologous sequences to identify new genes and validate models predicted intrinsically. Curated and unreviewed protein sequences from related species can be readily found on online database, such as Pfam[152] and in the case of plants, cDNAs, proteins, gene models and assembles for a number of plant species are available via the Phytozome[153] portal, Ensembl Plants[154] and also plaBi (www.plabipd.de), which actively tracks publications on new plantae reference genomes. The majority of aligners today generally use Burrows-Wheeler transform (BWT) or hashing to align short read sequences to a reference genome. BWT[155] based aligners, e.g. Burrows-Wheeler Aligner (BWA)[156] and SAMtools[113], and Bowtie2[157], and are efficient for mapping WGS reads to the genome, including repetitive reads[80]. Other aligners, such as GMAP[158], can be used for splice-aware mapping ESTs and transcripts to an assembly, from multiple sources, e.g. tissues and species, also incorporating predicted microexons into the subsequently mapped gene models. Another aligner, Exonerate[159] can carry out pairwise alignments using dynamic programming, that can identify splice junction and model intronic regions, combined with heuristic methods, similar to those employed by fast aligners like BLAST[160], for efficient complex alignment of cDNA or protein sequences. These alignments are then filtered to remove low quality alignments, that could lead to false positives, and redundant alignments, which would increase run times and unbalance weighting of the alignment evidence.

Comparers can then be used with the intrinsic and extrinsic evidence, to compare and contrast all forms of evidence, and decide the likelihood of each gene model. Producing a final consensus gene model that can have a confidence value assigned to it. Examples of comparers include MAKER[161] and EvidenceModeler (EVM)[162], with MAKER providing the option to run an automated pipeline that generates intrinsic and extrinsic inputs, as well as comparing predicted gene models. Both MAKER and EVM provide curation and quality metrics for gene predictions. MAKER uses an Annotation Edit Distance (AED), where sensitivity and specificity of all alignments are considered for each gene model, and proteome domain content to determine the most likely gene model to include in the final annotation. EVM deconstructs gene predictions from different evidences into their separate structural components, and uses a non-stochastic weighted evidence calculation, with the confidence (weight) of each piece of evidence provided by the user to produce consensus gene models. PASA can then further be used to update EVM gene models, including UTRs and alternatively spliced isoforms.

After high confidence gene models have been generated, they can be functionally annotation to provide an idea of their biological role. This can be performed using tools such as IntrProScan[163–165] and Blast2GO[166], to search for known functional motifs in the gene models and assign a predicted functional ontology to the gene.

Finally, automated annotations can be manually curated. This is a lengthy process that is often carried out between multiple researchers, each focusing on gene families of particular interest, to curate gene models and their functional pathways using the evidence from RNA-seq and orthologous sequences of related species[167].

1.6 Thesis objectives

The overall aim of this thesis is to explore the application of state of the art NGS and associated analyses to an orphan crop species, to investigate the relatively rare trait of dioecy and how this can be used to improve our understanding of sex evolution. The structure of the main thesis will be as follows: Chapter 2 will explore the generation of an improved reference genome for Chinese hamster ovary cells, as part of a collaboration involving my BBSRC iCASE industrial partner, Eagle Genomics Ltd, UK. Chinese hamster ovary cells are the industry standard for therapeutic protein production, but the current reference genome has not undergone any revisions since it was first published eight years ago and may be considered not fit for purpose. Since then sequencing and bioinformatics techniques have rapidly moved on, and now make it possible to generate

vastly improved references to what was possible a decade ago, this chapter will focus on assessing the current state of the art compared to what was previously possible. Chapter 3 will begin the main theme of the paper, applying NGS technology to an orphan crop to generate a genomic reference that can be used to elucidate the underlying mechanisms of sex determination in the crop and learn more about the phylogenetic trends of mating systems. For this purpose, I will investigate the genome of *D. rotundata* as part of an international collaboration with the Iwate Biotechnology Research Centre (IBRC), Japan and the Institute of Tropical Agriculture (IITA), Nigeria. The end aim of this collaboration being the generation of a reference genome of sufficient quality to begin deciphering the mechanisms of dioecy in this species. Continuing on from this, Chapter 3 will look at building references for two more yam species, *D. tokoro* and *D. alata*, through collaborations with IBRC and IITA, respectively. From this I will begin to explore the evolution of sex determination in the *Dioscorea* genus and create the foundations for further work to elucidate the ancestral state of separate sexes and sex determination. Chapter 4 will contain the general conclusions of this work, showing that advances in NGS and bioinformatics make it possible for smaller labs to conduct studies that would have been unthinkable for the human genome project and later advent of NGS. As such, it is now feasible to explore important neglected species and look at fundamental biological questions. Finally, I will give future perspectives on how the framework I've produced can be applied to investigate evolution of sex and other sex determination systems in *Dioscorea* and the impact this work will have on food security, industry, and our understand of sex.

Chapter 2

An improved genomic reference for Chinese hamster ovary cell lines

The work contained in this chapter was carried out as part of a collaboration between my PhD iCASE industrial partner (Eagle Genomics Ltd), the Wellcome Trust Sanger Institute, European Bioinformatics Institute, Horizon Discovery Group plc and myself at the Earlham Institute. Cell lines and financing of the project were provided by Horizon Discovery Group plc, sequencing was carried out at the Wellcome Trust Sanger Institute, and genome assembly and annotation was performed as a joint effort by Eagle Genomics Ltd and the European Bioinformatics Institute. As part of this, I carried out first pass assembly on the genome, contributed to further assemblies, performed quality control and evaluation of assemblies/annotation, comparative genomics, investigation of glutamine synthetase knockout, and wrote the manuscript for this work (unpublished).

2.1 Abstract

Chinese hamster ovary (CHO) cells are one of the most important cell lines for the development and production of therapeutic proteins. While genomic references (assembly and gene predictions) for CHO are available, these are suboptimal in terms of completeness, contiguity and accuracy for many biomolecular applications. Here I present a new open genomic reference that represents a substantial improvement over previous efforts. Generated through iterative combining of short read sequencing, HiRise scaffolding, BioNano optical mapping, and the latest assembly and annotation pipelines. Furthermore,

this approach that can be applied to other species for cost effective generation of reference quality assemblies. Our 2.35 Gbp Horizon Discovery CHO-K1 (glutamine null) reference sequence most notably validates previously reported structural synteny to mouse, with vastly improved contiguity, and contains a more comprehensive gene set than previous genomic references for CHO. Availability of the Eagle-Horizon CHOK1GS_HD genome in Ensembl will aid adoption and analysis of the genome, enhancing work into the production of vital therapeutic proteins and studies utilising this cell line.

2.2 Introduction

Since their commercial introduction over 30 years ago, cell lines originally derived from the epithelia of Chinese hamster (*Cricetulus griseus*) ovaries (CHO) have become the most commonly used mammalian hosts for industrial production of recombinant proteins, such as therapeutic glycoproteins[168, 169]. These recombinant proteins cannot yet be synthesised chemically and have a major role in the development of therapies for hard to treat disease, of which cancer and haematological conditions make up the majority of targets for currently approved therapeutic recombinant proteins[170]. Furthermore, the titre yield of these proteins has increased by over 100-fold since the introduction of CHO. However, this increase in yield has thus far been mainly driven by improvements in cell culture media[171, 172]. Compared to other mammalian cell lines, e.g. HeLa cells derived from cervical cancer of the late Henrietta Lacks, CHO cells remain the industry standard for the production of recombinant proteins. This is in part due to their metabolic plasticity, resistance to viral infection, ease of maintenance in serum free suspension and their ability to perform post-translational modification of recombinant proteins that are compatible with the human immune system[173].

CHO cell lines are made up of a number of related lineages, e.g. CHO-K1, CHO-S, CHO-DG44 and CHO-DXB11, and are primarily distinguished by their response to different cell culture conditions. Of these lineages, CHO-K1 and CHO-S are two of the oldest, having been directly derived from the immortalised host cell line. The main phenotypic difference between these two cell lines is that CHO-S grows better in single-cell suspension culture. From CHO-K1 the CHO-DXB11 line was derived through chemical deletion/mutagenesis of dihydrofolate reductase (DHFR) alleles; an enzyme which reduces dihydrofolic acid to the tetrahydrofolic acid that is required in the synthesis of purines and pyrimidines[174]. In turn giving rise to the methotrexate/DHFR clonal screening system that is employed in the double DHFR deletion found in CHO-DG44[175]. Allowing for

easy screening of successfully transformed cells.

Whilst advances in genome engineering through Zinc-finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs), meganucleases, and more recently, clustered regularly interspaced short palindromic repeats (CRISPR), have created new opportunities to directly optimise the CHO cell lines[176]. The success of these genome engineering techniques depends heavily on the quality of a suitable genomic reference (primary genome assembly and the gene predictions), against which guide RNAs, in the case of CRISPR/Cas9 for example, can be precisely designed[177].

Prior to the work contained in this thesis chapter, only three Chinese hamster genomes had been released publicly. The first of these, assembled from a CHO-K1 cell line, was released in 2011 and is generally considered the community standard; henceforth referred to as "CriGri_1.0"[178]. A genome of *C. griseus* strain 17A/GY (CHO_17A/GY), notable in that sequencing was performed on flow-sorted chromosomes enabling contig-to-chromosome assignments, was published in 2013[179]. Finally, a further *C. griseus* genome of unknown strain was also published in 2013[180]. In addition to these public reference sequences, several commercial groups have generated private (unpublished) genomes, and other community sequencing efforts are underway.

From a review of the literature I could find no published reports detailing specific limitations of these existing CHO genomic resources, for use in designing targets for genome engineering; that requires a high quality reference sequence as previously mentioned. However, genome sequencing, assembly and annotation methods have all advanced greatly since 2011 when CriGri_1.0 was released[178]. Examples of such advances include increased read length, improved accuracy of reads, optical mapping, and new assembly and annotation algorithms that produce greater contiguity and more complete gene models[100, 123, 131, 181–183].

Furthermore, as the CHO genome is inherently unstable, due to it having been immortalised, any given CHO cell line and/or passage may differ significantly from the original CHO-K1 cells used to generate CriGri_1.0[184, 185]. This means that reagents designed *in silico* against the reference may not work as intended *in vitro*, with potential off-target effects being one such concern. This problem is further impeded by licensing terms associated with the majority of manufacture ready CHO cell lines, restricting further modification and/or incurring royalties.

In an effort to circumvent these issues and improve the productivity of CHO as a valuable tool in the production of recombinant proteins, I present the CHOK1GS_HD genomic reference (genome assembly and gene predictions) as part of a multiparty

collaborative effort led by Horizon Discovery Ltd. CHOK1GS_HD was generated using a combination of recent sequencing and assembly techniques, and represents a considerable improvement over CriGri_1.0, in terms of completeness, contiguity, and accuracy. The CHO-K1 glutamine sythetase (GS) null cell line from which the reference is based is the current gold standard for commercially available bioproduction cell lines, and is available from Horizon Discovery under accessible terms with minimal licensing restrictions. This CHO-K1 GS null cell line allows for rapid production and screening of recombinant clones through use of the Methionine sulfoximine/GS screening system, that has been shown to be more stringent than the before mentioned Methotrexate/DHFR system[186].

As such, the utilisation of this improved reference for CHO and CHO-K1 GS null cell line will further enable genome-wide screening for CHO genes of interest, thereby stimulating innovation in bioproduction and ultimately driving down the cost of biotherapeutic manufacturing. The CHOK1GS_HD raw reads and assembly are freely available to browse and download via Ensembl[187] (release 90 onwards; CHOK1GS_HD), ENA and GenBank (GCA_900186095.1). Furthermore, I demonstrate the current state of the art in genomic reference generation, using a workflow that can readily be applied to other species.

2.3 Methods

2.3.1 Generation of the Horizon CHO-K1 GS null cell line

The CHOK1GS_HD genomic reference is based on the Horizon Discovery CHO-K1 GS null cell line, with gDNA samples from cell culture extracted and prepared by Horizon Discovery Ltd. The GS null line was generated through gene knockout (KO) of both alleles of the GS gene in CHO-K1 cells (ATCC CCL-61) purchased from the European Collection of Cell Cultures (ECCC). CHO-K1 stock was prepared by single cell dilution and selection of several clones based upon the similarity of growth characteristics compared to the parental CHO-K1 cells. A suitable clone was selected for targeting, alongside identification of optimal conditions for antibiotic selection following transfection tests, recombinant adeno-associated virus (rAAV) transduction tests and antibiotic death curves. The rAAV targeting strategy followed that of Liu *et al*, 2010[188], achieving functional knock-out of GS through the targeted deletion of exon six (coding exon five) in the GS gene. After successful knockout of the first allele, the selection marker was removed using Cre recombinase and the second allele was targeted using the same strategy. The knockout

was genotypically validated by PCR and phenotypically validated using real-time PCR (RT-PCR). Further details of the CHO-K1 GS null cell line derivation are publicly available (www.horizondiscovery.com/bioproduction/cho-cells/cell-line-derivation).

2.3.2 Sequencing of the Horizon CHO-K1 GS null cell line

Primary data for the assembly and annotation of the genome was generated by the Wellcome Trust Sanger Institute, through *de novo* sequencing, optical mapping, and RNA sequencing of the Horizon CHO-K1 GS null cell line. Two PCR-free PE Illumina fragment libraries were prepared from the cell line and these were both sequenced on an Illumina HiSeq 2500 sequencing platform; the first generating 536,058,733 125 bp read pairs and the second 154,007,001 250 bp read pairs. Dovetail Genomics (www.dovetailgenomics.com), prepared Chicago libraries from the cell line and sequenced them on the Illumina HiSeq 2500 in rapid run mode, as a commercial service. Optical mapping of the genome was carried out at the Wellcome Trust Sanger Institute using the Bionano Irys (www.bionanogenomics.com/products/irys).

2.3.3 Assembly of the Horizon CHO-K1 GS genome

The genome was assembled from the sequence data using a multi-step pipeline, with much of the complexity arising from availability of both 250-bp and 125-bp reads (Figure 2.1). In brief, raw 250 bp and 125 bp PE reads were subject to quality filtering using Kontaminant[115] version 2.0, by myself, using k-mer libraries for library adaptor sequences, PhIX (sequencing control; spike-in), chloroplast, *Xanthomonas campestris* and *Escherichia coli*. I did this to remove contamination from the sequences prior to assembly. The default k-value of 21 was used for screening and filtering of sequences as this allowed for 4.39 x 1,012 potential k-mers, making the majority of k-mers observed likely to be unique.

Further to this, I checked quality filtered PE reads with FASTQC[111] version 0.11.4 to see the overall sequence quality and then KAT[137] version 2.1.1 to estimate k-mer size and content, using the default k-value of 27 (Figure 7.3). Reads were not trimmed or filtered for quality beyond this. Two *de novo* assemblies were then carried out on the filtered reads, the first I performed *De novo* assembly with DISCOVAR *de novo*-52488[123] using the DiscovarExp option and the 250-bp reads. The reads represented an estimated 32.7-fold genome coverage of the first-pass assembly. D. Barrell (Eagle Genomics Ltd) then performed an additional *De novo* assembly with SGA[189] version

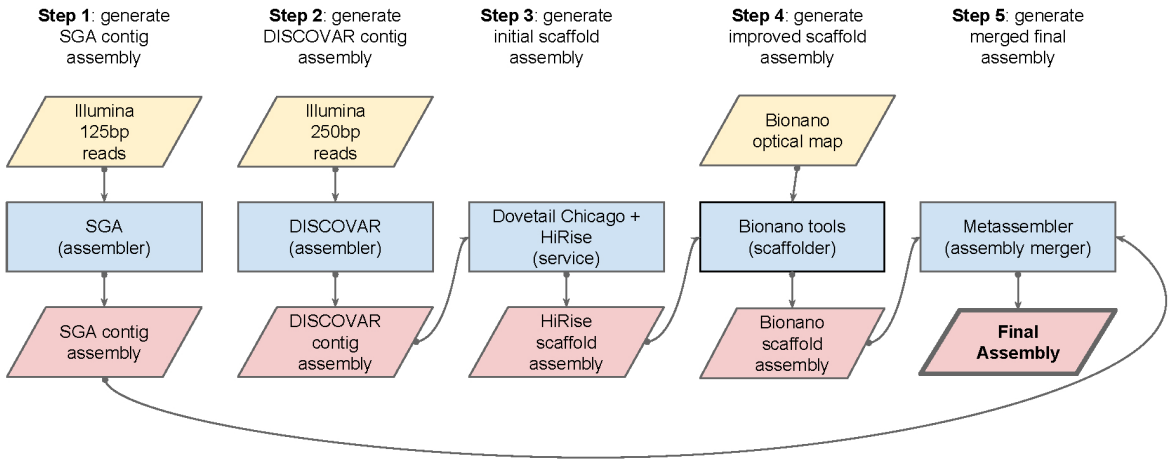

Figure 2.1: Overview of the multi-step pipeline used to create the CHOK1GS_HD genome assembly.

0.10.15, using the combined set of 250 bp and 125 bp PE reads. Scaffolding of the DISCOVAR *de novo* assembly was then performed as a commercial service by Dovetail Genomics, using data from the Chicago libraries with HiRise[94].

Following this, D. Barrell generated super-scaffolds following the sewing machine pipeline, using the HiRise scaffolds and BioNano genome map with stitch.pl[131]. Finally, a meta-assembly was then performed, by D. Barrell, using Metassembler[190] to combine the SGA assembly and super-scaffolds. I then carried out post-processing of the assembly using bioawk (www.github.com/lh3/bioawk) with the fastx option, to remove all contigs smaller than 2 kb as k-mer spectra analysis did not show these to introduce many unique (non-erroneous) k-mers into the assembly (Figure 7.4).

2.3.4 Genome Assembly: Quality Assessment

I generated basic quality assessment statistics for the final assembly using QUAST[135] version 2.3. I then carried out assembly quality control steps using KAT[191] version 2.1.1 to determine the k-mer distribution of the PE reads and to map unique k-mer abundance from the reads to the genome, in order to identify missing or erroneous k-mers. Afterwards, I performed further quality control and contamination screening using Blobtools[192] and MegaBLAST[193], with a minimum sequence length of 500 bp, coverage of 97%, e-value of 1e-150, and ribosomal RNA (rRNA) sequences from the SILVA database[194]. Contamination screening with Blobtools found 0.01% of reads not mapping to the genome, however these reads were unlikely contaminant as they did not

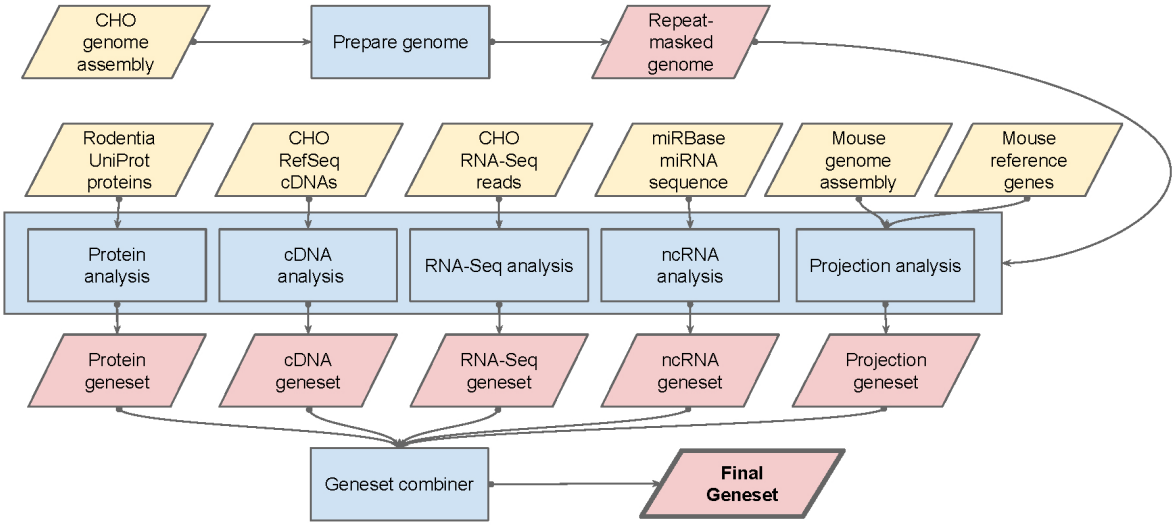

Figure 2.2: Overview of the Ensembl Genebuild gene prediction pipeline.

have hits to SILVA rRNA database. In comparison, the CHO_17A/GY chromosome sorted reference genome showed 0.31% reads unmapped, and while again none of these had hits to the SILVA rRNA database, the GC and coverage profile of these reads was far less consistent and scattered (Figure 7.1,7.2). I then ran BUSCO[139] version 3.0 vertebrata on the CHOK1GS_HD genome to identify homologues to core eukaryotic genes, as an initial assessment of the completeness of the protein coding space. Finally, I aligned the mitochondrial DNA scaffold of CHOK1GS_HD to both the CHO mitochondria reference NC_007936 and the mitochondria scaffold of CriGri_1.0, using BWA-0.7.15-r1140[156] aln, and called variants with Samtools-1.8[113].

2.3.5 Gene Prediction: Ensembl Genebuild

Gene prediction was performed using an updated version of the Ensembl pipeline, described by Aken, *et al*[195], by the Ensembl Genebuild team at the European Bioinformatics Institute (Figure 2.2). In brief, the Ensembl Genebuild team prepared the genome by masking its repeat regions using

RepeatMasker[144] version 4.0.5 (parameters '-nolow -species "rodents" -engine "wublast"'), Dust[196] and TRF[197]. From this, 34% of the genome was identified as repeat sequences. The repeat-masked genome was then used for all subsequent steps, and total repeat counts were extracted and processed from the MySQL database produced using a custom awk script (Table 7.2).

Next, the Genebuild team generated five independent gene sets and aligned them to

the genome in a splice aware manner using GenBlast[198](Table 7.1). These included a subset of UniProt[199] (April 2016 release) proteins that were selected to provide a broad, targeted coverage of the rodent proteome, and aligned with GenBlast, using a minimum cut-off of 50% coverage, identity and e-value of e-20, and the exon repair option. RNA-seq data generated from CHO-K1 GS null cell line that was used for validation of gene models and identification of splice sites. These RNA-seq reads were aligned to the genome using BWA[156], with a tolerance of 50% mismatch to allow for intron identification via split read alignment. Initial models generated from the BWA alignments were further refined using Exonerate[159].

Protein coding models identified using BLAST alignments of the longest open reading frame against the UniProt vertebrate PE one and two data sets. Whole genome alignments generated against the *Mus musculus* reference genome (GRCm38) using LastZ[200], and syntenic regions identified were used to map protein coding annotation from the GENCODE[201] M11 gene set. Finally, small ncRNAs obtained using a combination of BLAST and Infernal[202]/RNAfold[203].

All gene sets were then compared and classified in layers based on available evidence, from the most preferred to least preferred. In general, the highest layers contained the sets of evidence that were considered the most trustworthy in terms of both alignment/mapping quality and relevance to the CHO-K1. Finally, pseudogenes were then annotated by looking for genes with evidence of frame-shifting or lying in repeat dense regions.

2.3.6 Comparative genomics

Protein sequences for CHOK1GS_HD, CriGri_1.0, and eight other species obtained from Ensembl: *Canis familiaris* (GCA_000002285.2), *Cavia porcellus* (GCA_000151735.1), *Homo sapiens* (GCA_000001405.27), *Mesocricetus auratus* (GCA_000349665.1), *M. musculus* (GCA_000001635.8) and *Rattus norvegicus* (GCA_000001895.4), were used in comparative analysis. OrthoFinder[204] was used to identify groups of orthologs between the CHOK1GS_HD gene models and these species. Gene enrichment analysis of orthologous gene families was then performed with AgriGO[205], using the Fisher[206] statistical method, a minimum of five mapping entries, and the false discovery rate was adjusted using the Benjamini–Hochberg procedure[207]. The results of the gene enrichment analysis were filtered for redundant Gene Ontology (GO) terms using REVIGO[208], selecting for a medium (0.7) similarity, using the *Mus musculus* GO database with SimRel[209] semantic measures.

Syntenic blocks between coding regions of CHOK1GS_HD and mouse were identified

through the CoGe platform[210] , through use of SynMAP[211] with BLASTZ[212] alignments, DAGchainer[213] (options -D 30 and -A 2). From these blocks, the syntenic gene pair synonymous rate change was then calculated by CodeML[214].

To further validate GS knockout in the CHO-K1 GS null cell line, coding sequences (CDS) were obtained for GS in *Mus musculus* (ENSMUSG00000026473), CHOK1GS_HD (ENSCGRG00001012857) and CriGri_1.0 (ENSCGRG00000008714). These were then aligned to the CriGri_1.0 reference genome using BLAT-v36[215] and the alignments visualised using theIntegrated Genomics Viewer[216]. Afterwards, multiple sequence alignment (MAS) of the corresponding GS protein sequences for each orthologs was carried using PRALINE[217] with default settings.

2.4 Results

2.4.1 The CHOK1GS_HD genome assembly

Primary sequence data generated from the Horizon Discovery CHO-K1 GS null cell line was assembled *de novo* to create the CHOK1GS_HD genome assembly, as described in the materials and methods section. In brief, we adopted a hybrid multi-step pipeline and assessed the assembly quality through evaluation of the contiguity and completeness of the assembly at the end of each step (Figure 2.1). The quality of our assembly improved with each step of the assembly process, with some steps having more impact than others. Based on the k-mer spectra of the 125 bp and 250 bp reads, we predicted the genome size to be 1.8 Gbp, with a 0.80% heterozygous rate, and 56% coverage of the total genome using both sets of PE reads. These reads were then used to generate the first-pass assemblies using SGA and DISCOVAR *de novo*, of which the quality and completeness SGA was measured to be below that of DISCOVAR *de novo* (Table 2.1).

The quality of the DISCOVAR *de novo* assembly I generated is remarkable given that it is far below the minimum 60% coverage recommended, based on an estimated coverage of 32.7% based of 250 bp PE reads. Furthermore, the BUSCO vertebrata gene set showed 63.7% complete and 24.9% fragmented BUSCOs present in the assembly, with only 11.3% of genes not detected. Dovetail HiRise scaffolding of the DISCOVAR *de novo* then greatly improved N50 and contiguity. An even higher N50 was achieved when BioNano optical mapping was used to produce super-scaffolds. Furthermore, the longest scaffold increased from 1.65 Mb to 157.32 Mb with Dovetail scaffolds, and then to 224.80 Mb in the BioNano super-scaffolds. Whilst the combining of the SGA assembly

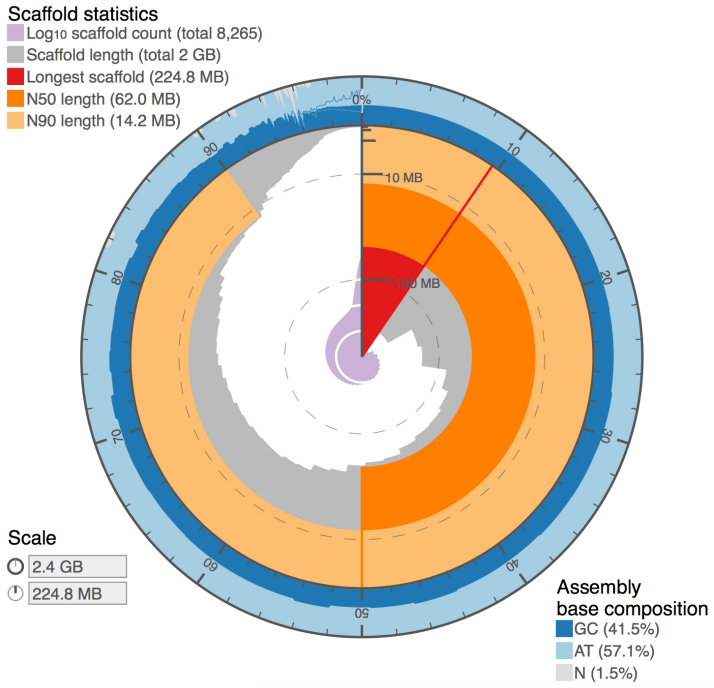

Figure 2.3: Visualisation of the CHOK1GS_HD genome assembly, showing total genome size, distribution of scaffold lengths from largest to smallest going clockwise across the plot, and GC/AT coverage across scaffolds.

and super-scaffolds to complete the hybrid assembly only marginally improved the mean scaffold length and N50, the number of scaffolds was reduced by 21% suggesting smaller scaffolds had been merged and redundant (duplicate) scaffolds removed. All scaffolds below 2 kb in the hybrid assembly were then removed, as k-mer spectra analysis of PE read alignments did not show these to contribute any new k-mers and added unnecessary redundancy to the assembly (Figure 7.4). Removal of these smaller scaffolds gave a final reference assembly with ten fold less scaffolds and identical N50. The GC content remained at 41.5% throughout all assembly iterations (Table 2.1).

The final CHOK1GS_HD genome assembly has a total length of 2.35 Gb, comprised of 8,264 scaffolds with an N50 of 62 Mb (Figure 2.3). CHOK1GS_HD is an improvement in contiguity and completeness over all previous CHO assemblies (including ChiGri_1.0), but of still less contiguous than the high quality human and mouse genome assemblies (Table 2.1,7.7). Total scaffold length between CHOK1GS_HD and CriGri_1.0 showed minor variation of ~30 Mbp, which may be due to differences in the target size and assembly pipelines used. Transposable element content of the two CHO-K1 genomes was relatively similar, with the only notable difference being a lower proportion of the

CHOK1GS_HD genome masked by low-complexity repeat sequences (Table 7.2). Comparison of CHOK1GS_HD and CriGri_1.0 Ensembl annotations shows CHOK1GS_HD to have 20,978 protein coding genes predicted, but less non-coding genes with far fewer long non-coding genes predicted (Table 7.3).

In addition to the nuclear genome, the mitochondrial 16,284 bp genome sequence was also assembled as part of CHOK1GS_HD. Compared to a recently reconstructed host CHO mitochondria sequence, the same variants as the CriGri_1.0 mtDNA sequence were observed (Table 7.6)[218]. This is possibly due to the recent (2011) derivation of the Horizon Discovery CHO-K1 GS null cell line from the same CHO-K1 ATCC cell line sequenced for the CriGri_1.0 assembly, as both CHOK1GS_HD and CriGri_1.0 mtDNA sequences showed no variants (SNPs or Indels) when aligned to each other.

2.4.2 Identification of new orthologous groups in CHO

I investigated gene orthology between CHO-K1 and related species, as an assessment of the assembled and annotated gene space, and identified 13,156 gene families with an orthologous relationship between CHOK1GS_HD, CriGri_1.0, *C. familiaris*, *C. porcellus*, *H. sapiens*, *M. auratus*, *M. musculus* and *R. norvegicus* (Figure 4.2). Of which, 10,360 families contained only single copy orthologs, and 16,866 orthogroups were shared between both CHOK1GS_HD and CriGri_1.0. Of all genes in CHOK1GS_HD, 97.8% could be assigned to 20,372 orthologous gene families, compared to 19,036 (97%) in CriGri_1.0. Of these, 615 gene families showed orthology only between CHOK1GS_HD and CriGri_1.0. Additionally, 1,055 orthogroups were found in CHOK1GS_HD, but not CriGRi_1.0 which had 203 orthogroups present in all references apart from CHOK1GS_HD. These counts were reduced to 111 (111 genes), 597 (599 genes) and 51 (53 genes), when selecting orthogroups with assigned annotated gene ontology (GO) terms, respectively. Gene enrichment analysis showed 35 (27 non-redundant) GO terms to be significantly (Benjamini–Hochberg adjusted p-value [padj] < 0.05) enriched in the CHO only gene families when compared to those conserved between all species (Table 7.4). Interestingly, several of the most significantly enriched GO terms were related to pathways associated with olfactory receptors and sense of smell (Figure 2.5). This could be associated with the improved assembly quality, as this gene family is known to be highly duplicated and therefore hard to assemble[219]. While no GO terms were found to be significantly enriched in either CHO individually, there were 298 (85 non-redundant) significantly enriched GO terms in the CHOK1GS_HD orthogroups that were not shared with CriGri_1.0 (Table 7.8). The top three of these represent metabolic and biological process

pathways responsible for cell fate and development, and RNA transcription (Figure 2.6). Conversely, no gene ontology terms were seen to be enriched in the orthgoroups found in CriGRi_1.0, but not CHOK1GS_HD (Table 7.5).

2.4.3 Improved genome synteny to mouse

To investigate macrosynteny between CHOK1GS_HD and *M. musculus*, and find out more about the evolutionary history of this CHO cell line and it's karyotype, I carried out whole-genome syntenic dotplot analysis (Figure 3.6). Remarkably, a number of CHOK1GS_HD super-scaffolds cover near whole chromosomes of *M. musculus*, indicating a high level of accuracy in the gene order of the assembly. For example, 126.7 Mb super-scaffold_2 covers the majority of the 156.5 Mb *M. musculus* chromosome 4. The synteny between these two chromosomes reflects previous findings using reciprocal chromosome painting between *C. griseus* and *M. musculus*[220, 221]. This suggests that super-scaffold_2 maybe CHO-K1 chromosome C or D, based on known chromosome rearrangements from BAC-based physical maps[222, 223]. These maybe unique features of CHO-K1 GS null's karyotype or potential mis-assemblies. It should also be noted that the karyotype in CHO has been shown to be unstable, and to vary within and between populations[185, 195]. Further study into the karyotype and population structure of the CHO-K1 GS null host cell line would be beneficial for further improvement of this cell line and also could be used to join super-scaffolds and achieve a chromosome level reference.

2.4.4 Confirmation of complete Glutamine Synthetase knock-out

One of the major hurdles of recombinant protein production is the generation and selection of recombinant clones that provide sufficient yield and growth rate. Two of the most common screening systems employed in CHO are the Methotrexate/Dihydrofolate reductase (MTX/DHFR) and Methionine Sulfoximine/Glutamine synthetase (MSX/GS), whereby the presence of MTX inhibits DHFR and MSX inhibits GS; both vital enzymes for cellular metabolism[173]. However, these systems lead to over-expression of endogenous DHFR or GS, and inhibitors can affect the quality and production of recombinant protein production[224]. Alternatively, CHO cell lines deficient in DHFR, and more recently GS, have been developed[186, 225]. The later of which was reported by Liachun, *et al*, utilising zinc-finger nuclease (ZFN) mRNAs to target the fifth coding exon of the CHO

GS gene, resulting in a GS knockouts showing no GS protein expression and glutamine-dependent growth[186]. Furthermore, use of the GS knockouts have been shown to improve stringency of screening compared to DHFR knockouts and drug inhibition screening systems[186, 226].

To this end, Horizon Discovery Ltd generated the CHO-K1 GS null cell line, following an rAAV targeting strategy based on that used by Liu, *et al*[188]. Whilst the success of this knockout has been reported previously through PCR and phenotyping (www.horizondiscovery.com/bioproduction/cho-cells/cell-line-derivation), comparison of CHOK1GS_HD with the *M. musculus* and CriGri_1.0 genomes further validates the successful knockout of GS (Figure 2.8).

Table 2.1: Comparison of assembly metrics, data type and assembler used for different iterations of the CHO genome assembly, and other publicly available *C. griseus* assemblies.

Assembly	**#scaffolds**	**Longest scaffold (Mbp)**	**Total length (Mbp)**	**GC (%)**	**N50 (Kb)**	**Avg. #N's per 100 kbp**	**BUSCOs (%)**
SGA	7,229,676	0.064	2,256.22	41.45	5	0	21.00
Dioscovar *de novo*	877,181	1.65	2,410.21	41.45	158	63.59	79.20
Dioscovar *de novo* + Dovetail	821,041	157.32	2,415.46	41.45	34,102	265.27	95.80
Dioscovar *de novo* + Dovetail + Bionano	820,943	224.8	2,443.56	41.45	61,985	1,411.88	95.30
Dioscovar *de novo* + Dovetail + Bionano + SGA	151,867	224.83	2,428.16	41.45	62,039	1,410.29	95.30
Dioscovar *de novo* + Dovetail + Bionano + SGA + > 2 kb Scaffolds	8,265	224.83	2,358.16	41.45	62,039	1,452.08	95.60
CriGri_1.0	109,151	8.77	2,383.17	41.37	1,165	3,419.18	93.10
BGI *C. griseus*	52,710	8.32	2,351.10	41.39	1,571	2,501.81	94.70
CHO_17A/GY	28,749	14.65	2,332.77	41.27	1,236	10,450.23	93.00
H. sapiens GRCh38	194	248.95	3,099.75	40.86	145,138	4,964.97	92.60
M. musculu GRCm38	66	195.47	2,730.87	41.67	130,694	2,859.46	95.30

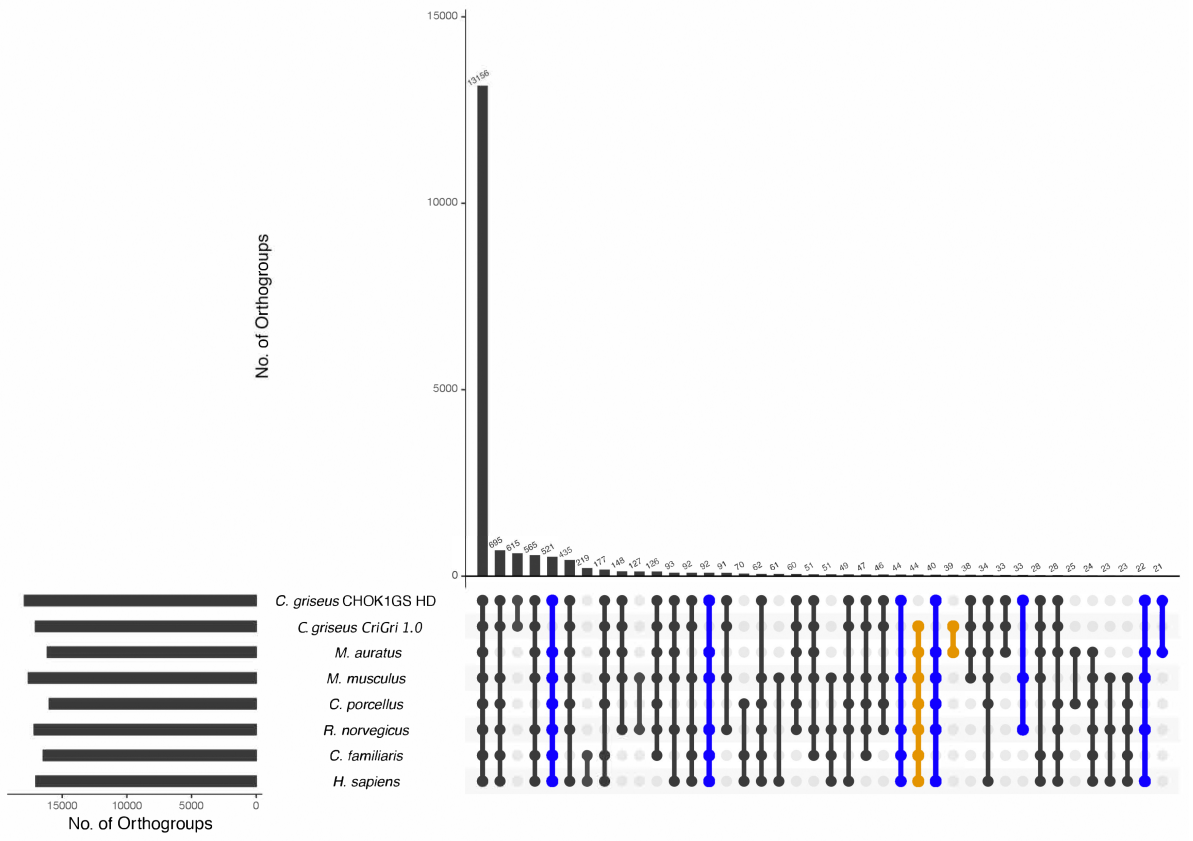

Figure 2.4: UpSet plot showing conserved orthogroups among CHOK1GS_HD, CriGri_1.0, *C. familiaris*, *Cavia porcellus*, *H. sapiens*, *M. auratus*, *M. musculus*, and *R. norvegicus*. Nodes (blue) below the bar chart are orthogroups present in CHOK1GS_HD and the other species, but not CriGri_1.0, and nodes (gold) are orthgoroups present in CriGri_1.0, but not CHOK1GS_HD.

Figure 2.5: Hierarchical tree graph of CHO only enriched GO terms, in the biological process category, related to olfactory receptor and sense of smell. Boxes in the graph represent GO terms labelled by their GO ID, term definition and statistical information. Boxes with significant terms (padj < 0.05) are filled with a nine-level colour gradient from yellow to red, to indicate their level of statistical significance (padj values displayed in box). Boxes with non-significant terms are white.

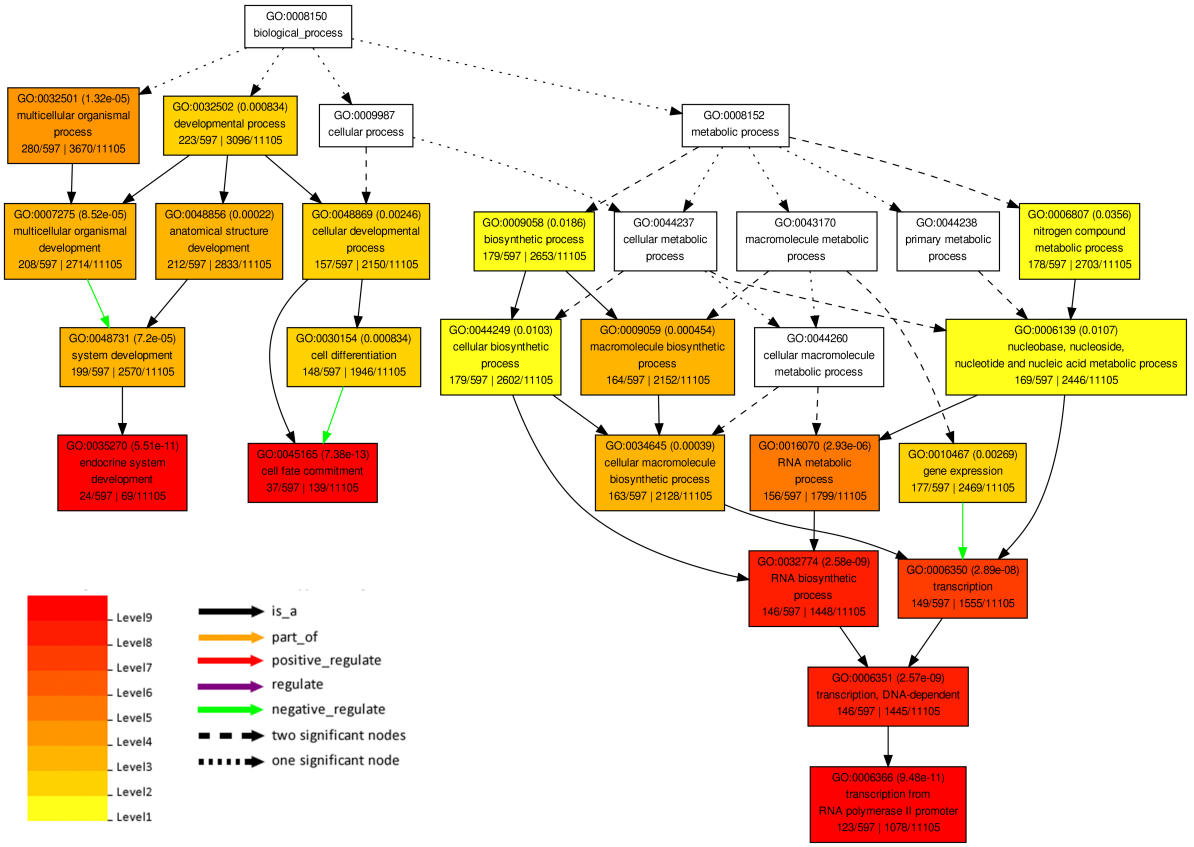

Figure 2.6: Hierarchical tree graph of the three most significantly enriched GO terms in the biological process category, observed in CHOK1GS_HD and every other genome studied apart from CriGri_1.0. Boxes in the graph represent GO terms labelled by their GO ID, term definition and statistical information. Boxes with significant terms (padj < 0.05) are filled with a nine-level colour gradient from yellow to red, to indicate their level of statistical significance (padj values displayed in box). Boxes with non-significant terms are white.

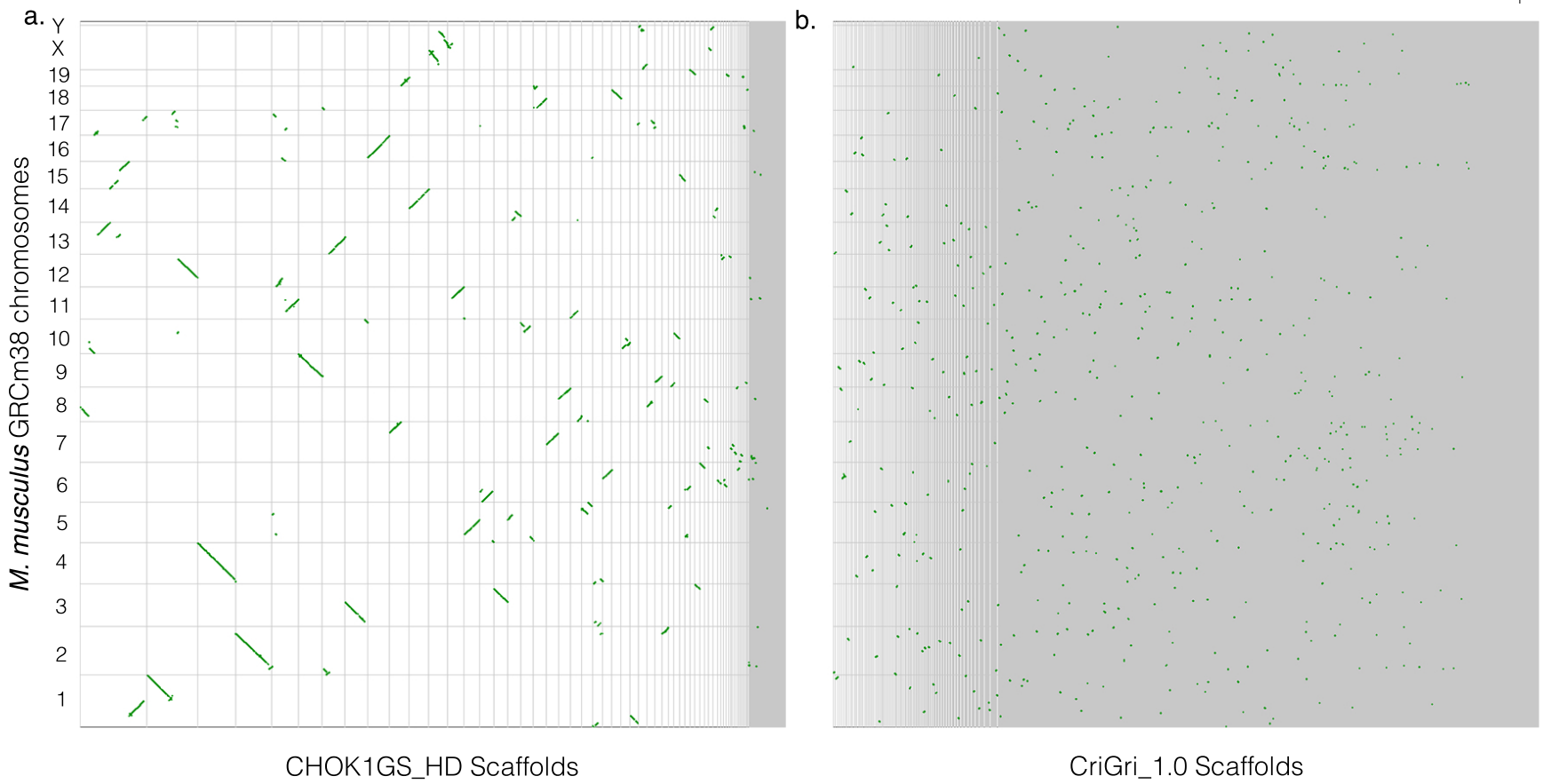

Figure 2.7: Whole genome synteny between *M. musculus* GRCm38 chromosomes and the two CHO scaffold assemblies; a. CHOK1GS_HD, b. CriGri_1.0. Syntelogs have been coloured based on their synonymous rate change.

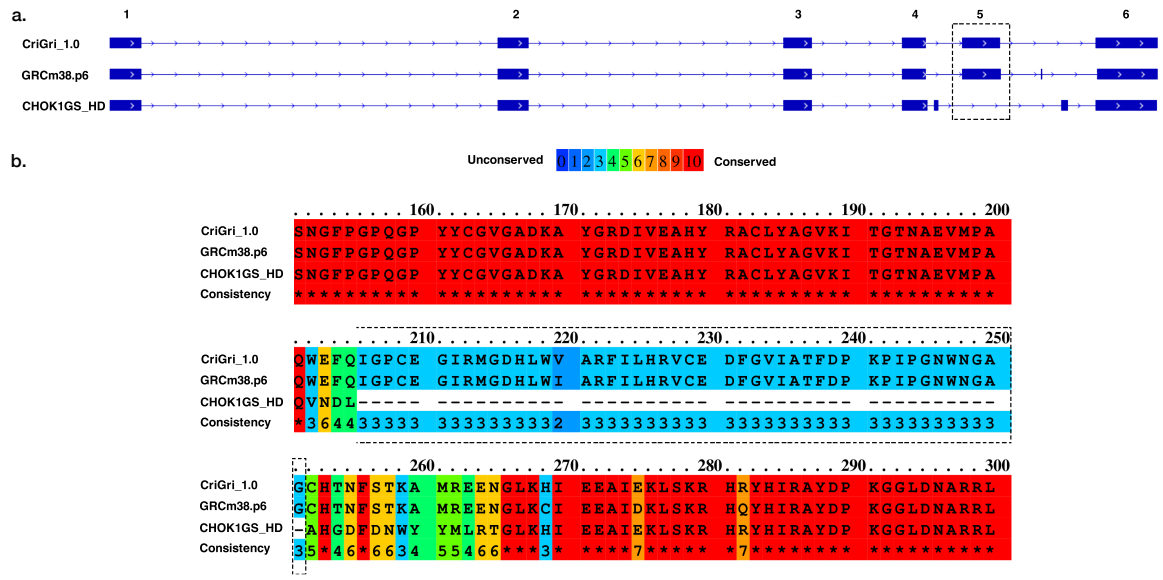

Figure 2.8: Confirmation of GS knockout in CHO-K1 GS null through (a) alignment of CHOK1GS_HD and GRCm38.p6 (GS) CDS DNA sequences to CriGri_1.0 reference, showing deletion of the fifth coding exon and flanking LoxP sites in CHOK1GS_HD, and (b) MAS of GS protein sequences between CriGri_1.0, GRCm38.p6 and CHOK1GS_HD.

2.5 Discussion

Over the past 30 years the pharmaceutical industry has considerably redesigned every part of the bioproduction process, improving productivity. However, during this time the CHO cell line, arguably one of the greatest potential source of efficiency improvements, has remained relatively unchanged since it's immortalisation by T. Puck in 1957[168]. Genome editing has become an exciting and ever progressing technology for engineering cell lines for therapeutic protein production. However, this requires a high-quality genomic reference and availability of the cell line that corresponds to the reference. Advances in long read sequence and genome mapping make it now possible to achieve a more structurally accurate model of the genome, with far longer (chromosome or near) scaffolds, for a fraction of the cost and time previously involved[128].

To this end, we have generated a new reference for CHO, specifically the Horizon Discovery CHO-K1 GS cell line, that will allow improved accuracy and confidence in the use of this cell line. To achieve this, a hybrid assembly was generating using short and long reads, and genome mapping. Generating a reference which a higher N50, more comprehensive annotation and vastly improved resolution of synteny to related

species. This combination of Dovetail[94] and Bionano[131] to successfully join scaffolds has similar been reported elsewhere in a number of eukaryote species[182, 227–229]. Improvements to the gene space from the use of genome mapping have also been observed in the *M. musculus* GRCh38 reference, compared to previous iterations[230]. While the improved completeness and contiguity could be validated, assessment of accuracy using more traditional methods, such as REAPR[136], were not possible due to inaccessibility of Chicago reads from Dovetail that would of added long range information to check for inversions and duplications in super-scaffolds. However, through increased macro synteny with *M. musculus*, I observed a higher degree of long range accuracy than CriGri_1.0.

When comparing the annotations of CHOK1GS_HD and CriGri_1.0, with several other rodent genomes, I observed a number of differences between these. CHOK1GS_HD had orthogroups with significant enrichment of GO terms for pathways associated with cell fate and development that were not present in the CriGri_1.0. Both CHO annotations showed a significant enrichment of orthogroups associated with olfactory receptors, compared to the other rodent genomes studied. Olfactory receptors represent one of the most diverse and the largest multigene family of terrestrial vertebrates, with some species having thousands of olfactory receptor genes[219]. Significant differences in the diversity and proportions of olfactory receptor families have been observed between distant and related species[219]. However, while this finding may be biological, olfactory receptor genes have also been shown to be difficult to assemble and as a result have produced conflicting reports in the literature[231]. Copy number variants can also make it difficult to recover the complete repertoire of olfactory receptor genes with short read sequencing alone[219]. It's therefore possible the expansion of olfactory receptors observed in CHO maybe technical, rather than showing evidence of adaptive evolution in *C. griseus*. Vishwanathan, *et al*'s transcriptome study of two *C. griseus* tissues and six CHO cell lines found the majority of transcripts associated with olfactory transduction pathways were not recovered in the transcriptome assemblies[232]. In order to validate whether or not these expansions in CHO are 'real', we would need to move beyond the cell line to the host organism and carry out both intra and interspecies surveys of olfactory receptors at the population level, taking into account the life histories of both *C. griseus* and other closely related rodent species. However, compound effects of assembly and annotation quality of the species studied would still have an impact on any natural variance that maybe observed.

Differences in the number of non-coding genes between CHOK1GS_HD and CriGri_1.0, in particularly long non-coding genes, were also seen. The majority of these

are likely technical, as both assemblies are derived from the same cell line and are potentially due to changes made to the Ensembl full annotation pipeline between the release of CriGri_1.0 and CHOK1GS_HD. Closely related species with recent Ensembl releases/updates show a similarly low number of long non-coding genes, such as *Peromyscus maniculatus bairdii* (GCA_000500345.1) with 28 long non-coding genes (3,962 non-coding total) and *Mesocricetus auratus* (GCA_000349665.1) with 23 long non-coding genes (3,720 non-coding total). The assembly of CHOK1GS_HD may also have better captured of full length genes that were otherwise fragmented in CriGri_1.0 and not validated as coding. As such, technical differences between the two genomes should explain the majority of variance observed, but some potential biological factors inherent of the CHO-K1 cell line, such as passage and copy number variances, could account for a small amount of this variation. Furthermore, as CHO-K1 is a single tissue cell line, it would also be difficult to find supporting evidence for the loci of genes that are expressed in a tissue specific manner (outside of the ovaries), thus potentially leaving a large number of genes as 'non-coding'. The reduction in non-coding genes observed in CHOK1GS_HD is potentially an under representation compared to the other genome though and attention should be paid to this in any updates to the reference genome.

Leading on from this, alignment of the CHOK1GS_HD mtDNA scaffold to the *C. griseus* mitochondrial genome showed identical variants to those observed in the CriGri_1.0 reference, when compared to the Kelly *et al* reconstructed host CHO mitochondrial. This is particularly interesting as Kelly, *et al*, observed 197 mutations across 22 commercial CHO cell lines[218]. Of which, all 17 CHO-K1 cell lines showed between three and 12 variants, included at least one SNP in both the tNRAval and16S rRNA genes. This indicates a degree of heteroplasmy in CHO cell lines and the lack of observed variation in CHO-K1 GS null maybe be due to the recent derivation (2011) of the cell line from the ATCC CCL-61 CHO-K1 host that CriGri_1.0 was assembled from. However, as CHO-K1 ATCC was used by Kelly, *et al*, and showed a different mitochondria phenotype, it's possible this is a different ATCC line than the European Collection of Authenticated Cell Cultures ATCC CCL-61 used in a development of the Horizon Discovery CHO-K1 GS null. It should also be noted that the previously mentioned study found the (Partridge *et al*, 2007) GenBank NC_007936 *C. griseus* mtDNA sequence to be derived from either CHO-K1 or CHO AL cell lines, and not *C. griseus*[218, 233, 234]. Alignment of NC_007936 to the CHOK1GS_HD mtDNA sequence shows no variants, and is indeed likely derived from the CHO-K1 ATCC CCL-61 cell line or CHO AL cell lines, which may share the same mitochondrial genome. Finally, I also further validated the successful

knockout of GS in the CHOK1GS_HD host cell line, originally generated through use of a rAAV targeting strategy[188].

While CHOK1GS_HD represents an improvement over the previous public reference(s), further work to investigate the karyotype of CHOK1GS_HD and to join super-scaffolds to generate chromosomes, perhaps with linked-reads from Hi-C[94] or 10x Genomics[96], is required. This is particularly important in addressing the issues of CHO genome instability, in particular clonal instability, raised here[184] and here[185]. Additional accompanying long reads from single molecule sequencing technologies, like those offered by Oxford Nanopore[100] and PacBio[181], would also be useful for gap filling and reconstruction of highly repetitive regions.

Finally, I believe that utilisation of CHOK1GS_HD will allow for more guided genome engineering, aided discovery of novel endogenous promotors and cis-elements that will provide greater control over transcription initiation compared to viral alternatives. Providing improved insight and control over cellular longevity and metabolism, that will ultimately increase the productivity and quality of recombinant proteins[235–237]. The assembly and annotation workflow presented here is also useable towards generation of assemblies in other species, that may lack an existing reference, and is a good display of the current state of the art and improvements that have been made to sequencing technologies and related informatics tools over the last eight years. Furthermore, import of CHOK1GS_HD into the Ensembl platform will ensure accessibility and improve adoption of this resource in the community, increasing confidence and accuracy in future studies utilising this cell line and improved reference genome.

Chapter 3

Annotation and exploration of the first yam genome

The work contained in this chapter was used towards an open access publication, reported in BMC Biology **doi: 10.1186/s12915-017-0419-x*****, of which I am a co-first author and primarily worked with our collaborators at the Iwate Biotechnology Research Center, Japan. All findings reported in this chapter are directly derived from my own work and the implication of this on the overall finding of the paper, by my collaborators, is discussed in brief at the end of the chapter. A copy of this paper is provided in Appendix B.***

3.1 Abstract

White Guinea yam (*Dioscorea rotundata* Poir.) is a common staple food that has contributed enormously to the subsistence and socio-cultural life of millions of people principally in West and Central Africa. Here I assembled and annotated the 594 Mb genome of a heterozygous line of *D. rotundata*. The genome sequence combined with RNA-seq and homologous protein mapping predicted a total of 26,198 genes. Amongst these genes I observed an expansion of bulb-type mannose specific binding lectins, that could be potentially involved in tuber defence. Phylogenetic analysis and comparative genomics of *Dioscorea* suggests the genus diverged early on from other sequenced monocotyledonous species and has a unique evolutionary history. Through use of the genome and QTL mapping by whole genome resequencing of bulked segregants, our collaborators identified a genomic region and candidate genes associated with female heterogametic (male=ZZ, female=ZW) sex determination, opening up genomic avenues for the genetic improvement

of this socio-economically important but understudied crop.

3.2 Introduction

Yam is a collective name for tuber-bearing crops belonging to the monocotyledonous *Dioscorea* genus under the family Dioscoreaceae. The genus is composed of two branching clades that consist of 10 major subclades, containing over 600 species, with those belonging to the later B branching clade spread throughout tropical and sub-tropical regions of the world[62, 70]. Thought to have been consumed in West Africa and Asia since 50,000 BC, with cultivation starting around 3,000 BC, yam has become the third most important tuber crop in these continents[69–71]. Of these, white Guinea yam (*D. rotundata*) is the most popular species in West and Central Africa; the dominant region for yam production in the world. In 2016, approximately 94% of the 65 million tons of yam produced globally came from the West-African countries[72]. Yams, in particular *D. rotundata*, are important to the region as a major source of food, income, and as an integral part of the socio-cultural life[76]. So much so that this region is often referred to as the 'Civilization of the yam', capturing the tight link between West African societies and yam cultivation[74, 75]. Yams are not only important in terms of food security, but also as major producers of steroid precursors and other compounds of potential medicinal value[238, 239].

Despite its considerable importance, the white Guinea yam has long been regarded as a neglected 'orphan' crop and its cultivation is constrained by several factors that need to be addressed by any programs aiming to improve its production. Botanical seeds are seldom used as starting materials, with yams commonly propagated clonally using small whole tubers (referred to as 'seed yams') or tuber pieces. Due to obligate outcrossing, each plant is highly heterozygous and a cultivar is often composed of genetically diverse clones. Yams are also annual climbers that requires stakes for support, and are highly vulnerable to a plethora of pests and diseases. Furthermore, the entire genus is characterised by dioecy, the presence of separate male and female plants, which is a rare trait found in only 5-6% of angiosperms and is thought to be synapomorphic[40, 62]. However, little is known about the genomic nature of dioecy in *Dioscorea*.

Therefore, improvements in traits associated with tuber yield and quality, staking reduction as well as resistance/tolerance to disease and nematodes, and an improved understanding of sex determination are much needed. The genetics of *Dioscorea* has not been studied in depth and yam improvement has relied mainly on clonal selection

with limited application of cross breeding via Marker Assisted Selection (MAS); stressing the need for DNA markers showing association with traits of interest. Generation of a reference genome for this species will contribute significantly towards broadening our knowledge of the genetics of the white Guinea yam, improve our understanding of the genus and of the evolution of dioecy as a whole, and will ultimately accelerate breeding improved yams.

To this end, an international collaboration was established and led by The International Institute of Tropical Agriculture (IITA), to generate genetic and genomic tools for accelerating yam breeding. As part of this I assisted with the assembly, annotation of a diploid genome of *D. rotundata*, and downstream analysis investigating the evolutionary history and sex determination of this orphan crop species.

3.3 Methods

3.3.1 Evaluation of the genomic assembly completeness

Sequencing and assembly was carried out by collaborators at the Iwate Biotechnology Research Center, Japan (IBRC) and detailed here in short. A single-plant, referred to as "TDr96_F1", from the progeny of the open-pollinated *D. rotundata* breeding line TDr96/00629 was used. This individual was found to be diploid ($2n = 2\times = 40$), based on the mitotic chromosome number within root meristem cells, and the total genome size was estimated to be 570 Mb by flow cytometry (FCM) analysis. Fresh leaf material from TDr96_F1 was used to generate PE libraries and eight types of mate-pair (MP) jump libraries, with insert sizes of 2, 3, 4, 5, 6, 8, 20, and 40 kb. These were sequenced using Illumina MiSeq and HiSeq 2500 platforms. An additional 100-kb jump bacterial artificial chromosome (BAC) library was also generated and subject to BAC-end Sanger sequencing. Assembly of the genome using reads filtered and cleaned, using the FASTX toolkit[114], was performed with ALLPATHS-LG[125] and SSPACE PREMIUM[126] for BAC-end scaffolding. RAD-based linkage maps of F1 progeny and parental lines were then used for anchoring of scaffolds into pseudo-molecules.

To evaluate the completeness of the *D. rotundata* genome assembly, the assembly was checked for the presence of 248 highly conserved core eukaryotic genes using CEGMA[138] version 2.4 with default parameters. To further assess the completeness of the genome, the successor to CEGMA, Benchmarking Universal Single-Copy Orthologs (BUSCO)[139], was used to check for the presence of 956 BUSCOs with version 1.1.b1 using the early

access plant dataset.

3.3.2 Annotation of transposable elements

Repetitive sequences were predicted using RepeatModeler 1.0.8[141] and masked with RepeatMasker 4.0.5[141]. Using the National Center for Biotechnology Information (NCBI) database, one of three options was used: interspersed RepeatModeler-based, interspersed Rebase-based, and Low complexity repeats: "nolow" (does not mask low complexity sequences), "nolow, species Viridiplantae", and "noint" (no interspersed repeats, masks complex/simple repeats), respectively. Repeat element content and other statistics were compared between the *D. rotundata* and *A. thaliana* TAIR10[240], *B. distachyon v3.1*[241], and *O. sativa* v7_JGI 323[242] genomes using the RepeatModeled and RepeatMasked references (Table 1).

3.3.3 Prediction of protein-coding genes

An AUGUSTUS[145] set of *ab inito* gene models, from here on referred to as the “legacy”, were generated previously by S. Natsume at IBRC, using the legacy repeat-masked reference genome and three approaches: *ab inito*, *ab inito* supported by evidence-based prediction, and evidence-based prediction. *ab inito* prediction was carried out with FGENESH 3.1.1[243]. The *ab inito* supported by evidence-based prediction was performed with AUGUSTUS 3.0.3 using the maize5 training set (the default monocot set; improved over original ’maize’ set) and a hint file as the gene model support information. To construct the hint file, TopHat 2.0.11[244] was used to align RNA-seq reads from tuber, flower (young), leaf (young), stem, leaf (old), and flower (old) samples to the *D. rotundata* reference genome, and Cufflinks 2.2.1[151] was used to generate gene models from these data.

The evidence-based predictions using the Program to Assemble Spliced Alignments (PASA) were generated in a Trinity assembled transcriptome from the RNA-seq data. JIGSAW 3.2.9 was used to select and combine the gene models obtained using the three approaches with the weighting values assigned to the results from FGENESH, AUGUSTUS, and PASA of 10, 3, and 3, respectively[149, 150, 245]. In total, 21,882 consensus gene models were predicted.

I then further improved upon these gene models using the MAKER pipeline (Figure 3.1)[246]. Publicly available ESTs and protein sequences from related plant species were aligned to the genome using GMAP and Exonerate 2.2.0, respectively[158, 159]. *De novo*

and reference-guided transcripts were assembled from RNA-seq data from 18 tissues (tuber top, middle and bottom, young and old flowers, young and old leaves, stem, axillary bud, flower bud, rachis, lower and upper inflorescence, pulvinus, petiole, root and spine) using Bowtie 1.1.1, Trinity 2.0.6 and SAMtools 1.2.0, and Trinity 2.0.6 and TopHat 2.1.0, respectively[113, 247, 248]. Both sets of assembled transcripts were used to build a comprehensive transcript database using PASA (Additional file 1: Table S13). High-quality non-redundant transcripts from PASA were used to generate a training set for AUGUSTUS 3.1. Gene models were predicted twice using the genome, improved repeat sequences, assembled transcripts, EST and protein alignments, the AUGUSTUS training set, and a legacy set of 21,882 gene models obtained previously using MAKER 2.31.6, retaining all legacy gene models or querying them with new evidence and discarding those that could not be validated. From both MAKER runs, 21,894 and 76,449 gene models were predicted, respectively.

A consensus set of gene models from both MAKER outputs was obtained, by S. Natsume, using JIGSAW 3.2.9 at a 1:1 ratio. In total, 26,198 consensus gene models were predicted in the *D. rotundata* genome. The corresponding amino acid sequences were also predicted for these gene models. To confirm these gene models, the RNA-seq reads were aligned to the coding sequences (CDS) of the predicted genes using BWA with default parameters. Accordingly, 85.8% of the gene models could be aligned by at least a single RNA-seq read[249].

Functional annotation of the amino acid sequences was performed using the in-house pipeline, AnnotF, which compares Blast2GO and InterProScan functional terms[165, 166].

3.3.4 Gene expression and enrichment

Enrichment of tissue-specific genes was predicted using TopHat 2.1.0 to align RNA-seq data from each of the 12 tissues to the genome, with one biological replicate for each tissue. HTSeq 0.6.1 was used to generate raw counts[250]. Then the Bioconductor package DESeq2 1.14.1 was used to compare raw counts of the three tuber tissues against all the other nine tissues to determine tissue specific enriched gene expression based on a log2 fold change > 0 and Benjamini-Hochberg adjusted p value < 0.05[207, 251].

Gene expression across pseuochromosomes was visualised using Bioconductor package KaryoploteR[252]. Gene enrichment analysis of orthology clusters was performed with GOATOOLS, using the Holm significance test, and the false discovery rate was adjusted p value using the Benjamini-Hochberg procedure[253, 254]. The list of enriched genes

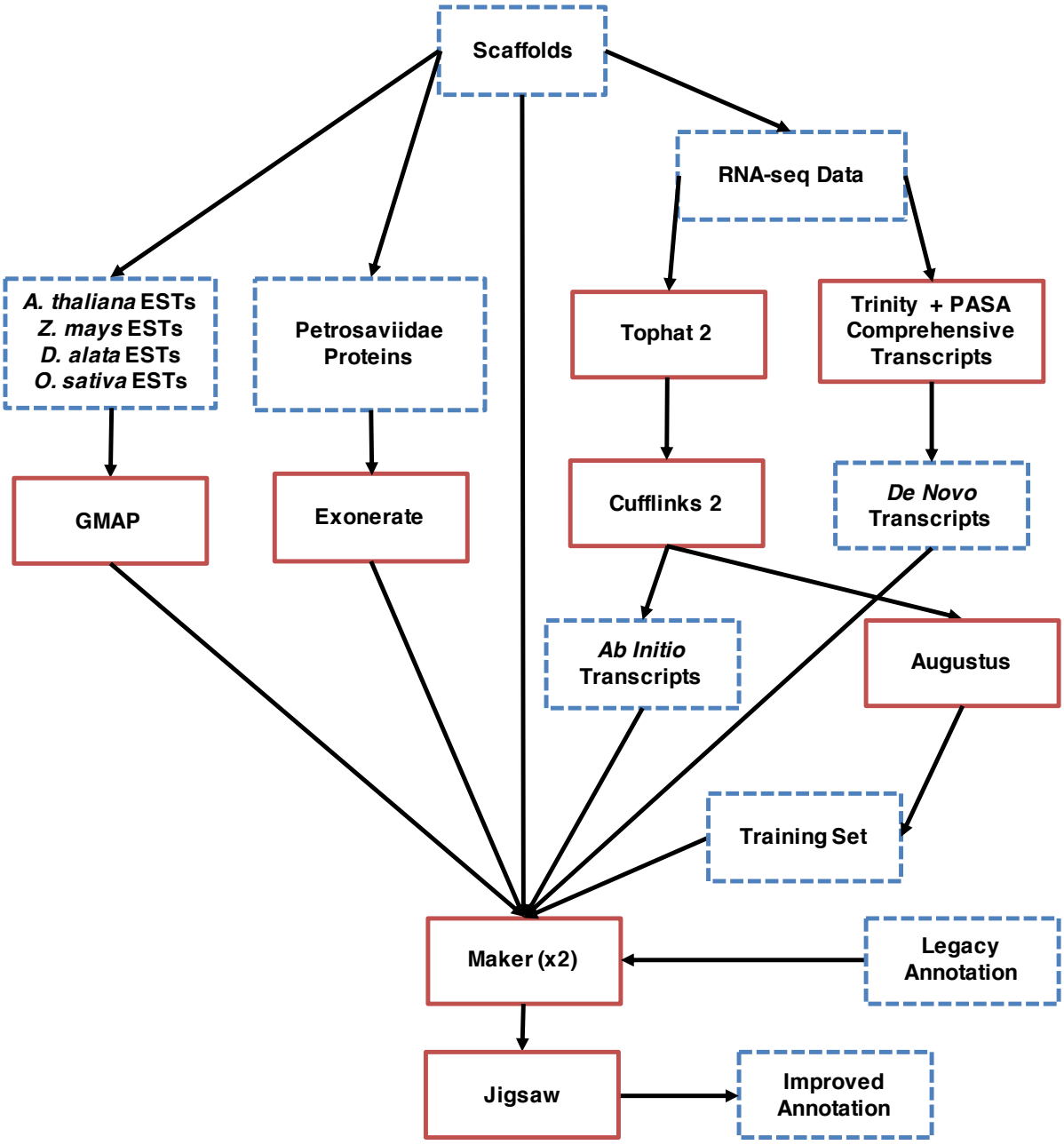

Figure 3.1: An outline of the annotation pipeline used, with inputs/out (blue, dashed boxes) and programs used (red, solid line boxes). **This figure has also been published in Tamiru, M., Natsume, S., Takagi, H., White, B., Yaegashi, H., Shimizu, M., *et al*, 2017 and appears here with permission.**

was filtered for redundant GO terms using REVIGO[208].

3.3.5 Comparative genomics

Pairwise orthologous relationships were determined with Inparanoid[255–258] using the longest protein-coding isoform for each gene in *Arabidopsis thaliana*(TAIR10), *Oryza sativa japonica* (v7.0), *Brachypodium distachyon* (v3.1), *Musa acuminata* (v2)[259], *Elaeis guineensis* (EG5)[260], and *Phoenix dactylifera* (DPV01)[261]. Orthologous groups across all seven species were determined using Multiparanoid[255]. Sequences for the 12 classes of lectins were obtained from UniProt for the proteomes of *A. thaliana* (up000006548), *B. distachyon* (up000008810), and *O. sativa* (up000059680)[262]. Protein alignments for B-lectin class protein sequences from all three of these species and *D. rotundata* were generated using the program 'Multiple Alignment using Fast Fourier Transform' (MAFFT)[263]. Maximum likelihood trees were constructed based on the concatenated alignments of all 378 B-lectin proteins using RAxML 8.0.2 with 1,000 bootstraps[264]. For the species phylogeny, 190 1:1 orthologous genes, based on best reciprocal hits using blastp (BLAST 2.3.0) hits with -evalue 1e6 -use_sw_tback parameters, from *D. rotundata* to *Aegilops tauschii**, *Ananas comosus*[265], *Arabidopsis thaliana*, *Beta vulgaris*[266], *Brachypodium distachyon*, *Carica papaya*^ ,*Chenopodium quinoa*[227], *Elaeis guineensis*[260], *Ipomoea nil*[267], *Malus domestica*[268], *Musa acuminata*[259], *Nelumbo nucifera*[269], *Oropetium thomaeum*[270], *Oryza sativa* Japonica, *Panicum hallii*†, *Phalaenopsis equestris*[271], *Phoenix dactylifera*[261], *Setaria viridis*†, *Sorghum bicolor*[272], *Zostera marina**, *Saccharum officinarum*^, *Spirodela polyrhiza*[273],*Populus nigra*^, *Vitis vinifera*^, and *Amborella trichopoda*[274] as an outgroup, were used to generate protein sequence alignments with MAFT, using the longest protein isoform (*UniProt[262], ^PlantGBD[275], †Phytozome[153]). Multiple sequence alignment for Debranching Enzyme 1 between the 26 species are provided in Supplementary Data 1, as support for the analysis. These species were chosen as they cover the majority of monocot clades that have well defined reference sequences available, in order to investigate the phylogentic position of Dioscoreaceae. Maximum likelihood trees were constructed based on the concatenated alignments of 2381 orthologous protein-coding genes using RAxML 8.2.8 with a JTT + Γ model and 1000 bootstraps. SynMAP[211] using BLASTZ[212] alignments, DAGchainer[213] (options -D 30 and -A 2), and no merging of syntenic blocks were used, as part of the CoGe platform, to identify syntenic blocks between the hard-masked pseudo-chromosomes of *D. rotundata* and scaffolds/contigs of *O. sativa* Japonica (A123v1.0)[211–213, 276], *S. polyrhiza* (v0.01), and *P. dactylifera*

L. (v3)[276]. A syntenic path assembly was then carried out on each of the same three species in SynMap using synteny between the scaffolds/contigs against *D. rotundata* pseudo-molecules. The syntenic path assembly is a reference-guided assembly that uses the synteny between two species to order and orientate contigs. This approach highlights regions of conservation that were otherwise too shuffled to be clearly observed. Self-self synteny analysis of *D. rotundata* pseudo-chromosomes was carried out using SynMap Last alignments with default parameters and syntenic gene pair synonymous rate change calculated by CodeML[214].

3.4 Results

3.4.1 Gene prediction and genomic content

I assessed the completeness of the *D. rotundata* assembly by checking for the presence of 248 highly conserved core eukaryotic genes using CEGMA and confirmed the presence of 243 (98%) at least partially present genes (Table 3.1). Similarly, I further validated this finding through use of the successor to CEMGA, BUSCO, and observed 94% of 956 benchmarking universal single-copy orthologs (BUSCOs) with at least one complete single-copy present in the assembly (Table 3.2). With the majority of core orthologs

Table 3.1: Assessment of the completeness of *D. rotundata* genome assembly using the 248 most highly-conserved Core Eukaryotic Genes by CEGMA. **This table has also been published in Tamiru, M., Natsume, S., Takagi, H., White, B., Yaegashi, H., Shimizu, M., *et al*, 2017 and appears here with permission.**

Scaffold Size	Scaffold N50	No. of eukaryotic genes identified	% Completeness
594 Mb	2.12 Mb	226	91.13% (complete*)
594 Mb	2.12 Mb	243	97.98% (partial*)

*"'Complete' refers to those predicted proteins in the set of 248 CEGs that when aligned to the HMM for the KOG for that protein-family, give an alignment length that is 70% of the protein length. I.e. if CEGMA produces a 100 amino acid protein, and the alignment length to the HMM to which that protein should belong is 110, then we would say that the protein is "complete" (91% aligned). If a protein is not complete, but if it still exceeds a pre-computed minimum alignment score, then we call the protein 'partial'. Note that a protein that is deemed to be 'Complete' will also be included in the set of Partial matches".

predicted to be present in the assembly, I could hypothesise that a large portion of the coding regions of the genome should be contained within it. With this in mind,

Table 3.2: Assessment of the completeness of *D. rotundata* genome assembly using 956 benchmarking universal single-copy orthologs (BUSCO) by BUSCOv1.1.b1 using the early access plant dataset. **This table has also been published in Tamiru, M., Natsume, S., Takagi, H., White, B., Yaegashi, H., Shimizu, M., *et al*, 2017 and appears here with permission.**

BUSCO Type	No. of BUSCOs	% of BUSCOs
Complete Single-copy	900	94
Complete Duplicated	137	14
Fragmented	25	2.6
Missing	31	3.2

I predicted genes and repetitive elements, to generate the final TDr96_F1 reference genome sequence. To construct reliable gene models, I followed the MAKER pipeline using RNA-seq data from 18 samples representing different *D. rotundata* tissues and combined the data with publicly available ESTs and homologous protein sequences from related angiosperm species (Figure 3.1). This resulted in the prediction of 26,198 genes, of which 22,477 (85.8%) are supported by alignment of the RNA-seq data to the MAKER transcripts.

I then compared the white Guinea yam genome sequence metrics with those of *Arabidopsis thaliana* (eudicot), *Brachypodium distachyon* (monocot), and *Oryza sativa* (monocot) as these are considered to be good quality references for well chararcterised plant model species (Table 3.3). Remarkably, the GC contents of the total genome and exons of protein-coding genes in white Guinea yam were 35.8% and 44.1%, respectively. This result is close to that of *Arabidopsis*, and much lower than Poales species *Brachypodium* and *Oryza*. An average of 6.03 exons and 4.03 introns were annotated per gene. Roughly half of the genome was represented by interspersed sequence (274.5 Mb), a major component of which was long terminal repeat (LTR) sequences (135.7 Mb).

3.4.2 Comparative genomics and phylogentics

With the gene models established, I investigated the gene orthology of *D. rotundata* and the three other species mentioned previously. I identified 5,557 *D. rotundata* genes with a 1:1:1:1 orthologous relationship between the high-quality *B. distachyon*, *O. sativa*, and *A. thaliana* gene models (Figure 3.2). This number was reduced to 2,795 genes when I included Arecales (*Elaeis guineensis*, *P. dactylifera*) and Zingiberales (*M. acuminata*) in

Table 3.3: Characteristics of nuclear genome sequence in *D. rotundata* and other angiosperms. **This table has also been published in Tamiru, M., Natsume, S., Takagi, H., White, B., Yaegashi, H., Shimizu, M., *et al*, 2017 and appears here with permission.**

FeatureValue	***D. rotundata***(v0.1)	***A. thaliana***(TAIR10)	***B. distachyon***(v3.1)	***O. sativa***(v7_JGI 323)
Total length (Mbp)	594.23	119.67	271.16	374.47
GC (%)	35.83	36.06	46.4	43.57
Number of scaffolds (≥ 0 bp)	4723	7	10	14
Number of scaffolds (≥ 1000 bp)	4704	7	10	14
Largest scaffold (Mbp)	13.61	30.43	75.07	43.27
N50 (Mbp)	2.12	23.46	59.13	29.96
N75 (Mbp)	0.77	19.7	48.59	28.44
Number of Ns per 100 kb	282.45a	155.6	155.85	44.13
Ambiguous bases	1,413,029	–	–	–
Number of genes	26,198	27,416	34,310	42,189
Exons				
Number	158,059	141,044	154,104	178,353
Average number per gene	6.03	5.14	4.49	4.25
Total length (Mbp)	42.43	33.49	39.01	46.85
Average size (bp)	268.43	237.46	253.15	262.7
Average GC (%)	44.08	43.7	51.02	51.12
Introns				
Number	105,663	86,212	85,484	94,345
Average number per gene	4.03	3.14	2.49	2.25
Total length (Mbp)	83.12	17.87	47.7	53.34
Average size (bp)	630.33	157.25	398.18	391.23
Average GC (%)	32.37	32.45	38.29	37.2
Transposable elementsb				
% Total interspersed	46.07	13.32	37.39	44.4
Total interspersed total length (Mbp)	274.51	15.94	101.39	166.27
% Short interspersed nuclear elements (SINEs)	0.02	0.17	0.38	0.88
SINEs total length (Mbp)	0.13	0.2	1.02	3.31
% Long interspersed nuclear elements (LINEs)	2.43	1.07	2.91	1.29
LINEs total length (Mbp)	14.46	1.29	7.9	4.83
% Long terminal repeat (LTR) elements	22.82	6.35	19.31	21.09
LTR elements total length (Mbp)	135.71	7.61	52.36	78.98
% DNA elements	6.7	3.08	7.11	16.7
DNA elements total length (Mbp)	39.83	3.69	19.27	62.82
% Unclassified	14.2	2.64	7.68	4.36

[a]Number of Ns per 100 kb using the *D. rotundata* broken scaffolds. [b]Transposable elements were identified by masking the genomes using RepeatModeler and RepeatMasker, with the same parameters across all species.

my analysis.

