Next-generation re-sequencing as a tool for rapid bioinformatic screening of presence and absence of genes and accessory chromosomes across isolates of Zymoseptoria tritici

McDonald, Megan C.; Williams, Angela H.; Milgate, Andrew; Pattemore, Julie A.; Solomon, Peter S.; Hane, James K.

DOI:
10.1016/j.fgb.2015.04.012

Fungal Genetics and Biology 79 (2015) 71–75

Contents lists available at ScienceDirect

Fungal Genetics and Biology

journal homepage: www.elsevier.com/locate/yfgbi

Next-generation re-sequencing as a tool for rapid bioinformatic screening of presence and absence of genes and accessory chromosomes across isolates of *Zymoseptoria tritici*

Megan C. McDonald [a], Angela H. Williams [b], Andrew Milgate [c], Julie A. Pattemore [d], Peter S. Solomon [a], James K. Hane [b,e,*]

[a] Plant Science Division, Research School of Biology, The Australian National University, Canberra, Australia
[b] CCDM Bioinformatics, Centre for Crop and Disease Management, Department of Environment and Agriculture, Curtin University, Perth, Australia
[c] NSW Department of Primary Industries, Wagga Wagga Agricultural Institute, Wagga Wagga, Australia
[d] Charles Sturt University, Wagga Wagga, Australia
[e] Curtin Institute for Computation, Curtin University, Perth, Australia

ARTICLE INFO

Article history:
Received 6 January 2015
Revised 10 April 2015
Accepted 13 April 2015

Keywords:
Zymoseptoria tritici
Accessory chromosome
Next-generation sequencing
Comparative genomics
Presence–absence variation

ABSTRACT

The wheat pathogen *Zymoseptoria tritici* possesses a large number of accessory chromosomes that may be present or absent in its genome. The genome of the reference isolate IPO323 has been assembled to a very high standard and contains 21 full length chromosome sequences, 8 of which represent accessory chromosomes. The IPO323 reference, when combined with low-cost next-generation sequencing and bioinformatics, can be used as a powerful tool to assess the presence or absence of accessory chromosomes. We present an outline of a range of bioinformatics techniques that can be applied to the analysis of presence–absence variation among accessory chromosomes across 13 novel isolates of *Z. tritici*.

1. Introduction

The establishment of whole-genome sequence (WGS) resources for several important fungal species has been a significant milestone in the field of fungal biology (Grigoriev, 2013). In the study of the wheat pathogen *Zymoseptoria tritici* (syn. *Mycosphaerella graminicola*, *Septoria tritici*), access to WGS resources for the representative isolate of this species, IPO323 (Goodwin et al., 2011), has enabled the juxtaposition of additional bioinformatic data derived from the transcriptome (Brunner et al., 2013; Kellner et al., 2014; Yang et al., 2013a) and proteome (Yang et al., 2013b). These combined resources have led to insights into the genome biology of *Z. tritici* (Croll et al., 2013; Goodwin et al., 2011; Kellner et al., 2014; Morais do Amaral et al., 2012; Torriani et al., 2011; Yang et al., 2013a) (reviewed in (McDonald et al., 2015) and Testa et al., 2015). Furthermore, these resources have also enabled molecular plant pathologists to routinely adopt reverse genetics approaches, greatly accelerating the accumulation of knowledge of plant–microbe interactions at a molecular level (Perez-Nadales et al., 2014).

An important feature of *Z. tritici* is its bipartite set of chromosomes, distinguished by being either essential for growth (core) or accessory (syn. dispensable) (Croll and McDonald, 2012; Goodwin et al., 2011). The WGS assembly of *Z. tritici* IPO323 comprises 21 near-chromosome-length sequences of a total length of 39.9 Mb and was selected as the reference isolate partly due to its large number of accessory chromosomes (ACs) (Goodwin et al., 2011). Alternate isolates of *Z. tritici* have been observed to lack ACs found in IPO323 (an Algerian isolate IPO95052 and isolates described in Croll et al. (2013)) (Goodwin et al., 2011). *Z. tritici* ACs have been observed to have lower gene density, higher repetitive DNA content and depleted G:C base compositions (Croll et al., 2013; Kellner et al., 2014) – all hallmarks of hotspots of gene innovation within fungal genomes (Croll and McDonald, 2012). Consequently, the ACs of *Z. tritici* may play important roles in plant pathogenicity and their presence could potentially be used as markers for pathogenicity phenotypes – as is the case in other fungal plant-pathogens possessing ACs including *Fusarium oxysporum* f. sp. *lycopersici* (Ma et al., 2010) and *Fusarium solani*

