Long read assembly of the pandemic strain of *Austropuccinia psidii* (myrtle rust) reveals an unusually large (gigabase sized) and repetitive fungal genome.

[1,2]Peri A Tobias* (peri.tobias@sydney.edu.au), [3]Benjamin Schwessinger* (benjamin.schwessinger@anu.edu.au), [4]Cecilia H Deng (cecilia.deng@plantandfood.co.nz), [4]Chen Wu (chen.wu@plantandfood.co.nz), [5]Chongmei Dong (chongmei.dong@syney.edu.au), [6]Jana Sperschneider (jana.sperschneider@anu.edu.au), [3]Ashley Jones (ashley.jones@anu.edu.au), [7]Grant R Smith (grant.smith@plantandfood.co.nz), [8]Josquin Tibbits (josquin.tibbits@ecodev.vic.gov.au), [9]David Chagné (david.chagne@plantandfood.co.nz, [5]Robert F Park (robert.park@sydney.edu.au)

* authors for correspondence

Affiliations

[1] School of Life and Environmental Sciences, University of Sydney, Camperdown, NSW 2006, Australia, [2] Plant & Food Research Australia, Sydney, Australia, [3]Australia Research School of Biology, the Australian National University, Acton, ACT 2601, [4]The New Zealand Institute for Plant and Food Research Limited, Private Bag 92169, Auckland 1142, New Zealand, [5]Plant Breeding Institute, University of Sydney. Private Bag 4011 Narellan, NSW 2567, Australia, [6] Biological Data Science Institute, The Australian National University, Canberra, ACT, 2600, Australia, [7]The New Zealand Institute for Plant and Food Research Limited, Private Bag 4704, Christchurch Mail Centre, Christchurch 8140, New Zealand, [8]Agriculture Victoria

Department of Jobs, Precincts and Regions, Bundoora, Vic 3083, Australia, [9]The New Zealand Institute for Plant & Food Research, Private Bag 11600, Palmerston North 4442, New Zealand

Abstract

Austropuccinia psidii, originating in South America, is a globally invasive plant pathogen causing rust disease on Myrtaceae. Several biotypes are recognized with the most widely distributed pandemic strain spreading throughout the Asia-Pacific and Oceania regions within the last decade. *Austropuccinia psidii* has a broad host range (currently 480 myrtaceous species), making it a particularly dangerous plant pathogen. In the nine years since the pandemic biotype was first found in Australia in 2010, the pathogen has caused the near extinction of at least three species, the decline of at least one keystone species, and negatively affected commercial production of several Myrtaceae. To enable molecular and genomic studies into *A. psidii* pathogenicity, we assembled a highly contiguous genome for the pandemic biotype based on PacBio sequence data and scaffolding with Hi-C technology. With an estimated haploid genome size of just over 1 Gbp, it is the largest assembled fungal genome to date. We found the *A. psidii* genome to have a lower GC content (33.8 %), greatly expanded telomeric and intergenic regions and more repetitive regions (> 90 %) compared to other genomes of species in the Pucciniales, however numbers of predicted coding regions (18,875) are comparable. Most of the increase in genome size is caused by a recent expansion of transposable elements belonging to the Gypsy superfamily. Post-inoculation mRNA sequence capture from a susceptible host provides expression evidence for 10,613 predicted coding genes, including 210 of the 367 putative effectors. The completeness of the genome provides a greatly needed resource for strategic approaches to combat disease spread.