Non-redundant gene orthology terms 'cell cycle', 'photosynthesis', 'plastid' and 'hydrolase activity, acting on ester bonds' were amongst those found to be significantly (padj < 0.05) enriched in these conserved genes (Table 7.11). Within this set of conserved genes, the most predominant were 186 orthologus groups containing genes with pentatricopeptide repeat (PPR) domains (Table 7.9). PPRs represent one of the largest gene families in terrestrial plants, with multiple genus specific expansions and contractions having been observed[277–279]. Despite their diversity and size, role of PPRs is not yet well document. However, in studies of *Arabdiopsis* and other module organisms, PPRs have been shown to be major mediators of organelle post-transcriptional control,

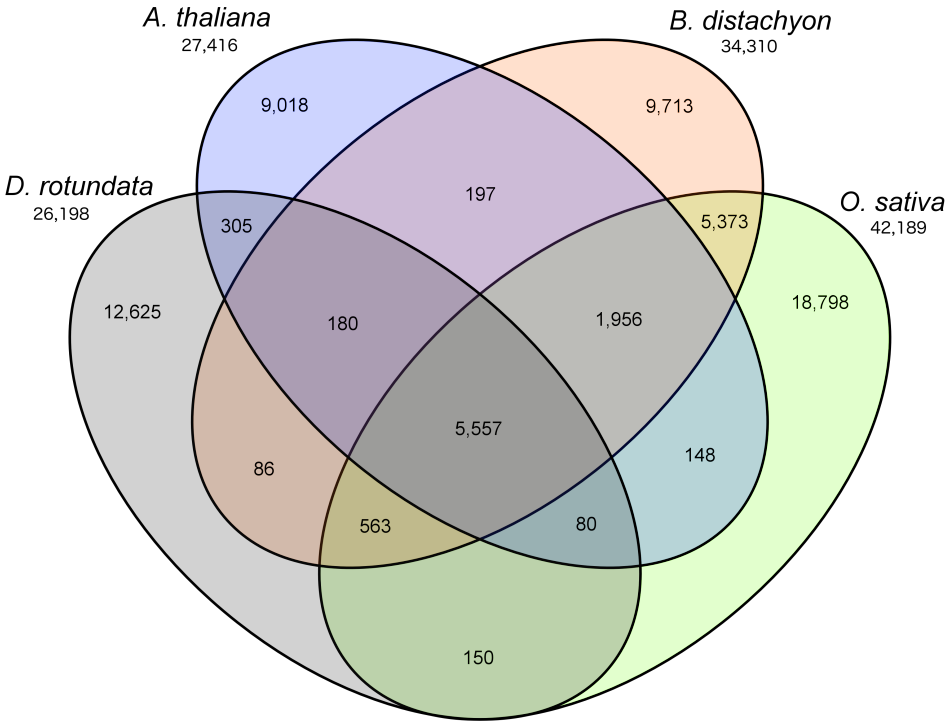

Figure 3.2: Venn diagram showing conserved and unique genes at 1:1 correspondence among *D. rotundata*, *A. thaliana*, *B. distachyon*, and *O. sativa*. Total gene counts in each genome are given below the species name. **This figure has also been published in Tamiru, M., Natsume, S., Takagi, H., White, B., Yaegashi, H., Shimizu, M., *et al*, 2017 and appears here with permission.**

serving a role in modulating the expression of RNAs in chloroplast and mitochondria[277]. Similarly to the conserved orthologs, the 869 orthologs present in all species but *D. rotundata*, also had PPRs (49) as their most represented group, in addition to eight heat shock proteins as the second most represented (Table 7.10).

Compared to orthologs shared between the other species, I obtained no evidence of orthologous relationships for 12,625 *D. rotundata* genes in *B. distachyon*, *O. sativa*, and *A. thaliana*, or for 11,348 *D. rotundata* genes when looking at all seven species (Table 7.12). Whilst this number of genes is considerably high, representing 48% of genes predicted in *D. rotundata*, similar numbers of species specific genes can be seen in each of the other species studied. This is potentially due to the large evolutionary distance between these species, as well as the completeness of gene models used. Partial gene models are unlikely to have high sequence similarly to their orthologous counterparts. As *D. rotundata* is thought to sit towards the base of the monocots, further studies including more closely related basal species would likely further reduce the number of these orphan

genes. This can be seen in the decrease from 12,625 to 11,348 genes with the additional of Arecale and Zingiberale species. The availability of the *D. rotundata* genome will also aid accurate identification of the number of genes specific to *D. rotundata* and the lineage as a whole, in future studies.

Of these 11,348 orphan genes, 3,260 were expressed in the tuber tissues; a tissue type not shared with the other species examined. Non-redundant gene ontology terms 'intracellular organelle', 'protein binding', and 'ion binding' were significantly (padj $<$ 0.05) enriched among *D. rotundata* genes that showed no orthology to the other species, but not among the conserved genes (Table 7.12). The *D. rotundata* genes without orthologs in the other species included 68 genes encoding proteins with lectin domains that maybe involved in defence against microbial pathogens, nematodes, and insects, accounting for 31% of the 216 lectin-coding genes functionally annotated in *D. rotundata*. Among the 12 subfamilies of lectins, the bulb-type lectin (snowdrop lectin; B-lectin) family contributed the largest share (110) of genes in *D. rotundata* (Table 3.4).

Table 3.4: Number of lectin class genes among four angiosperm species. **This table has also been published in Tamiru, M., Natsume, S., Takagi, H., White, B., Yaegashi, H., Shimizu, M., *et al*, 2017 and appears here with permission.**

FamilySpecies	***Brachypodium distachyon***	***Oryza sativa***	***Arabidopsis thaliana***	***Dioscorea rotundata***
B-lectin	93	122	53	110
Lectin-legB	54	80	54	40
Jacalin	24	32	50	5
Phloem	17	17	32	11
Lectin-C	1	2	1	1
Chitin-bind-1	16	12	10	7
Ricing-B-Lectin	18	11	3	6
Gal-lectin	11	20	18	10
Gal-Binding-Lectin	12	9	8	7
Calretriculin	15	9	9	7
EEA	2	0	0	2
LysM	17	15	12	10
Total	**309**	**267**	**199**	**216**

Phylogenetic analysis of the B-lectin genes in *D. rotundata* (110 genes; 51 unique), *B. distachyon*, *O. sativa*, and *A. thaliana* revealed two expansions of B-lectin genes in *D. rotundata* (Figure 3.3). The first expansion (blue band) consisted of 22 receptor-like serine threonine-protein kinases, which are thought to play a role in signalling and the activation of plant defence mechanisms[280]. The second expansion (red band) consisted of 28 mannose-binding lectins sharing high similarity with *Dioscorea batatas* tuber lectin DB1 (Accession number: AB178475). DB1 has been show to confer insecticidal properties against cotton bollworm (*Helicoverpa armigera*), and studies in transgenic tobacco and

rice plants expressing DBI demonstrated it also confers resistance against green-peach aphid and brown plant hopper, respectively[281–283]. Of these mannose-binding lectin genes in white Guinea yam, 16 did not have orthologs in any of the six other species, and eight showed enriched expression (Benjamini-Hochberg adjusted p-value [padj] $<$ 0.05) in tubers.

In the phylogentic tree, *D. rotundata* did not group with any species, including *Musa* of Zingiberales, *Phoenix* and *Elaeis* of Arecales, and *Oryza* and *Brachypodium* of Poales, suggesting that Dioscorea diverged independently from these taxa in the monocotyledons early on (Figure 3.4).

To investigate potential genome duplication in *D. rotundata*, I performed genome-wide syntenic dot plot analysis of *D. rotundata* against itself (Figure 3.5), which revealed no indication of genome duplication. Nevertheless, I observed 946 gene clusters composed of duplicated genes in *D. rotundata*. Of these, 145 duplicate gene clusters were observed only in *D. rotundata*.

To investigate macrosynteny between *D. rotundata* and related species, I carried out whole-genome syntenic dot plot analysis against the genomes of *Oryza sativa*, *Spirodela polyrhiza*, and *Phoenix dactylifera*. At the chromosomal level, it was difficult to observe synteny conservation between these species. However, I performed a syntenic path assembly of the scaffolds from these species against *D. rotundata*-masked pseudochromosomes (to align and orient them) and found a large proportion of the genomes to be conserved, suggesting that the *D. rotundata* genome has undergone many recombination events after divergence from the other species (Figure 3.6).

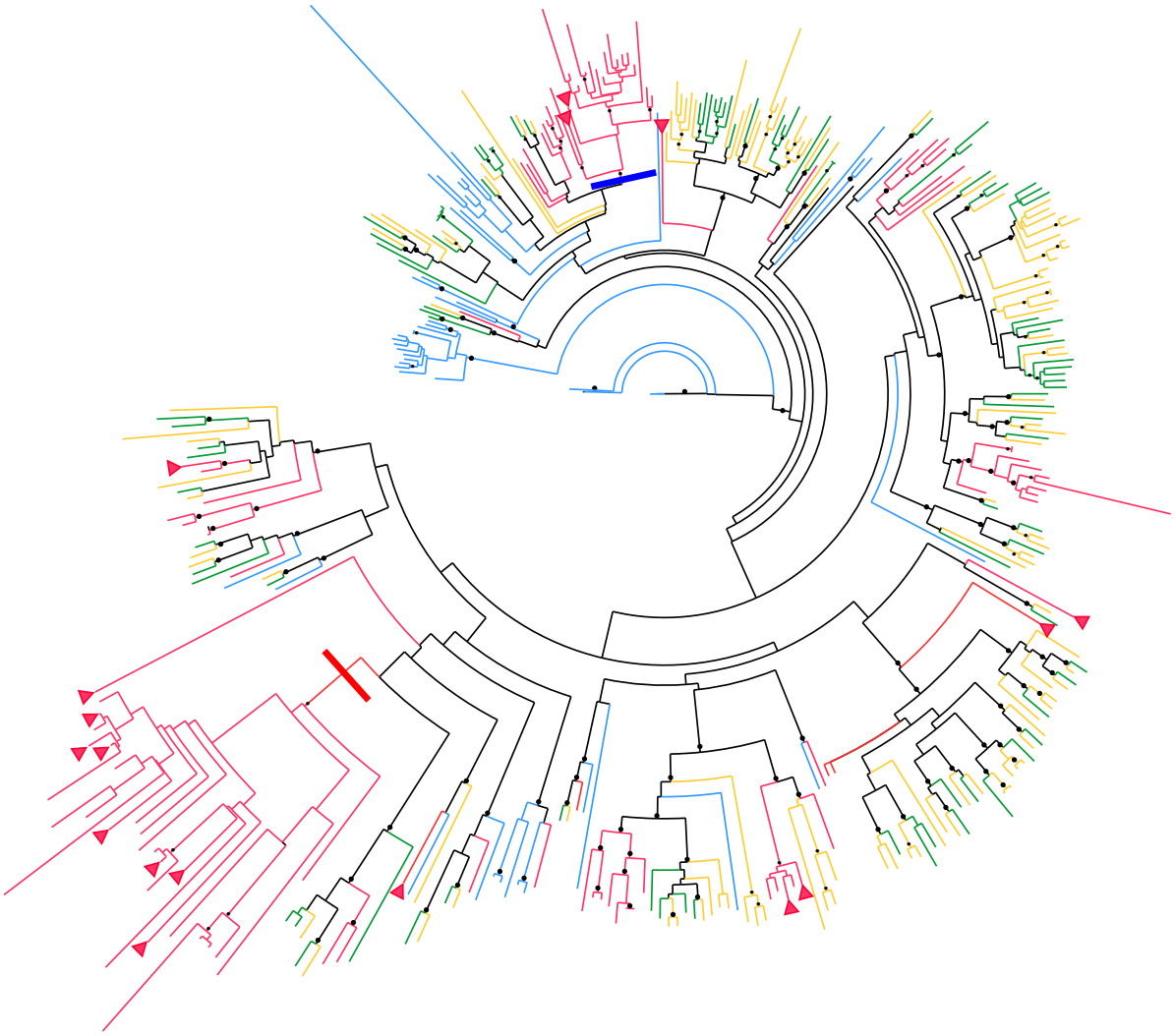

Figure 3.3: Phylogenetic analysis of the relationships of mannose-specific bulb-type lectin proteins in *D. rotundata* (red), *A. thaliana* (blue), *B. distachyon* (green), and *O. sativa* (orange). Arrowheads represent bulb-type lectins observed to have enriched expression in tubers. High confidence bootstrap values (1000 replicates) are represented at the nodes of the tree as dots. Thick red and blue lines show two root branches of -specific expanded genes. **This figure has also been published in Tamiru, M., Natsume, S., Takagi, H., White, B., Yaegashi, H., Shimizu, M., *et al*, 2017 and appears here with permission.**

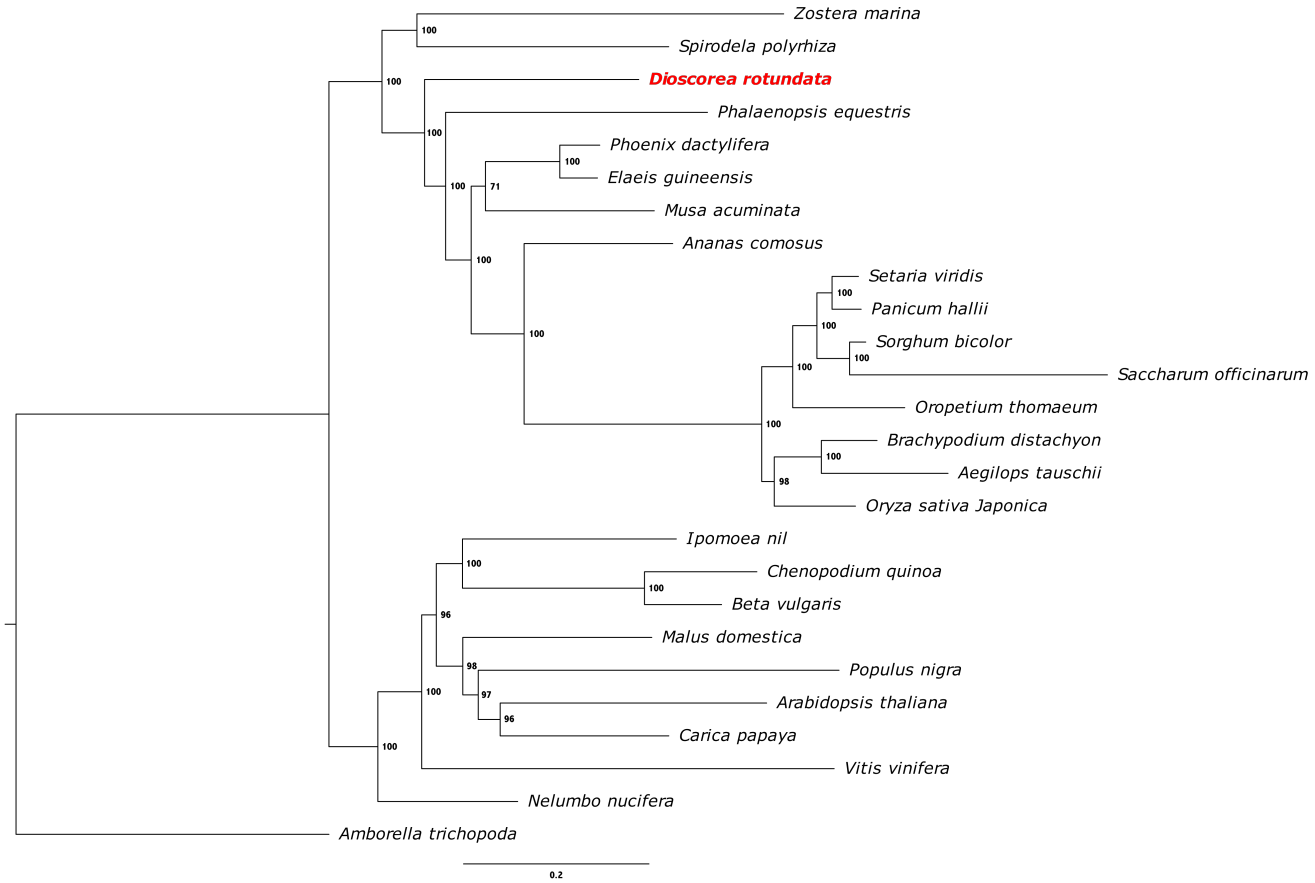

Figure 3.4: Maximum likelihood tree of 26 angiosperm species based on 190 orthologous protein-coding genes. The bootstrap values across 1000 resamplings are shown. The scale bar represents the mean number of substitutions per site.**This figure has been reproduced with persmission from Tamiru, M., Natsume, S., Takagi, H., White, B., Yaegashi, H., Shimizu, M., *et al*, 2017.**

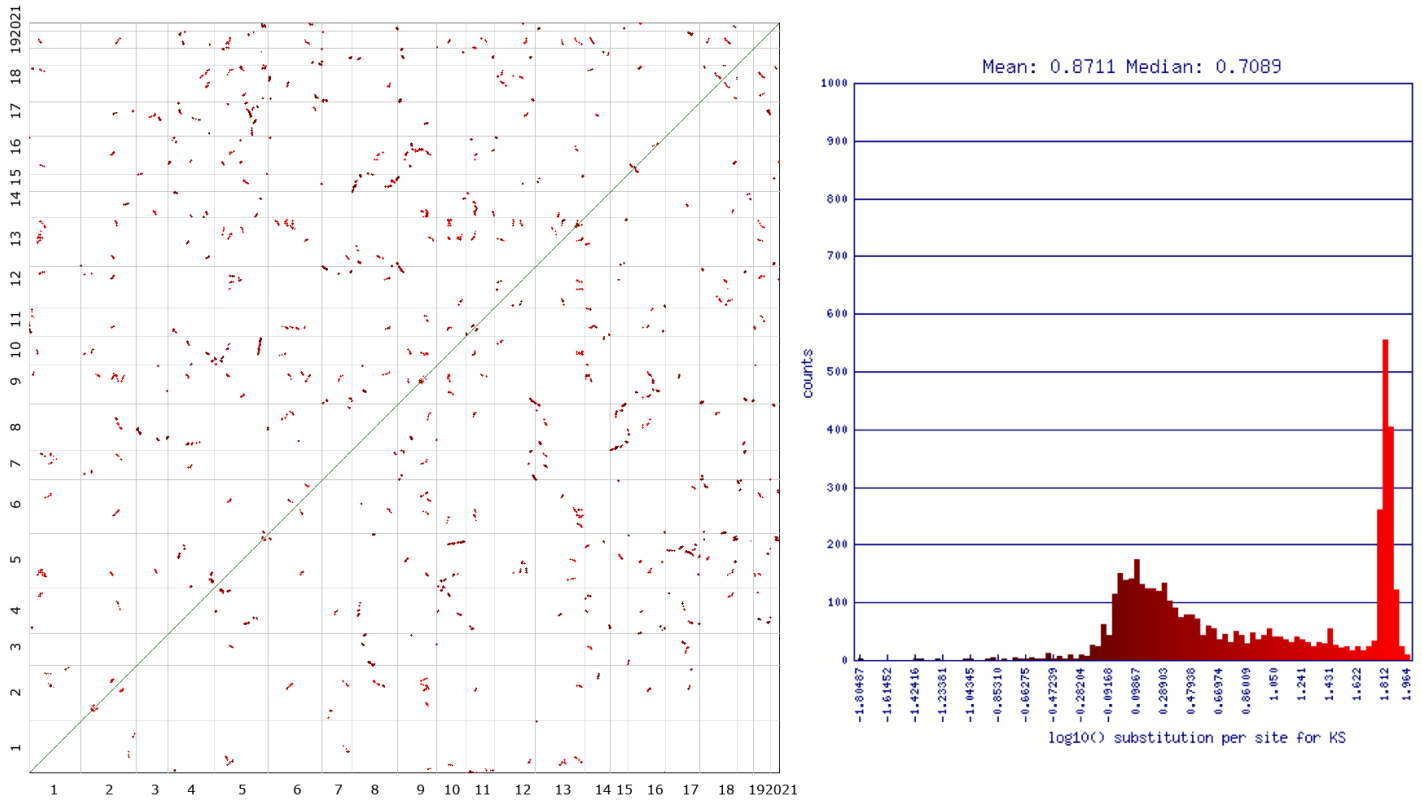

Figure 3.5: Self-self syntenic dotplot and synonymous substitution histogram of *D. rotundata* pseudo-chromosomes show no large scale genome duplication. Dotplot axis are labeled with pseudo-chromosome number. Syntelogs have been coloured based on their synonymous (KS) rate change. **This figure has also been published in Tamiru, M., Natsume, S., Takagi, H., White, B., Yaegashi, H., Shimizu, M., *et al*, 2017 and appears here with permission.**

3.4.3 Tissues specific gene expression

RNA-seq analysis of tuber tissues found 4,004 genes to have enriched expression in the tuber. The top 50 highly expressed genes included genes encoding starch synthases and branching enzymes, as well as three carbonic anhydrase-coding genes. BLASTP (https://blast.ncbi.nlm.nih.gov) analysis showed that these carbonic anhydrase-coding genes shared high identity (avg. 76%) with genes encoding *Dioscorea japonica* dioscorin precursors; this tuber storage protein has carbonic anhydrase activity and multiple isoforms[284].

Of the genes mapped to the genome, significant (padj < 0.05) up regulation was observed in 2,038 (9.9%) of genes, and down regulation in 2,820 (14%), when tuber expression was compared to the other tissues (Figure 3.7). The top 10 genes with the highest change in significant log2 fold change (LFC), expected down regulation was observed in Chlorophyll A-B binding protein (Dr07513), a light-harvesting complex, and up-regulation in Mycolic acid cyclopropane synthase (Dr16361) involved in lipid biosynthesis (Table 7.13). Conversely, when looking at the significant (padj < 0.05) LFC in flower and related tissues, 1,577 (7.7%) genes were observed to be up-regulated and 1,671 (8.1%) down regulated, compared to the other tissues (Figure 3.8). Of the 10 genes with the highest significant change in LFC, three are transcription factors (Dr08673, Dr09858 and Dr11587) and a Phosphatidylethanolamine-binding protein (Dr04574), a class of that's been shown to be involved in regulating development of inflorescence (Table 7.14)[285].

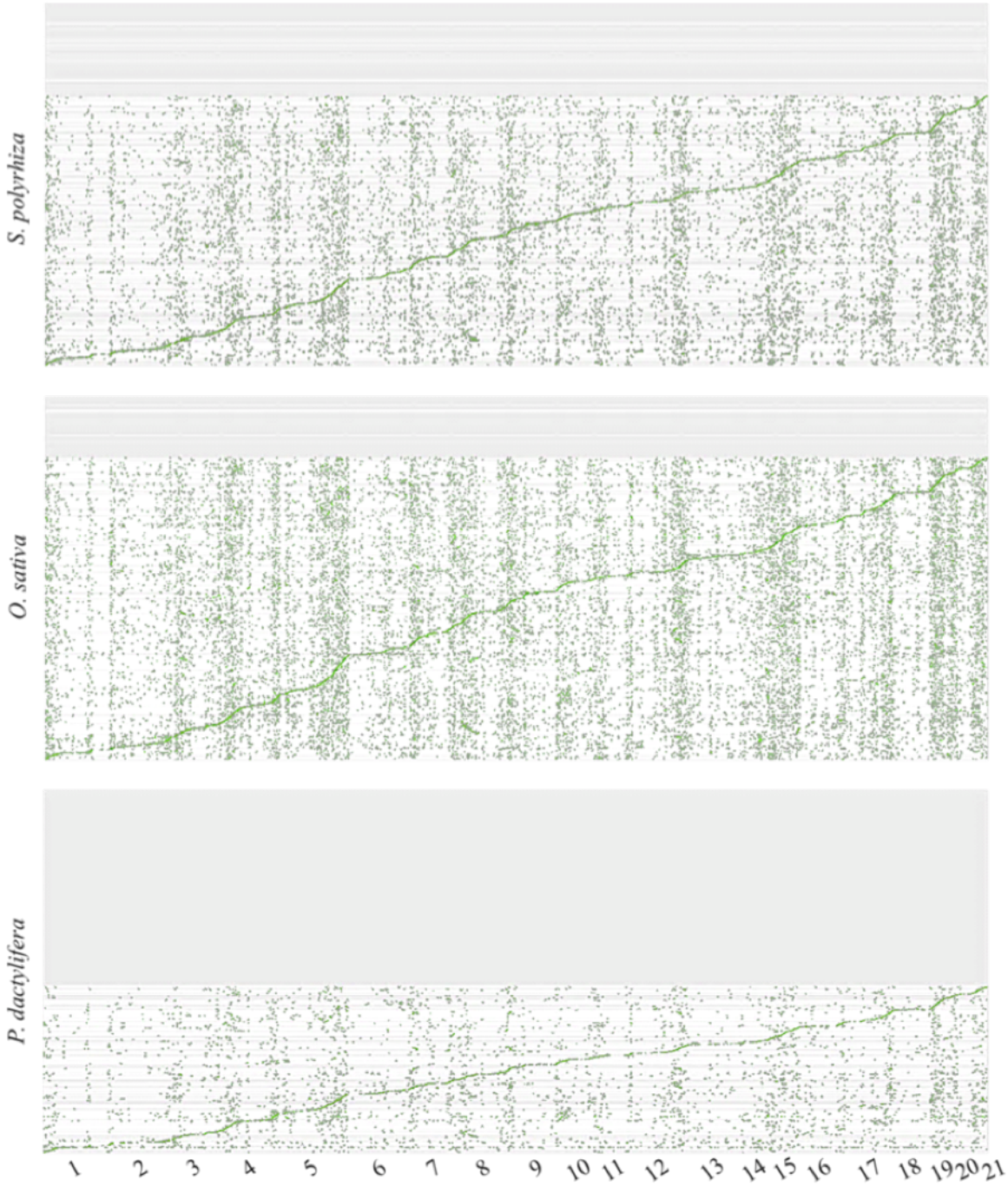

Figure 3.6: SyMAP dotplot analysis of whole genome synteny between scaffolds of three monocot species: *S. polyrhiza*, *O. sativa* and *P. dactylifera*, and *D. roundata* pseudo-chromosomes. Scaffolds were aligned and orientated to *D. rotundata* pseudo-chromosomes. Dots represent regions of sequence similarity between the two genomes, clustering of dots into horizontal lines indicates shared syntenic or orthologous blocks derived from a common ancestor. Scaffolds with no synteny are represented by the grey regions at the top of the dotplots. **This figure has also been published in Tamiru, M., Natsume, S., Takagi, H., White, B., Yaegashi, H., Shimizu, M., *et al*, 2017 and appears here with permission.**

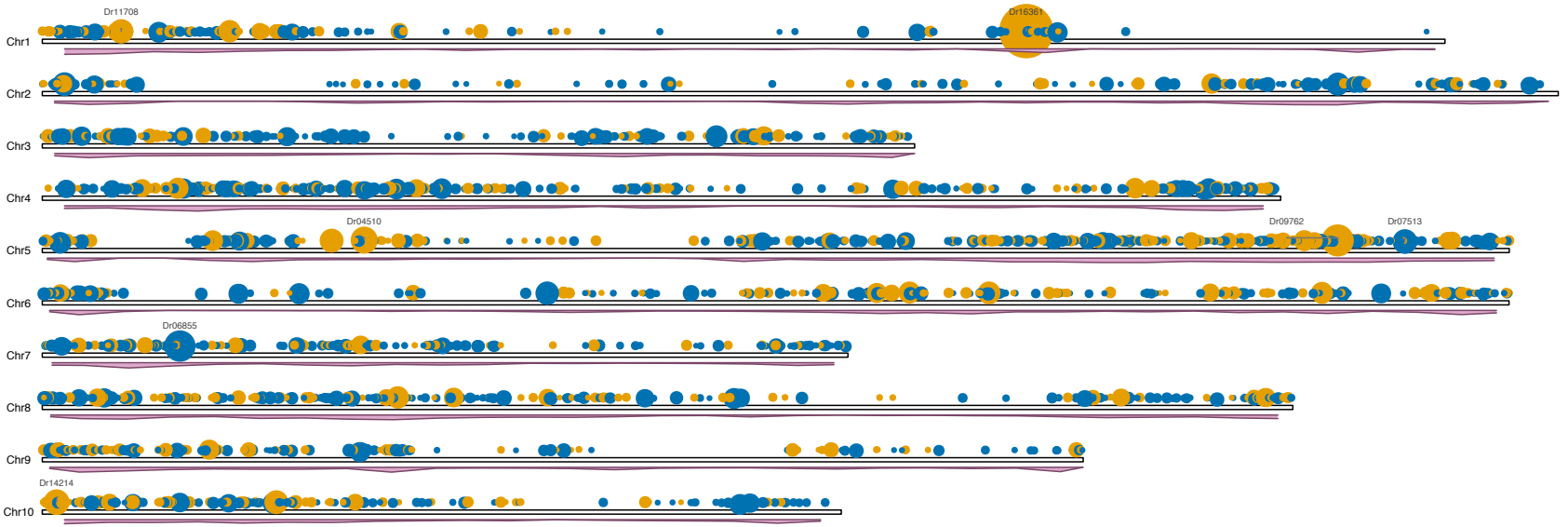

Figure 3.7: Visualisation of tuber tissues gene expression across pseudo-chromosomes. Each circle above the pseudo-chromosome ideogram represents an expressed gene, with the size of the circle indicating the padj significance of the LFC in expression compared to the other tissue; a bigger circle meaning greater significance. Colours of each circle represent a LFC < 0 (blue) or LFC > 0 (yellow). The top 10 genes with greatest increase or decrease in LFC compared to the other tissues are labeled above their corresponding position. Below each ideogram is a plot of gene density across the pseudo-chromosome.

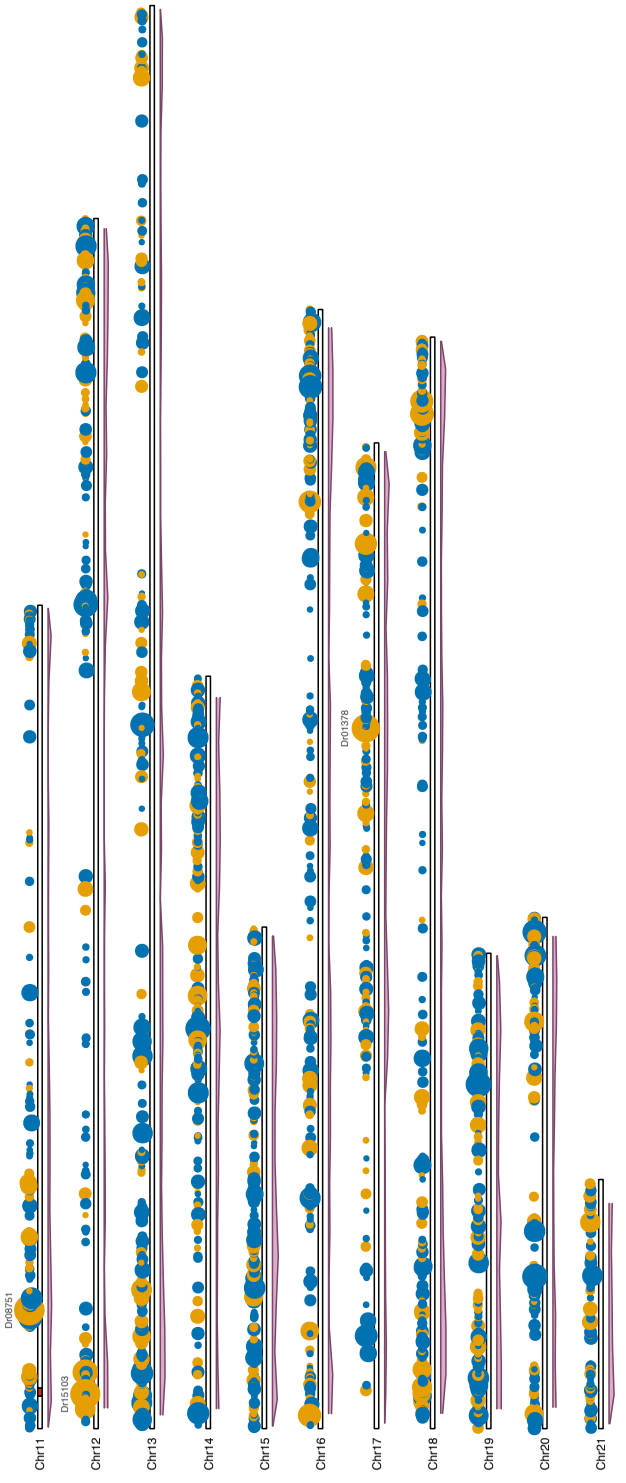

Figure 3.7: Continued.

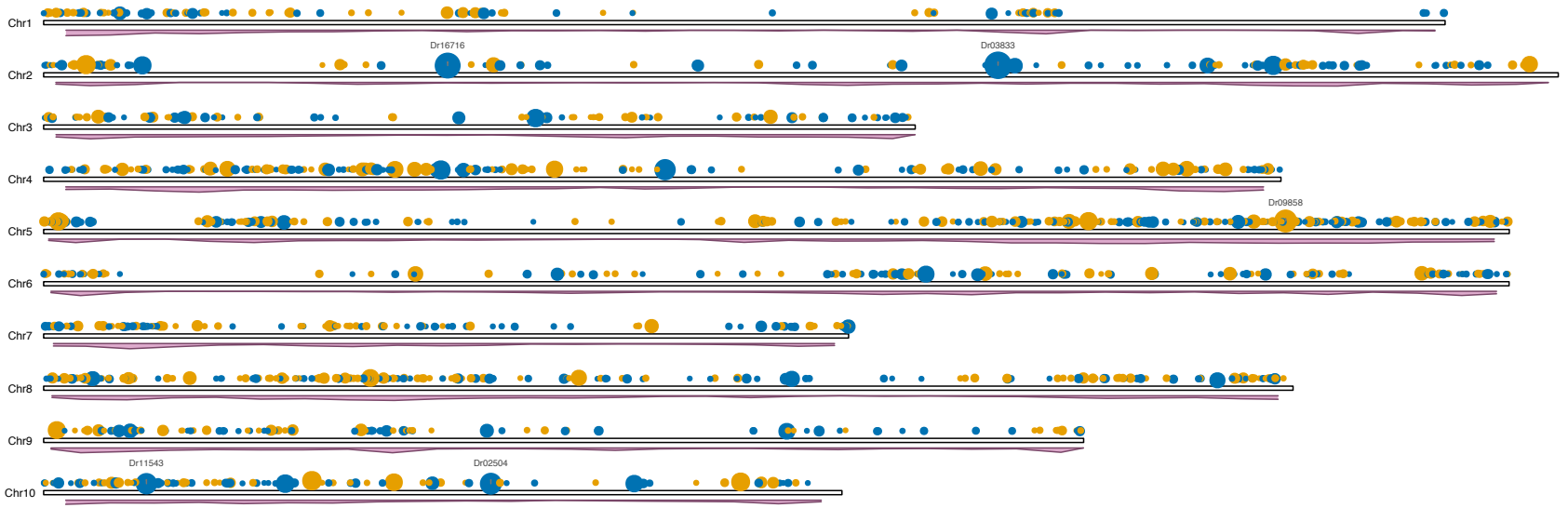

Figure 3.8: Visualisation of flower and related tissues gene expression across pseudo-chromosomes. Each circle above the pseudo-chromosome ideogram represents an expressed gene, with the size of the circle indicating the padj significance of the LFC in expression compared to the other tissue; a bigger circle meaning greater padj significance. Colours of each circle represent a LFC < 0 (blue) or LFC > 0 (yellow). Below each ideogram is a plot of gene density across the pseudo-chromosome.

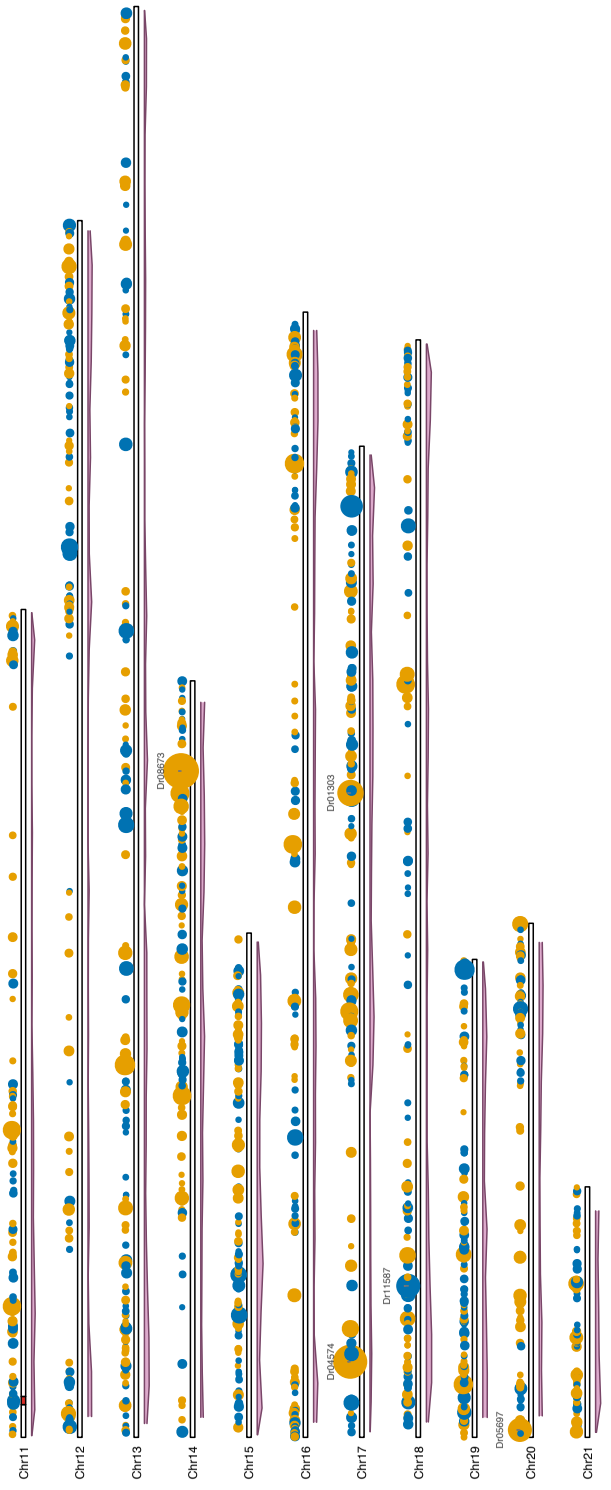

Figure 3.8: Continued.

3.5 Discussion

White Guinea yam sustains the lives of at least 60 million people in West and Central Africa. Despite this and the vital role it plays in the socio-cultural life of the society for centuries, genetic improvement of the crop has remained slow. This is partly due to our limited knowledge of its genetics and lack of modern genetic and genomic tools to support breeding programs. The reference genome I analysed with our collaborators should aid in bridging this gap and in bringing this orphan crop species into the genomics age. A large number of molecular markers such as SSRs, INDELs and SNPs can now be developed for various applications including, but not limited to, diversity analysis, linkage mapping, genome-wide association analysis, genomic selection, and MAS. Additionally, the reference for white Guinea yam is now publicly available as part of the Ensembl Genomes platform [187]. Phylogenetic analysis of conserved genes revealed that *Dioscorea* does not form a clade with other monocotyledonous species belonging to Alismatales, Asparagales, Poales, Arelales and Zinziberales, and the eudicots, suggesting that *Dioscorea* has diverged earlier from other sequenced monocotyledonous species. Not only this, but comparative analysis has highlighted a number of novel expansions in white Guinea yam. A lineage-specific expansion of genes encoding lectin and LRR proteins was predicted and these may be related to the defence of its starchy storage organ against microbe pathogens, nematodes and insects that attack yam tuber; a suitable source of nutrition not only for humans but also for other organisms. One of the most unique opportunities, and perhaps one of the most important aspects of crop improvement in white Guinea yam, is the study of dioecy. Through use of the genome assembly and annotation, our collaborators predict that the white Guinea yam is female heterogametic (male=ZZ, female=ZW). Identifying a sex-linked DNA marker (sp16) located within the putative female-specific region of the W chromosome (FSW), spanning only 161 kb of pseudo-chromosome 11 (Figure 3.9).

While I could not investigate differential expression, due to lack of the minimal number (triplicate) of biological repeats to infer statistical significance, I did observe that none of the genes with the highest tissue specific expression were in the FSW in the tuber or flower tissues (Figure 3.10, 3.11). However, Dr08751, located 1.6 Mb upstream of this on pseudo-chromosome 11, was found to be one of the third most overexpressed gene in tuber compared to other tissues. This gene lacks an InterPro functional annotation, but has GO terms associated with mitochondrion (GO:0005739) and transferase activity (GO:0016772). Further work on sex specific expression in the FSW would benefit from a time course study of male and female inflorescence at different

point of development, to better understand the regulation of genes involved in sex differentiation. In addition, comparison of the intergenic and intragenic gene space of FSW with other related dioecious species would assist with improving our understanding of the evolutionary history of sex determination in *Dioscorea*. Especially given that dioecy is the norm in *Dioscorea* species and available reports suggest that the male is usually male heterogametic sex (XY) in the genus and plants in general, as with related species, *D. tokoro*[23]. As such, the findings presented here indicate a possible transition of sex determination system, potentially caused by strong genetic conflicts among Z, W and organelle genome[29]. This is an exciting topic for further investigation and is explored in the next chapter of my thesis.

The availability of a reference genome for *D. rotundata* facilitates the application of NGS-based technologies for accelerating plant breeding and will contribute a solid basis for understanding the origin of white Guinea yam from its wild progenitor species widely distributed in West and Central Africa. The list of genes potentially linked to sex determination, that only occur in female individuals, will form a basis of future study of molecular mechanism of white Guinea yam sex determination (Table 3.5). Overall, the *D. rotundata* genome sequence represents an important milestone towards significantly increasing the role yam plays in ensuring food security for resource-poor households in Africa and beyond.

Table 3.5: List of genes predicted within the female specific (W-linked) region on pseudo-chromosome 11 identified by QTL-seq. **This table has also been published in Tamiru, M., Natsume, S., Takagi, H., White, B., Yaegashi, H., Shimizu, M., *et al*, 2017 and appears here with permission.**

Gene ID	mRNA Size	Interproscan IPR Description
Dr18978	1,318	NULL
Dr18979	680	Peptidase C48, SUMO/Sentrin/Ubl1
Dr18980	1,768	NULL
Dr18981	9,645	NULL
Dr18982	595	Nucleoporin Nup186/Nup192/Nup205
Dr18983	8,930	NULL
Dr18984	690	Exostosin-like
Dr18985	11,675	Cation/H+ exchanger
Dr18986	2,497	Cyclophilin-like peptidyl-prolyl cis-trans isomerase
Dr18987	630	Pentatricopeptide repeat

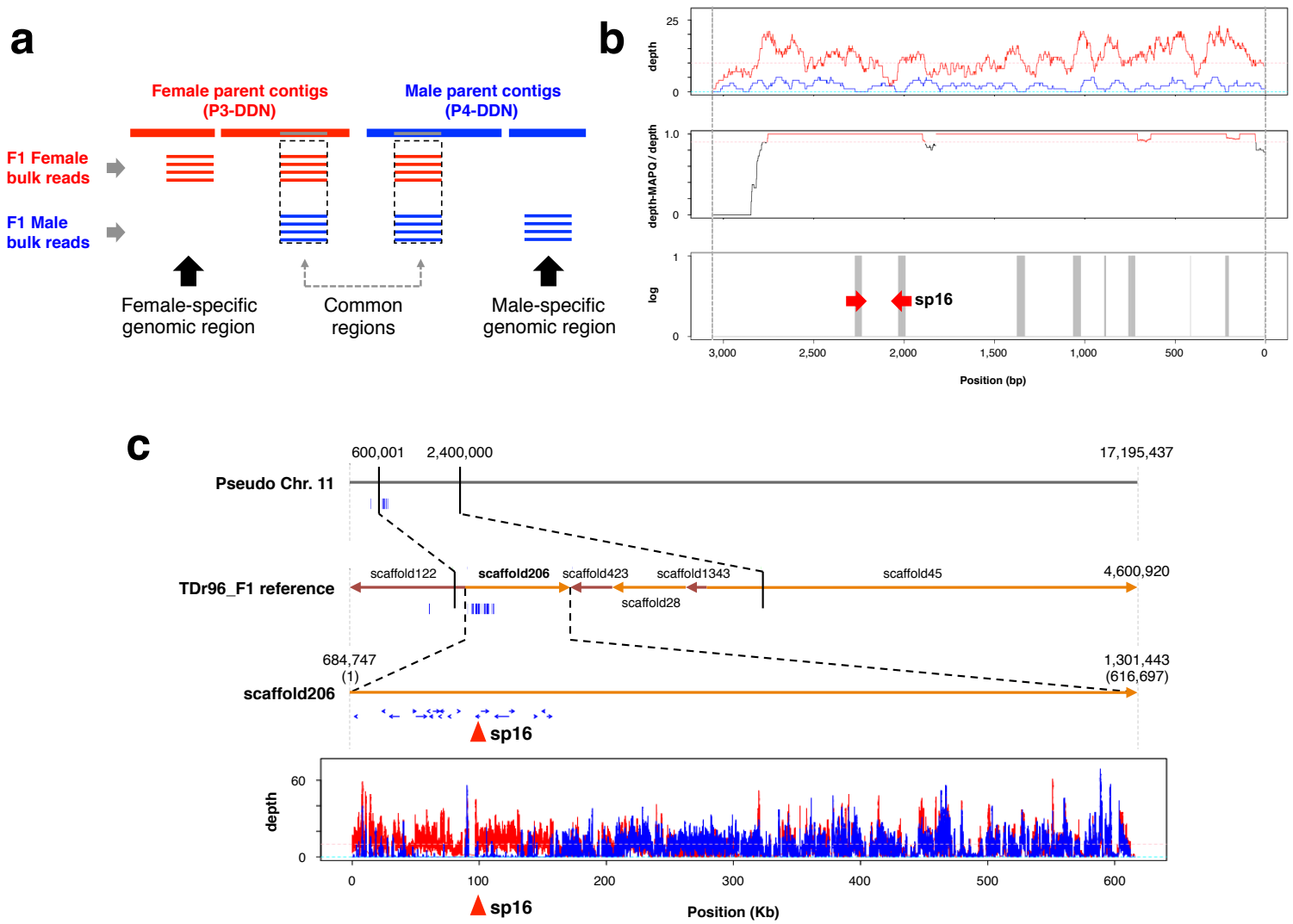
a
Female parent contigs
(P3-DDN)
Male parent contigs
(P4-DDN)
F1 Female
bulk reads
F1 Male
bulk reads
Female-specific
genomic region
Common
regions
Male-specific
genomic region
b
depth
depth-MAPQ / depth
log
sp16
Position (bp)
c
Pseudo Chr. 11
600,001
2,400,000
17,195,437
TDr96_F1 reference
scaffold122
scaffold206
scaffold423
scaffold1343
scaffold45
scaffold28
4,600,920
684,747
(1)
1,301,443
(616,697)
scaffold206
sp16
depth
Position (Kb)

Figure 3.9: Identification of FSW. **a.** Overview of the method used by our collaborators to identify the W-linked region, through *de novo* assembly of two male and female parents. Followed by mapping of bulked reads of their male and female progeny, that aligned to either male, female or common regions of the assembly. Female parental contigs that were mapped only with reads belonging to the F1 female bulk corresponded to FSW. Sequence reads mapped to such positions were identified by their high MAPQ scores (=60). **b.** An example of a female-specific contig (contig Female917_flattened_line_87512_3057). Alignment depths of F1 female bulk (red) and F1 male bulk (blue) are shown (top). Frequency of reads mapped with MAPQ score = 60. The red line corresponds to genomic regions that were covered by short reads, > 90% of which had a MAPQ score of 60 (middle). A genomic region that is covered only by female reads (not by male reads) and > 90% of mapped reads had MAPQ score = 60 (indicated by gray bars) (bottom). Red arrowheads indicate the positions of PCR primers for the DNA marker sp16. **c.** Location of the FSW region. Thick gray horizontal line denotes pseudo-chromosome 11 (top), scaffolds on chromosome 11 (middle), and scaffold206 (bottom). The thin blue lines shown under the first, second, and third horizontal lines indicate the positions of female contigs (P3-DDN) specifically mapped by F1 female bulk reads. The square box at the bottom indicates alignment depth of reads of F1 female bulk (red) and F1 bulk of progeny in which sp16 amplification was not observed (sp16-minus) (blue) to scaffold206. Red triangles indicate the position of DNA marker sp16. **This figure has also been published in Tamiru, M., Natsume, S., Takagi, H., White, B., Yaegashi, H., Shimizu, M., *et al*, 2017 and appears here with permission.**

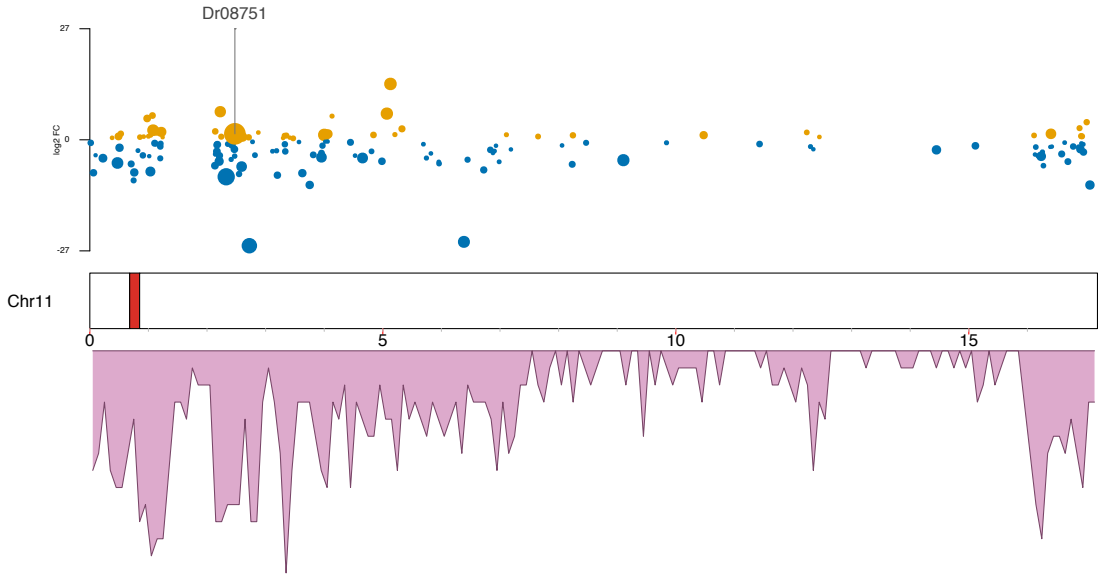

Figure 3.10: Visualisation of tuber tissues gene expression across pseudo-chromosome 11 (Mb). Each circle above the pseudo-chromosome ideogram represents an expressed gene, with the size of the circle indicating the padj significance of the LFC in expression compared to the other tissue; a bigger circle meaning greater padj significance. Colours of each circle represent a LFC < 0 (blue) or LFC > 0 (yellow). Below each ideogram is a plot of gene density across the pseudo-chromosome.

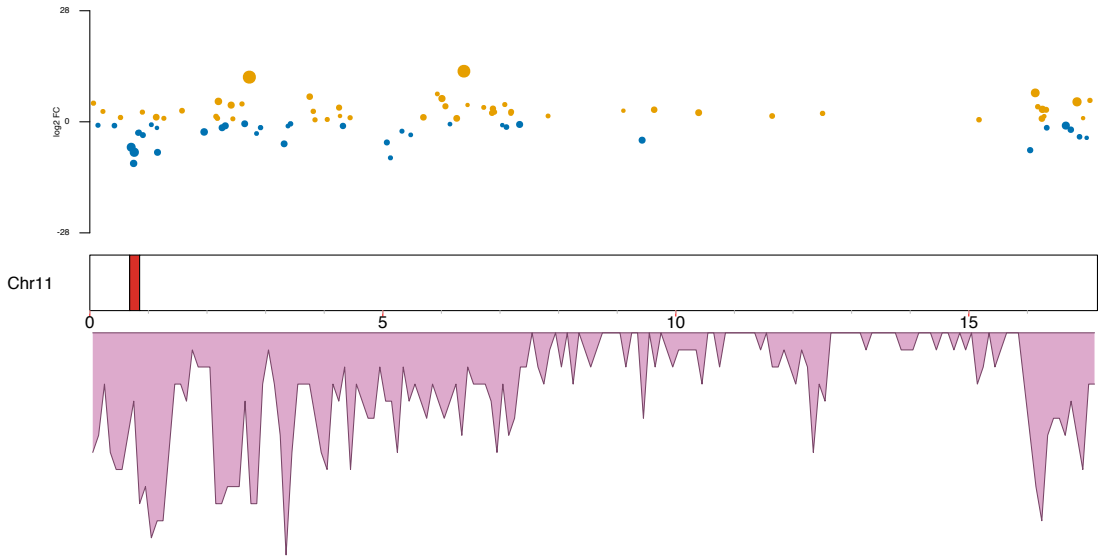

Figure 3.11: Visualisation of flower tissues gene expression across pseudo-chromosome 11 (Mb). Each circle above the pseudo-chromosome ideogram represents an expressed gene, with the size of the circle indicating the padj significance of the LFC in expression compared to the other tissue; a bigger circle meaning greater padj significance. Colours of each circle represent a LFC < 0 (blue) or LFC > 0 (yellow). Below each ideogram is a plot of gene density across the pseudo-chromosome.

Chapter 4

Exploration of the Oni-dokoro genome and evolution of sex in *Dioscorea*

The work contained in this chapter on *D. tokoro* was carried out as part of an on-going collaboration between myself and the Iwate Biotechnology Research Centre, which we plan to submit for publication by the end of 2018. As part of this work, I have carried out assesment of the assembly, whole genome annotation and comparative genomics in *D. tokoro*. Work relating to *D. alata* was part of a now finalised collaboration with the International Institute of Tropical Agriculture, in which I carried out the assembly, whole genome annotation and assesment of *D. alata*. Finally, all work related to the evolution of sex determination in *Dioscorea* are my own original hypotheses, outside of these collaborative efforts.

4.1 Abstract

The monocotyledon genus Dioscoreaceae, with over 600 species, is composed of 10 major clades with species belonging to branching clade B being the most studied and important economically. Dioecy, the presence of separate male and female plants, is present throughout the genus. Our previous study on *Dioscorea rotundata*, belonging to the Enantiophyllum clade of the Old World lineages belonging to branching clade B, identified a locus associated with female heterogametic (ZW) sex. As dioecy is most likely to evolve through gynodioecy, rather than androdioecy, and the XY sex chromosome

system is more prevalent in plants and throughout *Dioscorea*, it is exciting to observe ZW sex determination in *D. rotundata*. Therefore, to further investigate the evolutionary history of dioecy in *Dioscorea*, with our collaborators at IBRC we produced a reference genome for the previously reported XY species and early branching clade A species, *D. tokoro*. I compared gene orthology and phylogeny across 26 angiosperm species and confirmed the phylogenetic position of *Dioscorea*, sitting towards the base on the Monocotyledons as has been previously reported.

I then generated a first pass reference for another related species, *D. alata*, that has ill-defined sex determination but belongs to the same Enantiophyllum clade as *D. rotundata*. My findings also confirm the ancestor of *D. tokoro* to have likely split from *D. rotundata* and *D. alata* early on, forming the basal Stenophora clade. Comparison of the sex determination regions of *D. rotundata* and *D. tokoro* with *D. alata* show unique evolutionary history between the species. Further investigations with other *Dioscorea* species will improve our understanding of sex in the genus and the evolutionary mechanisms of sex determination as a whole.

4.2 Introduction

Our previous collaborative work on *D. rotundata*, one of the most cultivated yam species, found that sex determination in this species is most likely female heterogametic (male=ZZ, female=ZW) and identified a sex-linked DNA marker to the FSW, a 161 kb loci on pseudochromosome 11, that can distinguish genotypically male and female individuals[32, 286]. This marker can be used for sexing of *D. rotundata* at the seedling stage, which will accelerate breeding programs for improvement of this important staple crop. Additionally, the first reference sequence publicly available for this clade is now widely accessible via the Ensembl Plants platform[187].

The ZW sex system of *D. rotundata* is particularly interesting as the related species, Oni-dokoro (*D. tokoro*), has been observed to be male heterogametic (male=XY, female=XX), indicating transition of sex determination in the genus[64, 287]. The name Oni-dokoro comes from the Japanese for demon (oni) and is used to express the contrast of *D. tokoro's* wide and round leaves compared to related species Hime-dokoro (*D. tenuipes*), where 'hime' means Princess, that has long slender leaves (personal communication; S. Natsume, IBRC). Native to East Asia, the rhizome of wild *D. tokoro* has traditionally been consumed in northern Japan (specifically, Iwate prefecture), is used in traditional medicine and is thought to contain compounds of potential pharmaceutical

importance[238, 288, 289]. Additionally, this species has a fast growth rate and is easy to cross, making it an ideal species for study[64].

Another important dioecious yam species, *D. alata* (commonly known as purple yam) is one of the most widely distributed and cultivated yams. This species gets its name from the purple-fleshed tubers associated with it; however white tubers are also produced by some cultivars. This pigmentation is a result of high levels of anthocyanin, which are thought to play a role in defence and to also have potential health benefits, increasing the market value of purple tubers from this species[290]. Thought to belong to the same Enantiophyllum clade of the clade B Old World lineages as *D. rotundata*, there has been little study to date on the genome of *D. alata* and investigation of sex determination in the species has thus far been limited to a small number of cytological studies, of which early studies reported X0 sex determination. However, newer studies now dispute this finding[63, 291–294]. Furthermore, this clade contains several other yam species, such as *D. cayenensis*, that along with *D. rotundata* and *D. alata* make up the majority of global yam production[291]. While *D. tokoro* belongs to the predominantly subtropical and rhizomatous early branching A clade lineage, Stenophora, having split early in the Eocene period, ~48.2 Mya[295].

Dioecy is thought to most likely evolve through gynodioecy, rather than androdioecy, and the XY sex chromosome system is more prevalent in angiosperms[23]. It is therefore exciting to observe ZW sex determination in *D. rotundata*, especially with another XY sex determination system present in the same genus[23]. In mammals and other well characterised models, sex is controlled by a master regulator gene, such as *Sry* in mammals, however master regulators have been considered unlikely in angiosperms, as these are thought to require at least two regulators for maleness and femaleness[29, 32, 296]. Angiosperms have been shown to have a wide diversity of environmental and genetic sex determination mechanisms, similar to that seen in reptiles and fish[29, 297–299]. Despite this, there has been little study to date comparing plant species in the same genus with different sex determination mechanisms.

By studying the genomes of Enantiophyllum (*D. rotundata* and *D. alata*) and Stenophora (*D. tokoro*) species, we can begin to explore the ancestral state of sex in the clade Enantiophyllum and produce hypotheses on the drivers behind the differentiation of different sex determination mechanisms in this genus as a whole. Here I present a draft reference genome for *D. tokoro* and compare the proto sex chromosomes of this species with *D. rotundata*, as part of an on-going collaborative effort with the IBRC, Japan. Later, I then compare both species with a first pass assembly of *D. alata*, as part

of a collaboration with the IITA, Nigeria, as a starting point to unraveling the unique evolutionary history of sex determination within this genus.

4.3 Methods

4.3.1 Genome assessment and repeat annotation

Our collaborators at IBRC generated the whole genome assembly of a single *D. tokoro* phenotypically male individual using the ALLPATHS-LG[122] workflow, with 250 bp paired-end reads, and 2, 4, 6 kb mate-pair reads and 20 kb mate pair libraries, sequenced on the Illumina MiSeq. To evaluate completeness the assembly was checked for the presence of 1,440 Benchmarking Universal Single-Copy Orthologs (BUSCO), using the embryophyta_odb9 dataset and BUSCO version 3.0[139]. I also reran BUSCO with the same settings on *D. rotundata*, for direct comparison of the two genome using the updated (since the previous work on *D. rotundata*) embryophyta_odb9 dataset.

Repetitive sequences, including transposons, were predicted using a combination of RepeatModeleler-1.0.8[141], TransposonPSI-08222010 (http://transposonpsi.sourceforge.net), and LTRharvest[300] and LTRdigest[301], both part of the Genometools-1.5.9[302] package. Repeats were first modelled with RepeatModeler and then masked with RepeatMasker-4.0.7[144] using the National Centre for Biotechnology Information (NCBI) database, one of three other options was used to generate interspersed RepeatModeler-based, interspersed Rebase-based, and Low complexity repeats: “nolow”, “nolow, species Viridiplantae”, and “noint”, respectively. TransposonPSI was ran to further identify transposon open reading frames with Blast-2.6.0[303] and only sequences longer than 30 bp were retained. LTRharvester was used on the assembly to find LTR retrotransposons and then LTRdigest was ran using Hmmer-3.1b2[304], and the complete set hmm files from the Gypsy Database Collection[305] of mobile genetic elements to annotate features of LTRs found. These were then filtered for LTR candidates that didn’t have domain hits and to extract full-length elements. All outputs were combined and USEARCH-9.2.64[306] was used to cluster 22,953 identified repeat sequences into 13,257 clusters and remove redundant sequences to generate a final library of repetitive elements with a minimum of 80% sequence identity. Repeats in the combined repeat library were finally classified using the RepeatModeler tool, RepeatClassifier.

4.3.2 Prediction of protein-coding genes

An initial set of gene models was first generated with the MAKER-2.31.6[246] pipeline, using the assembled scaffolds. First, *ab inito* evidence-based prediction was performed by S. Natsume, IBRC, with AUGUSTUS[145], using the previously generated *Dioscorea rotundata* training set. I then used paired-end 75 bp RNA-seq reads from mature bud (male and female), flower (male and female), leaf, stem, reproductive shoot apex at three different developmental stages (male and female), rhizome (bud, root, stem and storage organ), root apex, vegetative shoot for reference guided assembly of transcripts using Bowtie-1.1.1[247], Trinity-2.0.6[149] and SAMtools-1.2.0[113], and TopHat-2.1.0[244]. Publicly available EST(s) and/or CDS of *Dioscorea alata*[78], *D. rotundata*[307], and several other *Dioscorea* species obtained from NCBI, *Spirodela polyrhiza*[273], *Oryza sativa* Japonica[308], were provided as 'alternative EST' evidence to MAKER. Protein sequences from *D. rotundata*[307] and a set of 465 Reviewed/57,375 Unreviewed, UniProt[262] non-fragmented proteins classed under Petrosaviidae taxonomy and not Poales, were further included as alternative protein evidence to MAKER. Finally, the combined repeat library was also included in the MAKER run to guide repeatmasking of the genome. The MAKER[246] annotation pipeline produced 19,275 gene models.

EvidenceModeler-r20120625_patch_v0.1[162] (EVM) was then used to build upon and improve this initial set of gene models, using the following additional inputs. StringTie[309] assembled transcripts generated by, S. Natsume, IBRC, using RNA-seq reads of all tissues and genome alignment with HISAT2[310]. In addition to these, I generated a *de novo* Trinity assembly of all RNA-seq data was combined with the reference guided assembly to build a comprehensive transcript database using Program to Assemble Spliced Alignments (PASA)[150]. High-quality non-redundant transcripts from PASA were then used to generate a training set for AUGUSTUS 3.1. As the *D. tokoro* transcripts only produced a relatively small amount of training data, the *D. rotundata* training set was incorporated into this following the 'artifical monster gene trick' outlined in the AUGUSTUS manual, to produce a set of *ab initio* gene predictions and the hint file for the EVM annotation. An additional set of *ab initio* predictions were generated with GeneMark-ES-4.33[311]. Protein sequences from *D. rotundata* and *S. polyrhiza* were combined with a set of all protein sequences under the taxonomy '*Lilipodia*' from UniProt (1,674,275 sequences), and these were aligned to the genome using Exonerate 2.2.0 with --score 500. Alignments were then filtered for a minimum of 70% coverage. In addition to this, 132,336 transcripts from *D. rotundata*[307], *D. alata*[78], *S. polyrhiza*[273],*O. sativa*[312], and *D. japonica* and *D. nipponica* obtained from the National Centre for

Biotechnology Information, were used in the MAKER run were also aligned to the genome and filtered using the same method.

EVM was then ran on the repeatmasked genome using the following weights and inputs: ABINITIO_PREDICTION MAKER gene models 5, ABINITIO_PREDICTION AUGUSTUS hint file 3, ABINITIO_PREDICTION GeneMark hint file 1, TRANSCRIPT StringTie transcripts 7, TRANSCRIPT Comprehensive transcripts 7, TRANSCRIPT Exonerate EST matches 1, and PROTEIN Exonerate protein sequence matches 5. Finally, PASA was ran three times on the EVM gene models, using the comprehensive transcript database, to add UTR annotations, correct consensus predictions and add alternative spliced isoforms. Producing a final set of 29,471 gene models and 31,283 alternative isoform transcripts.

Functional annotation of the amino acid sequences was performed using the in-house pipeline, AnnotF, which compares Blast2GO[166] and [165] functional terms. An additional round of manual functional annotation was carried out using Diamond-0.9.18[313] to blastp, with default settings and a minimum coverage of 80%, all available UniProt protein sequences under taxonomy 'Viridiplantae', taking the GO terms associated with the top hit for each gene model and incorporating this into the final functional annotation.

4.3.3 Comparative genomics

Protein sequences from *D. tokoro* and 25 other angiosperm species with comprehensive gene sets:*Elaeis guineensis*[260], *Phoenix dactylifera*[261], *Musa acuminata*[314], *Panicum hallii**, *Setaria viridis**, *Sorghum bicolor*[272], *Oropetium thomaeum*[270], *Aegilops tauschii*[315], *Brachypodium distachyon*[241], *Oryza sativa* Japonica[308], *Ananas comosus*[265], *Dioscorea rotundata*[307], *Phalaenopsis equestris*[271], *Spirodela polyrhiza*[273], *Zostera marina*[316], *Vitis vinifera*[317], *Carica papaya*[318], *Malus domestica*[268], *Arabidopsis thaliana*[240], *Ipomoea nil*[267], *Olea europaea*[319], *Chenopodium quinoa*[227], *Beta vulgaris*[266], *Nelumbo nucifera*[269], *Amborella trichopoda*[274], using the longest protein isoforms for each gene, were compared with OrthoFinder[204] using Diamond-0.8.37[313], DLCpar-0.9.1[320], MCL-12.068[321] and RAxML-8.2.9[264], to identify orthogroups (*these references were produced by the US Department of Energy Joint Genome Institute from and obtained via Phytozome[153]). I selected these species for a good spread across both the monocots and eudicots (Amborella as an outgroup to these), including *S. polyrhiza* and *Z. marina* from the Alismatales, and *P. equestris* (Asparagales) in particular, as these had been published after my *D. rotundata* work and represent families thought to be more closely related to Dioscoreaceae. Gene enrichment

analysis of orthologous gene families was performed with agriGO v2.0[322], using the Hypergeometric statistical test, a minimum of five mapping entries, $p > 0.05$, and the false discovery rate was adjusted using the Benjamini–Hochberg procedure[207].

For the species phylogeny, all 33 single-copy orthologous genes from 22 species, from the Orthofinder output, were used to generate multiple protein sequence alignments with MAFT[153] (Supplementary Table 7.35). These 22 species were selected to cover the majority of taxa present in the gene orthology analysis, while maintaining a reasonable number of single-copy orthogroups; of which all single-copy orthogroups present were used. Maximum likelihood trees were constructed based on the concatenated alignments using RAxML-8.2.8[264] with a JTT + Γ model and 1000 bootstraps. The split-network was then generated by Spectre[323], using FlatNetJoining (FlatNJ)[324] to construct a flat split network from the multiple protein sequence alignments, with default settings.

SynMAP[211] using LAST[325] alignments, DAGchainer[213] relative gene order (default options -D 20 and -A 5), and no merging of syntenic blocks were used as part of the CoGe platform[210] to identify syntenic blocks between the coding regions of *D. tokoro* and *D. rotundata*, and syntenic gene pair synonymous rate change calculated by CodeML[214] with a cut off of 1.7(log10) to remove noise. With the same method applied to self-synteny of *D. tokoro* and *D. rotundata*, with a CodeML cut off of 1.6(log10).

4.3.4 Sequencing and assembly of *D. alata*

A diploid *D. alata* male individual was sequenced to a depth of 120x coverage using a PCR-free paired-end Illumina fragment library, generated from gDNA extracted from leaf tissue of a single individual, and 250 bp paired-end reads on the Illumina HiSeq 2500 sequencing platform, in Rapid-Run mode, by the Earlham Institute, UK, pipelines team. I subjected raw FASTQ[111] reads to quality filtering using Kontaminant 2.0[115], using k-mer libraries for library adaptors, PhIX and *E. coli*. This was done to remove contamination from the sequences prior to assembly. A k-value of 21 was used for screening and filtering of sequences. The quality filtered paired-end reads were assembled using Discovar *de novo*[326] using the DiscovarExp option. Quality control steps were taken using KAT 2.0-alpha[191] to map unique k-mer content from the reads to the genome to identify missing or made up content and CEGMA 2.5[138] was run to identify homologues to core eukaryotic genes. Following assembly, bioawk (https://github.com/lh3/bioawk) with the fastx option was used to remove all contigs smaller than 2 kb, as k-mer spectra analysis showed these to not provide many new k-mers and only added to the number of contigs and repetition represent. I analysed the assembly for the presence of 248 highly

conserved core eukaryotic genes by CEGMA[138], using assemblies of 1, 2 and 3 kb minimum contig sizes, and observed no change in the presence of cegma between 1 and 2 kb cut-offs, mirroring the the k-mer spectra result, and a slight decrease from 212 to 211 complete cegma present at 3 kb (Table 7.19). Additionally I used 956 BUSCOs[139] on the 2 kb cut-off assembly, using the BUSCO early plantae release and found 845 (88%) of these with at least one complete single-copy present in the assembly, however there were a greater number of duplicated copies compared to the other two *Dioscorea* genomes (Table 7.20). Both CEGMA And BUSCO showed the assembly to have the majority of core protein genes to be at least partially assembled, indicating a reasonably complete (at least partial) gene set present in the assembly. The average GC content across the genome was seen to be similarly low, compared to the other *Dioscorea* species, at 36.05%. The final draft assembly contained 57,706 contigs with an N50 of 19.3 kb and a total assembly length of 620.9 Mb. The genome assembly is publicly available through NCBI GenBank (Accession: CZHE00000000.2).

The protein-coding genes of *D. alata* were predicted following the same MAKER workflow and datasets as those used for *D. rotundata*, with the addition of the *D. rotundata* protein sequences, resulting in 40,055 gene models.

4.3.5 Evolution of sex determination in *Dioscorea*

Protein sequences from *D. alata* and all species used in the *D. tokoro* gene orthology and enrichment study, using the longest protein isoforms for each gene, were compared with OrthoFinder[204] using Diamond-0.8.37[313], DLCpar-0.9.1[320], MCL-12.068[321] and RAxML-8.2.9[264], to identify groups of orthologus between species.

For the species phylogeny, all 30 single-copy orthologus genes from 22 species, from the Orthofinder output, were used to generate multiple protein sequence alignments with MAFT[153]. Maximum likelihood trees were constructed based on the concatenated alignments using RAxML-8.2.8[264] with a JTT + Γ model and 1000 bootstraps. The split-network was then generated by Spectre[323], using FlatNetJoining (FlatNJ)[324] to construct a flat split network from the multiple protein sequence alignments, with default settings.

SynMAP[211] using BLASTZ[212] alignments, DAGchainer[213] relative gene order (default options -D 20 and -A 5), and no merging of syntenic blocks were used as part of the CoGe platform[210] to identify syntenic blocks between the coding regions of *D. alata*, *D. tokoro* and *D. rotundata*, and syntenic gene pair synonymous rate change calculated by CodeML[214]. Further work investigating the branch specific evolutionary

rate (dN/dS) was carried out using MAFFT alignments and PAML, the rate of dN/dS was considered using orthogroups with significant maximum likelihood in the PAML model for uniform selective pressure among sites (M0), compared to variable selective pressure but no positive selection (M1), that had no null values present in the dN/dS under M0 of any branch.

4.4 Results

4.4.1 Genome Assembly and Annotation

Assembly of *D. tokoro* by S. Natsume, IBRC, produced a genome 370.87 Mb in length, that was a bit lower than the predicted 390 Mb, based on flowcytometry on another *D. tokoro* individual using rice as a control, and 467 Mb through k-mer spectra analysis of the ALLPATHS-LG assembly (Table 4.1; unpublished, S. Natsume, *et al*, IBRC). This would suggest that regions of the genome are missing from the assembly, potentially these are hard to sequence regions, such as repeat dense or regions of extremes in GC content. As such, genome completeness was assessed by BUSCO, of which 79.3% of the 1,440 embryophyta_odb9 BUSCOs were complete in the genome, compared to 90.7% in *D. rotundata*, with the remainder either fragmented or missing (Table 7.16,7.17). It's therefore likely that some of the gene space is also missing from the assembly, and that a similar approach to known hard to sequence regions, such as the FSW, would benefit from additional long range sequencing technologies.

I predicted genes and repeats using the *D. tokoro* reference genome sequence. To construct reliable gene models, I ran a first pass assembly using the MAKER pipeline, which only produced 19,275 gene models. As the number of gene models was 10,196 lower than *D. rotundata* and most other model plant species, I used the output from this in an EVM run with additional alignments of publicly available homologous sequence from related species. Later using PASA for correction and prediction of isoforms of the EVM gene/transcript models, in order to achieve a more comprehensive set of gene models[150, 246]. This resulted in 29,471 protein coding gene models and 31,283 transcripts. The difference in result between MAKER and EVM could be due to the quality metrics used to evaluate gene models, the weighted consensus evaluation model used in EVM has more flexibility, in that the weight of each piece of evidence is user definable, compared to the MAKER AED. Additional protein alignments used in the EVM annotation may have also provided the additional evidence required to validate

more gene models.

Gene density was found to generally increased toward the ends of pseudo-chromosomes, while repeat density increased towards the center of each pseudo-chromosome and is particularly pronounced on pseudo-chromosome 5, as would be expected given the heterochromatic nature of potential centromeric regions (Figure 4.1a-c, Table 4.1)[23]. Of this gene set, 73.4% of the 1,440 embryophyta_odb9 BUSCOs were present, representing the majority of those predicted within the current assembly (Table 7.18). I compared the genome sequence metrics of *D. tokoro* to our previously published *D. rotundata* reference, as this is the closest publicly available species. From this I found that *D. tokoro* has a total assembly size ~37.6 % smaller than *D. rotundata*, however I also observed a lower proportion of the genome to be represented by interspersed repeat sequence, accounting for ~140 Mb of the difference in genome size between both species. Whilst this may be biological, it's possible this could be due to an assembly issue as repeat dense regions are difficult to assemble and *D. tokoro* did not benefit from the BAC-end sequencing that was used in *D. rotundata*. Both species have similarly low GC content, and number of exons in genes was similar to *D. rotundata*, although the average and total exon size was smaller (Figure 4.1d., Table 4.1). Finally, investigation of potential genome duplication events with self-synteny did not show evidence for any largely duplicated regions in *D. tokoro* (Figure 4.1e.).

4.4.2 Gene orthology prediction and phylogentic inference

I compared gene orthology across 26 angiosperm species, including *D. tokoro* and *D. rotundata*, and identified 5,005 conserved orthogroups across all species, with a 23,531 orthogroups present in total (Figure 4.2). These 5,005 conserved orthogroups showed significant enrichment (Hypergeometric test, Benjamini–Hochberg procedure; $q < 0.05$) for 76 plant slim GO terms, when compared to all orthogroups of *D. tokoro*. These orthogroups are enriched for house keeping functions associated with key processes including 'cellular protein metabolic processes', 'RNA binding', and 'intracellular part' and several organelles (Figure 4.6, 4.7, and Table 7.21). Of these, only one single-copy orthogroup (OG0010694) was present in all species, with functional annotation for tRNA sulfurtransferase activity. From the total set of orthogroups, *D. tokoro* was observed to have 10,862 orthogroups, containing 19,671 genes, and 9,800 orphan genes that could not be assigned to any orthogroups. Of the 9,800 orphan genes, only 968 had associated GO terms and these showed no significant enrichment (Hypergeometric test, Benjamini–Hochberg procedure; $q < 0.05$) compared to all functionally annotated genes

in *D. tokoro*. These included nine subtilisin-like proteases, three zinc finger, and three mechanosensitive ion channel coding genes, forming the largest gene families (Table). Of these, the subtilisin-like proteases are particularly interesting as they are associated with biotic and abiotic stress responses, potentially unique to *D. tokoro's* life history as genus specific expansions have been indentified across the angiosperms and have likely evolved in response to external stressors[327]. The number of orphan orthogroups and genes in *D. tokoro* is similar to the closest relative studied, *D. rotundata*, which had 10,005 orthogroups (19,660 genes). Between both species, 8,682 orthogroups (16,447 genes in *D. tokoro* and 17,123 in *D. rotundata*) were shared, and 32 *D. tokoro* specific orthogroups (219 genes) and 84 *D. rotundata* specific orthogroups (571 genes). The 32 orthogroups only observed in *D. tokoro* were associated with nine plant slim GO terms, including 'primary metabolic process', 'binding' and 'cell', however none of these showed significant (Hypergeometric test, Benjamini–Hochberg procedure; $q < 0.05$) plant slim GO term enrichment compare to GO terms of the 5,005 conserved orthogroups (Table 7.22). These orthogroups consisted of a diverse set of single gene families, including nuclear ribonucleoprotein, isocitrate lyase and phosphoinositide phosphatase suppressor of actin-1 (SAC1) genes (Table 7.31). Of these, SAC1 has been shown to be vital for cellular organisation and regulation of lipid storage in *A. thaliana*[328]. In comparison, the largest gene family observed in the *D. rotundata* specific orthogroups in this study were composed of 24 pentatricopeptide repeats (Table 7.32). This family of genes represent one of the largest and most diverse in terrestrial plants[277].

Between the species studied, 328 orthogroups were observed to be conserved in all species except for *D. tokoro*. Using the *D. rotundata* annotation to compared these orthogroups to those conserved in all species studied, 23 plant slim GO terms were found to be significantly enriched (Hypergeometric test, Benjamini–Hochberg procedure; $q < 0.05$), including response to abiotic, biotic, and external stimulus, and plastid (Figure 4.8, 4.9, and Table 7.24). Within these orthogroups, families of nitrilase homologs are particularly interesting as nitrilases enzymes are found throughout the angiosperms and have been shown to play a role in defence against pathogens and herbivores through catabolism of cyanide, and cyanogenic glycosides and glucosinolates (Table 7.25)[329]. Hydrogen cyanide is well documented in *Dioscorea* species, including *D. roundata*[330, 331]. However, saponins, which *D. tokoro* is well known for, represents another similar method of deterring herbivores with bitter taste and toxicity in *Dioscorea*[288]. It's therefore possible that some nitrilases have either been lost in *D. tokoro*, due to it's evolution of another class of anti-nutrient compounds for defence of its rhizome.

Between both *D. tokoro* and *D. rotundata* there were 174 orthogroups not present in the 24 other species studied, of these the plant slim GO term for 'hydrolase activity' was seen to be significantly enriched ($q < 0.05$) when compared to orthogroups conserved across all 26 species (Table 7.34, 7.23). Of these, three orthogroups (OG0017758, OG0021859 and G0017758) contained genes with lipase associated domains that could possibly play a role as storage proteins for yam tuber/rhizomes, as this class of genes are associated with energy storage in other species, such as lipase 'patatin' in potato[3, 332].

Only 66.7% of genes could be assigned to orthogroups in *D. tokoro*, compared to 75% in *D. rotundata*, however more orthogroups (46.2%) from all species studied were present in *D. tokoro* than in *D. rotundata*, with only 42.5% of orthogroups present.

In *D. rotundata*, I previously observed an expansion of bulb-type lectin (snowdrop lectin; B-lectin) gene families when compared to *Arabidopsis thaliana*, *Brachypodium distachyon* and *Oryza sativa*[307]. These B-lectin genes were present in 18 orthogroups in this study, containing 127 genes, of which 14 (58 genes) were present in *D. tokoro*. Only two groups showed expansion in *D. rotundata* compared to *D. tokoro* (Table 7.30). Indicating that these expansions maybe specific to the *D. rotundata* lineage or could have been lost in *D. tokoro*.