* Corresponding author at: Department of Environment & Agriculture, Curtin University, GPO Box U1987, Perth, Western Australia 6845, Australia. Tel.: +61 8 9266 1726.
E-mail address: James.Hane@curtin.edu.au (J.K. Hane).

http://dx.doi.org/10.1016/j.fgb.2015.04.012
1087-1845/

(Coleman et al., 2009). It is important to note that unlike *Fusarium* spp. ACs have never been directly associated with virulence towards particular wheat cultivars, however genes on accessory chromosomes have been shown to be under accelerated evolution and are highly expressed *in planta,* both hallmarks for pathogenicity related genes (Kellner et al., 2014; Stukenbrock et al., 2010). *Z. tritici* ACs were originally proposed to have originated via horizontal transfer from an unknown source, which was followed by degeneration and extensive recombination with core chromosomes (Goodwin et al., 2011; Stukenbrock et al., 2011). Early reports also noted widespread paralogy between genes on accessory and core chromosomes (Goodwin et al., 2011), however these speculations have since been disputed (Kellner et al., 2014).

Next-generation sequencing (NGS), now a widely used technique for low-cost, high-throughput nucleotide sequencing, has been widely applied to several fungal species including *Z. tritici*. Since the WGS assembly was completed using Sanger sequencing before NGS was available, NGS has only been applied to the reference isolate for the sequencing of its transcriptome (RNA-Seq). However NGS can also be a powerful tool when applied to the re-sequencing of alternate *Z. tritici* isolates with variable pathogenicity on wheat (McDonald et al., in preparation). Since the reference genome assembly of *Z. tritici* IPO323 is of very highquality in near-complete chromosome-length sequences and represents most known ACs, it is possible to use short-length NGS reads aligned to the IPO323 reference for intra-species comparisons across multiple *Z. tritici* isolates without the need for further genome assembly. We present a case-study illustrating the application of NGS genome re-sequencing of multiple *Z. tritici* isolates and its use in rapidly determining whether ACs are present or absent in novel isolates of *Z. tritici.*

A complementary approach is the comparison of *de novo* assemblies of NGS reads from alternate isolates to the IPO323 reference. This can serve as a useful means of capturing sequence data (contigs or genes) where Illumina reads do not map reliably. While IPO323 contains a large number of ACs, there is also potential for novel isolates to possess additional sequences in the form of extra ACs or large insertion mutations that are not present in IPO323. *De novo* assembly of this data, via tools such as SPAdes (Bankevich et al., 2012) can produce large contig or scaffold sequences, which can then be probed with tools such as BLAST (Altschul et al., 1990) for coding genes that are located on an alternative chromosome when compared to IPO323.

Additionally, the NGS re-sequencing examples presented in this study highlight the idiosyncrasies of the *Z. tritici* ACs in contrast to its core chromosomes and are also useful at gene-level resolution to rapidly identify genes that are present, absent or mutated across isolates.

2. Methods

The various methods presented in this study are summarised in Fig. 1, with details of miscellaneous methods and scripts available in Supplementary Data 1.

Genomic DNA of several *Z. tritici* isolates was sequenced via the Illumina HiSeq 2000 platform. Standard paired-end sequence libraries with 100 bp read lengths and approximate insert size of 250 bp were generated for the following isolates: WAI221, WAI56, WAI332, WAI147, WAI320, WAI321, WAI322, WAI323, WAI324, WAI326, WAI327, WAI328 and WAI329. Short sequence reads were trimmed for adapter, primer and low quality sequences via Cutadapt v1.1 (homopolymer/polyN < 5 bp,>Q30, discard reads < 50 bp) (Martin, 2011).