Key words: myrtle rust, pandemic strain, long read, genome assembly, Myrtaceae, transposable elements

Background

The globally invasive fungal plant pathogen, *Austropuccinia psidii* (G. Winter) Beenken *comb. nov.* [1], was first reported in 1884 by Winter [2] in South America on guava (*Psidium guajava*) and named *Puccinia psidii*. This pathogen is the causal agent of rust disease on Myrtaceae, with guava rust, eucalyptus rust, ‘ōhi’a rust and myrtle rust used as disease names. Several biotypes are recognized [3] with only the pandemic strain [4] currently believed to be present in the Asia-Pacific and Oceania regions [5,6]. Although only a relatively recent arrival in these geographic regions: for example it was first detected in Hawaii (2005), China (2009), Australia (2010), New Caledonia (2013) and New Zealand (2017) [7], the disease is rapidly spreading and causing devastating impacts to natural vegetation communities [8,9]. *Austropuccinia psidii* is a biotrophic pathogen and has a macrocyclic life cycle but predominates in the asexual, dikaryotic (two haploid nuclei in one cell) urediniospore state [10]. Germination of wind-borne urediniospores requires the presence of free water, darkness and optimal temperatures between 15 and 25˚ C, with cuticular penetration occurring in susceptible hosts within 12 hours [11]. During extended cold periods (~8 □) the pathogen can overwinter in a dormant state within the host plant [12]. The pathogen causes disease symptoms on the new foliage, stems and buds [13,14] of a wide range of perennial plants (480 known susceptible species globally)[15]. Repeated infections can lead to crown loss and tree mortality [16]. Globally, there are 5,950 described myrtaceous species [17], of which 2,250 are endemic to Australia, where myrtle rust has been particularly damaging [18]. For example, two previously wide-spread east Australian

species; *Rhodamnia rubescens*, *Rhodomyrtus psidioides*, and one recently described species, *Lenwebbia* sp. Main Range have now been listed as critically endangered owing to repeated infection causing death [16,19]. Myrtle rust has also caused the decline of at least one keystone species, *Melaleuca quinquenervia* [16], and affected commercial production of Myrtaceae such as the tea-tree (*Melaleuca alternifolia*) and lemon myrtle (*Backhousia citriodora*) [7]**.**

Critical to combatting this pathogen is understanding the mechanisms it uses to parasitize its many hosts. Successful fungal infection is known to be caused by a suite of secreted proteins known as effectors, that manipulate susceptible hosts and debilitate plant defenses [20]. Prediction of the effector complement and studies into the biology of the infection process have been hampered by the lack of a high-quality *A. psidii* genome. An early draft assembly using short sequence reads of the pandemic biotype of *A. psidii*, predicted the genome size at 103 – 145 Mbp [21]. More recently the draft assembly of the South African biotype using linked-reads sequence data (Chromium technology, 10x Genomics, Pleasanton, CA, USA) [22] suggested a haploid genome size of 1.2 Gbp [23]. To address the discrepancy in the published genome size for this pathogen, and to improve on the previous *A. psidii* assemblies, we optimized the use of Pacific BioScience SMRT® cells (PacBio, Menlo Park, California, USA) for long read sequencing. We produced a highly contiguous genome assembly followed by scaffolding the contigs with chromosomal conformation capture sequencing, Hi-C technology [24]. The haploid genome for *A. psidii* is here determined at 1.02 Gbp over 66 contigs and is deposited at the European Nucleotide Archive (ENA) at EMBL-EBI under the following project: PRJEB34084.

The *A. psidii* genome is particularly large compared with the average fungal genome, around 44.2 Mbp [25], and relative to its Myrtaceae hosts, around 300 - 650 Mbp [26,27]. With the resource of a highly contiguous *A. psidii* genome assembly, we computationally annotated

18,875 predicted coding regions. Post-inoculation mRNA sequence capture from a susceptible host provided evidence for expression of 10,613 predicted coding genes, including 210 of the 367 putative effectors. We determined a high proportion of repetitive sequences (> 90 %), predominated by Class I long terminal repeat (LTR) transposable elements of the Gypsy superfamily [28]. Telomeric regions appear expanded in *A. psidii* (>1000 nt in 21 scaffolds) and predicted telomerase transcripts are present during infection indicating a role in early pathogenesis. While predicted gene numbers appear comparable to other genomes of species in the Pucciniales we found greatly expanded intergenic regions. The overall GC content is 33.8 %, with 41.5 % present in coding sequences, making *A. psidii* a particularly AT-rich genome. The high contiguity of the genome, combined with the unique genomic features identified in this study, provide essential resources to understand and address the spread of this invasive plant pathogen.