I constructed a phylogenetic tree based on the alignment of 33 orthologous protein-coding genes with single-copy orthogroups across *D. tokoro* and 21 other species from the gene orthology study (Figure 4.3). Both trees place Dioscoreales as one of the closest lineages to the base of the Monocotyledons as has been previously reported[333]. Within the taxa, *D. tokoro* is thought to have split from *D. rotundata* earlier on, becoming a member of the Stenophora clade, and the ancestor of *D. rotundata* to have undergone several speciation events before forming the Enantiophyllum clade[238, 292, 334, 335]. While both the phylogenetic tree and split-network support this, inclusion of additional *Dioscorea* species from other clades and use of an outgroup will assist with reconstructing the evolutionary history of the genus at whole genome level.

To infer the past genome duplication in *Dioscorea*, I performed genome-wide dot plot analysis of *D. tokoro* against itself, and also against *D. rotundata*, from which I could observe no recent genome duplication based on relative gene order (Figure 4.1e, 4.4). However, there are small duplicated areas in *D. tokoro* present on pseudo-chromosome 2, and syntenic blocks at the start and end of pseudo-chromosome 3 are duplicated in pseudo-chromosome 8 and 15 of *D. rotundata* (highlighted with arrows in Figure 4.4). Based on synonymous mutation rate change, these duplication events are relatively old. In all, there is a large degree of conserved synteny in both species.

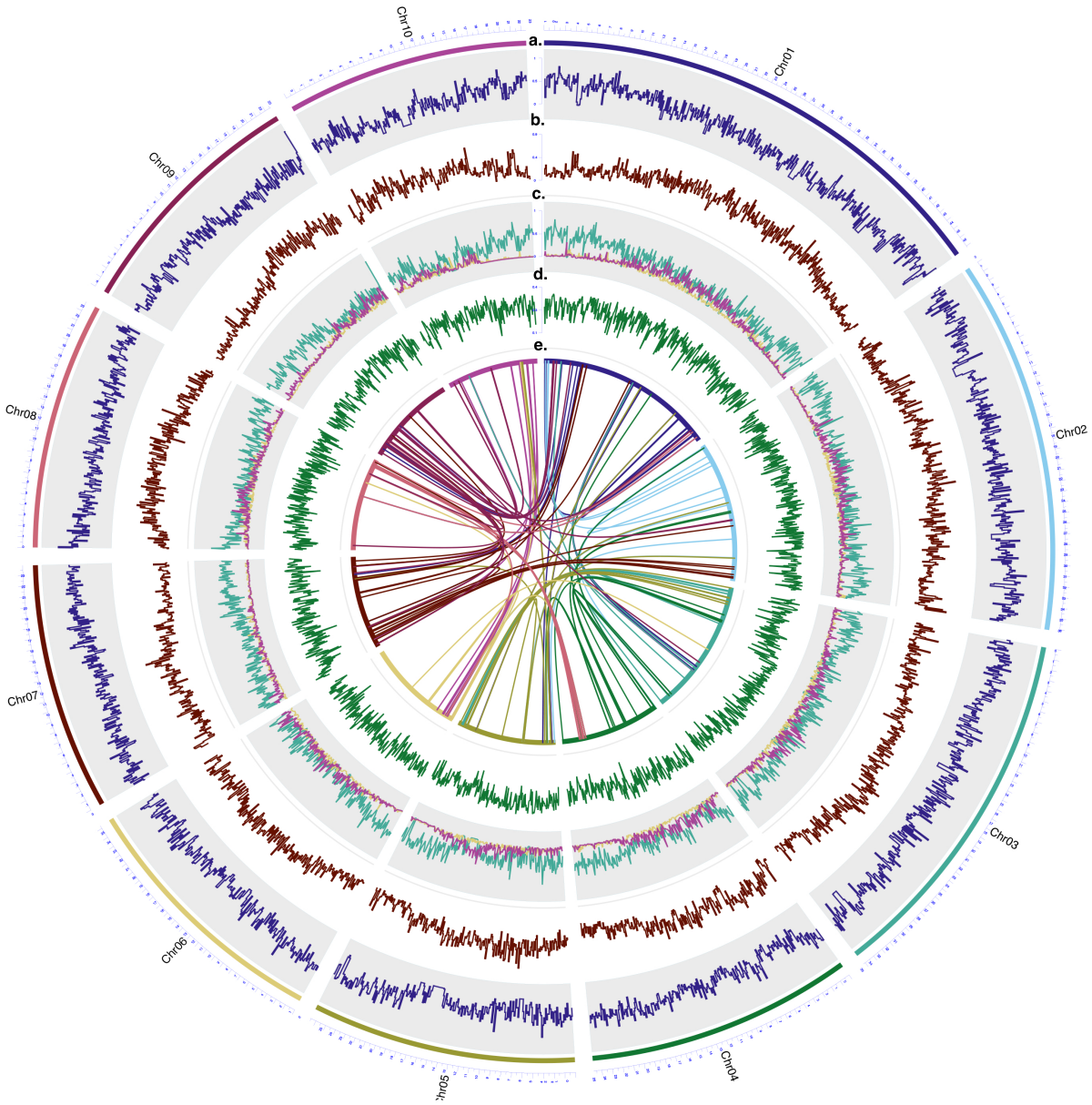

Figure 4.1: Genome features of *D. tokoro*, showing coverage in 100 kb windows across pseudo-chromosomes of **a.** genes, **b.** interspersed repeats, **c.** genes (green), copia LTR (blue) and gypsie LTR (pink), and **d.** GC%. **e.** Self synteny blocks of CDS > 10 kb.

Table 4.1: *Dioscorea tokoro* genome scaffold assembly summary and comparison with *D. rotundata*.

Feature	**Value**		
	***D. tokoro* EVM**	***D. tokoro* MAKER**	***D. rotundata* v0.1**
Total length (Mbp)	370.87	370.87	594.23
GC (%)	38.22	38.22	35.83
No. scaffolds (>= 0 bp)	11,063	11,063	4723
No. scaffolds (>= 1000 bp)	10,903	10,903	4,704
Largest scaffold (Mbp)	2.60	2.60	13.61
N50	0.31	0.31	2.12
N75	0.12	0.12	0.77
No. N's per 100 kb	19,445	19,445	282.45*
No. genes	29,471	19,275	26,198
Exons**			
Number	153,291	105,027	158,059
Average no. per gene	5.20	5.44	6.03
Total length (Mbp)	30.76	23.17	42.43
Average size (bp)	200.64	220.61	268.43
Average GC (%)	47.09	45.30	44.08
Introns			
Number	123,820	66,477	105,663
Average no. per gene	4.20	3.44	4.03
Total length (Mbp)	83.72	48.72	83.12
Average size (bp)	676.18	568.18	630.33
Average GC (%)	35.40	32.55	32.37
Transposible Elements			
Total interspersed (%)	36.47	36.47	46.07
Total interspersed total length (Mbp)	135.27	135.27	274.51
SINEs (%)	0.08	0.08	0.02
SINEs total length (Mbp)	0.28	0.28	0.13
LINEs (%)	2.75	2.75	2.43
LINEs total length (Mbp)	10.21	10.21	14.46
LTR elements (%)	26.94	26.94	22.82
LTR elements total length (Mbp)	99.91	99.91	135.71
DNA elements (%)	1.76	1.76	6.70
DNA elements total length (Mbp)	6.51	6.51	39.83
Unclassified (%)	4.95	4.95	14.20
Unclassified total length (Mbp)	18.36	18.36	84.38

*Number of Ns per 100 kb using the *D. rotundata* broken scaffolds. **For genes with multiple transcripts, gene models were flattened and the longest exons were used to obtain these figures.

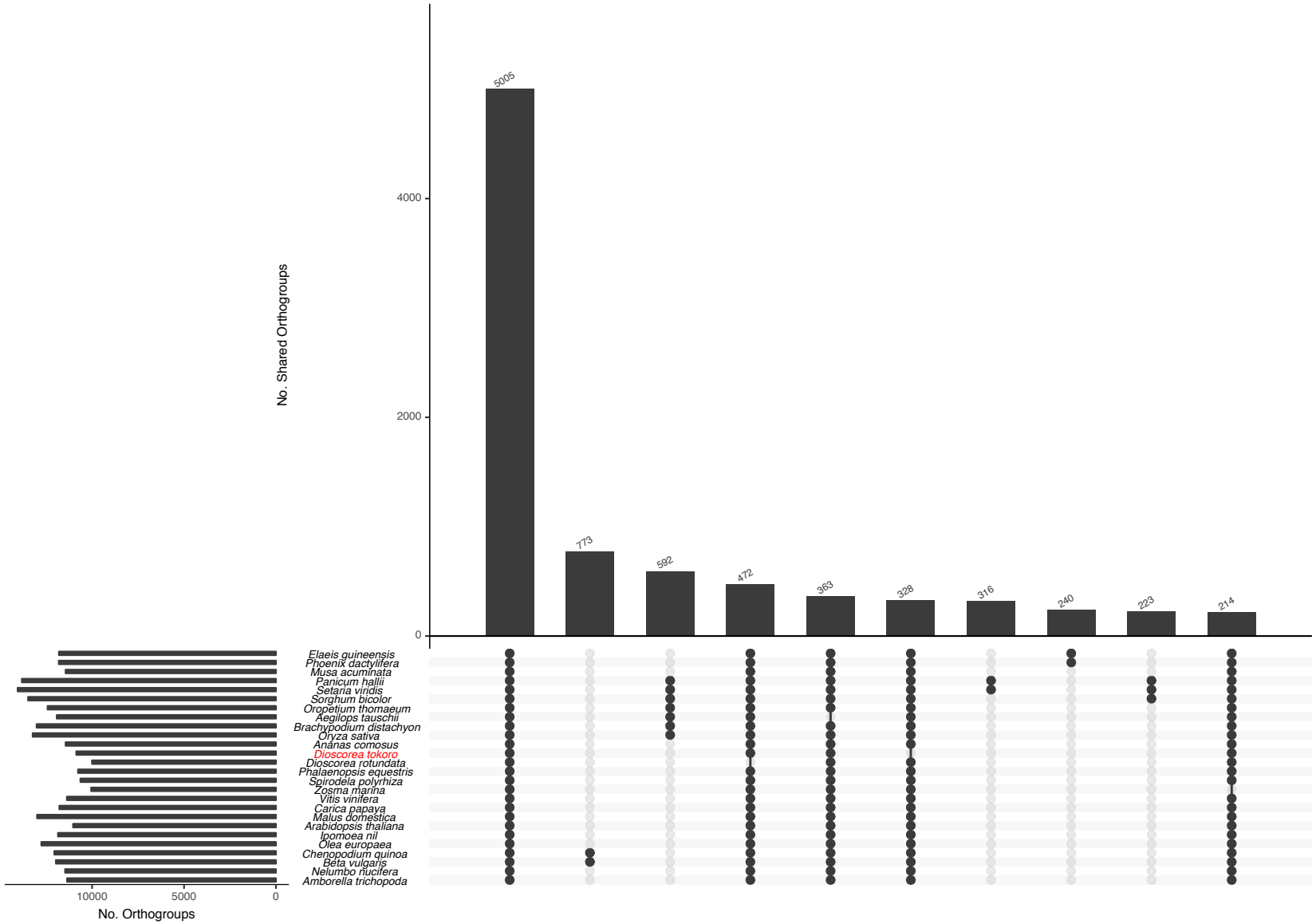

Figure 4.2: UpSet plot showing the 10 most frequent intersects of orthogroups present in *D. tokoro* and 25 other angiosperm species.

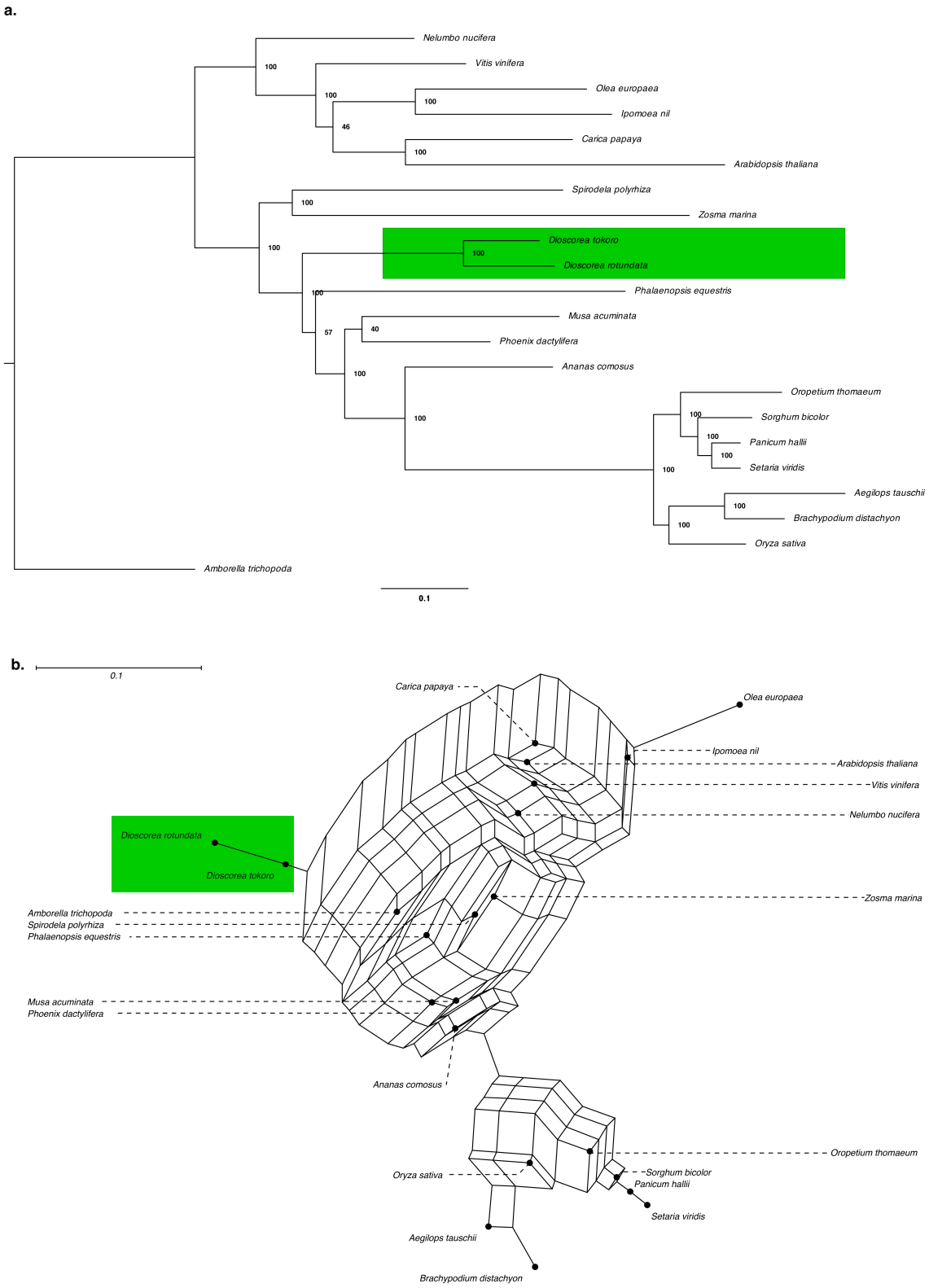

Figure 4.3: Phylogenetic relationships between *D. tokoro* and 21 other angiosperm species from this study, based on alignment of 33 single-copy orthologs. **a.** Bipartiain tree generated by RAxML maximum likelihood analysis, with confidence intervals from 1,000 bootstrap resamplings shown. **b.** Split network generated by Spectre using the flat net joining method.

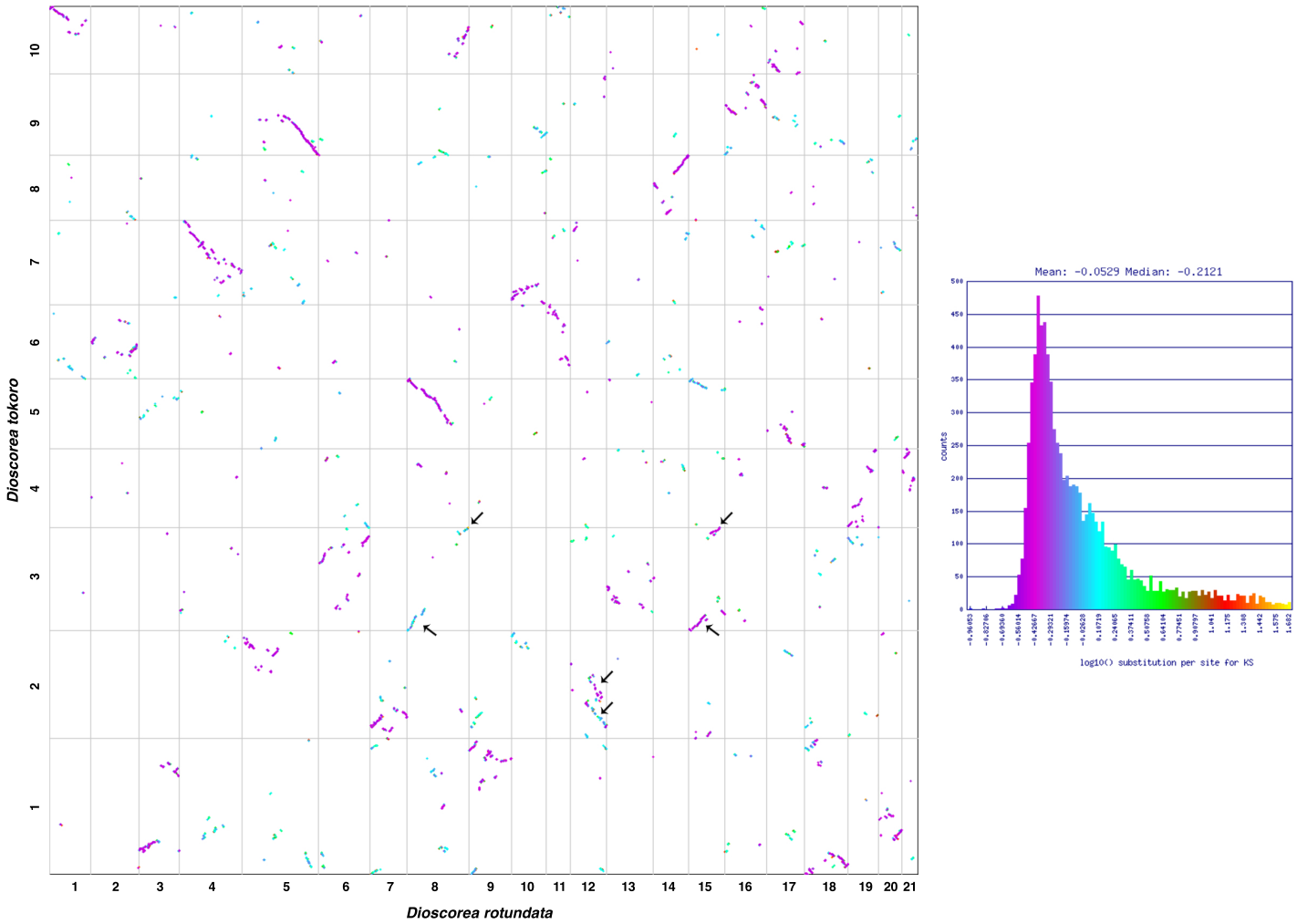

Figure 4.4: SynMAP syntenic dotplot, and synonymous substitution histogram, of *D. tokoro* and *D. rotundata* CDS pseudo-chromosomes show regions of macro synteny and no recent large scale genome duplication event. Dotplot axis are labeled with pseudo-chromosome number. Syntelogs have been coloured based on their synonymous (KS) rate change. Region of duplication events of interest are highlighted with arrows. Red lines represent positive regulation of downstream GO terms.

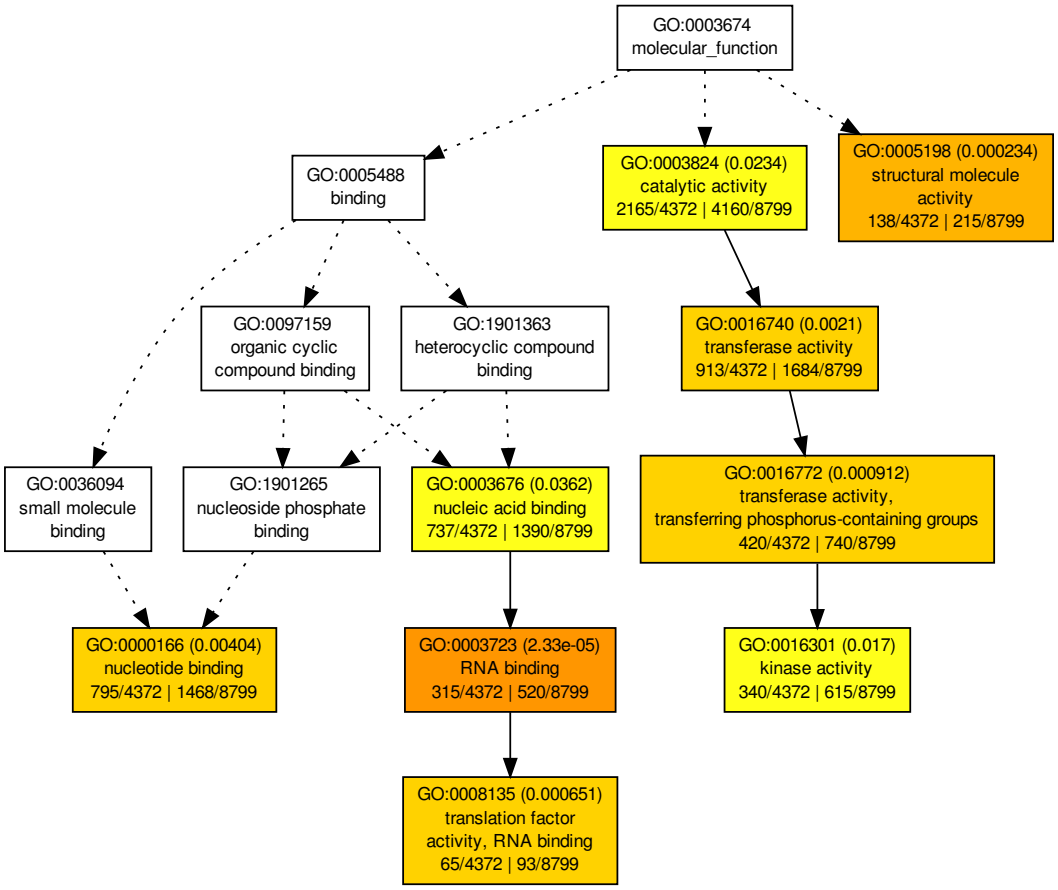

Figure 4.5: Hierarchical tree graph of the three most significantly enriched plant slim GO terms belonging to the 'molecular function' category, conserved between *D. tokoro* and 25 other angiosperm species, when compared to all orthogroups in *D. tokoro*. Boxes in the graph represent GO terms labeled by their GO ID, term definition and statistical information. The significant term (adjusted P $\leq$ 0.05) are marked with colour, while non-significant terms are shown as white boxes. Colour of the boxes indicates level of statistical significance, from yellow (least significant) to red (most significant). Solid, dashed, and dotted lines represent two, one and zero enriched terms at both ends connected by the line, respectively.

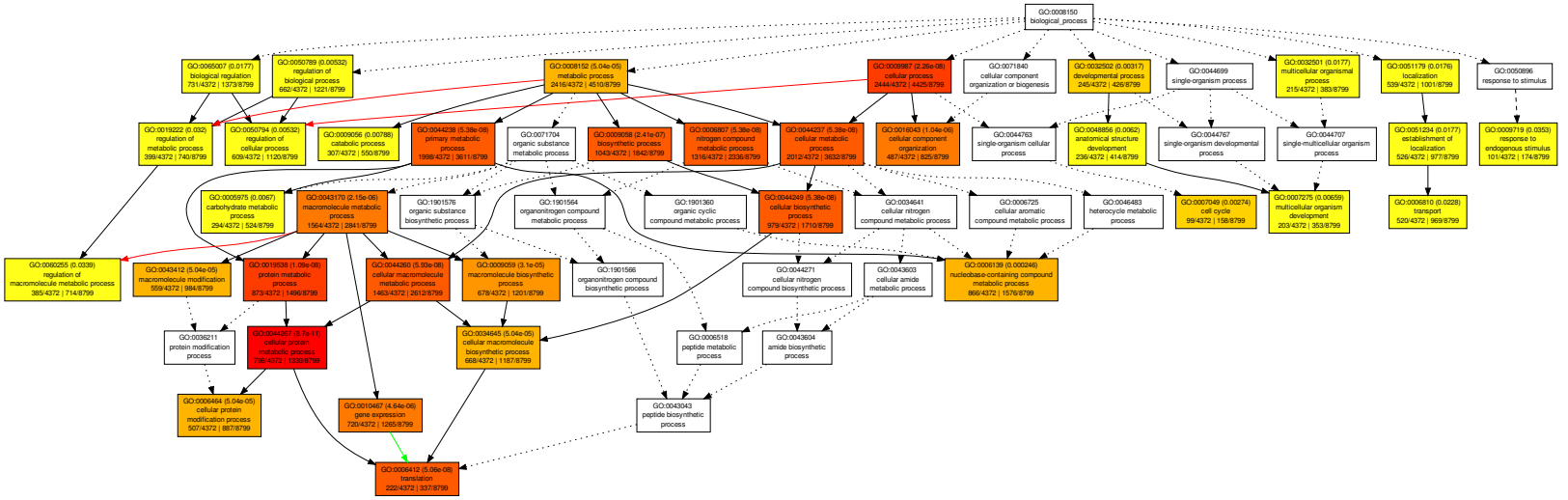

Figure 4.6: Hierarchical tree graph of the three most significantly enriched plant slim GO terms belonging to the 'biological process' category, conserved between *D. tokoro* and 25 other angiosperm species, when compared to all orthogroups in *D. tokoro*. Boxes in the graph represent GO terms labeled by their GO ID, term definition and statistical information. The significant term (adjusted P $\leq$ 0.05) are marked with colour, while non-significant terms are shown as white boxes. Colour of the boxes indicates level of statistical significance, from yellow (least significant) to red (most significant). Solid, dashed, and dotted lines represent two, one and zero enriched terms at both ends connected by the line, respectively.

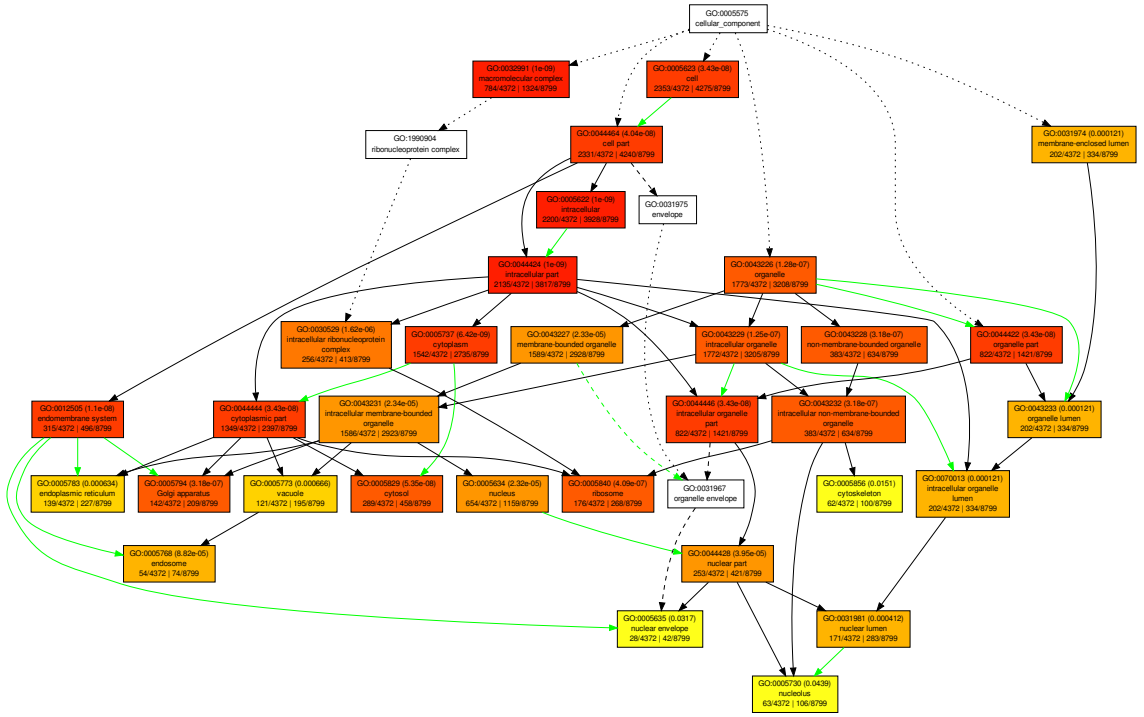

Figure 4.7: Hierarchical tree graph of the three most significantly enriched plant slim GO terms belonging to the 'cellular component' category, conserved between *D. tokoro* and 25 other angiosperm species, when compared to all orthogroups in *D. tokoro*. Boxes in the graph represent GO terms labeled by their GO ID, term definition and statistical information. The significant term (adjusted $P \leq 0.05$) are marked with colour, while non-significant terms are shown as white boxes. Colour of the boxes indicates level of statistical significance, from yellow (least significant) to red (most significant). Solid, dashed, and dotted lines represent two, one and zero enriched terms at both ends connected by the line, respectively. Green lines represent negative regulation of downstream GO terms.

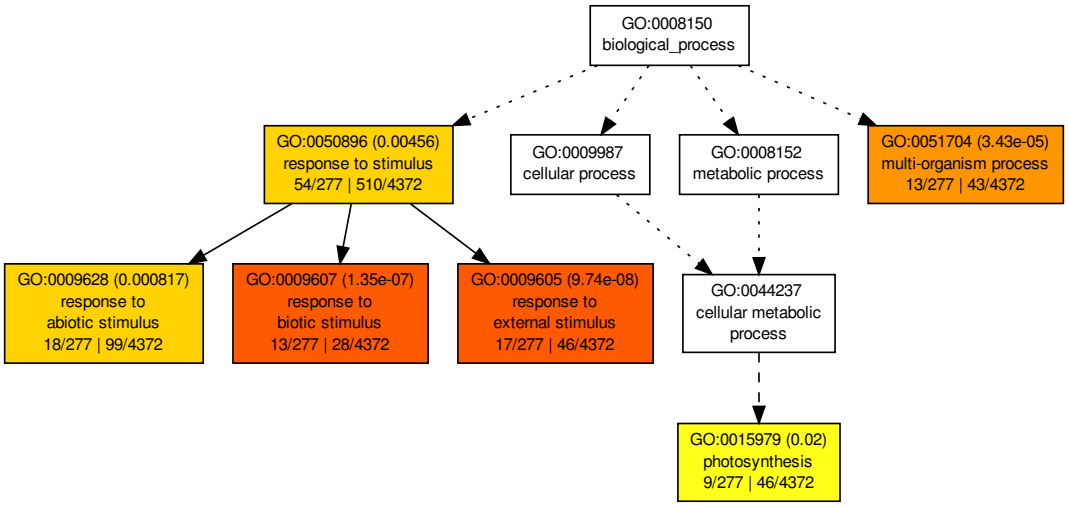

Figure 4.8: Hierarchical tree graph of the three most significantly enriched plant slim GO terms belonging to the 'biological process' category, comparing all orthogroups conserved between *D. tokoro* and 25 other angiosperm species, compared with all orthogroups conserved in all species expect for *D. tokoro*. Boxes in the graph represent GO terms labeled by their GO ID, term definition and statistical information. The significant term (adjusted P $\leq$ 0.05) are marked with colour, while non-significant terms are shown as white boxes. Colour of the boxes indicates level of statistical significance, from yellow (least significant) to red (most significant). Solid, dashed, and dotted lines represent two, one and zero enriched terms at both ends connected by the line, respectively.

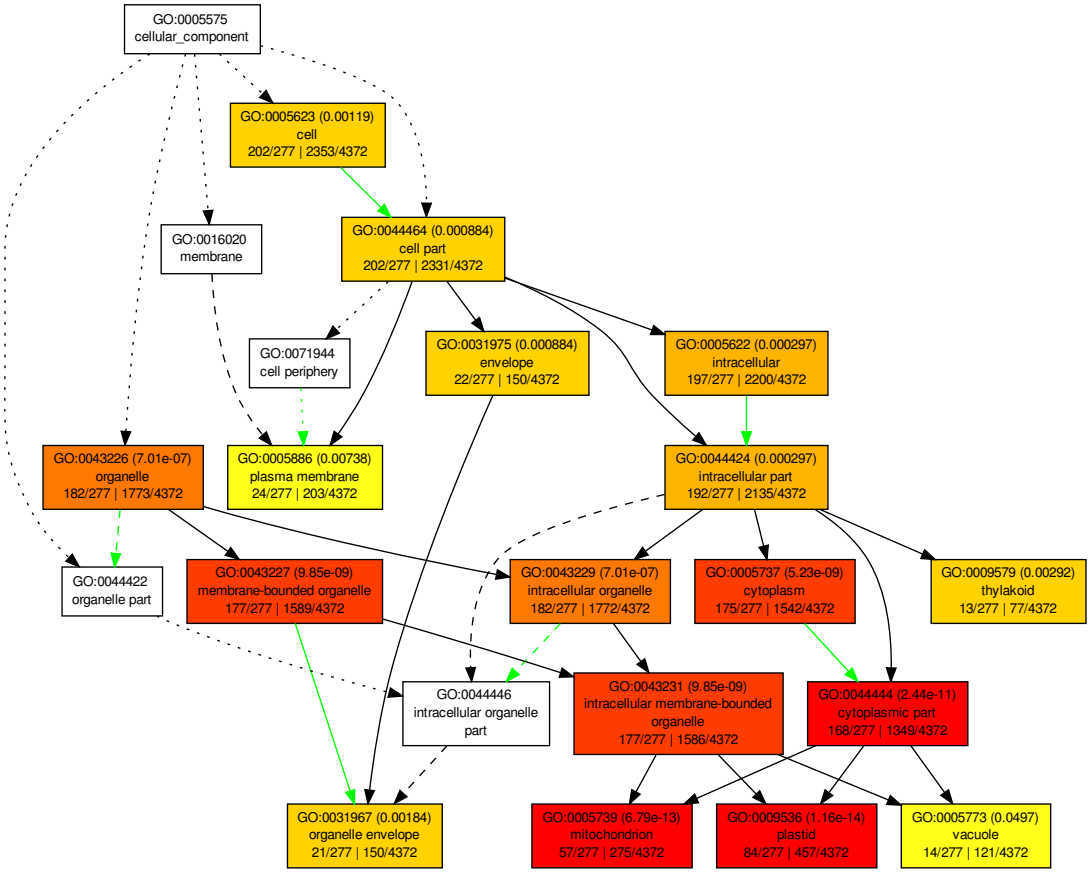

Figure 4.9: Hierarchical tree graph of the three most significantly enriched plant slim GO terms belonging to the 'cellular component' category, comparing all orthogroups conserved between *D. tokoro* and 25 other angiosperm species, compared with all orthogroups conserved in all species expect for *D. tokoro*. Boxes in the graph represent GO terms labeled by their GO ID, term definition and statistical information. The significant term (adjusted P $\leq$ 0.05) are marked with colour, while non-significant terms are shown as white boxes. Colour of the boxes indicates level of statistical significance, from yellow (least significant) to red (most significant). Solid, dashed, and dotted lines represent two, one and zero enriched terms at both ends connected by the line, respectively. Green lines represent negative regulation of downstream GO terms.

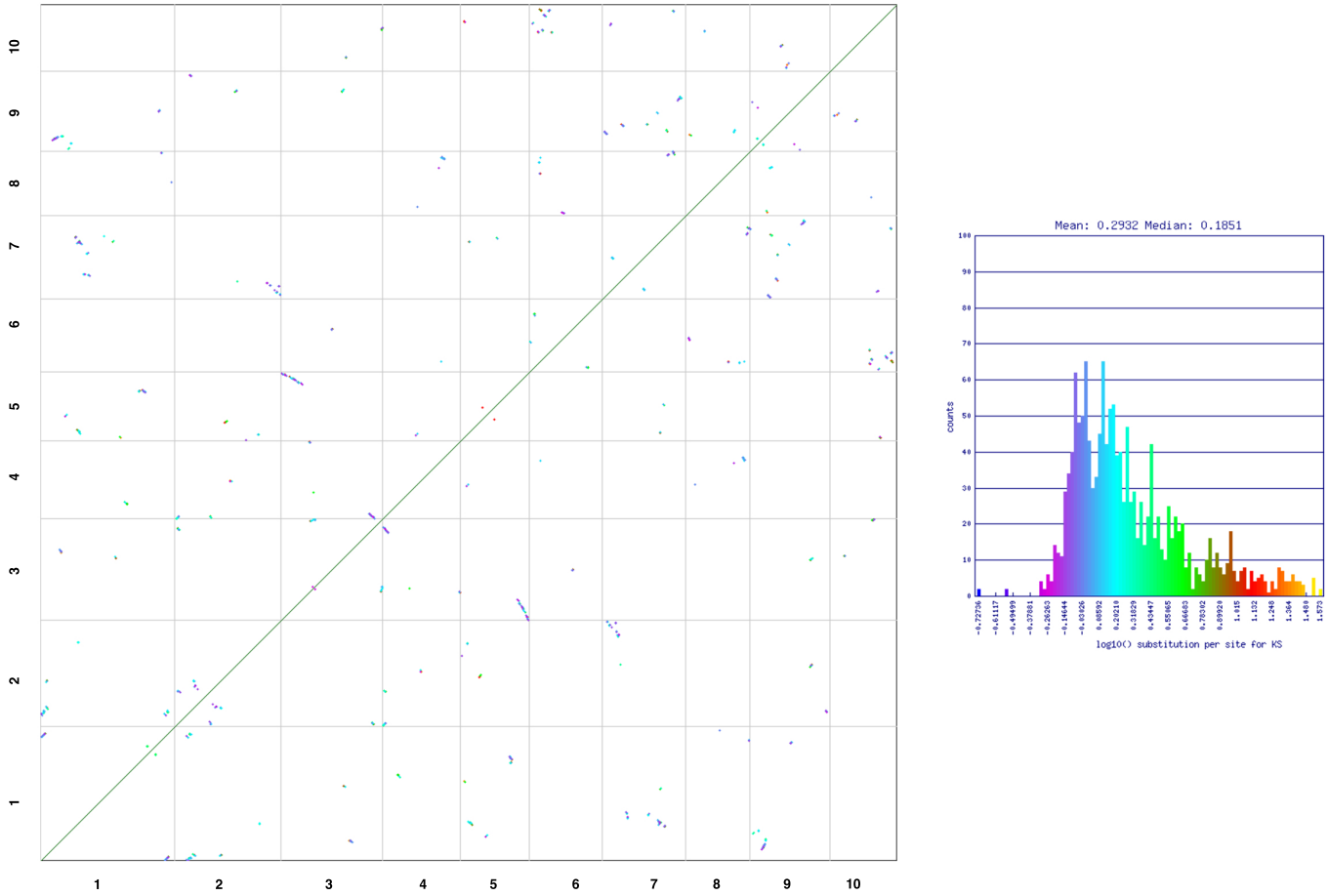

Figure 4.10: SynMAP self synteny dotplot, and synonymous substitution histogram, of *D. tokoro* CDS pseudo-chromosomes show no recent large scale genome duplication event. Dotplot axis are labeled with pseudo-chromosome number. Syntelogs have been coloured based on their synonymous (KS) rate change.

4.4.3 Evolution of sex determination in *Dioscorea*

Sex chromosomes in dioecious species are thought to emerge through recombination of suppression within the MSY/FSW, containing two or more sterility genes, that extends the evolutionary strata across PAR regions and eventually the near entirety of the chromosome[23]. As such, in the sex chromosomes of *D. tokoro* and *D. rotundata*, I would expect to see a higher evolutionary rate that reflects decreased recombination and relaxation of selection or positive selection, compared to the PAR and autosomal regions of the genome. In this particular case, recombination suppression is expected to be the main driver of an increased evolutionary rate in the sex chromosomes, specifically within the evolutionary strata.

Through the use of QTL-Seq, similarly to our previous *D. rotundata* study, we were able to identify a putative MSY sex determination loci in *D. tokoro* on pseudo-chromosome 3[307]. The MSY was found to be far larger than *D. rotundata*, at 21.4 Mb in length compared to 161 Kb, and to contain 1,265 genes, far more than the 57 within the FSW of *D. rotundata*.

I began to investigate the evolution of sex determination in these two species by comparing syntenic blocks between the pseudo-chromosomes of both *D. tokoro* and *D. rotundata* with the other species' proto-sex pseudo-chromosome, focusing on comparing the PAR regions as no synteny could be observed in the *D. rotundata* FSW sex determination region to *D. tokoro* MSY (Figure 4.10). The evolutionary rate (Kn/Ks) across the *D. tokoro* MSY CDS, compared with Kn/Ks of syntenic PAR regions of *D. rotundata*, were marginally significantly increased (Wilcoxon signed-rank test[336]; $p < 0.05$)(Figure 4.11). Conversely, comparison of CDS Kn/KS with *D. rotundata* PAR regions, to pseudo-autosomes, showed no significant difference in mutation rate (Figure 4.12).

In the *D. tokoro* proto-sex pseudo-chromosome, the combined repeat coverage was found to be higher than any of the pseudo-autosomes and significantly (Wilcoxon signed-rank test; $p < 0.05$) greater than the average combined repeat coverage of the pseudo-autosomes (Table 7.267.27). In comparison, the *D. rotundata* proto-sex pseudo-chromosome had the second highest content of combined copia and gypsy LTR(s) at 29.19%, behind that of pseudo-chromosome 2 (29.95%), that was significantly (one sample T-test[337]; $p < 0.05$) greater than the average content of the pseudo-autosomes (Table 7.28, 7.29). Gene and total exon content were also significantly reduced (Wilcoxon signed-rank test; $p < 0.05$), with a steep reduction in gene and exon density compared to other pseudo-molecules. This could indicate a maturing sex chromosome as the low density of CDS and high repeat content are consistent with the pseudogensiation of genes

and insertion of transposable elements[36, 338–340]. However, both sets of proto-sex pseudo-chromosomes have no cytologically observed heteromorphism, suggesting that they are in the early proto-sex stages of evolving into sex chromosomes. In addition to the absence of observed evidence for suppression of recombination in the sex determination regions. Given this, I hypothesise that the the FSW has either 1) recently been acquired by *D. rotundata*, after divergence from *D. tokoro*, or 2) the FSW is more representative of the ancestral state of sex in *Dioscorea*, having been lost in *D. tokoro* that is now MSY; both resulting in the turnover of sex determination.

In order to further investigate the evolutionary history of the *D. rotundata* FSW, compared with the *D. tokoro* MSY and to learn more about sex determination evolution in *Dioscorea*, I sequenced and assembled the genome of the related Enantiophyllum species, *D. alata*. I investigated the conservation of the FSW in *D. alata*, assessing if it shares similar sex determination to *D. rotundata*, and potentially other Enantiophyllum species, the most socio-economically important, also giving us the opportunity to test evolutionary rate of the three species to learn more about the age of the sex determination loci of *D. rotundata* and *D. tokoro*.

Briefly, a diploid phenotypically female *D. alata* individual was sequenced to a depth of 120x coverage using paired-end Illumina sequencing, this was then assembled with Discovar *de novo*[326] using the DiscovarExp option, and annotation was carried out similarly to *D. rotundata* with the addition of the *D. rotundata* protein sequences. This produced a first pass assembly with a total length of 620.9 Mb, similar in size to *D. rotundata* and N50 of 19.3 kb, far less than the scaffold and mapped assembly of *D. rotundata* and *D. tokoro*. In total, I predicted 40,055 gene models, far higher than either other *Dioscorea* genome and potentially an indication of many partial gene models, due to the reduced contiguity and lack of long read sequences to sequence through repetitive regions. However, the genome is of sufficient quality for comparisons with the other species in this study.

Phylogenetic reconstruction using the alignment of 30 orthologous protein-coding genes with single-copy orthogroups across *D. alata*, *D. tokoro*, *D. rotundata* and 19 other species showed *D. alata* to be closer related to *D. rotundata* than *D. tokoro*, as has been previously reported (Figure 4.13)[291, 341].

Comparison of CDS synteny between *D. alata* and, *D. rotundata* and *D. tokoro*, showed microsynteny with *D. alata* that was used for Kn/Ks calculations. However, synteny at chromosome level was heavily obscured and likely not possible due to the fragmented nature of the *D. alata* assembly. Further investigation of gene with 1:1

orthology between *D. tokoro* and *D. rotundata*, and *D. alata*, found the majority of genes in the PAR and sex determination loci to not have 1:1 orthology between species (Table 7.36). Additionally, I observed 2,351 genes with 1:1 orthology in *D. tokoro* and 3,445 in *D. rotundata*, with *D. alata*.

From the gene orthology, I also observed evidence of gene duplications in the sex determination regions of both *D. rotundata* and *D. tokoro*. In the FSW of *D. rotundata* I observed 4/53 orthogroups to be expanded. Of these orthogroups, a duplication of two putative ATP-binding cassette (ABC) transporter type-1 genes (Dr15771 and Dr190331) were observed in the FSW. ABC transporter genes have a number of associated of functions in land plants, such as pathogen resistance, tolerance to abiotic stresses, and anther and pollen development, and are often seen to be expanded in plants, potentially play a pivotal role in adaptation to land[342]. Additionally, I also found only one gene in the orthogroups of the FSW that was not conserved in either other *Dioscorea*, a putative Histone H3 (Dr19016) gene that would be interesting for further study into the epigenetic regulation of sex in *D. rotundata*. Of the expanded orthogroups in *D. tokoro* MSY, I observed 36/1131 to be expanded. These genes included a potential duplication of putative Alpha/beta hydrolase fold-1 genes (Dt20773 and Dt22585, and Dt23082 belonging to another orthogroup), a superfamily associated with regulation of pathways, including metabolic processes and development[343].

I then extended this analysis using the orthogroups of 2,794 genes with 1:1:1 orthology between *D. tokoro*, *D. rotundata*, and *D. alata* to investigate the branch specific evolutionary rate (dN/dS) of the proto-sex pseudo-chromosomes of *D. tokoro* and *D. rotundata*. For this analysis, I specifically compared the dN/dS of orthologs located on the proto-sex pseudo-chromosomes to orthologs on their respective autosomes. In order to take into account differences in evolutionary history that may influence the analyses, the comparison was performed on each branch independently. For each comparison, orthologs on the proto-sex pseudo-chromosomes showed a significant difference in dN/dS, that was seen to be significantly less (Wilcoxon signed-rank test; $p < 0.05$), through alternative hypothesis testing, in all branches for both species' proto-sex pseudo-chromosomes.

From these results we can consider that both species' proto-sex chromosomes have a slower evolutionary rate than their respective autosomes and that this is conserved between all branches studied. While in heterogametic sex chromosomes we may expect to see a higher evolutionary rate, due to the loss of selective pressure caused by suppression of recombination, as there has been no evidence found for recombination suppression and that these sex chromosomes are potentially still relatively new, these result are

unsurprising.

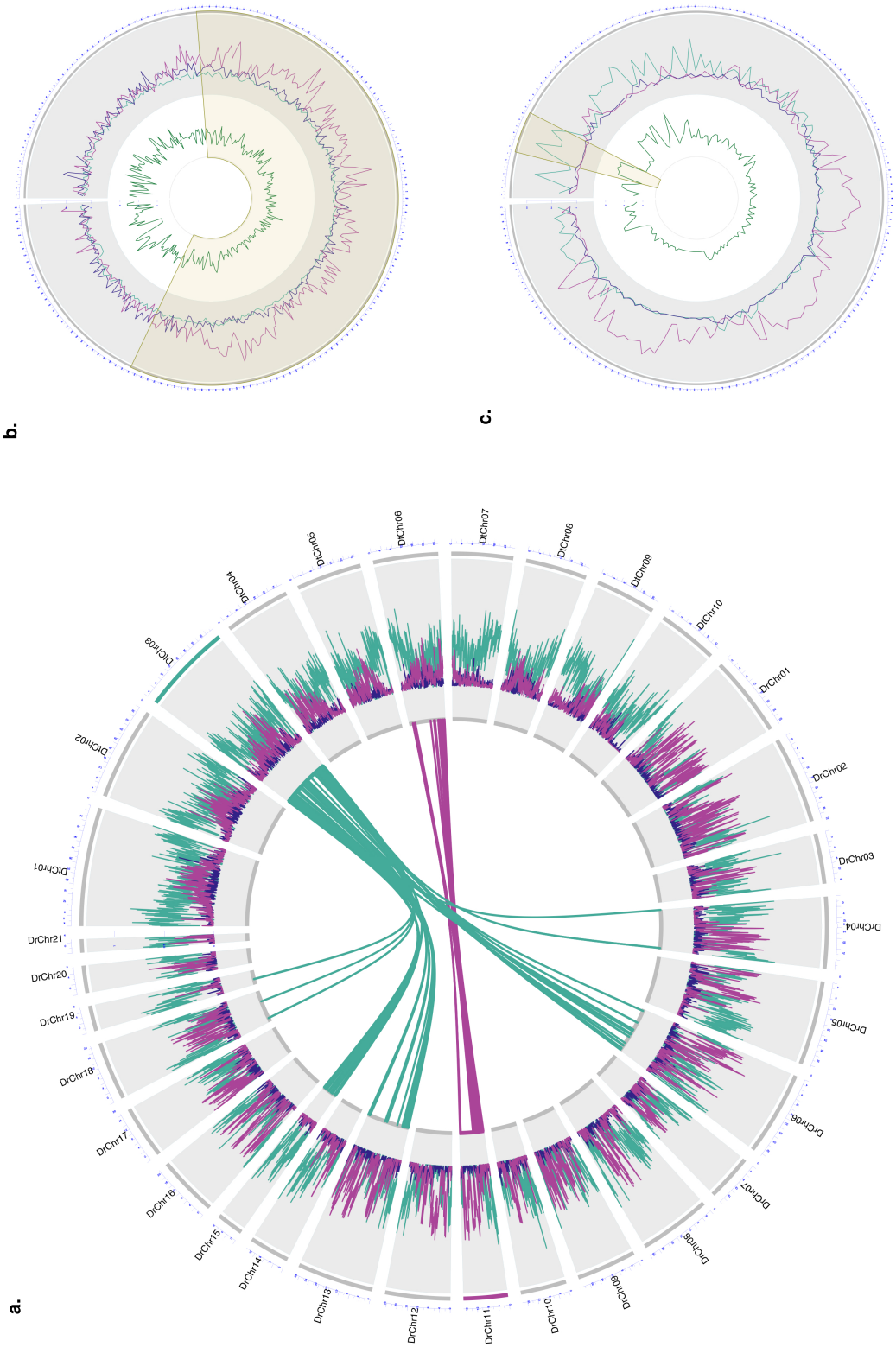
b.
c.
a.
DrChr06
DrChr07
DrChr08
DrChr09
DrChr10
DrChr01
DrChr02
DrChr03
DrChr04
DrChr05
DrChr06
DrChr07
DrChr08
DrChr09
DrChr10
DrChr11
DrChr12
DrChr13
DrChr14
DrChr15
DrChr16
DrChr17
DrChr18
DrChr19
DrChr20
DrChr21
DrChr01
DrChr02
DrChr03
DrChr04
DrChr05

Figure 4.10: Comparison of *D. tokoro* and *D. rotundata* pseudo-chromosomes. **a.** Coverage in 100 kb windows, across both genomes, of genes (green), copia LTR (blue), and gypsie (LTR) are show on the outer circle. The inner links show syntenic blocks between the pseudo-chromosomes of both *D. tokoro* (green) and *D. rotundata* (pink), using syntenic blocks > 10 kb. **b.** *D. tokoro* proto-sex pseudo-chromosome 3, with coverage in 100 kb windows of genes (green), copia LTR (blue), and gypsie (LTR) on the outer circle, and density of genes with orthologs in *D. rotundata* on the inner circle. **c.** *D. rotundata* proto-sex pseudo-chromosome 11 with coverage in 100 kb windows of genes (green), copia LTR (blue), and gypsie (LTR) on the outer circle, and density of genes with orthologs in *D. rotundata* on the inner circle. Putative sex determination loci are highlighted in both **b** and **c**.

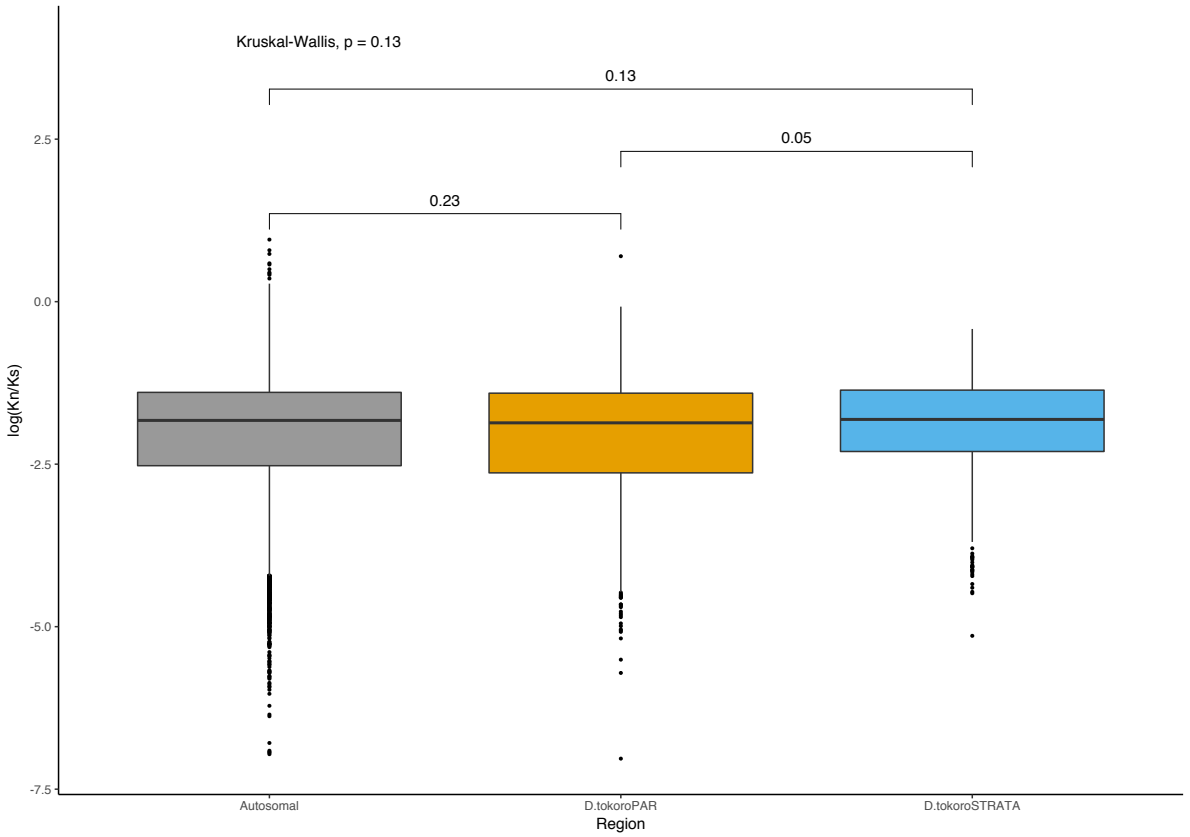

Figure 4.11: Boxplot showing log non-synonymous/synonymous mutation rate (Kn/Ks) of CDS synteny between *D. rotundata* and *D. tokoro*, within the putative sex determination loci (blue) and PAR regions (orange) of *D. tokoro*, and shared autosomal regions with no synteny to (grey) of *D. rotundata* proto-sex chromosome. Significance values of Wilcoxon signed-rank tests ($p < 0.05$) are shown above plots for each region tested and Kruskal−Wallis across all regions is also shown ($p < 0.05$).

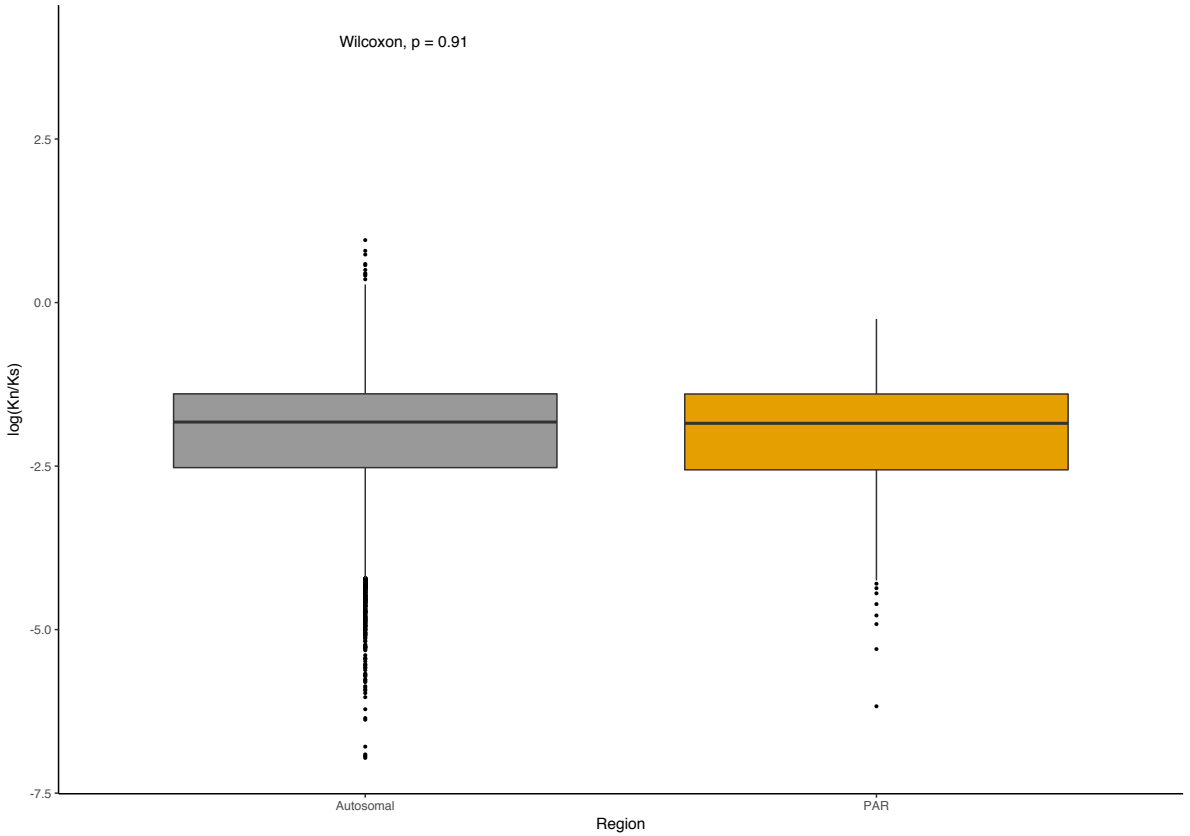

Figure 4.12: Boxplot showing log non-synonymous/synonymous mutation rate (Kn/Ks) of CDS synteny between *D. rotundata* and *D. tokoro*, within the PAR regions (orange) *D. rotundata* and shared autosomal regions with no synteny to (grey) of *D. tokoro* proto-sex chromosome. No CDS synteny was reported in the Significance values of Wilcoxon signed-rank tests ($p < 0.05$) are shown.

4.5 Discussion

Yam is a staple crop of great cultural and socioeconomic importance, belonging to the Monocotyledon *Dioscorea* genus of over 600 species, belonging to 10 major clades[62, 70, 71, 291]. The entire *Dioscorea* genus is also characterised by dioecy[40]. Our previous study on *D. rotundata* determined that sex determination in this species is most likely female heterogametic (male=ZZ, female=ZW) and identified a sex-linked DNA marker that can genetically distinguish male and female individuals[286].

In this chapter, we have validated previous findings of male heterogametic sex determination in *D. tokoro* and have observed a potential MSY on pseudo-chromosome 3; the putative proto-sex chromosome[64, 287]. We have generated a 370.87 Mb reference genome for the more basal Stenophora species, *D. tokoro*, with 29,471 protein coding genes annotated (Table 4.1). While smaller than *D. rotundata*, the difference in size is mainly due to an increase in repeat content in *D. rotundata* that may be biological or due to the need for additional long range information, such as the BAC-end sequencing used in *D. rotundata* or a long read technology like PacBio or Nanopore to capture repetitive regions that could be collapsed in the assembly.

Comparative analysis of gene orthogroups found 5,005 orthogroups to be conserved between *D. tokoro* and 25 other angiosperm species, with 174 orthogroups specific to *Dioscorea*(Figure 4.3). These *D. rotundata* specific orthogroups showed GO-term enrichment for 'hydrolase activity'. Phylogentic inference of 33 single-copy orthogroups between 22 of these species provided further support for the relatively basel origin of *Dioscorea* in the monocots, compared to the commelinids (Figure 4.3).

Investigation of previously observed expansions of bulb-type lectin (snowdrop lectin; B-lectin) gene in *D. rotundata* showed that these to either be lineage specific to *D. rotundata* or lost in *D. tokoro* (Table 7.30). As the biggest expansions of these B-lectin genes was found to have high sequence similarity to the *Dioscorea batatas* tuber lectin DB1 (accession number AB178475), and GO-term enrichment associated with defence, it is plausible these expansions are adaptations to pest defence of the tuberous tissues in *D. rotundata*, that may otherwise absent in the predominantly rhizomatous Stenophora clade of *D. tokoro* [62, 281, 282, 307].

When investigating the syntenic relationship of *D. tokoro* and *D. rotundata*, large regions of synteny were observed with striking synteny conserved between the pseudo-autosome of *D. rotundata* and *D. tokoro*, but no syntenic conservation of the *D. rotundata* sex determination region (Figure 4.4,4.10). Given the lack of synteny between the *D.*

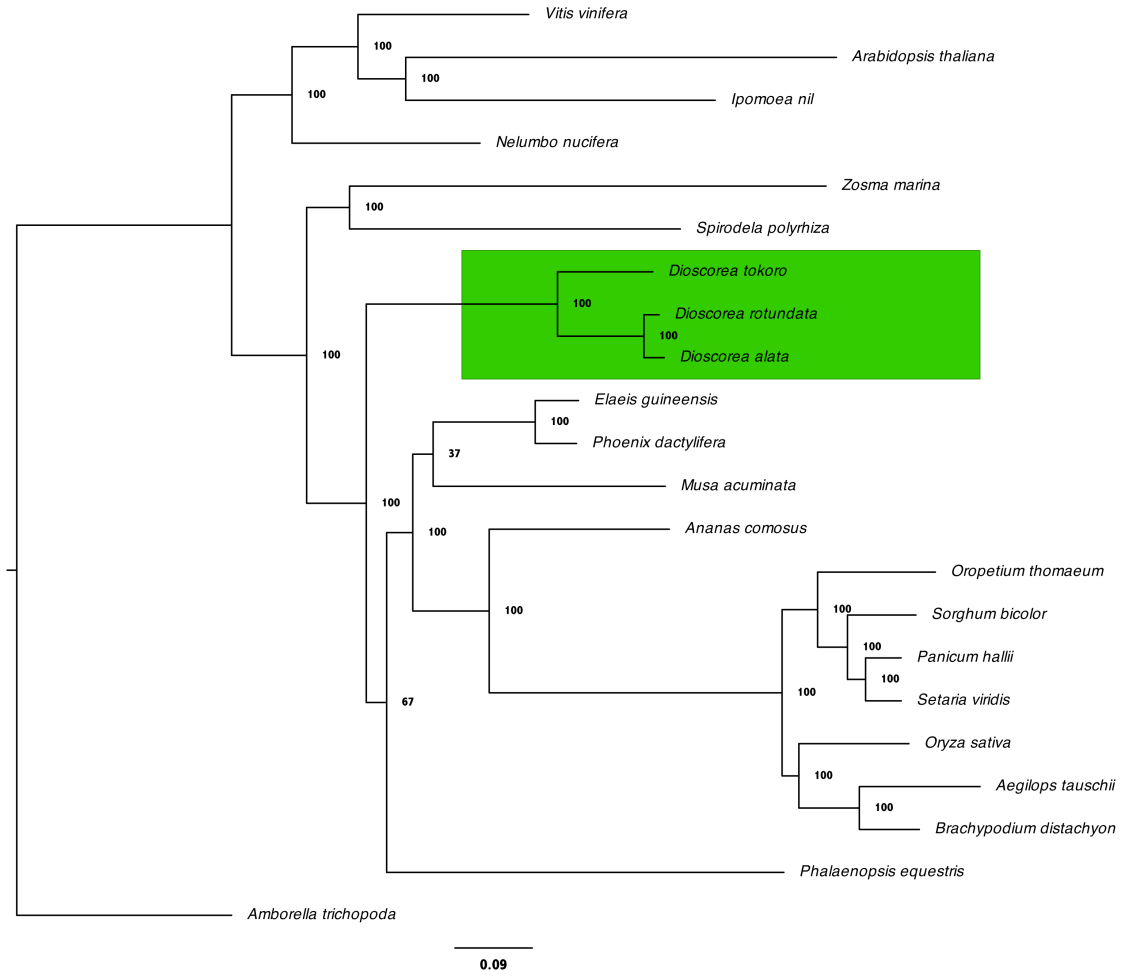

Figure 4.13: Phylogenetic relationships between *D. alata* and the 22 other angiosperm species from this study, based on alignment of 30 single-copy orthologs. **a.** Bipartiain tree generated by RAxML maximum likelihood analysis, with confidence intervals from 1,000 bootstrap resamplings shown. **b.** Split network generated by Spectre using the flat net joining method.

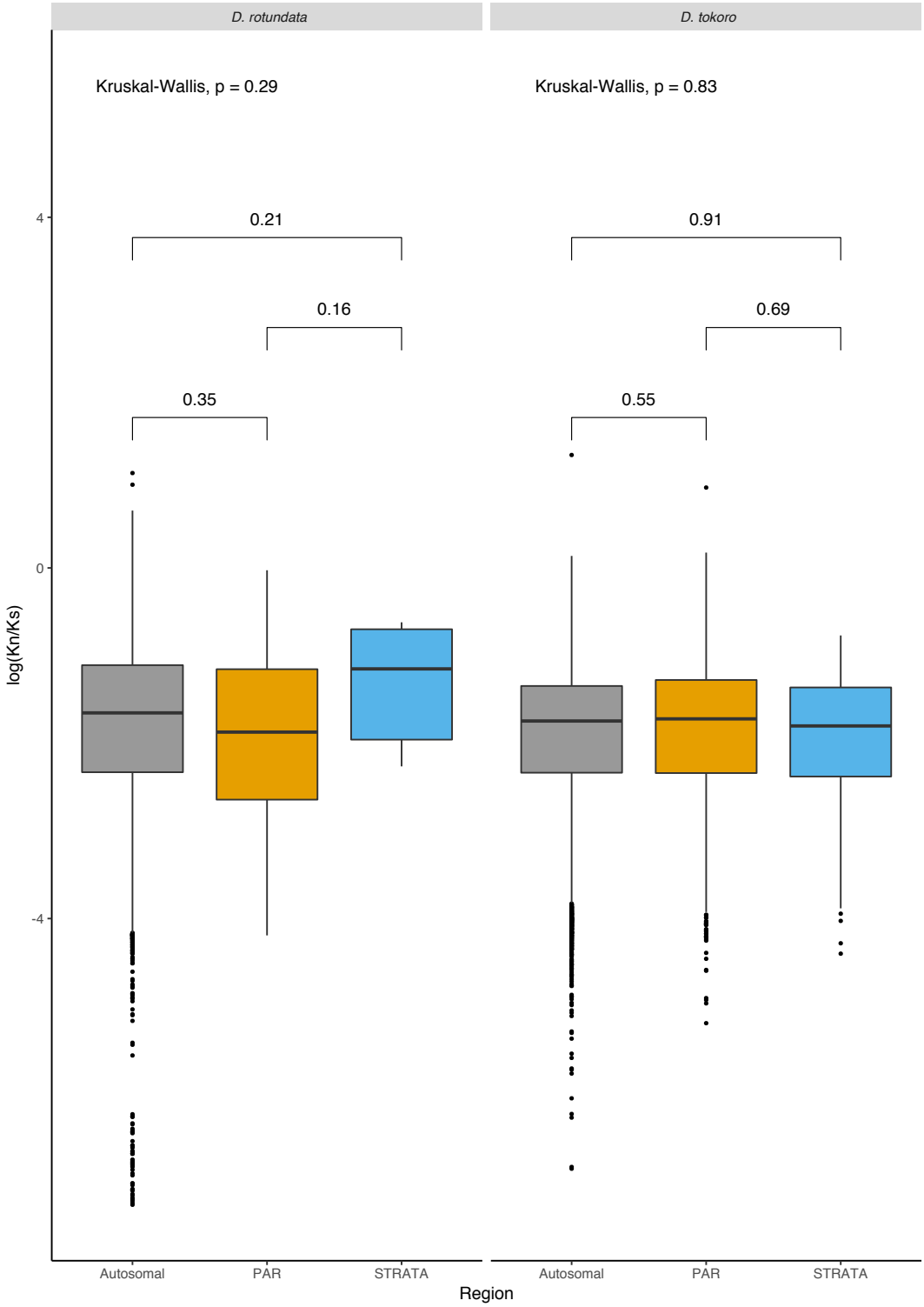

Figure 4.14: Boxplot showing log non-synonymous/synonymous mutation rate (Kn/Ks) of *D. alata* CDS synteny with *D. rotundata* and *D. tokoro*, within the putative sex determination loci (blue), PAR regions (orange) and autosomes (grey). Significance values of Wilcoxon signed-rank tests ($p < 0.05$) are shown above plots for each region tested and Kruskal–Wallis across all regions of each species also shown ($p < 0.05$).

rotundata FSW and *D. tokoro*, and autosomal placement of *D. tokoro* MSY synteny in *D. rotundata*, I propose that sex determination in *D. tokoro* evolved after the divergence of Stenophora and Enantiophyllum clades.

Based on the lower repeat and relatively unchanged coding coverage in *D. tokoro*, compared to *D. rotundata*, it would be useful for further studies to investigate pseudogene content and the age of the pseudo-autosomes and evolutionary strata within *Dioscorea* to determine if *D. rotundata* is now in the 'shrinkage phase' of sex chromosome evolution, while *D. tokoro* may be in the earlier 'expansion phase' of it's own independent evolutionary trajectory[36, 60, 340, 344–346]. As regardless of divergence time, the rate of evolution of sex chromosomes is not universal[29].

In order to investigate further the evolutionary rate of the sex determination loci between the two species, I also generated a first-pass assembly of the Enantiophyllum species, *D. alata*. Unlike *D. tokoro*, there was a degree of synteny observed between *D. alata* and the *D. rotundata* FSW, further study with additional Enantiophyllum species and an improved *D. alata* will be required to investigate if the FSW represents the ancestral state of sex in Enantiophyllum and if this region controls sex determination in the clade. This would also allow us to gain a better understanding on sex determination in *D. rotundata* and *D. alata*, as no putative sterility genes have been identified in *D. rotundata* yet and the sex determination of *D. alata* is still unknown. Synteny conservation to the MSY of *D. tokoro* was observed in *D. alata*, as well as *D. rotundata*, supports the proposal of this sex determination system evolving prior to formation of the Enantiophyllum clade.

Investigation of the branch specific evolutionary rate of *D. rotundata* and *D. tokoro* proto-sex chromosomes, compared to their respective autosomes, showed these to have significantly lower (Wilcoxon signed-rank test; $p < 0.05$) evolutionary rate in all branches. This result maybe consistent with recently emerging sex determination, whereby the proto-sex chromosomes do not have wide spread recombination of suppression that would impact selection and recombination rate.

However, these calculations were based on the full lengths of genes and could therefore be masking codon specific changes in dN/dS. Further studies using more species from other clades and outgroups from different genera of *Dioscorea*, would be needed to study the ancestral state of sex in the genus and to further investigate at what point transitions in sex determination occurred.

Due to the nature of dioecy to evolve through gynodioecy, rather than androdioecy, the presence of both XY and ZW sex determination in *Dioscorea* suggests a transition

in sex determination from XY to ZW that is supported through the synteny observed between *D. tokoro* and *D. rotundata* sex determination loci[23]. As the majority of other species in the genus are thought to be dioecious, with no validated heteromorphic sex chromosomes reported to date, *Dioscorea* represents an exciting opportunity to study the mechanisms behind the diversity and evolutionary drivers of sex. By studying the genome of additional yam species, we could use the pre-existing reference genomes to investigate the ancestral pseudo-autosomal state of the Enantiophyllum clade and even extend this to the genus as a whole, with the availability of an appropriate outgroup[238, 292, 334, 335].

Additional future work could also be aimed at developing a capture test to study the evolution and population structure of sex in *Dioscorea* species showing geographic cline, such as *D. tokoro*, which has been observed to have geographical variation in allelic frequency across Japan[67]. Making *Dioscorea* an exciting model to improve our understanding of sex evolution and diversity of sex determination in angiosperms, and the broader consequences of this, with regards to recombination and speciation[36].

Chapter 5

General conclusions

The advent of NGS has brought with it many opportunities that were previously not feasible, allowing generation of references for many neglected species, in particular orphan crops. Use of NGS has also been fundamental in improving our understanding of sex determination and sex evolution.

With my iCASE industrial partner, Eagle Genomics Ltd, I've shown the potential of advances in NGS to produce vastly improved genomics references compared to what a possible eight years ago when the first CHO genome was published[178]. Through generation of a new reference for CHO, specifically the Horizon Discovery CHO-K1 GS cell line called 'CHOK1GS_HD', by hybrid assembly using short and long reads, and genome mapping. I observed this improved reference to have a higher N50, more comprehensive annotation, vastly improved resolution of synteny to related species and further validated the GS knock-out of the cell line[178, 179]. Through investigation of the gene orthology I observed significant enrichment of orthogroups associated with olfactory receptors, that while maybe biologically relevant to the host or cell line sequenced, could be indicative of issues in previous reference genomes in assembling this highly duplicated family of genes[219, 231]. While this indicates improvements through the use of Chicago libraries and Bionano to resolve these copy-number errors, the reference could still benefit from long read technologies, such as those developed by Oxford Nanopore[100] and PacBio[181], to further resolve gaps and repeated regions. Furthermore, the lack of chromosome assignments and pseudo-chromosome assembly of the scaffolds are an issue for studies looking at large scale structural changes or collinearity of whole chromosomes.

While un-shown attempts were made to bin scaffolds into chromosomes, and even reconstruct these based on synteny to *M. musculus*, differences in the expected order based

on previous chromosomes painting studies made this difficult. This discrepancy indicates that the cell line may have a different karyotype to what was previously reported[220–223]. This may not be unexpected, as the karyotype in CHO has been shown to be unstable and to vary within and between populations[185, 195]. This coupled with the previous evidence from literature highlight the importance of generating high quality genomes for cell lines of industrial relevant.

Furthermore, inaccessibility of the sequences for the markers used in these studies further hindered my ability to assign scaffolds or build the pseudo-chromosomes, using these previously published findings. The genomic reference would therefore benefit from conformation capture techniques to join super-scaffolds into pseudo-chromosomes perhaps with linked-reads from Hi-C[94] or 10x Genomics[96]. In addition to the application of 10X Genomics sequencing methodology that would make it possible to resolve the different haplotypes[347].

After completion of the work associated with CHOK1GS_HD, a separate reference for *C. griseus* was published by Rudd, *et al*, 2018, using a similar assembly methodology to the workflow we used to generate CHOK1GS_HD[348]. In this study the sequencing was mainly carried out using paired-end and mate pair Illumina sequencing, with additional long read sequencing by PacBio. The authors performed multiple iterative assemblies, using ALLPATHS-LG[122] and SOAPdenovo2[349], that were, similarly to our work, merged to create a hybrid assembly with Metassembler[190]. As with my findings, this merged assembly also showed a large increase in contiguity. Annotation of the genome was carried out using a similar workflow to what I later used in *D. rotundata*, using MAKER[161]. However, one of the key differences between our CHOK1GS_HD and the reference described by the authors, is the use of chromosome sorted-libraries, that used the previously mentioned chromosomes paints by Yang, *et al*, 2000, to assign scaffolds to chromosomes[220]. Given the public availability of the author's reference, it would be plausible to use this with CHOK1GS_HD to assign chromosomes to CHOK1GS_HD scaffolds based on synteny to the host genome. This would also be useful for investigating potentially the potentially novel karyotypic difference that I observed, based on unexpected synteny with *M. musculus*.

In comparison to the final reference described by the authors, CHOK1GS_HD shows a higher N50 of 62 Mb compared to 20 Mb, and the longest scaffold present is 224 Mb, compared to 80 Mb. However, the contiguity of the newly published assembly is higher than CHOK1GS_HD, with 1,829 scaffolds present, opposed to 8,265 in CHOK1GS_HD. The author's reference also presents more genes, 24,686 compared to 20,978 in CHOK1GS_HD,

while such variation is possibly a result of the difference in assembly quality and annotation methods used, leading to multiple partial genes that could be biologically relevant as CHO-K1 may have selectively lost genes after immortalisation.

An updated comparison of the cell line and host, through gene orthology analysis and synteny of CHOK1GS_HD and this new *C. griseus* may provide more biologically relevant insight into the cell line. Regardless, my work shows that this workflow can successfully improve upon previously published references and can readily be applied to other perhaps unreferenced species, such as orphan crops. Highlighting the striking advances in NGS over the last couple of decades. This work also validated the use of quality control techniques and comparative analysis pipelines that I took forward into the rest of my thesis towards building genomic resources in *Dioscorea*.

Yam (*Dioscorea*) is a staple crop of great cultural and socioeconomic significance to Africa, the Americas, the Caribbean, South Pacific and Asia. Belonging to the genus Dioscoreaceae, with over 600 species, yams are primarily cultivated in West and Central Africa, where they are an important staple root crop[350]. In 2016, 65 Mt of yams were produced globally, with a gross product value of over 16 billion USD, of which Japan produced 159,800 t of yam worth 42.3 million USD[72]. Yams also act as major producers of steroid precursors and other compounds of pharmaceutical and industrial value[238, 350]. Examples of these compounds include diosgenin, that has an estimated market value of about $500 million USD, and shikimic acid, the base material for antiviral drug Tamiflu, which has been observed in yam at levels similar to the current crop (*Illicium verum*) used for the production of shikimic acid[238]. Additionally, the entire genus is characterised by dioecy, the presence of separate male and female plants, which is a rare trait found in only 5-6% of angiosperms and is thought to be synapomorphic[40, 350]. Despite this, not many genomic resources for yam existed before our work, and there was little known about sex determination and the evolution of sex in this genus. Hindering previous efforts to breed improved cultivars and investigate high value compounds in this important orphan crop[238].

Through our collaboration with the Terauchi Group, IBRC, on *D. rotundata*, the predominantly most cultivated species, we generated a reference genome for this species. As part of this, I carried out quality control of the assembly and performed MAKER whole genome annotation to generate this genomic resource for *D. rotundata*. The genome was found to have a total length of 594 Mb, that was close to the expected size by flow-cytometry, and 26,198 gene models were generated from this. I investigated the evolutionary history of *Dioscorea* in the monocots through comparison with related

species that had available reference sequences and found it to be evolutionary distant, sitting towards the base of the monocots. I also observed an expansion of B-lectin genes that potentially play a role in defence of the tuber.

Through the use of this reference, our collaborators were able to determined that this species is most likely female heterogametic (male=ZZ, female=ZW) and identified a sex-linked DNA marker that can distinguish genotypical male and female individuals[307]. Allowing the development of a marker for sex identification of *D. rotundata* at the seedling stage, which can be used to accelerate breeding programs for improvement of this important staple crop. As well as generating the first reference sequence publicly available for this genus on the Ensembl Plants platform[187].

The ZW sex system of *D. rotundata* is particularly interesting as the related species, *D. tokoro*, has been observed to be male heterogametic (male=XY, female=XX)[64]. Which through another collaboration, we have built a reference genome that is smaller than *D. rotundata* at 370 Mb, although k-mer analysis and flow-cytometry of *D. tokoro* predict this species to have a smaller genome, potentially due to lower repeat content than *D. rotundata*. Through use of an updated annotation workflow, I predicted 29,471 gene models in this species. While this is more than *D. rotundata* and maybe biologically significant, it's possible that the switch from MAKER to EVM and use of additional evidence that was not previously available has allowed us to identify additional genes that may not have been validated in the *D. rotundata* annotation. Our annotation would however likely gain from the availability of more comprehensive transcriptome data sets across development time points of more tissues, as well as the application of long range RNA-seq strategies, such as ISO-seq, that would lead to more accurate gene models and capture of splice variants[351].