NGS reads were aligned to the reference genome, initially as per a previous study (Croll et al., 2013). Paired end Illumina libraries were aligned to the IPO323 reference assembly via bowtie2 v 2.1.0 (parameters: – sensitive – end-to-end) (Langmead and Salzberg, 2012) producing alignment outputs that were converted to BAM format via SAMtools v0.1.19 (Li et al., 2009). The percentage of IPO323 chromosomes covered by NGS reads of novel isolates was calculated via BEDtools genomeCoverageBed (requiring at least 10× coverage) using BAM alignments generated in this study (Quinlan and Hall, 2010). The percentage of IPO323 genes covered by NGS reads of novel isolates was similarly calculated using BEDtools coverageBED using BAM alignments (this study) and gene annotation GFF data downloadable from JGI Mycocosm (Goodwin et al., 2011; Grigoriev et al., 2013) and Ensemblfungi (Kersey et al., 2014). This data was used to determine presence–absence variation (PAV) of ACs across isolates. Regional coverage and PAV of IPO323 chromosomes by genes, repeats and NGS alignments from alternate *Z. tritici* isolates were visualised using Circos (Krzywinski et al., 2009), pooling counts of each respective dataset within 100 Kb increments (Fig. 2).

BAM alignments were also processed with GATK v1.5-20 (DePristo et al., 2011; McKenna et al., 2010; Van der Auwera et al., 2013) UnifiedGenotyper (-stand_call_conf 50.0-stand_emit_conf 10.0-dcov 50 – genotype_likelihoods_model BOTH) to determine sites of mutation across *Z. tritici* isolates relative to the IPO323 reference, producing outputs describing chromosome coordinates of single nucleotide polymorphisms (SNPs) and insertion/deletion mutations (indels) in variant call format (VCF). VCF data was also compared to JGI GFF gene annotations via BEDtools coverageBED as above, for the purpose of determining the number of mutations present between isolates at each *Z. tritici* IPO323 locus. For the purpose of presenting this data concisely, we have summarised the gene-level mutation rate (SNPs and indels) at the chromosome-level (Fig. 3, Supplementary Table 1).

To demonstrate how novel isolate assemblies can be applied to PAV analysis, *de novo* assembly was performed on all isolates with SPAdes v3.30 with all paired reads and unpaired reads that lost their mate during quality trimming (-k 21,33,55,77 – careful) (Bankevich et al., 2012). A FASTA file of genomic DNA gene sequences was generated from a modified version of the gene annotations provided in GFF3 format at Ensemblfungi (Kersey et al., 2014). This FASTA file was used as the queries for local BLASTN searches. *De novo* SPAdes assembled contigs were converted to local BLAST databases with BLAST+ v2.2.27 (Camacho et al., 2009) (makeblastdb-in file.fasta-input_type fasta-dbtype nucl-parse_seqids-out isolate-title isolate.spades.database). BLASTN searches were run on each gDNA gene region (-evalue 1e-3-outfmt '6 qseqid sseqid pident length qlen qstart qend slen sstart send bitscore evalue'). The presence of each gene in *de novo* assemblies relative to IPO323 was determined if the top BLAST hit represented 50% of the total length of the IPO323 gene. The python script used to generate FASTA files, parse BLAST results and generate presence/absence files for each *de novo* assembly is available in Supplementary Data 1. This dataset represents a gene-level PAV analysis and furthermore the *de novo* assembled sequences that did not match to the IPO323 genome could be used to further investigate novel ACs not present in IPO323. For the purposes of presenting this data, we have summarised the gene-level PAV analysis at both the chromosome-level (Fig. 2A, Supplementary Table 2) and at the gene-level (Fig. 2B). The R script and example data used to generate Fig. 2B are also available in Supplementary Data 1.

3. Results & discussion

At the chromosome-level, alignment of NGS genome data of various *Z. tritici* isolates to the IPO323 reference assembly indicates

