Editors summary

A method for rapid cloning of plant disease-resistance genes could provide sustainable genetic solutions to crop pests and pathogens in place of agrichemicals.

Accelerated cloning of a potato late blight–resistance gene using RenSeq and SMRT sequencing

Kamil Witek[1,4], *Florian Jupe*[1,3,4], *Agnieszka I Witek*[1], *David Baker*[2], *Matthew D Clark*[2], *and Jonathan DG Jones*[1]

[1]The Sainsbury Laboratory, Norwich Research Park, Norwich, United Kingdom

[2]The Genome Analysis Centre (TGAC), Norwich Research Park, Norwich, United Kingdom

[3] Current address: Plant Biology Laboratory, Salk Institute for Biological Studies, La Jolla, CA, United States

[4] These authors contributed equally to this work

Authors to whom correspondence should be addressed: Kamil Witek; kamil.witek@tsl.ac.uk or Jonathan DG Jones; jonathan.jones@tsl.ac.uk

Author: abstract should be <159 words

Global yields of potato and tomato crops are reduced owing to potato late blight disease, which is caused by *Phytophthora infestans*. Although most commercial potato varieties are susceptible to blight, wild potato relatives are not and are therefore a potential source of *Resistance to P. infestans* (*Rpi*) genes. Resistance breeding has exploited *Rpi* genes from closely related tuber-bearing potato relatives, but is laborious and slow [1–3]. Here we report that the wild, diploid non-tuber-bearing *Solanum americanum* harbors multiple *Rpi* genes. We combine *R* gene sequence capture (RenSeq[4]) with single-molecule real-time SMRT sequencing (SMRT RenSeq) to clone *Rpi-amr3i* . This technology should enable *de novo* assembly of complete nucleotide-binding, leucine-rich repeat receptor (NLR) genes, their regulatory elements and complex multi-NLR loci from uncharacterized germplasm. SMRT RenSeq can be applied to rapidly clone multiple *R* genes for engineering pathogen-resistant crops.

Plants have complex defense systems that prevent or reduce disease if activated promptly upon pathogen detection. Timely activation of defense mechanisms requires pathogen detection using either cell surface receptors or intracellular immune receptors of the nucleotide-binding, leucine-rich repeat (NLR, aka NB-LRR or NBS-LRR) protein family[5-7]. Most disease resistance (*R*) genes encode NLR proteins, which directly or indirectly recognize pathogen effectors, and can carry TIR domain or a coiled-coil (CC) signaling domain at their N-termini (TIR-NLRs or CC-NLRs)[5]. Plant reference genome sequences can be mined for NLR genes, and biotinylated RNA sequence capture libraries synthesized to clone NLRs from related, unsequenced taxa (RenSeq)[4]. RenSeq on genomic DNA from previously unsequenced parents, and from bulked resistant and susceptible genotypes, followed by short-read sequencing, assembly and comparison, enables genetic mapping of NLR-type *R* genes. It also enables (re)annotation of NLR complements, and identification of transcribed NLRs (cDNA RenSeq[4,8]). However, large copy numbers and highly repetitive coding sequences impair accurate *de novo* assembly of complete NLR genes using short reads[4,8]

Many *Solanum* species have been assessed for genetic variation in resistance to *P. infestans*, but resistance is rare[3]. The hexaploid *S. nigrum* is reported as resistant, and *S. americanum* is a likely diploid ancestor of *S. nigrum*[9]. To clone novel *Resistance to Phytophthora infestans* (*Rpi)* genes from the wild potato relative *S. americanum*, we screened thirteen diploid *S. americanum* accessions from three European seed collections for late blight resistance (Supplementary Table 1). We assessed pathogen susceptibility to four highly virulent *P. infestans* isolates (88069, 06_3928A, EC1 and MP324) using detached leaf assays (DLAs). Accession 954750186 (working name SP2271; Supplementary Table 1) was susceptible to all isolates tested and supported both mycelial growth and sporulation. All other accessions remained resistant, with either no visible sign of infection or only small sites of local hypersensitive response (HR) at the site of *P. infestans* inoculation (Supplementary Figure 1a). We crossed resistant *S. americanum* accessions to the susceptible SP2271 accession. Heterozygous F_1 progeny were self-pollinated, and 60 to 100 plants per F_2 were screened for response to inoculation with 06_3928A and EC1. Six F_1 crosses segregated in a ratio that indicated the presence of a single (semi) dominant resistance gene (fitting 3:1 or 2:1), and six crosses had either a 15:1 segregation or all plants remained resistant, which might indicate two or more unlinked *Rpi* genes (Supplementary Table 1).