By comparative genomic and phylogenetic analyses with more closely related species than previously used, I again confirmed the position of *Dioscorea* within the monocots. Investigation of the previously observed B-lectin expansion in *D. rotundata* found this to be absent in *D. tokoro*. As these B-lectins had high similarity to tuber specific genes, it's possible that these are an adaptation of *D. rotundata* to defence of its tuber, and the rhizomatous nature of *D. tokoro* may explain their absence. Furthermore, an expansion of novel resistance genes in *D. rotundata*, using the reference genome, have also been observed (unpublished; E. Baggs, Earlham Institute/UC Berkeley). While not explored in my thesis, Burkill, 1960, proposed that Dioscoreales were rhizomatous when they split from the Liliales and that adaptation to the harsher environments of the tropics and subtropics drove the evolution of tubers from these, around the time of the pangaean rift

when Dioscoreales began spreading west out of Asia[65]. Further work into the tuber specific genes and expression of *D. rotundata* and other branching clade B species with the rhizome specific genes and related expression in *D. tokoro* would be interesting for exploring the evolution of tubers. As well as the discovery of genes of importance in the development of tubers, that could have direct application to improving tuber yield and stress resistance in not just *D. rotundata*, but other tuber bearing crops too, such as potato.

Through the use of QTL-seq we were able to validate previous findings of male heterogametic sex determination in *D. tokoro* and also to locate a potential MSY region for this species[64, 287]. Moreover, with our collaborators we have identified a potential sex determination gene, Dt19188, that segregates strongly between male and female individuals (personal communication, S. Natusume, IBRC). This gene has been found to be expressed in the flowers and rhizome of male individuals, and encodes a small RNA that has orthology to aof-MIR2275c, a small RNA that has been associated with sex determination in asparagus[60]. It is fascinating to observe this, as it could indicate a potentially similar sex determination system to that seen in persimmon, whereby sex is controlled through epigenetic factors[27, 352]. The role of epigenetics in sex has been under-explored in plants and this finding provides an exciting opportunity to investigate the epigenetics of sex in *Dioscorea*. This should also be considered in related species, as it maybe indicative of sex in the Stenophora clade or potentially throughout the genus. Future studies into the role of epigenetics and sex determination could take the form of combined bisulphite and small RNA sequencing, to investigate differential methylation between developmental stages of stamen and carpal development[353]. Such an approach would not only provide insight into the regulation of sex determination and associated pathways, but could also be used to identify potential sex determination genes. A recent bisulphite study in *Populus balsamifera* found that a single candidate gene in the sex determination region, PbRR9, is potentially a master regulator of sex in poplar and also can be used to determine the sex of an individual based on the methylation pattern of the gene alone[354]. Further to this, through the use of third generation sequencing with PacBio and/or Nanopore, it is now possible to directly detect base modifications without the need for chemical treatments that potentially introduce biases and also are limited in the scope of the modifications that can be detected[355].

It is particularly exciting to observe both ZW and XY sex determination systems in *Dioscorea*, as this indicates turnover of sex determination systems within the genus. Given that dioecy is most likely to evolve through populations of coexisting females

and cosexuals, rather than males and hermaphrodites, into a population with XY sex determination; the most prevalent sex determination system in angiosperms[23]. We can hypothesise that the XY sex determination of *D. tokoro* may precede that of ZW *D. rotundata*.

In order to further investigate the evolution of sex determination and evolutionary rate of this between *D. rotundata* and *D. tokoro*, I generated a first-pass assembly of the Enantiophyllum species, *D. alata*, as part of a collaboration with IITA. While the completeness of the assembly was not as high as *D. rotundata* or *D. tokoro*, it was of sufficient quality to begin exploring synteny and gene orthology between these species. I observed no synteny between the *D. rotundata* FSW and *D. tokoro*, conversely *D. tokoro* MSY showed autosomal synteny in *D. rotundata*, and both FSW and MSY had microsynteny to *D. alata*. Using all three species I investigated the branch specific evolutionary rate of *D. rotundata* and *D. tokoro* proto-sex chromosomes, compared to their respective autosomes and found these to have a significantly lower (Wilcoxon signed-rank test; $p < 0.05$) evolutionary rate in all branches.

From these results, unobserved heteromorphic sex chromosomes, and the repeat and gene density of the FSW and MSY of *D. rotundata* and *D. tokoro*, I propose that both species have early stage proto-sex chromosomes. Whereby the current sex determination systems present have recently been gained and that XY sex determination in *D. tokoro* has evolved after the divergence of Stenophora and Enantiophyllum clades (branching clades A and B) less than ~48.2 Mya[295]. Additionally, *D. alata* will soon have an improved publicly available genome sequence, that will have improved correctness, completeness and contiguity compared to my first pass assembly, through long read sequencing and genome mapping (Personal communication; Bhattacharjee R., International Institute of Tropical Agriculture, Nigeria). This improved assembly will be useful in investigating sex determination in *D. alata*, to determine if it shares the same ZW sex determination as *D. rotundata* and as a consequence potentially other Enantiophyllum species.

Until now, there has been little study to date comparing plant species in the same genus with different sex determination systems, making *Dioscorea* a unique opportunity to investigate the turnover of sex determination[23]. The lack of previous studies may be due to the, until recently, prohibitive nature of sequencing and assembling genomes of related dioecious species at a sufficient resolution for in-depth exploration of sex determination loci. This is in part due to the costs involved, expertise required and the need for assembly of repetitive regions that are indicative of evolutionary strata. However, with the increasing throughput and accuracy of NGS technologies, especially

those offering long-range information for correct assembly of large repeats, and steep decline in associated costs, it's now feasible for single labs to consider genus-wide surveys of sex determination. The approach I have demonstrated here, in our generation of three genomic references in the previously under explored *Dioscorea* through the combination of short and long-read technologies, and use of QTL-seq, could readily be applied to other species of *Dioscorea* or extended to other genuses of interest. Having more complete and accurate gene models, due to improvements in assembly, also opens up the use of community omics resources, such as Ensembl, to further investigate genes families of interest in the context of an ever growing number of nearby and distantly related individuals for improved evolutionary insight[154].

Further studies into the evolution of sex determination of *Dioscorea* should consider using additional related species, alongside *D. rotundata* and *D. tokoro*. These could potentially include, the improved *D. alata* reference, *D. bulbifera*, *D. polystachya*, *D. japonica*, *D. communis*, and an outgroup, *Tacca leontopetaloides*, that is an important food source to the tropics that is also widely grown ornamentally[341, 350].

The choice of these species would enable study of sex determination evolution and the ancestral state of sex in this important genus, representing a several of the major clades in *Dioscorea* and *T. leontopetaloides* (Figure 5.1). Syntenic and collinear comparisons between all species would highlight regions of micro and macro-synteny between sex determination loci and corresponding autosomes, that could reveal sex specific linages of the genus and their divergence from the ancestral state of sex in Dioscoreaceae. Furthermore, characterisation of differential expression of floral buds in these species' sex specific gene expression could lead to the identification of sex determination loci in *D. alata*, *D. bulbifera*, *D. polystachya*, *D. japonica* and *D. communis*, as well assisting with the development of genetic markers of direct interest to the breeding community. Additionally, with reference genomes and gene models of comparable quality to *D. rotundata* and *D. tokoro*, these expression studies could also be extended to knockout experiments to investigate the effects of loss of function of putative sex determination genes on the development of stamen and carpal in these species. Providing additional insight into the function of these genes and also their regulation.

In turn, improving our understanding of sex determination evolution in *Dioscorea* and plants as a whole, as well as the consequences of this on recombination and in driving speciation[36]. In terms of fundamental research, the findings of this thesis will enable further study into the phenotypic implications of dioecy and control of inflorescence.

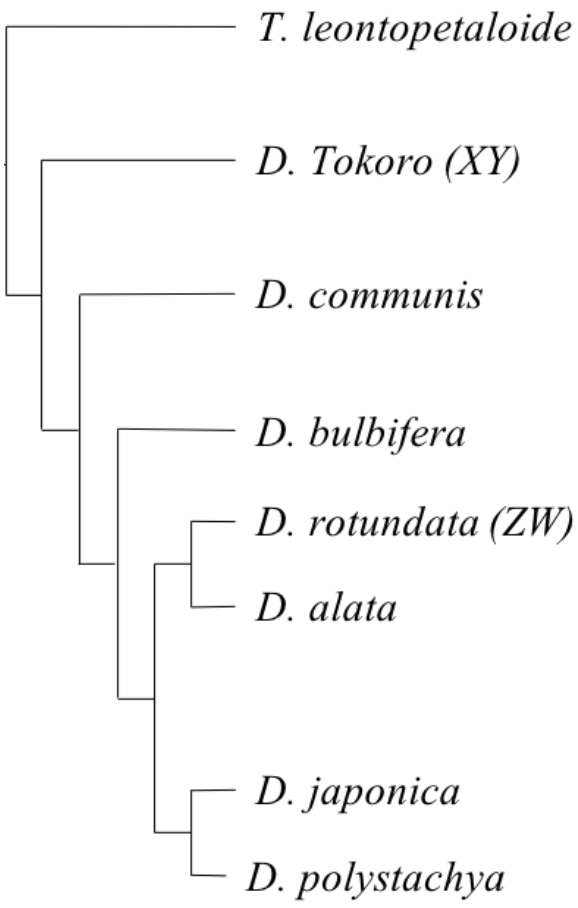

Figure 5.1: Phylogenic representation of the Dioscoreaceae species that could be used to explore the ancestral state of sex and sex determination in *Dioscorea*.

References

[1] Deborah Charlesworth. Evolution of plant breeding systems. *Current Biology*, 16(17):R726–R735, 2006.

[2] Spencer C. H. Barrett. Evolution of mating systems: outcrossing versus selfing. *In Losos JB, ed; The Princeton Guide to Evolution.*, pages 356–62, 2014.

[3] Kent E. Holsinger. Reproductive systems and evolution in vascular plants. *Proc Natl Acad Sci USA*, 97(13):7037–7042, 2000.

[4] John Maynard Smith and John Haigh. The hitch-hiking effect of a favourable gene. *Genetical Research*, 23(1):23–35, 1974.

[5] J. H. Gillespie. Genetic drift in an infinite population. the pseudohitchhiking model. *Genetics*, 155(2):909–919, 2000.

[6] Deborah Charlesworth and John H. Willis. The genetics of inbreeding depression. *Nature Reviews Genetics*, 10:783, 2009.

[7] H. J. Muller. The relation of recombination to mutational advance. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 1(1):2–9, 1964.

[8] Matthew Hartfield and Sylvain Glémin. Hitchhiking of deleterious alleles and the cost of adaptation in partially selfing species. *Genetics*, 196(1):281, 2014.

[9] Hideki Innan and Wolfgang Stephan. Distinguishing the hitchhiking and background selection models. *Genetics*, 165(4):2307, 2003.

[10] B. Charlesworth, M. T. Morgan, and D. Charlesworth. The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134(4):1289, 1993.

[11] Michael J. McDonald, Daniel P. Rice, and Michael M. Desai. Sex speeds adaptation by altering the dynamics of molecular evolution. *Nature*, 531:233, 2016.

[12] Matthew R. Goddard. Sex accelerates adaptation. *Nature*, 531:176, February 2016.

[13] W. G. Hill and Alan Robertson. The effect of linkage on limits to artificial selection. *Genetical Research*, 8(3):269–294, 1966.

[14] Paul N Pearson. *Red Queen Hypothesis*. American Cancer Society, 2001.

[15] Leigh Van Valen. A new evolutionary law. *Evol Theo*, 1:1–30, 1973.

[16] Levi T. Morran, Olivia G. Schmidt, Ian A. Gelarden, Raymond C. Parrish, and Curtis M. Lively. Running with the red queen: Host-parasite coevolution selects for biparental sex. *Science*, 333(6039):216, 2011.

[17] Nadia D. Singh, Dallas R. Criscoe, Shelly Skolfield, Kathryn P. Kohl, Erin S. Keebaugh, and Todd A. Schlenke. Fruit flies diversify their offspring in response to parasite infection. *Science*, 349(6249):747, 2015.

[18] Lesley A. Boyd, Christopher Ridout, Donal M. O'Sullivan, Jan E. Leach, and Hei Leung. Plantpathogen interactions: disease resistance in modern agriculture. *Trends in Genetics*, 29(4):233–240, 2013.

[19] Matthew A. Parker. Pathogens and sex in plants. *Evolutionary Ecology*, 8(5):560–584, 1994.

[20] Marc T. J. Johnson, Stacey D. Smith, and Mark D. Rausher. Plant sex and the evolution of plant defenses against herbivores. *Proc Natl Acad Sci USA*, 106(43):18079, 2009.

[21] Indrė Žliobaitė, Mikael Fortelius, and Nils C. Stenseth. Reconciling taxon senescence with the red queens hypothesis. *Nature*, 552:92, 2017.

[22] Jeffrey D. Karron, Christopher T. Ivey, Randall J. Mitchell, Michael R. Whitehead, Rod Peakall, and Andrea L. Case. New perspectives on the evolution of plant mating systems. *aob*, 109(3):493–503, 2011.

[23] Deborah Charlesworth. Plant sex chromosomes. *Annual Review of Plant Biology*, 67(1):397–420, 2016.

[24] Brian Charlesworth and Deborah Charlesworth. A model for the evolution of dioecy and gynodioecy. *The American Naturalist*, 112(988):975–997, 1978.

[25] S. Kumar, R. Kumari, and V. Sharma. Genetics of dioecy and causal sex chromosomes in plants. *Journal of Genetics*, 93(1):241–277, 2014.

[26] Rómulo Sobral, Helena G. Silva, Leonor Morais-Cecílio, and Maria M. R. Costa. The quest for molecular regulation underlying unisexual flower development. *Frontiers in Plant Science*, 7:160, 2016.

[27] Takashi Akagi, Isabelle M. Henry, Ryutaro Tao, and Luca Comai. A y-chromosome-encoded small rna acts as a sex determinant in persimmons. *Science*, 346(6209):646, 2014.

[28] Maria F. Torres, Lisa S. Mathew, Ikhlak Ahmed, Iman K. Al-Azwani, Robert Krueger, Diego Rivera, Yasmin A. Mohamoud, Andrew G. Clark, Karsten Suhre, and Joel A. Malek. Genus-wide sequencing supports a two-locus model for sex-determination in phoenix. *bioRxiv*, 2018. doi: https://doi.org/10.1101/245514.

[29] Doris Bachtrog, Judith E. Mank, Catherine L. Peichel, Mark Kirkpatrick, Sarah P. Otto, Tia-Lynn Ashman, Matthew W. Hahn, Jun Kitano, Itay Mayrose, Ray Ming, Nicolas Perrin, Laura Ross, Nicole Valenzuela, Jana C. Vamosi, and Consortium The Tree of Sex. Sex determination: Why so many ways of doing it? *PLOS Biology*, 12(7):e1001899, 2014.

[30] Hèléne Adam, Myriam Collin, Frèdèrique Richaud, Thierry Beulè, David Cros, Alphonse Omorè, Leifi Nodichao, Bruno Nouy, and James W. Tregear. Environmental regulation of sex determination in oil palm: current knowledge and insights from other species. *Annals of Botany*, 108(8):1529–1537, 2011.

[31] Masaki Shimamura. Marchantia polymorpha : Taxonomy, phylogeny and morphology of a model system. *Plant and Cell Physiology*, 57(2):230–256, 2016.

[32] Deborah Charlesworth. Plant sex chromosome evolution. *Journal of Experimental Botany*, 64(2):405–420, 2013.

[33] Ray Ming, Jianping Wang, Paul H. Moore, and Andrew H. Paterson. Sex chromosomes in flowering plants. *American Journal of Botany*, 94(2):141–150, 2007.

[34] Ray Ming, Abdelhafid Bendahmane, and Susanne S. Renner. Sex chromosomes in land plants. *Annu. Rev. Plant Biol.*, 62(1):485–514, 2011.

[35] Deborah Charlesworth. Plant contributions to our understanding of sex chromosome evolution. *New Phytologist*, 208(1):52–65, 2015.

[36] Alison E. Wright, Rebecca Dean, Fabian Zimmer, and Judith E. Mank. How to make a sex chromosome. *Nature Communications*, 7:12087, 2016.

[37] Hans Ellegren. Sex-chromosome evolution: recent progress and the influence of male and female heterogamety. *Nature Reviews Genetics*, 12:157, 2011.

[38] Roman Hobza, Radim Cegan, Wojciech Jesionek, Eduard Kejnovsky, Boris Vyskot, and Zdenek Kubat. Impact of repetitive elements on the y chromosome formation in plants. *Genes*, 8(11):302, 2017.

[39] Aleksandra Grabowska-Joachimiak and Andrzej Joachimiak. C-banded karyotypes of two silene species with heteromorphic sex chromosomes. *Genome*, 45(2):243–252, 2002.

[40] S. S. Renner. The relative and absolute frequencies of angiosperm sexual systems: dioecy, monoecy, gynodioecy, and an updated online database. *Am J Bot*, 101(10):1588–1596, 2014.

[41] Yusuke Kazama, Kotaro Ishii, Wataru Aonuma, Tokihiro Ikeda, Hiroki Kawamoto, Ayako Koizumi, Dmitry A. Filatov, Margarita Chibalina, Roberta Bergero, Deborah Charlesworth, Tomoko Abe, and Shigeyuki Kawano. A new physical mapping approach refines the sex-determining gene positions on the *Silene latifolia* y-chromosome. *Scientific Reports*, 6:18917, 2016.

[42] Marc Krasovec, Michael Chester, Kate Ridout, and Dmitry A. Filatov. The mutation rate and the age of the sex chromosomes in *Silene latifolia*. *Current Biology*, 28(11):1832–1838.e4, 2018.

[43] Nicolas Blavet, Hana Blavet, Radim Čegan, Niklaus Zemp, Jana Zdanska, Bohuslav Janoušek, Roman Hobza, and Alex Widmer. Comparative analysis of a plant pseudoautosomal region (par) in *Silene latifolia* with the corresponding *S. vulgaris* autosome. *BMC Genomics*, 13(1):226, 2012.

[44] Jennifer F. Hughes and David C. Page. The biology and evolution of mammalian y chromosomes. *Annu. Rev. Genet.*, 49(1):507–527, 2015.

[45] Yasuyuki Onodera, Itaru Yonaha, Satoshi Niikura, Seishi Yamazaki, and Tetsuo Mikami. Monoecy and gynomonoecy in Spinacia oleracea l.: Morphological and genetic analyses. *Scientia Horticulturae*, 118(3):266–269, 2008.

[46] Tomohiro Kudoh, Mitsuhiko Takahashi, Takayuki Osabe, Atsushi Toyoda, Hideki Hirakawa, Yutaka Suzuki, Nobuko Ohmido, and Yasuyuki Onodera. Molecular insights into the non-recombining nature of the spinach male-determining region. *Molecular Genetics and Genomics*, 293(2):557–568, 2018.

[47] William H. Wadlington and Ray Ming. Development of an x-specific marker and identification of yy individuals in spinach. *Theoretical and Applied Genetics*, 131(9):1987–1994, 2018.

[48] Roberta Bergero, Suo Qiu, and Deborah Charlesworth. Gene loss from a plant sex chromosome system. *Current Biology*, 25(9):1234–1240, 2015.

[49] P. Pucholt, A.-C. Rönnberg-Wästljung, and S. Berlin. Single locus sex determination and female heterogamety in the basket willow (*Salix viminalis* l.). *Heredity*, 114:575, 2015.

[50] Matthew J. Aylott, E. Casella, I. Tubby, N. R. Street, P. Smith, and Gail Taylor. Yield and spatial supply of bioenergy poplar and willow short-rotation coppice in the uk. *New Phytologist*, 178(2):358–370, 2008.

[51] Xiaogang Dai, Quanjun Hu, Qingle Cai, Kai Feng, Ning Ye, Gerald A. Tuskan, Richard Milne, Yingnan Chen, Zhibing Wan, Zefu Wang, Wenchun Luo, Kun Wang, Dongshi Wan, Mingxiu Wang, Jun Wang, Jianquan Liu, and Tongming Yin. The willow genome and divergent evolution from poplar after the common genome duplication. *Cell Research*, 24:1274, 2014.

[52] G. A. Tuskan, S. DiFazio, S. Jansson, J. Bohlmann, I. Grigoriev, U. Hellsten, N. Putnam, S. Ralph, S. Rombauts, A. Salamov, J. Schein, L. Sterck, A. Aerts, R. R. Bhalerao, R. P. Bhalerao, D. Blaudez, W. Boerjan, A. Brun, A. Brunner, V. Busov, M. Campbell, J. Carlson, M. Chalot, J. Chapman, G.L. Chen, D. Cooper, P. M. Coutinho, J. Couturier, S. Covert, Q. Cronk, R. Cunningham, J. Davis, S. Degroeve, A. Déjardin, C. dePamphilis, J. Detter, B. Dirks, I. Dubchak, S. Duplessis, J. Ehlting, B. Ellis, K. Gendler, D. Goodstein, M. Gribskov, J. Grimwood, A. Groover, L. Gunter, B. Hamberger, B. Heinze, Y. Helariutta, B. Henrissat,

D. Holligan, R. Holt, W. Huang, N. IslamFaridi, S. Jones, M. Jones-Rhoades, R. Jorgensen, C. Joshi, J. Kangasjärvi, J. Karlsson, C. Kelleher, R. Kirkpatrick, M. Kirst, A. Kohler, U. Kalluri, F. Larimer, J. Leebens-Mack, J.-C. Leplé, P. Locascio, Y. Lou, S. Lucas, F. Martin, B. Montanini, C. Napoli, D. R. Nelson, C. Nelson, K. Nieminen, O. Nilsson, V. Pereda, G. Peter, R. Philippe, G. Pilate, A. Poliakov, J. Razumovskaya, P. Richardson, C. Rinaldi, K. Ritland, P. Rouzé, D. Ryaboy, J. Schmutz, J. Schrader, B. Segerman, H. Shin, A. Siddiqui, F. Sterky, A. Terry, C.J. Tsai, E. Uberbacher, P. Unneberg, J. Vahala, K. Wall, S. Wessler, G. Yang, T. Yin, C. Douglas, M. Marra, G. Sandberg, Y. Van de Peer, and D. Rokhsar. The genome of black cottonwood, *Populus trichocarpa* (torr. & gray). *Science*, 313(5793):1596, 2006.

[53] Jing Hou, Ning Ye, Defang Zhang, Yingnan Chen, Lecheng Fang, Xiaogang Dai, and Tongming Yin. Different autosomes evolved into sex chromosomes in the sister genera of *Salix* and *Populus*. *Scientific Reports*, 5:9076, 2015.

[54] Tongming Yin, Stephen P. DiFazio, Lee E. Gunter, Xinye Zhang, Michell M. Sewell, Scott A. Woolbright, Gery J. Allan, Collin T. Kelleher, Carl J. Douglas, Mingxiu Wang, and Gerald A. Tuskan. Genome structure and emerging evidence of an incipient sex chromosome in *Populus*. *Genome Research*, 18(3):422–430, 2008.

[55] Hanno Schaefer and Susanne S. Renner. Phylogenetic relationships in the order cucurbitales and a new classification of the gourd family (cucurbitaceae). *Taxon*, 60(1):122–138, 2011.

[56] Aretuza Sousa, Jörg Fuchs, and Susanne S. Renner. Cytogenetic comparison of heteromorphic and homomorphic sex chromosomes in coccinia (cucurbitaceae) points to sex chromosome turnover. *Chromosome Research*, 25(2):191–200, 2017.

[57] Takashi Akagi, Isabelle M. Henry, Haruka Ohtani, Takuya Morimoto, Kenji Beppu, Ikuo Kataoka, and Ryutaro Tao. A y-encoded suppressor of feminization arose via lineage-specific duplication of a cytokinin response regulator in kiwifruit. *Plant Cell*, 30(4):780, 2018.

[58] Shengxiong Huang, Jian Ding, Dejing Deng, Wei Tang, Honghe Sun, Dongyuan Liu, Lei Zhang, Xiangli Niu, Xia Zhang, Meng Meng, Jinde Yu, Jia Liu, Yi Han, Wei Shi, Danfeng Zhang, Shuqing Cao, Zhaojun Wei, Yongliang Cui, Yanhua Xia, Huaping Zeng, Kan Bao, Lin Lin, Ya Min, Hua Zhang, Min Miao, Xiaofeng

Tang, Yunye Zhu, Yuan Sui, Guangwei Li, Hanju Sun, Junyang Yue, Jiaqi Sun, Fangfang Liu, Liangqiang Zhou, Lin Lei, Xiaoqin Zheng, Ming Liu, Long Huang, Jun Song, Chunhua Xu, Jiewei Li, Kaiyu Ye, Silin Zhong, Bao-Rong Lu, Guanghua He, Fangming Xiao, Hui-Li Wang, Hongkun Zheng, Zhangjun Fei, and Yongsheng Liu. Draft genome of the kiwifruit Actinidia chinensis. *Nature Communications*, 4:2640, 2013.

[59] Qiong Zhang, Chunyan Liu, Yifei Liu, Robert VanBuren, Xiaohong Yao, Caihong Zhong, and Hongwen Huang. High-density interspecific genetic maps of kiwifruit and the identification of sex-specific markers. *DNA Research*, 22(5):367–375, 2015.

[60] Alex Harkess, Jinsong Zhou, Chunyan Xu, John E. Bowers, Ron Van der Hulst, Saravanaraj Ayyampalayam, Francesco Mercati, Paolo Riccardi, Michael R. McKain, Atul Kakrana, Haibao Tang, Jeremy Ray, John Groenendijk, Siwaret Arikit, Sandra M. Mathioni, Mayumi Nakano, Hongyan Shan, Alexa Telgmann-Rauber, Akira Kanno, Zhen Yue, Haixin Chen, Wenqi Li, Yanling Chen, Xiangyang Xu, Yueping Zhang, Shaochun Luo, Helong Chen, Jianming Gao, Zichao Mao, J. Chris Pires, Meizhong Luo, Dave Kudrna, Rod A. Wing, Blake C. Meyers, Kexian Yi, Hongzhi Kong, Pierre Lavrijsen, Francesco Sunseri, Agostino Falavigna, Yin Ye, James H. Leebens-Mack, and Guangyu Chen. The asparagus genome sheds light on the origin and evolution of a young y chromosome. *Nature Communications*, 8(1):1279, 2017.

[61] Mai Mitoma, Lei Zhang, Itaru Konno, Shunpei Imai, Satoru Motoki, and Akira Kanno. A new dna marker for sex identification in purple asparagus. *Euphytica*, 214(9):154, 2018.

[62] Paul Wilkin, Peter Schols, Mark W. Chase, Kongkanda Chayamarit, Carol A. Furness, Suzy Huysmans, Franck Rakotonasolo, Erik Smets, Chirdsak Thapyai, and Alan W. Meerow. A plastid gene phylogeny of the yam genus, dioscorea: Roots, fruits and madagascar. *Systematic Botany*, 30(4):736–749, 2005.

[63] F. W. Martin. Sex ratio and sex determination in *Dioscorea*. *J Heredity*, 57(3):95–99, 1966.

[64] Ryohei Terauchi and Günter Kahl. Mapping of the *Dioscorea tokoro* genome: Aflp markers linked to sex. *Genome*, 42(4):752–762, 1999.

[65] I. H. Burkill. The organography and the evolution of dioscoreaceae, the family of the yams. *Journal of the Linnean Society of London, Botany*, 56(367):319–412, September 1960.

[66] Mimi Li, Q.-Q. Yan, Xiaoqin Sun, Y.-M. Zhao, Y.-F. Zhou, and Y.-Y. Hang. A preliminary study on pollination biology of three species in *Dioscorea* (dioscoreaceae). *Life Science Journal*, 11:436–444, 2014.

[67] Terauchi Ryohei. Genetic diversity and population structure of *Dioscorea tokoro* makino, a dioecious climber. *Plant Species Biology*, 5(2):243–253, 1990.

[68] Lizabeth R. Caddick, Paul Wilkin, Paula J. Rudall, Terry A. J. Hedderson, and Mark W. Chase. Yams reclassified: A recircumscription of dioscoreaceae and dioscoreales. *Taxon*, 51(1):103–114, 2002.

[69] Mas Yamaguchi. Yam. In *World Vegetables: Principles, Production and Nutritive Values*, pages 139–147. Springer Netherlands, Dordrecht, 1983.

[70] Amit Kumar Srivastava, Thomas Gaiser, Heiko Paeth, and Frank Ewert. The impact of climate change on yam (*Dioscorea alata*) yield in the savanna zone of west africa. *Agriculture, Ecosystems & Environment*, 153(0):57–64, 2012.

[71] Z. K. Muthamia, A. B. Nyende, E. G. Mamati, M. E. Ferguson, and J. Wasilwa. Determination of ploidy among yam (*Dioscorea* spp.) landraces in kenya by flow cytometry. *African Journal of Biotechnology*, 13(3):394–402, 2014.

[72] Food and Agriculture Organization of the United Nations. http://www.fao.org/faostat/en/data. cited 14th Aug 2018.

[73] Wireko-Manu Faustina Dufie, Ibok Oduro, William Otoo Ellis, Robert Asiedu, and Bussie Maziya-Dixon. Potential health benefits of water yam (*Dioscorea alata*). *Food & function*, 4(10):1496–501, 2013.

[74] D. G. Coursey. The civilizations of the yam: Interrelationships of man and yams in africa and the indo-pacific region. *Archaeology & Physical Anthropology in Oceania*, 7(3):215–233, 1972.

[75] Edward S. Ayensu and D. G. Coursey. Guinea yams: The botany, ethnobotany, use and possible future of yams in west africa. *Economic Botany*, 26(4):301–318, 1972.

[76] Jude Ejikeme Obidiegwu and Emmanuel Matthew Akpabio. The geography of yam cultivation in southern nigeria: Exploring its social meanings and cultural functions. *Journal of Ethnic Foods*, 4(1):28–35, 2017.

[77] Achebe Chinua. Things fall apart. *Ch. Achebe*, pages 1–117, 1958.

[78] Satya S. Narina, Ramesh Buyyarapu, Kameswara R. Kottapalli, Alieu M. Sartie, Mohamed I. Ali, Asiedu Robert, Mignouna J. D. Hodeba, Brian L. Sayre, and Brian E. Scheffler. Generation and analysis of expressed sequence tags (ests) for marker development in yam (Dioscorea alata l.). *BMC genomics*, 12(1):100, 2011.

[79] Sara Goodwin, John D. McPherson, and W. Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17:333, 2016.

[80] Rasmus Nielsen, Joshua S. Paul, Anders Albrechtsen, and Yun S. Song. Genotype and snp calling from next-generation sequencing data. *Nat Rev Genet*, 12(6):443–451, 2011.

[81] Martin Trick, Nikolai Adamski, Sarah Mugford, Cong-Cong Jiang, Melanie Febrer, and Cristobal Uauy. Combining snp discovery from next-generation sequencing data with bulked segregant analysis (bsa) to fine-map genes in polyploid wheat. *BMC Plant Biology*, 12(1):14, 2012.

[82] Eman K. Al-Dous, Binu George, Maryam E. Al-Mahmoud, Moneera Y. Al-Jaber, Hao Wang, Yasmeen M. Salameh, Eman K. Al-Azwani, Srinivasa Chaluvadi, Ana C. Pontaroli, Jeremy DeBarry, Vincent Arondel, John Ohlrogge, Imad J. Saie, Khaled M. Suliman-Elmeer, Jeffrey L. Bennetzen, Robert R. Kruegger, and Joel A. Malek. De novo genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). *Nat Biotech*, 29(6):521–527, 2011.

[83] Chibuikem I. N. Unamba, Akshay Nag, and Ram K. Sharma. Next generation sequencing technologies: The doorway to the unexplored genomics of non-model plants. *Frontiers in Plant Science*, 6:1074, 2015.

[84] Adessi C. Kawashima E. Mayer P. Mermod J.J. and Turcatti G. Methods of nucleic acid amplification and sequencing. pct patent application. *WO2000018957*, 2000.

[85] Keith R. Mitchelson. *New high throughput technologies for DNA sequencing and genomics*, volume 2. Elsevier, 2011.

[86] Carl W. Fuller, Lyle R. Middendorf, Steven A. Benner, George M. Church, Timothy Harris, Xiaohua Huang, Stevan B. Jovanovich, John R. Nelson, Jeffery A. Schloss, David C. Schwartz, and Dmitri V. Vezenov. The challenges of sequencing by synthesis. *Nature Biotechnology*, 27:1013, 2009.

[87] H. P. J. Buermans and J. T. den Dunnen. Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1842(10):1932–1941, 2014.

[88] André E. Minoche, Juliane C. Dohm, and Heinz Himmelbauer. Evaluation of genomic high-throughput sequencing data generated on illumina hiseq and genome analyzer systems. *Genome Biology*, 12(11):R112, 2011.

[89] Douglas W. Fadrosh, Bing Ma, Pawel Gajer, Naomi Sengamalay, Sandra Ott, Rebecca M. Brotman, and Jacques Ravel. An improved dual-indexing approach for multiplexed 16s rrna gene sequencing on the illumina miseq platform. *Microbiome*, 2(1):6, 2014.

[90] Marcel Margulies, Michael Egholm, William E. Altman, Said Attiya, Joel S. Bader, Lisa A. Bemben, Jan Berka, Michael S. Braverman, Yi-Ju Chen, Zhoutao Chen, Scott B. Dewell, Lei Du, Joseph M. Fierro, Xavier V. Gomes, Brian C. Godwin, Wen He, Scott Helgesen, Chun He Ho, Gerard P. Irzyk, Szilveszter C. Jando, Maria L. I. Alenquer, Thomas P. Jarvie, Kshama B. Jirage, Jong-Bum Kim, James R. Knight, Janna R. Lanza, John H. Leamon, Steven M. Lefkowitz, Ming Lei, Jing Li, Kenton L. Lohman, Hong Lu, Vinod B. Makhijani, Keith E. McDade, Michael P. McKenna, Eugene W. Myers, Elizabeth Nickerson, John R. Nobile, Ramona Plant, Bernard P. Puc, Michael T. Ronan, George T. Roth, Gary J. Sarkis, Jan Fredrik Simons, John W. Simpson, Maithreyan Srinivasan, Karrie R. Tartaro, Alexander Tomasz, Kari A. Vogt, Greg A. Volkmer, Shally H. Wang, Yong Wang, Michael P. Weiner, Pengguang Yu, Richard F. Begley, and Jonathan M. Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437:376, 2005.

[91] Jonathan M. Rothberg, Wolfgang Hinz, Todd M. Rearick, Jonathan Schultz, William Mileski, Mel Davey, John H. Leamon, Kim Johnson, Mark J. Milgrew, Matthew Edwards, Jeremy Hoon, Jan F. Simons, David Marran, Jason W. Myers, John F. Davidson, Annika Branting, John R. Nobile, Bernard P. Puc, David Light, Travis A. Clark, Martin Huber, Jeffrey T. Branciforte, Isaac B. Stoner, Simon E. Cawley, Michael Lyons, Yutao Fu, Nils Homer, Marina Sedova, Xin

Miao, Brian Reed, Jeffrey Sabina, Erika Feierstein, Michelle Schorn, Mohammad Alanjary, Eileen Dimalanta, Devin Dressman, Rachel Kasinskas, Tanya Sokolsky, Jacqueline A. Fidanza, Eugeni Namsaraev, Kevin J. McKernan, Alan Williams, G. Thomas Roth, and James Bustillo. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475:348–352, 2011.

[92] Weixing Feng, Sen Zhao, Dingkai Xue, Fengfei Song, Ziwei Li, Duojiao Chen, Bo He, Yangyang Hao, Yadong Wang, and Yunlong Liu. Improving alignment accuracy on homopolymer regions for semiconductor-based sequencing technologies. *BMC Genomics*, 17(7):521, 2016.

[93] Steven R. Head, H. Kiyomi Komori, Sarah A. LaMere, Thomas Whisenant, Filip Van Nieuwerburgh, Daniel R. Salomon, and Phillip Ordoukhanian. Library construction for next-generation sequencing: Overviews and challenges. *BioTechniques*, 56(2):61–77, 2015.

[94] Nicholas H. Putnam, Brendan L. O'Connell, Jonathan C. Stites, Brandon J. Rice, Marco Blanchette, Robert Calef, Christopher J. Troll, Andrew Fields, Paul D. Hartley, Charles W. Sugnet, David Haussler, Daniel S. Rokhsar, and Richard E. Green. Chromosome-scale shotgun assembly using an *in vitro* method for long-range linkage. *Genome Research*, 26(3):342–350, 2016.

[95] Wen-Biao Jiao, Gonzalo Garcia Accinelli, Benjamin Hartwig, Christiane Kiefer, David Baker, Edouard Severing, Eva-Maria Willing, Mathieu Piednoel, Stefan Woetzel, Eva Madrid-Herrero, Bruno Huettel, Ulrike Hümann, Richard Reinhard, Marcus A. Koch, Daniel Swan, Bernardo Clavijo, George Coupland, and Korbinian Schneeberger. Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. *Genome Research*, 27(5):778–786, 2017.

[96] Grace X. Y. Zheng, Billy T. Lau, Michael Schnall-Levin, Mirna Jarosz, John M. Bell, Christopher M. Hindson, Sofia Kyriazopoulou-Panagiotopoulou, Donald A. Masquelier, Landon Merrill, Jessica M. Terry, Patrice A. Mudivarti, Paul W. Wyatt, Rajiv Bharadwaj, Anthony J. Makarewicz, Yuan Li, Phillip Belgrader, Andrew D. Price, Adam J. Lowe, Patrick Marks, Gerard M. Vurens, Paul Hardenbol, Luz Montesclaros, Melissa Luo, Lawrence Greenfield, Alexander Wong, David E. Birch, Steven W. Short, Keith P. Bjornson, Pranav Patel, Erik S. Hopmans, Christina Wood, Sukhvinder Kaur, Glenn K. Lockwood, David Stafford, Joshua P.

Delaney, Indira Wu, Heather S. Ordonez, Susan M. Grimes, Stephanie Greer, Josephine Y. Lee, Kamila Belhocine, Kristina M. Giorda, William H. Heaton, Geoffrey P. McDermott, Zachary W. Bent, Francesca Meschi, Nikola O. Kondov, Ryan Wilson, Jorge A. Bernate, Shawn Gauby, Alex Kindwall, Clara Bermejo, Adrian N. Fehr, Adrian Chan, Serge Saxonov, Kevin D. Ness, Benjamin J. Hindson, and Hanlee P. Ji. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature Biotechnology*, 34:303, 2016.

[97] Jacob O. Kitzman. Haplotypes drop by drop. *Nature Biotechnology*, 34:296, 2016.

[98] Anthony Rhoads and Kin Fai Au. Pacbio sequencing and its applications. *Genomics, Proteomics & Bioinformatics*, 13(5):278–289, 2015.

[99] V. Dominguez Del Angel, E. Hjerde, L. Sterck, S. Capella-Gutierrez, C. Notredame, O. Vinnere Pettersson, J. Amselem, L. Bouri, S. Bocs, C. Klopp, J. F. Gibrat, A. Vlasova, B. L. Leskosek, L. Soler, M. Binzer-Panchal, and H. P. Y. Lantz. Ten steps to get started in genome assembly and [version annotation, 1 and referees: 2 approved]. *F1000 Research*, 2018.

[100] Miten Jain, Hugh E. Olsen, Benedict Paten, and Mark Akeson. The oxford nanopore minion: delivery of nanopore sequencing to the genomics community. *Genome Biology*, 17(1):239, 2016.

[101] M. Jain, J. R. Tyson, M. Loose, C. L. C. Ip, D. A. Eccles, J. O'Grady, S. Malla, R. M. Leggett, O. Wallerman, H. J. Jansen, V. Zalunin, Brown Birney, E., B. L., T. P. Snutch, H. E. Olsen, MinION Analysis, and Reference Consortium. Minion analysis and reference consortium: Phase 2 data release and analysis of r9.0 chemistry [version 1; referees: 1 approved, 2 approved with reservations]. *F1000Research*, 6:760, 2017.

[102] Franka J. Rang, Wigard P. Kloosterman, and Jeroen de Ridder. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biology*, 19(1):90, 2018.

[103] Shiguo Zhou, Jill Herschleb, and David C. Schwartz. Chapter 9 a single molecule system for whole genome analysis. In Keith R. Mitchelson, editor, *Perspectives in Bioanalysis*, volume 2, pages 265–300. Elsevier, 2007.

[104] Somes K. Das, Michael D. Austin, Matthew C. Akana, Paru Deshpande, Han Cao, and Ming Xiao. Single molecule linear analysis of dna in nano-channel labeled with sequence specific fluorescent probes. *Nucleic Acids Research*, 38(18):e177–e177, 2010.

[105] Alexandra C. Nica and Emmanouil T. Dermitzakis. Expression quantitative trait loci: present and future. *Philos Trans R Soc Lond B Biol Sci*, 368(1620), 2013.

[106] Rajeev K. Varshney, Ryohei Terauchi, and Susan R. McCouch. Harvesting the promising fruits of genomics: Applying genome sequencing technologies to crop breeding. *PLOS Biology*, 12(6):e1001883, 2014.

[107] Emma M. Quinn, Paul Cormican, Elaine M. Kenny, Matthew Hill, Richard Anney, Michael Gill, Aiden P. Corvin, and Derek W. Morris. Development of strategies for snp detection in rna-seq data: Application to lymphoblastoid cell lines and evaluation using 1000 genomes data. *PLoS ONE*, 8(3):e58815, 2013.

[108] Daniel R. Garalde, Elizabeth A. Snell, Daniel Jachimowicz, Botond Sipos, Joseph H. Lloyd, Mark Bruce, Nadia Pantic, Tigist Admassu, Phillip James, Anthony Warland, Michael Jordan, Jonah Ciccone, Sabrina Serra, Jemma Keenan, Samuel Martin, Luke McNeill, E. Jayne Wallace, Lakmal Jayasinghe, Chris Wright, Javier Blasco, Stephen Young, Denise Brocklebank, Sissel Juul, James Clarke, Andrew J. Heron, and Daniel J. Turner. Highly parallel direct rna sequencing on an array of nanopores. *Nature Methods*, 15:201, 2018.

[109] Ewan Birney. A useful analogy i used this week: Sequencing, analysing and interpreting genomes is routine in the same way the us navy âroutinelyâ lands planes on aircraft carriers. it might happen regularly by well trained crew with the right equipment but it is not an easy thing to do. www.twitter.com/ewanbirney/status/1040144488948281344. Twitter Post, 2018.

[110] B. Ewing and P. Green. Base-calling of automated sequencer traces using phred. ii. error probabilities. *Genome Res*, 8, 1998.

[111] S. Andrews. *FastQC: A Quality Control Tool for High Throughput Sequence Data;*. 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

[112] Peter J. A. Cock, Christopher J. Fields, Naohisa Goto, Michael L. Heuer, and

Peter M. Rice. The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. *Nucleic Acids Research*, 38(6):1767–1771, 2010.

[113] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[114] A. Gordon and G. J. Hannon. Fastx-toolkit fastq/a short-reads pre-processing tools (unpublished). *(unpublished) http://hannonlab. cshl. edu/fastx_toolkit*, 2010. *Fastx − toolkit*.

[115] R. Ramirez-Gonzalez. Kontaminant, a k-mer based contamination screening and filtering tool. The Genome Analysis Centre, 2013.

[116] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

[117] Vivien Marx. The big challenges of big data. *Nature*, 498:255, 2013.

[118] E. W. Myers, G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo, and M. J. Flanigan. A whole-genome assembly of drosophila. *Science*, 287, 2000.

[119] Sergey Koren, Brian P. Walenz, Konstantin Berlin, Jason R. Miller, Nicholas H. Bergman, and Adam M. Phillippy. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, 27(5):722–736, 2017.

[120] Zhenyu Li, Yanxiang Chen, Desheng Mu, Jianying Yuan, Yujian Shi, Hao Zhang, Jun Gan, Nan Li, Xuesong Hu, Binghang Liu, Bicheng Yang, and Wei Fan. Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. *Briefings in Functional Genomics*, 11(1):25–37, 2012.

[121] Phillip E. C. Compeau, Pavel A. Pevzner, and Glenn Tesler. How to apply de bruijn graphs to genome assembly. *Nature Biotechnology*, 29:987, 2011.

[122] Jonathan Butler, Iain MacCallum, Michael Kleber, Ilya A. Shlyakhter, Matthew K. Belmonte, Eric S. Lander, Chad Nusbaum, and David B. Jaffe. Allpaths: De novo assembly of whole-genome shotgun microreads. *Genome Research*, 18(5):810–820, 2008.

[123] A. Mi Kim, Jae-Sung Rhee, H. Tae Kim, S. Jung Lee, Ah-Young Choi, Beom-Soon Choi, Ik-Young Choi, and C. Young Sohn. Evaluation of discovar de novo using a mosquito sample for cost-effective short-read genome assembly. *BMC Genomics*, 17(1):187, 2016. 1471-2164.

[124] Kristoffer Sahlin, Rayan Chikhi, and Lars Arvestad. Assembly scaffolding with pe-contaminated mate-pair libraries. *Bioinformatics*, 32(13):1925–1932, 2016.

[125] S. Gnerre, I. Maccallum, D. Przybylski, F. J. Ribeiro, J. N. Burton, and B. J. Walker. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci USA*, 108, 2011.

[126] M. Boetzer, C. V. Henkel, H. J. Jansen, D. Butler, and W. Pirovano. Scaffolding pre-assembled contigs using sspace. *Bioinformatics*, 27, 2011.

[127] Chen-Shan Chin, Paul Peluso, Fritz J. Sedlazeck, Maria Nattestad, Gregory T. Concepcion, Alicia Clum, Christopher Dunn, Ronan O'Malley, Rosa Figueroa-Balderas, Abraham Morales-Cruz, Grant R. Cramer, Massimo Delledonne, Chongyuan Luo, Joseph R. Ecker, Dario Cantu, David R. Rank, and Michael C. Schatz. Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*, 13:1050, 2016.

[128] Fritz J. Sedlazeck, Hayan Lee, Charlotte A. Darby, and Michael C. Schatz. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nature Reviews Genetics*, 19(6):329–346, 2018.

[129] D. Paulino, R. L. Warren, B. P. Vandervalk, A. Raymond, S. D. Jackman, and I. Birol. Sealer: a scalable gap-closing application for finishing draft genomes. *BMC Bioinformatics*, 16, 2015.

[130] Adam C. English, Stephen Richards, Yi Han, Min Wang, Vanesa Vee, Jiaxin Qu, Xiang Qin, Donna M. Muzny, Jeffrey G. Reid, Kim C. Worley, and Richard A. Gibbs. Mind the gap: Upgrading genomes with pacific biosciences rs long-read sequencing technology. *PLOS ONE*, 7(11):e47768, 2012.

[131] Jennifer M. Shelton, Michelle C. Coleman, Nic Herndon, Nanyan Lu, Ernest T. Lam, Thomas Anantharaman, Palak Sheth, and Susan J. Brown. Tools and pipelines for bionano data: molecule assembly pipeline and fasta super scaffolding tool. *BMC Genomics*, 16(1):734, 2015.

[132] D. Earl, K. Bradnam, J. St John, A. Darling, D. Lin, J. Fass, H. O. Yu, V. Buffalo, D. R. Zerbino, M. Diekhans, N. Nguyen, P. N. Ariyaratne, W. K. Sung, Z. Ning, M. Haimel, J. T. Simpson, N. A. Fonseca, T. R. Docking, I. Y. Ho, D. S. Rokhsar, R. Chikhi, D. Lavenier, G. Chapuis, D. Naquin, N. Maillet, M. C. Schatz, D. R. Kelley, A. M. Phillippy, and S. Koren. Assemblathon 1: a competitive assessment of *de novo* short read assembly methods. *Genome Res*, 21:2224–41, 2011.

[133] Keith R. Bradnam, Joseph N. Fass, Anton Alexandrov, Paul Baranay, Michael Bechner, Inanç Birol, Sébastien Boisvert, Jarrod A. Chapman, Guillaume Chapuis, Rayan Chikhi, Hamidreza Chitsaz, Wen-Chi Chou, Jacques Corbeil, Cristian Del Fabbro, T. Roderick Docking, Richard Durbin, Dent Earl, Scott Emrich, Pavel Fedotov, Nuno A. Fonseca, Ganeshkumar Ganapathy, Richard A. Gibbs, Sante Gnerre, Élénie Godzaridis, Steve Goldstein, Matthias Haimel, Giles Hall, David Haussler, Joseph B. Hiatt, Isaac Y. Ho, Jason Howard, Martin Hunt, Shaun D. Jackman, David B. Jaffe, Erich D. Jarvis, Huaiyang Jiang, Sergey Kazakov, Paul J. Kersey, Jacob O. Kitzman, James R. Knight, Sergey Koren, Tak-Wah Lam, Dominique Lavenier, François Laviolette, Yingrui Li, Zhenyu Li, Binghang Liu, Yue Liu, Ruibang Luo, Iain MacCallum, Matthew D. MacManes, Nicolas Maillet, Sergey Melnikov, Delphine Naquin, Zemin Ning, Thomas D. Otto, Benedict Paten, Octávio S. Paulo, Adam M. Phillippy, Francisco Pina-Martins, Michael Place, Dariusz Przybylski, Xiang Qin, Carson Qu, Filipe J. Ribeiro, Stephen Richards, Daniel S. Rokhsar, J. Graham Ruby, Simone Scalabrin, Michael C. Schatz, David C. Schwartz, Alexey Sergushichev, Ted Sharpe, Timothy I. Shaw, Jay Shendure, Yujian Shi, Jared T. Simpson, Henry Song, Fedor Tsarev, Francesco Vezzi, Riccardo Vicedomini, Bruno M. Vieira, Jun Wang, Kim C. Worley, Shuangye Yin, Siu-Ming Yiu, Jianying Yuan, Guojie Zhang, Hao Zhang, Shiguo Zhou, and Ian F. Korf. Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species. *GigaScience*, 2(1):10, 2013.

[134] Mark Yandell and Daniel Ence. A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet*, 13(5):329–342, 2012.

[135] Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. Quast: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075, 2013.

[136] Martin Hunt, Taisei Kikuchi, Mandy Sanders, Chris Newbold, Matthew Berriman,

and Thomas D. Otto. Reapr: a universal tool for genome assembly evaluation. *Genome Biology*, 14(5):R47, 2013.

[137] Daniel Mapleson, Gonzalo Garcia Accinelli, George Kettleborough, Jonathan Wright, and Bernardo J. Clavijo. Kat: a k-mer analysis toolkit to quality control ngs datasets and genome assemblies. *Bioinformatics*, 33(4):574–576, 2016.

[138] G. Parra, K. Bradnam, and I. Korf. Cegma: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, 23, 2007.

[139] F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov. Busco: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31, 2015.

[140] Bernardo J. Clavijo, Luca Venturini, Christian Schudoma, Gonzalo Garcia Accinelli, Gemy Kaithakottil, Jonathan Wright, Philippa Borrill, George Kettleborough, Darren Heavens, Helen Chapman, James Lipscombe, Tom Barker, Fu-Hao Lu, Neil McKenzie, Dina Raats, Ricardo H. Ramirez-Gonzalez, Aurore Coince, Ned Peel, Lawrence Percival-Alwyn, Owen Duncan, Josua Trösch, Guotai Yu, Dan M. Bolser, Guy Namaati, Arnaud Kerhornou, Manuel Spannagl, Heidrun Gundlach, Georg Haberer, Robert P. Davey, Christine Fosker, Federica Di Palma, Andrew L. Phillips, A. Harvey Millar, Paul J. Kersey, Cristobal Uauy, Ksenia V. Krasileva, David Swarbreck, Michael W. Bevan, and Matthew D. Clark. An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. *Genome Research*, 27(5):885–896, 2017.

[141] R. Hubley & P. Green A.F.A. Smit. Repeatmodeler 1.0.8. Open 4.0, 2008-2015.

[142] Zhirong Bao and Sean R. Eddy. Automated *de novo* identification of repeat sequence families in sequenced genomes. *Genome Research*, 12(8):1269–1276, 2002.

[143] Alkes L. Price, Neil C. Jones, and Pavel A. Pevzner. *De novo* identification of repeat families in large genomes. *Bioinformatics*, 21, 2005.

[144] R. Hubley & P. Green A.F.A. Smit. Repeatmasker 4.0.5. Open 4.0, 2008-2015.

[145] M. Stanke, R. Steinkamp, S. Waack, and B. Morgenstern. Augustus: a web server for gene finding in eukaryotes. *Nucleic Acids Res*, 32(Web Server issue):W309–W312, 2004.

[146] Vardges Ter-Hovhannisyan, Alexandre Lomsadze, Yury O. Chernoff, and Mark Borodovsky. Gene prediction in novel fungal genomes using an *ab initio* algorithm with unsupervised training. *Genome Research*, 18(12):1979–1990, 2008.

[147] Alexandre Lomsadze, Paul D. Burns, and Mark Borodovsky. Integration of mapped rna-seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Research*, 42(15):e119–e119, 2014.

[148] Alexandre Lomsadze, Karl Gemayel, Shiyuyun Tang, and Mark Borodovsky. Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes. *Genome Research*, 28(7):1079–1089, 2018.

[149] B. J. Haas, A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, and J. Bowden. De novo transcript sequence reconstruction from rna-seq using the trinity platform for reference generation and analysis. *Nat Protoc*, 8(8):1494–1512, 2013.

[150] B. J. Haas, A. L. Delcher, S. M. Mount, J. R. Wortman, and R. K. Smith. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res*, 31(19):5654–5666, 2003.

[151] Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R. Kelley, Harold Pimentel, Steven L. Salzberg, John L. Rinn, and Lior Pachter. Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nature Protocols*, 7:562, 2012.

[152] R. D. Finn, J. Mistry, B. Schuster-Böckler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, and R. Durbin. Pfam: clans, web tools and services. *Nucleic Acids Res*, 34, 2006.

[153] David M. Goodstein, Shengqiang Shu, Russell Howson, Rochak Neupane, Richard D. Hayes, Joni Fazo, Therese Mitros, William Dirks, Uffe Hellsten, Nicholas Putnam, and Daniel S. Rokhsar. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research*, 40(D1):D1178–D1186, 2012.

[154] Dan Bolser, Daniel M. Staines, Emily Pritchard, and Paul Kersey. Ensembl plants: Integrating tools for visualizing, mining, and analyzing plant genomics data. In *Plant Bioinformatics: Methods and Protocols*, pages 115–140. Springer New York, 2016.

[155] Michael Burrows and David Wheeler. A block-sorting lossless data compression algorithm. *DIGITAL SRC RESEARCH REPORT*, 1994.

[156] H. Li and R. M. Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25, 2009.

[157] B. Langmead and S. L. Salzberg. Fast gapped-read alignment with bowtie 2. *Nat Methods*, 9, 2012.

[158] T. D. Wu and C. K. Watanabe. Gmap: a genomic mapping and alignment program for mrna and est sequences. *Bioinformatics*, 21(9):1859–1875, 2005.

[159] Guy St C. Slater and Ewan Birney. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6(1):1–11, 2005.

[160] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215, 1990.

[161] B. L. Cantarel, I. Korf, S. M. C. Robb, G. Parra, E. Ross, B. Moore, C. Holt, A. Sanchez Alvarado, and M. Yandell. Maker: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res*, 18, 2008.

[162] Brian J. Haas, Steven L. Salzberg, Wei Zhu, Mihaela Pertea, Jonathan E. Allen, Joshua Orvis, Owen White, C. Robin Buell, and Jennifer R. Wortman. Automated eukaryotic gene structure annotation using evidencemodeler and the program to assemble spliced alignments. *Genome Biology*, 9(1):R7, 2008.

[163] E. Quevillon, V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler, and R. Lopez. Interproscan: protein domains identifier. *Nucl Acids Res*, 33, 2005.

[164] N. Mulder and R. Apweiler. Interpro and interproscan: tools for protein sequence classification and comparison. *Methods Mol Biol*, 396, 2007.

[165] P. Jones, D. Binns, H. Y. Chang, M. Fraser, W. Li, and C. McAnulla. Interproscan 5: genome-scale protein function classification. *Bioinformatics*, 30(9):1236–1240, 2014.

[166] A. Conesa, S. Götz, J. M. García-Gómez, J. Terol, M. Talón, and M. Robles. Blast2go: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21(18):3674–3676, 2005.

[167] Prashant S. Hosmani, Teresa Shippy, Sherry Miller, Joshua B. Benoit, Monica Munoz-Torres, Mirella Flores, Lukas A. Mueller, Helen Wiersma-Koch, Tom D'elia, Susan J. Brown, and Surya Saha. A quick guide for student-driven community genome annotation. *ArXiv e-prints*, 2018. DOI: https://arxiv.org/abs/1805.03602.

[168] Theodore T. Puck, Steven J. Cieciura, and Arthur Robinson. Genetics of somatic mammalian cells : Iii. long-term cultivation of euploid cells from human and animal subjects. *The Journal of Experimental Medicine*, 108(6):945–956, 1958.

[169] Marie-Eve Lalonde and Yves Durocher. Therapeutic glycoprotein production in mammalian cells. *Journal of Biotechnology*, 251:128–140, 2017.

[170] H. A. Daniel Lagassé, Aikaterini Alexaki, Vijaya L. Simhadri, Nobuko H. Katagiri, Wojciech Jankowski, Zuben E. Sauna, and Chava Kimchi-Sarfaty. Recent advances in (therapeutic protein) drug development. *F1000Research*, 6:113, 2017.

[171] Ashok D. Bandaranayake and Steven C. Almo. Recent advances in mammalian protein production. *FEBS letters*, 588(2):253–260, 2013.

[172] F. M. Wurm. Production of recombinant protein therapeutics in cultivated mammalian cells. *Nat Biotechnol*, 22, 2004.

[173] Tingfeng Lai, Yuansheng Yang, and Say Kong Ng. Advances in mammalian cell line development technologies for recombinant protein production. *Pharmaceuticals*, 6(5):579–603, 2013.

[174] F. M. Huennekens. Folic acid coenzymes in the biosynthesis of purines and pyrimidines. In Robert S. Harris, Ira G. Wool, John A. Loraine, and Kenneth V. Thimann, editors, *Vitamins & Hormones*, volume 26, pages 375–394. Academic Press, 1969.

[175] M. Florian Wurm. Cho quasispecies - implications for manufacturing processes. *Processes*, 1(3), 2013.

[176] Lise Marie Grav, Karen Julie la Cour Karottki, Jae Seong Lee, and Helene Faustrup Kildegaard. Application of crispr/cas9 genome editing to improve recombinant protein production in cho cells. In *Heterologous Protein Production in CHO Cells: Methods and Protocols*, pages 101–118. Springer New York, 2017.

[177] Xin Luo, Min Li, and Bing Su. Application of the genome editing tool crispr/cas9 in non-human primates. *Zoological research*, 37(4):214–219, July 2016.

[178] Xun Xu, Harish Nagarajan, Nathan E. Lewis, Shengkai Pan, Zhiming Cai, Xin Liu, Wenbin Chen, Min Xie, Wenliang Wang, Stephanie Hammond, Mikael R. Andersen, Norma Neff, Benedetto Passarelli, Winston Koh, H. Christina Fan, Jianbin Wang, Yaoting Gui, Kelvin H. Lee, Michael J. Betenbaugh, Stephen R. Quake, Iman Famili, Bernhard O. Palsson, and Jun Wang. The genomic sequence of the chinese hamster ovary (cho)-k1 cell line. *Nature Biotechnology*, 29:735, July 2011.

[179] Karina Brinkrolf, Oliver Rupp, Holger Laux, Florian Kollin, Wolfgang Ernst, Burkhard Linke, Rudolf Kofler, Sandrine Romand, Friedemann Hesse, Wolfgang E. Budach, Sybille Galosy, Dethardt Müller, Thomas Noll, Johannes Wienberg, Thomas Jostock, Mark Leonard, Johannes Grillari, Andreas Tauch, Alexander Goesmann, Bernhard Helk, John E. Mott, Alfred Pühler, and Nicole Borth. Chinese hamster genome sequenced from sorted chromosomes. *Nature Biotechnology*, 31:694, August 2013.

[180] Nathan E. Lewis, Xin Liu, Yuxiang Li, Harish Nagarajan, George Yerganian, Edward O'Brien, Aarash Bordbar, Anne M. Roth, Jeffrey Rosenbloom, Chao Bian, Min Xie, Wenbin Chen, Ning Li, Deniz Baycin-Hizal, Haythem Latif, Jochen Forster, Michael J. Betenbaugh, Iman Famili, Xun Xu, Jun Wang, and Bernhard O. Palsson. Genomic landscapes of chinese hamster ovary cell lines as revealed by the cricetulus griseus draft genome. *Nature Biotechnology*, 31:759, July 2013.

[181] Richard J. Roberts, Mauricio O. Carneiro, and Michael C. Schatz. The advantages of smrt sequencing. *Genome Biology*, 14(6):405, 2013.

[182] Karen M. Moll, Peng Zhou, Thiruvarangan Ramaraj, Diego Fajardo, Nicholas P. Devitt, Michael J. Sadowsky, Robert M. Stupar, Peter Tiffin, Jason R. Miller, Nevin D. Young, Kevin A. T. Silverstein, and Joann Mudge. Strategies for optimizing bionano and dovetail explored through a second reference quality assembly for the legume model, medicago truncatula. *BMC Genomics*, 18(1):578, 2017.

[183] Bronwen L. Aken, Sarah Ayling, Daniel Barrell, Laura Clarke, Valery Curwen, Susan Fairley, Julio Fernandez Banet, Konstantinos Billis, Carlos García Girón, Thibaut Hourlier, Kevin Howe, Andreas Kähäri, Felix Kokocinski, Fergal J. Martin,

Daniel N. Murphy, Rishi Nag, Magali Ruffier, Michael Schuster, Y. Amy Tang, Jan-Hinnerk Vogel, Simon White, Amonida Zadissa, Paul Flicek, and Stephen M. J. Searle. The ensembl gene annotation system. *Database*, 2016:093–093, 2016.

[184] M. Florian Wurm and J. Maria Wurm. Cloning of cho cells, productivity and genetic stability–a discussion. *Processes*, 5(2), 2017.

[185] Dorai Haimanti, Corisdeo Susanne, Ellis Dawn, Kinney Cherylann, Chomo Matt, Hawley-Nelson Pam, Moore Gordon, J. Betenbaugh Michael, and Ganguly Subinay. Early prediction of instability of chinese hamster ovary cell lines expressing recombinant antibodies and antibody/-fusion proteins. *Biotechnol. Bioeng.*, 109(4):1016–1030, 2012.

[186] Fan Lianchun, Kadura Ibrahim, E. Krebs Lara, C. Hatfield Christopher, M. Shaw Margaret, and C. Frye Christopher. Improving the efficiency of cho cell line generation using glutamine synthetase gene knockout cells. *Biotechnol. Bioeng.*, 109(4):1007–1015, 2012.

[187] Paul Julian Kersey, James E. Allen, Alexis Allot, Matthieu Barba, Sanjay Boddu, Bruce J. Bolt, Denise Carvalho-Silva, Mikkel Christensen, Paul Davis, Christoph Grabmueller, Navin Kumar, Zicheng Liu, Thomas Maurel, Ben Moore, Mark D. McDowall, Uma Maheswari, Guy Naamati, Victoria Newman, Chuang Kee Ong, Michael Paulini, Helder Pedro, Emily Perry, Matthew Russell, Helen Sparrow, Electra Tapanari, Kieron Taylor, Alessandro Vullo, Gareth Williams, Amonida Zadissia, Andrew Olson, Joshua Stein, Sharon Wei, Marcela Tello-Ruiz, Doreen Ware, Aurelien Luciani, Simon Potter, Robert D. Finn, Martin Urban, Kim E. Hammond-Kosack, Dan M. Bolser, Nishadi DeÂ Silva Kevin L. Howe, Nicholas Langridge, Gareth Maslen, Daniel Michael Staines, and Andrew Yates. Ensembl genomes 2018: an integrated omics infrastructure for non-vertebrate species. *Nucleic Acids Research*, 46(D1):D802–D808, 2018.

[188] Liu Pei-Qi, M. Chan Edmond, J. Cost Gregory, Zhang Lin, Wang Jianbin, C. Miller Jeffrey, Y. Guschin Dmitry, Reik Andreas, C. Holmes Michael, E. Mott John, N. Collingwood Trevor, and D. Gregory Philip. Generation of a triple-gene knockout mammalian cell line using engineered zinc-finger nucleases. *Biotechnol. Bioeng.*, 106(1):97–105, December 2009.

[189] Jared T. Simpson and Richard Durbin. Efficient *de novo* assembly of large genomes using compressed data structures. *Genome Research*, 22(3):549–556, 2012. PMC.

[190] Alejandro Hernandez Wences and Michael C. Schatz. Metassembler: merging and optimizing *de novo* genome assemblies. *Genome Biology*, 16(1):207, 2015.

[191] Daniel Mapleson, Gonzalo Garcia Accinelli, George Kettleborough, Jonathan Wright, and Bernardo J. Clavijo. Kat: a k-mer analysis toolkit to quality control ngs datasets and genome assemblies. *Bioinformatics*, 2016.

[192] Sujai Kumar, Martin Jones, Georgios Koutsovoulos, Michael Clarke, and Mark Blaxter. Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated gc-coverage plots. *Frontiers in Genetics*, 4(237):237, 2013.

[193] Ying Chen, Weicai Ye, Yongdong Zhang, and Yuesheng Xu. High speed blastn: an accelerated megablast search tool. *Nucleic Acids Research*, 43(16):7762–7768, 2015.

[194] Christian Quast, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. The silva ribosomal rna gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, 41(D1):D590–D596, 2013.

[195] Vcelar Sabine, Jadhav Vaibhav, Melcher Michael, Auer Norbert, Hrdina Astrid, Sagmeister Rebecca, Heffner Kelley, Puklowski Anja, Betenbaugh Michael, Wenger Till, Leisch Friedrich, Baumann Martina, and Borth Nicole. Karyotype variation of cho host cell lines over time in culture characterized by chromosome counting and chromosome painting. *Biotechnol. Bioeng.*, 115(1):165–173, 2017.

[196] Aleksandr Morgulis, E. Michael Gertz, Alejandro A. Schäffer, and Richa Agarwala. A fast and symmetric dust implementation to mask low-complexity dna sequences. *Journal of Computational Biology*, 13(5):1028–1040, 2006.

[197] G. Benson. Tandem repeats finder: a program to analyze dna sequences. *Nucleic Acids Research*, 27(2):573–580, 1999.

[198] Rong She, Jeffrey Shih-Chieh Chu, Bora Uyar, Jun Wang, Ke Wang, and Nansheng Chen. genblastg: using blast searches to build homologous gene models. *Bioinformatics*, 27(15):2141–2143, 2011.

[199] The UniProt Consortium. Uniprot: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1):D158–D169, January 2017.

[200] Robert S. Harris. *Improved pairwise alignment of genomic DNA*. The Pennsylvania State University. 0549431705.

[201] Jennifer Harrow, France Denoeud, Adam Frankish, Alexandre Reymond, Chao-Kung Chen, Jacqueline Chrast, Julien Lagarde, James G. R. Gilbert, Roy Storey, David Swarbreck, Colette Rossier, Catherine Ucla, Tim Hubbard, Stylianos E. Antonarakis, and Roderic Guigo. Gencode: producing a reference annotation for encode. *Genome biology*, 7 Suppl 1:S4.1–9, 2006.

[202] Eric P. Nawrocki, Diana L. Kolbe, and Sean R. Eddy. Infernal 1.0: inference of rna alignments. *Bioinformatics*, 25(10):1335–1337, 2009.

[203] Ronny Lorenz, Stephan H. Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F. Stadler, and Ivo L. Hofacker. Viennarna package 2.0. *Algorithms for Molecular Biology*, 6(1):26, 2011.

[204] David M. Emms and Steven Kelly. Orthofinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, 16(1):157, 2015.

[205] Zhou Du, Xin Zhou, Yi Ling, Zhenhai Zhang, and Zhen Su. agrigo: a go analysis toolkit for the agricultural community. *Nucleic Acids Research*, 38(Web Server issue):W64–W70, 2010.

[206] R. A. Fisher. On the interpretation of x2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922.

[207] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B Methodol*, 57(1):289–300, 1995.

[208] F. Supek, M. Bošnjak, N. Škunca, and T. Šmuc. Revigo summarizes and visualizes long lists of gene ontology terms. *PLoS One*, 6(7):e21800, 2011.

[209] Andreas Schlicker, Francisco S. Domingues, Jörg Rahnenführer, and Thomas Lengauer. A new measure for functional similarity of gene products based on gene ontology. *BMC Bioinformatics*, 7(1):302, 2006.

[210] Eric Lyons, Brent Pedersen, Josh Kane, Maqsudul Alam, Ray Ming, Haibao Tang, Xiyin Wang, John Bowers, Andrew Paterson, Damon Lisch, and Michael Freeling. Finding and comparing syntenic regions among arabidopsis and the outgroups papaya, poplar, and grape: Coge with rosids. *Plant Physiology*, 148(4):1772–1781, 2008.

[211] E. Lyons. The value of nonmodel genomes and an example using synmap within coge to dissect the hexaploidy that predates the rosids. *Trop Plant Biol*, 1(3):181–190, 2008.

[212] Scott Schwartz, W. James Kent, Arian Smit, Zheng Zhang, Robert Baertsch, Ross C. Hardison, David Haussler, and Webb Miller. Human-mouse alignments with blastz. *Genome Research*, 13(1):103–107, 2003.

[213] Brian J. Haas, Arthur L. Delcher, Jennifer R. Wortman, and Steven L. Salzberg. Dagchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics*, 20(18):3643–3646, 2004.

[214] Z. Yang. Paml 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol*, 24(8):1586–1591, 2007.

[215] W. James Kent. Blat–the blast-like alignment tool. *Genome Research*, 12(4):656–664, 2002.

[216] James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, and Jill P. Mesirov. Integrative genomics viewer. *Nature biotechnology*, 29(1):24–26, 2011.

[217] V. A. Simossis and J. Heringa. Praline: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. *Nucleic Acids Research*, 33(Web Server issue):W289–W294, 2005.

[218] Paul S. Kelly, Colin Clarke, Alan Costello, Craig Monger, Justine Meiller, Heena Dhiman, Nicole Borth, Michael J. Betenbaugh, Martin Clynes, and Niall Barron. Ultra-deep next generation mitochondrial genome sequencing reveals widespread heteroplasmy in chinese hamster ovary cells. *Metabolic Engineering*, 41:11–22, 2017.

[219] Yuki Dehara, Yasuyuki Hashiguchi, Kazumi Matsubara, Tokuma Yanai, Masahito Kubo, and Yoshinori Kumazawa. Characterization of squamate olfactory receptor

genes and their transcripts by the high-throughput sequencing approach. *Genome Biology and Evolution*, 4(4):602–616, 2012.

[220] F. Yang, P. C. M. O'Brien, and M. A. Ferguson-Smith. Comparative chromosome map of the laboratory mouse and chinese hamster defined by reciprocal chromosome painting. *Chromosome Research*, 8(3):219–227, 2000.

[221] Svetlana A. Romanenko, Polina L. Perelman, Natalya A. Serdukova, Vladimir A. Trifonov, Larisa S. Biltueva, Jinhuan Wang, Tangliang Li, Wenhui Nie, Patricia C. M. OBrien, Vitaly T. Volobouev, Roscoe Stanyon, Malcolm A. Ferguson-Smith, Fengtang Yang, and Alexander S. Graphodatsky. Reciprocal chromosome painting between three laboratory rodent species. *Mammalian Genome*, 17(12):1183–1192, 2006.

[222] Yihua Cao, Shuichi Kimura, Joon-young Park, Miyuki Yamatani, Kohsuke Honda, Hisao Ohtake, and Takeshi Omasa. Chromosome identification and its application in chinese hamster ovary cells. *BMC Proceedings*, 5(8):O8, 2011.

[223] Yihua Cao, Shuichi Kimura, Takayuki Itoi, Kohsuke Honda, Hisao Ohtake, and Takeshi Omasa. Construction of bac-based physical map and analysis of chromosome rearrangement in chinese hamster ovary cell lines. *Biotechnology and Bioengineering*, 109(6):1357–1367, 2012.

[224] Tang Danming, Lam Cynthia, Louie Salina, Hoi Kam Hon, Shaw David, Yim Mandy, Snedecor Brad, and Misaghi Shahram. Supplementation of nucleosides during selection can reduce sequence variant levels in cho cells using gs/msx selection system. *Biotechnol. J.*, 13(1):1700335, 2018.

[225] G. Urlaub and L. A. Chasin. Isolation of chinese hamster cell mutants deficient in dihydrofolate reductase activity. *Proceedings of the National Academy of Sciences of the United States of America*, 77(7):4216–4220, July 1980.

[226] Soo Min Noh, Seunghyeon Shin, and Gyun Min Lee. Comprehensive characterization of glutamine synthetase-mediated selection for the establishment of recombinant cho cells producing monoclonal antibodies. *Scientific Reports*, 8(1):5361, 2018.

[227] David E. Jarvis, Yung Shwen Ho, Damien J. Lightfoot, Sandra M. Schmöckel, Bo Li, Theo J. A. Borm, Hajime Ohyanagi, Katsuhiko Mineta, Craig T. Michell, Noha Saber, Najeh M. Kharbatia, Ryan R. Rupper, Aaron R. Sharp, Nadine Dally,

Berin A. Boughton, Yong H. Woo, Ge Gao, Elio G. W. M. Schijlen, Xiujie Guo, Afaque A. Momin, Sónia Negrão, Salim Al-Babili, Christoph Gehring, Ute Roessner, Christian Jung, Kevin Murphy, Stefan T. Arold, Takashi Gojobori, C. Gerard van der Linden, Eibertus N. van Loo, Eric N. Jellen, Peter J. Maughan, and Mark Tester. The genome of *Chenopodium quinoa*. *Nature*, 542:307, 2017.

[228] Q. Liu, S. Chang, G. L. Hartman, and L. L. Domier. Assembly and annotation of a draft genome sequence for glycine latifolia, a perennial wild relative of soybean. *Plant J*, 2018.

[229] Jeramiah J. Smith, Nataliya Timoshevskaya, Chengxi Ye, Carson Holt, Melissa C. Keinath, Hugo J. Parker, Malcolm E. Cook, Jon E. Hess, Shawn R. Narum, Francesco Lamanna, Henrik Kaessmann, Vladimir A. Timoshevskiy, Courtney K. M. Waterbury, Cody Saraceno, Leanne M. Wiedemann, Sofia M. C. Robb, Carl Baker, Evan E. Eichler, Dorit Hockman, Tatjana Sauka-Spengler, Mark Yandell, Robb Krumlauf, Greg Elgar, and Chris T. Amemiya. The sea lamprey germline genome provides insights into programmed genome rearrangement and vertebrate evolution. *Nature Genetics*, 50(2):270–277, 2018.

[230] Valerie A. Schneider, Tina Graves-Lindsay, Kerstin Howe, Nathan Bouk, Hsiu-Chuan Chen, Paul A. Kitts, Terence D. Murphy, Kim D. Pruitt, Françoise Thibaud-Nissen, Derek Albracht, Robert S. Fulton, Milinn Kremitzki, Vincent Magrini, Chris Markovic, Sean McGrath, Karyn Meltz Steinberg, Kate Auger, William Chow, Joanna Collins, Glenn Harden, Timothy Hubbard, Sarah Pelan, Jared T. Simpson, Glen Threadgold, James Torrance, Jonathan M. Wood, Laura Clarke, Sergey Koren, Matthew Boitano, Paul Peluso, Heng Li, Chen-Shan Chin, Adam M. Phillippy, Richard Durbin, Richard K. Wilson, Paul Flicek, Evan E. Eichler, and Deanna M. Church. Evaluation of grch38 and *de novo* haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*, 27(5):849–864, 2017.

[231] Ming-Shan Wang, He-Chuan Yang, Newton O. Otecko, Dong-Dong Wu, and Ya-Ping Zhang. Olfactory genes in tibetan wild boar. *Nature Genetics*, 48:972, 2016.

[232] Nandita Vishwanathan, Andrew Yongky, Kathryn C. Johnson, HsuYuan Fu, Nitya M. Jacob, Huong Le, Faraaz N. K. Yusufi, Dong Yup Lee, and WeiShou Hu.

Global insights into the chinese hamster and cho cell transcriptomes. *Biotechnology and Bioengineering*, 112(5):965–976, 2015.

[233] Michael A. Partridge, Mercy M. Davidson, and Tom K. Hei. The complete nucleotide sequence of chinese hamster (*Cricetulus griseus*) mitochondrial dna. *DNA Sequence*, 18(5):341–346, 2007.

[234] Xun Xu, Harish Nagarajan, Nathan E. Lewis, Shengkai Pan, Zhiming Cai, Xin Liu, Wenbin Chen, Min Xie, Wenliang Wang, Stephanie Hammond, Mikael R. Andersen, Norma Neff, Benedetto Passarelli, Winston Koh, H. Christina Fan, Jianbin Wang, Yaoting Gui, Kelvin H. Lee, Michael J. Betenbaugh, Stephen R. Quake, Iman Famili, Bernhard O. Palsson, and Jun Wang. The genomic sequence of the chinese hamster ovary (cho)-k1 cell line. *Nature Biotechnology*, 29(8):735–741, 2011.

[235] Romanova Nadiya and Noll Thomas. Engineered and natural promoters and chromatinmodifying elements for recombinant protein expression in cho cells. *Biotechnol. J.*, 13(3):1700232, 2018.

[236] Yoshihiro Kaneko, Ryuji Sato, and Hideki Aoyagi. Evaluation of chinese hamster ovary cell stability during repeated batch culture for large-scale antibody production. *Journal of Bioscience and Bioengineering*, 109(3):274–280, 2010.

[237] Hooman Hefzi, Kok Siong Ang, Michael Hanscho, Aarash Bordbar, David Ruckerbauer, Meiyappan Lakshmanan, Camila A. Orellana, Deniz Baycin-Hizal, Yingxiang Huang, Daniel Ley, Veronica S. Martinez, Sarantos Kyriakopoulos, Natalia E. Jiménez, Daniel C. Zielinski, Lake-Ee Quek, Tune Wulff, Johnny Arnsdorf, Shangzhong Li, Jae Seong Lee, Giuseppe Paglia, Nicolas Loira, Philipp N. Spahn, Lasse E. Pedersen, Jahir M. Gutierrez, Zachary A. King, Anne Mathilde Lund, Harish Nagarajan, Alex Thomas, Alyaa M. Abdel-Haleem, Juergen Zanghellini, Helene F. Kildegaard, Bj G. Voldborg, Ziomara P. Gerdtzen, Michael J. Betenbaugh, Bernhard O. Palsson, Mikael R. Andersen, Lars K. Nielsen, Nicole Borth, Dong-Yup Lee, and Nathan E. Lewis. A consensus genome-scale reconstruction of chinese hamster ovary cell metabolism. *Cell Systems*, 3(5):434–443.e8, 2016.

[238] Elliott J. Price, Paul Wilkin, Viswambharan Sarasan, and Paul D. Fraser. Metabolite profiling of *Dioscorea* (yam) species reveals underutilised biodiversity and renewable sources for high-value compounds. *Scientific Reports*, 6:29136, 2016.

[239] Chih-Chun Wen, Hui-Ming Chen, and Ning-Sun Yang. Chapter 6 - developing phytocompounds from medicinal plants as immunomodulators. In Lie-Fen Shyur and Allan S. Y. Lau, editors, *Advances in Botanical Research*, volume 62, pages 197–272. Academic Press, 2012.

[240] E. Huala, A. W. Dickerman, M. Garcia-Hernandez, D. Weems, L. Reiser, F. LaFond, D. Hanley, D. Kiphart, M. Zhuang, W. Huang, L. A. Mueller, D. Bhattacharyya, D. Bhaya, B. W. Sobral, W. Beavis, D. W. Meinke, C. D. Town, C. Somerville, and S. Y. Rhee. The arabidopsis information resource (tair): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res*, 29(1):102–105, 2001.

[241] John P. Vogel, David F. Garvin, Todd C. Mockler, Jeremy Schmutz, Dan Rokhsar, Michael W. Bevan, Kerrie Barry, Susan Lucas, Miranda Harmon-Smith, and Kathleen Lail. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature*, 463(7282):763–768, 2010. 0028-0836.

[242] Shu Ouyang, Wei Zhu, John Hamilton, Haining Lin, Matthew Campbell, Kevin Childs, FranÃ§oise Thibaud-Nissen, Renae L. Malek, Yuandan Lee, Li Zheng, Joshua Orvis, Brian Haas, Jennifer Wortman, and C. Robin Buell. The tigr rice genome annotation resource: improvements and new features. *Nucleic Acids Research*, 35(Database issue):D883–D887, 2006.

[243] Asaf A. Salamov and Victor V. Solovyev. *Ab initio* gene finding in *Drosophila* genomic dna. *Genome Research*, 10(4):516–522, 2000.

[244] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg. Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*, 14(4):1–13, 2013.

[245] J. E. Allen and S. L. Salzberg. Jigsaw: integration of multiple sources of evidence for gene prediction. *Bioinformatics*, 21(18):3596–3603, 2005.

[246] C. Holt and M. Yandell. Maker2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, 12(1):1–14, 2011.

[247] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol*, 10(3):1–10, 2009.

[248] H. Li. A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993, 2011.

[249] H. Li and R. Durbin. Fast and accurate long-read alignment with burrows-wheeler transform. *Bioinformatics*, 26(5):589–595, 2010.

[250] S. Anders, P. T. Pyl, and W. Huber. Htseq a python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169, 2014.

[251] Simon Anders and Wolfgang Huber. Differential expression of rna/-seq data at the gene levelâthe deseq package. *Heidelberg, Germany: European Molecular Biology Laboratory (EMBL)*, 2012.