Analyses and results

Biological sample from an Australian isolate

Austropuccinia psidii urediniospores, from a single pustule isolate, were used for all workflows including flow cytometry, Hi-C libraries and high molecular weight DNA extraction, as described in the methods of this manuscript.

Predicting genome size with flow cytometry and kmer analysis

To guide the quantity of long read sequence required to assemble the *A. psidii* genome, the nuclear DNA content of *A. psidii* cells was estimated using flow cytometry. Details for all experimental approaches are described in the methods section of this manuscript. Flow cytometry estimated the 1C (single nuclei) genome content at 1.42 pg (1,388.91 Mbp) using the

host genome (*Eucalyptus grandis*) as the internal reference. As an additional approach, we ran a kmer analysis using Jellyfish (v.2.2.6), to predict the diploid genome size at 2,100 Mbp. From these results we estimated a haploid genome of around 1 Gbp and sequenced first 21 and then a further 8 PacBio SMRT® cells (18 X RSII plus 11 X Sequel).

Long read genome assembly with Canu (v1.6)

After converting RSII data to conform with the Sequel output of subreads.bam, we obtained 162 GigaBytes (GB) of data. Details on data pre-processing and assembly are available in the methods section of this manuscript and scripts are available at the myrtle-rust-genome-assembly github repository [29]. An initial diploid assembly with Canu (v1.6), using data from 21 SMRT cells, produced an assembly size of 1.52 Gbp (Table 1). The contig numbers for this assembly were high and completeness was low (79.8% complete and 7.0% fragmented) based on Benchmarking Universal Single-Copy Orthologs (BUSCO) (v3.1.0) in genome mode [30]. We therefore sequenced eight additional SMRT cells and assembled the diploid genome of *A. psidii* at 1.98 Gbp, with 13,361 contigs, using Canu (v1.6) [31] long read sequence assembler. The Canu corrected and trimmed reads, at 35 times coverage, were used in the final construction of contigs (Table 1).

Table 1. Comparative statistics for the *Austropuccinia psidii* raw sequence and assembly data from 21 and then 29 PacBio SMRT® cells using diploid parameters in Canu v1.6 [31]**.**

Raw data statistics	21 SMRT® cells	29 SMRT® cells
Raw read numbers	2 696 004	5 798 431
Raw bases	36 099 416 447	72 410 490 565
Corrected mean length (bp)	11 733	21 811
Corrected N50 (bp)	47 581	46 751
Corrected and trimmed reads	2 048 322	1 853 980

Corrected and trimmed (bp)	23 774 178 167	35 041 589 693
Assembly statistics		
Assembly parameters	diploid	diploid
Node (contig) count	22 474	13 361
Edge count	762 342	512 781
Total length (bp)	1 520 827 096	**1 985 401 333**
N50 (bp)	96 784	287 498
Shortest (bp)	1 004	1 003
Lower quartile (bp)	27 605	37 228
Median node (bp)	47 237	75 798
Upper quartile node (bp)	83 201	178 631
Longest node (bp)	1 037 919	2 557 042
Assembly time (aprox.)	2.5 months	4.5 months