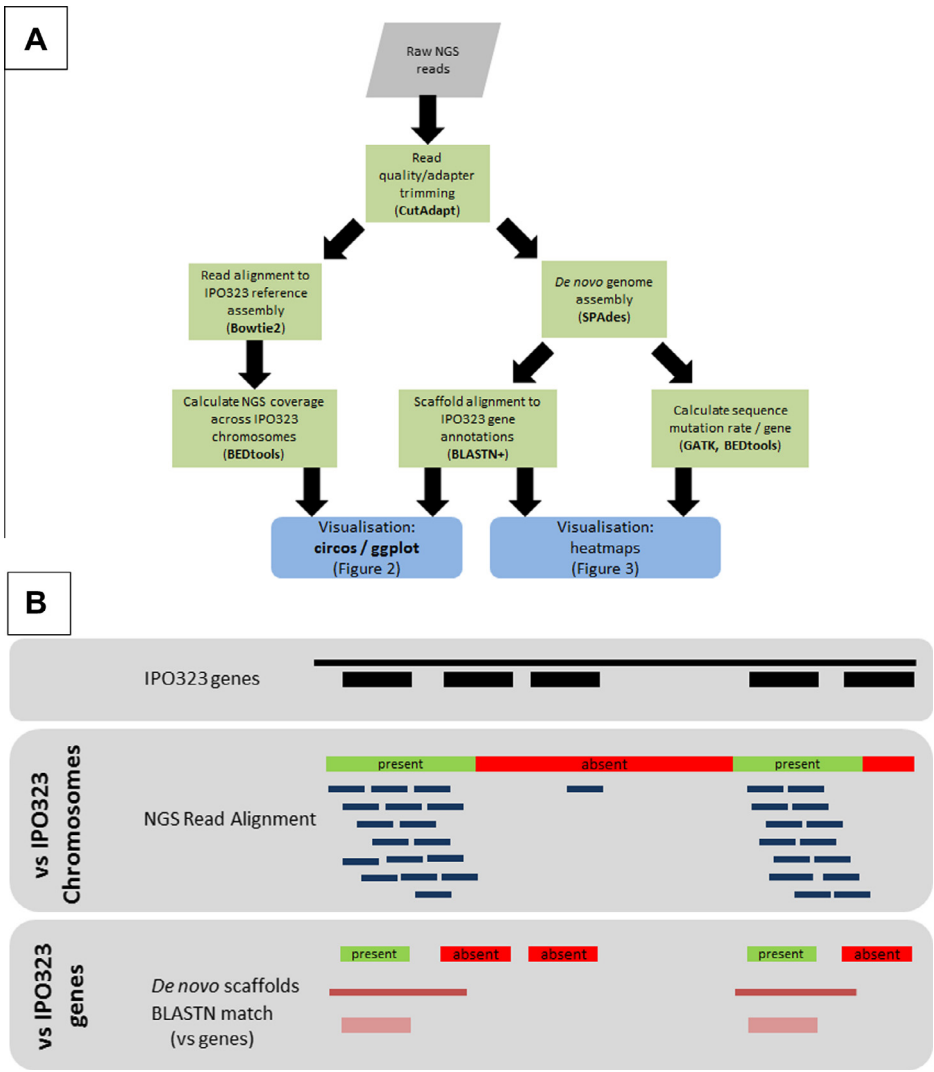

Fig. 1. Overview of techniques used to assess accessory chromosome presence and absence across multiple isolates of *Z. tritici* relative to the reference isolate IPO323. (A) Flow chart summarising the progression from raw NGS data from alternate isolates through to the visualisations presented in this study. (B) Example of differing presence/absence variation patterns detected by NGS read-alignment and *de novo* assembly-alignment based approaches.

the following PAV profiles: AC14 is absent in WAI323; AC15 is absent in WAI147, WAI322 and WAI328; AC16 is absent in WAI147 and WAI320; AC17 is absent in WAI322 and WAI320; AC18 is absent in all novel isolates; AC19 appears to be absent in WAI329, AC20 is clearly absent in WAI56 and possibly WAI147; AC21 is absent in WAI56, WAI147, WAI320, WAI324 and WAI329 (Figs. 2 and 3). In Fig. 2, it should be noted that percent coverage values may be >0% where ACs are genuinely absent, due to misalignment of short reads to the reference AC sequences, often in regions of repetitive DNA. However the decreased coverage indicative of PAV in ACs is clearly distinguishable when compared to coverage values for core chromosomes 1–13.

At the gene-level, analysis of NGS alignments to IPO323 gene regions indicate that ACs 14–21 are typified by distinctive patterns of presence and absence across various isolates and variable (often increased) rates of gene mutation (Fig. 3, Supplementary Table 1). In particular we observe AC17 and AC21 to frequently contain the highest rate of mutations per gene across most isolates. The PAV patterns observed at the chromosome-level are also strongly corroborated by gene-level PAVs (summarised by chromosome in Figs. 2 and 3 and Supplementary Table 2), which exhibit far lower percentage coverage (due to reduced background) for the aforementioned ACs as these are restricted to genes and exclude DNA repeats. This gene-level information can also be readily adapted for the purpose of discovering novel pathogenicity genes, which have a PAV or mutation profile that correlates with a known phenotype across isolates of *Z. tritici*.