[252] Bernat Gel and Eduard Serra. karyoploter: an r/bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics*, 33(19):3088–3090, 2017.

[253] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[254] Debra Klopfenstein, Brent Pedersen, Patrick Flick, Kenta Sato, Fidel Ramirez, Jeff Yunes, Chris Mungall, and Haibao Tang. Goatools: Tools for gene ontology. *Zenodo*, 2015.

[255] Andrey Alexeyenko, Ivica Tamas, Gang Liu, and Erik L. L. Sonnhammer. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics*, 22(14):e9–e15, 2006.

[256] Ann-Charlotte Berglund, Erik Sjölund, Gabriel Östlund, and Erik L. L. Sonnhammer. Inparanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic Acids Research*, 36(suppl_1):D263–D266, 2008.

[257] K. P. O'Brien, M. Remm, and E. L. Sonnhammer. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res*, 33(Database issue):D476–80, 2005.

[258] M. Remm, C. E. Storm, and E. L. Sonnhammer. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol*, 314(5):1041–52, 2001.

[259] Angelique D'Hont, France Denoeud, Jean-Marc Aury, Franc-Christophe Baurens, Francoise Carreel, Olivier Garsmeur, Benjamin Noel, Stephanie Bocs, Gaetan Droc, Mathieu Rouard, Corinne Da Silva, Kamel Jabbari, Celine Cardi, Julie Poulain, Marlene Souquet, Karine Labadie, Cyril Jourda, Juliette Lengelle, Marguerite Rodier-Goud, Adriana Alberti, Maria Bernard, Margot Correa, Saravanaraj Ayyampalayam, Michael R. McKain, Jim Leebens-Mack, Diane Burgess, Mike Freeling, Didier Mbeguie-A-Mbeguie, Matthieu Chabannes, Thomas Wicker, Olivier Panaud, Jose Barbosa, Eva Hribova, Pat Heslop-Harrison, Remy Habas, Ronan Rivallan, Philippe Francois, Claire Poiron, Andrzej Kilian, Dheema Burthia, Christophe Jenny, Frederic Bakry, Spencer Brown, Valentin Guignon, Gert Kema, Miguel Dita, Cees Waalwijk, Steeve Joseph, Anne Dievart, Olivier Jaillon, Julie Leclercq, Xavier Argout, Eric Lyons, Ana Almeida, Mouna Jeridi, Jaroslav Dolezel, Nicolas Roux, Ange-Marie Risterucci, Jean Weissenbach, Manuel Ruiz, Jean-Christophe Glaszmann, Francis Quetier, Nabila Yahiaoui, and Patrick Wincker. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature*, 488(7410):213–217, 2012.

[260] R. Singh, M. Ong-Abdullah, E. T. Low, M. A. Manaf, R. Rosli, and R. Nookiah. Oil palm genome sequence reveals divergence of interfertile species in old and new worlds. *Nature*, 500(7462):335–339, 2013.

[261] I. S. Al-Mssallem, S. Hu, X. Zhang, Q. Lin, W. Liu, and J. Tan. Genome sequence of the date palm phoenix dactylifera l. *Nat Commun*, 4:2274, 2013.

[262] Rolf Apweiler, Amos Bairoch, Cathy H. Wu, Winona C. Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, Maria J. Martin, Darren A. Natale, Claire O'Donovan, Nicole Redaschi, and Lai/-Su L. Yeh. Uniprot: the universal protein knowledgebase. *Nucleic Acids Research*, 32:D115–D119, 2004.

[263] K. Katoh and D. M. Standley. Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*, 30(4):772–780, 2013.

[264] A. Stamatakis. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 2014.

[265] Ray Ming, Robert VanBuren, Ching Man Wai, Haibao Tang, Michael C. Schatz, John E. Bowers, Eric Lyons, Ming-Li Wang, Jung Chen, Eric Biggers, Jisen Zhang, Lixian Huang, Lingmao Zhang, Wenjing Miao, Jian Zhang, Zhangyao Ye, Chenyong Miao, Zhicong Lin, Hao Wang, Hongye Zhou, Won C. Yim, Henry D. Priest, Chunfang Zheng, Margaret Woodhouse, Patrick P. Edger, Romain Guyot, Hao-Bo Guo, Hong Guo, Guangyong Zheng, Ratnesh Singh, Anupma Sharma, Xiangjia Min, Yun Zheng, Hayan Lee, James Gurtowski, Fritz J. Sedlazeck, Alex Harkess, Michael R. McKain, Zhenyang Liao, Jingping Fang, Juan Liu, Xiaodan Zhang, Qing Zhang, Weichang Hu, Yuan Qin, Kai Wang, Li-Yu Chen, Neil Shirley, Yann-Rong Lin, Li-Yu Liu, Alvaro G. Hernandez, Chris L. Wright, Vincent Bulone, Gerald A. Tuskan, Katy Heath, Francis Zee, Paul H. Moore, Ramanjulu Sunkar, James H. Leebens-Mack, Todd Mockler, Jeffrey L. Bennetzen, Michael Freeling, David Sankoff, Andrew H. Paterson, Xinguang Zhu, Xiaohan Yang, J. Andrew C. Smith, John C. Cushman, Robert E. Paull, and Qingyi Yu. The pineapple genome and the evolution of cam photosynthesis. *Nature Genetics*, 47:1435, 2015.

[266] Juliane C. Dohm, André E. Minoche, Daniela Holtgrŵe, Salvador Capella-Gutiérrez, Falk Zakrzewski, Hakim Tafer, Oliver Rupp, Thomas Rosleff Sörensen, Ralf Stracke, Richard Reinhardt, Alexander Goesmann, Thomas Kraft, Britta Schulz, Peter F. Stadler, Thomas Schmidt, Toni Gabaldón, Hans Lehrach, Bernd Weisshaar, and Heinz Himmelbauer. The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). *Nature*, 505:546, 2013.

[267] Atsushi Hoshino, Vasanthan Jayakumar, Eiji Nitasaka, Atsushi Toyoda, Hideki Noguchi, Takehiko Itoh, Tadasu Shin-I, Yohei Minakuchi, Yuki Koda, Atsushi J. Nagano, Masaki Yasugi, Mie N. Honjo, Hiroshi Kudoh, Motoaki Seki, Asako Kamiya, Toshiyuki Shiraki, Piero Carninci, Erika Asamizu, Hiroyo Nishide, Sachiko Tanaka, Kyeung-Il Park, Yasumasa Morita, Kohei Yokoyama, Ikuo Uchiyama, Yoshikazu Tanaka, Satoshi Tabata, Kazuo Shinozaki, Yoshihide Hayashizaki, Yuji Kohara, Yutaka Suzuki, Sumio Sugano, Asao Fujiyama, Shigeru Iida, and Yasubumi Sakakibara. Genome sequence and analysis of the japanese morning glory *Ipomoea nil*. *Nature Communications*, 7:13295, 2016.

[268] Riccardo Velasco, Andrey Zharkikh, Jason Affourtit, Amit Dhingra, Alessandro

Cestaro, Ananth Kalyanaraman, Paolo Fontana, Satish K. Bhatnagar, Michela Troggio, Dmitry Pruss, Silvio Salvi, Massimo Pindo, Paolo Baldi, Sara Castelletti, Marina Cavaiuolo, Giuseppina Coppola, Fabrizio Costa, Valentina Cova, Antonio Dal Ri, Vadim Goremykin, Matteo Komjanc, Sara Longhi, Pierluigi Magnago, Giulia Malacarne, Mickael Malnoy, Diego Micheletti, Marco Moretto, Michele Perazzolli, Azeddine Si-Ammour, Silvia Vezzulli, Elena Zini, Glenn Eldredge, Lisa M. Fitzgerald, Natalia Gutin, Jerry Lanchbury, Teresita Macalma, Jeff T. Mitchell, Julia Reid, Bryan Wardell, Chinnappa Kodira, Zhoutao Chen, Brian Desany, Faheem Niazi, Melinda Palmer, Tyson Koepke, Derick Jiwan, Scott Schaeffer, Vandhana Krishnan, Changjun Wu, Vu T. Chu, Stephen T. King, Jessica Vick, Quanzhou Tao, Amy Mraz, Aimee Stormo, Keith Stormo, Robert Bogden, Davide Ederle, Alessandra Stella, Alberto Vecchietti, Martin M. Kater, Simona Masiero, Pauline Lasserre, Yves Lespinasse, Andrew C. Allan, Vincent Bus, David Chagné, Ross N. Crowhurst, Andrew P. Gleave, Enrico Lavezzo, Jeffrey A. Fawcett, Sebastian Proost, Pierre Rouzé, Lieven Sterck, Stefano Toppo, Barbara Lazzari, Roger P. Hellens, Charles-Eric Durel, Alexander Gutin, Roger E. Bumgarner, Susan E. Gardiner, Mark Skolnick, Michael Egholm, Yves Van de Peer, Francesco Salamini, and Roberto Viola. The genome of the domesticated apple (*Malus x domestica* borkh.). *Nature Genetics*, 42:833, 2010.

[269] Ray Ming, Robert VanBuren, Yanling Liu, Mei Yang, Yuepeng Han, Lei-Ting Li, Qiong Zhang, Min-Jeong Kim, Michael C. Schatz, Michael Campbell, Jingping Li, John E. Bowers, Haibao Tang, Eric Lyons, Ann A. Ferguson, Giuseppe Narzisi, David R. Nelson, Crysten E. Blaby-Haas, Andrea R. Gschwend, Yuannian Jiao, Joshua P. Der, Fanchang Zeng, Jennifer Han, Xiang Jia Min, Karen A. Hudson, Ratnesh Singh, Aleel K. Grennan, Steven J. Karpowicz, Jennifer R. Watling, Kikukatsu Ito, Sharon A. Robinson, Matthew E. Hudson, Qingyi Yu, Todd C. Mockler, Andrew Carroll, Yun Zheng, Ramanjulu Sunkar, Ruizong Jia, Nancy Chen, Jie Arro, Ching Man Wai, Eric Wafula, Ashley Spence, Yanni Han, Liming Xu, Jisen Zhang, Rhiannon Peery, Miranda J. Haus, Wenwei Xiong, James A. Walsh, Jun Wu, Ming-Li Wang, Yun J. Zhu, Robert E. Paull, Anne B. Britt, Chunguang Du, Stephen R. Downie, Mary A. Schuler, Todd P. Michael, Steve P. Long, Donald R. Ort, J. William Schopf, David R. Gang, Ning Jiang, Mark Yandell, Claude W. dePamphilis, Sabeeha S. Merchant, Andrew H. Paterson, Bob B. Buchanan, Shaohua Li, and Jane Shen-Miller. Genome of the long-living sacred lotus (*Nelumbo nucifera* gaertn.). *Genome Biology*, 14(5):R41, 2013.

[270] Robert VanBuren, Doug Bryant, Patrick P. Edger, Haibao Tang, Diane Burgess, Dinakar Challabathula, Kristi Spittle, Richard Hall, Jenny Gu, Eric Lyons, Michael Freeling, Dorothea Bartels, Boudewijn Ten Hallers, Alex Hastie, Todd P. Michael, and Todd C. Mockler. Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. *Nature*, 527:508, 2015.

[271] Jing Cai, Xin Liu, Kevin Vanneste, Sebastian Proost, Wen-Chieh Tsai, Ke-Wei Liu, Li-Jun Chen, Ying He, Qing Xu, Chao Bian, Zhijun Zheng, Fengming Sun, Weiqing Liu, Yu-Yun Hsiao, Zhao-Jun Pan, Chia-Chi Hsu, Ya-Ping Yang, Yi-Chin Hsu, Yu-Chen Chuang, Anne Dievart, Jean-Francois Dufayard, Xun Xu, Jun-Yi Wang, Jun Wang, Xin-Ju Xiao, Xue-Min Zhao, Rong Du, Guo-Qiang Zhang, Meina Wang, Yong-Yu Su, Gao-Chang Xie, Guo-Hui Liu, Li-Qiang Li, Lai-Qiang Huang, Yi-Bo Luo, Hong-Hwa Chen, Yves Van de Peer, and Zhong-Jian Liu. The genome sequence of the orchid phalaenopsis equestris. *Nature Genetics*, 47:65, 2014.

[272] A. H. Paterson, J. E. Bowers, R. Bruggmann, I. Dubchak, J. Grimwood, and H. Gundlach. The *Sorghum bicolor* genome and the diversification of grasses. *Nature*, 457(7229):551–556, 2009.

[273] W. Wang, G. Haberer, H. Gundlach, C. GläÃer, T. Nussbaumer, M. C. Luo, A. Lomsadze, M. Borodovsky, R. A. Kerstetter, J. Shanklin, D. W. Byrant, T. C. Mockler, K. J. Appenroth, J. Grimwood, J. Jenkins, J. Chow, C. Choi, C. Adam, X. H. Cao, J. Fuchs, I. Schubert, D. Rokhsar, J. Schmutz, T. P. Michael, K. F. X. Mayer, and J. Messing. The *Spirodela polyrhiza* genome reveals insights into its neotenous reduction fast growth and aquatic lifestyle. *Nature Communications*, 5:3311, 2014.

[274] S. Chamala, A. S. Chanderbali, J. P. Der, T. Lan, B. Walts, and V. A. Albert. Assembly and validation of the genome of the nonmodel basal angiosperm *Amborella*. *Science*, 342, 2013.

[275] Qunfeng Dong, Shannon D. Schlueter, and Volker Brendel. Plantgdb, plant genome database and analysis tools. *Nucleic Acids Research*, 32(suppl1):D354–D359, 2004.

[276] Lyons Eric and Freeling Michael. How to usefully compare homologous plant genes and chromosomes as dna sequences. *The Plant Journal*, 53(4):661–673, 2008.

[277] Sam Manna. An overview of pentatricopeptide repeat proteins and their applications. *Biochimie*, 113:93–99, 2015.

[278] Manisha Sharma and Girdhar K. Pandey. Expansion and function of repeat domain proteins during stress and development in plants. *Frontiers in Plant Science*, 6:1218, 2016.

[279] Haitao Xing, Xiaokang Fu, Chen Yang, Xiaofeng Tang, Li Guo, Chaofeng Li, Changzheng Xu, and Keming Luo. Genome-wide investigation of pentatricopeptide repeat gene family in poplar and their expression analysis in response to biotic and abiotic stresses. *Scientific Reports*, 8(1):2817, 2018.

[280] S. Y. Jiang, Z. Ma, and S. Ramachandran. Evolutionary history and stress regulation of the lectin superfamily in higher plants. *BMC Evol Biol*, 10(1):1–24, 2010.

[281] A. J. Afzal, A. J. Wood, and D. A. Lightfoot. Plant receptor-like serine threonine kinases: roles in signaling and plant defense. *Mol Plant Microbe Interact*, 21(5):507–517, 2008.

[282] Y. Ohizumi, M. Gaidamashvili, S. Ohwada, K. Matsuda, J. Kominami, and S. Nakamura-Tsuruta. Mannose-binding lectin from yam (*Dioscorea batatas*) tubers with insecticidal properties against *Helicoverpa armigera* (lepidoptera: Noctuidae). *J Agric Food Chem*, 57(7):2896–2902, 2009.

[283] S. Yoshimura, M. Komatsu, K. Kaku, M. Hori, T. Ogawa, and K. Muramoto. Production of transgenic rice plants expressing dioscorea batatas tuber lectin 1 to confer resistance against brown planthopper. *Plant Biotechnol*, 29(5):501–504, 2012.

[284] Y. L. Xue, T. Miyakawa, Y. Sawano, and M. Tanokura. Cloning of genes and enzymatic characterizations of novel dioscorin isoforms from *Dioscorea japonica*. *Plant Sci*, 183:14–19, 2012.

[285] M. J. Banfield and R. L. Brady. The structure of *Antirrhinum centroradialis* protein (cen) suggests a role as a kinase regulator11. *Journal of Molecular Biology*, 297(5):1159–1170, 2000.

[286] Muluneh Tamiru, Shinsuke Yamanaka, Chikako Mitsuoka, Pachakkil Babil, Hiroko Takagi, Antonio Lopez-Montes, Alieu Sartie, Robert Asiedu, and Ryohei Terauchi. Development of genomic simple sequence repeat markers for yam. *Crop Science*, 55(5):2191–2200, 2015.

[287] R. Terauchi and G. Kahl. Sex determination in *Dioscorea tokoro*, a wild yam species. In C. C. Ainsworth, editor, *Sex determination in plants*. BIOS Scientific Publishers, 1999.

[288] Manami Oyama, Tetsuo Tokiwano, Satoru Kawaii, Yasunori Yoshida, Kouichi Mizuno, Keimei Oh, and Yuko Yoshizawa. Protodioscin, isolated from the rhizome of *Dioscorea tokoro* collected in northern japan is the major antiproliferative compound to hl/-60 âšleukemic cells. *Current Bioactive Compounds*, 13(2):170–174, 2017.

[289] Yang Fei, Dan Ye, XiaoFen Fan, and FengQin Dong. Effect of *Dioscorea tokoro* makino extract on hyperuricemia in mice. *Tropical Journal of Pharmaceutical Research*, 15(9):1883–1887, 2016.

[290] Zhi-Gang Wu, Wu Jiang, Nitin Mantri, Xiao-Qing Bao, Song-Lin Chen, and Zheng-Ming Tao. Transciptome analysis reveals flavonoid biosynthesis regulation and simple sequence repeats in yam (*Dioscorea alata* l.) tubers. *BMC Genomics*, 16(1):1–12, 2015.

[291] Marie Florence Sandrine Ngo Ngwe, Denis Ndoumou Omokolo, and Simon Joly. Evolution and phylogenetic diversity of yam species (*Dioscorea* spp.): Implication for conservation and agricultural practices. *PLOS ONE*, 10(12):e0145364, 2015.

[292] M. F. S. Ngo-Ngwe, S. Joly, M. Bourge, S. Brown, and D. N. Omokolo. Nuclear dna content analysis of four cultivated species of yams (*Dioscorea* spp.) from cameroon. *Journal of Plant Breeding and Genetics*, 2(2):87–95, 2014.

[293] K. Abraham and P. Gopinathan Nair. Polyploidy and sterility in relation to sex in *Dioscorea alata* l. (dioscoreaceae). *Genetica*, 83(2):93–97, January 1991.

[294] K. Abraham, A. Nemorin, V. Lebot, and Gemma Arnau. Meiosis and sexual fertility of autotetraploid clones of greater yam *Dioscorea alata* l. *Genetic Resources and Crop Evolution*, 60(3):819–823, 2013.

[295] Viruel Juan, Segarra-Moragues José Gabriel, Raz Lauren, Forest Félix, Wilkin Paul, Sanmartín Isabel, and Catalán Pilar. Late cretaceous-early eocene origin of yams (*Dioscorea*, dioscoreaceae) in the laurasian palaearctic and their subsequent oligocene-miocene diversification. *J. Biogeogr.*, 43(4):750–762, 2016.

[296] Kenichi Kashimada and Peter Koopman. Sry: the master switch in mammalian sex determination. *Development*, 137(23):3921, 2010.

[297] Pokorná. Martina and Kratochvíl Lukáš. Phylogeny of sex-determining mechanisms in squamate reptiles: are sex chromosomes an evolutionary trap? *Zoological Journal of the Linnean Society*, 156(1):168–183, 2009.

[298] C. E. Holleley, S. D. Sarre, D. O'Meally, and A. Georges. Sex reversal in reptiles: Reproductive oddity or powerful driver of evolutionary change? Sex Dev, 2016.

[299] Y. Kobayashi, Y. Nagahama, and M. Nakamura. Diversity and plasticity of sex determination and differentiation in fishes. Sex Dev, 2013.

[300] David Ellinghaus, Stefan Kurtz, and Ute Willhoeft. Ltrharvest, an efficient and flexible software for *de novo* detection of ltr retrotransposons. *BMC Bioinformatics*, 9(1):18, 2008.

[301] Sascha Steinbiss, Ute Willhoeft, Gordon Gremme, and Stefan Kurtz. Fine-grained annotation and classification of de novo predicted ltr retrotransposons. *Nucleic Acids Research*, 37(21):7002–7013, 2009.

[302] G. Gremme, S. Steinbiss, and S. Kurtz. Genometools: A comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 10(3):645–656, 2013.

[303] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, 25, 1997.

[304] Jaina Mistry, Robert D. Finn, Sean R. Eddy, Alex Bateman, and Marco Punta. Challenges in homology search: Hmmer3 and convergent evolution of coiled-coil regions. *Nucleic Acids Research*, 41(12):e121–e121, July 2013.

[305] Carlos Llorens, Ricardo Futami, Laura Covelli, Laura Domínguez-Escribá, Jose M. Viu, Daniel Tamarit, Jose Aguilar-Rodríguez, Miguel Vicente-Ripolles, Gonzalo Fuster, Guillermo P. Bernet, Florian Maumus, Alfonso Munoz-Pomer, Jose M. Sempere, Amparo Latorre, and Andres Moya. The gypsy database (gydb) of mobile genetic elements: release 2.0. *Nucleic Acids Research*, 39:D70–D74, 2011.

[306] Robert C. Edgar. Search and clustering orders of magnitude faster than blast. *Bioinformatics*, 26(19):2460–2461, 2010.

[307] Muluneh Tamiru, Satoshi Natsume, Hiroki Takagi, Benjamen White, Hiroki Yaegashi, Motoki Shimizu, Kentaro Yoshida, Aiko Uemura, Kaori Oikawa, Akira Abe, Naoya Urasaki, Hideo Matsumura, Pachakkil Babil, Shinsuke Yamanaka, Ryo Matsumoto, Satoru Muranaka, Gezahegn Girma, Antonio Lopez-Montes, Melaku Gedil, Ranjana Bhattacharjee, Michael Abberton, P. Lava Kumar, Ismail Rabbi, Mai Tsujimura, Toru Terachi, Wilfried Haerty, Manuel Corpas, Sophien Kamoun, GÃŒnter Kahl, Hiroko Takagi, Robert Asiedu, and Ryohei Terauchi. Genome sequencing of the staple food crop white guinea yam enables the development of a molecular marker for sex determination. *BMC Biology*, 15(1):86, 2017.

[308] S. A. Goff, D. Ricke, T. . H. Lan, G. Presting, R. Wang, M. Dunn, J. Glazebrook, A. Sessions, P. Oeller, and H. Varma. A draft sequence of the rice genome (*Oryza sativa* l. ssp. japonica). *Science*, 296(5565):92–100, 2002.

[309] Mihaela Pertea, Geo M. Pertea, Corina M. Antonescu, Tsung-Cheng Chang, Joshua T. Mendell, and Steven L. Salzberg. Stringtie enables improved reconstruction of a transcriptome from rna-seq reads. *Nature Biotechnology*, 33:290, 2015.

[310] Daehwan Kim, Ben Langmead, and Steven L. Salzberg. Hisat: a fast spliced aligner with low memory requirements. *Nature Methods*, 12:357, 2015.

[311] A. Lomsadze, V. Ter-Hovhannisyan, Y. O. Chernoff, and M. Borodovsky. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucl Acids Res*, 33, 2005.

[312] Yoshihiro Kawahara, Melissa de la Bastide, John P. Hamilton, Hiroyuki Kanamori, W. Richard McCombie, Shu Ouyang, David C. Schwartz, Tsuyoshi Tanaka, Jianzhong Wu, Shiguo Zhou, Kevin L. Childs, Rebecca M. Davidson, Haining Lin, Lina Quesada-Ocampo, Brieanne Vaillancourt, Hiroaki Sakai, Sung Shin Lee, Jungsok Kim, Hisataka Numa, Takeshi Itoh, C. Robin Buell, and Takashi Matsumoto. Improvement of the *Oryza sativa* nipponbare reference genome using next generation sequence and optical map data. *Rice*, 6(1):4, 2013.

[313] Benjamin Buchfink, Chao Xie, and Daniel H. Huson. Fast and sensitive protein alignment using diamond. *Nature Methods*, 12:59, November 2014.

[314] Angelique D'Hont, France Denoeud, Jean-Marc Aury, Franc-Christophe Baurens, Francoise Carreel, Olivier Garsmeur, Benjamin Noel, Stephanie Bocs, Gaetan Droc, Mathieu Rouard, Corinne Da Silva, Kamel Jabbari, Celine Cardi, Julie Poulain, Marlene Souquet, Karine Labadie, Cyril Jourda, Juliette Lengelle, Marguerite Rodier-Goud, Adriana Alberti, Maria Bernard, Margot Correa, Saravanaraj Ayyampalayam, Michael R. McKain, Jim Leebens-Mack, Diane Burgess, Mike Freeling, Didier Mbeguie-A-Mbeguie, Matthieu Chabannes, Thomas Wicker, Olivier Panaud, Jose Barbosa, Eva Hribova, Pat Heslop-Harrison, Remy Habas, Ronan Rivallan, Philippe Francois, Claire Poiron, Andrzej Kilian, Dheema Burthia, Christophe Jenny, Frederic Bakry, Spencer Brown, Valentin Guignon, Gert Kema, Miguel Dita, Cees Waalwijk, Steeve Joseph, Anne Dievart, Olivier Jaillon, Julie Leclercq, Xavier Argout, Eric Lyons, Ana Almeida, Mouna Jeridi, Jaroslav Dolezel, Nicolas Roux, Ange-Marie Risterucci, Jean Weissenbach, Manuel Ruiz, Jean-Christophe Glaszmann, Francis Quetier, Nabila Yahiaoui, and Patrick Wincker. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature*, 488(7410):213–217, 2012.

[315] Jizeng Jia, Shancen Zhao, Xiuying Kong, Yingrui Li, Guangyao Zhao, Weiming He, Rudi Appels, Matthias Pfeifer, Yong Tao, Xueyong Zhang, Ruilian Jing, Chi Zhang, Youzhi Ma, Lifeng Gao, Chuan Gao, Manuel Spannagl, Klaus F. X. Mayer, Dong Li, Shengkai Pan, Fengya Zheng, Qun Hu, Xianchun Xia, Jianwen Li, Qinsi Liang, Jie Chen, Thomas Wicker, Caiyun Gou, Hanhui Kuang, Genyun He, Yadan Luo, Beat Keller, Qiuju Xia, Peng Lu, Junyi Wang, Hongfeng Zou, Rongzhi Zhang, Junyang Xu, Jinlong Gao, Christopher Middleton, Zhiwu Quan, Guangming Liu, Jian Wang, International Wheat Genome Sequencing Consortium, Huanming Yang, Xu Liu, Zhonghu He, Long Mao, and Jun Wang. *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature*, 496(7443):91–95, 2013.

[316] Jeanine L. Olsen, Pierre Rouzé, Bram Verhelst, Yao-Cheng Lin, Till Bayer, Jonas Collen, Emanuela Dattolo, Emanuele De Paoli, Simon Dittami, Florian Maumus, Gurvan Michel, Anna Kersting, Chiara Lauritano, Rolf Lohaus, Mats Töpel, Thierry Tonon, Kevin Vanneste, Mojgan Amirebrahimi, Janina Brakel, Christoffer Boström, Mansi Chovatia, Jane Grimwood, Jerry W. Jenkins, Alexander Jueterbock, Amy Mraz, Wytze T. Stam, Hope Tice, Erich Bornberg-Bauer, Pamela J. Green, Gareth A. Pearson, Gabriele Procaccini, Carlos M. Duarte, Jeremy Schmutz,

Thorsten B. H. Reusch, and Yves Van de Peer. The genome of the seagrass *Zostera marina* reveals angiosperm adaptation to the sea. *Nature*, 530:331, 2016.

[317] The French-Italian Public Consortium for Grapevine Genome Characterization. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, 449:463, 2007.

[318] Ray Ming, Shaobin Hou, Yun Feng, Qingyi Yu, Alexandre Dionne-Laporte, Jimmy H. Saw, Pavel Senin, Wei Wang, Benjamin V. Ly, Kanako L. T. Lewis, Steven L. Salzberg, Lu Feng, Meghan R. Jones, Rachel L. Skelton, Jan E. Murray, Cuixia Chen, Wubin Qian, Junguo Shen, Peng Du, Moriah Eustice, Eric Tong, Haibao Tang, Eric Lyons, Robert E. Paull, Todd P. Michael, Kerr Wall, Danny W. Rice, Henrik Albert, Ming-Li Wang, Yun J. Zhu, Michael Schatz, Niranjan Nagarajan, Ricelle A. Acob, Peizhu Guan, Andrea Blas, Ching Man Wai, Christine M. Ackerman, Yan Ren, Chao Liu, Jianmei Wang, Jianping Wang, Jong-Kuk Na, Eugene V. Shakirov, Brian Haas, Jyothi Thimmapuram, David Nelson, Xiyin Wang, John E. Bowers, Andrea R. Gschwend, Arthur L. Delcher, Ratnesh Singh, Jon Y. Suzuki, Savarni Tripathi, Kabi Neupane, Hairong Wei, Beth Irikura, Maya Paidi, Ning Jiang, Wenli Zhang, Gernot Presting, Aaron Windsor, Rafael Navajas-Perez, Manuel J. Torres, F. Alex Feltus, Brad Porter, Yingjun Li, A. Max Burroughs, Ming-Cheng Luo, Lei Liu, David A. Christopher, Stephen M. Mount, Paul H. Moore, Tak Sugimura, Jiming Jiang, Mary A. Schuler, Vikki Friedman, Thomas Mitchell-Olds, Dorothy E. Shippen, Claude W. dePamphilis, Jeffrey D. Palmer, Michael Freeling, Andrew H. Paterson, Dennis Gonsalves, Lei Wang, and Maqsudul Alam. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* linnaeus). *Nature*, 452(7190):991–996, 2008.

[319] Turgay Unver, Zhangyan Wu, Lieven Sterck, Mine Turktas, Rolf Lohaus, Zhen Li, Ming Yang, Lijuan He, Tianquan Deng, Francisco Javier Escalante, Carlos Llorens, Francisco J. Roig, Iskender Parmaksiz, Ekrem Dundar, Fuliang Xie, Baohong Zhang, Arif Ipek, Serkan Uranbey, Mustafa Erayman, Emre Ilhan, Oussama Badad, Hassan Ghazal, David A. Lightfoot, Pavan Kasarla, Vincent Colantonio, Huseyin Tombuloglu, Pilar Hernandez, Nurengin Mete, Oznur Cetin, Marc Van Montagu, Huanming Yang, Qiang Gao, Gabriel Dorado, and Yves Van de Peer. Genome of wild olive and the evolution of oil biosynthesis. *Proc Natl Acad Sci USA*, 114(44):E9413, 2017.

[320] Yi-Chieh Wu, Matthew D. Rasmussen, Mukul S. Bansal, and Manolis Kellis. Most parsimonious reconciliation in the presence of gene duplication, loss, and deep coalescence using labeled coalescent trees. *Genome Research*, 24(3):475–486, 2013.

[321] A. J. Enright, S. Van Dongen, and C. A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30(7):1575–1584, 2002.

[322] Tian Tian, Yue Liu, Hengyu Yan, Qi You, Xin Yi, Zhou Du, Wenying Xu, and Zhen Su. agrigo v2.0: a go analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Research*, 45(Web Server issue):W122–W129, 2017.

[323] Sarah Bastkowski, Daniel Mapleson, Andreas Spillner, Taoyang Wu, Monika Balvočiūté, and Vincent Moulton. Spectre: a suite of phylogenetic tools for reticulate evolution. *Bioinformatics*, 34(6):1056–1057, 2018.

[324] Monika Balvočūté, Andreas Spillner, and Vincent Moulton. Flatnj: A novel network-based approach to visualize evolutionary and biogeographical relationships. *Systematic Biology*, 63(3):383–396, 2014.

[325] Szymon M. Kie, Raymond Wan, Kengo Sato, Paul Horton, and Martin C. Frith. Adaptive seeds tame genomic sequence comparison. *Genome Research*, 21(3):487–493, 2011.

[326] R. Rebecca Love, Neil I. Weisenfeld, David B. Jaffe, Nora J. Besansky, and Daniel E. Neafsey. Evaluation of discovar *de novo* using a mosquito sample for cost-effective short-read genome assembly. *BMC Genomics*, 17(1):187, 2016.

[327] Joana Figueiredo, Marta Sousa Silva, and Andreia Figueiredo. Subtilisin-like proteases in plant defence: the past, the present and beyond. *Molecular Plant Pathology*, 19(4):1017–1028, 2018.

[328] Ruiqin Zhong, David H. Burk, C. Joseph Nairn, Alicia Wood-Jones, 3rd Morrison, W Herbert, and Zheng-Hua Ye. Mutation of sac1, an *Arabidopsis* sac domain phosphoinositide phosphatase, causes alterations in cell morphogenesis, cell wall synthesis, and actin organization. *The Plant cell*, 17(5):1449–1466, 2005.

[329] Andrew J. M. Howden and Gail M. Preston. Nitrilase enzymes and their role in plant-microbe interactions. *Microbial biotechnology*, 2(4):441–451, 2009.

[330] P. S. Shajeela, Veerabahu Mohan, L. Louis Jesudas, and P. Dr. Tresina. Nutritional and antinutritional evaluation of wild yam (*Dioscorea* spp.). *Tropical and Subtropical Agroecosystems*, 14:723–730, 2011.

[331] Megh Raj Bhandari and Jun Kawabata. Bitterness and toxicity in wild yam (*Dioscorea* spp.) tubers of nepal. *Plant foods for human nutrition (Dordrecht, Netherlands)*, 60:129–35, 2005.

[332] B. Krischner and H. Hahn. Patatin, a major soluble protein of the potato (solanum tuberosum l.) tuber is synthesized as a larger precursor. *Planta*, 168(3):386–389, September 1986.

[333] Kate L. Hertweck, Michael S. Kinney, Stephanie A. Stuart, Olivier Maurin, Sarah Mathews, Mark W. Chase, Maria A. Gandolfo, and J. Chris Pires. Phylogenetics, divergence times and diversification from three genomic partitions in monocots. *Botanical Journal of the Linnean Society*, 178(3):375–393, 2015.

[334] Olivier Maurin, A. Muthama Muasya, Pilar Catalan, Eugene Z. Shongwe, Juan Viruel, Paul Wilkin, and Michelle van der Bank. Diversification into novel habitats in the africa clade of *Dioscorea* (dioscoreaceae): erect habit and elephantâs foot tubers. *BMC Evolutionary Biology*, 16(1):238, 2016.

[335] Akira Kawabe, Naohiko T. Miyashita, and Ryohei Terauchi. Phylogenetic relationship among the section stenophora in the genus *Dioscorea* based on the analysis of nucleotide sequence variation in the phosphoglucose isomerase (pgi) locus:. *Genes & Genetic Systems*, 72(5):253–262, 1997.

[336] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[337] STUDENT. The probable error of a mean. *Biometrika*, 6(1):1–25, 1908.

[338] Deborah Charlesworth. Evolution of recombination rates between sex chromosomes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1736), 2017.

[339] Deborah Charlesworth. Does sexual dimorphism in plants promote sex chromosome evolution? *Environmental and Experimental Botany*, 146:5–12, 2018.

[340] Melissa Toups, Paris Veltsos, and John R. Pannell. Plant sex chromosomes: Lost genes with little compensation. *Current Biology*, 25(10):R427–R430, 2015.

[341] Ricardo S. Couto, Aline C. Martins, Mônica Bolson, Rosana C. Lopes, Eric C. Smidt, and João Marcelo A. Braga. Time calibrated tree of *Dioscorea* (dioscoreaceae) indicates four origins of yams in the neotropics since the eocene. *Botanical Journal of the Linnean Society*, page boy052, 2018.

[342] Jae-Ung Hwang, Won-Yong Song, Daewoong Hong, Donghwi Ko, Yasuyo Yamaoka, Sunghoon Jang, Sojeong Yim, Eunjung Lee, Deepa Khare, Kyungyoon Kim, Michael Palmgren, HwanÂ Su Yoon, Enrico Martinoia, and Youngsook Lee. Plant abc transporters enable many unique aspects of a terrestrial plant's lifestyle. *Molecular Plant*, 9(3):338–355, 2016.

[343] Jeffrey T. Mindrebo, Charisse M. Nartey, Yoshiya Seto, Michael D. Burkart, and Joseph P. Noel. Unveiling the functional diversity of the alpha-beta hydrolase fold in plants. *Current opinion in structural biology*, 41:233–246, 2017.

[344] Janka Puterova, Zdenek Kubat, Eduard Kejnovsky, Wojciech Jesionek, Jana Cizkova, Boris Vyskot, and Roman Hobza. The slowdown of y chromosome expansion in dioecious *Silene latifolia* due to dna loss and male-specific silencing of retrotransposons. *BMC Genomics*, 19(1):153, 2018.

[345] Doris Bachtrog. Plant sex chromosomes: A non-degenerated y? *Current Biology*, 21(18):R685–R688, 2011.

[346] Satoshi Takahata, Takumi Yago, Keisuke Iwabuchi, Hideki Hirakawa, Yutaka Suzuki, and Yasuyuki Onodera. Comparison of spinach sex chromosomes with sugar beet autosomes reveals extensive synteny and low recombination at the male-determining locus. *Journal of Heredity*, 107(7):679–685, 2016.

[347] Billy T. Lau, John M. Bell, Hanlee P. Ji, Christina Wood-Bouwens, Li C. Xia, Stephanie U. Greer, Ian D. Connolly, and Melanie H. Gephart. Chromosome-scale mega-haplotypes enable digital karyotyping of cancer aneuploidy. *nar*, 45(19):e162, 2017.

[348] Oliver Rupp, Madolyn L. MacDonald, Shangzhong Li, Heena Dhiman, Shawn Polson, Sven Griep, Kelley Heffner, Inmaculada Hernandez, Karina Brinkrolf, Vaibhav Jadhav, Mojtaba Samoudi, Haiping Hao, Brewster Kingham, Alexander

Goesmann, Michael J. Betenbaugh, Nathan E. Lewis, Nicole Borth, and Kelvin H. Lee. A reference genome of the chinese hamster based on a hybrid assembly strategy. *Biotechnology and Bioengineering*, 115(8):2087–2100, 2018.

[349] Ruibang Luo, Binghang Liu, Yinlong Xie, Zhenyu Li, Weihua Huang, Jianying Yuan, Guangzhu He, Yanxiang Chen, Qi Pan, Yunjie Liu, Jingbo Tang, Gengxiong Wu, Hao Zhang, Yujian Shi, Yong Liu, Chang Yu, Bo Wang, Yao Lu, Changlei Han, David W. Cheung, Siu-Ming Yiu, Shaoliang Peng, Zhu Xiaoqian, Guangming Liu, Xiangke Liao, Yingrui Li, Huanming Yang, Jian Wang, Tak-Wah Lam, and Jun Wang. Soapdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, 1(1):1–6, 2012.

[350] Juan Viruel, Félix Forest, Ovidiu Paun, Mark W. Chase, Dion Devey, Ricardo Sousa Couto, José Gabriel Segarra-Moragues, Pilar Catalán, and Paul Wilkin. A nuclear xdh phylogenetic analysis of yams (*Dioscorea*: Dioscoreaceae) congruent with plastid trees reveals a new neotropical lineage. *Botanical Journal of the Linnean Society*, 187(2):232–246, 2018.

[351] A. Mi Kim, Jae-Sung Rhee, H. Tae Kim, S. Jung Lee, Ah-Young Choi, Beom-Soon Choi, Ik-Young Choi, and C. Young Sohn. Alternative splicing profile and sex-preferential gene expression in the female and male pacific *Abalone haliotis* discus hannai. *Genes*, 8(3), 2017.

[352] Takashi Akagi, Isabelle M. Henry, Takashi Kawai, Luca Comai, and Ryutaro Tao. Epigenetic regulation of the sex determination gene megi in polyploid persimmon. *Plant Cell*, 28(12):2905, 2016.

[353] Nazmul Haque and Masamichi Nishiguchi. Bisulfite sequencing for cytosine-methylation analysis in plants. In *RNAi and Plant Gene Function Analysis: Methods and Protocols*, pages 187–197. Humana Press, Totowa, NJ, 2011.

[354] Katharina Bräutigam, Raju Soolanayakanahally, Marc Champigny, Shawn Mansfield, Carl Douglas, Malcolm M. Campbell, and Quentin Cronk. Sexual epigenetics: gender-specific methylation of a gene in the sex determining region of *Populus balsamifera*. *Scientific reports*, 7:45388–45388, 2017.

[355] Daniela Barros-Silva, J. C. Marques, Rui Henrique, and Carmen Jerónimo. Profiling dna methylation based on next-generation sequencing approaches: New insights and clinical applications. *Genes*, 9(9):429, 2018.

List of abbreviations

Annotation Edit Distance (AED)
Bacteria artificial chromosomes (BAC)
Base pair (bp)
Benchmarking Universal Single Copy Orthologs (BUSCO)
Burrows-Wheeler transform (BWT)
Cell-free Hi-C for Assembly and Genome Organization (Chicago)
Chinese hamster ovary (CHO)
Chromatin immunoprecipitation followed by sequencing (ChIP-seq)
Clustered regularly interspaced short palindromic repeats (CRISPR/Cas9)
Copy number variants (CNVs)
Core Eukaryotic Genes Mapping Approach (CEGMA)
Cricetulus griseus strain 17A/GY (CHO_17A/GY)
Dihydrofolate reductase (DHFR)
Environmental sex determination (ESD)
European Collection of Cell Cultures (ECCC)
EvidenceModeler (EVM)
Expression quantitative trait locus (eQTLs)
Expression sequence tags (ESTs)
Female heterogametic (ZW)
Female-specific region of the W chromosome (FSW)
FlatNetJoining (FlatNJ)
Fluorescently labelled single nucleotides (dNTPs)
Gene knockout (KO)
Gene ontology (GO)
Genome-wide association studies (GWAS)
Genotypic sex determination (GSD)

Gigabase (Gb)
Glutamine sythetase (GS)
Haploid-diploidy (UV)
Inserts and deletions (indels)
International Institute for Tropical Agriculture (IITA)
Iwate Biotechnology Research Institute (IBRC)
Kilobases (Kb)
Log2 fold change (LFC)
Long interspersed nuclear elements (LINEs)
Male Growth Inhibitor (MeGI)
Male heterogametic (XY)
Male-specific region of the Y chromosomes (MSY)
Marker Assisted Selection (MAS)
Megabases (Mb)
Million years ago (Mya)
Multiple samples to be pooled in same solution with unique indices (multiplexed) Multiple sequence alignment (MAS)
Next Generation Sequencing (NGS)
Oppressor of MeGI (OGI)
Program to Assemble Spliced Alignments (PASA)
Quantitative trait locus (QTL)
Recognition of Errors in Assemblies using Paired Reads (REAPR)
Recombinant adeno-associated virus (rAAV)
Restriction site associated DNA sequencing (RAD-seq)
Short interspersed nuclear elements (SINEs)
Short sequence repeats (SSRs)
Single molecule, real-time sequencing (SMRT)
Single nucleotide polymorphism (SNPs)
Transcription activator-like effector nucleases (TALENs)
Transposable elements (TEs)
Uniform selective pressure among sites (M0)
Untranslated regions (UTRs)
Variable selective pressure but no positive selection (M1)
Zero-mode waveguides (ZMWs)
Zinc-finger nucleases (ZFNs)

Appendix A

Table 7.1: Input data used for Ensembl Genebuild annotation.

Taxa	**Type**	**Name**	**Accession**
Cricetulus griseus	Genome Assembly	CHOK1GS_HD	GCA_900186095.1
Mus musculus	Genome assembly	GCRm38	GCA_000001635.7
Mus musculus	Reference genes	Ensembl	88_38
Rodentia	Proteins	UniProt release 2016_04	2016_04
Cricetulus griseus	RefSeq cDNAs	RefSeq	Release 76
All	MicroRNA sequences	miRBase	Release 21
All	RNA families	Rfam	Version 12.0
Cricetulus griseus	RNA-Seq reads	ENA BioProject	PRJEB14303

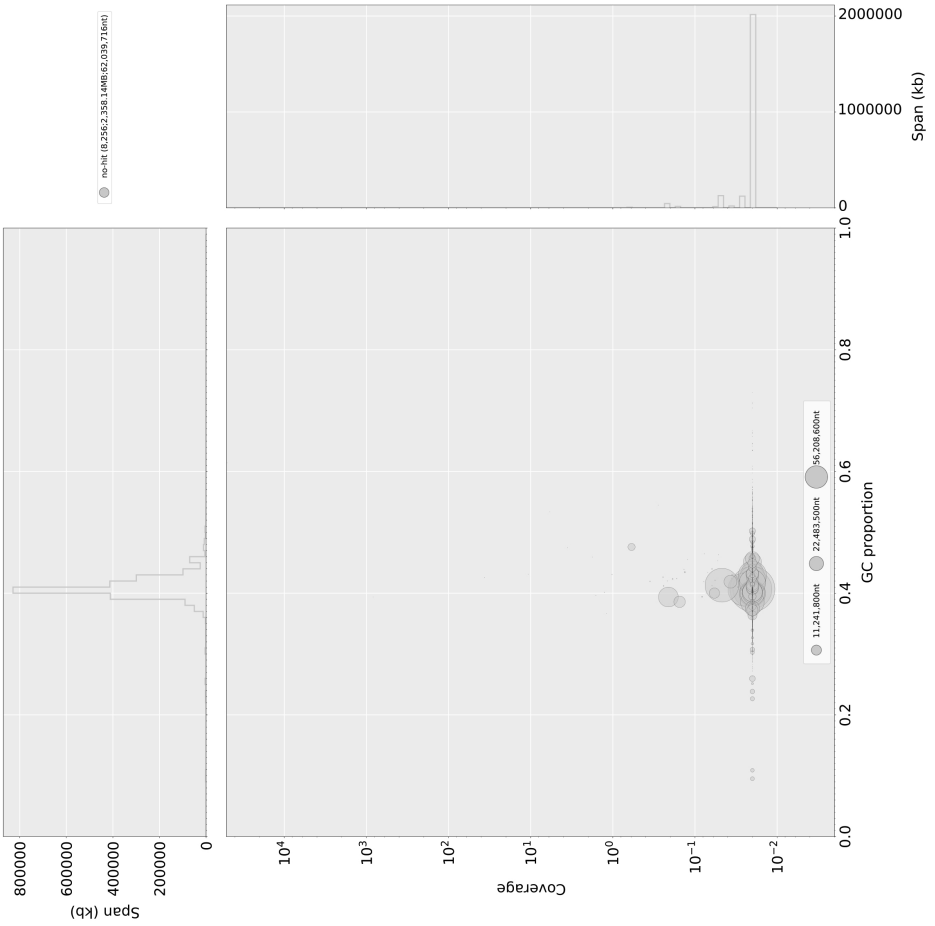

Figure 7.1: Visualisation of patterns of differential coverage signatures between SILVA rRNA sequences in the CHOK1GS_HD assembly.

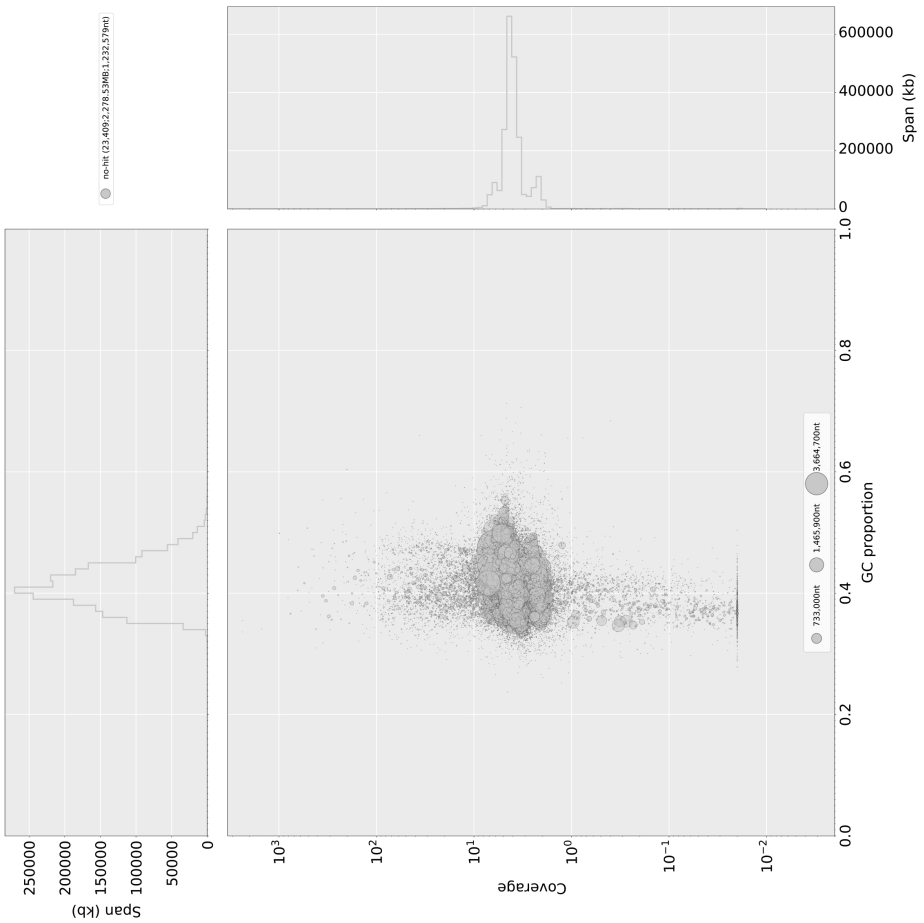

Figure 7.2: Visualisation of patterns of differential coverage signatures between SILVA rRNA sequences in the CHO_17A/GY assembly.

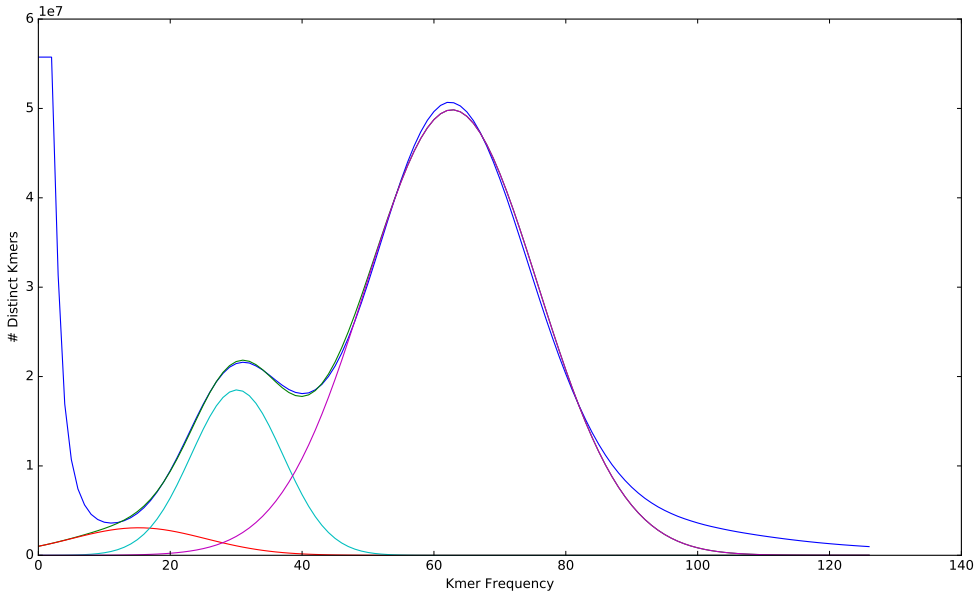

Figure 7.3: K-mer spectra of unfiltered 125 bp and 250 bp paired-end reads. Lines represent the histrogram (blue), overall fitted distribution of k-mers (green), and fit distributions for homozygous peaks 1 (red), 2 (turquoise), and 3 (purple).

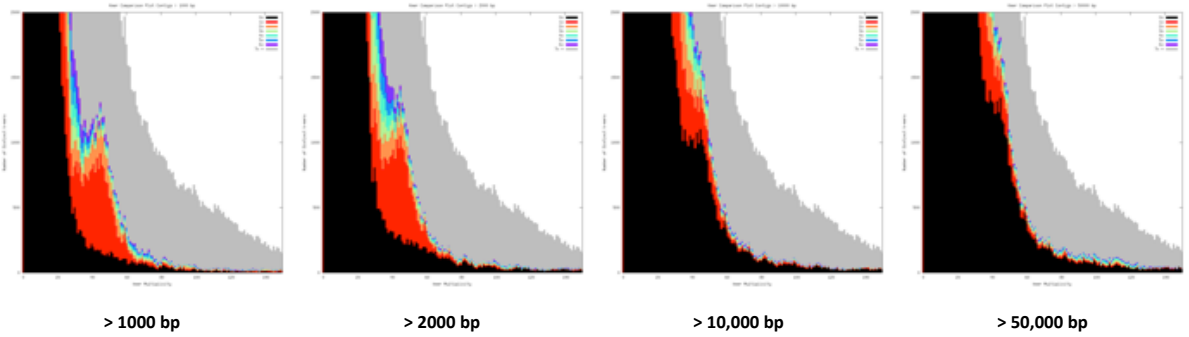

Figure 7.4: K-mer spectra analysis of CHO-K1 first pass DISCOVAR *de novo* assembly using paired-end reads mapped to the assembly at different minimum scaffold size cut offs. The black area of the graph represents content observed in the reads but not in the assembly.

Table 7.2: Table for repeats in CHOK1GS_HD and CriGri_1.0. Values are shown in Mbp.

Class	CHOK1GS_HD	CriGri_1.0
Type I Transposons/LINE	347.04	337.11
Type I Transposons/SINE	239.34	240.45
Type II Transposons	30.80	31.00
LTRs	226.13	221.14
Tandem repeats	89.15	93.49
Satellite repeats	4.45	4.03
Dust	95.62	167.03
RNA repeats	1.98	1.97
Other repeats	8.19	8.28
Unknown	4.38	4.17
Total	1,047.09	1,108.64

Table 7.3: Comparison of Ensembl pipeline annotation reults for both CHO genome.

Assembly	CHOK1GS_HD	CriGri_1.0
Gene prediction	Ensembl	Ensembl
Coding genes	20,824	19,617
Non coding genes	4,142	6,605
Small non coding genes	3,346	3,273
Long non coding genes	22	2,563
Misc non coding genes	774	769
Pseudogenes	106	446
Gene transcripts	32,575	34,472

Table 7.4: Enriched GO terms found in CHO only orthogroups.

Term ID	Description	log10 p-value
GO:0000413	protein peptidyl-prolyl isomerization	-25.4815
GO:0046942	carboxylic acid transport	-2.3188
GO:0050890	cognition	-1.8861
GO:0051606	detection of stimulus	-10.5229
GO:0006457	protein folding	-11.7959
GO:0044275	cellular carbohydrate catabolic process	-7.1135
GO:0006091	generation of precursor metabolites and energy	-1.9208
GO:0019538	protein metabolic process	-0.3468
GO:0007186	G-protein coupled receptor signaling pathway	-5.0269
GO:0006412	translation	-9.9586
GO:0015849	organic acid transport	-2.2218
GO:0030529	intracellular ribonucleoprotein complex	-12.5229
GO:0031224	intrinsic component of membrane	-0.4089
GO:0043228	non-membrane-bounded organelle	-0.4089
GO:0000786	nucleosome	-6.3872
GO:0032993	protein-DNA complex	-3.0706
GO:0043232	intracellular non-membrane-bounded organelle	-0.4089
GO:0003676	nucleic acid binding	-0.7959
GO:0003735	structural constituent of ribosome	-35.9208
GO:0003755	peptidyl-prolyl cis-trans isomerase activity	-23.0862
GO:0004871	signal transducer activity	-1.699
GO:0004984	olfactory receptor activity	-24.3768
GO:0005198	structural molecule activity	-15.4685
GO:0060089	molecular transducer activity	-1.699
GO:0016853	isomerase activity	-8.5229
GO:0003723	RNA binding	-0.8539

Table 7.5: Enriched GO terms found in orthogroups shared between CriGri_1.0 and at least one other species, but not CHOK1GS_HD.

Term ID	Description	log10 p-value
GO:0003824	catalytic activity	-0.585
GO:0016740	transferase activity	-0.585
GO:0043167	ion binding	-0.585
GO:0046872	metal ion binding	-0.585
GO:0016787	hydrolase activity	-0.4949

Table 7.6: SNPs and Indels of CHOK1GS_HD aligned to CHO mitocondria reference.

Chrom	Position	Ref	Cons
MT	9306	A	G
MT	11399	C	A
MT	12683	G	A
MT	13436	A	-T/-T
MT	13456	C	+A/+A
MT	15717	G	-C/-C
MT	15888	G	T

Table 7.7: Full comparison of assembly metrics, data type and assembler used for different iterations of the CHO genome assembly, and other publicly available *C. griseus* assemblies. Scaffolds referes to contigs in the case of SGA, where no scaffolding was performed.

Assembly	# contigs (>= 0 bp)	# contigs (>= 0.05 Mbp)	Total length (>= 0 bp)	Total length (>= 0.50 Mbp)	#scaffolds	Longest scaffold (Mbp)	Total length (Mbp)	GC (%)	N50 (Kb)	# N's per 100 kbp	BUSCOs Present (%)
SGA	7,229,676	4	399.13	0.22	774,852	0.064	2,256.22	41.45	5	0	21.00
Dioscovar *de novo*	877,181	10823	260.63	1,832.26	169,068	1.65	2,410.21	41.45	158	63.59	79.20
Dioscovar *de novo* + Dovetail	821,041	433	261.15	2,286.37	112,927	157.32	2,415.46	41.45	34,102	265.27	95.80
Dioscovar *de novo* + Dovetail + Bionano	820,943	335	263.96	2,314.47	112,829	224.8	2,443.56	41.45	61,985	1,411.88	95.30
Dioscovar *de novo* + Dovetail + Bionano + SGA	151,867	337	245.35	2,315.43	88,765	224.83	2,428.16	41.45	62,039	1,410.29	95.30
Dioscovar *de novo* + Dovetail + Bionano + SGA + No < 2 kb Scaffolds	8,265	337	235.81	2,315.43	8,265	224.83	2,358.16	41.45	62,039	1,452.08	95.60
CriGri$_{1}$.0	109,151	4360	239.97	2,246.51	55,010	8.77	2,383.17	41.37	1,165	3,419.18	93.10
BGI *C. griseus*	52,710	2515	236.01	2,294.87	20,918	8.32	2,351.10	41.39	1,571	2,501.81	94.70
CHO_17A/GY	28,749	3391	233.27	2,216.75	28,749	14.65	2,332.77	41.27	1,236	10,450.23	93.00
H. sapiens GRCh38	194	78	309.97	3,098.61	194	248.95	3,099.75	40.86	145,138	4,964.97	92.60
M. musculu GRCm38	66	43	273.08	2,730.23	66	195.47	2,730.87	41.67	130,694	2,859.46	95.30

Table 7.8: Enriched GO terms found in orthogroups shared between CHOK1GS_HD and at least one other species, but not CriGri_1.0.

Term_ID	Description	log10 p-value
GO:0032501	multicellular organismal process	-4.8861
GO:0032502	developmental process	-3.0809
GO:0045165	cell fate commitment	-12.1308
GO:0065007	biological regulation	-1.4685
GO:0033108	mitochondrial respiratory chain complex assembly	-1.585
GO:0014855	striated muscle cell proliferation	-2.4437
GO:0007219	Notch signaling pathway	-5.3565
GO:0008283	cell proliferation	-1.4089
GO:0001508	action potential	-1.5528
GO:0043536	positive regulation of blood vessel endothelial cell migration	-2.7212
GO:0000122	negative regulation of transcription from RNA polymerase II promoter	-4.9586
GO:0050890	cognition	-2.7212
GO:0051606	detection of stimulus	-5.0269
GO:0019222	regulation of metabolic process	-3.0862
GO:0007530	sex determination	-3.3279
GO:0031128	developmental induction	-1.4815
GO:0060021	palate development	-1.585
GO:0003158	endothelium development	-3.9586
GO:0003008	system process	-4.6021
GO:0035270	endocrine system development	-10.2596
GO:0051674	localization of cell	-1.3872
GO:0048518	positive regulation of biological process	-2
GO:0010467	gene expression	-2.5686

GO:0007186	G-protein coupled receptor signaling pathway	-3.0044
GO:0003002	regionalization	-7.5528
GO:0072006	nephron development	-4.6778
GO:0055123	digestive system development	-1.6383
GO:0009058	biosynthetic process	-1.7212
GO:0048736	appendage development	-2.1135
GO:0060173	limb development	-2.1135
GO:0006807	nitrogen compound metabolic process	-1.4437
GO:0001655	urogenital system development	-3.284
GO:0048663	neuron fate commitment	-9.4318
GO:0071156	regulation of cell cycle arrest	-2.6778
GO:0007389	pattern specification process	-7.041
GO:0031018	endocrine pancreas development	-3.284
GO:0035239	tube morphogenesis	-4.301
GO:0050794	regulation of cellular process	-2.1938
GO:0060541	respiratory system development	-2.1308
GO:0031016	pancreas development	-3
GO:0051239	regulation of multicellular organismal process	-2.9208
GO:0035295	tube development	-4.2147
GO:0001501	skeletal system development	-2.0269
GO:0001894	tissue homeostasis	-1.3372
GO:0043009	chordate embryonic development	-4.0555
GO:0030278	regulation of ossification	-1.6778
GO:0048732	gland development	-4.4437
GO:0007422	peripheral nervous system development	-1.6778

GO:0045597	positive regulation of cell differentiation	-4.3279
GO:0051171	regulation of nitrogen compound metabolic process	-5.7959
GO:0050789	regulation of biological process	-1.9586
GO:0021510	spinal cord development	-9.301
GO:0090100	positive regulation of transmembrane receptor protein serine/threonine kinase signaling pathway	-1.7959
GO:0090090	negative regulation of canonical Wnt signaling pathway	-2.1487
GO:0007166	cell surface receptor signaling pathway	-3.0044
GO:0043010	camera-type eye development	-3.8539
GO:0032722	positive regulation of chemokine production	-1.4202
GO:0007517	muscle organ development	-1.7212
GO:0014032	neural crest cell development	-2.3872
GO:0042698	ovulation cycle	-1.4089
GO:0033002	muscle cell proliferation	-1.8539
GO:0007492	endoderm development	-2.3665
GO:0007423	sensory organ development	-3.7696
GO:0007399	nervous system development	-5.5376
GO:0009790	embryo development	-4.2076
GO:0009888	tissue development	-3.4815
GO:0050905	neuromuscular process	-1.4949
GO:0007498	mesoderm development	-2.7212
GO:0003007	heart morphogenesis	-2.7959
GO:0010628	positive regulation of gene expression	-6.4949
GO:0007165	signal transduction	-2.3768

GO:0007417	central nervous system development	-6.2147
GO:0042552	myelination	-1.5528
GO:0007272	ensheathment of neurons	-1.4815
GO:0009892	negative regulation of metabolic process	-2.7696
GO:0003702	(obsolete) RNA polymerase II transcription factor activity	-6.2518
GO:0003705	transcription factor activity, RNA polymerase II distal enhancer sequence-specific binding	-6.2518
GO:0004871	signal transducer activity	-2.1024
GO:0004930	G-protein coupled receptor activity	-6.2291
GO:0016782	transferase activity, transferring sulfur-containing groups	-2.2924
GO:0030528	(obsolete) transcription regulator activity	-7.9208
GO:0043565	sequence-specific DNA binding	-8.284
GO:0060089	molecular transducer activity	-2.1024
GO:0003682	chromatin binding	-2.1675
GO:0008146	sulfotransferase activity	-1.9208

Table 7.9: List of PPRs conserved between the seven species.

D. rotundata	*A. thaliana*	*B. distachyon*
O. sativa	*E. guineensis*	*P. dactylifera*
M. acuminata		
Dr03551	AT1G59720	Bradi1g54840
LOC_Os07g09370	LOC105046423	LOC103712051
Ma10_g09870		
Dr01916	AT1G08610	Bradi3g28747

LOC_Os10g33700	LOC105033622	LOC103716082
Ma11_g14980		
Dr10623	AT5G43120	Bradi1g70670
LOC_Os03g10980	LOC105044680	LOC103707390
Ma04_g07030		
Dr13510	AT3G29290	Bradi1g70330
LOC_Os03g11310	LOC105055996	LOC103709263
Ma04_g14270		
Dr04443	AT1G02420	Bradi1g70637
LOC_Os03g11020	LOC105059773	LOC103705387
Ma05_g08760		
Dr02888	AT3G04260	Bradi3g28060
LOC_Os10g32540	LOC105053651	LOC103723700
Ma05_g22970		
Dr00590	AT2G20710	Bradi3g30100
LOC_Os10g35760, LOC_Os10g35750	LOC105042850	LOC103712094
Ma05_g23170		
Dr19913	AT1G62350	Bradi1g65120
LOC_Os03g18620	LOC105057161	LOC103696433
Ma08_g01690		
Dr10261	AT5G08310	Bradi2g31210
LOC_Os05g24930	LOC105053725	LOC103712455
Ma05_g10640		
Dr00932	AT4G14820	Bradi4g15307
LOC_Os04g14450	LOC105042305	LOC103718081
Ma00_g04220, Ma08_g27470		
Dr16228	AT2G33760	Bradi2g50550
LOC_Os12g08140	LOC105043669	LOC103721708
Ma04_g20460		
Dr10696	AT5G64320	Bradi3g24450
LOC_Os10g22370	LOC105056202	LOC103713375
Ma04_g03610		
Dr24397	AT1G26460	Bradi1g22510

LOC_Os07g40800	LOC105055974	LOC103709240, LOC103708246
Ma01_g07080		
Dr13752, Dr13263	AT1G71490	Bradi3g13540, Bradi1g52620
LOC_Os08g02040, LOC_Os07g19400	LOC105036355, LOC105049791	LOC103724073, LOC103719495
Ma11_g17830		
Dr14089	AT3G57430	Bradi5g01440
LOC_Os04g02850	LOC105044242	LOC103701583
Ma08_g15220		
Dr04441	AT1G13040	Bradi1g17190
LOC_Os07g48850	LOC105044971	LOC103701141
Ma03_g08310		
Dr12768	AT2G39380, AT3G09520, AT3G55150	Bradi5g06870
LOC_Os11g01050, LOC_Os12g01040	LOC105045752, LOC105057006	LOC103702815, LOC103706626
Ma09_g03090, Ma06_g36330, Ma10_g05510		
Dr06523	AT3G20730	Bradi5g22640
LOC_Os04g53630	LOC105052646	LOC103703482
Ma11_g24190		
Dr12962	AT1G80150	Bradi3g39320
LOC_Os08g39050	LOC105038992	LOC103718943
Ma04_g09760		
Dr21506	AT2G35030	Bradi1g51400
LOC_Os11g24530	LOC105039093	LOC103718518
Ma07_g21000		
Dr11620	AT5G27460	Bradi3g54100
LOC_Os02g55310	LOC105060569	LOC103719090
Ma07_g12650		
Dr08937	AT1G06710	Bradi5g14060
LOC_Os04g41140	LOC105051420	LOC103717529
Ma05_g17800		

Dr00433	AT1G60770	Bradi1g49750
LOC_Os04g25410	LOC105047099, LOC105047731	LOC103715940, LOC103715138
Ma01_g09490		
Dr10599	AT1G25360	Bradi1g22990
LOC_Os07g39910	LOC105057922, LOC105060730	LOC103705338, LOC103718592
Ma03_g31320		
Dr02052	AT5G12100	Bradi3g35257
LOC_Os08g31110	LOC105032139, LOC105045245, LOC105032138	LOC103724022, LOC103698576, LOC103699462
Ma01_g19430		
Dr21963	AT1G19720	Bradi3g35210
LOC_Os03g59264	LOC105041986	LOC103712578
Ma05_g24450		
Dr10683	AT4G26680	Bradi3g03280
LOC_Os10g21920	LOC105032794	LOC103714917
Ma01_g06210		
Dr07198	AT3G04130	Bradi3g47936
LOC_Os02g38950	LOC105045426	LOC103718684
Ma01_g19170		
Dr14998	AT1G01970	Bradi1g18720
LOC_Os07g46730	LOC105045324	LOC103722851
Ma11_g18730		
Dr11209	AT5G03800	Bradi1g56040
LOC_Os07g07620	LOC105034986	LOC103718370
Ma04_g29200		
Dr09364	AT2G27800	Bradi2g49720
LOC_Os01g54380	LOC105040964	LOC103721272
Ma06_g22690		
Dr09107	AT3G25970	Bradi1g21907
LOC_Os08g28830	LOC105047750	LOC103721026
Ma06_g36440		
Dr07998	AT1G74900	Bradi3g51690
LOC_Os02g45590	LOC105055946	LOC103707645

Ma04_g10770		
Dr06719	AT2G15630, AT1G09680	Bradi5g12557, Bradi5g13435
LOC_Os04g40130, LOC_Os04g38930	LOC105056471	LOC103706834, LOC103712911
Ma06_g32300, Ma08_g19310		
Dr13844	AT1G79540	Bradi4g08330
LOC_Os09g02260	LOC105060988	LOC103716239
Ma01_g00470		
Dr15431	AT3G21470	Bradi2g56860
LOC_Os01g65840	LOC105034219	LOC103717186
Ma07_g24830		
Dr10294, Dr12374	AT1G79080, AT1G09900	Bradi3g46270
LOC_Os02g35750	LOC105041810, LOC105032037	LOC103721349
Ma05_g26910, Ma04_g38360		
Dr04561	AT3G49170	Bradi2g32290
LOC_Os05g12130	LOC105040701	LOC103703045
Ma06_g15810		
Dr01328	AT1G53330	Bradi2g32660
LOC_Os05g11700	LOC105040591	LOC103707348
Ma10_g21460		
Dr10589	AT2G16880	Bradi1g71082
LOC_Os03g10420	LOC105057943	LOC103720762
Ma01_g15870		
Dr20196	AT2G06000	Bradi5g11010
LOC_Os04g36840	LOC105041853	LOC103704922, LOC103697477
Ma03_g07900		
Dr13868	AT5G18390	Bradi3g40880
LOC_Os08g41380	LOC105048893	LOC103709441
Ma11_g13890		
Dr07191	AT3G13880	Bradi2g61387
LOC_Os01g72930	LOC105051725	LOC103704879

Ma06_g18950		
Dr13344	AT1G77010	Bradi4g44720
LOC_Os03g63260	LOC105037526	LOC103722223
Ma03_g22940		
Dr00599	AT1G09820	Bradi3g29181
LOC_Os10g34310	LOC105034568	LOC103719133
Ma03_g01250		
Dr08990, Dr12896, Dr12894	AT3G27750	Bradi2g00270, Bradi2g00245, Bradi5g13310
LOC_Os04g39970	LOC105057077, LOC105052840, LOC105034378	LOC103714745, LOC103700440
Ma09_g04330, Ma06_g38260, Ma08_g19090		
Dr10966	AT3G09040	Bradi2g37360
LOC_Os04g43430	LOC105054468	LOC103703917
Ma04_g28590		
Dr20186	AT3G18110	Bradi1g46560
LOC_Os06g09880	LOC105053071	LOC103703300
Ma10_g31140		
Dr15761	AT4G16835	Bradi3g16967
LOC_Os08g05750	LOC105054749	LOC103712428, LOC103696236
Ma03_g16820		
Dr16913	AT3G02330	Bradi3g32282
LOC_Os10g40920	LOC105050663	LOC103702444
Ma04_g09500		
Dr00056	AT1G07590	Bradi3g16227
LOC_Os08g06500	LOC105061335	LOC103705034
Ma09_g23660		
Dr00305	AT3G08820	Bradi4g44712
LOC_Os12g01850, LOC_Os11g01836	LOC105045919	LOC103714412
Ma05_g31870		

Dr12437	AT2G02980	Bradi1g05610
LOC_Os03g58100	LOC105041736	LOC103718806
Ma04_g22720		
Dr02236	AT1G30610	Bradi2g46360
LOC_Os01g48380	LOC105049056	LOC103719614
Ma03_g24700		
Dr03677	AT1G28690	Bradi1g61520
LOC_Os03g25380	LOC105049864	LOC103723658
Ma10_g08830		
Dr17728, Dr22720, Dr2193	AT3G51320	Bradi2g39487
LOC_Os05g01635	LOC105061162	LOC103707216
Ma03_g14060		
Dr15744	AT3G53170	Bradi1g77530
LOC_Os03g02430	LOC105054102	LOC103723676
Ma11_g20120		
Dr15037	AT5G46580	Bradi1g01021
LOC_Os03g63910	LOC105053878	LOC103704682
Ma11_g00040		
Dr11610, Dr09981	AT1G08070	Bradi1g64170, Bradi5g13537
LOC_Os03g19980	LOC105045112, LOC105038828	LOC103723163, LOC103717604
Ma03_g08290, Ma05_g17640		
Dr04059	AT5G15280	Bradi1g73810
LOC_Os03g07220	LOC105041752	LOC103718825
Ma06_g01060		
Dr08768	AT4G11860	Bradi1g34980
LOC_Os06g49800	LOC105034131	LOC103710635
Ma09_g21240		
Dr00308	AT4G19440	Bradi3g19897
LOC_Os08g19310	LOC105052253, LOC105045913	LOC103714407
Ma11_g22050		
Dr09686	AT4G18975	Bradi1g25360
LOC_Os07g36180	LOC105050238	LOC103695916

Ma09_g27390		
Dr07515	AT4G39952	Bradi5g12610
LOC_Os04g38980	LOC105056857	LOC103716129
Ma10_g30340		
Dr08269	AT5G61990	Bradi1g51377
LOC_Os07g20510	LOC105036528	LOC103707812
Ma08_g00540		
Dr13418	AT1G76280	Bradi5g02047
LOC_Os09g29790	LOC105032890	LOC103717709
Ma09_g28750		
Dr11982	AT5G56310	Bradi2g09441
LOC_Os01g15530	LOC105057267	LOC103709663
Ma09_g17020		
Dr09451	AT5G04780	Bradi4g42160
LOC_Os12g06070	LOC105052093	LOC103714260
Ma09_g09040		
Dr08848	AT5G27270	Bradi1g50250
LOC_Os06g02120	LOC105060646	LOC103695800
Ma03_g14090		
Dr11704	AT4G22760	Bradi3g16240
LOC_Os08g06490	LOC105059396	LOC103717401
Ma03_g11090		
Dr01129	AT1G30290	Bradi4g01620
LOC_Os12g42120	LOC105059241	LOC103695846
Ma05_g02420		
Dr04736	AT4G21065	Bradi1g58370
LOC_Os07g05560	LOC105046723	LOC103710531
Ma09_g26630		
Dr16430	AT3G46790	Bradi2g54927
LOC_Os01g62910	LOC105056879	LOC103722550, LOC103716203
Ma02_g21010		
Dr11315	AT4G14050	Bradi2g25180
LOC_Os11g43934	LOC105033180	LOC103709405
Ma08_g00890		

Dr21693	AT3G26540	Bradi5g10630
LOC_Os04g35650	LOC105038480	LOC103713823
Ma06_g25470		
Dr21510	AT5G10690	Bradi4g30897
LOC_Os09g26190	LOC105039617	LOC103718395
Ma04_g02160		
Dr01885	AT1G33350	Bradi3g48960
LOC_Os02g40750	LOC105044169	LOC103715434
Ma05_g09870		
Dr19390	AT5G14770	Bradi3g55920
LOC_Os02g57800	LOC105054862	LOC103705544
Ma11_g10980		
Dr07988	AT4G16390	Bradi1g69827
LOC_Os03g11670	LOC105056047	LOC103707571
Ma07_g28260		
Dr07676	AT3G48250	Bradi2g61970
LOC_Os01g73950	LOC105051245	LOC103717565
Ma08_g29100		
Dr11621	AT5G09450	Bradi5g11860
LOC_Os04g37720	LOC105058470	LOC103723172
Ma10_g24880		
Dr01247	AT3G49240	Bradi2g42630
LOC_Os11g24570	LOC105042725	LOC103714206
Ma06_g28300		
Dr03210	AT2G34400	Bradi3g03440
LOC_Os05g24150	LOC105049298	LOC103713111
Ma05_g04710		
Dr13293	AT3G61170	Bradi2g57140
LOC_Os01g66160	LOC105037194	LOC103719486
Ma11_g24830		
Dr15908	AT1G77405	Bradi1g09470
LOC_Os03g52620	LOC105053466	LOC103697883
Ma01_g16850		
Dr02295	AT4G35850	Bradi2g36217
LOC_Os05g05320	LOC105039811	LOC103719701

Ma08_g06620		
Dr03344	AT2G03380	Bradi1g47170
LOC_Os06g08650	LOC105050742	LOC103710447
Ma10_g23870		
Dr16727	AT2G17120	Bradi1g76177, Bradi4g37090
LOC_Os09g37600, LOC_Os03g04110	LOC105055192	LOC103701084
Ma03_g00470		
Dr04470	AT2G29760, AT3G15930	Bradi2g04737
LOC_Os01g08120	LOC105035889, LOC105047860	LOC103704100, LOC103711226
Ma05_g31270		
Dr11765	AT1G02060	Bradi3g02140
LOC_Os02g02950	LOC105035673	LOC103697383
Ma03_g15480		
Dr13883	AT5G66500	Bradi1g22070
LOC_Os03g42650	LOC105049608	LOC103715219
Ma06_g01450		
Dr04360	AT1G61870	Bradi2g52250
LOC_Os01g58080	LOC105046876	LOC103717141
Ma06_g02110		
Dr14532	AT4G30700	Bradi1g47160
LOC_Os06g08660	LOC105032313	LOC103707467
Ma06_g00210		
Dr09453	AT5G04810	Bradi5g26670
LOC_Os04g58780	LOC105035288	LOC103721419
Ma07_g03390		
Dr19401	AT5G16860	Bradi4g40340
LOC_Os05g23960, LOC_Os04g14130	LOC105038615	LOC103708580
Ma05_g17870		
Dr01319	AT2G39620, AT4G18750	Bradi2g06450
LOC_Os01g10800	LOC105041471, LOC105046298	LOC103704081, LOC103707385
Ma10_g13920		

Dr16412	AT2G37230	Bradi3g01116
LOC_Os02g02020	LOC105033422	LOC103719875
Ma01_g04400		
Dr18323	AT3G23020	Bradi1g74185
LOC_Os03g06710	LOC105053421	LOC103697863
Ma08_g27620		
Dr12672	AT3G25060	Bradi3g32220
LOC_Os10g39460	LOC105051140	LOC103710845
Ma08_g16460		
Dr02389	AT3G53700	Bradi4g14080
LOC_Os03g40020	LOC105046909	LOC103717152
Ma06_g23830		
Dr02008	AT3G16890	Bradi4g04840
LOC_Os12g37100	LOC105040049	LOC103712493
Ma02_g03690		
Dr03358	AT4G31850	Bradi1g09357
LOC_Os10g28600	LOC105060160	LOC103707470
Ma07_g16270		
Dr00260	AT4G28010	Bradi1g21250
LOC_Os07g42880	LOC105042482	LOC103707178
Ma03_g30820		
Dr08014	AT5G60960	Bradi1g22060
LOC_Os07g41260	LOC105058105	LOC103704205
Ma02_g08160		
Dr09089	AT4G20740	Bradi1g55540
LOC_Os07g08180	LOC105057738	LOC103714692
Ma05_g17250		
Dr01908	AT3G23330	Bradi1g20737
LOC_Os07g39090	LOC105033605	LOC103717598
Ma08_g15000		
Dr03198	AT3G42630	Bradi3g40470
LOC_Os08g40870	LOC105039823	LOC103724305
Ma10_g02870		
Dr06505	AT2G17033	Bradi3g01770
LOC_Os02g02770	LOC105055155	LOC103701150

Ma07_g04270		
Dr13769	AT3G02650	Bradi2g57630
LOC_Os01g67210	LOC105042274, LOC105032638	LOC103720570
Ma07_g02940		
Dr14115	AT1G18900, AT1G74750	Bradi4g00490
LOC_Os12g44170	LOC105054119	LOC103721968
Ma08_g20130		
Dr07732	AT5G50280	Bradi1g27140
LOC_Os05g22870	LOC105038743	LOC103720297
Ma08_g29250		
Dr12436	AT1G74600	Bradi4g30190
LOC_Os09g24680	LOC105058314	LOC103716405
Ma05_g19540		
Dr10717	AT1G20230	Bradi2g27440
LOC_Os05g30710	LOC105044013	LOC103721154
Ma05_g11260		
Dr15662	AT3G62470, AT3G62540, AT5G14820	Bradi2g28390
LOC_Os05g28720	LOC105045045	LOC103700962
Ma09_g15870		
Dr04223	AT1G10910	Bradi2g41260
LOC_Os01g37870	LOC105037320	LOC103712971
Ma03_g27170		
Dr16978	AT3G13770	Bradi1g36010
LOC_Os06g41040	LOC105050501	LOC103702307
Ma03_g11550		
Dr10197	AT4G10590, AT4G10570	Bradi3g20790
LOC_Os10g07270	LOC105034613	LOC103722919
Ma04_g28400		
Dr22661, Dr17671	AT5G03560	Bradi5g13440, Bradi4g21200
LOC_Os08g44650, LOC_Os06g20354	LOC105034399, LOC105059059	LOC103698744, LOC103720703
Ma07_g18250, Ma05_g20820		