Here, we describe identification and cloning of one of these *Rpi* gene derived from the Mexican *S. americanum* accession 944750095 (working name SP1102;

Supplementary Table 1). We generated a mapping population and positioned the underlying resistance gene on Ch 4: 3.5-8.5Mb, using short-read RenSeq combined with bulked segregant analysis (BSA) (Supplementary Figure 1b, Supplementary data, Supplementary File 1). The underlying resistance gene locus carries three NLR clusters in both the potato and tomato reference genomes; the *R2/Rpi-blb3* cluster and the uncharacterized clusters C17 and C18. The syntenic region in potato hosts 30, seven and ten NLRs, respectively[4,10]. Using a backcrossed F_2 population and markers derived from RenSeq and whole genome shotgun (WGS) data, we positioned the gene to the C18 NLR cluster (see Supplementary Figure 1c and Supplementary Methods).

Identification and cloning of *R* gene candidates is slow and expensive. To accelerate *Rpi* gene cloning, and to remove any need for BAC or fosmid libraries construction, we refined RenSeq to capture and sequence fragments of up to 3.2 kb, which is the average NLR gene length. We previously used RenSeq to target the complete NLR complement in several Solanaceae with 500 bp-insert libraries and various short-read Illumina sequencing platforms. However, high copy number and sequence similarity between paralogs make it difficult to assemble *de novo* the full NLR repertoire from short reads[4,10].

Here we explored NLR enrichment in combination with long read single-molecule real time (SMRT) sequencing using the PacBio RSII platform[11-13] (SMRT RenSeq). We used our *Solanum* NLR bait library[4,10] to capture the NLR complement from two DNA libraries of different size (1.5 kb and 2.5 kb fragments), derived from the same resistant SP1102 accession, and sequenced these on one and two SMRT cells, respectively. As PacBio reads from P5-C3 or later chemistries average more than 10 kb, most SMRT RenSeq molecules have multiple sequence passes from which a consensus Reads of Inserts (ROI) sequence is generated. The resultant 70,600 ROIs were analysed with an NLR-specific[4,10] motif alignment and search tool (MAST[14]) and NLR-parser[15]. More than 21,500 ROIs (30%) derive from NLRs, and 1030 (~5%) reads harbor even full length NLR coding sequences. Individual ROIs, including those containing full length NLRs, were *de novo* assembled, resulting in contigs that harbor 322 full length and 293 partial NLRs, with a coverage of 5 – 70x (Supplementary Table 2).

We compared short-read NLR assemblies with long-read assemblies, and confirmed the fragmented reconstruction of the NLR-complement from short-read sequencing data; 41% CLC Assembly Cell (CLC, www.clcbio.com) or 43% SPAdes assembler[16]-derived contigs had 99% coverage by a SMRT-derived contig, but only 21% (or 29%)

revealed full length NLRs. The central positions of matching aligned CLC or SPAdes NLR assemblies to long-read contigs (Figure 1a) further confirmed the superiority of SMRT RenSeq by also capturing >1 kb of flanking promoter and terminator sequences. This additional sequence information revealed potential regulatory elements, and enabled the PCR amplification of the gene including promoter and terminator sequences for complementation (Figure 1a-c). We found 327 sequences assembled from long-read data for which no high confidence short-read assembly is present. Reciprocally, we found just four short-read contigs that were not assembled using SMRT RenSeq data. Inspection revealed corresponding PacBio ROI reads within the SMRT dataset, however the generated contigs had read coverage below the threshold (Supplementary Methods and Supplementary Table 3). We manually confirmed five short-read SPAdes contigs as chimeric sequences, in which parts of the short read chimera matched different long-read contigs (Supplementary Figure 2)**.** These comparisons between long and short-read assemblies (SPAdes and CLC, Supplementary Methods) clearly showed that SMRT RenSeq generates reads that can be assembled into complete genes including their regulatory elements. Capture and sequencing of long fragments can resolve any repetitive gene family or structural genome variation[13] by spanning repeat-rich regions with long reads.