Together these two methods can confidently predict both whole-chromosome absence (via read mapping) as well as single gene deletions in both core chromosomes and ACs. However these methods are not without their limitations. Due to the higher percentage of repetitive sequence content in ACs, it is challenging to accurately determine the exact sequences or break points in absent sequences, especially when mapping NGS data. Similarly using only short-read NGS data, *de novo* genome assemblies will remain highly fragmented, which limits the discovery of "novel" complete ACs and/or genes. Improvements to this method could involve limiting spurious read mappings to repetitive regions by initial masking of repeat families – reviewed in (Jurka et al., 2011). Alternatively, pre-existing annotations of repetitive regions for isolate IPO323 (Dhillon et al., 2014) could be used to exclude these regions with BEDtools (Quinlan and Hall, 2010), in order to remove unreliable regions containing known repeats from consideration.

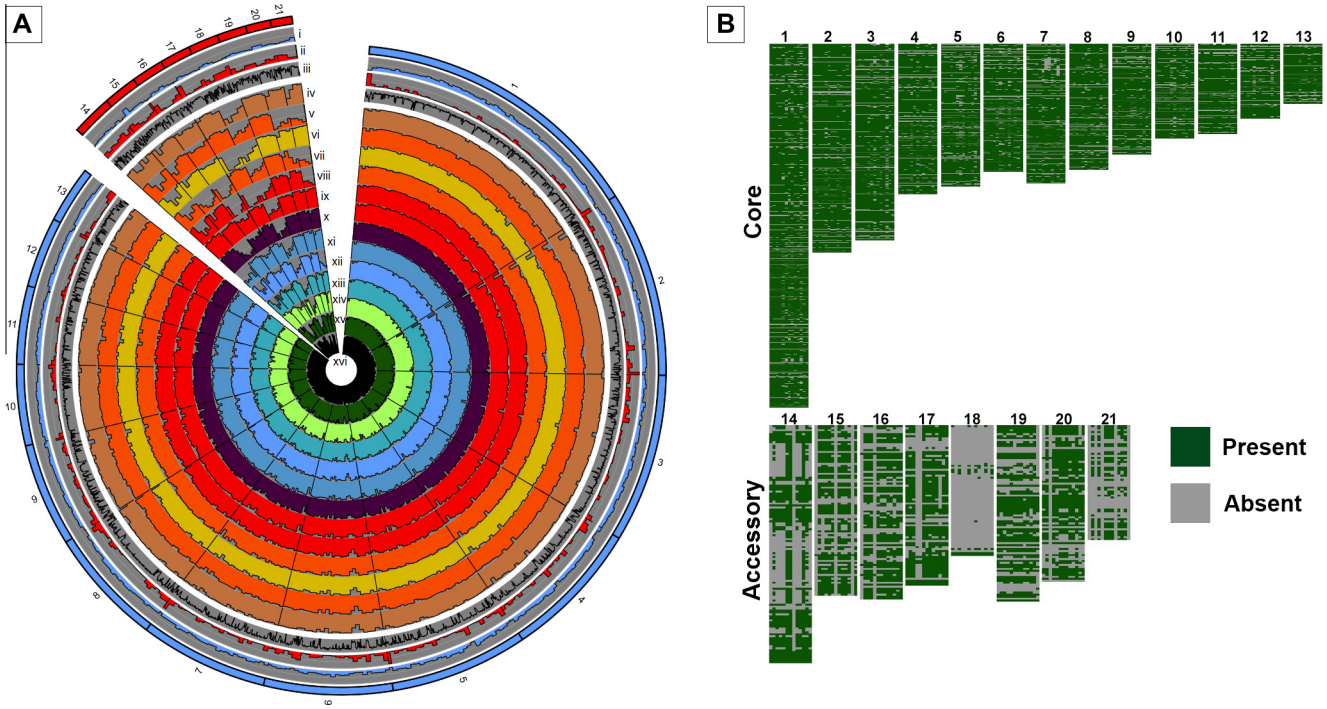

Fig. 2. Summary of *Z. tritici* chromosome comparative genomics across multiple isolates relative to the reference isolate IPO323. (A) Circos representation of presence/absence variation relative to IPO323 chromosomes (core = blue, dispensable = red), displaying the percentage (0–100%) of 100 kb windows containing: (i) gene-coding regions, (ii) repetitive DNA (via RepeatMasker/Repbase ("fungi")), (iii) %G:C content; % of window covered by aligned NGS reads (>10×) to isolate, (iv) WAI221, (v) WAI56, (vi) WAI332, (vii) WAI147, (viii) WAI320, (ix) WAI321, (x) WAI322, (xi) WAI323, (xii) WAI324, (xiii) WAI326, (xiv) WAI327, (xv) WAI328 and (xvi) WAI329. (B) R ggPlot representation of the presence or absence of coding genes across novel isolates relative to IPO323 core and accessory chromosomes. Only genes annotated in IPO323 are shown – arranged in matrices according to their order on their respective chromosomes – in which a gene is represented by one block in green if present or grey if absent. Genes were defined as present if 50% of their total length was covered by the top BLASTN match versus the *de novo* assembled genomes of thirteen isolates, sorted in columns from left to right: WAI147, WAI221, WAI56, WAI332, WAI320, WAI321, WAI322, WAI323, WAI324, WAI326, WAI327, WAI328 and WAI329.