Dr00481	AT3G16610	Bradi2g62740
LOC_Os01g74600	LOC105036235	LOC103698290
Ma05_g11530		
Dr09691, Dr00858	AT1G55890, AT3G13160	Bradi2g52016, Bradi2g52030
LOC_Os01g57630	LOC105035895	LOC103712627
Ma10_g10310		
Dr09718	AT4G39530	Bradi2g62180
LOC_Os08g25280	LOC105052721	LOC103709934
Ma01_g14980		
Dr11148	AT5G06400	Bradi4g07790
LOC_Os12g27060	LOC105057916	LOC103706603
Ma07_g07790		
Dr00869	AT5G50390	Bradi3g14600
LOC_Os08g04400	LOC105052655	LOC103703546
Ma03_g24250		
Dr12507	AT2G17670	Bradi3g01817
LOC_Os02g02740	LOC105055184	LOC103701157
Ma01_g11380		
Dr21292	AT2G41720	Bradi1g54007
LOC_Os07g11280	LOC105045536	LOC103708030
Ma07_g22690		
Dr08001	AT1G69290	Bradi2g52170
LOC_Os01g57900	LOC105056104	LOC103710275
Ma09_g06870		
Dr19070	AT4G21300	Bradi1g06766
LOC_Os03g56850	LOC105033905	LOC103709824
Ma04_g38430		
Dr15832	AT4G17616, AT1G03100	Bradi1g26970, Bradi3g27027
LOC_Os10g28665, LOC_Os07g31310	LOC105053369, LOC105047213	LOC103715009
Ma04_g03000, Ma05_g20530		
Dr07322	AT4G20770	Bradi3g05000
LOC_Os02g07050	LOC105049154	LOC103719625

Ma08_g06860		
Dr19306	AT5G08490	Bradi5g10954, Bradi3g00880
LOC_Os02g18810	LOC105038694	LOC103711103
Ma03_g18250		
Dr00114	AT2G20540	Bradi2g24060
LOC_Os05g36350	LOC105037996	LOC103704406
Ma05_g28510		
Dr04113	AT3G46610	Bradi4g39100
LOC_Os12g18640	LOC105038736	LOC103720303
Ma03_g17830		
Dr10632	AT5G67570	Bradi1g15900
LOC_Os05g25060	LOC105057954	LOC103722764
Ma07_g27680		
Dr19814	AT1G71210	Bradi3g02670
LOC_Os02g03530	LOC105054926	LOC103706396
Ma07_g16970		
Dr17305	AT2G18940	Bradi1g31370
LOC_Os05g19380	LOC105039276	LOC103701530
Ma02_g06830		
Dr13673	AT3G63370	Bradi5g12970
LOC_Os04g39410	LOC105060773	LOC103711434
Ma03_g04060		
Dr12741	AT1G16480	Bradi3g37252
LOC_Os11g45410	LOC105045697	LOC103716520
Ma08_g02410		
Dr02258	AT1G53600	Bradi3g14380
LOC_Os08g03676	LOC105037953	LOC103705668
Ma05_g28940		
Dr06199, Dr05628	AT2G32230	Bradi3g10160, Bradi5g27596
LOC_Os04g59600, LOC_Os02g17360	LOC105037047, LOC105040834	LOC103702617, LOC103696071
Ma10_g22060, Ma03_g23070		
Dr06567	AT1G73710	Bradi1g28220

LOC_Os07g28900	LOC105035192	LOC103720934
Ma08_g07100		
Dr20432	AT1G05670	Bradi1g26330
LOC_Os07g32900	LOC105042441	LOC103707767
Ma08_g04680		
Dr14069	AT1G03780	Bradi1g26680
LOC_Os07g32390	LOC105050790, LOC105061566	LOC103707173, LOC103721945
Ma08_g19570		
Dr07647	AT5G48730	Bradi1g48080
LOC_Os06g07550	LOC105051223	LOC103718446
Ma01_g08590		
Dr17089	AT5G08510	Bradi1g49477
LOC_Os06g03570	LOC105059308	LOC103716720
Ma01_g12010		
Dr18651	AT4G37170	Bradi1g68870
LOC_Os03g13230	LOC105058660	LOC103705863
Ma01_g15480		
Dr01430	AT2G15690	Bradi2g14520
LOC_Os04g09530	LOC105042916	LOC103721816
Ma04_g15330, Ma04_g10450		
Dr09697	AT5G42310	Bradi1g28870
LOC_Os07g36390	LOC105050369	LOC103707165
Ma09_g11680		
Dr17300	AT3G61360	Bradi1g33710
LOC_Os06g47950	LOC105039669	LOC103718057
Ma10_g06300		
Dr09987	AT1G66345	Bradi2g04397
LOC_Os01g07610	LOC105058797	LOC103711574
Ma03_g21590		
Dr11601	AT5G18475	Bradi1g45780
LOC_Os02g26890	LOC105060580	LOC103716179
Ma05_g05280		
Dr04424	AT4G01990, AT1G02370	Bradi2g11510

LOC_Os01g19490	LOC105037070, LOC105045017	LOC103700987, LOC103700989
Ma11_g04800		
Dr08802	AT1G02150	Bradi5g17360
LOC_Os04g46010	LOC105053926, LOC105043771	LOC103704609, LOC103710792
Ma05_g15960		
Dr12327	AT2G30780	Bradi2g25590
LOC_Os05g33760	LOC105060460	LOC103712322
Ma08_g05770, Ma03_g26120		
Dr20082	AT2G41080	Bradi2g55230
LOC_Os03g60200	LOC105035947	LOC103697509
Ma10_g22420		
Dr15456	AT1G15510	Bradi2g15700
LOC_Os05g49920	LOC105041666	LOC103697277
Ma06_g03330		
Dr16916	AT1G68930	Bradi3g00900
LOC_Os02g01610	LOC105060293	LOC103712772
Ma04_g26200		
Dr00824, Dr14060	AT5G06540, AT5G40405	Bradi2g48180, Bradi1g45070
LOC_Os01g51790	LOC105047803, LOC105061500, LOC105045439	LOC103720535, LOC103717324, LOC103707685
Ma07_g00030, Ma09_g16050		
Dr02571	AT1G80880	Bradi3g19530
LOC_Os08g17080	LOC105038169	LOC103723814
Ma07_g29100		
Dr00589	AT4G21705	Bradi3g30090
LOC_Os07g40750	LOC105036045	LOC103698308
Ma04_g22740		
Dr09692	AT3G46870	Bradi1g25240
LOC_Os07g36450	LOC105050296	LOC103723871
Ma02_g15860		

Dr09113	AT1G07740	Bradi2g17702
LOC_Os05g47510	LOC105045335	LOC103722857
Ma08_g09660		
Dr06055	AT4G37380	Bradi1g59070
LOC_Os07g02280	LOC105033508	LOC103700909
Ma04_g34650		
Dr15492	AT5G09950	Bradi2g42310
LOC_Os01g40720	LOC105049115	LOC103697638
Ma09_g19050		
Dr09119	AT4G14170	Bradi2g25480
LOC_Os05g33920	LOC105034789	LOC103706079
Ma05_g29860		
Dr04053	AT5G39680	Bradi2g15830
LOC_Os05g49740	LOC105041747	LOC103718824
Ma02_g09270		
Dr12603	AT3G50420	Bradi5g25817
LOC_Os04g57670	LOC105059205	LOC103724310
Ma10_g08090		
Dr20128	AT5G59600	Bradi1g66810
LOC_Os03g16450	LOC105044362	LOC103715699
Ma08_g21550		
Dr03262	AT5G44230	Bradi1g33680
LOC_Os06g47920	LOC105039665	LOC103718042
Ma05_g25110		
Dr18428	AT1G80270	Bradi2g10850
LOC_Os01g17320, LOC_Os12g07260	LOC105034235	LOC103697964
Ma08_g22450		
Dr02234	AT3G14580	Bradi5g21920
LOC_Os04g52725	LOC105045345, LOC105059831	LOC103720547
Ma03_g25570		
Dr02920	AT3G22150	Bradi5g20810
LOC_Os04g51350	LOC105060832	LOC103718748
Ma08_g15890		

Dr19832	AT3G05340	Bradi2g36120
LOC_Os05g05490	LOC105057552	LOC103707249
Ma09_g15770		
Dr00937	AT4G14850	Bradi4g16267
LOC_Os11g37330	LOC105042300	LOC103712570
Ma04_g03010		
Dr08713	AT5G62370	Bradi3g30440
LOC_Os10g02650	LOC105054344	LOC103705656
Ma04_g20060		
Dr19290	AT1G10270	Bradi2g54360, Bradi1g31760
LOC_Os05g30240	LOC105034475, LOC105049814	LOC103696374
Ma02_g02700		
Dr10092	AT4G34830	Bradi3g22720
LOC_Os10g10170	LOC105044162	LOC103719903, LOC103715437
Ma08_g15110		
Dr12109	AT5G52630	Bradi2g45250
LOC_Os01g46230	LOC105033112	LOC103715934
Ma04_g03660		
Dr16211	AT5G46100	Bradi2g46220
LOC_Os01g48140	LOC105043498	LOC103722929
Ma09_g06440		
Dr20562	AT3G54980	Bradi2g43300
LOC_Os01g42620	LOC105043700	LOC103695471
Ma05_g04970		
Dr14056	AT3G53360	Bradi5g15860
LOC_Os06g30940	LOC105046848	LOC103719157
Ma01_g03730		
Dr03754, Dr16939	AT1G55860, AT1G70320	Bradi4g07997
LOC_Os09g07900	LOC105032451, LOC105051145, LOC105050650, LOC105060538	LOC103710828, LOC103721479, LOC103702442

Ma10_g17060, Ma04_g06280		
Dr15649	AT5G48910	Bradi1g03670
LOC_Os03g60690	LOC105045672, LOC105044998	LOC103701138, LOC103716523
Ma08_g16990, Ma03_g05840		
Dr15098	AT2G15820	Bradi3g52580
LOC_Os02g47360	LOC105053548	LOC103716647
Ma08_g15840		
Dr19828	AT3G06920	Bradi1g53347
LOC_Os07g14530	LOC105052765	LOC103715087
Ma08_g06500		
Dr11812	AT2G02150	Bradi1g37736
LOC_Os06g36910	LOC105054980	LOC103712243
Ma01_g01820		
Dr15409	AT1G43980	Bradi5g01406
LOC_Os03g32090	LOC105037516	LOC103722233
Ma03_g22960		

Table 7.10: List of PPRs and heat shock proteins conserved between all species apart from *D. rotundata*

A. thaliana	*B. distachyon*	*O. sativa*
E. guineensis	*P. dactylifera*	*M. acuminata*
AT5G16420	Bradi1g49460	LOC_Os06g03530
LOC105054625	LOC103719072	Ma09_g08510
AT5G42450	Bradi2g54530	LOC_Os01g62220
LOC105045731	LOC103705936	Ma02_g17930
AT2G13600	Bradi5g10620	LOC_Os04g35610
LOC105051765	LOC103720527	Ma01_g06700
AT2G35130	Bradi3g05233	LOC_Os02g07360
LOC105059866	LOC103713193	Ma07_g03620
AT4G13650	Bradi2g22840	LOC_Os12g36620
LOC105032299	LOC103718361	Ma04_g28000
AT1G56690	Bradi1g64080	LOC_Os03g20190

LOC105044957	LOC103700938	Ma10_g02760
AT4G37480	Bradi1g07830	LOC_Os03g55360
LOC105058515	LOC103700810	Ma06_g32470
AT4G38010	Bradi5g14000	LOC_Os04g41120
LOC105051423	LOC103717527	Ma03_g02560
AT3G17830	Bradi1g69564	LOC_Os03g12236
LOC105041099	LOC103695837	Ma09_g29620
AT5G61800	Bradi5g08500	LOC_Os04g32870
LOC105060369	LOC103697172	Ma10_g29770
AT1G56570	Bradi2g11440	LOC_Os01g19380
LOC105044976	LOC103717605	Ma10_g00820
AT3G14730	Bradi2g53377	LOC_Os01g60250
LOC105052328	LOC103699614	Ma07_g15610
AT4G02820	Bradi4g06130	LOC_Os12g34340
LOC105045382	LOC103708019	Ma01_g10410
AT2G22070	Bradi1g07090	LOC_Os03g56400
LOC105033012	LOC103715225	Ma05_g19500
AT3G62190	Bradi1g20140	LOC_Os07g44310
LOC105061501	LOC103707686	Ma05_g23410
AT4G11690	Bradi1g27730	LOC_Os05g50690
LOC105052240	LOC103697884	Ma06_g36080
AT5G66631	Bradi2g04161	LOC_Os01g07340
LOC105055769	LOC103711858	Ma10_g03680
AT2G01390	Bradi1g14760	LOC_Os10g21470
LOC105057973	LOC103718955	Ma09_g19490
AT1G11290	Bradi1g51700	LOC_Os06g02200
LOC105059257	LOC103723863	Ma02_g06650
AT1G61770	Bradi2g13367	LOC_Os12g15590
LOC105033692	LOC103708009	Ma11_g18460
AT2G35720	Bradi3g30390	LOC_Os10g36370
LOC105043633	LOC103705124	Ma08_g10900
AT3G49142	Bradi1g68478	LOC_Os03g13830
LOC105053762	LOC103714630	Ma11_g00910
AT3G47530	Bradi3g37120	LOC_Os08g33700
LOC105056600	LOC103705268	Ma10_g09150

AT4G32430	Bradi1g18310	LOC_Os07g47370
LOC105040166	LOC103708823	Ma05_g05780
AT3G58590	Bradi2g50160	LOC_Os01g55070
LOC105043708	LOC103702891	Ma01_g21490
AT1G10330	Bradi4g21120	LOC_Os05g18950
LOC105048543	LOC103701858	Ma06_g18510
AT5G08305	Bradi3g27630	LOC_Os10g30590
LOC105042766	LOC103717077	Ma11_g01680
AT2G32630	Bradi1g18220	LOC_Os07g47470
LOC105045278	LOC103715273	Ma11_g10560
AT1G71460	Bradi4g40440	LOC_Os12g10184
LOC105051659	LOC103704868	Ma03_g18230
AT1G34160	Bradi1g45290	LOC_Os12g17080
LOC105056732	LOC103705344	Ma08_g18310
AT5G11310	Bradi3g33450	LOC_Os01g32170
LOC105039693	LOC103722717	Ma04_g15790
AT1G26900	Bradi1g20570	LOC_Os08g38610
LOC105041718	LOC103716799	Ma08_g16920
AT4G19890	Bradi4g14010	LOC_Os11g39360
LOC105048626	LOC103701885	Ma01_g05800
AT5G13230	Bradi2g13297	LOC_Os03g32620
LOC105051070	LOC103708067	Ma10_g18390
AT3G48810	Bradi2g38481	LOC_Os05g04160
LOC105057315	LOC103709672	Ma08_g10570
AT5G52850	Bradi1g00491	LOC_Os03g64370
LOC105050906	LOC103715537	Ma07_g15650
AT1G03540	Bradi1g36120	LOC_Os06g40860
LOC105045407	LOC103695712	Ma10_g04670
AT5G54660	Bradi2g20767	LOC_Os05g42120
LOC105050143	LOC103709974	Ma08_g03540
AT4G35130	Bradi1g75980	LOC_Os03g04390
LOC105061197	LOC103703825	Ma11_g01100
AT2G41000	Bradi2g10686	LOC_Os01g17040
LOC105034385	LOC103722284	Ma10_g01780
AT4G30825	Bradi4g30570	LOC_Os09g25550

LOC105032320	LOC103707435	Ma05_g04370
AT2G03880	Bradi3g40131	LOC_Os06g27790
LOC105049394	LOC103711594	Ma07_g02060
AT2G33680	Bradi2g13137	LOC_Os01g27650
LOC105052895	LOC103718801	Ma05_g24630
AT4G25270	Bradi1g01363	LOC_Os03g63560
LOC105054268	LOC103702899	Ma05_g14410
AT5G02860	Bradi1g22890	LOC_Os07g40120
LOC105044503	LOC103707391	Ma06_g18300
AT3G59040	Bradi3g14950	LOC_Os08g09270
LOC105033126	LOC103705591	Ma08_g07030
AT2G27610	Bradi3g55520	LOC_Os05g38190
LOC105035804	LOC103720572	Ma11_g24500
AT1G03560	Bradi5g23200	LOC_Os07g30930
LOC105045441	LOC103699260	Ma05_g00790
AT5G52640	Bradi5g02037	LOC_Os04g01740
LOC105059479	LOC103715800	Ma08_g15090
AT3G02010	Bradi2g03730	LOC_Os01g01115
LOC105056858	LOC103716120	Ma09_g07810
AT2G17210	Bradi3g60830	LOC_Os02g58620
LOC105055337	LOC103701176	Ma00_g03640
AT1G62260	Bradi1g45170	LOC_Os06g12360
LOC105048611	LOC103701884	Ma10_g29690
AT2G42920	Bradi1g14160	LOC_Os03g43470
LOC105055732	LOC103696603	Ma07_g01600
AT5G37570	Bradi5g09540	LOC_Os04g33840
LOC105052138	LOC103714263	Ma09_g02540
AT3G24000	Bradi4g33270	LOC_Os10g30760
LOC105049925	LOC103695988	Ma06_g38050
AT1G22960	Bradi4g42330	LOC_Os12g05640
LOC105038239	LOC103703086	Ma06_g37370
AT5G39350	Bradi3g50980	LOC_Os02g44480
LOC105034597	LOC103705577	Ma04_g21350

Table 7.11: Non-redundant gene ontology terms for 2,795 genes significantly (after FDR correction) enriched in *D. rotundata* with orthologous genes identified in *A. thaliana*, *B. distachyon*, *O. sativa*, *E. guineensis*, *P. dactylifera* and *M. acuminata*. **This table has been reproduced with persmission from Tamiru, M., Natsume, S., Takagi, H., White, B., Yaegashi, H., Shimizu, M., *et al*, 2017.**

Term_ID	Description	log10 p-value
GO:0001071	nucleic acid binding transcription factor activity	-6.6144
GO:0003674	molecular_function	-6.6144
GO:0003700	sequence-specific DNA binding transcription factor activity	-6.6144
GO:0003824	catalytic activity	-6.6144
GO:0005488	binding	-6.6144
GO:0016788	hydrolase activity, acting on ester bonds	-6.6144
GO:0019825	oxygen binding	-6.6144
GO:0016787	hydrolase activity	-6.6144
GO:0097159	organic cyclic compound binding	-6.6144
GO:0003676	nucleic acid binding	-6.6144
GO:1901363	heterocyclic compound binding	-6.6144
GO:0003723	RNA binding	-6.6144
GO:0004518	nuclease activity	-6.6144
GO:0005575	cellular_component	-6.6144
GO:0009536	plastid	-6.6144
GO:0043226	organelle	-6.6144
GO:0044464	cell part	-6.6144
GO:0009579	thylakoid	-6.6144
GO:0044424	intracellular part	-6.6144
GO:0005622	intracellular	-6.6144
GO:0044444	cytoplasmic part	-6.6144
GO:0005777	peroxisome	-6.6144
GO:0042579	microbody	-6.6144

GO:0043227	membrane-bounded organelle	-6.6144
GO:0000003	reproduction	-6.6144
GO:0008150	biological_process	-6.6144
GO:0009987	cellular process	-6.6144
GO:0015979	photosynthesis	-6.6144
GO:0016043	cellular component organization	-6.6144
GO:0032501	multicellular organismal process	-6.6144
GO:0044699	single-organism process	-6.6144
GO:0044767	single-organism developmental process	-6.6144
GO:0071840	cellular component organization or biogenesis	-6.6144
GO:0044238	primary metabolic process	-6.6144
GO:0009056	catabolic process	-6.6144
GO:0006807	nitrogen compound metabolic process	-6.6144
GO:0071704	organic substance metabolic process	-6.6144
GO:1901360	organic cyclic compound metabolic process	-6.6144
GO:0007049	cell cycle	-6.6144
GO:0044707	single-multicellular organism process	-6.6144
GO:0044260	cellular macromolecule metabolic process	-6.6144
GO:0044237	cellular metabolic process	-6.6144
GO:0006629	lipid metabolic process	-6.6144
GO:0043170	macromolecule metabolic process	-6.6144
GO:0046483	heterocycle metabolic process	-6.6144
GO:0006725	cellular aromatic compound metabolic process	-6.6144
GO:0090304	nucleic acid metabolic process	-6.6144

GO:0019538	protein metabolic process	-6.6144
GO:0006259	DNA metabolic process	-6.6144
GO:0008152	metabolic process	-6.6073
GO:0032502	developmental process	-6.426
GO:0009058	biosynthetic process	-6.007
GO:0006810	transport	-5.6345
GO:0051179	localization	-5.6345
GO:0005829	cytosol	-4.2076
GO:0043603	cellular amide metabolic process	-4.1273
GO:0006412	translation	-4.1273
GO:0006091	generation of precursor metabolites and energy	-4.1273
GO:0006518	peptide metabolic process	-4.1273
GO:1901564	organonitrogen compound metabolic process	-4.1273
GO:0044249	cellular biosynthetic process	-4.1273
GO:1901566	organonitrogen compound biosynthetic process	-4.1273
GO:0016740	transferase activity	-3.9173
GO:0036094	small molecule binding	-3.844
GO:0000166	nucleotide binding	-3.844
GO:1901265	nucleoside phosphate binding	-3.844
GO:0044267	cellular protein metabolic process	-3.633
GO:0003774	motor activity	-2.8082
GO:0016817	hydrolase activity, acting on acid anhydrides	-2.8082
GO:0019748	secondary metabolic process	-2.6951
GO:0005975	carbohydrate metabolic process	-2.6511
GO:0044710	single-organism metabolic process	-2.5812
GO:0005215	transporter activity	-2.4349

GO:0016020	membrane	-1.9399
GO:0008135	translation factor activity, nucleic acid binding	-1.413
GO:0040029	regulation of gene expression, epigenetic	-1.2464
GO:0008219	cell death	-0.7823
GO:0007154	cell communication	-0.7244
GO:0009719	response to endogenous stimulus	-0.7089
GO:0005737	cytoplasm	-0.6889
GO:0005623	cell	-0.6391
GO:0030312	external encapsulating structure	-0.589
GO:0005618	cell wall	-0.5201
GO:0043412	macromolecule modification	-0.4177
GO:0031975	envelope	-0.3499
GO:0005635	nuclear envelope	-0.3499
GO:0003682	chromatin binding	-0.0462

Table 7.13: The top 10 genes with greatest increase or decrease in LFC, in tuber tissues, compared to the other tissues. Genes lacking functionally annotated appear in the table as 'NULL'.

Gene	Chr	log2FoldChange	pvalue	padj	Interproscan IPR Description
Dr16361	Chr1	4.926	4.49E-121	7.87E-117	Mycolic acid cyclopropane synthase
Dr09762	Chr5	3.802	5.38E-45	3.14E-41	Alpha-D-phosphohexomutase
Dr06855	Chr7	-8.403	5.81E-43	2.04E-39	Glutamine synthetase
Dr08751	Chr11	1.453	3.40E-36	8.50E-33	NULL
Dr15103	Chr12	3.562	4.89E-35	1.07E-31	Protein of unknown function DUF4079
Dr04510	Chr5	7.543	7.36E-33	1.43E-29	NULL
Dr01378	Chr17	3.273	2.62E-31	4.60E-28	Ribosomal protein S24e
Dr14214	Chr10	1.354	7.44E-30	1.19E-26	Dcp1-like decapping
Dr07513	Chr5	-9.880	8.43E-28	1.14E-24	Chlorophyll A-B binding protein
Dr11708	Chr1	2.101	4.13E-27	4.82E-24	Enolase

Table 7.12: Non-redundant gene ontology terms for 11,348 genes significantly (after FDR correction) enriched in *D. rotundata* with no orthologous genes identfied in *A. thaliana*, *B. distachyon*, *O. sativa*, *E. guineensis*, *P. dactylifera* and *M.acuminata*. **This table has been reproduced with persmission from Tamiru, M., Natsume, S., Takagi, H., White, B., Yaegashi, H., Shimizu, M., *et al*, 2017.**

Term ID	Description	log10 p-value
GO:0005575	cellular_component	-3.9145
GO:0044424	intracellular part	-3.9145
GO:0044444	cytoplasmic part	-3.9145
GO:0044464	cell part	-3.9145
GO:0003674	molecular_function	-3.9145
GO:0005488	binding	-3.9145
GO:0043226	organelle	-3.2246
GO:0043229	intracellular organelle	-3.2246
GO:0008150	biological_process	-2.675
GO:0005515	protein binding	-2.4422
GO:0003824	catalytic activity	-1.6261
GO:0043167	ion binding	-0.6668
GO:0016740	transferase activity	-0.1605
GO:0097159	organic cyclic compound binding	-0.1395
GO:1901363	heterocyclic compound binding	-0.1395

Table 7.14: The top 10 genes with greatest increase or decrease in LFC, in flower and related tissues, compared to the other tissues.

Gene	Chr	log2FoldChange	pvalue	padj	Interproscan IPR Description
Dr08673	Chr14	12.445	4.50E-48	3.87E-44	Transcription factor
Dr04574	Chr17	12.857	8.93E-44	5.12E-40	PEBP
Dr03833	Chr2	-2.545	3.24E-34	1.39E-30	Ribosomal protein L13e
Dr16716	Chr2	-6.028	1.29E-30	3.16E-27	Leucine-rich repeat
Dr01303	Chr17	9.620	1.22E-27	2.62E-24	Probable transposase
Dr09858	Chr5	9.621	8.39E-25	1.44E-21	Transcription factor
Dr11587	Chr18	-6.412	1.40E-23	2.13E-20	Transcription factor
Dr05697	Chr20	7.100	1.86E-23	2.46E-20	NULL
Dr02504	Chr10	-10.720	8.53E-23	1.05E-19	Armadillo-like helical
Dr11543	Chr10	-10.574	7.54E-22	8.66E-19	Isocitrate and isopropylmalate dehydrogenases family

Table 7.15: Top 50 highest expressed genes observed to be enriched in tuber (padj < 0.05).

Gene name	log2FoldChange	pvalue	padj	Interpro Description
Dr17256	5.104876489	2.83E-05	0.000281719	Cytochrome P450|Cytochrome P450, E-class, group I|Cytochrome P450, conserved site
Dr24140	5.467783158	8.60E-07	1.39E-05	Transposase, MuDR, plant|Zinc finger, SWIM-type|MULE transposase domain|Zinc finger, PMZ-type
Dr15196	6.751437946	2.65E-16	3.74E-14	Probable transposase, Ptta/En/Spm, plant
Dr13213	5.297880809	6.55E-08	1.46E-06	DnaJ domain|DnaJ domain, conserved site
Dr11460	5.479438854	8.14E-09	2.34E-07	Oxoglutarate/ iron-dependent dioxygenase |Non-haem dioxygenase N-terminal domain|Isopenicillin N synthase-like

Dr19009	5.066575556	6.39E-05	0.000549598	Myc-type, basic helix-loop-helix (bHLH) domain|Transcription factor MYC/MYB N-terminal
Dr04692	5.127486982	0.001195591	0.00647645	Glycoside hydrolase, family 18, catalytic domain|Glycoside hydrolase, superfamily|Glycoside hydrolase, catalytic domain|Glycoside hydrolase, chitinase active site
Dr24198	5.51029429	0.000158958	0.001193717	Glycoside hydrolase, family 18, catalytic domain|Glycoside hydrolase, superfamily|Glycoside hydrolase, catalytic domain|Glycoside hydrolase, chitinase active site

Dr01881	5.723567042	4.15E-07	7.33E-06	SANT/Myb domain|AWPM-19-like|Homeodomain-like|Myb domain
Dr10787	5.051371422	0.000357752	0.002363409	Major intrinsic protein|Aquaporin-like|Major intrinsic protein, conserved site
Dr08902	6.364370905	1.48E-14	1.40E-12	
Dr15474	5.423516695	2.34E-07	4.40E-06	Terpene synthase, N-terminal domain|Terpene synthase, metal-binding domain|Terpenoid cyclases/protein prenyltransferase alpha-alpha toroid|Terpenoid synthase
Dr01919	7.487599803	2.29E-44	1.03E-40	Domain of unknown function DUF640
Dr02868	5.134873175	3.59E-06	4.70E-05	UDP-glucuronosyl/UDP-glucosyltransferase
Dr14344	5.745390255	1.97E-15	2.33E-13	Starch synthase, catalytic domain

Dr19425	6.470708618	3.96E-14	3.48E-12	Alpha crystallin/Hsp20 domain|HSP20-like chaperone
Dr04510	7.313079258	8.31E-35	1.65E-31	
Dr02722	5.413860555	4.72E-06	5.98E-05	Leucine-rich repeat|Leucine-rich repeat-containing N-terminal, type 2|Leucine-rich repeat, typical subtype
Dr05218	4.937285295	2.02E-05	0.000210524	
Dr11281	5.603830633	8.80E-05	0.000722665	Haem peroxidase, plant/ fungal/ bacterial| Plant peroxidase|Haem peroxidase|Peroxidases heam-ligand binding site|Peroxidase, active site
Dr01855	7.288238604	2.57E-07	4.78E-06	Carbonic anhydrase, alpha-class|Alpha carbonic anhydrase

Dr00831	6.602757503	5.68E-08	1.29E-06	Cytochrome P450|Cytochrome P450, E-class, group I|Cytochrome P450, conserved site
Dr08729	6.303965966	2.07E-11	1.05E-09	
Dr05434	4.984628559	5.50E-07	9.43E-06	
Dr15590	5.202020083	6.22E-05	0.000536881	
Dr14834	5.041289482	0.001509401	0.007878368	

Dr17018	5.622814559	1.01E-19	2.77E-17	Glycoside hydrolase, family 13|Glycosyl hydrolase, family 13, catalytic domain|Alpha-amylase, C-terminal all beta|Glycoside hydrolase, family 13, N-terminal|Glycoside hydrolase, superfamily|Immunoglobulin E-set|Immunoglobulin-like fold|Glycosyl hydrolase, family 13, all-beta|Glycoside hydrolase, catalytic domain|Glycosyl hydrolase, family 13, subfamily, catalytic domain|1,4-alpha-glucan-branching enzyme
Dr06853	6.185377296	1.24E-07	2.53E-06	
Dr16306	8.881850522	4.49E-22	2.06E-19	PEBP

Dr07061	5.293842437	1.13E-16	1.74E-14	Dienelactone hydrolase
Dr05500	5.216728548	0.000296699	0.002006725	
Dr00194	5.174136145	2.68E-12	1.65E-10	
Dr19408	9.258803292	2.69E-12	1.65E-10	Bulb-type lectin domain
Dr17669	5.302613435	9.07E-05	0.000743219	Glycoside hydrolase, family 5|Glycoside hydrolase, superfamily|Glycoside hydrolase, catalytic domain
Dr14784	6.683144364	2.87E-23	1.66E-20	
Dr02895	5.332659501	3.06E-11	1.52E-09	Biopterin transport-related protein BT1|Major facilitator superfamily domain, general substrate transporter
Dr08991	5.100227647	6.35E-05	0.000546622	
Dr15578	5.353001783	4.97E-08	1.15E-06	UDP-glucuronosyl/ UDP-glucosyltransferase
Dr01858	11.00473989	4.45E-17	7.31E-15	Carbonic anhydrase, alpha-class|Alpha carbonic anhydrase

Dr01856	9.171749592	3.17E-14	2.86E-12	Carbonic anhydrase, alpha-class|Alpha carbonic anhydrase
Dr01640	6.004178235	4.65E-13	3.45E-11	Ferredoxin--NADP reductase| Oxidoreductase FAD/NAD(P)-binding|Riboflavin synthase-like beta-barrel|Ferredoxin reductase-type FAD-binding domain
Dr01642	11.05917496	1.43E-14	1.37E-12	Ferredoxin--NADP reductase| Oxidoreductase FAD/NAD(P)-binding|Riboflavin synthase-like beta-barrel|Ferredoxin reductase-type FAD-binding domain
Dr10113	5.923243533	0.000635421	0.003803391	Bulb-type lectin domain
Dr15345	5.894619543	7.62E-05	0.000638671	

Dr04333	5.315041875	3.87E-06	5.02E-05	Haem peroxidase, plant/ fungal/ bacterial| Plant peroxidase|Haem peroxidase|Peroxidases heam-ligand binding site
Dr22133	5.290544037	5.93E-05	0.000514771	Zinc finger, double-stranded RNA binding|Zinc finger C2H2-type/integrase DNA-binding domain|Zinc finger, C2H2-like|Zinc finger, C2H2
Dr15854	5.988544415	1.74E-08	4.56E-07	
Dr24177	5.328245463	1.92E-17	3.41E-15	
Dr20590	5.46486886	2.63E-05	0.000264019	
Dr13471	7.091600436	2.49E-12	1.55E-10	

Table 7.25: Subset of *D. rotundata* enriched genes found in orthogroups conserved between all species except for *D. tokoro*

Gene	Description
Dr04974	abctransporter
Dr06081	ribosomebiogenesisgtp-bindingprotein
Dr21682	smallrnadegradingnuclease5-like
Dr21995	protein
Dr07569	gprotein-coupledreceptor
Dr07838	imidazoleglycerol-phosphatedehydratase

Dr00289	armrepeatprotein
Dr00754	phosphotyrosylphosphataseactivatorprotein
Dr00841	adeninenucleotidealphahydrolases-likeprotein
Dr10692	chitin-induciblegibberellin-responsiveprotein
Dr11684	protein
Dr12793	nitrilasehomolog1-like
Dr14410	protein
Dr14802	chloroplastphotosystemii10kdaprotein
Dr03453	uncharacterizedprotein
Dr03519	alphabetafoldfamilyprotein
Dr03468	tousled-likekinase
Dr15168	pap-specificphosphatasehal2-like
Dr15491	histidineproteinmethyltransferase1homolog
Dr16764	alpha-mannosidase-likeprotein
Dr04141	proline-richfamilyprotein
Dr17490	glycosyltransferasefamily1protein
Dr17663	nucleoporinautopeptidase
Dr04425	proteinhighchlorophyllfluorescent107
Dr17730	dna-directedrnapolymerasesandiiikdapolypeptide
Dr19093	u2snrnpauxiliarysmall
Dr05404	phosphoglyceratemutase-likeprotein
Dr05363	multiplechloroplastdivisionsite1
Dr06624	map4kinase
Dr07368	protein
Dr22070	transferringglycosyl
Dr22140	trna-specificadenosinedeaminase1-like
Dr00572	splicingfactor3bsubunit2
Dr00694	abctransporterfamilyprotein
Dr00890	replicationfactorcsubunit1-like
Dr23385	rnabindingprotein
Dr10459	uncharacterizedprotein
Dr10460	uncharacterizedprotein
Dr10454	peptidyl-trnahydrolase

Dr10593	ribosomalrnalargesubunitmethyltransferasen
Dr11440	ubiquitincarboxyl-terminalhydrolase22-like
Dr11518	2-dehydro-3-deoxyphosphooctonatealdolase
Dr01745	stressresponsivealpha-betabarreldomainprotein
Dr02153	protein
Dr02418	importinalpha-2
Dr02451	mateeffluxfamilyprotein
Dr02449	calcineurin-likemetallo-phosphoesterase-likeprotein
Dr12791	nitrilasehomolog1-like
Dr12796	nitrilasehomolog1-like
Dr12799	nitrilasehomolog1-like
Dr13318	uncharacterizedprotein
Dr13372	tubulinfoldingcofactorb

Table 7.31: Subset of oprhan genes in comparion of orthogroups between *D. rotundata* and 25 other angiosperm species. List contains subtilisin-like protease, zinc finger protein zat9-like and mechanosensitive ion channel protein 6-like coding gene models.

Gene	Description
Dt01547	subtilisin-like protease
Dt11082	zinc finger protein zat9-like
Dt11233	subtilisin-like protease
Dt11773	subtilisin-like protease
Dt16476	subtilisin-like protease
Dt16846	zinc finger protein zat9-like
Dt18170	zinc finger protein zat9-like
Dt18183	subtilisin-like protease
Dt19792	mechanosensitive ion channel protein 6-like
Dt21589	subtilisin-like protease
Dt23914	mechanosensitive ion channel protein 6-like
Dt24196	subtilisin-like protease

Dt24425	subtilisin-like protease
Dt25839	subtilisin-like protease
Dt29208	mechanosensitive ion channel protein 6-like

Table 7.32: List of 84 orthgroups and genes only observed in *D. rotundata* when compared with *D. tokoro* and with 25 other angiosperm species.

Orthogroup	Gene
OG0005291	Dr00524, Dr01282, Dr04544, Dr04643, Dr08154, Dr08189, Dr10434, Dr13474, Dr14462, Dr18799, Dr19710, Dr19884, Dr20678, Dr21210, Dr21565, Dr21576, Dr21793, Dr21866, Dr22231, Dr22370, Dr22637, Dr22638, Dr23042, Dr23488, Dr23489, Dr23565, Dr23714, Dr23809, Dr23848, Dr24148, Dr24165, Dr24189, Dr24513, Dr24545, Dr24733, Dr24861, Dr25117, Dr25891, Dr25900, Dr25910, Dr25951, Dr26018, Dr26062, Dr26100
OG0010421	Dr00154, Dr03955, Dr04005, Dr04111, Dr04563, Dr05552, Dr06452, Dr06831, Dr08078, Dr08505, Dr08522, Dr11662, Dr12159, Dr13150, Dr13610, Dr14290, Dr15196, Dr17710, Dr18089, Dr18425, Dr19289, Dr21846, Dr22115, Dr22338, Dr22718, Dr23207, Dr23323, Dr24888
OG0010654	Dr00570, Dr00687, Dr00836, Dr00928, Dr02557, Dr04361, Dr06247, Dr07168, Dr08390, Dr08405, Dr08742, Dr15083, Dr15142, Dr15232, Dr16070, Dr18701, Dr19132, Dr21244, Dr23204, Dr24082, Dr24609, Dr25279, Dr25418, Dr25528, Dr25655, Dr25737, Dr26032

OG0011091	Dr00610, Dr01325, Dr03738, Dr04644, Dr05188, Dr05853, Dr05928, Dr08515, Dr12131, Dr17400, Dr19673, Dr19674, Dr20557, Dr21350, Dr22160, Dr22905, Dr23357, Dr23872, Dr23950, Dr24342, Dr24780, Dr25085, Dr25914, Dr26015
OG0011092	Dr03912, Dr10304, Dr17378, Dr17846, Dr18830, Dr19366, Dr19573, Dr20218, Dr20352, Dr21197, Dr21211, Dr21932, Dr22163, Dr22242, Dr22460, Dr22929, Dr23036, Dr23149, Dr24511, Dr24643, Dr25568, Dr25658, Dr25681, Dr25711
OG0011440	Dr02850, Dr03774, Dr04556, Dr04751, Dr06498, Dr06528, Dr06972, Dr08028, Dr08180, Dr09635, Dr11859, Dr13154, Dr15351, Dr15385, Dr20980, Dr22123, Dr25041, Dr25063, Dr25315, Dr25839, Dr26044
OG0011663	Dr00853, Dr02529, Dr02717, Dr05580, Dr06114, Dr10443, Dr17376, Dr17751, Dr18399, Dr19725, Dr20558, Dr21129, Dr21452, Dr23218, Dr23604, Dr24278, Dr24279, Dr24752, Dr25184
OG0011664	Dr03887, Dr12648, Dr18814, Dr18845, Dr19276, Dr20293, Dr20460, Dr21403, Dr21422, Dr21770, Dr21774, Dr21965, Dr22433, Dr22456, Dr22587, Dr23979, Dr24076, Dr24740, Dr25148
OG0011812	Dr01435, Dr02090, Dr02098, Dr02102, Dr03028, Dr04239, Dr07070, Dr08326, Dr10979, Dr11226, Dr11795, Dr12549, Dr19767, Dr24043, Dr24800, Dr24927, Dr25159, Dr25465

OG0011947	Dr00161, Dr18010, Dr18555, Dr19247, Dr20534, Dr21376, Dr22619, Dr22640, Dr22801, Dr22893, Dr23309, Dr23331, Dr23525, Dr23801, Dr23893, Dr24467, Dr24817
OG0012098	Dr00014, Dr00420, Dr01020, Dr03990, Dr07847, Dr08210, Dr13757, Dr13992, Dr14845, Dr18387, Dr19037, Dr19318, Dr20033, Dr24963, Dr25310, Dr25416
OG0012442	Dr02325, Dr05919, Dr08923, Dr10185, Dr10353, Dr11092, Dr12005, Dr12213, Dr15724, Dr17787, Dr18790, Dr19762, Dr23006, Dr25878
OG0012626	Dr01104, Dr03745, Dr05119, Dr12117, Dr12241, Dr12852, Dr14990, Dr15166, Dr15864, Dr21332, Dr22648, Dr25982, Dr26160
OG0012627	Dr02517, Dr07683, Dr09117, Dr11152, Dr11478, Dr12506, Dr15189, Dr22085, Dr22158, Dr23990, Dr24491, Dr24650, Dr25325
OG0012840	Dr02523, Dr02657, Dr06518, Dr07105, Dr08191, Dr10266, Dr15870, Dr16545, Dr18800, Dr22262, Dr22877, Dr23496
OG0013111	Dr06520, Dr17836, Dr20287, Dr20509, Dr22986, Dr23308, Dr23566, Dr23657, Dr24013, Dr24233, Dr24525
OG0013113	Dr17884, Dr20267, Dr21484, Dr21586, Dr22182, Dr22276, Dr23035, Dr23326, Dr23558, Dr23804, Dr25102
OG0013415	Dr01767, Dr06403, Dr06726, Dr07029, Dr07259, Dr07677, Dr23610, Dr23912, Dr25202, Dr25676
OG0014280	Dr01186, Dr01649, Dr01654, Dr08541, Dr20764, Dr21213, Dr22960, Dr23506
OG0014960	Dr01241, Dr01682, Dr05411, Dr06767, Dr13989, Dr14161, Dr15396

OG0014962	Dr02676, Dr04870, Dr06511, Dr15862, Dr18369, Dr23792, Dr23905
OG0014976	Dr18567, Dr21541, Dr21922, Dr22313, Dr23694, Dr24391, Dr24750
OG0015734	Dr00516, Dr00528, Dr14768, Dr18358, Dr21330, Dr24414
OG0015736	Dr01049, Dr05775, Dr10286, Dr12243, Dr23590, Dr25861
OG0015739	Dr04803, Dr05532, Dr07949, Dr15373, Dr23449, Dr25219
OG0015741	Dr06609, Dr20531, Dr22127, Dr23026, Dr23693, Dr25588
OG0015749	Dr16777, Dr23064, Dr23578, Dr23641, Dr23922, Dr25822
OG0015750	Dr17360, Dr21386, Dr21470, Dr21626, Dr21820, Dr25503
OG0016565	Dr02243, Dr06447, Dr23688, Dr23724, Dr24916
OG0016566	Dr02513, Dr14840, Dr17049, Dr22592, Dr23822
OG0016567	Dr02744, Dr03172, Dr21238, Dr24449, Dr25278
OG0016573	Dr05029, Dr05030, Dr05031, Dr05033, Dr05167
OG0016575	Dr06516, Dr11007, Dr12148, Dr19969, Dr22369
OG0016592	Dr20477, Dr21924, Dr22816, Dr23079, Dr24452
OG0016593	Dr22921, Dr22997, Dr23434, Dr24370, Dr24940
OG0017716	Dr02624, Dr18459, Dr20223, Dr24640
OG0017724	Dr05361, Dr05541, Dr16594, Dr18708
OG0017725	Dr05430, Dr11219, Dr12283, Dr18445
OG0017727	Dr06391, Dr11024, Dr19861, Dr25609
OG0017739	Dr10468, Dr23450, Dr25472, Dr25549
OG0017744	Dr13520, Dr16522, Dr16528, Dr24396
OG0017749	Dr15692, Dr18436, Dr19389, Dr23253
OG0017753	Dr17883, Dr19574, Dr20343, Dr22404

OG0019409	Dr00137, Dr15844, Dr16734
OG0019410	Dr00185, Dr12236, Dr20508
OG0019414	Dr00593, Dr13007, Dr23918
OG0019416	Dr01040, Dr04636, Dr20447
OG0019436	Dr04662, Dr07953, Dr15114
OG0019439	Dr05423, Dr14720, Dr26045
OG0019443	Dr06657, Dr11656, Dr17356
OG0019446	Dr07137, Dr20900, Dr20901
OG0019451	Dr07898, Dr25533, Dr25923
OG0019453	Dr08093, Dr14819, Dr23221
OG0019455	Dr10469, Dr13009, Dr21355
OG0019470	Dr14029, Dr14034, Dr14044
OG0019473	Dr14501, Dr21937, Dr25113
OG0019477	Dr15896, Dr25921, Dr25999
OG0019483	Dr19580, Dr19601, Dr23551
OG0019491	Dr22569, Dr22572, Dr25794
OG0021757	Dr00529, Dr05768
OG0021758	Dr00594, Dr23226
OG0021764	Dr01291, Dr15595
OG0021766	Dr02015, Dr21909
OG0021770	Dr02589, Dr15386
OG0021790	Dr05234, Dr05236
OG0021796	Dr06558, Dr25865
OG0021799	Dr06732, Dr09081
OG0021803	Dr07737, Dr19984
OG0021805	Dr08109, Dr09145
OG0021808	Dr08547, Dr09053
OG0021822	Dr11322, Dr11323
OG0021830	Dr12485, Dr17344
OG0021835	Dr12881, Dr17691
OG0021874	Dr16714, Dr20979
OG0021877	Dr17184, Dr17185
OG0021883	Dr17875, Dr24907
OG0021887	Dr18788, Dr18792
OG0021892	Dr19110, Dr19113

OG0021896	Dr19529, Dr24687
OG0021900	Dr19758, Dr19759
OG0021907	Dr20549, Dr25758
OG0021913	Dr23325, Dr25046
OG0021915	Dr24028, Dr25246
OG0021916	Dr24303, Dr25648

Table 7.33: List of 32 *D. rotundata* orthgroups and genes not shared with 26 other angiosperm species.

Orthogroup	Gene
OG0005750	Dt00652.1, Dt02422.1, Dt02518.1, Dt03411.1, Dt03945.1, Dt03946.1, Dt04035.1, Dt06514.1, Dt07202.1, Dt07362.1, Dt07607.1, Dt08471.1, Dt08796.1, Dt09849.1, Dt10525.1, Dt10749.1, Dt10755.1, Dt11546.1, Dt12170.1, Dt12685.1, Dt13287.1, Dt13293.1, Dt13346.1, Dt15287.1, Dt16073.1, Dt17202.1, Dt17645.1, Dt17673.1, Dt17720.1, Dt18474.1, Dt18563.1, Dt20294.1, Dt20433.1, Dt20960.1, Dt21454.1, Dt21716.1, Dt26701.1, Dt27230.1, Dt27253.1, Dt27259.1, Dt27754.1, Dt28083.1

OG0006661	Dt00330.1, Dt00353.1, Dt00500.1, Dt00501.1, Dt01885.1, Dt02318.1, Dt02569.1, Dt03417.1, Dt04422.1, Dt05156.1, Dt08031.1, Dt08941.1, Dt09216.1, Dt09676.1, Dt09885.1, Dt10007.1, Dt10153.1, Dt10880.1, Dt13663.1, Dt14051.1, Dt15709.1, Dt16046.1, Dt16197.1, Dt18085.1, Dt18126.1, Dt18509.1, Dt19225.1, Dt19813.1, Dt19884.1, Dt21763.1, Dt22090.1, Dt22720.1, Dt23150.1, Dt23321.1, Dt23640.1, Dt23759.1, Dt26342.1, Dt26873.1, Dt27107.1
OG0011313	Dt01792.1, Dt01919.1, Dt02151.1, Dt02686.1, Dt03569.1, Dt05359.1, Dt05988.1, Dt07396.1, Dt07698.1, Dt08575.1, Dt09226.1, Dt10721.1, Dt16386.1, Dt19410.1, Dt19761.1, Dt20265.1, Dt20956.1, Dt21293.1, Dt24902.1, Dt25081.1, Dt26504.1, Dt28450.1
OG0011948	Dt00514.1, Dt00662.1, Dt00853.1, Dt05421.1, Dt06065.1, Dt10393.1, Dt14873.1, Dt19091.1, Dt20388.1, Dt21496.1, Dt22568.1, Dt24590.1, Dt25551.1, Dt26585.1, Dt28129.1, Dt28493.1, Dt28941.1
OG0013794	Dt13704.1, Dt13705.1, Dt13707.1, Dt13709.1, Dt13711.1, Dt13712.1, Dt13713.1, Dt13714.1, Dt13716.1
OG0014289	Dt05271.1, Dt05550.1, Dt10616.1, Dt14173.1, Dt14246.1, Dt16776.1, Dt19912.1, Dt24997.1
OG0015758	Dt00918.1, Dt10685.1, Dt17700.1, Dt24762.1, Dt24899.1, Dt27923.1

OG0015759	Dt01818.1, Dt14391.1, Dt14945.1, Dt20479.1, Dt21035.1, Dt21969.1
OG0016597	Dt02505.1, Dt11936.1, Dt14521.1, Dt25011.1, Dt25635.1
OG0016599	Dt03631.1, Dt27526.1, Dt27783.1, Dt28091.1, Dt28988.1
OG0016600	Dt03986.1, Dt17805.1, Dt28436.1, Dt28736.1, Dt29175.1
OG0016602	Dt07228.1, Dt07231.1, Dt07232.1, Dt07233.1, Dt07234.1
OG0017765	Dt00268.1, Dt07894.1, Dt20447.1, Dt23601.1
OG0017766	Dt00723.1, Dt11776.1, Dt24159.1, Dt28751.1
OG0017770	Dt02473.1, Dt07459.1, Dt13422.1, Dt26646.1
OG0017796	Dt23108.1, Dt28111.1, Dt28382.1, Dt29126.1
OG0019494	Dt00259.1, Dt17620.1, Dt19687.1
OG0019510	Dt07889.1, Dt10578.1, Dt20636.1
OG0021927	Dt00909.1, Dt04764.1
OG0021936	Dt05420.1, Dt18724.1
OG0021939	Dt06050.1, Dt23492.1
OG0021942	Dt08158.1, Dt18046.1
OG0021945	Dt10930.1, Dt20876.1
OG0021946	Dt11097.1, Dt28753.1
OG0021947	Dt11449.1, Dt13594.1
OG0021951	Dt12693.1, Dt12694.1
OG0021954	Dt13903.1, Dt28799.1
OG0021965	Dt18960.1, Dt24549.1
OG0021974	Dt23292.1, Dt23338.1
OG0021975	Dt24628.1, Dt27024.1
OG0021989	Dt27959.1, Dt29296.1
OG0021993	Dt28565.1, Dt28686.1

Table 7.16: Assessment of the completeness of *D. tokoro* genome assembly using 1,440 benchmarking universal single-copy orthologs (BUSCO) by BUSCOv3.0.0 using the embryophyta_odb9 (30 species) dataset.

BUSCO Type	No. of BUSCOs	% of BUSCOs
Complete Single-copy	1112	77.2
Complete Duplicated	30	2.1
Fragmented	63	4.4
Missing	235	16.3

Table 7.17: Assessment of the completeness of *D. rotundata* v0.1 genome assembly using 1,440 benchmarking universal single-copy orthologs (BUSCO) by BUSCOv3.0.0 using the embryophyta_odb9 (30 species) dataset.

BUSCO Type	No. of BUSCOs	% of BUSCOs
Complete Single-copy	1261	87.6
Complete Duplicated	45	3.1
Fragmented	42	2.9
Missing	92	6.4

Table 7.34: List of 174 orthgroups and genes only observed in *D. tokoro* and *D. rotundata*, when comapred to not shared with 24 other angiosperm species.

Orthogroup	*D. rotundata*	*D. tokoro*

OG0000919	Dr03997,	Dr04004,	Dr04567,	Dt01920.1, Dt02153.1, Dt02688.1,
	Dr06679,	Dr08476,	Dr09236,	Dt16388.1, Dt16447.1, Dt19760.1,
	Dr10249,	Dr10393,	Dr10394,	Dt21296.1, Dt27887.1
	Dr12675,	Dr13106,	Dr13716,	
	Dr13720,	Dr13993,	Dr14470,	
	Dr16789,	Dr16790,	Dr16809,	
	Dr16810,	Dr17682,	Dr17696,	
	Dr17747,	Dr17831,	Dr17856,	
	Dr17859,	Dr17872,	Dr17897,	
	Dr17989,	Dr17990,	Dr18498,	
	Dr18508,	Dr18582,	Dr18585,	
	Dr18608,	Dr18711,	Dr18737,	
	Dr18977,	Dr19278,	Dr19536,	
	Dr19537,	Dr19541,	Dr20282,	
	Dr20547,	Dr20865,	Dr21130,	
	Dr21147,	Dr21186,	Dr21488,	
	Dr21497,	Dr21597,	Dr21608,	
	Dr21795,	Dr22147,	Dr22170,	
	Dr22206,	Dr22266,	Dr22291,	
	Dr22307,	Dr22412,	Dr22481,	
	Dr22527,	Dr22547,	Dr22628,	
	Dr22632,	Dr22687,	Dr22829,	
	Dr22839,	Dr22852,	Dr22862,	
	Dr22881,	Dr23185,	Dr23192,	
	Dr23245,	Dr23305,	Dr23322,	
	Dr23346,	Dr23399,	Dr23513,	
	Dr23547,	Dr23574,	Dr23624,	
	Dr23838,	Dr24060,	Dr24262,	
	Dr24266,	Dr24620,	Dr24706,	
	Dr24724,	Dr25096,	Dr25268,	
	Dr25542,	Dr25698,	Dr25716,	
	Dr25736,	Dr25768,	Dr25772,	
	Dr25817, Dr25867, Dr26115			

OG0002800	Dr01048, Dr23821, Dr26106	Dr01137, Dr24902,	Dr01390, Dr25389,	Dt00456.1, Dt00506.1, Dt01057.1, Dt01928.1, Dt02037.1, Dt02139.1, Dt02741.1, Dt03572.1, Dt03910.1, Dt05262.1, Dt05881.1, Dt06466.1, Dt06799.1, Dt06848.1, Dt07020.1, Dt08098.1, Dt08934.1, Dt09017.1, Dt09961.1, Dt10878.1, Dt11437.1, Dt11487.1, Dt11949.1, Dt12124.1, Dt12430.1, Dt12886.1, Dt13095.1, Dt13615.1, Dt15799.1, Dt16487.1, Dt16710.1, Dt16860.1, Dt17615.1, Dt17866.1, Dt18908.1, Dt19081.1, Dt19672.1, Dt19902.1, Dt20469.1, Dt21311.1, Dt21598.1, Dt22211.1, Dt22721.1, Dt24284.1, Dt25070.1, Dt25148.1, Dt27131.1, Dt27132.1, Dt27153.1, Dt27340.1, Dt27868.1, Dt28276.1, Dt28342.1, Dt28537.1, Dt28602.1, Dt28939.1
OG0003571	Dr00099, Dr02504, Dr04497, Dr06645, Dr12513, Dr15795, Dr19533, Dr21280, Dr22023, Dr23156	Dr00752, Dr03828, Dr06572, Dr10714, Dr13395, Dr18388, Dr20456,	Dr00946, Dr03911, Dr06644, Dr12006, Dr15560, Dr18409, Dr20970,	Dt00040.1, Dt01055.1, Dt01452.1, Dt01974.1, Dt02713.1, Dt03193.1, Dt03967.1, Dt05689.1, Dt05704.1, Dt06861.1, Dt07307.1, Dt08425.1, Dt08928.1, Dt09242.1, Dt09833.1, Dt12485.1, Dt16992.1, Dt17346.1, Dt17635.1, Dt18920.1, Dt19482.1, Dt20323.1, Dt20450.1, Dt21086.1, Dt21087.1, Dt22141.1, Dt22755.1, Dt23385.1, Dt23386.1, Dt25354.1, Dt27536.1

OG0005292	Dr07623, Dr22474, Dr25025	Dt00439.1, Dt00774.1, Dt01115.1, Dt01449.1, Dt01515.1, Dt01591.1, Dt01918.1, Dt02090.1, Dt02588.1, Dt05388.1, Dt06816.1, Dt06843.1, Dt06879.1, Dt07006.1, Dt08189.1, Dt09718.1, Dt10948.1, Dt11470.1, Dt11943.1, Dt11947.1, Dt12508.1, Dt12964.1, Dt14422.1, Dt14691.1, Dt16256.1, Dt16385.1, Dt17205.1, Dt18117.1, Dt18637.1, Dt20446.1, Dt20868.1, Dt20897.1, Dt22753.1, Dt23609.1, Dt25575.1, Dt25878.1, Dt26094.1, Dt27023.1, Dt27044.1, Dt28625.1, Dt28965.1
OG0008814	Dr03995, Dr08819, Dr13114, Dr13980, Dr18067, Dr18539, Dr18825, Dr19244, Dr19249, Dr20241, Dr20840, Dr21192, Dr21558, Dr21806, Dr22407, Dr22486, Dr22898, Dr23484, Dr23512, Dr23698, Dr23775, Dr23844, Dr24088, Dr24299, Dr24784, Dr25161, Dr25484, Dr25896	Dt00058.1, Dt00546.1, Dt15304.1, Dt16311.1, Dt27357.1
OG0009875	Dr00857, Dr00927, Dr01017, Dr01190, Dr01608, Dr04471, Dr05448, Dr07021, Dr07605, Dr08473, Dr09090, Dr11215, Dr17769, Dr18322, Dr19047, Dr19951, Dr19981, Dr21305, Dr22054, Dr22714, Dr24579	Dt06640.1, Dt07156.1, Dt07740.1, Dt16826.1, Dt16827.1, Dt18745.1, Dt19905.1, Dt20111.1, Dt23708.1

OG0010968	Dr00266, Dr02500, Dr04621, Dr05295, Dr05387, Dr06877, Dr14291, Dr14927, Dr15105, Dr17366, Dr17526, Dr18430, Dr18451, Dr20616, Dr21285, Dr21417, Dr21859, Dr22029, Dr22503, Dr22908, Dr23378, Dr23563, Dr24004, Dr24711	Dt05941.1
OG0011312	Dr02555, Dr07277, Dr18047, Dr18501, Dr19348, Dr19356, Dr19360, Dr20077, Dr20471, Dr20826, Dr21867, Dr22260, Dr24003, Dr24595, Dr24628, Dr25160, Dr25556, Dr25598	Dt01968.1, Dt11526.1, Dt14231.1, Dt22794.1
OG0011439	Dr00245, Dr01310, Dr04013, Dr04965, Dr05544, Dr05737, Dr08220, Dr08484, Dr15152, Dr15381, Dr19292, Dr19311, Dr25109, Dr25147, Dr25511, Dr25728, Dr25904, Dr25960, Dr26003, Dr26054	Dt24996.1
OG0011546	Dr15685	Dt00345.1, Dt00421.1, Dt01116.1, Dt01511.1, Dt04327.1, Dt05487.1, Dt07722.1, Dt09322.1, Dt14657.1, Dt16560.1, Dt17264.1, Dt18947.1, Dt24600.1, Dt25810.1, Dt26368.1, Dt26791.1, Dt26894.1, Dt27366.1, Dt28924.1
OG0012447	Dr23925, Dr24602	Dt03675.1, Dt04410.1, Dt08477.1, Dt09955.1, Dt10456.1, Dt10850.1, Dt12263.1, Dt12265.1, Dt13726.1, Dt15048.1, Dt17124.1, Dt21407.1
OG0012624	Dr00451, Dr00453, Dr02791, Dr04562, Dr04962, Dr05032, Dr05193, Dr05194, Dr18994	Dt05278.1, Dt14573.1, Dt16649.1, Dt28298.1

OG0012630	Dr15865, Dr17832, Dr18039, Dr18063, Dr18549, Dr20331, Dr20488, Dr21731, Dr22287, Dr23646, Dr23727, Dr24006	Dt07318.1
OG0012841	Dr03325, Dr04664, Dr05447, Dr07220, Dr11485, Dr24992	Dt07147.1, Dt10776.1, Dt12496.1, Dt13096.1, Dt20813.1, Dt28827.1
OG0013106	Dr01326, Dr01327, Dr06667, Dr07275, Dr11399, Dr21029, Dr23762, Dr24166, Dr26027	Dt10991.1, Dt12284.1
OG0013108	Dr01640, Dr01642, Dr11920, Dr12625, Dr13996, Dr17070, Dr17071, Dr17074, Dr17732, Dr24911	Dt26268.1
OG0013112	Dr14823	Dt06530.1, Dt11795.1, Dt20883.1, Dt22172.1, Dt23399.1, Dt23980.1, Dt24464.1, Dt25406.1, Dt28040.1, Dt29191.1
OG0013416	Dr03224, Dr07797, Dr08485, Dr19328, Dr21328, Dr23165, Dr24683, Dr25336, Dr25597	Dt24130.1
OG0013781	Dr00817, Dr00818, Dr03963, Dr03964, Dr15025, Dr25765	Dt19086.1, Dt23455.1, Dt28886.1
OG0013782	Dr00839, Dr04893, Dr08103, Dr19967	Dt02571.1, Dt16317.1, Dt19543.1, Dt27672.1, Dt27789.1
OG0013783	Dr02604, Dr04135, Dr08503, Dr24435	Dt03664.1, Dt08094.1, Dt09310.1, Dt15613.1, Dt20196.1
OG0013786	Dr10795, Dr19821, Dr19822, Dr19823, Dr19833, Dr22214, Dr22218	Dt02446.1, Dt17667.1
OG0013789	Dr24244	Dt01569.1, Dt02161.1, Dt06792.1, Dt07742.1, Dt15887.1, Dt17309.1, Dt23909.1, Dt23910.1
OG0014278	Dr00549, Dr10481, Dr25101, Dr25264, Dr25441, Dr25940	Dt18274.1, Dt26321.1
OG0014279	Dr00907, Dr06454, Dr11469, Dr15015, Dr19152, Dr19940, Dr25154	Dt15669.1

OG0014284	Dr12119	Dt01909.1, Dt02298.1, Dt07444.1, Dt14268.1, Dt24562.1, Dt24849.1, Dt28060.1
OG0014964	Dr05406, Dr07724	Dt00303.1, Dt10992.1, Dt15259.1, Dt27352.1, Dt29201.1
OG0014965	Dr06316, Dr06317	Dt10649.1, Dt10650.1, Dt10653.1, Dt10655.1, Dt10656.1
OG0014968	Dr06830, Dr16512, Dr22028, Dr24422, Dr25436	Dt04442.1, Dt17509.1
OG0014972	Dr14829, Dr21787, Dr22492, Dr22780, Dr23340, Dr24512	Dt00522.1
OG0015743	Dr08144, Dr18779, Dr24961	Dt02328.1, Dt27736.1, Dt29106.1
OG0015756	Dr26141	Dt04343.1, Dt12490.1, Dt21734.1, Dt22389.1, Dt28821.1
OG0016562	Dr00259, Dr04022, Dr11447, Dr22603	Dt21732.1
OG0016569	Dr03086, Dr20022	Dt15324.1, Dt23323.1, Dt29318.1
OG0016571	Dr04330, Dr04331, Dr04459, Dr12626	Dt07081.1
OG0016579	Dr09498	Dt06954.1, Dt26893.1, Dt28122.1, Dt28160.1
OG0016586	Dr14280	Dt19606.1, Dt19607.1, Dt26169.1, Dt26170.1
OG0017710	Dr01101, Dr19139	Dt05367.1, Dt09032.1
OG0017711	Dr01178, Dr15591	Dt18123.1, Dt29008.1
OG0017714	Dr02048	Dt13654.1, Dt13656.1, Dt13659.1
OG0017717	Dr02675, Dr06563	Dt10745.1, Dt28765.1
OG0017719	Dr03648, Dr03649	Dt18026.1, Dt18226.1
OG0017721	Dr03987, Dr18215	Dt24853.1, Dt24857.1
OG0017722	Dr04485, Dr04525	Dt10735.1, Dt10739.1
OG0017730	Dr07627, Dr25626	Dt13047.1, Dt17434.1
OG0017737	Dr09527, Dr22476	Dt05661.1, Dt05664.1
OG0017738	Dr10337	Dt01865.1, Dt01866.1, Dt01869.1
OG0017742	Dr12364, Dr20067	Dt06385.1, Dt06386.1
OG0017745	Dr14086	Dt18319.1, Dt18322.1, Dt18324.1
OG0017747	Dr15147, Dr24503	Dt23430.1, Dt26615.1
OG0017756	Dr18961	Dt02534.1, Dt09317.1, Dt28773.1

OG0017757	Dr19384, Dr25040	Dt22104.1, Dt27467.1
OG0017758	Dr19680	Dt05268.1, Dt21622.1, Dt21623.1
OG0017760	Dr23443, Dr24300	Dt00669.1, Dt07213.1
OG0019418	Dr01114, Dr02221	Dt08644.1
OG0019419	Dr01264	Dt22715.1, Dt27426.1
OG0019420	Dr01268	Dt12588.1, Dt24364.1
OG0019421	Dr02070, Dr02073	Dt23423.1
OG0019428	Dr03312	Dt25793.1, Dt27878.1
OG0019431	Dr04212, Dr04213	Dt04044.1
OG0019432	Dr04275	Dt14062.1, Dt27576.1
OG0019433	Dr04352, Dr14705	Dt07111.1
OG0019438	Dr04858	Dt00283.1, Dt00286.1
OG0019441	Dr06489	Dt11074.1, Dt15233.1
OG0019442	Dr06553, Dr06554	Dt28992.1
OG0019448	Dr07250	Dt23267.1, Dt28837.1
OG0019449	Dr07658, Dr25898	Dt28172.1
OG0019450	Dr07807, Dr07810	Dt17759.1
OG0019457	Dr10823, Dr10825	Dt13904.1
OG0019458	Dr10824, Dr11710	Dt15599.1
OG0019464	Dr12271, Dr25024	Dt01996.1
OG0019465	Dr12583	Dt21423.1, Dt22986.1
OG0019468	Dr13442	Dt14728.1, Dt23728.1
OG0019469	Dr13535	Dt05089.1, Dt05091.1
OG0019472	Dr14403, Dr14405	Dt21788.1
OG0019475	Dr15097	Dt06355.1, Dt20490.1
OG0019479	Dr17796, Dr19077	Dt13064.1
OG0019484	Dr19769	Dt24423.1, Dt24426.1
OG0021759	Dr00726	Dt25060.1
OG0021760	Dr00727	Dt25059.1
OG0021761	Dr00838	Dt02479.1
OG0021762	Dr00842	Dt27435.1
OG0021763	Dr01170	Dt18521.1
OG0021767	Dr02337	Dt05701.1
OG0021769	Dr02554	Dt28076.1
OG0021772	Dr02932	Dt16343.1

OG0021773	Dr02974	Dt09290.1
OG0021774	Dr03027	Dt08026.1
OG0021775	Dr04001	Dt11700.1
OG0021776	Dr04049	Dt28386.1
OG0021777	Dr04179	Dt04074.1
OG0021778	Dr04282	Dt14046.1
OG0021779	Dr04320	Dt17284.1
OG0021780	Dr04415	Dt29459.1
OG0021781	Dr04452	Dt08993.1
OG0021784	Dr04560	Dt27304.1
OG0021785	Dr04912	Dt12854.1
OG0021787	Dr05093	Dt06774.1
OG0021789	Dr05218	Dt04981.1
OG0021791	Dr05302	Dt04899.1
OG0021792	Dr05311	Dt13854.1
OG0021794	Dr05921	Dt17998.1
OG0021795	Dr06173	Dt17396.1
OG0021800	Dr06780	Dt05114.1
OG0021801	Dr07372	Dt07443.1
OG0021806	Dr08171	Dt23299.1
OG0021807	Dr08537	Dt28999.1
OG0021810	Dr09217	Dt24411.1
OG0021811	Dr09225	Dt01599.1
OG0021812	Dr09526	Dt18780.1
OG0021813	Dr09721	Dt25401.1
OG0021814	Dr09866	Dt01148.1
OG0021815	Dr09928	Dt25300.1
OG0021818	Dr10910	Dt02936.1
OG0021819	Dr11273	Dt16320.1
OG0021820	Dr11274	Dt09427.1
OG0021823	Dr11418	Dt17484.1
OG0021824	Dr11517	Dt25724.1
OG0021825	Dr11771	Dt15863.1
OG0021826	Dr12041	Dt07217.1
OG0021829	Dr12322	Dt26179.1

OG0021831	Dr12568	Dt21421.1
OG0021832	Dr12586	Dt22984.1
OG0021833	Dr12599	Dt17253.1
OG0021836	Dr12908	Dt15016.1
OG0021837	Dr13018	Dt02066.1
OG0021839	Dr13384	Dt22603.1
OG0021841	Dr13408	Dt26612.1
OG0021842	Dr13484	Dt19416.1
OG0021844	Dr13526	Dt11652.1
OG0021846	Dr13674	Dt14370.1
OG0021848	Dr13867	Dt02601.1
OG0021849	Dr13956	Dt14082.1
OG0021850	Dr13959	Dt14085.1
OG0021851	Dr14133	Dt04194.1
OG0021853	Dr14376	Dt12545.1
OG0021854	Dr14380	Dt12541.1
OG0021855	Dr14785	Dt22378.1
OG0021856	Dr14837	Dt08564.1
OG0021857	Dr14903	Dt11885.1
OG0021858	Dr14942	Dt07328.1
OG0021859	Dr15084	Dt06334.1
OG0021860	Dr15193	Dt08081.1
OG0021861	Dr15303	Dt05110.1
OG0021863	Dr15783	Dt19531.1
OG0021866	Dr16086	Dt08433.1
OG0021867	Dr16132	Dt27668.1
OG0021868	Dr16151	Dt29198.1
OG0021869	Dr16217	Dt26099.1
OG0021870	Dr16341	Dt19622.1
OG0021871	Dr16597	Dt26285.1
OG0021873	Dr16620	Dt09763.1
OG0021875	Dr16770	Dt06610.1
OG0021879	Dr17261	Dt17160.1
OG0021881	Dr17620	Dt24650.1
OG0021882	Dr17640	Dt10347.1

OG0021885	Dr18134	Dt16465.1
OG0021888	Dr18936	Dt18387.1
OG0021889	Dr18949	Dt26977.1
OG0021890	Dr18969	Dt11359.1
OG0021895	Dr19519	Dt17534.1
OG0021898	Dr19732	Dt05794.1
OG0021899	Dr19748	Dt05810.1
OG0021901	Dr19857	Dt10069.1
OG0021902	Dr19862	Dt28896.1
OG0021904	Dr20173	Dt22646.1
OG0021905	Dr20326	Dt15490.1
OG0021908	Dr21098	Dt20578.1
OG0021909	Dr22093	Dt16125.1
OG0021911	Dr22598	Dt00130.1
OG0021914	Dr23907	Dt21635.1
OG0021918	Dr24673	Dt28030.1
OG0021921	Dr25584	Dt03011.1
OG0021923	Dr26055	Dt07786.1

Table 7.35: List of the 33 conserved genes between the 22 angiosperm species used to generate Figure 4.3.

Aegilops tauschii	***Amborella trichopoda***
EMT17105	evm_27.model.AmTr_v1.0_scaffold00045.225
EMT05400	evm_27.model.AmTr_v1.0_scaffold00013.76
EMT32094	evm_27.model.AmTr_v1.0_scaffold00001.239
EMT16855	evm_27.model.AmTr_v1.0_scaffold00024.269
EMT21000	evm_27.model.AmTr_v1.0_scaffold00002.300
EMT09619	evm_27.model.AmTr_v1.0_scaffold00103.69
EMT28862	evm_27.model.AmTr_v1.0_scaffold00034.41
EMT14703	evm_27.model.AmTr_v1.0_scaffold00057.205
EMT11689	evm_27.model.AmTr_v1.0_scaffold00019.6
EMT11947	evm_27.model.AmTr_v1.0_scaffold00067.225
EMT00740	evm_27.model.AmTr_v1.0_scaffold00064.38
EMT03164	evm_27.model.AmTr_v1.0_scaffold00163.5

EMT28169	evm_27.model.AmTr_v1.0_scaffold00006.57
EMT08744	evm_27.model.AmTr_v1.0_scaffold00001.157
EMT12933	evm_27.model.AmTr_v1.0_scaffold00057.273
EMT09041	evm_27.model.AmTr_v1.0_scaffold00045.90
EMT11615	evm_27.model.AmTr_v1.0_scaffold00116.16
EMT32187	evm_27.model.AmTr_v1.0_scaffold00069.116
EMT29045	evm_27.model.AmTr_v1.0_scaffold00026.15
EMT13874	evm_27.model.AmTr_v1.0_scaffold00071.127
EMT31384	evm_27.model.AmTr_v1.0_scaffold00185.9
EMT23251	evm_27.model.AmTr_v1.0_scaffold00105.25
EMT21873	evm_27.model.AmTr_v1.0_scaffold00019.172
EMT16726	evm_27.model.AmTr_v1.0_scaffold00079.39
EMT05902	evm_27.model.AmTr_v1.0_scaffold00029.298
EMT18819	evm_27.model.AmTr_v1.0_scaffold00025.394
EMT33602	evm_27.model.AmTr_v1.0_scaffold00002.215
EMT18210	evm_27.model.AmTr_v1.0_scaffold00029.31
EMT33004	evm_27.model.AmTr_v1.0_scaffold00066.232
EMT30519	evm_27.model.AmTr_v1.0_scaffold00137.24
EMT04067	evm_27.model.AmTr_v1.0_scaffold00068.157
EMT03910	evm_27.model.AmTr_v1.0_scaffold00106.57
EMT11416	evm_27.model.AmTr_v1.0_scaffold00022.388

Ananas comosus	***Arabidopsis thaliana***
Aco000675.1	AT3G17030.1
Aco003669.1	AT1G67320.3
Aco018356.1	AT4G35250.1
Aco003959.1	AT5G21930.1
Aco023108.1	AT3G28460.1
Aco019066.1	AT3G08010.1
Aco000869.1	AT1G49540.2
Aco020258.1	AT1G55040.1
Aco021912.1	AT5G57930.2
Aco014586.1	AT3G04460.1
Aco000300.1	AT5G54080.1
Aco004540.1	AT5G20990.1

Aco020812.1	AT2G01120.2
Aco011984.1	AT2G17020.1
Aco000653.1	AT1G50940.1
Aco006380.1	AT3G17465.1
Aco001988.1	AT3G54480.1
Aco015560.1	AT5G62140.1
Aco011552.1	AT2G07690.1
Aco004809.1	AT5G63620.1
Aco025404.1	AT5G39590.1
Aco021246.1	AT2G01320.3
Aco006504.1	AT2G18710.1
Aco022244.1	AT3G18390.1
Aco016871.1	AT3G46960.1
Aco020989.1	AT3G63140.1
Aco012872.1	AT2G44760.1
Aco010641.1	AT1G50320.1
Aco004752.1	AT5G36170.1
Aco013352.1	AT3G49725.1
Aco023254.1	AT5G44635.1
Aco017519.1	AT5G51430.1
Aco000294.1	AT1G51310.1

Brachypodium distachyon	***Carica papaya***
Bradi1g02910.1.p	evm.model.supercontig_9.65
Bradi5g07930.1.p	evm.model.supercontig_43.82
Bradi3g42580.1.p	evm.model.supercontig_20.183
Bradi1g72790.1.p	evm.model.supercontig_126.19
Bradi1g09450.1.p	evm.model.supercontig_14.62
Bradi3g48480.1.p	evm.model.supercontig_125.25
Bradi3g39160.1.p	evm.model.supercontig_65.71
Bradi1g52630.3.p	evm.model.supercontig_48.69
Bradi3g56890.2.p	evm.model.supercontig_136.43
Bradi3g28320.1.p	evm.model.supercontig_306.4
Bradi1g52291.1.p	evm.model.supercontig_131.80
Bradi5g24930.1.p	evm.model.supercontig_34.224

Bradi2g46720.1.p	evm.model.supercontig_65.90
Bradi1g76050.1.p	evm.model.supercontig_20.128
Bradi1g02630.1.p	evm.model.supercontig_1133.1
Bradi2g08840.2.p	evm.model.supercontig_146.23
Bradi1g18980.1.p	evm.model.supercontig_223.13
Bradi3g51640.1.p	evm.model.supercontig_3.335
Bradi3g54190.1.p	evm.model.supercontig_397.6
Bradi3g13345.1.p	evm.model.supercontig_47.14
Bradi5g25130.2.p	evm.model.supercontig_14.105
Bradi2g01610.1.p	evm.model.supercontig_65.118
Bradi3g19100.1.p	evm.model.supercontig_75.6
Bradi2g17390.1.p	evm.model.supercontig_3.507
Bradi3g04550.1.p	evm.model.supercontig_112.72
Bradi1g54030.1.p	evm.model.supercontig_370.4
Bradi5g17407.1.p	evm.model.supercontig_22.25
Bradi5g26060.1.p	evm.model.supercontig_809.1
Bradi1g25310.1.p	evm.model.supercontig_123.43
Bradi1g10007.2.p	evm.model.supercontig_129.63
Bradi2g31580.2.p	evm.model.supercontig_30.68
Bradi1g32317.2.p	evm.model.supercontig_3.324
Bradi1g72967.1.p	evm.model.supercontig_84.62

Dioscorea rotundata	***Dioscorea tokoro***
Dr16640	Dt03030.1
Dr09860	Dt01158.1
Dr12523	Dt11047.1
Dr12482	Dt03319.1
Dr15903	Dt00579.1
Dr22354	Dt05642.1
Dr18690	Dt03396.1
Dr02054	Dt13646.1
Dr13953	Dt15086.1
Dr10939	Dt22174.1
Dr06340	Dt12740.1
Dr02492	Dt25841.1

Dr19800	Dt13562.1
Dr18776	Dt26330.1
Dr12287	Dt25705.1
Dr15631	Dt09142.1
Dr02075	Dt23420.1
Dr08646	Dt19199.1
Dr14480	Dt18981.1
Dr05481	Dt03461.1
Dr06469	Dt22478.1
Dr04992	Dt19126.1
Dr11060	Dt22023.1
Dr17511	Dt26677.1
Dr19433	Dt02995.1
Dr02116	Dt12458.1
Dr08370	Dt28962.1
Dr25596	Dt22990.1
Dr13777	Dt18087.1
Dr14658	Dt10262.1
Dr01259	Dt02101.1
Dr11810	Dt15565.1
Dr12112	Dt08661.1

Ipomoea nil	***Musa acuminata***
INIL14g41253.t1	GSMUA_Achr6P20470_001
INIL08g13856.t1	GSMUA_Achr6P27530_001
INIL15g23839.t1	GSMUA_Achr10P07300_001
INIL01g14025.t1	GSMUA_Achr8P30620_001
INIL11g09796.t1	GSMUA_AchrUn_randomP00220_001
INIL12g21841.t1	GSMUA_Achr6P13440_001
INIL00g40193.t1	GSMUA_Achr8P22940_001
INIL06g36961.t1	GSMUA_Achr11P13790_001
INIL08g27175.t1	GSMUA_Achr7P27400_001
INIL11g26654.t1	GSMUA_Achr5P12950_001
INIL01g36761.t1	GSMUA_Achr10P00150_001
INIL12g21833.t1	GSMUA_Achr11P15140_001

INIL07g34522.t1	GSMUA_Achr9P21730_001
INIL06g38466.t1	GSMUA_Achr5P11170_001
INIL11g09930.t1	GSMUA_Achr11P18320_001
INIL13g07841.t1	GSMUA_Achr5P21030_001
INIL09g30182.t1	GSMUA_AchrUn_randomP14870_001
INIL11g26638.t1	GSMUA_Achr11P03150_001
INIL08g38811.t1	GSMUA_Achr10P30670_001
INIL02g10395.t1	GSMUA_Achr5P25890_001
INIL01g25544.t1	GSMUA_Achr1P19930_001
INIL07g34546.t2	GSMUA_Achr10P20030_001
INIL13g28674.t1	GSMUA_Achr4P12710_001
INIL03g15030.t1	GSMUA_Achr3P26360_001
INIL07g06230.t1	GSMUA_Achr9P03710_001
INIL08g13635.t1	GSMUA_Achr4P00420_001
INIL09g30090.t1	GSMUA_Achr4P28930_001
INIL10g12117.t1	GSMUA_Achr2P03790_001
INIL01g25690.t1	GSMUA_Achr3P21580_001
INIL02g40700.t1	GSMUA_Achr1P27060_001
INIL05g21782.t1	GSMUA_Achr6P13960_001
INIL05g22986.t1	GSMUA_Achr1P16820_001
INIL08g27205.t2	GSMUA_Achr1P22830_001

Nelumbo nucifera	***Olea europaea***
XP_010259273.1	OE6A046656P1
XP_010250659.1	OE6A082939P1
XP_010242859.1	OE6A007212P1
XP_010261895.1	OE6A030625P3
XP_010272157.1	OE6A077798P2
XP_010257188.1	OE6A091687P1
XP_010279373.1	OE6A077595P3
XP_010241885.1	OE6A095803P1
XP_010273620.1	OE6A010919P2
XP_010264914.1	OE6A014806P3
XP_010279197.1	OE6A078624P1
XP_010240846.1	OE6A061078P3

XP_010279414.1	OE6A088392P2
XP_010266924.1	OE6A025269P3
XP_010259654.1	OE6A067429P1
XP_010247570.1	OE6A072434P1
XP_019054266.1	OE6A024002P2
XP_010255555.1	OE6A051547P1
XP_010245540.1	OE6A014354P1
XP_010272510.1	OE6A040590P1
XP_010246950.1	OE6A097096P2
XP_010243793.1	OE6A106191P1
XP_010268356.1	OE6A049514P1
XP_010265125.1	OE6A088642P1
XP_010256687.1	OE6A021983P1
XP_010255955.1	OE6A036235P1
XP_010275474.1	OE6A054217P3
XP_010249669.1	OE6A011795P1
XP_010258507.1	OE6A112601P1
XP_010248316.1	OE6A069298P1
XP_010260836.1	OE6A058187P1
XP_010264264.1	OE6A071633P1
XP_010257320.1	OE6A100405P1

Oropetium thomaeum	***Oryza sativa***
Oropetium_20150105_24806A	LOC_Os03g61700.1
Oropetium_20150105_08801A	LOC_Os07g22400.2
Oropetium_20150105_14522A	LOC_Os08g44000.1
Oropetium_20150105_00256A	LOC_Os03g08070.1
Oropetium_20150105_07486A	LOC_Os03g52640.2
Oropetium_20150105_02729A	LOC_Os02g39740.1
Oropetium_20150105_20316A	LOC_Os08g38570.1
Oropetium_20150105_21236A	LOC_Os07g22024.1
Oropetium_20150105_04106A	LOC_Os02g50010.1
Oropetium_20150105_19912A	LOC_Os10g32960.1
Oropetium_20150105_16848A	LOC_Os06g01360.1
Oropetium_20150105_01964A	LOC_Os04g56620.1

Oropetium_20150105_11570A	LOC_Os01g49010.1
Oropetium_20150105_22666A	LOC_Os03g04270.1
Oropetium_20150105_24829A	LOC_Os03g61920.1
Oropetium_20150105_09759A	LOC_Os01g14830.1
Oropetium_20150105_10311A	LOC_Os07g46555.1
Oropetium_20150105_21146A	LOC_Os02g45460.1
Oropetium_20150105_04513A	LOC_Os02g55410.1
Oropetium_20150105_13685A	LOC_Os08g01760.1
Oropetium_20150105_01979A	LOC_Os04g56790.1
Oropetium_20150105_06978A	LOC_Os01g03144.1
Oropetium_20150105_13892A	LOC_Os08g15460.1
Oropetium_20150105_03971A	LOC_Os05g47850.1
Oropetium_20150105_14394A	LOC_Os02g06500.1
Oropetium_20150105_04761A	LOC_Os07g11110.1
Oropetium_20150105_12275A	LOC_Os04g46100.1
Oropetium_20150105_02069A	LOC_Os04g57930.1
Oropetium_20150105_13573A	LOC_Os07g36250.1
Oropetium_20150105_07429A	LOC_Os03g51790.1
Oropetium_20150105_03288A	LOC_Os05g14590.1
Oropetium_20150105_15633A	LOC_Os06g45830.1
Oropetium_20150105_00241A	LOC_Os03g07850.1

Panicum hallii	***Phalaenopsis equestris***
Pahal.I01323.1	XP_020582237.1
Pahal.G00897.1	XP_020591204.1
Pahal.F00410.1	XP_020582231.1
Pahal.I00629.1	XP_020582120.1
Pahal.I01828.1	XP_020577338.1
Pahal.A02491.1	XP_020581413.1
Pahal.F01574.1	XP_020576240.1
Pahal.B01517.1	XP_020577913.1
Pahal.A03346.1	XP_020576918.1
Pahal.I03097.1	XP_020579009.1
Pahal.J01112.1	XP_020592921.1
Pahal.G02688.1	XP_020592276.1

Pahal.E02098.1	XP_020599620.1
Pahal.I00308.1	XP_020574219.1
Pahal.A01090.1	XP_020576822.1
Pahal.E03547.1	XP_020587815.1
Pahal.B04833.1	XP_020596701.1
Pahal.A02906.1	XP_020574142.1
Pahal.A03705.1	XP_020586983.1
Pahal.F00219.1	XP_020589099.1
Pahal.G02707.1	XP_020576338.1
Pahal.E03798.1	XP_020592864.1
Pahal.F02211.1	XP_020572356.1
Pahal.C02041.1	XP_020599364.1
Pahal.A00372.1	XP_020571087.1
Pahal.B01391.1	XP_020584826.1
Pahal.G01796.1	XP_020599331.1
Pahal.G02811.1	XP_020570778.1
Pahal.B04238.1	XP_020590227.1
Pahal.I01906.1	XP_020576917.1
Pahal.C00852.1	XP_020589926.1
Pahal.C00341.1	XP_020584992.1
Pahal.I00613.1	XP_020590686.1

Phoenix dactylifera	***Setaria viridis***
XP_008809273.1	Sevir.9G023000.1.p
XP_008809555.1	Sevir.7G079700.1.p
XP_017696803.1	Sevir.6G252900.1.p
XP_008797345.1	Sevir.9G524700.1.p
XP_017702466.1	Sevir.9G099600.1.p
XP_008778393.1	Sevir.1G232900.1.p
XP_008789255.1	Sevir.6G201900.1.p
XP_008777783.1	Sevir.2G115700.1.p
XP_008810393.1	Sevir.1G316000.1.p
XP_008785165.1	Sevir.9G227200.1.p
XP_008784438.1	Sevir.4G002300.3.p
XP_008787719.1	Sevir.3G026600.1.p

XP_008787657.1	Sevir.5G276200.1.p
XP_008776320.1	Sevir.9G554100.1.p
XP_008800859.1	Sevir.9G025600.1.p
XP_008790485.1	Sevir.5G060900.1.p
XP_008813420.1	Sevir.2G431500.1.p
XP_008783353.1	Sevir.1G275800.1.p
XP_008784797.1	Sevir.1G358900.1.p
XP_008803674.1	Sevir.6G008000.1.p
XP_008776884.1	Sevir.3G028700.1.p
XP_008811977.1	Sevir.5G081900.2.p
XP_008789707.1	Sevir.6G102800.1.p
XP_008787651.1	Sevir.3G159200.1.p
XP_008798964.1	Sevir.1G078500.1.p
XP_008790965.1	Sevir.2G085800.1.p
XP_008786174.1	Sevir.7G197600.1.p
XP_008811076.1	Sevir.3G018200.1.p
XP_008788225.1	Sevir.2G355000.1.p
XP_008785720.1	Sevir.9G106800.1.p
XP_008783215.1	Sevir.3G084700.1.p
XP_008796946.1	Sevir.4G279700.1.p
XP_008801937.1	Sevir.9G526000.1.p

Sorghum bicolor	***Spirodela polyrhiza***
Sobic.001G025900.1.p	Spipo6G0021500
Sobic.006G051000.3.p	Spipo13G0025200
Sobic.007G173800.1.p	Spipo1G0076600
Sobic.001G484400.1.p	Spipo15G0047300
Sobic.001G099000.1.p	Spipo2G0061900
Sobic.004G211600.1.p	Spipo24G0014000
Sobic.007G220300.1.p	Spipo0G0054900
Sobic.002G112600.1.p	Spipo22G0011500
Sobic.004G248400.3.p	Spipo4G0069400
Sobic.001G224600.1.p	Spipo17G0046800
Sobic.010G001600.1.p	Spipo8G0047800
Sobic.006G251800.1.p	Spipo4G0042900

Sobic.003G259600.1.p	Spipo14G0038800
Sobic.001G514000.1.p	Spipo17G0026800
Sobic.001G023200.1.p	Spipo3G0038200
Sobic.003G113400.1.p	Spipo27G0015000
Sobic.002G404200.1.p	Spipo3G0056400
Sobic.004G283100.1.p	Spipo1G0096000
Sobic.004G331600.1.p	Spipo28G0010200
Sobic.007G009100.1.p	Spipo9G0052000
Sobic.006G255000.1.p	Spipo1G0060800
Sobic.003G092800.1.p	Spipo19G0025400
Sobic.007G090537.1.p	Spipo28G0016700
Sobic.009G223200.1.p	Spipo25G0001300
Sobic.004G048800.1.p	Spipo6G0076300
Sobic.002G077000.1.p	Spipo6G0004600
Sobic.006G166500.2.p	Spipo5G0050800
Sobic.006G265900.1.p	Spipo1G0044300
Sobic.002G331100.1.p	Spipo9G0047300
Sobic.001G105700.1.p	Spipo3G0109500
Sobic.009G087000.1.p	Spipo8G0051100
Sobic.010G223400.1.p	Spipo1G0081200
Sobic.001G486500.3.p	Spipo3G0096600

Vitis vinifera	***Zosma marina***
GSVIVT01035001001	Zosma24g00920.1
GSVIVT01029087001	Zosma2g01350.1
GSVIVT01023933001	Zosma57g00770.1
GSVIVT01035474001	Zosma240g00160.1
GSVIVT01032566001	Zosma4g00500.1
GSVIVT01032129001	Zosma185g00120.1
GSVIVT01011835001	Zosma10g01730.1
GSVIVT01000088001	Zosma156g00350.1
GSVIVT01023615001	Zosma401g00200.1
GSVIVT01023441001	Zosma215g00280.1
GSVIVT01014248001	Zosma222g00210.1
GSVIVT01037501001	Zosma184g00210.1

GSVIVT01011806001	Zosma68g00550.1
GSVIVT01023957001	Zosma109g00210.1
GSVIVT01001984001	Zosma48g00860.1
GSVIVT01017688001	Zosma24g01150.1
GSVIVT01000659001	Zosma342g00070.1
GSVIVT01010637001	Zosma1g02110.1
GSVIVT01001046001	Zosma55g00150.1
GSVIVT01022421001	Zosma87g00040.1
GSVIVT01009862001	Zosma240g00050.1
GSVIVT01011781001	Zosma214g00430.1
GSVIVT01032310001	Zosma218g00170.1
GSVIVT01008296001	Zosma452g00030.1
GSVIVT01024809001	Zosma7g01800.1
GSVIVT01011082001	Zosma8g01310.1
GSVIVT01018294001	Zosma259g00170.1
GSVIVT01026149001	Zosma19g00280.1
GSVIVT01019055001	Zosma349g00110.1
GSVIVT01002450001	Zosma114g00430.1
GSVIVT01003253001	Zosma177g00240.1
GSVIVT01018659001	Zosma85g00550.1
GSVIVT01025274001	Zosma55g00350.1

Table 7.18: Assessment of the completeness of *D. tokoro* gene model set using 1,440 benchmarking universal single-copy orthologs (BUSCO) by BUSCOv3.0.0 using the embryophyta_odb9 (30 species) dataset.