We assessed the effect of fragment length, and performed a SMRT RenSeq experiment on the size-selected 3-4 kb gDNA library from the susceptible line SP2271, that on average, was 1 kb larger than the library from the resistant line SP1102. 34,300 ROI were from NLRs (60%), and 5,689 reads (10%) harbored full-length NLR coding sequences. Similar to the R parent, we assembled *de novo* individual ROIs and identified 401 complete and 245 partial NLRs (Supplementary Table 2). While we identified similar numbers of NLR genes using this library, the increase in average fragment length from 2.5 kb to 3.5 kb increased the average flanking sequence length from 1,341 bp to 2,702 bp (Figure 1b). Fragments with 2.5kb (R parent) and 3.5kb (S parent) also enabled assembly of multi-NLR contigs, for example an *Sw5* (five NLRs) or an *R2-like* locus (five NLRs) in R and S parents, respectively (Figure 1a, Supplementary Figure 3 and Supplementary File 2).

Because ROI reads have higher error rates than the Illumina short-reads[17], we assessed how this affects the NLR gene assembly. First, we corrected the assembled C18 NLR encoding contigs from both parents using high-quality Illumina short-reads (see Supplementary data for methods). We then mapped the original ROI reads back to these QC'd assemblies and identified an accuracy of more than 98.6% (average 99.6). We did not observe a link between ROI length (500 bp to

4000 bp) and percentage of errors (Supplementary Figure 4a). Overall, we identified 171 single-nucleotide errors within the assembled contigs, comprising a pool of 276,000 bases. These include three substitutions, and 145 (85%) single nucleotide insertions or deletions (23, 13.5%), occurring within homopolymer regions. We found that 52% of the errors were located in regions with ROI coverage lower than 3x e.g. at the edges of contigs. ROI coverage over 20x ensured high quality assemblies with 99.98% accuracy (Supplementary Figure 4b). Subsequent candidate gene cloning steps are not affected by this error rate, and potential false-positive frame-shift mutations are revealed by cDNA RenSeq.

The availability of a comprehensive *S. americanum* NLR complement allowed us to study this family in more detail. In contrast to findings in Brassicaceae and Triticeae, *Solanum* spp. have relatively few NLR proteins with fusions to non-NLR domains[18]. We found 28 NLRs to be fused to a "domain of unknown function" DUF3542 and two cases of fusion to ASF1-like histone chaperone. Single fusions were found for a RNA polymerase 3, DUF659, a protein tyrosine phosphatase domain and several other domains (Supplementary Table 4). Despite capturing long flanking sequences, we could not find any inverted paired NLRs[18-20]. The phylogenetic tree of SP2271 NLRs (Figure 2), constructed from the NB-ARC domains of complete NLRs (Supplementary Files 3-5), shows a Solanaceae typical structure, including a TIR-NLR to CC-NLR ratio of 1:4.7, opposite to that found in Brassicaceae[4]. We also observed an expansion of certain clades (e.g. CNL-10) and emergence of a novel clade CNL-17, compared to tomato and potato NLR phylogenetic trees[8,10].