IPO323 Chromosome	Chr. Length	# Genes	Genes / Mb	Average gene length (Kb)
1	6,088,797	1,980	325	1.6
2	3,860,111	1,136	294	1.7
3	3,505,381	1,071	306	1.6
4	2,880,011	821	285	1.6
5	2,861,803	778	272	1.7
6	2,674,951	692	259	1.6
7	2,665,280	766	287	1.5
8	2,443,572	689	282	1.6
9	2,142,475	604	282	1.6
10	1,682,575	516	307	1.7
11	1,624,292	488	300	1.6
12	1,462,624	408	279	1.6
13	1,185,774	330	278	1.5
14	773,098	114	148	1.3
15	639,501	86	135	1.7
16	607,044	88	145	1.4
17	584,099	78	134	1.2
18	573,698	64	112	1.4
19	549,847	87	158	1.2
20	472,105	79	167	1.4
21	409,213	58	142	1.4

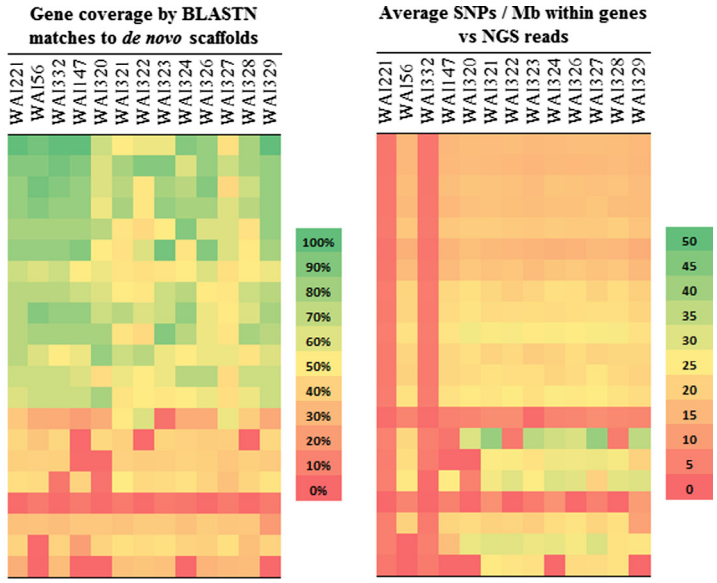

Fig. 3. Heat-map summary of gene-based sequence comparisons across *Z. tritici* isolates relative to the reference isolate IPO323. Accessory chromosomes of IPO323 have been italicised. The differences in various characteristics between core and accessory chromosomes are summarised, including chromosome length, gene density and mutation rates across novel isolates. Genome sequences of novel isolates were *de novo* assembled using SPAdes. Presence of accessory chromosomes was indicated by the percentage of IPO323 protein-coding genes on each chromosome that were ⩾50% conserved based on BLASTN matches across the length of IPO323 genes. Mutation rates within gene regions were based on BAM alignments of NGS reads and determined via GATK, relative to the genes of IPO323 and presented here as an average across genes on each chromosome.

4. Conclusion

In summary, the application of low-cost NGS and simple bioinformatic workflows can be a powerful tool for rapidly assaying PAVs and mutation rates across novel *Z. tritici* isolates, relative to the core and accessory chromosomes of the reference isolate IPO323. Thirteen isolates, which vary in their pathogenicity profiles on wheat, have been presented in this study as examples illustrating the application of these techniques. The correlation of these profiles with PAV and mutation data will be further investigated in subsequent studies.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fgb.2015.04.012.

References

Altschul, S.F. et al., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Bankevich, A. et al., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455–477.