BUSCO Type	No. of BUSCOs	% of BUSCOs
Complete Single-copy	1057	70.8
Complete Duplicated	38	2.6
Fragmented	122	8.5
Missing	261	18.1

Table 7.19: Assessing the completeness of three *D. alata* assemblies with different minimum contig size cut offs using the 248 most highly-conserved Core Eukaryotic Genes by CEGMA.

Cut off (bp)	No. complete	Complete (%)	No. partial	Partial (%)
1000	212	85.48	241	97.18
2000	212	85.48	241	97.18
3000	211	85.08	241	97.18

*"'Complete' refers to those predicted proteins in the set of 248 CEGs that when aligned to the HMM for the KOG for that protein-family, give an alignment length that is 70% of the protein length. I.e. if CEGMA produces a 100 amino acid protein, and the alignment length to the HMM to which that protein should belong is 110, then we would say that the protein is complete (91% aligned). If a protein is not complete, but if it still exceeds a pre-computed minimum alignment score, then we call the protein 'partial'..... Note that a protein that is deemed to be 'Complete' will also be included in the set of Partial matches.".

Table 7.20: Assessment of the completeness of *D. alata* gene model set using 956 benchmarking universal single-copy orthologs (BUSCO) by BUSCOv3.0.0 using the early plantae release dataset.

BUSCO Type	No. of BUSCOs	% of BUSCOs
Complete Single-copy	845	88
Complete Duplicated	247	25
Fragmented	58	6.0
Missing	53	5.5

Table 7.21: Gene ontology terms for all significantly enriched orthogroups conserved across 26 angiosperm species, when compared with all orthogroups of *D. tokoro*.

GO term	Ontology	Description	Number in input list	Number in BG/Ref	p-value	q
GO:0044267	P	cellular protein metabolic process	798	1330	1.2e-12	8.7e-11
GO:0019538	P	protein metabolic process	873	1496	3.1e-10	1.1e-08
GO:0009987	P	cellular process	2444	4425	9.5e-10	2.3e-08
GO:0006412	P	translation	222	337	2.9e-09	5.1e-08
GO:0006807	P	nitrogen compound metabolic process	1316	2336	6.1e-09	5.4e-08
GO:0044249	P	cellular biosynthetic process	979	1710	5.7e-09	5.4e-08
GO:0044238	P	primary metabolic process	1998	3611	6.1e-09	5.4e-08
GO:0044237	P	cellular metabolic process	2012	3632	3.8e-09	5.4e-08
GO:0044260	P	cellular macromolecule metabolic process	1463	2612	7.5e-09	5.9e-08
GO:0009058	P	biosynthetic process	1043	1842	3.4e-08	2.4e-07
GO:0016043	P	cellular component organization	487	825	1.6e-07	1e-06
GO:0043170	P	macromolecule metabolic process	1564	2841	3.6e-07	2.1e-06
GO:0010467	P	gene expression	720	1265	8.5e-07	4.6e-06
GO:0009059	P	macromolecule biosynthetic process	678	1201	6.1e-06	3.1e-05
GO:0034645	P	cellular macromolecule biosynthetic process	668	1187	1.1e-05	5e-05
GO:0006464	P	cellular protein modification process	507	887	1.3e-05	5e-05
GO:0008152	P	metabolic process	2416	4510	1.2e-05	5e-05
GO:0043412	P	macromolecule modification	559	984	1.3e-05	5e-05
GO:0006139	P	nucleobase-containing compound metabolic process	866	1576	6.6e-05	0.00025
GO:0007049	P	cell cycle	99	158	0.00077	0.0027
GO:0032502	P	developmental process	245	426	0.00094	0.0032
GO:0050789	P	regulation of biological process	662	1221	0.0017	0.0053
GO:0050794	P	regulation of cellular process	609	1120	0.0017	0.0053
GO:0048856	P	anatomical structure development	236	414	0.0021	0.0062
GO:0007275	P	multicellular organism development	203	353	0.0023	0.0066
GO:0005975	P	carbohydrate metabolic process	294	524	0.0025	0.0067
GO:0009056	P	catabolic process	307	550	0.003	0.0079
GO:0051179	P	localization	539	1001	0.0069	0.018
GO:0065007	P	biological regulation	731	1373	0.0077	0.018
GO:0032501	P	multicellular organismal process	215	383	0.0077	0.018
GO:0051234	P	establishment of localization	526	977	0.0076	0.018
GO:0006810	P	transport	520	969	0.01	0.023
GO:0019222	P	regulation of metabolic process	399	740	0.015	0.032
GO:0060255	P	regulation of macromolecule metabolic process	385	714	0.016	0.034
GO:0009719	P	response to endogenous stimulus	101	174	0.017	0.035
GO:0003723	F	RNA binding	315	520	8e-07	2.3e-05
GO:0005198	F	structural molecule activity	138	215	1.6e-05	0.00023
GO:0008135	F	translation factor activity, RNA binding	65	93	6.7e-05	0.00065
GO:0016772	F	transferase activity, transferring phosphorus-containing groups	420	740	0.00013	0.00091
GO:0016740	F	transferase activity	913	1684	0.00036	0.0021
GO:0000166	F	nucleotide binding	795	1468	0.00084	0.004
GO:0016301	F	kinase activity	340	615	0.0041	0.017
GO:0003824	F	catalytic activity	2165	4160	0.0065	0.023
GO:0003676	F	nucleic acid binding	737	1390	0.011	0.036
GO:0044424	C	intracellular part	2135	3817	6.1e-11	1e-09
GO:0005622	C	intracellular	2200	3928	2.4e-11	1e-09
GO:0032991	C	macromolecular complex	784	1324	5.5e-11	1e-09
GO:0005737	C	cytoplasm	1542	2735	5.2e-10	6.4e-09
GO:0012505	C	endomembrane system	315	496	1.1e-09	1.1e-08
GO:0044446	C	intracellular organelle part	822	1421	6.3e-09	3.4e-08
GO:0044444	C	cytoplasmic part	1349	2397	5.8e-09	3.4e-08
GO:0044422	C	organelle part	822	1421	6.3e-09	3.4e-08
GO:0005623	C	cell	2353	4275	5e-09	3.4e-08
GO:0044464	C	cell part	2331	4240	8.2e-09	4e-08
GO:0005829	C	cytosol	289	458	1.2e-08	5.3e-08
GO:0043229	C	intracellular organelle	1772	3205	3.1e-08	1.3e-07
GO:0043226	C	organelle	1773	3208	3.4e-08	1.3e-07
GO:0005794	C	Golgi apparatus	142	209	9.9e-08	3.2e-07
GO:0043232	C	intracellular non-membrane-bounded organelle	383	634	1e-07	3.2e-07
GO:0043228	C	non-membrane-bounded organelle	383	634	1e-07	3.2e-07
GO:0005840	C	ribosome	176	268	1.4e-07	4.1e-07
GO:0030529	C	intracellular ribonucleoprotein complex	256	413	6e-07	1.6e-06
GO:0005634	C	nucleus	654	1159	9e-06	2.3e-05
GO:0043227	C	membrane-bounded organelle	1589	2928	9.5e-06	2.3e-05
GO:0043231	C	intracellular membrane-bounded organelle	1586	2923	1e-05	2.3e-05
GO:0044428	C	nuclear part	253	421	1.8e-05	3.9e-05
GO:0005768	C	endosome	54	74	4.1e-05	8.8e-05
GO:0031974	C	membrane-enclosed lumen	202	334	6.4e-05	0.00012
GO:0043233	C	organelle lumen	202	334	6.4e-05	0.00012
GO:0070013	C	intracellular organelle lumen	202	334	6.4e-05	0.00012
GO:0031981	C	nuclear lumen	171	283	0.00023	0.00041
GO:0005783	C	endoplasmic reticulum	139	227	0.00036	0.00063
GO:0005773	C	vacuole	121	195	0.00039	0.00067
GO:0005856	C	cytoskeleton	62	100	0.0092	0.015
GO:0005635	C	nuclear envelope	28	42	0.02	0.032
GO:0005730	C	nucleolus	63	106	0.029	0.044

Table 7.22: Gene ontology terms for all orthogroups only observed in *D. tokoro*, when compared with orthogroups of 25 other angiosperm species.

GO term	Ontology	Description	Number in input list	Number in BG/Ref	p-value	q
GO:0044238	P	primary metabolic process	6	1998	0.58	0.72
GO:0009987	P	cellular process	7	2444	0.61	0.72
GO:0044237	P	cellular metabolic process	6	2012	0.58	0.72
GO:0008152	P	metabolic process	6	2416	0.72	0.72
GO:0005488	F	binding	7	2549	0.65	0.65
GO:0005623	C	cell	5	2353	0.81	0.81
GO:0044464	C	cell part	5	2331	0.81	0.81
GO:0005622	C	intracellular	5	2200	0.77	0.81
GO:0044424	C	intracellular part	5	2135	0.75	0.81

Table 7.23: Gene ontology terms for significantly enriched orthogroups only observed in *D. tokoro* and *D. rotundata*, when compared with conserved orthogroups of 24 other angiosperm species.

GO term	Ontology	Description	Number in input list	Number in BG/Ref	p-value	q
GO:0016787	F	hydrolase activity	51	738	2.5e-05	0.00046

Table 7.24: Gene ontology terms for all significantly enriched orthogroups observed in 25 angiosperm species and not *D. tokoro*, whe compared with those conserved between *D. tokoro* and the 25 angiosperm species.

GO term	Ontology	Description	Number in input list	Number in BG/Ref	p-value	q
GO:0009605	P	response to external stimulus	17	46	1.9e-09	9.7e-08
GO:0009607	P	response to biotic stimulus	13	28	5.2e-09	1.4e-07
GO:0051704	P	multi-organism process	13	43	2e-06	3.4e-05
GO:0009628	P	response to abiotic stimulus	18	99	6.3e-05	0.00082
GO:0050896	P	response to stimulus	54	510	0.00044	0.0046
GO:0015979	P	photosynthesis	9	46	0.0023	0.02
GO:0009536	C	plastid	84	457	3.2e-16	1.2e-14
GO:0005739	C	mitochondrion	57	275	3.8e-14	6.8e-13
GO:0044444	C	cytoplasmic part	168	1349	2e-12	2.4e-11
GO:0005737	C	cytoplasm	175	1542	5.8e-10	5.2e-09
GO:0043231	C	intracellular membrane-bounded organelle	177	1586	1.5e-09	9.9e-09
GO:0043227	C	membrane-bounded organelle	177	1589	1.6e-09	9.9e-09
GO:0043229	C	intracellular organelle	182	1772	1.5e-07	7e-07
GO:0043226	C	organelle	182	1773	1.6e-07	7e-07
GO:0044424	C	intracellular part	192	2135	7.8e-05	0.0003
GO:0005622	C	intracellular	197	2200	8.2e-05	0.0003
GO:0031975	C	envelope	22	150	0.00029	0.00088
GO:0044464	C	cell part	202	2331	0.00029	0.00088
GO:0005623	C	cell	202	2353	0.00043	0.0012
GO:0031967	C	organelle envelope	21	150	0.00071	0.0018
GO:0009579	C	thylakoid	13	77	0.0012	0.0029
GO:0005886	C	plasma membrane	24	203	0.0033	0.0074
GO:0005773	C	vacuole	14	121	0.023	0.05

Table 7.26: Total lengths of gene, coding, gypsie LTR and copia LTR, across pseudomolecules of *D. tokoro*

Pseudomolecule no.	Gene len	% of pseudo-chromosome	Exon len	% of pseudo-chromosome	Gypsie LTR len	% of pseudo-chromosome	Copia LTR len	% of pseudo-chromosome
01	15782833	34.21	4737096	10.27	5215394	11.30	2664546	5.78
02	11977215	33.11	3587121	9.92	4150838	11.48	2268832	6.27
03	11818129	33.58	3406598	9.68	4497328	12.78	2263227	6.43
04	8899001	34.53	2677614	10.39	3371017	13.08	1459684	5.66
05	8340483	32.66	2209648	8.65	3032603	11.87	1685105	6.60
06	8364716	33.08	2535310	10.03	3056177	12.09	1658633	6.56
07	9368985	39.11	3018531	12.60	1374926	5.74	1345329	5.62
08	7381112	30.85	2013332	8.41	2403123	10.04	1658774	6.93
09	8580225	36.33	2669155	11.30	2577908	10.91	1311375	5.55
10	7599847	32.40	2245977	9.58	1532452	6.53	1172310	5.00
Average	9811255	33.99	2910038	10.08	3121177	10.58	1748782	6.04

Table 7.27: Total counts of genes, exons, gypsie LTR and copia LTR(s), across pseudomolecules of *D. tokoro*

Pseudomolecule no.	Gene	Exon	Gypsie LTR	Copia LTR
01	3584	20127	11710	6826
02	2837	15475	9477	5589
03	2714	14696	9746	5139
04	2074	11341	7629	3571
05	1841	9726	6763	3933
06	1947	10756	7319	4012
07	2216	12377	4740	4082
08	1717	9259	5319	3580
09	2124	11771	7644	3627
10	1791	9684	3213	2355

Table 7.28: Total lengths of gene, coding, gypsie LTR and copia LTR, across pseudochromosomes of *D. rotundata*

Pseudomolecule no.	Gene len	% of pseudo-chromosome	Exon len	% of pseudo-chromosome	Gypsie LTR len	% of pseudomolecules	Copia LTR len	% of pseudo-chromosome
01	4132482	13.15	1453686	4.62	8154691	25.94	1134998	3.61
02	5418954	15.95	1708074	5.03	7508842	22.10	1763537	5.19
03	4910292	25.12	1609480	8.23	2577256	13.18	897831	4.59
04	7126633	25.68	2696090	9.72	4259648	15.35	1138979	4.10
05	9015125	27.43	3458351	10.52	4823626	14.67	1432633	4.36
06	6083040	18.51	1844037	5.61	7396513	22.50	1530698	4.66
07	4258951	23.59	1529072	8.47	2456747	13.61	906220	5.02
08	7110457	25.38	2677733	9.56	4122392	14.71	1222435	4.36
09	4708070	20.18	1497890	6.42	4596611	19.71	1042195	4.47
10	4006527	22.38	1397449	7.81	1958802	10.94	1031852	5.76
11	2787026	16.21	885060	5.15	4205568	24.46	813280	4.73
12	4150391	16.42	1474139	5.83	4922657	19.48	1312517	5.19
13	5002308	16.83	1559048	5.25	6848640	23.04	1485075	5.00
14	4346291	27.66	1358182	8.64	1471757	9.37	729346	4.64
15	4070504	38.86	1742957	16.64	253437	2.42	330693	3.16
16	4951717	21.19	1704951	7.29	4714358	20.17	1088892	4.66
17	4283328	20.81	1578456	7.67	3239229	15.74	987063	4.79
18	4822906	21.16	1753111	7.69	3673198	16.11	1113100	4.88
19	3889775	39.18	1570811	15.82	339592	3.42	365247	3.68
20	2737075	25.63	926654	8.68	1405532	13.16	503327	4.71
21	1964815	37.76	799504	15.37	378406	7.27	193339	3.72
Average	4751270	23.77	1677368	8.57	3776548	15.59	1001107	4.54

Table 7.29: Total counts of genes, exons, gypsie LTR and copia LTR(s), across pseudomolecules of *D. rotundata*

Pseudomolecule no.	Gene	Exon	Gypsie LTR	Copia LTR
01	920	5320	8701	1730
02	1049	6074	8242	2544
03	881	5460	2977	1354
04	1390	9621	4738	1699
05	1677	11950	5452	2144
06	1126	7480	7932	2325
07	820	5035	2804	1315
08	1350	9124	4635	1918
09	934	5857	5197	1567
10	777	5250	2472	1504
11	523	3351	4334	1142
12	786	5206	5366	1909
13	1036	5755	7379	2111
14	764	5177	1804	1206
15	799	5576	441	619
16	915	6038	5251	1635
17	845	5698	3650	1467
18	948	6130	4136	1655
19	669	5559	487	563
20	507	3582	1676	809
21	370	2409	491	291

Table 7.30: Orthogroups of previously identified B-lectin genes in *D. rotundata* compared to *Dioscorea tokoro*.

Orthogroup	Dioscorea rotundata	Dioscorea tokoro
OG0000005	31	6
OG0000052	20	7
OG0000094	14	8
OG0000102	12	0
OG0000152	3	3
OG0000332	10	9
OG0001314	1	2
OG0002180	20	3
OG0002758	1	0
OG0003693	1	1
OG0004255	2	2
OG0007457	1	1
OG0010291	2	1
OG0010639	5	10
OG0010744	1	0
OG0011262	1	1
OG0016586	1	4
OG0018766	1	0

Table 7.36: Comparison of the total number of genes with 1:1 orthology in the autosomal, PAR and sex determination loci (FSW/MSY) regions, between *D. tokoro* and *D. rotundata* and with *D. alata*.

	Total No.	**No. 1:1 with *D.rotundata***	**No. 1:1 with *D.alata***
D. tokoro			
Autosome	20131	3241	2061
PAR	1449	128	207
MSY	1265	282	83
Total	22845	3651	2351
	Total No.	**No. 1:1 with *D.tokoro***	**No. 1:1 with *D.alata***
D. rotundata			
Autosome	18563	3801	3364
PAR	466	80	74
FSW	57	10	7
Total	19086	3891	3445

Appendix B

Tamiru et al. BMC Biology (2017) 15:86
DOI 10.1186/s12915-017-0419-x

BMC Biology

RESEARCH Open Access

Genome sequencing of the staple food crop white Guinea yam enables the development of a molecular marker for sex determination

Muluneh Tamiru[1†], Satoshi Natsume[1†], Hiroki Takagi[1†], Benjamen White[2†], Hiroki Yaegashi[1†], Motoki Shimizu[1†], Kentaro Yoshida[3], Aiko Uemura[1], Kaori Oikawa[1], Akira Abe[1], Naoya Urasaki[4], Hideo Matsumura[5], Pachakkil Babil[6], Shinsuke Yamanaka[7], Ryo Matsumoto[7], Satoru Muranaka[7], Gezahegn Girma[8], Antonio Lopez-Montes[8], Melaku Gedil[8], Ranjana Bhattacharjee[8], Michael Abberton[8], P. Lava Kumar[8], Ismail Rabbi[8], Mai Tsujimura[9], Toru Terachi[9], Wilfried Haerty[2], Manuel Corpas[2], Sophien Kamoun[10], Günter Kahl[11^], Hiroko Takagi[7*], Robert Asiedu[8*] and Ryohei Terauchi[1,12*]

Abstract

Background: Root and tuber crops are a major food source in tropical Africa. Among these crops are several species in the monocotyledonous genus *Dioscorea* collectively known as yam, a staple tuber crop that contributes enormously to the subsistence and socio-cultural lives of millions of people, principally in West and Central Africa. Yam cultivation is constrained by several factors, and yam can be considered a neglected "orphan" crop that would benefit from crop improvement efforts. However, the lack of genetic and genomic tools has impeded the improvement of this staple crop.

Results: To accelerate marker-assisted breeding of yam, we performed genome analysis of white Guinea yam (*Dioscorea rotundata*) and assembled a 594-Mb genome, 76.4% of which was distributed among 21 linkage groups. In total, we predicted 26,198 genes. Phylogenetic analyses with 2381 conserved genes revealed that *Dioscorea* is a unique lineage of monocotyledons distinct from the Poales (rice), Arecales (palm), and Zingiberales (banana). The entire *Dioscorea* genus is characterized by the occurrence of separate male and female plants (dioecy), a feature that has limited efficient yam breeding. To infer the genetics of sex determination, we performed whole-genome resequencing of bulked segregants (quantitative trait locus sequencing [QTL-seq]) in F1 progeny segregating for male and female plants and identified a genomic region associated with female heterogametic (male = ZZ, female = ZW) sex determination. We further delineated the W locus and used it to develop a molecular marker for sex identification of Guinea yam plants at the seedling stage.

(Continued on next page)

* Correspondence: hiroko55105@gmail.com; r.asiedu@cgiar.org; terauchi@ibrc.or.jp
†Equal contributors
^Deceased
[7]Japan International Research Center for Agricultural Sciences, Tsukuba, Japan
[8]International Institute of Tropical Agriculture, Ibadan, Nigeria
[1]Iwate Biotechnology Research Center, Kitakami, Japan
Full list of author information is available at the end of the article

(Continued from previous page)

Conclusions: Guinea yam belongs to a unique and highly differentiated clade of monocotyledons. The genome analyses and sex-linked marker development performed in this study should greatly accelerate marker-assisted breeding of Guinea yam. In addition, our QTL-seq approach can be utilized in genetic studies of other outcrossing crops and organisms with highly heterozygous genomes. Genomic analysis of orphan crops such as yam promotes efforts to improve food security and the sustainability of tropical agriculture.

Keywords: Yam, *Dioscorea*, Whole-genome sequence, Dioecy, Sex determination

Background

Yam is a collective name for tuber-bearing crops belonging to the monocotyledonous *Dioscorea* genus in the family Dioscoreaceae of the order Dioscoreales. This genus contains approximately 450 species which are primarily distributed in tropical and subtropical regions worldwide [1]. Among the Dioscoreaceae, three minor genera are monoecious (having male and female flowers on a plant), but the entire genus *Dioscorea* is characterized by dioecy (the presence of separate male and female plants), a feature shared by only 5–6% of angiosperms [2]. The origin of *Dioscorea* is supposed to be in the Late Cretaceous (~80 Mya [3]), suggesting that the origin of dioecy dates back to this time. Approximately 10 *Dioscorea* species have been independently domesticated in West Africa, Southeast Asia, and the Pacific and Caribbean islands [4]. *D. rotundata* is the most popular species in West and Central Africa, the main region for yam production worldwide, which contributed approximately 96% of the 63 million tons of yam produced globally in 2013 (Additional file 1: Table S1 and Additional file 2: Figure S1). *D. rotundata* (white Guinea yam) and *D. cayenensis* (yellow Guinea yam) represent a major source of food and income in this region, as well as an integral part of the socio-cultural life. This geographical region is often referred to as the "civilization of the yam," reflecting the West African societies that are tightly linked to yam cultivation [5, 6].

Despite its considerable regional importance, Guinea yam has long been regarded as an "orphan" crop, as it is not traded around the world, and it has attracted little attention from researchers and little investment. Guinea yam cultivation is constrained by several factors. Seeds are seldom used as starting materials; instead, yams are commonly propagated clonally using small whole tubers (referred to as "seed yams") or tuber pieces. Yam is an annual climber that requires stakes for support and is highly vulnerable to a plethora of pests and diseases. Therefore, an understanding of yam genetics and a systematic improvement of yam based on crossbreeding for traits associated with tuber yield and quality, a reduced requirement for staking, and resistance/tolerance to disease and nematodes are urgently needed. Genetic analysis of *Dioscorea* has been constrained by the small number of available genetic markers. Furthermore, *Dioscorea* cultivars are highly heterozygous due to their obligate outcrossing. This heterozygosity renders genetic analysis approaches commonly used in inbreeding species, e.g., linkage analysis using the segregating progeny of an F2 generation and recombinant inbred lines (RILs), inapplicable to yam.

The International Institute of Tropical Agriculture (IITA) has a global mandate for yam research and development within the CGIAR Consortium [7]. We initiated a yam genomics program several years ago as part of an IITA-coordinated international collaboration. To generate genetic and genomic tools for yam breeding, we sequenced and assembled a highly heterozygous diploid genome of *D. rotundata*. We used this genome sequence and genetic resources to identify a locus associated with sex determination, which we used to develop a diagnostic marker for sex identification at the seedling stage. These genomic resources broaden our knowledge of Guinea yam genetics and provide a platform for implementing genomics-assisted breeding by marker-assisted selection (MAS) in this important staple crop.

Results

Whole-genome sequencing (WGS) and assembly

To generate a *D. rotundata* genome sequence, an individual plant, TDr96_F1, was selected from the progeny in the open-pollinated *D. rotundata* breeding line TDr96/00629 (Fig. 1a, b). As TDr96_F1 never flowered during the current study period, we could not determine its sex. While *D. rotundata* is characterized by different ploidy levels (2× and 3×) with a basic chromosome number of 20 [8, 9], we found TDr96_F1 to be diploid ($2n = 2\times = 40$) based on the mitotic chromosome number within root meristem cells (Fig. 1c). We estimated the genome size of TDr96_F1 to be 570 Mb by flow cytometry (FCM) analysis (Fig. 1d).

We used total DNA from fresh leaf samples to prepare a paired-end (PE) library and eight types of mate-pair (MP) jump libraries with insert sizes of 2, 3, 4, 5, 6, 8, 20, and 40 kb and sequenced the PE and MP jump libraries on Illumina sequencers. We also generated a 100-kb jump bacterial artificial chromosome (BAC) library, from which 9984 clones were subjected to BAC-end Sanger

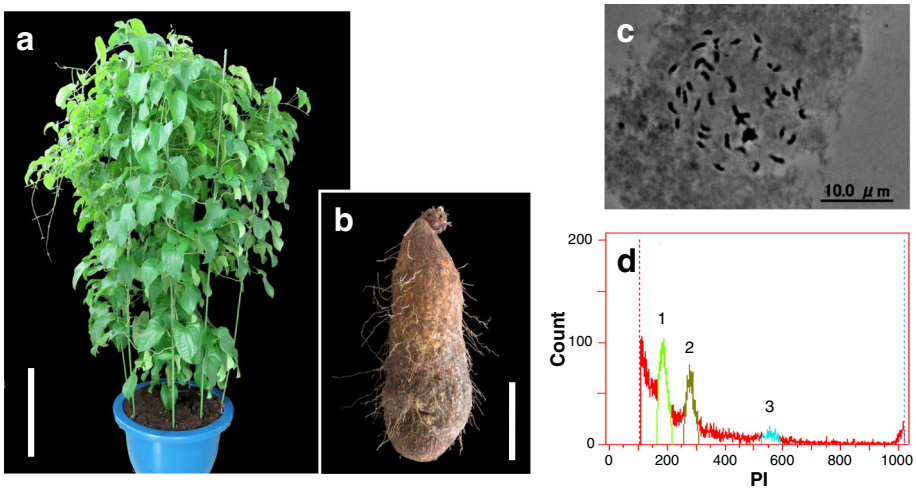

Fig. 1 Determination of ploidy level and genome size in *Dioscorea rotundata* plant TDr96_F1. **a** TDr96_F1 plant grown in a greenhouse at Iwate Biotechnology Research Center (IBRC), Japan. Bar = 50 cm. **b** TDr96_F1 tuber. Bar = 10 cm. **c** Diploid somatic chromosomes at metaphase stage obtained from TDr96_F1 root tips ($2n = 2\times = 40$). **d** FCM histogram of propidium iodide (PI)-stained nuclei from *D. rotundata* (TDr96_F1) and rice (*Oryza sativa* L.). Rice (genome size = 380 Mb) served as an internal reference standard. 1 = G1 (*O. sativa*), 2 = G1 (*D. rotundata*), and 3 = G2 (*D. rotundata*), where G1 and G2 represent the Gap 1 and Gap 2 phases of the cell cycle, respectively

sequencing, resulting in PE reads corresponding to a 0.46-Gb sequence with ~ 0.8× genome coverage (Additional file 1: Table S2 and Additional file 2: Figure S2). In total, we generated 85.14 Gb of sequencing reads, representing ~ 149.4× coverage of the estimated 570-Mb genome (Additional file 1: Table S2). Using *k*-mer analysis-based genome size estimation [10] of TDr96_F1 PE reads with ALLPATHS-LG [11] (see below), we found that the genome size was roughly 579 Mb, which is similar to the size estimated by FCM (Additional file 2: Figure S3). The PE and MP jump reads were used for de novo assembly with the ALLPATHS-LG assembler [11], which provides good performance even for highly heterozygous genomes [12]. Further scaffolding with SSPACE software using the 100-kb jump reads [13] (Additional file 2: Figure S4) generated 4723 scaffolds with a total length of 594 Mb, i.e., 2.6% and 4.2% longer than the genome size estimated by *k*-mer (579 Mb) and FCM (570 Mb) analyses, respectively. We estimated the scaffold N50 to be 2.12 Mb (longest scaffold: 13.6 Mb), with approximately 93.9% of the assembly represented by 586 scaffolds longer than 100 kb (Additional file 1: Table S3). From ALLPATHS-LG output, we judged that more than 1.4 million sites were potentially heterozygous (Table 1). This assembly is hereafter referred to as the "TDr96_F1 reference genome."

We assessed the quality of our assembly by investigating the presence of 248 highly conserved core eukaryotic genes with the Core Eukaryotic Genes Mapping Approach (CEGMA) [14] and confirmed the presence of 243 (98%) of those genes (Additional file 1: Table S4). Similarly, 94% of 956 Benchmarking Universal Single-Copy Orthologs (BUSCOs) [15] were present in at least one complete single copy in the assembly (Additional file 1: Table S5). Since the TDr96_F1 reference genome was generated from total genomic DNA, it also contained organelle-derived sequences. Alignment of the TDr96_F1 PE reads to the published *D. rotundata* chloroplast genome sequence [16] showed that 14.7% of the total PE reads were derived from the chloroplast genome (Additional file 1: Table S6). We also isolated mitochondrial DNA from TDr96_F1 leaves, sequenced this DNA using PE reads with Illumina MiSeq, and generated a 564-kb de novo assembly comprising 76 scaffolds (Additional file 2: Figure S5). Among PE reads, 1.25% represented mitochondrial sequences.

Generation of pseudo-chromosomes by anchoring scaffolds onto a linkage map

We developed a genetic map of *D. rotundata* using 150 F1 individuals obtained from a cross between two heterozygous breeding lines, TDr97/00917 (P1, female) and TDr99/02627 (P2, male), using restriction site associated DNA (RAD)-tags as DNA markers [17] (Additional file 2: Figures S6, S7) and the pseudo-testcross method [18, 19]. We aligned RAD-tags to TDr96_F1 scaffold sequences and selected DNA markers heterozygous in P1 and homozygous in P2, as well as markers heterozygous in P2 and homozygous in P1, resulting in 1326 and 1272 markers for P1 and P2 heterozygous sites, respectively (Additional file 1: Table S7 and Additional file 2: Figure S8). We then calculated the recombination fraction (*rf*)

Table 1 Characteristics of nuclear genome sequence in *Dioscorea rotundata* and other angiosperms

Feature	Value			
	D. rotundata (v0.1)	*A. thaliana* (TAIR10)	*B. distachyon* (v3.1)	*O. sativa* (v7_JGI 323)
Total length (Mbp)	594.23	119.67	271.16	374.47
GC (%)	35.83	36.06	46.40	43.57
Number of scaffolds (≥ 0 bp)	4723	7	10	14
Number of scaffolds (≥ 1000 bp)	4704	7	10	14
Largest scaffold (Mbp)	13.61	30.43	75.07	43.27
N50 (Mbp)	2.12	23.46	59.13	29.96
N75 (Mbp)	0.77	19.70	48.59	28.44
Number of Ns per 100 kb	282.45[a]	155.60	155.85	44.13
Ambiguous bases	1,413,029	–	–	–
Number of genes	26,198	27,416	34,310	42,189
Exons				
Number	158,059	141,044	154,104	178,353
Average number per gene	6.03	5.14	4.49	4.25
Total length (Mbp)	42.43	33.49	39.01	46.85
Average size (bp)	268.43	237.46	253.15	262.70
Average GC (%)	44.08	43.70	51.02	51.12
Introns				
Number	105,663	86,212	85,484	94,345
Average number per gene	4.03	3.14	2.49	2.25
Total length (Mbp)	83.12	17.87	47.70	53.34
Average size (bp)	630.33	157.25	398.18	391.23
Average GC (%)	32.37	32.45	38.29	37.20
Transposable elements[b]				
% Total interspersed	46.07	13.32	37.39	44.40
Total interspersed total length (Mbp)	274.51	15.94	101.39	166.27
% Short interspersed nuclear elements (SINEs)	0.02	0.17	0.38	0.88
SINEs total length (Mbp)	0.13	0.20	1.02	3.31
% Long interspersed nuclear elements (LINEs)	2.43	1.07	2.91	1.29
LINEs total length (Mbp)	14.46	1.29	7.90	4.83
% Long terminal repeat (LTR) elements	22.82	6.35	19.31	21.09
LTR elements total length (Mbp)	135.71	7.61	52.36	78.98
% DNA elements	6.70	3.08	7.11	16.7
DNA elements total length (Mbp)	39.83	3.69	19.27	62.82
% Unclassified	14.20	2.64	7.68	4.36
Unclassified total length (Mbp)	84.38	3.16	20.84	16.32

[a]Number of Ns per 100 kb using the *D. rotundata* broken scaffolds
[b]Transposable elements were identified by masking the genomes using RepeatModeler and RepeatMasker, with the same parameters across all species

between the RAD markers to generate linkage maps. If the pairwise *rf* value of two RAD markers on the same scaffold exceeded 0.25, the scaffold was divided halfway between the markers because they were likely misassembled (see explanation in Additional file 2: Figure S9 and Additional file 1: Table S8). Two linkage maps, P1-Map and P2-Map, were generated based on the segregation pattern of the selected markers in the F1 progeny (Additional file 2: Figure S10), to which *D. rotundata* scaffolds were anchored using the 100-bp DNA sequences of RAD-tags. We combined the two maps using shared scaffolds (Additional file 2: Figures S11, S12), which allowed the ~ 454-Mb sequence (representing 76.4% of the assembly) to be anchored onto 21 linkage groups (LGs) to

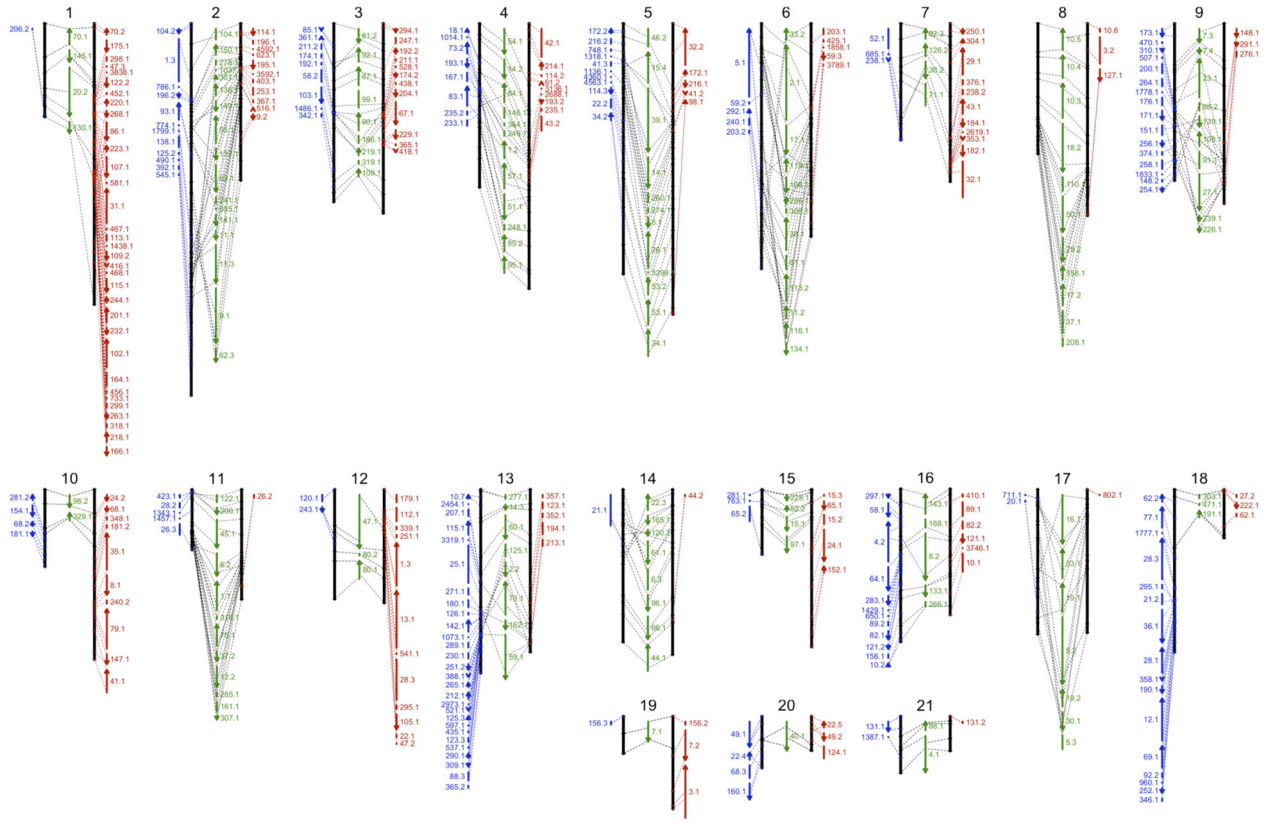

Fig. 2 Integrated genetic and physical map of *D. rotundata*. Approximately 76.4% of *D. rotundata* scaffold sequences were anchored using a RAD-based genetic map generated with 150 F1 individuals obtained from a cross between TDr97/00917 (P1, female) and TDr99/02627 (P2, male). The 21 chromosome-scale pseudo-molecules are numbered from 1 to 21. Markers are located according to genetic distance (cM). *Black lines* represent the 21 P1 and P2 linkage groups (LGs), and scaffolds anchored to P1 and P2 LGs are shown in *red* and *blue*, respectively. Scaffolds shared between the P1 and P2 LGs are shown in *green*. *Numbers* and *arrows* indicate scaffolds and their orientation, respectively

construct chromosome-scale pseudo-molecules (Fig. 2 and Additional file 1: Table S9). Smaller LGs could not be unequivocally mapped; hence, 21 LGs were obtained, whereas 20 LGs are expected based on the basic chromosome number. We validated the quality of our assembly by comparing the pseudo-molecule sequence with newly sequenced PE sequence reads having an average insert size of ~ 100 kb obtained from the BACs. Among the 315 BAC clones for which sequences of both ends could be mapped onto the assembly, 265 (84.1%) had both pairs in the same scaffold in the correct orientation, with an average distance of 116 kb (Additional file 1: Table S10 and Additional file 2: Figure S13), confirming the quality of our assembly. We compared the de novo assembled scaffolds to linkage information about the RAD markers, finding that 75% of RAD markers on the same scaffolds had $rf < 0.25$, and 73.5% of the scaffolds were retained without the need for splitting (Additional file 1: Table S11). The remaining 26.5% of the scaffolds had an $rf > 0.25$ and were divided into two or more scaffolds to solve the inconsistency between assembly and linkage information.

Guinea yam gene prediction and comparative genomics

We predicted genes and transposons using the TDr96_F1 reference genome sequence. To construct reliable gene models, we followed the MAKER pipeline using RNA-seq data from 18 samples representing various *D. rotundata* tissues (Additional file 1: Tables S12, S13) and combined the data with publicly available expressed sequence tags (ESTs) and homologous protein sequences from related angiosperm species (Additional file 2: Figure S14). This resulted in the prediction of 26,198 genes (Table 1 and Additional file 3), 22,477 (85.8%) of which are supported by RNA-seq data.

We compared Guinea yam genome sequence metrics with those of *Arabidopsis thaliana* (dicot), *Brachypodium distachyon* (monocot), and *Oryza sativa* (monocot) (Table 1). Interestingly, the GC contents of the total genome and exons of protein-coding genes in Guinea yam were 35.8% and 44.1%, respectively; these values are close to those of *Arabidopsis* and much lower than those of the Poales species *Brachypodium* and *Oryza* (Table 1). We annotated an average of 6.03 exons and 4.03 introns per gene. Roughly half of the genome was represented

by an interspersed sequence (274.5 Mb), a major component of which was long terminal repeat (LTR) sequences (135.7 Mb) (Table 1).

We identified 5557 *D. rotundata* genes with a 1:1:1:1 orthologous relationship to the high-quality *B. distachyon*, *O. sativa*, and *A. thaliana* gene models (Fig. 3a and Additional files 4, 5). This number was reduced to 2795 genes when we included Arecales (*Elaeis guineensis*, *Phoenix dactylifera*) and Zingiberales (*Musa acuminata*) in our analysis (Additional files 6, 7). We constructed a phylogenetic tree based on the alignment of 2381 orthologous protein-coding genes in the five monocotyledonous species (Fig. 3b). *D. rotundata* did not group with any species in the tree, including *Musa* of Zingiberales, *Phoenix* and *Elaeis* of Arecales, and *Oryza* and *Brachypodium* of Poales, suggesting that *Dioscorea* diversified independently from these taxa in monocotyledons.

For 12,625 *D. rotundata* genes, no orthologs or paralogs were found in *B. distachyon*, *O. sativa*, or *A. thaliana*, and 11,348 *D. rotundata* genes had no clear homologs in any of the six species shown in Fig. 3a and Additional file 8. Of these 11,348 genes without homologs, 3422 were expressed in tuber tissues, a tissue type not shared with the other species examined.

Non-redundant Gene Ontology (GO) terms "intracellular organelle", "protein binding", and "ion binding" were significantly enriched among *D. rotundata* genes that showed no orthology to the other species, but not among the conserved genes (Additional files 9, 10). *D. rotundata* genes without orthologs in the other species included 68 genes encoding proteins with lectin domains that are involved in defense against microbial pathogens, nematodes, and insects, accounting for 31% of the 216 lectin-coding genes functionally annotated in *D. rotundata*. Among the 12 subfamilies of lectins [20], the bulb-type lectin (snowdrop lectin; B-lectin) family contributed the largest share (110) of genes in *D. rotundata* (Additional file 1: Table S14). Phylogenetic analysis of the B-lectin genes in *D. rotundata* (110 genes; 51 unique), *B. distachyon*, *O. sativa*, and *A. thaliana* revealed two expansions of B-lectin genes in *Dioscorea* (Fig. 3c). The first expansion (blue band) consisted of 22 receptor-like serine/threonine-protein kinases, which are thought to play a role in signaling and the activation of plant defense mechanisms [21]. The second expansion (red band) consisted of 28 mannose-binding lectins sharing high similarity with *Dioscorea batatas* tuber lectin DB1 (accession number AB178475). DB1 has insecticidal properties against cotton bollworm (*Helicoverpa armigera*), and studies in transgenic tobacco and rice plants expressing *DB1* demonstrated that it also confers resistance against green peach aphid and brown plant hopper, respectively [22–24]. Of these mannose-binding lectin genes in Guinea yam, 16 did not have orthologs in any of the six other species examined, and two showed enriched expression (Benjamini–Hochberg [25] adjusted *P* value [padj] < 0.05) in tubers.

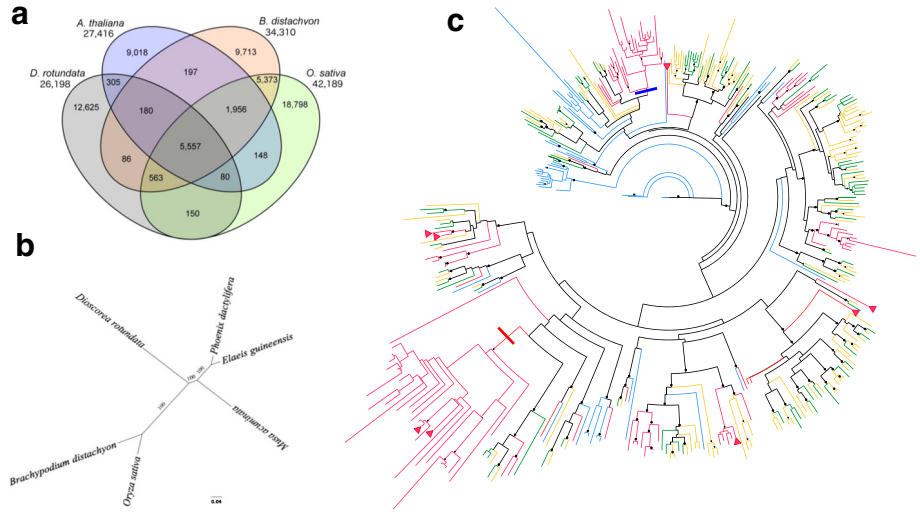

Fig. 3 Comparative genomics of *Dioscorea rotundata* and other angiosperm species. **a** Venn diagram showing conserved and unique genes at 1:1 correspondence among *D. rotundata*, *Arabidopsis thaliana*, *Brachypodium distachyon*, and *Oryza sativa*. Total gene counts in each genome are given below the species name. **b** Maximum likelihood tree of *D. rotundata*, *B. distachyon*, *O. sativa*, *Elaeis guineensis*, *Musa acuminata*, and *Phoenix dactylifera* based on 2381 orthologous protein-coding genes. The bootstrap values across 1000 resamplings are shown. The scale bar represents the mean number of substitutions per site. **c** Phylogenetic analysis of the relationships of mannose-specific bulb-type lectin proteins in *D. rotundata* (*red*), *A. thaliana* (*blue*), *B. distachyon* (*green*), and *O. sativa* (*orange*). *Arrowheads* represent bulb-type lectins observed to have enriched expression in tubers. High confidence bootstrap values (1000 replicates) are represented at the nodes of the tree as *dots*. *Thick red* and *blue lines* show two root branches of *D. rotundata*-specific expanded genes

RNA-seq analysis comparing three tuber tissues to all other nine tissues (Additional file 1: Table S12) revealed that 2023 genes were enriched in tubers. The top 50 highly expressed (padj < 0.05) genes included genes encoding starch synthases and branching enzymes, as well as three carbonic anhydrase-encoding genes. Basic Local Alignment Search Tool (BLASTP) (https://blast.ncbi.nlm.nih.gov) analysis showed that these carbonic anhydrase-encoding genes shared high identity (average 76%) with genes encoding *Dioscorea japonica* precursors of dioscorin, a tuber storage protein that has carbonic anhydrase activity and exists in multiple isoforms [26] (Additional file 11).

To infer the past genome duplication in *D. rotundata*, we performed genome-wide dot plot analysis of *D. rotundata* against itself (Additional file 2: Figure S15), which revealed no indication of genome duplication. Nevertheless, we observed 946 paralogous gene clusters composed of duplicated genes in *D. rotundata*. Of these, 145 duplicate clusters of paralogous genes were observed only in *D. rotundata*. To investigate macrosynteny between *D. rotundata* and related species, we carried out whole-genome syntenic dot plot analysis against the genomes of *Oryza sativa*, *Spirodela polyrhiza*, and *Phoenix dactylifera*. At the chromosomal level, it was difficult to observe synteny conservation between these species. To assess microsynteny conservation, we performed a syntenic path assembly [27] of the scaffolds from these species against *D. rotundata*-masked pseudo-chromosomes (see Methods). The reordering and re-orientation of the scaffolds relative to *D. rotundata* pseudo-molecules identified large proportions of the genomes to be conserved at the microsyntenic level (Additional file 2: Figure S16). This suggested that the *D. rotundata* genome has undergone many recombination events after its divergence from the other species.

Whole-genome resequencing of F1 bulk segregants identifies a genomic region associated with sex determination in *D. rotundata*

We previously developed a next generation sequencing (NGS)-based method for bulked segregant analysis (BSA) for quantitative trait locus (QTL) mapping in rice, named QTL-seq [28]. To our knowledge, this method has not been applied in species with highly heterozygous genomes. The majority of *Dioscorea* species, including *D. rotundata*, are mostly dioecious, with separate male and female plants (Fig. 4a), making it interesting to understand the genetic mechanism of sex determination in this genus. From a cross between two *D. rotundata* breeding accessions, TDr97/00917 (P3, female) and TDr97/00777 (P4, male), we generated an F1 population of 253 individuals in 2014 that segregated for male, female, monoecious (male and female flowers on the same plant), and non-flowering types (Additional file 1: Table S15). For QTL-seq analysis (see Additional file 2:

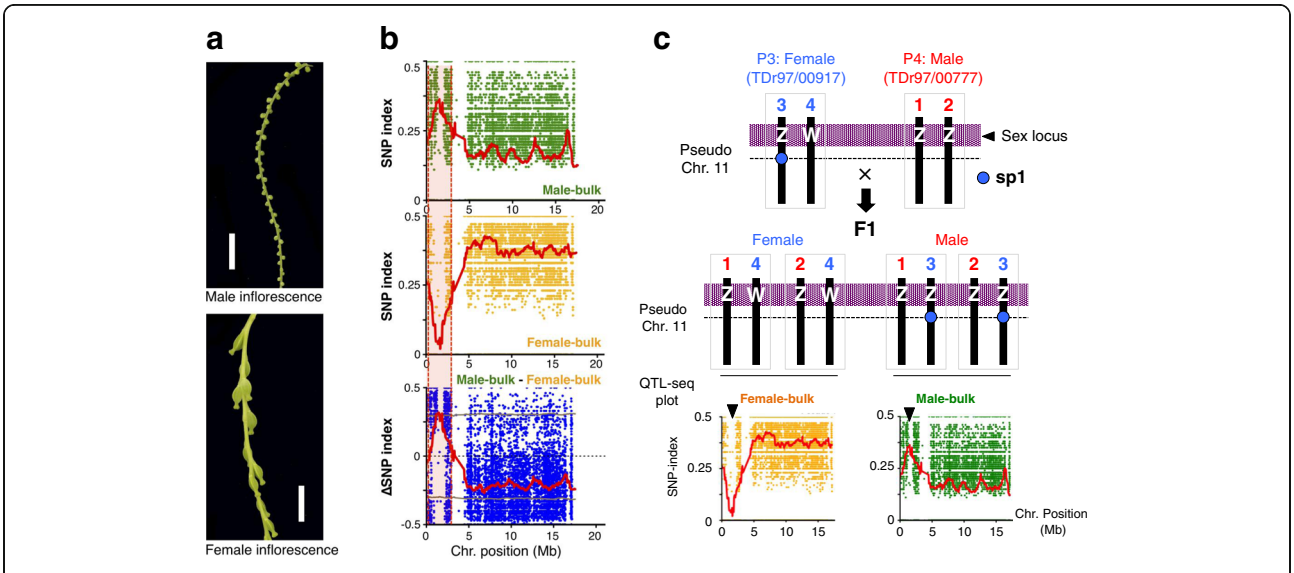

Fig. 4 QTL-seq-based analysis of sex determination in *D. rotundata*. **a** Male and female inflorescences of *D. rotundata*. Bars = 10 mm. **b** SNP-index and ΔSNP-index plots generated for pseudo-chromosome 11 (see Fig. 2). DNA samples from 50 male and 50 female F1 individuals were pooled to prepare the male and female bulks, respectively. *Green*, *yellow*, and *blue dots* represent SNP-index values at all SNP positions, and *red lines* denote the sliding window average SNP-index values at 1-Mb intervals with 50-kb increments. *Horizontal brown lines* in the ΔSNP-index plot represent the 95% confidence limit. The candidate genomic region presumably associated with sex determination is indicated by a *pink background*. **c** Schematic diagram showing the possible genotypes of female (P3, TDr97/00917) and male (P4, TDr97/00777) parents as well as their F1 progeny segregating for female and male. Genotypes of sex-determination locus are indicated as ZW or ZZ. The position of the cleaved amplified polymorphic sequence (CAPS) marker, sp1, is indicated by a *dashed line*. Sister chromatids are indicated by numbers

Figure S17 for details), we sequenced two DNA bulks representing male and female plants, each from 50 individuals, generating 7.9- and 7.3-Gb sequences, which provided 13.9× and 12.7× coverage of the predicted *D. rotundata* genome, respectively (Additional file 1: Table S16). We also resequenced the genome of the female parent (P3) and generated a P3 reference sequence (P3-Ref) by replacing TDr96_F1 nucleotides with P3 nucleotides at all different sites between the two genotypes. Likewise, we generated the male parent (P4) reference sequence (P4-Ref) by aligning P4 sequence reads to TDr96_F1 and replacing TDr96_F1 nucleotides with those of P4 at all different sites. We then separately aligned sequence reads obtained from F1 male-bulk and female-bulk DNA to the P3- and P4-Ref sequences. To identify single-nucleotide polymorphism (SNP) markers associated with the F1 gender phenotype, thus potentially suggesting candidate sex-determining gene(s), we focused on SNPs that segregated in the F1 progeny either as SNPs homozygous in the female parent (P3) but heterozygous in the male parent (P4), or vice versa. We could then identify genomic regions with SNPs heterozygous in one parent whose alleles were differentially transmitted to the two sexes in the F1, suggesting Y or W linkage, respectively. This is similar to mapping by backcrossing, but does not require a BC1 generation using inbred lines. Scanning the entire genome identified a single region, from 0.65 Mb to 2.35 Mb on pseudo-chromosome 11, whose SNP-index values (the frequency of short reads aligned to a particular position of the genome with SNPs different from the reference sequence [28]) differed for male and female bulks in the second category of SNPs just described (Fig. 4b and Additional file 2: Figure S18).

We identified a sex-linked region with category 2 SNP markers that are heterozygous in the female parent (P3) but homozygous in the male parent (P4) (Additional file 2: Figure S18), suggesting that the male sex is determined by the homozygous (designated ZZ) state of the locus responsible for sex determination, whereas that of the female sex is determined by the heterozygous (ZW) (or hemizygous: Z-) state of this locus (Fig. 4c). Genotyping of the F1 individuals used for bulk sequencing using the cleaved amplified polymorphic sequence (CAPS) marker sp1 developed within the candidate genomic region revealed significant co-segregation between the sp1 marker and the sex of the individual ($P = 1.913e\text{-}14$, Fisher's exact test). This analysis confirmed that the genomic region identified by QTL-seq is indeed associated with sex determination (Fig. 5a, b and Additional file 2: Figure S19). The switch of sp1 male and female marker genotypes in the F1 progeny occurred because the marker genotype was heterozygous in the female parent (Fig. 4c).

As the TDr96_F1 plant never flowered, we were unable to determine its sex based on flower phenotype and therefore could not directly characterize its genotype (ZZ or ZW) at the candidate sex locus. To identify the genomic regions linked to Z and W, we assembled the P3 (female) and P4 (male) genomes de novo using their PE reads with the DISCOVAR De Novo assembler [29], generating P3-DDN (female, N50 = 3.3 kb) and P4-DDN (male, N50 = 2.7 kb) reference sequences (Fig. 6a, Additional file 1: Table S17 and Additional file 2: Figure S20). We separately mapped short reads derived from bulked DNA from 50 male and 50 female F1 progeny (P3 × P4) to P3-DDN and P4-DDN and looked for unique P3-DDN (female) genomic regions (presumably corresponding to the W-linked region) that were specifically mapped by F1 female-bulk reads but not by F1 male-bulk reads. The 1345 regions (sizes from 1 to 129 bp) totaling 15,390 bp conformed to this pattern (Additional file 2: Figure S21). We ordered these fragments by size and found that the N20 value was 42 bp. Conversely, we found only 435 regions (total size

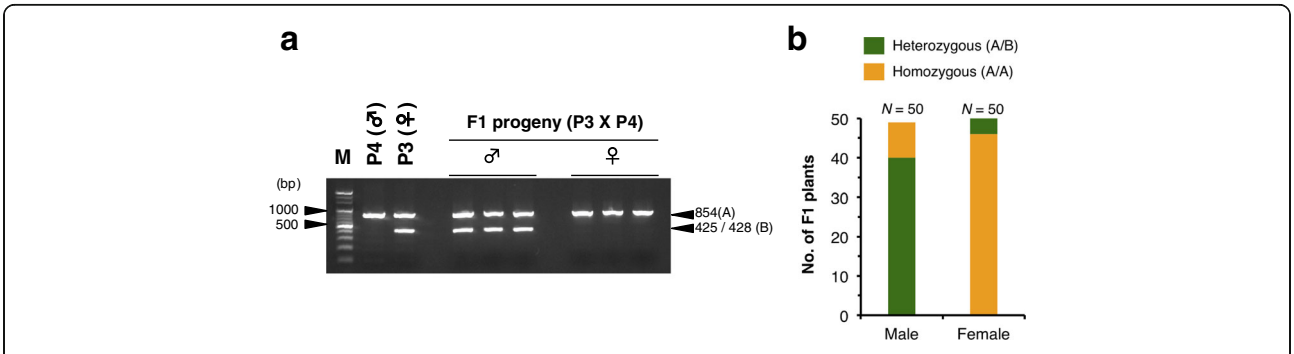

Fig. 5 A CAPS marker developed on pseudo-chromosome 11 co-segregates with sex in F1 progeny derived from a cross between female (P3) and male (P4) parents. **a** Agarose gel electrophoresis of the CAPS marker, sp1, for the parents and F1 progeny segregating for male and female phenotypes. This marker segregates for a non-cleaved band (854 bp) indicated as (A) and cleaved bands (425 bp + 428 bp) indicated as (B). **b** Frequency of the sp1 genotypes (A/B or A/A) among the F1 progeny segregating for male (50 plants) and female (50 plants). There is a statistically significant association between A/B sp1 genotype and male and between A/A sp1 genotype and female (Fisher's exact test: $P = 1.913e\text{-}14$)

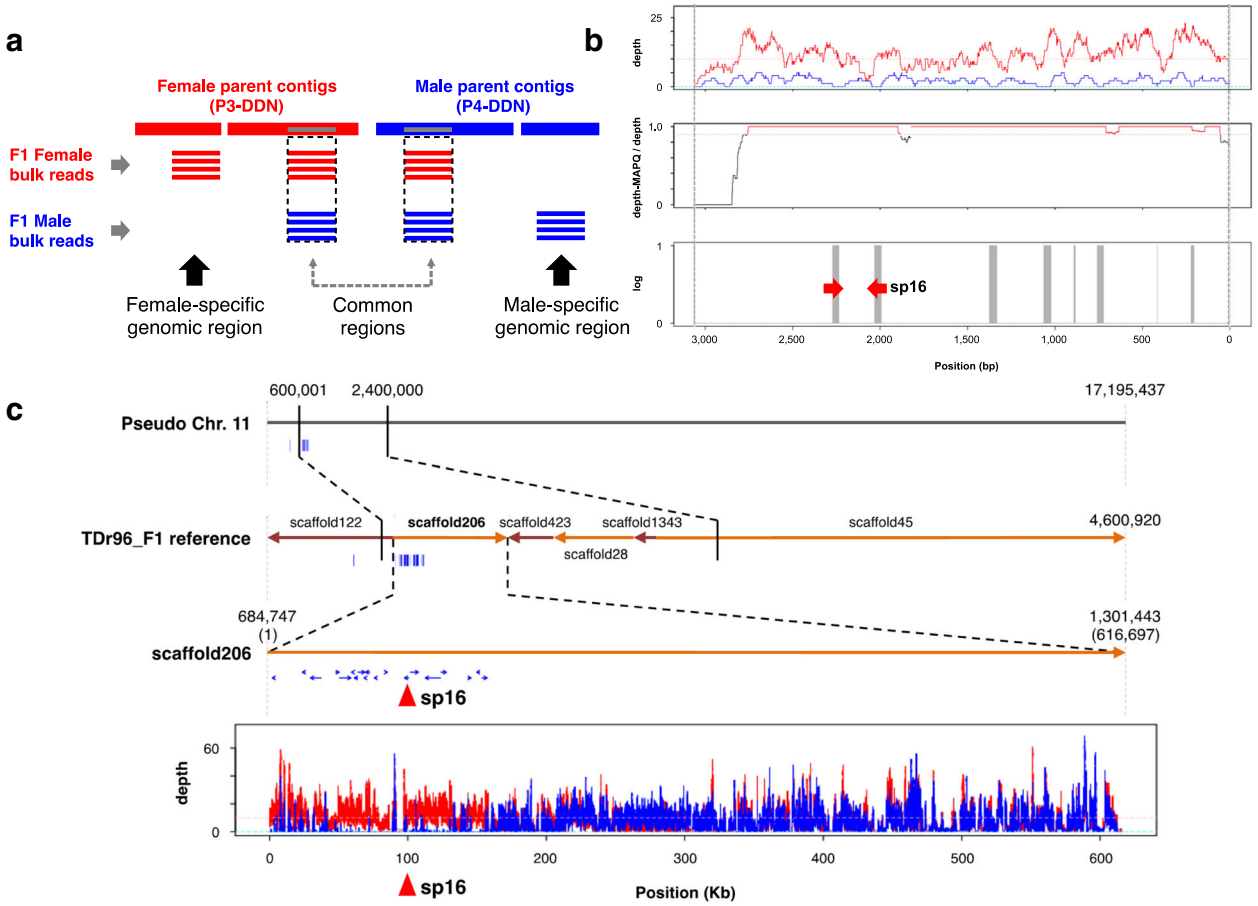

Fig. 6 Identification of female-specific putative W-linked genomic region. **a** Schematic diagram of the method used to identify the female-specific putative W-linked genomic region. De novo assembled genome sequences of female (P3-DDN) and male (P4-DDN) parents were combined to serve as a reference sequence. Short reads of bulked DNA from F1 female and F1 male progeny were separately mapped onto this combined reference sequence. The majority of reads mapped to two duplicated homologous locations in the reference genome (indicated as "common regions"), which gave low MAPQ scores (<60) in the BWA alignment. Female parental contigs that were mapped only with reads belonging to the F1 female bulk corresponded to female-specific genomic regions. Sequence reads mapped to such positions were identified by their high MAPQ scores (=60). **b** An example of a female-specific contig (contig Female917_flattened_line_87512_3057). Alignment depths of F1 female bulk (*red*) and F1 male bulk (*blue*) are shown (*top*). Frequency of reads mapped with MAPQ score = 60. The *red line* corresponds to genomic regions that were covered by short reads, > 90% of which had a MAPQ score of 60 (middle). A genomic region that is covered only by female reads (not by male reads) and > 90% of mapped reads had MAPQ score = 60 (indicated by *gray bars*) (*bottom*). *Red arrowheads* indicate the positions of PCR primers for the DNA marker sp16. **c** Location of the female-specific genomic region. *Thick gray horizontal line* denotes pseudo-chromosome 11 (*top*), scaffolds on chromosome 11 (*middle*), and scaffold206 (*bottom*). The *thin blue lines* shown under the *first*, *second*, and *third horizontal lines* indicate the positions of female contigs (P3-DDN) specifically mapped by F1 female bulk reads. The *square box* at the bottom indicates alignment depth of reads of F1 female bulk (*red*) and F1 bulk of progeny in which sp16 amplification was not observed (sp16-minus) (*blue*) to scaffold206. *Red triangles* indicate the position of DNA marker sp16

3775 bp) of P4-DDN (male) mapped by the F1-male bulk but not by the F1-female bulk (Additional file 2: Figure S21). The large size difference between female-specific P3-DDN regions (total 15,390 bp) and male-specific P4-DDN regions (total 3775 bp) suggested that the ZW female genome has additional DNA sequences not present in the ZZ male. We hypothesize that the recovery of small male-specific P4-DDN regions may have occurred by chance. We focused on 36 female-specific contigs of P3-DDN that contained DNA fragments larger than 42 bp (Fig. 6b and Additional file 2: Figure S21). When we used the 36 contigs as BLASTN queries against the TDr96_F1 reference genome, 20 were located on scaffold206 (667.8 kb) on pseudo-chromosome 11 (Fig. 6c, Additional file 1: Table S18), suggesting that P3-DDN contigs with female-specific regions were indeed located within the sex-linked region identified by QTL-seq (Fig. 4b). We developed a PCR primer pair for one such P3-DDN contig (Fig. 6b; Female917_flattened_line_87512_3057) harboring female-specific regions; we named this DNA marker sp16. sp16 amplified a PCR fragment in the P3 female parent but not in the P4 male parent (Fig. 7a), demonstrating that this fragment was located in the female-specific region. An sp16 PCR

fragment was amplified in TDr96_F1, our reference genome plant (Fig. 7a), suggesting that this individual likely had the ZW genotype. In F1 progeny derived from a P3 × P4 cross, the sp16 fragment was amplified in all female plants, but it failed to be amplified in the majority of male individuals. Furthermore, sp16 fragments were amplified in monoecious as well as non-flowering progeny (Fig. 7a). We monitored flowering in all 249 F1 individuals in two consecutive seasons (2014 and 2015) and found that 194 plants showed consistent sex phenotypes. However, the remaining 55 plants showed changes in sex among male, female, and monoecious (Fig. 7b). Genotyping of all F1 individuals using sp16 revealed a striking pattern: 121 of the 125 plants that were consistent for male over the 2 years showed no PCR amplification of sp16, whereas all plants with the remaining phenotypes showed amplification of sp16 (Fig. 7b). A similar pattern was observed in another F1 family (TDr04-219 × P4) involving the same male parent, P4 (Fig. 7c). We also assayed 24 Guinea yam breeding accessions of known sex using the same marker (Fig. 8). All 10 female accessions, as well as three accessions that did not flower, showed amplification of sp16. Of the 11 male accessions genotyped, eight did not show amplification of sp16, whereas the remaining three did.

We hypothesized that the ZZ genotype stably gives rise to the male phenotype, whereas the ZW genotype results in unstable sex phenotypes; ZW mainly generates the female phenotype, but sometimes monoecious or male phenotypes depending on the environments. Therefore, some individuals of the F1 progeny derived from a cross between P3 and P4 might have been scored as male despite their genotype being ZW, which may have obscured our analysis, resulting in non-zero depth of male DNA bulk within the putative W-region (Fig. 6b). To address this possibility, we selected 50 ZZ plants from the F1 progeny based on their sp16 genotype and bulked and sequenced the DNA (sp16-minus bulk). The sp16-minus bulk reads, as well as female bulk reads, were separately mapped to the combined sequence of the TDr96_F1 reference genome and P4-DDN to identify the female-specific TDr96_F1 genomic region, as described in Fig. 6a. As shown in Fig. 6c and Additional file 2: Figure S20c, d, we successfully delineated the putative W-linked region mapped predominantly with female-only bulk DNA, representing an approximately 161-kb region of scaffold206 on pseudo-chromosome

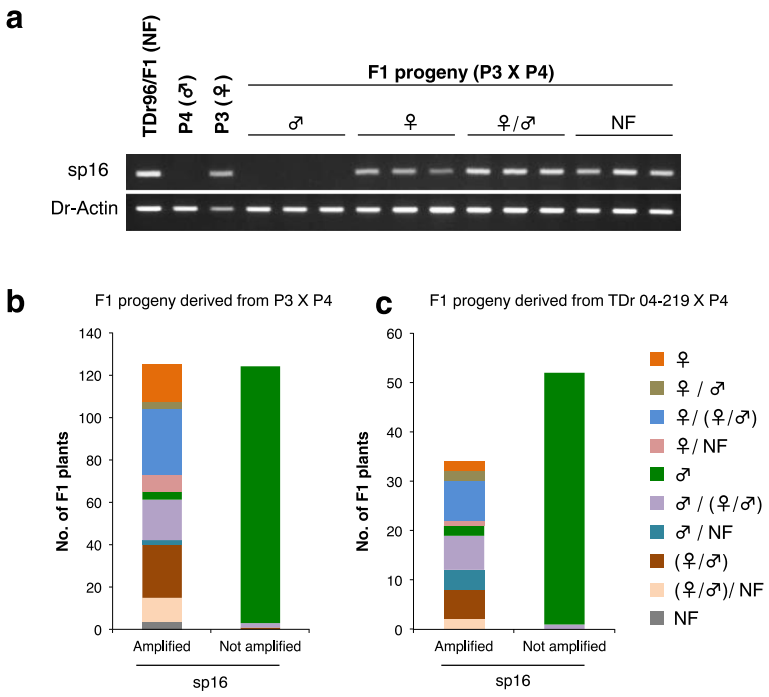

Fig. 7 DNA marker sp16 is located in a W-linked region. **a** Results of agarose gel electrophoresis of PCR products amplified by DNA marker sp16 (sp16). *Actin* from *D. rotundata* (Dr-Actin) served as a control to show that template DNA was present for all samples. *NF* non-flowering. **b** Bar graphs showing the correspondence of sp16 genotypes (sp16 PCR product Amplified or Not amplified) with the sex of F1 progeny derived from a cross between P3 and P4 and phenotyped over 2 years (2014 and 2015). Color codes indicate sex manifestation of the plants during the 2-year period, disregarding the yearly order (i.e., plants showing sex changes from male [2014] to female [2015] and female [2014] to male [2015] were combined and are indicated by ♀/♂). Monoecy is indicated by (♀/♂). *NF* non-flowering. **c** The same as **b** but for F1 progeny obtained from a separate cross involving TDr04-219 and P4

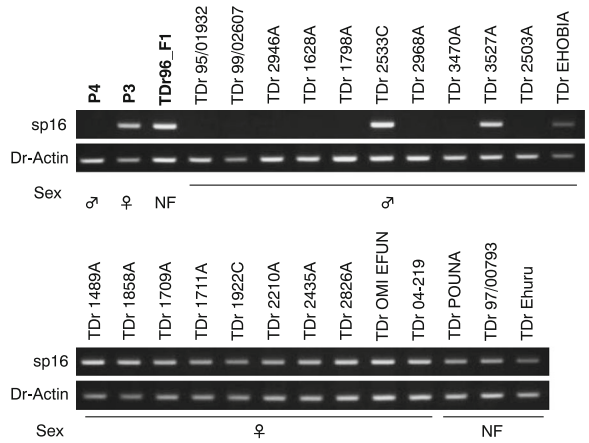

Fig. 8 Test of correspondence between sp16 genotypes and sex in 24 *D. rotundata* breeding accessions. Results of agarose gel electrophoresis of PCR products amplified using sp16 DNA markers are shown. Dr-Actin is a control indicating the presence of template DNA for all lines. *NF* non-flowering

11. This putative female-specific W-linked region contains ~ 10 predicted genes (Additional file 12).

Discussion

Molecular markers, such as simple sequence repeats (SSRs), indels, and SNPs, can, for the first time, be developed for various applications in Guinea yam, including linkage mapping, genome-wide association analysis, genomic selection, and MAS. We have already analyzed sequences containing SSR motifs in the genome and identified more than 22,000 candidates that can be used to design primers (Additional file 1: Table S19). We designed primer pairs for 1000 of these sequences and obtained the information necessary for their immediate use in genetic analyses (Additional file 13). SSR markers isolated from one *Dioscorea* species can be transferred to other species [30]. From a practical plant breeding point of view, the sp16 sex-linked marker should prove useful for selecting plantlets for crossing, substantially saving the space and labor required to grow plants and accelerating breeding programs. However, the sex-determination system may vary among *Dioscorea* species (see below), so the transferability of sex-linked DNA markers from *D. rotundata* to other species should be addressed in future studies.

Our identification of the locus underpinning an important trait by QTL-seq, using F1 progeny derived from highly heterozygous parents, opens up new avenues to WGS-based mapping of important traits in crops and tree species for which inbred lines are difficult to obtain and/or generation times are too long, impeding the use of conventional linkage analysis approaches.

Development of DNA markers linked to agronomically important traits and their use for MAS increase the role yam plays in ensuring food security for resource-poor households in Africa and beyond. The *D. rotundata* genome sequences reported here should also contribute to understanding the origin of Guinea yam and its domestication from its wild progenitor species, which are widely distributed in West and Central Africa.