To identify putative *Rpi-amr3* sequences from the SP1102 NLR repertoire, we searched for *S. americanum* C18 homologs in the PacBio assembly. We identified 14 complete (single ORF, full-length cds; Supplementary Figure 1d) and 10 pseudogenised NLRs (>80% identity >1kb). Mapping of cDNA RenSeq short-read data to these sequences[8,21] provided evidence for the expression of six candidates with a uniform cDNA coverage over the whole sequence (Supplementary Table 5). We further confirmed co-segregation of four of these sequences using gene specific markers (*Rpi-amr3a, Rpi-amr3i, Rpi-amr3j* and *Rpi-amr3k*). Therefore, SMRT-RenSeq enabled us to determine the full sequence of six co-segregating, expressed candidate *Rpi* genes rapidly. We PCR amplified the open reading frames of these six candidate NLRs and placed them under control of *35S* promoter in binary vector pICSLUS0003 (Supplementary File 6). These constructs were transiently expressed using *Agrobacterium* in *Nicotiana benthamiana* leaves, which were 48 hours later detached and inoculated with *P. infestans* isolate 88069. *P. infestans* growth was observed 6 dpi on GFP-infiltrated control leaves and all other constructs, except for the *R2* control and the candidate gene *Rpi-amr3i* (Figure 3a, Supplementary Figure 5). Transient delivery of candidate *Rpi-amr3i* cloned under its native promoter and terminator elements (1.9 kb 5' and 0.8 kb 3', Figure 1c) in the pICSLUS0001 binary vector (Supplementary File 7), followed by *P. infestans* infection, conferred similar resistance levels as the *35S:Rpi-amr3i* construct (Supplementary Figure 6). This confirmed functionality of the candidate gene *Rpi-amr3i*. We next created stable transgenic plants carrying *Rpi-amr3i* under the control of either *35S* promoter or its own promoter in diploid homozygous potato *Solynta Research line 26* (www.solynta.com). *Rpi-amr3i*-carrying plants showed resistance against all tested diverse *P. infestans* isolates (Figure 3b; Supplementary Figure 7 and Supplementary Table 6), in contrast to the five lines that were transformed with 35S:*Rpi-amr3a, b, j, k* or *l* paralogs (data not shown). Transgene expression levels were analysed with quantitative RT-PCR and showed mRNA levels similar to the parental accession for *Rpi-amr3i* for fully resistant lines (Supplementary Figure 8 and Supplementary data). This result confirms the functionality of *Rpi-amr3i* for resistance against multiple isolates of *P. infestans in planta*.

Rpi-amr3i has a single 2,664 bp ORF encoding 887 amino acids (Supplementary Figure 9) with motifs and domains typical for a CC-NLR protein; coiled-coil (CC, 1-115), nucleotide binding (NB-ARC 151-433) and leucine-rich repeats (LRR, Figure 1c). We identified 14 full length *Rpi-amr3*-like NLRs in the R parent, which share between 70% to 96% nucleotide identity. In the susceptible parent, the *Rpi-amr3* family contains 17 full-length members (Supplementary Figure 10b), and the closest paralog to the functional *Rpi-amr3i* shares 96.5% DNA identity but has an early stop codon after the NB-ARC domain rendering it non-functional (Supplementary Figure 5). Combined phylogenetic analyses of C18 NLRs of both parents identifies two subgroups and a distant relationship with the physically linked *R2/Rpi-blb3* family (Supplementary Figure 10a)[4], and show high similarity between the two accessions of *S. americanum* (Supplementary Figure 10b).

We expect that a single *R* gene could be mapped to sufficient resolution within two generations using SMRT-Renseq, provided >100 susceptible segregants can be identified. SMRT Renseq enabled *de novo* assembly and analysis of full NLR repertoires in a previously un-sequenced *S. americanum* accession. We rapidly cloned a broad-spectrum resistance gene against *P. infestans* from *S. americanum*, a diploid, homozygous, wild *Solanum* species with high genetic diversity, and thereby identify *S. americanum* as a novel source of resistance genes against *P. infestans*. The SMRT RenSeq method has the potential for use in investigating genetic variation for other important traits likely to involve known multigene families such as metabolic pathways (e.g. cytochromes P450, terpene cyclases) or transcription factors, especially if combined with mutagenesis (Steuernagel et al 2016, accompanying paper).