Brunner, P.C. et al., 2013. Coevolution and life cycle specialization of plant cell wall degrading enzymes in a hemibiotrophic pathogen. Mol. Biol. Evol. 30, 1337–1347.

Camacho, C. et al., 2009. BLAST+: architecture and applications. BMC Bioinformatics 10, 421.

Coleman, J.J. et al., 2009. The genome of *Nectria haematococca*: contribution of supernumerary chromosomes to gene expansion. PLoS Genet. 5, e1000618.

Croll, D., McDonald, B.A., 2012. The accessory genome as a cradle for adaptive evolution in pathogens. PLoS Pathogens 8, e1002608.

Croll, D. et al., 2013. Breakage-fusion-bridge cycles and large insertions contribute to the rapid evolution of accessory chromosomes in a fungal pathogen. PLoS Genet. 9, e1003567.

DePristo, M.A. et al., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43, 491–498.

Dhillon, B. et al., 2014. The landscape of transposable elements in the finished genome of the fungal wheat pathogen *Mycosphaerella graminicola*. BMC Genomics 15, 1132.

Goodwin, S.B. et al., 2011. Finished genome of the fungal wheat pathogen *Mycosphaerella graminicola* reveals dispensome structure, chromosome plasticity, and stealth pathogenesis. PLoS Genet. 7, e1002070.

Grigoriev, I.V., 2013. A changing landscape of fungal genomics. Ecological Genomics of Fungi. John Wiley & Sons, Inc., Hoboken, NJ, pp. 1–20.

Grigoriev, I.V. et al., 2013. MycoCosm portal: gearing up for 1000 fungal genomes. Nucl. Acids Res., gkt1183

Jurka, J. et al., 2011. Repetitive elements: bioinformatic identification, classification and analysis. eLS.

Kellner, R. et al., 2014. Expression profiling of the wheat pathogen *Zymoseptoria tritici* reveals genomic patterns of transcription and host-specific regulatory programs. Genome Biol. Evol. 6, 1353–1365.

Kersey, P.J. et al., 2014. Ensembl genomes 2013: scaling up access to genome-wide data. Nucl. Acids Res. 42, D546–D552.

Krzywinski, M. et al., 2009. Circos: an information aesthetic for comparative genomics. Genome Res. 19, 1639–1645.

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359.

Li, H. et al., 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079.

Ma, L.J. et al., 2010. Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. Nature 464, 367–373.

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. J. 17, 10–12.

McDonald, M.C. et al., 2015. Recent advances in the *Zymoseptoria tritici*–wheat interaction: insights from pathogenomics. Front. Plant Sci., 6

McKenna, A. et al., 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303.

Morais do Amaral, A. et al., 2012. Defining the predicted protein secretome of the fungal wheat leaf pathogen *Mycosphaerella graminicola*. PLoS One 7, e49904.

Perez-Nadales, E. et al., 2014. Fungal model systems and the elucidation of pathogenicity determinants. Fungal Genet. Biol. 70, 42–67.

Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842.

Stukenbrock, E.H. et al., 2010. Whole-genome and chromosome evolution associated with host adaptation and speciation of the wheat pathogen *Mycosphaerella graminicola*. PLoS Genet. 6, e1001189.

Stukenbrock, E.H. et al., 2011. The making of a new pathogen: insights from comparative population genomics of the domesticated wheat pathogen *Mycosphaerella graminicola* and its wild sister species. Genome Res. 21, 2157–2166.

Testa, A., Oliver, R., Hane, J., 2015. Overview of genomic and bioinformatic resources for *Zymoseptoria tritici*. Fungal Genet. Biol. 79, 13–16.

Torriani, S.F. et al., 2011. Evolutionary history of the mitochondrial genome in *Mycosphaerella* populations infecting bread wheat, durum wheat and wild grasses. Mol. Phylogenet. Evol. 58, 192–197.

Van der Auwera, G.A. et al., 2013. From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. Curr. Protocols Bioinformatics 11, 11 10 1–11 10 33.

Yang, F. et al., 2013a. Transcriptional reprogramming of wheat and the hemibiotrophic pathogen *Septoria tritici* during two phases of the compatible interaction. PLoS One 8, e81606.

Yang, F. et al., 2013b. Battle through signaling between wheat and the fungal pathogen *Septoria tritici* revealed by proteomics and phosphoproteomics. Mol. Cell. Proteomics 12, 2497–2508.