Our results suggest that the Guinea yam sex-determination system involves female heterogamy (male = ZZ, female = ZW). We identified two DNA markers, sp1 (linked to the putative Z-linked region) and sp16 (presumably located within the putative W-linked region, which in TDr96_F1 is presumed to be ZW, and spans only 161 kb). The chromosomes carrying the Z- and W-linked regions are probably not strongly differentiated, and diverged sequences corresponding to Z and W chromosomes were not recovered in our reference genome. Future work should test for structural differences, such as inversions, between the Z- and W-linked regions. Guinea yam sex determination is not, however, a simple genetic system. The consistent maleness of individuals with the ZZ genotype, based on the sp16 sequence, versus occasional maleness of ZW individuals, suggests that maleness is the default phenotype and that the W allele is dominant over Z and can, but does not always, suppress male organ development and feminize the flower. If the feminizing function of the W allele fails in a subset of flowers, the individual will be monoecious. ZW individuals can change sex over time (Fig. 7), indicating that the Z-suppressing function can be affected by the environment. Self-pollination between male and female flowers of ZW monoecious plants could become possible, which may allow inbred lines to be generated, allowing fixation of desired alleles of agronomically important traits. To make it practical, though, we may have to carefully monitor the level of inbreeding depression in *D. rotundata*. Dioecy is the norm in *Dioscorea* species, and previous reports suggest that males are usually the heterogametic sex (XY) in the genus [31, 32]. A genetic study of *D. tokoro* also confirmed an XY male system [19]. *D. tokoro* belongs to the section Stenophora, which is distantly related to the section Enantiophyllum, which contains *D. rotundata* [3]. Our data suggest that the sex-determination system has changed within the genus during the evolution, which could be an interesting topic for future studies. Once the *D. rotundata* sex-determination gene has been isolated, its comparison with another dioecious monocot species such as *Asparagus*, for which the sex-determination gene has been recently isolated [33], would be interesting.

Conclusions

Here, we sequenced the whole genome (594 Mb) of the dioecious tuber crop Guinea yam (*Dioscorea rotundata*) using a heterozygous individual and anchored the scaffolds to 21 linkage groups to generate pseudo-chromosomes. We exploited the genome sequence to map the sex-determination locus by QTL-seq using BSA of F1 progeny. This analysis revealed a genomic region on pseudo-chromosome 11 tightly linked to femaleness within a female heterogametic (ZZ = male, ZW = female) sex-determination system. This genome sequence will serve as a springboard towards gene mapping and discovery in yam (*Dioscorea* spp.) and genetic improvement of these important yet neglected staple crops.

Methods

Plant materials

The TDr96_F1 line used for WGS was selected from F1 progeny obtained from an open-pollinated *D. rotundata* breeding line (TDr96/00629) grown under field conditions in the experimental fields of the International Institute of Tropical Agriculture (IITA) in Nigeria. F1 seeds from TDr96/00629 and those obtained from the cross between the parental lines TDr97/00917 and TDr99/02627 used for RAD-seq were germinated on wet paper towels in darkness at 28 °C. After germination, the seeds were transferred to soil (Sakata Supermix A [34]) and grown at 30 °C with a 16-h/8-h photoperiod in a greenhouse at Iwate Biotechnology Research Institute (IBRC) in Japan. Fresh leaf samples were collected for DNA extraction. Additionally, to resequence the F1 progeny used for QTL-seq analysis, lyophilized leaf samples obtained from plants that were grown and phenotyped under field conditions at IITA were used for DNA extraction.

Determination of chromosome number and ploidy level

For chromosome observation, root tips of TDr96_F1 plants generated by in vitro propagation of nodal explants were sampled and fixed in acetic acid-alcohol (1:3 ratio) for 24 h without pretreatment. Fixed root tips were stained with a 1% aceto-carmine solution for 24 h. Samples were prepared by the squash method and analyzed under an Olympus BX50 optical microscope (Olympus Optical Co, Ltd., Tokyo, Japan [35]) at 400× magnification.

Estimation of *D. rotundata* genome size

The genome size of TDr96_F1 (*D. rotundata*) was estimated both by FCM and *k*-mer analyses. FCM analysis was carried out using nuclei prepared from fresh leaf samples of TDr96_F1 and a *japonica* rice (*Oryza sativa* L.) cultivar of known genome size (~380 Mb [36]), which served as an internal reference standard. Nuclei were isolated and stained with propidium iodide (PI) simultaneously and analyzed using a Cell Lab Quanta™ SC Flow Cytometer (Beckman Coulter, CA [37]) following the manufacturer's protocol. The ratio of G1 peak means [yam (281.7):rice (188.7) = 1.493] was used to estimate the genome size of *D. rotundata* to be ~ 570 Mb (380 Mb × 1.5). *k*-mer analysis-based genome size estimation [10] was performed with TDr96_F1 PE reads with an average size of ~ 230 bp and a total length of 16.77 Gb (16,771,579,510 bp) using ALLPATHS-LG [11]. *k*-mer frequency analysis, with the *k*-mer size set to 25, generated values for *k*-mer coverage (Kc = 25.66) and mean read length (Rl = 228.8), which were used to estimate the genome size of TDr96_F1 to 579 Mb as follows:

$$\begin{aligned}\text{Genome Size} &= \text{Total PE read length (bp)} \\ &\div \text{Read coverage (Rc)} \\ \text{Read coverage (Rc)} &= [k\text{–}mer \text{ coverage (Kc)} \times \text{Rl}] \\ &\div [\text{Rl} - k\text{-}mer \text{ length (k)} + 1]\end{aligned}$$

Whole-genome sequencing

For WGS, genomic DNA was extracted from fresh TDr96_F1 leaf samples using a NucleoSpin Plant II Kit according to the manufacturer's protocol (Macherey-Nagel GmbH & Co. KG [38]) with slight modifications. Homogenized samples were washed with 0.1 M 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES) buffer to remove contaminating polysaccharides. Just before use, 120 mg polyvinylpyrrolidone (PVP), 90 mg L-ascorbic acid, and 200 μl 2-mercaptoethanol (ME) were added to 10 ml HEPES buffer, and 1 ml of the mixture was used to wash each sample; washing was repeated three times. Additionally, 10 μl 2-ME and 5 μl of 30% polyethylene glycol (PEG)-20000 were added to 1 ml of PL1 buffer (provided with the NucleoSpin Plant II Kit), and twice the recommended volume of buffer (800 μl) was used for cell lysis. Libraries for PE short reads and MP jump reads of various insert sizes including 2, 3, 4, 5, 6, and 8 kb were constructed using an Illumina TruSeq DNA LT Sample Prep Kit and a Nextera Mate Pair Sample Prep Kit, respectively. The PE library was sequenced on the Illumina MiSeq platform, while the MP libraries were sequenced on the HiSeq 2500 platform. Library construction and sequencing of the 20- and 40-kb MP jump sequences were carried out by Eurofins Genomics (Operon [39]) and Lucigen [40], respectively. The 20-kb and 40-kb jump libraries were sequenced on the MiSeq and HiSeq 2500 platforms, respectively. BAC libraries were constructed by Lucigen, and BAC-end sequencing was carried out by Genaris [41] using Sanger sequencing. A total of 30,750 clones corresponding to

3072 Mb of sequence and 5.4× genome coverage were constructed. Of these, 9984 clones were used for BAC-end sequencing, generating a 13.6-Mb sequence in PE fasta format, which was converted to 50-bp PE short reads corresponding to a 0.46-Gb sequence and ~ 0.8× coverage of the estimated 470-Mb *D. rotundata* genome (Additional file 1: Table S2 and Additional file 2: Figure S2).

De novo assembly

All TDr96_F1 sequence reads in fastq format were filtered for quality using the FASTX-Toolkit version 0.0.13 [42]. For PE reads and MP short jump reads with insert sizes ranging from 2 to 8 kb, only those having sequence reads with a Phred quality score of ≥ 30 (i.e., ≥ 90% of the reads) were retained. Adapter trimming and removal of MP reads with the wrong insert sizes were performed using an in-house pipeline of scripts written in Perl and C++. Quality filtering of the long jump sequences (20-, 40-, and 100-kb insert sizes) was carried out by the suppliers. For the initial de novo assembly, short PE reads and MP jump reads with 2- to 40-kb insert sizes (Additional file 1: Table S2) were assembled using ALLPATHS-LG assembler version R49856 [11]. Further scaffolding of the assembly generated by ALLPATHS-LG was performed using the 100-kb jump MP fastq reads obtained by BAC-end sequencing and the SSPACE PREMIUM 2.3 scaffolding tool with default parameters [13].

Constructing organelle genome sequences

De novo assembly of the *D. rotundata* mitochondrial genome sequence was performed using mitochondrial DNA isolated from TDr96_F1 leaf samples according to the method of Terachi and Tsunewaki [43] with the following minor modifications. Fresh green leaves (ca. 150 g) were homogenized in 1.5 L of homogenization buffer containing 0.44 M mannitol, 50 mM Tris-HCl (pH 8.0), 3 mM ethylenediaminetetraacetic acid (EDTA), 5 mM 2-ME, 0.1% (w/v) bovine serum albumin, and 0.1% (w/v) PVP. Following DNaseI treatment, the mitochondrial fraction was collected from the interface between 1.30 M and 1.45 M of a sucrose gradient. Mitochondrial DNA was purified by EtBr/CsCl centrifugation at 80,000 rpm for 6 h at 20 °C in a Beckman TLA 100.3 rotor. The DNA band was collected and purified by ethanol precipitation. The resulting mitochondrial DNA (15 ng) was amplified using a REPLI-g Mini Kit (Qiagen, Cat. no. 150023) and used for library construction. The library was sequenced on an Illumina MiSeq sequencer, and the resulting PE reads were assembled de novo using DISCOVAR De Novo [29], generating *D. rotundata* mitochondria contigs. For scaffolding, MP reads with insert sizes of 2, 3, 4, 5, 6, 8, and 20 kb obtained from *D. rotundata* genomic DNA (gDNA) were aligned to the *D. rotundata* mitochondrial contigs. MP reads showing 100% alignment were selected and used for scaffolding of *D. rotundata* mitochondrial contigs by SSPACE [13] (Additional file 2: Figure S5). To reconstruct the *D. rotundata* chloroplast genome sequence, the PE reads of TDr96_F1 were aligned to the recently published *D. rotundata* chloroplast genome sequence [16] (GenBank ID = NC_024170.1) by Burrows-Wheeler alignment (BWA) [44], and chloroplast-derived sequences were identified, amounting to 5,403,420 reads (14.74% of the total size of PE reads generated for TDr96_F1 [Table 1]) matching the assembled 155.4-kb chloroplast genome of *D. rotundata*.

Evaluation of the completeness of the genomic assembly

To evaluate the completeness of the *D. rotundata* genome assembly, the assembly was checked for the presence of 248 highly conserved core eukaryotic genes [45] using CEGMA version 2.4 with default parameters [14] (Additional file 1: Table S4). To further assess the completeness of the genome, the successor to CEGMA, Benchmarking Universal Single-Copy Orthologs (BUSCO), was used to check for the presence of 956 BUSCOs with version 1.1.b1 [15] using the early access plant dataset (Additional file 1: Table S5).

Annotation of transposable elements (TEs)

Legacy repetitive sequences, including transposons, were predicted using CENSOR 4.2.29 [46] with the following options: show_simple, nofilter, and mode rough using the Munich Information Center for Protein Sequences (MIPS) Repeat Element Database [47]. Following identification, the repeat elements were classified using mipsREcat [47]. Repetitive sequences were later improved by remodeling using RepeatModeler 1.0.8 [48] and masked with RepeatMasker 4.0.5 [49]. Using the National Center for Biotechnology Information (NCBI) database, one of three other options was used to generate interspersed RepeatModeler-based, interspersed Rebase-based, and Low complexity repeats: "nolow", "nolow, species Viridiplantae", and "noint", respectively. Repeat element content and other statistics were compared between the *D. rotundata* and *A. thaliana* TAIR10 [50], *B. distachyon* v3.1 [51], and *O. sativa* v7_JGI 323 [52] genomes using the RepeatModeled and RepeatMasked references (Table 1).

RNA-seq

Total RNA was extracted using leaf, stem, flower, and tuber samples collected from a greenhouse-grown TDr96_F1 plant using a Plant RNeasy Kit (Qiagen [53]) with slight modifications. RLC buffer was used for lysis after the addition of 5 μl 30% PEG-20000 and 10 μl 2-ME to 1 ml of buffer. The RNA samples were treated

with DNase (Qiagen) to remove contaminating genomic DNA. Two micrograms of total RNA was used to construct complementary DNA (cDNA) libraries using a TruSeq RNA Sample Prep Kit V2 (Illumina) according to the manufacturer's instructions. The libraries were used for PE sequencing using 2× 100 cycles on the HiSeq 2500 platform in high-output mode. Illumina sequencing reads were filtered by Phred quality score, and reads with a quality score of ≥ 30 (≥90% of reads) were retained (Additional file 1: Table S12). Only one RNA-seq experiment was carried out per tissue/organ (indicated as sample in Additional file 1: Table S12).

Prediction of protein-coding genes

The legacy gene models were generated previously using the legacy repeat-masked reference genome and three approaches: ab initio, ab initio supported by evidence-based prediction, and evidence-based prediction. The ab initio prediction was carried out with FGENESH 3.1.1 [54]. The ab initio supported by evidence-based prediction was performed with AUGUSTUS 3.0.3 [55] using the maize5 training set and a hint file as the gene model support information. To construct the hint file, TopHat 2.0.11 [56] was used to align RNA-seq reads from tuber, flower (young), leaf (young), stem, leaf (old), and flower (old) samples to the *D. rotundata* reference genome, and Cufflinks 2.2.1 [57] was used to generate gene models from these data. The evidence-based predictions using the Program to Assemble Spliced Alignments (PASA) [58] were generated in a Trinity [59] assembled transcriptome from the RNA-seq data. JIGSAW 3.2.9 [60] was used to select and combine the gene models obtained using the three approaches with the weighting values assigned to the results from FGENESH, AUGUSTUS, and PASA of 10, 3, and 3, respectively. In total, 21,882 consensus gene models were predicted. These gene models were further improved upon using the MAKER [61] pipeline (Additional file 2: Figure S14). Publicly available ESTs and protein sequences from related plant species were aligned to the genome using GMAP [62] and Exonerate 2.2.0 [63], respectively. De novo and reference-guided transcripts were assembled from RNA-seq data from all 18 tissues using Bowtie 1.1.1 [64], Trinity 2.0.6 and SAMtools 1.2.0 [65], and Trinity 2.0.6 and TopHat 2.1.0, respectively. Both sets of assembled transcripts were used to build a comprehensive transcript database using PASA (Additional file 1: Table S13). High-quality non-redundant transcripts from PASA were used to generate a training set for AUGUSTUS 3.1. Gene models were predicted twice using the genome, improved repeat sequences, assembled transcripts, EST and protein alignments, the AUGUSTUS training set, and a legacy set of 21,882 gene models obtained previously using MAKER 2.31.6 [61], retaining all legacy gene models or querying them with new evidence and discarding those that could not be validated. From both MAKER runs, 21,894 and 76,449 gene models were predicted, respectively. A consensus set of gene models from both MAKER outputs was obtained using JIGSAW 3.2.9 [60] at a 1:1 ratio. In total, 26,198 consensus gene models were predicted in the *D. rotundata* genome. The corresponding amino acid sequences were also predicted for these gene models. To confirm these gene models, the RNA-seq reads were aligned to the CDSs (coding sequences) of the predicted genes using BWA [44] with default parameters. Accordingly, 85.8% of the gene models could be aligned by at least a single RNA-seq read. Functional annotation of the amino acid sequences was performed using the in-house pipeline, AnnotF, which compares Blast2GO [66] and InterProScan [67] functional terms.

Comparative genomics

Pairwise orthology relationships were determined with Inparanoid [68–70] using the longest protein-coding isoform for each gene in *Arabidopsis thaliana* (TAIR10) [50], *Oryza sativa japonica* (v7.0) [52], *Brachypodium distachyon* (v3.1) [71], *Musa acuminata* (v2) [72], *Elaeis guineensis* (EG5) [73], and *Phoenix dactylifera* (DPV01) [74]. Orthology clusters across all seven species were determined using Multiparanoid [75]. Sequences for the 12 classes of lectins were obtained from UniProt [76] for the proteomes of *A. thaliana* (up000006548), *B. distachyon* (up000008810), and *O. sativa* (up000059680). Protein alignments for B-lectin class protein sequences from all three of these species and *D. rotundata* were generated using the program Multiple Alignment using Fast Fourier Transform (MAFFT) [77]. Maximum likelihood trees were constructed based on the concatenated alignments of all 378 B-lectin proteins using RAxML [78] 8.0.2 with 1000 bootstraps. Enrichment of tuber-specific genes was detected using TopHat 2.1.0 to align RNA-seq data from each of the 12 tissues to the genome, with one biological replicate for each tissue. HTSeq 0.6.1 [79] was used to generate raw counts. Then the Bioconductor package DESeq2 1.14.1 [80] was used to compare raw counts of the three tuber tissues against all the other nine tissues (Additional file 1: Table S12) to determine tuber-enriched gene expression based on a log2 fold change > 0 and Benjamini–Hochberg [25] adjusted P value < 0.05.

Gene enrichment analysis of orthology clusters was performed with GOATOOLS [81], using the Holm significance test, and the false discovery rate was adjusted using the Benjamini–Hochberg procedure [25]. The list of enriched genes was filtered for redundant Gene

Ontology (GO) terms using REVIGO [82]. For the species phylogeny, protein alignments for each gene with a 1:1 orthologous relationship across all monocot species were generated with MAFFT using the longest protein isoform. Maximum likelihood trees were constructed based on the concatenated alignments of 2381 orthologous protein-coding genes using RAxML 8.2.8 [78] with a JTT + Γ model and 1000 bootstraps.

SynMAP [83] using BLASTZ [84] alignments, DAGchainer [85] (options -D 30 and -A 2), and no merging of syntenic blocks were used as part of the CoGe platform [86] to identify syntenic blocks between the hard-masked pseudo-chromosomes of *D. rotundata* and scaffolds/contigs of *Oryza sativa japonica* (A123v1.0), *Spirodela polyrhiza* (v0.01), and *Phoenix dactylifera* L. (v3). A syntenic path assembly was then carried out on each of the same three species in SynMap using synteny between the scaffolds/contigs against *D. rotundata* pseudo-molecules. The syntenic path assembly is a reference-guided assembly that uses the synteny between two species to order and orientate contigs. This approach highlights regions of conservation that were otherwise too shuffled to be clearly observed. Self-self synteny analysis of *D. rotundata* pseudo-chromosomes was carried out using SynMap Last alignments with default parameters and syntenic gene pair synonymous rate change calculated by CodeML [87].

RAD-based linkage mapping and scaffold anchoring

RAD-seq was performed as previously described [88] with a minor modification. Genomic DNA was digested with the restriction enzymes *Pac*I and *Nla*III to prepare libraries used to generate PE reads by Illumina HiSeq 2500 (Additional file 2: Figure S6). Approximately 822.7-Mb and 250.4-Mb sequence reads covering 22.9% and 5.3% of the estimated 504-Mb *D. rotundata* genome sequence, excluding gap regions, at average depths of 7.2× and 9.8× were generated for the parental lines and F1 individuals, respectively (Additional file 2: Figure S7).

Library preparation and sequencing

For library construction, 1 μg DNA obtained from the two parental lines (TDr97/00917-P1 and TDr99/02627-P2) and the 150 F1 individuals was digested with *Pac*I, which recognizes 5'-TTAATTAA-3', and a biotinylated adapter-1 was ligated to the digested DNA fragments. The adapter-1-ligated DNA fragments were digested with a second enzyme, *Nla*III (5'-CATG-3'). After collecting the biotinylated fragments using streptavidin-coated magnetic beads, adapter-2 was ligated to the *Nla*III-digested ends. The adapter-ligated DNA was amplified using primers containing sample-specific index sequences, adapter-1 (F) and adapter-2 (R) sequences, and sequences corresponding to the P7 and P5 primers for Illumina sequencing library preparation (Additional file 2: Figure S6). The PCR products were pooled in equal proportions, purified, and subjected to PE sequencing on the Illumina HiSeq 2500 platform. Detailed information about the primers (P7 and P5) used for Illumina library preparation are given in Additional file 1: Table S20.

Identification of parental line-specific heterozygous markers

RAD-tags were aligned to the *D. rotundata* reference genome using BWA [44]. The aligned data were converted to SAM/BAM files using SAMtools [65], and the RAD-tags with mapping quality < 60 or containing insertions/deletions in the alignment data were excluded from analysis. Low mapping positions including those with only a single RAD-tag and a mapping quality score of < 30 were also excluded. SNP-index values [28] were calculated at all SNP positions. For linkage mapping, two types of heterozygous markers (SNP-type and presence/absence-type) were identified (Additional file 2: Figure S8). The SNP-type heterozygous markers were defined based on SNP-index patterns of the parental line RAD-tags. For example, positions with SNP-index values ranging from 0.2 to 0.8 in P1 but homozygous in P2 with SNP-index values of either 0 or 1 were defined as P1-specific heterozygous SNPs. A similar procedure was followed to identify P2-specific heterozygous SNP markers. The selected markers were filtered using depth information at all positions. To increase the accuracy of the selected markers, their segregation (1:1 ratio) was confirmed in 150 F1 individuals obtained from a cross between P1 and P2. If the segregation ratio was out of the confidence interval ($P < 0.05$) hypothesized by the binomial distribution, $B(n = \text{number of individuals}, P = 0.5)$, the markers were excluded from further analysis. Only one marker was selected per 10-kb interval based on the number of F1 individuals represented and tag coverage. A total of 1105 and 990 P1- and P2-heterozygous SNP markers were selected, respectively (Additional file 1: Table S7).

The presence/absence-type markers were defined based on the alignment depth of parental line RAD-tags. First, genomic positions that could be aligned by RAD-tags from only one of the parental lines were identified. Additionally, aligned tags should be heterozygous for that particular region. Similar to the SNP-type markers, the segregation patterns of candidate presence/absence-type markers in the F1 progeny were confirmed, and only those that segregated at a 1:1 ratio (as confirmed by binomial distribution filter) were retained. In the F1 progeny, positions with sequencing depths of ≥ 3 and 0 were defined as heterozygous and homozygous, respectively.

For a given candidate position/marker, if the number of F1 individuals defined as homozygous or heterozygous was less than 120, the marker was excluded from further analysis. Only one heterozygous position was selected as a marker within a given 10-kb interval. In total, 221 and 282 positions were selected as P1- and P2-specific presence/absence-type heterozygous markers, respectively (Additional file 1: Table S7).

Linkage mapping

To developing parental line-specific linkage maps, P1-Map and P2-Map, recombination fraction (*rf*) values between all pairs of markers on a given scaffold were calculated for both parents using the recombination pattern of the 150 F1 individuals. To minimize incorrect mapping, scaffolds were divided at positions where *rf* values exceeded 0.25 from the initial marker position (Additional file 2: Figure S9). Only two flanking (distal) markers per scaffold were selected, corresponding to 477 and 493 P1- and P2-specific markers, respectively. These markers were used to develop P1 and P2 linkage maps according to the pseudo-testcross method [18] using the backcross model of R/qtl [89]. Due to the use of the pseudo-testcross method, the initial maps contained both the coupling and repulsion-type markers. Consequently, the genetic distance in linkage groups was larger than expected. To avoid the effect of repulsion-type markers when calculating genetic distances, these markers were converted to coupling-type markers. If a marker showed a high logarithm of odds (LOD) score and an *rf* value > 0.5, it was defined as repulsion type and was therefore converted to the coupling-type genotype. This conversion was carried out gradually by changing the threshold of the LOD score from 10 to 5, and then to 3. After converting all repulsion markers to coupling markers, linkage maps were developed using markers showing LOD score > 3 and *rf* value < 0.25. Accordingly, a total of 21 and 23 linkage groups, each with a minimum of three markers, were generated for P1- and P2-Maps, respectively (Additional file 1: Table S8 and Additional file 2: Figure S10).

Anchoring scaffolds

To develop chromosome-scale pseudo-molecules, TDr96_F1 scaffold sequences were anchored onto the two parental-specific linkage maps using the selected RAD markers. To combine the two maps, the number of scaffolds shared between all possible linkage group (LG) pairs corresponding to the two maps was determined (Additional file 2: Figures S11, S12). LG pairs that shared the largest number of scaffolds were combined using the same scaffolds. Each combined LG represented a pseudo-chromosome, which was designated/numbered according to the P1-Map LG designation (see Fig. 2 and Additional file 2: Figure S11). After combining the two maps to construct the pseudo-chromosomes, P1- and P2-specifc scaffolds were ordered according to their original order in their respective LGs. If the order of scaffolds could not be decided because the order was similar in both the P1- and P2-Maps, the order in P1-LG was adopted (Fig. 2). Finally, the ordered scaffolds were connected by 1000 nucleotides of "N" into a single fasta file for each pseudo-chromosome (Additional file 2: Figure S12).

QTL-seq analysis

DNA samples obtained from the two parental lines, TDr97/00917 (P3, female) and TDr97/00777 (P4, male), as well as samples pooled in equal amounts from 50 male (male-bulk) and 50 female (female-bulk) F1 individuals obtained from the cross between P3 and P4 were subjected to WGS. Libraries for sequencing were constructed from 1-μg DNA samples with a TruSeq DNA PCR-Free LT Sample Preparation Kit (Illumina) and were sequenced via 76 cycles on the Illumina NextSeq 500 platform. Short reads in which more than 20% of sequenced nucleotides exhibited a Phred quality score of < 20 were excluded from further analysis. To perform QTL-seq analysis of F1 progeny, two types of analyses are required. In the first analysis, the SNP index and ΔSNP index are calculated at P4-specific heterozygous positions. The second analysis is performed using P3-specific heterozygous positions. To identify P4-specific heterozygous positions, the P3 "reference sequence" was first developed by aligning P3 reads to the reference genome sequence of *D. rotundata* and replacing nucleotides of the *D. rotundata* reference genome sequence with nucleotides of P3 at all SNP positions showing an SNP index of 1 (Additional file 2: Figure S17c). SNP detection, calculation of SNP index, and replacement of SNPs were carried out via step 2 of QTL-seq pipeline version 1.4.4 [90]. Short reads obtained from both the male and female parents were then aligned to the "reference sequence" and heterozygous SNP positions between the two were extracted. A SNP was defined as heterozygous if the same position showed an SNP-index value ranging from 0.4 to 0.6 in one parent and a value of 0 in the second parent. Of the selected markers/positions, only those having enough depth in both parents (>15) were used for analysis of SNP-index values in the bulk-sequenced samples. P3-specific heterozygous positions were identified similarly using the P4 "reference sequence."

After identifying P4- and P3-specific heterozygous positions, the Illumina reads from the two bulk-sequenced samples (male and female bulks) were aligned to the reference sequences using BWA [44] and subjected to Coval filtering [91] as previously described. When the

P3 reference sequence was used for alignment, the SNP-index values were calculated only at all of the P4-specific heterozygous positions. By contrast, when the P4 reference sequence was used for alignment, the SNP-index values were calculated only at the P3-specific heterozygous positions. In both cases, positions with shallow depth (< 6) in either of the two samples were excluded from analysis. The ΔSNP index was calculated by subtracting the SNP-index values of the male bulk from those of the female bulk. To generate confidence intervals of the SNP-index value, an in silico test simulating the application of QTL-seq to DNA bulked from 50 randomly selected F1 individuals was performed as described previously [28] (Additional file 2: Figure S22). The simulation test was repeated 10,000 times depending on the alignment depth of short reads to generate confidence intervals. These intervals were plotted for all SNP positions analyzed. Finally, sliding window analysis was applied to SNP-index, ΔSNP-index, and confidence interval plots with a 1-Mb window size and a 50-kb increment to generate SNP-index graphs (Additional file 2: Figure S18).

Identification of putative W-region by de novo assembly of female and male parental genomes and mapping of bulked DNA from female and male F1 progeny

DNA samples obtained from the two parental lines, TDr97/00917 (P3, female) and TDr97/00777 (P4, male), were separately subjected to de novo assembly. Libraries for sequencing were prepared with a TruSeq DNA PCR-Free LT Sample Preparation Kit (Illumina) and were sequenced for 251 cycles on the Illumina MiSeq platform. Contigs were generated using the DISCOVAR De Novo assembler [29], resulting in P3-DDN and P4-DDN, respectively. Separately, whole-genome resequencing of bulked DNA was performed on bulked DNA samples obtained from 50 female F1 (Female-bulk.fastq) and 50 male F1 (Male-bulk.fastq) progeny, all derived from a cross between P3 and P4. Two reference sequences, P3-DDN and P4-DDN, were combined to generate P3-DDN/P4-DDN. Short reads from the female and male bulks were separately mapped to P3-DDN/P4-DDN using the alignment software BWA [44]. After mapping, the MAPQ scores of the aligned reads were obtained. Under our conditions, if a short read was mapped to a unique position of the reference sequence, the MAPQ score was 60, whereas if the read was mapped to multiple positions, MAPQ was < 60. Since two reference sequences (P3-DDN and P4-DDN) were fused to generate P3-DDN/P4-DDN, most genomic regions were represented twice. Therefore, most short reads mapped to two or more positions, leading to a MAPQ score < 60. The reads that mapped to the P3-DDN/P4-DDN with MAPQ = 60 were judged to be located in either P3- or P4-specific genomic regions. After finding these P3- or P4-specific genomic regions, the depth of short reads that covered the regions for Female-bulk.fastq and Male-bulk.fastq, respectively, was evaluated. If the depth of Female-bulk.fastq was high and the depth of Male-bulk.fastq was 0 or close to 0, such genomic regions were retained as putative W-regions (Fig. 6 and Additional file 2: Figure S20).

DNA markers linked to sex

The primer sequences used for amplification of sex-linked markers sp1 and sp16, as well as the control *Actin* gene fragment (Dr-Actin), were as follows:

PCR primers for sp1 fragment:

sp1-F; 5'-GATCTGGCTTCCTCCATCTTG-3'
sp1-R; 5'-GCTTGGGTGGTTAGTTTATTGTTTG-3'
PCR primers for sp16 fragment:
sp16-F; 5'-AATGTGTTTAACAGGGTGAATTC-3'
sp16-R; 5'-GAATTCAGCCGAATATACTTATTC-3'
PCR primers for Dr-Actin gene fragment:
Dr-Actin-F; 5'-CAGGGAAAAGATGACCCAAATC-3'
Dr-Actin-R; 5'-CCATCACCAGAATCCAGCAC-3'

PCR was performed using the following conditions: 30 cycles of 98 °C for 10 s, 55 °C for 30 s, and 72 °C for 1 min. For CAPS analysis of the sp1 marker, the amplified DNA was digested with *Eco*RI. All PCR products were electrophoresed on 1.5% agarose gels.

Identification of SSR markers

Approximately 4,932,582 bp simple sequence repeat (SSR) motif-containing sequences were predicted in the *D. rotundata* genome. Within this region, SSR sequences with enough flanking regions were identified and evaluated for use in primer design. Accordingly, 134,101 SSR-containing sequences, excluding those with single base repeats, were identified. The SSR information for these sequences was analyzed using GMATo [92] version 1.2 Build 20130106 with the following parameters: m (minimum motif length) = 2, x (maximum motif length) = 10, and r (minimum repeat number) = 10. The necessary information was obtained for 22,164 SSR-containing sequences in the assembled genome, 12,724 (57.4%) of which were anchored to the genetic map (Additional file 1: Table S19). Primer pairs were designed for 1000 of these sequences using Primer3 [93] software release 2.3.6 with the following parameters: product size = 100–500, primer length = 18–22 bp (optimum 20 bp), GC content = 40–60% (optimum 50%), and Tm = 57–63 (optimum = 60.0).

Additional files

Additional file 1: Tables S1–S20. Supplementary data including world yam production statistics [94] (**Table S1**), summary of genome sequence reads (**Table S2**), summary of genome assembly (**Table S3**), CEGMA result [95] (**Table S4**), BUSCO result (**Table S5**), summary of chloroplast genome assembly (**Table S6**), data on RAD-based linkage analysis and anchoring of scaffolds (**Tables S7–S9**), validation of genome assembly (**Tables S10, S11**), summary of RNA-seq data (**Table S12**), summary of assembly of transcripts (**Table S13**), number of lectin genes in the genomes of *D. rotundata* and three species (**Table S14**), segregation of sex in F1 derived from a cross between two accessions (**Table S15**), summary statistics of bulk DNA sequencing and its analysis (**Tables S16, S17**), BLAST result of female-specific region against TDr96_F1 reference genome (**Table S18**), summary of simple sequence repeats (**Table S19**), sequences of primers used for RAD-seq (**Table S20**). (PPTX 137 kb)

Additional file 2: Figures S1–S22. Supplementary figures including a summary of world yam production and photos of yam markets in West Africa (**Figure S1**), summary of BAC-end sequencing used for genome scaffolding (**Figure S2**), summary of k-mer analysis of Guinea yam genome (**Figure S3**), flowchart of Guinea yam genome assembly (**Figure S4**), summary of Guinea yam mitochondrial genome (**Figure S5**), flowchart of RAD-seq for linkage analysis (**Figure S6**), summary of RAD-seq analysis (**Figure S7**), summary of RAD-seq DNA markers used for linkage mapping and anchoring of scaffolds (**Figure S8**), procedure of linkage analysis and split of scaffolds depending on recombination fraction between RAD markers (**Figure S9**), RAD-seq-based linkage maps of *D. rotundata* generated by pseudo-testcross method (**Figure S10**), a matrix showing scaffolds shared between two linkage groups generated for two parents (**Figure S11**), schematic diagram for developing physical map of *D. rotundata* (**Figure S12**), frequency of distances of BAC-end sequences in the genome (**Figure S13**), scheme showing pipeline of genome annotation of *D. rotundata* (**Figure S14**), self-self syntenic dot plot of *D. rotundata* pseudo-chromosomes (**Figure S15**), SyMAP dot plot analysis of whole genome synteny between three monocot species (**Figure S16**), explanation of QTL-seq analysis to identify sex-linked genome regions in *D. rotundata* (**Figure S17**), QTL-seq results (**Figure S18**), sp1 DNA marker genotypes of F1 progeny and their association with sex (**Figure S19**), explanation of method for identification of putative W-region of D. rotundata genome (**Figure S20**), identification of female- and male-specific genomic regions (**Figure S21**), method of calculation of confidence interval of QTL-seq analysis (**Figure S22**). (PPTX 15700 kb)

Additional file 3: Supplemental dataset S1. List of the 26,198 protein coding genes predicted in the *D. rotundata* genome. (XLSX 1480 kb)

Additional file 4: Supplemental dataset S2. Gene orthology between four angiosperms showing presence, absence, and duplication between species. (XLSX 1250 kb)

Additional file 5: Supplemental dataset S3. Functional annotation of *D. rotundata* genes conserved between *D. rotundata* and at least one other angiosperm (*A. thaliana*, *B. distachyon*, and *O. sativa*). (XLSX 639 kb)

Additional file 6: Supplemental dataset S4. Gene orthology between seven angiosperms showing presence, absence, and duplication between species. (XLSX 2360 kb)

Additional file 7: Supplemental dataset S5. Functional annotation of *D. rotundata* genes conserved between *D. rotundata* and at least one other angiosperm (*A. thaliana*, *B. distachyon*, *O. sativa*, *E. guineensis*, *P. dactylifera*, and *M. acuminata*). (XLSX 356 kb)

Additional file 8: Supplemental dataset S6. Functional annotation of *D. rotundata* genes with no orthologous genes found in *A. thaliana*, *B. distachyon*, *O. sativa*, *E. guineensis*, *P. dactylifera*, and *M. acuminata*. (XLSX 720 kb)

Additional file 9: Supplemental dataset S7. Non-redundant Gene Ontology terms for 2795 genes significantly (after FDR correction) enriched in *D. rotundata* with orthologous genes identified in *A. thaliana*, *B. distachyon*, *O. sativa*, *E. guineensis*, *P. dactylifera*, and *M. acuminata*. (XLSX 35 kb)

Additional file 10: Supplemental dataset S8. Non-redundant Gene Ontology terms for 11,348 genes significantly (after FDR correction) enriched in *D. rotundata* with no orthologous genes identfied in *A. thaliana*, *B. distachyon*, *O. sativa*, *E. guineensis*, *P. dactylifera*, and *M. acuminata*. (XLSX 9 kb)

Additional file 11: Supplemental dataset S9. Top 50 highest expressed genes observed to be enriched in tuber. (XLSX 59 kb)

Additional file 12: Supplemental dataset S10. List of genes predicted within the female-specific (W-linked) region on pseudo-chromosome 11 identified by QTL-seq. (XLSX 464 kb)

Additional file 13: Supplemental dataset S11. New SSR markers developed from D. rotundata genome sequence. (XLSX 126 kb)

Acknowledgements
We thank Prof. John Vogel from the DOE Joint Genome Institute (Walnut Creek, CA, USA) for access to the *B. distachyon* v3.1 genome.
This article is dedicated to the memory of the late Professor Günter Kahl, who contributed enormously to the development of yam research.

Funding
This study (design of the study, collection, analysis, and interpretation of data, and manuscript writing) was funded by the Japan International Research Center for Agricultural Sciences (JIRCAS) and the International Institute of Tropical Agriculture (IITA) as a component of an international collaborative research project involving Iwate Biotechnology Research Center (IBRC), JIRCAS, and IITA entitled the "Use of genomic information and molecular tools for yam germplasm utilization and improvement for West Africa (EDITS-Yam)." Computations were partially performed on the National Institute of Genetics (NIG) supercomputer at ROIS (Research Organization of Information and Systems) National Institute of Genetics (Mishima, Shizuoka, Japan). Work performed at The Earlham Institute, Norwich, UK, was supported by strategic Biotechnology and Biological Sciences Research Council (BBSRC) funding (Institute Strategic Programme Grant BB/J004669/1 and iCASE Studentship BB/L017350/1) and by the Norwich Bioscience Institutes (NBI) Computing infrastructure for Science (CiS) group.

Availability of data and materials
The datasets supporting the conclusions of this article are available in the DNA Databank of Japan (DDBJ)/European Molecular Biology Laboratory (EMBL)/GenBank databases.
BioProject accession number: PRJDB3383
Raw sequences:

TDr96_F1 de novo assembly: DRX025239–DRX025248 (Additional file 1: Table S2)
RNA-seq 2013,2014: DRX040446–DRX040451 (Additional file 1: Table S12)
RNA-seq 2015: DRX057356–DRX057367 (Additional file 1: Table S12)
QTL-seq basic: DRX040452–DRX040455 (Additional file 1: Table S16)
Identification of putative W-region: DRX057354–DRX057355 (Additional file 1: Table S16)
sp16-minus bulk: DRX080879 (Additional file 1: Table S16)
P3 (TDr97/00917): DRX057369 (Additional file 1: Table S17)
P4 (TDr97/00777): DRX057368 (Additional file 1: Table S17)
TDr96_F1 mitochondrial genome: DRX057351 (Additional file 2: Figure S5)

Fasta:

TDr96_F1 reference genome: DF933857–DF938579
TDr96_F1 mitochondrion: LC219374–LC219449
TDr96_F1 Pseudo_Chromosome: BDMI01000001–BDMI01000021
P3-DDN: BDML01000001–BDML01615107
P4-DDN: BDMK01000001–BDMK01641416

Genome annotation
Gene/protein sequences and gff3 annotation files are publicly available at the following URL:
http://genome-e.ibrc.or.jp/home/bioinformatics-team/yam
Code availability
QTL-seq codes used in the work are publicly available at the following URL:
http://genome-e.ibrc.or.jp/home/bioinformatics-team/mutmap

Statement
The authors declare that the study observed all local, national, and international guidelines and legislation and the appropriate permissions.

Authors' contributions
MT performed the sequencing and biological analysis, SN assembled the genome sequence, HT performed linkage analysis, scaffold anchoring, and QTL-seq analysis, BW performed comparative genomics, HY performed linkage analysis and anchoring, MS performed sex linkage analysis, KY performed genetic studies, AU performed DNA sequencing, KO performed DNA sequencing and RNA-seq, AA performed sex linkage analysis, NU performed DNA sequencing, HM performed DNA sequencing, PB performed chromosome counting and FCM analysis, SY performed population studies, RM performed F1 segregation phenotyping of sex, SM performed F1 segregation phenotyping of sex, GG performed RNA-seq, ALM performed plant crosses, MG performed DNA sampling, RB performed DNA sampling, MA performed population studies, PLK performed DNA sampling, IR performed DNA sampling, MT performed mitochondrial extraction, TT performed mitochondrial extraction, WH performed comparative genomics, MC performed comparative genomics, SK analyzed and interpreted the results, GK analyzed and interpreted the results, HT conceived and supervised the entire study, RA conceived and supervised the entire study, and RT conceived and supervised the entire study. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]Iwate Biotechnology Research Center, Kitakami, Japan. [2]The Earlham Institute, Norwich, UK. [3]Kobe University, Kobe, Japan. [4]Okinawa Agricultural Research Center, Naha, Japan. [5]Shinshu University, Nagano, Japan. [6]Tokyo University of Agriculture, Tokyo, Japan. [7]Japan International Research Center for Agricultural Sciences, Tsukuba, Japan. [8]International Institute of Tropical Agriculture, Ibadan, Nigeria. [9]Kyoto Sangyo University, Kyoto, Japan. [10]The Sainsbury Laboratory, Norwich, UK. [11]University of Frankfurt, Frankfurt, Germany. [12]Kyoto University, Kyoto, Japan.

Received: 3 May 2017 Accepted: 10 August 2017
Published online: 19 September 2017

References
1. Wilkin P, Scholsb P, Chasea MW, Chayamaritc K, Furnessa CA, Huysmansb S, Rakotonasolod F, et al. A plastid gene phylogeny of the yam genus, *Dioscorea*: roots, fruits and Madagascar. Syst Bot. 2005;30:736–49.
2. Renner SS. The relative and absolute frequencies of angiosperm sexual systems: dioecy, monoecy, gynodioecy, and an updated online database. Am J Bot. 2014;101:1588–96.
3. Maurin O, Muasya M, Catalan P, Shongwe EZ, Viruel J, Wilkin P, van der Bank M. Diversification into novel habitats in the Africa clade of *Dioscorea* (Dioscoreaceae): erect habit and elephant's foot tubers. BMC Evol Biol. 2016;16:238.
4. Lebot V. Tropical root and tuber crops: cassava, sweet potato, yams and aroids (Crop Production Science in Horticulture Series 17). Wallingford: CABI Publishing; 2009. p. 405.
5. Coursey DG. The civilizations of the yam: interrelationships of man and yams in Africa and the Indo-Pacific region. Archeol Phys Anthropol Oceania. 1972;7:215–33.
6. Ayensu ES, Coursey DG. Guinea yams: the botany, ethnobotany, use and possible future of yams in West Africa. Econ Bot. 1972;26:301–18.
7. International Institute of Tropical Agriculture (IITA). http://www.iita.org. Accessed 1 Aug 2017.
8. Scarcelli N, Daïnou O, Agbangla C, Tostain S, Pham JL. Segregation patterns of isozyme loci and microsatellite markers show the diploidy of African yam *Dioscorea rotundata* (2*n* = 40). Theor Appl Genet. 2005;111:226–32.
9. Girma G, Hyma KE, Asiedu R, Mitchell SE, Gedil M, Spillane C. Next-generation sequencing based genotyping, cytometry and phenotyping for understanding diversity and evolution of guinea yams. Theor Appl Genet. 2014;127:1783–94.
10. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. Bioinformatics. 2011;27:764–70.
11. Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A. 2011;108:1513–18.
12. Earl D, Bradnam K, St John J, Darling A, Lin D, Fass J, et al. Assemblathon 1: a competitive assessment of de novo short read assembly methods. Genome Res. 2011;21:2224–41.
13. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics. 2011;27:578–9.
14. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 2007;23:1061–7.
15. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–12.
16. Mariac C, Scarcelli N, Pouzadou J, Barnaud A, Billot C, Faye A, et al. Cost-effective enrichment hybridization capture of chloroplast genomes at deep multiplexing levels for population genetics and phylogeography studies. Mol Ecol Resour. 2014;14:1103–13.
17. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, et al. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One. 2008;3:e3376.
18. Grattapaglia D, Sederoff R. Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a pseudo-testcross: mapping strategy and RAPD markers. Genetics. 1994;137:1121–37.
19. Terauchi R, Kahl G. Mapping of the *Dioscorea tokoro* genome: AFLP markers linked to sex. Genome. 1999;42:752–62.
20. Jiang SY, Ma Z, Ramachandran S. Evolutionary history and stress regulation of the lectin superfamily in higher plants. BMC Evol Biol. 2010;10:79.
21. Afzal AJ, Wood AJ, Lightfoot DA. Plant receptor-like serine threonine kinases: roles in signaling and plant defense. Mol Plant Microbe Interact. 2008;21:507–17.
22. Ohizumi Y, Gaidamashvili M, Ohwada S, Matsuda K, Kominami J, Nakamura-Tsuruta S, et al. Mannose-binding lectin from yam (*Dioscorea batatas*) tubers with insecticidal properties against *Helicoverpa armigera* (Lepidoptera: Noctuidae). J Agric Food Chem. 2009;57:2896–902.
23. Kato T, Hori M, Ogawa T, Muramoto K, Toriyama K. Expression of gene for *Dioscorea batatas* tuber lectin 1 in transgenic tobacco confers resistance to green-peach aphid. Plant Biotechnol. 2010;27:141–5.
24. Yoshimura S, Komatsu M, Kaku K, Hori M, Ogawa T, Muramoto K, et al. Production of transgenic rice plants expressing *Dioscorea batatas* tuber lectin 1 to confer resistance against brown planthopper. Plant Biotechnol. 2012;29:501–4.
25. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B Methodol. 1995;57(1):289–300.
26. Xue YL, Miyakawa T, Sawano Y, Tanokura M. Cloning of genes and enzymatic characterizations of novel dioscorin isoforms from *Dioscorea japonica*. Plant Sci. 2012;183:14–9.
27. Lyons E, et al. Using genomic sequencing for classical genetics in *E. coli* K12. PLoS One. 2011;6(2):e16717.
28. Takagi H, Abe A, Yoshida K, Kosugi S, Natsume S, Mitsuoka C, et al. QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. Plant J. 2013;74:174–83.
29. Love RR, et al. Evaluation of DISCOVAR de novo using a mosquito sample for cost-effective short-read genome assembly. BMC Genomics. 2016;17(1):187.
30. Tamiru M, Yamanaka S, Mitsuoka C, Babil P, Takagi H, Lopez-Montes A, et al. Development of genomic simple sequence repeat markers for Yam. Crop Sci. 2015;55:2191–200.
31. Martin FW. Sex ratio and sex determination in *Dioscorea*. J Heredity. 1966;57:96–9.
32. Terauchi R, Kahl G. Sex determination in *Dioscorea tokoro*, a wild yam species. In: Ainsworth CC, editor. Sex determination in plants. Oxford: BIOS Scientific Publishers; 1999. p. 163–71.

33. Murase K, Shigenobu S, Fujii S, Ueda K, Murata T, et al. MYB transcription factor gene involved in sex determination in *Asparagus officinalis*. Genes Cells. 2016;22:115–23.
34. Sakata Seed Co. http://www.sakataseed.co.jp/. Accessed 1 Aug 2017.
35. Olympus Co. http://www.olympus-global.com/en/. Accessed 1 Aug 2017.
36. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. Nature. 2005;436:793–800.
37. Beckman Coulter Co. https://www.beckmancoulter.com/. Accessed 1 Aug 2017.
38. Macherey-Nagel GmbH & Co. KG. http://www.mn-net.com. Accessed 1 Aug 2017.
39. Operon Co. http://www.operon.com/. Accessed 1 Aug 2017.
40. Lucigen Co. http://www.lucigen.com/. Accessed 1 Aug 2017.
41. Genaris Co. http://genebay.co.jp/. Accessed 11 Sept 2017.
42. Hannon laboratory. http://hannonlab.cshl.edu/fastx_toolkit/. Accessed 1 Aug 2017.
43. Terachi T, Tsunewaki K. The molecular basis of genetic diversity among cytoplasms of *Triticum* and *Aegilops*: 5. Mitochondrial genome diversity among Aegilops species having identical chloroplast genomes. Theor Appl Genet. 1986;73:175–81.
44. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26:589–95.
45. Ian Korf Lab. http://korflab.ucdavis.edu/datasets/genome_completeness/. Accessed 1 Aug 2017.
46. Kohany O, Gentles AJ, Hankus L, Jurka J. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics. 2006;7:474.
47. Nussbaumer T, Martis MM, Roessner SK, Pfeifer M, Bader KC, Sharma S, et al. MIPS PlantsDB: a database framework for comparative plant genome research. Nucleic Acids Res. 2013;41:D1144–51.
48. Smit AFA, Hubley R. RepeatModeler. Open-1.0. (2008–2015).
49. Smit AFA, Hubley R, Green P. RepeatMasker. Open-4.0. (2013–2015).
50. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, et al. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. 2012;40:D1202–10.
51. International Brachypodium Initiative. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature. 2010;463:763–8.
52. Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, et al. The TIGR Rice Genome Annotation Resource: improvements and new features. Nucleic Acids Res. 2007;35:D883–7.
53. Quiagen Co. https://www.qiagen.com. Accessed 1 Aug 2017.
54. Salamov AA, Solovyev VV. Ab initio gene finding in *Drosophila* genomic DNA. Genome Res. 2000;10:516–22.
55. Stanke M, Steinkamp R, Waack S, Morgenstern B. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res. 2004;32:W309–12.
56. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14:R36.
57. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012;7:562–78.
58. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith Jr RK, et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 2003;31:5654–66.
59. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013;8:1494–512.
60. Allen JE, Salzberg SL. JIGSAW: integration of multiple sources of evidence for gene prediction. Bioinformatics. 2005;21:3596–603.
61. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 2011;12:491.
62. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. 2005;21:1859–75.
63. Slater GS, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 2005;6:31.
64. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10:R25.
65. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27:2987–93.
66. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005;21:3674–76.
67. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30:1236–40.
68. Remm M, Storm CE, Sonnhammer EL. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol. 2001;314:1041–52.
69. O'Brien KP, Remm M, Sonnhammer EL. Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res. 2005;33:D476–80.
70. Berglund AC, Sjölund E, Östlund G, Sonnhammer EL. InParanoid 6: eukaryotic ortholog clusters with in paralogs. Nucleic Acids Res. 2008;36:D263–66.
71. Brachypodium distachyon v3.1 DOE-JGI, https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Bdistachyon/. Accessed 11 Sept 2017.
72. Droc G, Larivière D, Guignon V, Yahiaoui N, This D, Garsmeur O, et al. The banana genome hub. Database. 2013;2013:bat035.
73. Singh R, Ong-Abdullah M, Low ET, Manaf MA, Rosli R, Nookiah R, et al. Oil palm genome sequence reveals divergence of interfertile species in Old and New worlds. Nature. 2013;500:335–39.
74. Al-Mssallem IS, Hu S, Zhang X, Lin Q, Liu W, Tan J, et al. Genome sequence of the date palm *Phoenix dactylifera* L. Nat Commun. 2013;4:2274.
75. Alexeyenko A, Tamas I, Liu G, Sonnhammer EL. Automatic clustering of orthologs and in paralogs shared by multiple proteomes. Bioinformatics. 2006;22:e9–15.
76. The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017;45(D1):D158–69.
77. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80.
78. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–3.
79. Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics. 2014;31:166–9.
80. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
81. Tang H, Klopfenstein D, Pedersen B, Flick P, Sato K, Ramirez F, et al. GOATOOLS: Tools for Gene Ontology. Zenodo. 2015: http://doi.org/10.5281/zenodo.31628.
82. Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One. 2011;6:e21800.
83. Lyons E, et al. The value of nonmodel genomes and an example using SynMap within CoGe to dissect the hexaploidy that predates the rosids. Trop Plant Biol. 2008;1(3):181–90.
84. Schwartz S, Kent WJ, Smit A, et al. Human–mouse alignments with BLASTZ. Genome Res. 2003;13(1):103–7.
85. Haas BJ, Delcher AL, Wortman JR, Salzberg SL. DAGchainer: a tool for mining segmental genome duplications and synteny. Bioinformatics. 2004;20(18):3643–6.
86. Lyons E, Freeling M. How to usefully compare homologous plant genes and chromosomes as DNA sequences. Plant J. 2008;53(4):661–73.
87. Yang Z. PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24(8):1586–91.
88. Matsumura H, Miyagi N, Taniai N, Fukushima M, Tarora K, Shudo A, et al. Mapping of the gynoecy in bitter gourd (*Momordica charantia*) using RAD-seq analysis. PLoS One. 2014;9:e87138.
89. Broman KW, Wu H, Sen Ś, Churchill GA. R/qtl: QTL mapping in experimental crosses. Bioinformatics. 2003;19:889–90.
90. Iwate Biotechnology Research Center. http://genome-e.ibrc.or.jp/home/bioinformatics-team/mutmap. Accessed 1 Aug 2017.
91. Kosugi S, Natsume S, Yoshida K, MacLean D, Cano L, Kamoun S, et al. Coval: improving alignment quality and variant calling accuracy for next-generation sequencing data. PLoS One. 2013;8:e75402.
92. Wang X, Lu P, Luo Z. GMATo: a novel tool for the identification and analysis of microsatellites in large genomes. Bioinformation. 2013;9:541–44.
93. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. Primer3—new capabilities and interfaces. Nucleic Acids Res. 2012;40:e115.
94. FAO. http://faostat3.fao.org. Accessed 1 Aug 2017.
95. CEGMA page of Ian Korf Lab. http://korflab.ucdavis.edu/datasets/cegma/#SCT7. Accessed 1 Aug 2017.

Supplementary Data 1

Multiple sequence alignment of conserved Debranching Enzyme 1 between 26 angiosperm species.

>EMT14551_pep_supercontig_ASM34733v1_Scaffold44269_26834_31197__1_gene_F775_08059_transcript_EMT14551_gene_biotype_protein_coding_transcript_biotype_protein_coding_description_debranching_enzyme_1__Source_Projected_from_Arabidopsis_thaliana_AT4G31770_TAIR----------
MKIAVEGCMHGELDKVYDTMQRLEAAEGIKIDLLICCGDFQAVRNESDLQCVNVPPKF--
RTMNSFWKYYSGQAVAPYPTIFIGGNHEAANYLWELYYGGWAAPNIYFLGFA----------
GHHERPPYDNATIRSVYHVRHYDVLKLMHVKEPLDIFMSHDWPLGITEYGNWERLLREKPFFTEESVAPMGNKSMPIWKTNPYGI
KFCMVHQRKLGSESAAKLLNKLKPPYWFSAHLHCRFPAIIQHGEDGPTTKFLALDKCLPGRNFLQVIDIPSNPGPYEIQYDEEWLAI
TRRFNSAFPLTRMPCTIRNEELDIQDDRQWVRSKLNARGAKTFDFVQTAPPYDPSRPVYNPPIAVPCRNPQTESFLQFLELPYLLDS--
---------SNPGGVDINVSSSQAAP----------ALDNDDIE------LPD-----
EVEDDEDDEE----------->evm_27_model_AmTr_v1_0_scaffold00033_28MS----
LWACFNGVHTPVLEIDDPYKA--
DAEIKRSMRFFTTVGRFGWAPMISRITSTRLGLQLQISSLFNRVSMKIAIEGCMHGELDNVYSTLLYIEKVENTKIDLLICCGDFQAIR
NKDDLKSVNVKPKYWDNCMNSFWKYYSGKETAPIPTIFVGGNHEASNYLWELYYGGWTAPNIYFLGFAGVVKFGDIRIGGLSGIY
KARDYYSGHYERTPYNSNDIRSIYHVRQYDVYKLMQIEEKIDIFISHDWPLGITDHGSCKELIQQKPFFERE----------
IRERSLGSKPAAELLEKLKPAYWFSGHLHCKFPAIVQHGEDGPITKFLALDKCLPGRKFLQIIEIESNPGPFEIHYDEEWLAITLKYNPL
LPLTRKSAQLGGEEPGLQGYRERVKNQLMARGSKPFEFSPTVPAHDPESFSGDPSFQGHIRNPQIESFLQLLGLPYLLDS----------
NQESDAFLRSSSSLEFRGQNDQA------FDNDGDD-------DDV---
DDIEALAGGEF----------
>Aco015889_1_pacid_33051834_transcript_Aco015889_1_locus_Aco015889_ID_Aco015889_1_v3_annot_version_v3----------
----
MKIAVEGCMHGELDAVYATLRRLEQVENTKIDLLLCCGDFQSVL----------
RYEKSSRKLAVEEGKW----------LMKWNRRR----------
HFERPPYNDNDIRSIYHVRHYDVLKLMHVEEPIDIFLSHDWPLGITEYGNWEKLIKIKQHFTEE----------
VYSRTLGSKPAAELLNKLRPPFWFSAHLHCKFPAVIQHGKDGPITKFLALDKCLHGRRFLQIIDIESDPGPYEIQYDEEWLAITKKFN
NILPLTRKYFDGGVKQLDIQDCRQWVMRKLDERGAKPFDFVQTVPAYDPSQTFSNPFRTGHCRNPQTESLLQFLELPYLLDG--------
--ENVANTPSQSTKQLNPTD----------ALDDD----------DDV---DELEEHADSDDEE----------
>AT4G31770----------
MKIAIEGCMHGDLDNVYKTIQHYEQIHNTKVDLLLCCGDFQAVRNEKDMDSLNVPRKY--
REMKSFWKYYSGQEVAPIPTIFIGGNHEASNYLWELYYGGWAATNIYFLGFAGVVKFGNVRIGGLSGIYNERHYRSGHFERPPYNES
TIRSVYHVRDYDVQKLMQLEEPLDIFLSHDWPVGITDYGDSESLMRQKPYFRQE----------
IEEKTLGSKPAALLLEKLKPQYWFSAHLHCKFAAAVQHGNDGSVTKFLALDKCLPGKKFLQIIEIESEPGPFEVLYDEEWLAITRKFN
SIFPLTRRYTNVSTAG-TIQESREWVRKKLEERQFKPFEFARTVPAYNPSQRVFD-SIPEIPQNPQTLSLLELLGLPYLLDS----------
SPVTGERTDIPASLAPSDL--------PTYDSEEIP------IDDI---DEIEEMEEAKADDHT----------
RDDA----------
>Bv5_124850_aamx_t1_cDNAEvidence_91_3----------
MKIAVEGCMHGDLDNVYKTIIELERAENTKIDLLICCGDFQAVRNKKDLESLNVKPKY--
RQMNTFWKYYSGQVVAPVPTVFIGGNHEASNYLWELYYGGWAAPNIYFLGFAGVIKFGSIRIGGLSGIYNFRHYKLGHFERPPYNE
SDIRSIYHVREYDAHKLMQVEEPIDIFLSHDWPLGITDCGNWRELVRKKPFFEKE----------
IQERTLGSKPAAKLLENLKPSYWFSAHLHCKFSSLVQHGDDGPVTKFLALDKCLPGRKFLQVIDIESGPGPFEICYDEEWLAITRIFNP
AFPLTRKPADLRSVQVDKEGCRQWVRSKLNIRGAKPFEFSRTAPCYDPVQSVSNGPGAEDHRNPQTEALLKFLELPYLLDN----------
STASSELSSDSPSVSRGF----------YSNNEDIP------IDDV---DELEELAADDDDD----------
>Bradi1g10190----------
MKIAVEGCMHGELDKVYDTLRKLEEAEGVKIDLLICCGDFQAVRNESDLQCVNVPDKF--
RTMNSFWKYYSGQAVAPYPTIFIGGNHEASNYLWELYYGGWAAPNIYFLGFAGVVKFGNVRIGGLSGIHKQQHYYLGHHERPPYD
QSSIRSVYHVRHYDVLKLMHVKEPLDIFMSHDWPLGITEYGNWQNLIRDKKFFEEE----------
VNNRTLGSEPAAKLLNKLKPPYWFSAHLHCKFPAIIQHGEDGPTTKFLALDKCLPGRNFLQVIDIPSNPGPYEIQYDEEWLAITRRFN
SVFPLTWMRFTIRNEQLDIQDDRQWVRSKLNASGAKPFDFVQTAPPFDPSKPVSNPSLAVHCRNPQTESFLQFLELPYLLDS----------
SHSEGLDRNVSGSQAGH----------PFGDDSIE------LPD-----EVEDADDDE----------
>evm_model_supercontig_40_1_pacid_16419400_transcript_evm_model_supercontig_40_1_locus_evm_TU_supercontig_40_1_annot_version_ASGPBv0_4----------
MKIAVEGCMHGDLDKVYKTIQYMEQIHTTKIDLLLCCGDFQAVRNERDMESLSVPSKY--
LAMKSFWKYYSGQEVAPIPTIFIGGNHEASNYLWELYYGGWVAPNIYFLGFAGVVKFGNIRIGGLSGIYNQRHYRSGHFERPPYNEN
TVRSVYHIREYDVHKLLQVDEPIDIFLSHDWPLGITDYGDWKKLVQQKRHFEKE----------
IQERTLGSKPAAQLLEKLKPHYWFSAHLHCKFAAVVQHGESGPLTKFLALDKCLPGRKFLQIVEVESEPGPYELQYDEEWLAITRKF
NSIFPLTRRNADFGEKKKLKFPIRYSVK----------FPKTIVLLYRFC-----IILTHQRRRLERCRLKM----------
PLQEFGSSAHFCS----------SEIQELS----------
>Cqu_c17399_1_g002_1__2_46_----------
MKIAVEGCMHGDLDNVYKTIADLQREENTKIDLLICCGDFQAVRNKKDLESLNVKPKY--
RQMNTFWKYYSGEEVAPVPTIFIGGNHEASNYLWELYYGGWAAPNIFFLGFAGVVKFGNIRIGGLSGIYNFRHYKLGHFERPPYNES
DIRSIYHVREYDVHKLMQVQEPIDIFLSHDWPVGITDCGNWRELVRKKPFFEKE----------
IQERTLGSKPAAKLLEKLKPSYWFSAHLHCKFSALVQHGEGGSMTKFLALDKCLPNRKFLQIIEIESDPGPFEICYDEEWLAITRTFN
AAFPITARPADFQSVQVDIEGSRQCVRSKLNTRGVKPFEFTRTAPCHDPSQSVFNGSSAEHYRNPQTEALLKLLELPYLLDN----------
SIASREMSSYSPSISKAAM----------DVK------MED-----
AAASTASTSEANVAPPTAVEKSPYDLLKKSKLSVEEIVAKMLSLKKDDKPKPELRELVTQMFLNFVSLRQVVLRIIN
>augustus_masked_scaffold104_size1425670_processed_gene_0_200_mRNA_1_protein_AED_0_49_eAED_0_49_QI_130_0_88_0_9_1_0_88_0_8_10_320_395----------
MKIAIEGCMHGDLDKVYATLKRLEEEEKTKIDLLLCCGDFQAVRNLNDLESLNVKPKY--
RSMNSFWKYYSGQAVAPYPTIFIGGNHEASNYLWELV----------NFCVSTIKCVICSP--
LIGHFERPPYNESDIRSIYHVREYDVLKLMQIEEPIDIFISHDWPLGITEYGNWQKLLREKPFFKEE----------
VEKRSLGSRPAAELLAKLKPPYWFSAHLHCKFPAIIQHGEDGPVTKFLALDKCLPRRKFLQIVEIESEPGPFEIHFDEEWLTITRMFNS
SFPVTRRSARFVAEQLGKQAHQQWVRNKLSARGPKPFEFVRTVPSYDPDQPHPSTTLDGHCRNPQTEAFLQLLELPYLLDA----------

VAQSNTSVESSLHSADAHLHS------SYINDESID------
LDDVDYMDDMEELTTNEDA----------
>XP_010936250_1----------
MKIAVEGCMHGELDKVYATLKHLEEVENTKIDLLLCCGDFQAIRNENDLESLNVKPKY--
RNMNSFWKYYSGQAVAPYPTIFIGGNHEASNYLWELYYGGWAAPNIYFLGYSGVIKFGNIRIGGLSGIYKPRDYHSGHYEMPPYDT
NDIRSVYHVREYDVMKLKEVKGQIDIFMSHDWPLRITEYGNWEKLVRQKPFLRQE----------
VLDGTLGSVPAAELLKALQPHYWFSAHLHCKFPAIIQHGENGPVTKFLALDKCLRGRKFLQIVDIESDPGPYEIQYDEEWLAITRKF
NSIFPLTHKSVQLGG--LDEQDYQQWVRNKLNARGAKPFDFVQTVPSFDPSQPLSNRSLCGHIRNPQTQSLLQFLELPYLLAI----------
TAEANTPNVNDE----------LFDHEYVD------VDDV---DELEELVQVDDENT----------
>gi_1109300817_ref_XP_019175171_1__PREDICTED__lariat_debranching_enzyme_isoform_X2__Ipomoea_nil_----------
----------MKIAVEGCMHGDLDNVYATLLHLQDVEKIKIDLLICCGDFQAVRNEKDLESLNVPPKY--
KSMNSFWKYYSGEKVAPFPTIFIGGNHEASNYLWELYYGGWAAPQIYFLGFAGVIKFGNIRIGGLSGIYKSHHYHSGHYEKVPYNEL
DIRSIYHVREYDIHKLMQVEEPIDIFLSHDWPVGITDHGNLKSLLRQKPFFEQE----------
IQEGTLGSKPAAELLEKLRPSYWFSAHLHCKFAALVQHGE-
GSVTKFLALDKCLPGRNFLQVIEVESEPGPYELQYDEEWLAIMRKFNSILPLTIRHADYSNVQLDLQECRHFVRNKLQSRGAKPFDF
VRTVPCHDPRQPLANGVFSGHCRNPQTEALLQLLELEYLLDN----------MSESSSFG----------YGTEDIP------IDDV---
DDIDEPKEADASETE----------NEQI----------
>gi_657945350_ref_XP_008379412_1__PREDICTED__lariat_debranching_enzyme__Malus_x_domestica_----------
----------MKIAVEGCMHGDLDNVYRTLQHYEKLHDTTIDLLLCCGDFQAVRNEKDLDSLNVPHKF--
RSMNTFWKYYSGEAVAPYPTIFIGGNHEASNYLWELYYGGWAAPNIYFLGFAGVVKFGNLRIGGLSGIYKQPNFLKGHFERPPFND
STIRSVYHVREYDFHKLMQVEEPIDIFLSHDWPVGITDCGDWKKLVQQKPYFRQE----------
VQERKLGSKPAAQLLEKLKPPYWFSAHLHCKFAARVQHGEDGPVTNFLALDKCQPGRKFLQVIEIESEPGPYEIQYDEEWLEITRRF
NSIFPLTTRSANLWNVRLDKQECRQWVRSKLQARGARPFEFTQTVPPYNPSQSVTYGSFPGHVRSPQTESLLQFLELPYLLDN--------
--VSQSSEVLPSPPRGS----------VEENSEDIP------IDDV---DELEEDAVDCENENS----------
>Ma08_p27380----------
MKIAVEGCLHGEMDKVYDTIRHMEKVENIKIDLLICCGDFQAVRNENDLESLSCPPKY--
RSMNSFWKYYSGQEVAPVPTIFIGGNHEASNYLWELYYGGWAAPNVYFLGFSGVIKFGNIRIGGVSGIYKQGHYHLGHFERPPYNE
SDLRSVYHVREYDVMKLKDIKEPIDIFISHDWPVGIYEYGNSKRLVQRKPHFEEE----------
IRKRTLGSLPAAELLNQLKPHYWFSGHLHCNFAAVVQHREDGSVTKFLALDKCLPGRKFLQIVDVNSDPGPYEIQYDEEWLAITRK
FNSIFPLSRKTVHLRPEQLDKQDYREWVRNKLITRGARPFDFVQTVSPFDPSRPITRSSTSGHCRNPQTESLLQLLELPYLLDN---------
-VAEAGMPSQNPNFFNTKDAWNQ-----LSDDGNSAE------VGDV---
DELEELAEDEDGDA----------
>gi_719982089_ref_XP_010250324_1__PREDICTED__lariat_debranching_enzyme_isoform_X2__Nelumbo_nucifera_----------
----------MKIAVEGCMHGDLDNVYSTLLHLEKVENTKIDLLICCGDFQAVRNETDLESLNVKAKY--
RSMNSFWKYYSGEEVAPFPTIFIGGNHEASNYLWELYYGGWAAPNIFFLGFAGVIKFGNIRIGGLSGIYNARHYNLGHYERPPYNEN
DIRSIYHVREYDVYKLMQVEEPIDIFLSHDWPVGITDYGNWKELVRFKPFFEKE----------
IEERTLGSKAAADLLDKLKPPYWFSAHLHCKFAALVQHGEDGPVTKFLALDKCLPGRKFLQVVEIESKPGPYEIQYDEEWLAITKN
FNCIFPLTRRLVHLGNLQVNLQDYRQWVRDKIKTRGAKPFDFTRTLPIYDPSCPGSNGSIPGHQRNPQTESLLKFLELPYVLDS---------
--TSESNVLMHLISADANEVA--------LNDENEDAE------IEDV---DEMEELAEVGVVDAG----------
NDQ----------
>Oropetium_20150105_07412A_pacid_36018865_transcript_Oropetium_20150105_07412A_locus_Oropetium_20150105_07412_ID_Oro
petium_20150105_07412A_v1_0_annot_version_v1_0MFDVMLHWE----------
FPVQIAVEGCMHGELDTVYDTLRRLEEAEGIKIDLLLCCGDFQAVRNESDLQCLNVPHKF--
RSMNTFWKYYSGQAVAPYPTIFIGGNHEASNYLWELYYGGWAAPNIYFLGFAGVVKFGNIRI----------
GHYERPPYDEHTIRSVYHVRHYDVLKLMHVKEPLDIFLSHDWPLGITEYGDWQKLIRVKRHFEEE----------
VMNRALGSKPAAELLDKLKPPYWFAAHLHCQFPAIIQHGEGGPTTKFLALDKCLPRRNFLQVIDIPSNPGPYEIHYDEEWLAITRKF
NSFFPLTRMRFTMRYEQLDIQDDRQWVRSKLNTRGSKPFDFVQTAPSFDPSRRVSNHSIPVPCRNPQTESFLELLELPYLLDS----------
SKVVGDDQTESSLQPGQ----------APDNDDIE------LPD-----EVEDTVEDDE----------
>gi_1002247547_ref_XP_015627948_1__PREDICTED__lariat_debranching_enzyme__Oryza_sativa_Japonica_Group_----------
----------MKIAVEGCMHGELDKVYDTLRELEKAEGVKIDLLLCCGDFQAVRNENDLQCLNVKPRF--
REMKSFWKYYSGQAVAPYPTIFIGGNHEASNYLWELYYGGWAAPNIYFLGFAGVVKFGNIRIGGLSGIYKQQHYHLGHYERPPYNE
NTIRSVYHVRHYDVLKLMHVKEPLDIFMSHDWPLGITEYGNWQKLIREKRFFEEE----------
VNKRTLGSEPAARLLNKLKPPYWFSAHLHCKFPAVIQHGEGGPTTKFLALDKCLPRRGFLQVIDIPSGPGPHEIQYDEEWLAITRKFN
NVFSLTRMPFTMLDEQVDTQDDLQWVRNKLNARGAKPIDFVQTAASYDPSCQASNPSITVHCRNPQTESFLQLLNLPYLLDS--------
--SNSYGVSRNESSSQTGQ----------ALDSDDIE------LPD-----DEDDPADDDD----------
>Pahal_I01926_1_pacid_32509773_transcript_Pahal_I01926_1_locus_Pahal_I01926_ID_Pahal_I01926_1_v2_0_annot_version_v2_0------
----
MKIAVEGCMHGELDIVYDTLRRLEEAEGIKIDLLLCCGDFQAVRNKDDLRCVNVPLKY--
RAMNSFWKYYSGQAVAPYPTIFIGGNHEASNYLWELYYGGWAAPNIYFLGFSGVVKFGNIRIGGLSGIHKQHNYHSGHYERPPYNE
QTIRSVYHVRHYDVLKLMHVKEPLDIFLSHDWPLGITEYGNWQKLIRAKKHFEEE----------
VNNRTLGSKPAAELLNKLKPPYWFSAHLHCRFPAIIQHGENGPTTKFLALDKCLPGRNFLQVIDIPSNPGPHEIQYDEEWLAITRRFN
SIFPLTRRRFSIRDEQLDTQDDREWVRNKLNTRGVKPFDFVQTAPSFNPSNPVSNSSITRSCRNPQTESFLQLLELPYLLDS----------
SNSEEVDRNQSSSQPGN----------TLDDEDIE------LPD-----EDEDAIDE----------
>gi_1175875120_ref_XP_020591193_1__lariat_debranching_enzyme__Phalaenopsis_equestris_----------
----------MKIAVEGCLHGELDKVYATIKRLEEAENTKIDLLLCCGDFQAVRNEEDLKSLNVPPKY--
RHMNSFWKYYSGQVTAPYPTIFIGGNHEASNYLWELYYGGWAAPNIYFLGFAGVIKYGNIRIGGLSGIYKSKDYHLGHYERPPYDD
NSIRSVYHVRQYDVLRLMQIKEPIDIFLSHDWPLGITEYGNWEKLVQQKPFFKDE----------
VRLRTLGSKPAAELLDKLKPPYWFSAHLHCRFPAIIQHGENGPITKFLALDKCLPGRRFLQIFEIKSDPGPPEIQYDEEWLAITRKFNR
FFPLSREYFHLRSDHFESQDFRNWVRSQLNARGAKPFEFLKTMPSFDPNKP-SSALPSGHCRNPQTVSFLKLLELPYLLDI----------
KDETSTPKKITEFSSPLGIHSQKNLD-VEEGEDDVH------GDDV---
DELEELAACGNDEF----------

>XP_008778774_1----------
MNSFWKYYSGQAVAPYPTIFIGGNHEASNYLWELYYGGWAAPNIYFLGYSGVIKFGNIRIGGLSGIYKPREYHLGHYEMPPYDENDI
RSVYHVREYDVMKLKEVKGQIDIFMSHDWPLRITEYGNWEKLVRHKPFFRQE----------
VLDGTLGSVPAAELLKALQPRYWFSAHLHCKFPAIIQHGENGPVTKFLALDKCLRGRKFLQIVDIEADPGPYEVQYDEEWLAITRKF
NSIFPLTRKSVQLGG--LDKQEYQQWVRDKLNARGAKPFDFVPTVPSFDPSQALSNRSHCGHIRNPQTESLLQFLELPYLLDI----------
TAEANTPNVNDG----------LFDREYVD------LDDV---DELEELAQVDDEDT----------
>PUT_163a_Populus_nigra_5697_1_PlantGDB_assembled_Unique_Transcript_fragment_derived_from_Populus_nigra_mRNAs_Jan_27_
2008_based_on_GenBank_release_163_NF----------SRQQIQNQRV-----
RSMKIAIEGCMHGDLDKVYQTLKLIESQNGTKIDLLLCCGDFQAVRNERDMESLNVPLKY--
REMKSFWKYYSGREIAPVPTIFIGGNHEASNYLWELCYGGYAAPNIYFLGFAGVIKFGNIRIGGLSGIYNARNYRTGHHERAPYNES
SIRSVYHVREYDVHKLMQVEEPIDIFLSHDWPVGITDCGNWKQLVRYKPHFEKE----------
IQEKTLGSKAAA----------

>PUT_157a_Saccharum_officinarum_112443_1_PlantGDB_assembled_Unique_Transcript_fragment_derived_from_Saccharum_officinaru
m_mRNAs_Jan_28_2007_based_on_GenBank_release_157_------RWS-----PSPRLC-GTPTSAPRPAWLSRGAE------
RAGCAPLETAAGAPSCRAARQRPWTVP-
GTMKIAVEGCMHGELDIVYDTLRKLEEAEGVKIDLLLCCGDFQAVRNENDLQWVNVPHKY--
RTMNSFWKYYSGEAVAPYPTIFIGGNHEAFKYLWEMYYGRRAAPNIYFLGVAGGGKFGNIPNCGLAGKKTRSPFYRDQPGGP--------
----
RP----------
RGYQTGRX----------
>Sevir_9G108900_1_p_pacid_32653597_transcript_Sevir_9G108900_1_locus_Sevir_9G108900_ID_Sevir_9G108900_1_v1_1_annot_ver
sion_v1_1----------
MKIAVEGCMHGELDIVYDTLRRLEEAEWIKIDLLLCCGDFQAVRNTDDLRCVNVPLKY--
RNMNSFWKYYSGQAVAPYPTIFIGGNHEASNYLWELYYGGWAAPNIYFLGFAGVVKFGNIRIGGLSGIHKQHNYHSGHYERPPYNE
QTIRSVYHVRHYDVLKLMHVKEPLDIFLSHDWPLGITEYGNWQELIRAKNHFEAE----------
VNNRTLGSKPAAELLNKLKPPYWFSAHLHCRFPAIIQHGENGPTTKFLALDKCFRGRNFLQVIDIPSNPGPYEIHYDEEWLAITRRFN
SVFPLTQRRFTMRDEQLDTQDDRQWVRSKLNARGFKPFDFVQTAPSFNPSNPVSNSSITGSCRNPQTESFLQLLELPYLLDS----------
SNSEGVERNESSSQPGN----------TLGDEDIE------LPD-----EDEDAADDDE----------
>ref_XP_021306749_1__lariat_debranching_enzyme__Sorghum_bicolor_----------
MKIAVEGCMHGELDIVYDTLRKLEEAEGVKIDLLLCCGDFQAVRNENDLQCVNVPQKF--
RAMNSFWKYYSGEAVAPYPTIFIGGNHEASNYLWELYYGGWAAPNIYFLGFAGVVKFGNIRIGGLSGIYNKYRYHLGHYERPPYNE
DTIRSVYHVRHYDVLKLMHLKEPLDIFLSHDWPLGITEYGNWQKLISVKKHFEEE----------
VNNRTLGSKPAAELLNKLKPPYWFSGHLHCKFPAIIQHGKNGPTTKFLALDKCIPGRNFLQVIDIPSNPGPYEIQYDEEWLAITRKFN
SVFPLARTRFTMRDEQLDTQEDRQWVRSKLNTRGAKPFDFVQTAPSFNPSNTISKHSTTVCCRNPQTESFLQLLELPYLLDSSNSEG
HYLKKSNSEGFGRNESSSQPGN----------TLDDEDIE------LPD-----
EDEDALEDDE----------
>Spipo2G0106600_Lariat_debranching_enzyme----------
NQIAVEGCMHGELDNVYATIQHLEKVENIKIDLLICCGDFQAVRYQSDLNSLNVKPNY--
RKMNSFWKYYSGEEIAPYPTIFIGGNHEASNYLWELYYGGWAAPNIYFLGYAGVMKFGDIRIAGLSGIYNPRHYNLGHYERPPYNE
SDIRSIYHVREYDVFKLMQIKEPIDLFISHDWPLGITDFGDSDNLIRKKPYFRRE----------
IEERTLGSRAAAQLLDKLKPPYWFSAHLHCKFPAVIQHGENGPVTKFLALDKCLPRRQFLQILEIGSDPGPHEIMFDEEWLAITRKFN
SVFPLTRKPW-LGAQQDENQDHYQWIKDKLKARGGRPFEFIRTPPSSGNLCF-FFCDPLGHQRNPQTESLLEFLELPYLLDV----------
TTETSTVLQGNILCRHLSQYENG----LHYSYYSLKCVRCFFL----------
>PUT_169a_Vitis_vinifera_39861_1_PlantGDB_assembled_Unique_Transcript_fragment_derived_from_Vitis_vinifera_mRNAs_Jan_14_
2009_based_on_GenBank_release_169_NP----------SSPVFL----------LSAIEVALNPLKPIS-
ANMRIAVEGCMHGDLNNVYSTLRYLEEVENTKIDLLICCGDFQAVRNKKDLESLNVPPKY--
RSMNSFWKYYSRQEVAPFPTIFIGGNHEASNYLWELYYGGWAAPNIYFLGFAGVVKFGNIRIGGLSGIYNERHYHLRHYERPPYNE
RDIRSVYHVREYDVHKLMQVEEPIDIFLSHDWPCGITDHGNWKELVRYKPFFEKE----------
IQERTLGNKAAAELLGKLKPSYWFSAHLHCKFAAPCPNGGG----------W----------
SKLX----------

>Zosma96g00440_1_pacid_33177684_transcript_Zosma96g00440_1_locus_Zosma96g00440_ID_Zosma96g00440_1_v2_2_annot_version
_v2_2----------
MRIAVEGCMHGELDVVYGTLQDLERRENVKIDLLICCGDFQSVRNEEDLKSLNCPPNY--
RKMNSFHKYYSGDLIAPFPTIFIGGNHEASNYLWELYYGGWAAPNIYFLGFAGIIKFGNIRIGGLSGIYKATHYSMGHFERPPYNASD
IRSVYHVREYDVHKLMQIEEPIDIFISHDWPLGITDFGNWQKLIKQKSYFEKE----------
IRERTLGSRPAAQLLDKLKPPYWFSAHLHCKFPAIIQHGDNGPSTKFLALDKCLPNRQFLQIIEIGSDVGPFELQYDEEWLAITRKFHC
IIPLTRRPAQLRAQLTDIQENRQWVKNNITSKGTQPFDFARTSTS----------SGHCRNPQTELLLEFLDG----------
SMNTGLMGDDSSSLKHKDKAKQVVDDSYIYKTEEIF------LPD-----DDDESSDDEKNESE----------EIAG------
RDKAN----------N