Accession codes at GenBank: Sequencing data: PRJEB9916 (Individual run numbers within this project are detailed in Supplementary Data); *Rpi-amr3i* locus: KT373889

Acknowledgements

We would like to thank Sandy Knapp, Sophien Kamoun, Brande Wulff, Burkhard Steuernagel, Ghanasyam Rallapalli, Oliver Furzer and Nicolas Champouret for useful discussions, The Genome Analysis Centre (TGAC) for excellent PacBio service, Laurence Tomlinson and William Barrett for technical assistance, Pim Lindhout for Solynta potato lines, Sara Perkins and Justine Campling for maintenance of plants in the greenhouse, Matthew Smoker and Jody Pike for plant transformation, Vivianne Vleeshouwers for the *R2* construct, TSL Bioinformatics Support Team and specifically Christian Schudoma for bioinformatics assistance and Andrew Davies for photography. This work was financed from BBSRC Grants no. BB/H019820/1, BB/G02197X/1, BB/L011794/1, BB/L009757/1, BB/J010375/1 and the Gatsby charitable foundation. This research was supported in part by the NBI Computing infrastructure for Science (CiS) group through the provision of a High Performance Computing Cluster.

Author contribution

K.W., F.J. and J.D.G.J. designed the study. K.W and A.I.W. performed the experiments. K.W., F.J. and A.I.W. analyzed the data. M.D.C and D.B. contributed to SMRT RenSeq protocol development. K.W., F.J. and J.D.G.J. wrote the manuscript. K.W. and F.J. made equivalent contributions and should be considered joint first authors.

Competing financial interests

K.W. and J.D.G.J. have filed a US patent application 62/159,240 based on this work. M.D.C. owns shares in Pacific Biosciences of California.

References

1. Jones, J. D. G. *et al.* Elevating crop disease resistance with cloned genes. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **369,** 20130087–20130087 (2014).
2. Haverkort, A. J. *et al.* Societal Costs of Late Blight in Potato and Prospects of Durable Resistance Through Cisgenic Modification. *Potato Res.* **51,** 47–57 (2008).
3. Rodewald, J. & Trognitz, B. Solanumresistance genes against Phytophthora infestansand their corresponding avirulence genes. *Molecular Plant Pathology* **14,** 740–757 (2013).
4. Jupe, F. *et al.* Resistance gene enrichment sequencing (RenSeq) enables reannotation of the NB-LRR gene family from sequenced plant genomes and rapid mapping of resistance loci in segregating populations. *Plant J.* **76,** 530–544 (2013).
5. Dangl, J. L. & Jones, J. D. Plant pathogens and integrated defence responses to infection. *Nature* **411,** 826–833 (2001).
6. Jones, J. D. G. & Dangl, J. L. The plant immune system. *Nature* **444,** 323–329 (2006).
7. Dangl, J. L., Horvath, D. M. & Staskawicz, B. J. Pivoting the plant immune system from dissection to deployment. *Science* **341,** 746–751 (2013).
8. Andolfo, G. *et al.* Defining the full tomato NB-LRR resistance gene repertoire using genomic and cDNA RenSeq. **14,** 1–12 (2014).
9. Lebecka, R. Host–pathogen interaction between Phytophthora infestans and Solanum nigrum, S. villosum, and S. scabrum. *Eur J Plant Pathol* **120,** 233–240 (2007).
10. Jupe, F. *et al.* Identification and localisation of the NB-LRR gene family within the potato genome. *BMC Genomics* **13,** 75 (2012).
11. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323,** 133–138 (2009).
12. Sharon, D., Tilgner, H., Grubert, F. & Snyder, M. A single-molecule long-read survey of the human transcriptome. *Nat Biotechnol* **31,** 1009–1014 (2013).
13. Wang, M. *et al.* PacBio-LITS: a large-insert targeted sequencing method for characterization of human disease-associated chromosomal structural variations. *BMC Genomics* **16,** 214 (2015).
14. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucl. Acids Res.* **37,** W202–8 (2009).
15. Steuernagel, B., Jupe, F., Witek, K., Jones, J. D. G. & Wulff, B. B. H. NLR-parser: rapid annotation of plant NLR complements. *Bioinformatics* **31,** 1665–1667 (2015).
16. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19,** 455–477 (2012).
17. Jiao, X. A Benchmark Study on Error Assessment and Quality Control of CCS Reads Derived from the PacBio RS. *J Data Mining in Genom Proteomics* **04,** 1–12 (2013).
18. Cesari, S., Bernoux, M., Moncuquet, P., Kroj, T. & Dodds, P. N. A novel conserved mechanism for plant NLR protein pairs: the ‘integrated decoy’ hypothesis. *Front Plant Sci* **5,** 606 (2014).
19. Narusaka, M. *et al.* RRS1 and RPS4 provide a dual Resistance-gene system against fungal and bacterial pathogens. *Plant J.* **60,** 218–226 (2009).
20. Sarris, P. F. *et al.* A Plant Immune Receptor Detects Pathogen Effectors that Target WRKY Transcription Factors. *Cell* **161,** 1089–1100 (2015).
21. Rallapalli, G. *et al.* EXPRSS: an Illumina based high-throughput expression-

profiling method to reveal transcriptional dynamics. *BMC Genomics* **15,** 341 (2014).

Figures captions

Figure 1. Comparison of MiSeq and PacBio RenSeq read-based assemblies.
(**a**) Representation of the 33 kb long *Contig_7*, which is derived from assembled SP2271 ROI reads. The contig contains four complete *R2*-like NLR genes and one partial NLR and is compared to the corresponding contigs assembled using MiSeq reads (using the SPAdes algorithm). Predicted open reading frames (ORFs) are marked in purple, and black color above the contig indicates PacBio ROI read depth. The assembly covers a 4.6 kb intergenic region, which is not present in the target capture design. MiSeq read-derived contigs are shown below *Contig_7*. Correctness of the assembled contig was confirmed with WGS data (see Supplementary Figure 3). (**b**) Long DNA libraries allowed capture of promoter and terminator regions, not covered in the capture design. The length of the adjacent 5' region is depicted in a comparison between PacBio ROI assembled contigs 2.5kb (blue) and 3.5kb (dark pink) and MiSeq-derived contigs using the CLC (green) or SPAdes (maroon) assembler. (**c**) Details of the *Rpi-amr3i* contig, assembled using PacBio ROI reads, showing the ORF (mustard yellow), NLR typical protein domains: coiled-coil (CC; purple), nucleotide binding site (NB-ARC; blue) and leucine-rich repeats (LRR; green). ROI reads that assembled into the *Rpi-amr3* contig are shown in black. Picture drawn to scale.

Figure 2. The phylogenetic tree of SP2271 NLRs.
Full NB-ARC domains of 363 annotated NLR genes were used together with 31 functionally characterized plant *R* genes (red and blue font, CED4 used as outgroup) in a maximum likelihood analysis based on the Poisson (G+I) model. CC-NLR (CNL) clades are collapsed based on a bootstrap value higher than 79 and numerated. Bootstrap values of 70% and higher are shown for the main branches. The TIR-NLR (TNL) clade is drawn with a yellow background. Labels show the gene IDs. The tree is drawn to scale, with branch lengths proportional to the number of substitutions per site.

Figure 3. Candidate gene *Rpi-amr3i* confers full resistance against *Phytophthora infestan*s in transient complementation assays in *Nicotiana benthamiana* and in stable transgenic potato plants*.*

(**a**) Third leaves of *N. benthamiana* plants were infiltrated with the binary vector pICSLUS0003::*35S* overexpressing either the late blight resistance gene *R2* (positive control), one of six *Rpi-amr3* candidates or GFP (negative control). Leaves were inoculated with *P. infestans* strain 88069 24 hours post-infiltration. Only leaves infiltrated with *R2* and *Rpi-amr3i* (pictured) remained infection free, while *P. infestans* grew well on the remaining *Rpi-amr3* candidates and on the GFP control. The figure shows *Rpi-amr3a* whose phenotype was the same for all non-functional candidate genes (See Supplementary Figure 5). Photographs were taken 6 dpi. (**b**) Transgenic diploid potato “Line 26” (Solynta B.V.) which expresses *Rpi-amr3i* under the native regulatory elements is resistant to *P. infestans* isolates 88069 (top). The transgenic line displays no to weak HR at the spot of inoculation. In contrast, the control plants carrying non-functional candidate (*Rpi-amr3a,* bottom) show large necrotic lesions and sporulation. Each leaflet was inoculated with a droplet containing 500 zoospores; photographs were taken 6dpi. Scale bars indicate 1cm. These experiments were repeated three times with similar results.