We deduplicated the assembly contigs with the Purge Haplotigs pipeline [32] to produce primary contigs and secondary haplotigs of ~ 1 Gbp each (Table 2) as well as contigs labelled as assembly artefacts (34 MegaBytes). We then generated scaffolds from the assembly contigs using Hi-C [24] generated read data with ALLHIC (v.0.8.11) [33]. Initially, we produced 40 primary scaffolds but found several telomeres from the pre-scaffolded assembly embedded within scaffolds. We therefore curated to break these scaffolds at the start/end of telomeres for a final number of 66 primary (67 secondary) scaffolds. The primary genome assembly file, named APSI_primary.v1.fasta, has been deposited under the following ENA accession: ERZ1194194 (GCA_902702905). Locus tags are registered as APSI (for *Austropuccinia psidii*) and scaffolds identified as APSI_Pxxx (where x indicates scaffold number) for primary assembly. The diploid assembly file (APSI_v1.fa) is available from DOI:10.5281/zenodo.3567172 and incorporates the 67 secondary assembly scaffolds (APSI_Hxxx for secondary/haplotigs).

Table 2. Statistics for the Canu (v1.6) [31] ***A. psidii*** **primary and secondary assemblies after deduplication with Purge Haplotigs** [32] **(A) before (B), after Hi-C scaffolding.**

(A) Assembly	contigs	N50	L50	Largest	Bases	GC%
Primary	3 187	520 311	619	**2 559 506**	**1 018 086 732**	**33.80**
Secondary	8 626	159 727	1794	957 974	933 887 333	34.26

(B)	scaffolds	N50	L50	Largest	Bases	GC%
Primary	66	56 243 252	7	**124 273 364**	**1 018 398 822**	33.80
Secondary	67	52 409 407	7	89 073 602	934 744 333	34.26

We polished the primary assembly using the genomic consensus algorithm *arrow*, as described in the methods and github [29]. This process was repeated twice and genome completeness was assessed with BUSCO (v3.1.0) [30] and identified the presence of 94.7% conserved proteins for the primary assembly and 97.6% for the combined primary and secondary. The BUSCO analysis was run using protein mode following gene prediction and annotation, described in the methods.

The genome size expansion is driven by transposable elements

Our analysis of the *A. psidii* genome assembly provides evidence for very large and diverse repetitive regions that constitute greater than 91 % (932,575,815 bp) of the genome. When we repeat-masked the APSI_primary and secondary genomes with published fungi and viridiplantae repeat databases (in RepBase) [34], we found very low percentages of matches (Table 3). The *A. psidii* genome is largely novel repeat regions, prompting further analysis with the REPET pipeline [35] to characterise transposable elements (TEs).

Table 3. Repeat masking the *Austropuccinia psidii* (APSI) genome using the RepBase library databases for fungi and viridiplantae indicates that the genome is composed of novel repeats.

Masked with library	fungi	viridiplantae
% of genome masked in APSI	19.63% (19.89 %)	6.97% (6.99 %)
% of genome masked as Retroelements	17.3% (17.71 %)	4.91% (5.01 %)
% of genome masked as Gypsy/DIRS1	16.63% (17.14 %)	4.31% (4.44 %)

Using the REPET pipeline, we show that the massive repeat expansion in *A. psidii* is mostly driven by TEs belonging to the Gypsy superfamily (Figure 1). Most of the expansion was driven by a limited amount of TE families at discrete time points (Figure 2). For example, the extensive expansion observed around 78% family level identity is mostly driven by a single TE family. This is similar to early expansions (30-50% family level identity) which are driven by only a limited amount of TE families and in contrast to the most recent expansion (80-90% family level identity), which is driven by a more diverse set of TE families including several Class II superfamilies, such as those belonging to the TIR order.

Figure 1. *Austropuccinia psidii* has a high repeat content driven by expansion of the Gypsy superfamily. Repetitive element annotation on primary scaffolds. Top panel show percentages of genome coverage for all repetitive elements and different subcategories with some overlap. These include transposable elements (TEs) of class I (RNA retrotransposons) and class II (DNA transposons), simple sequence repeats (SSR), and unclassifiable repeats (no Cat). Bottom panels show percentages of genome coverage of Class I and Class II TEs categorized to class, order, and superfamily levels wherever possible. Repetitive elements were identified using the REPET pipeline, and classifications were inferred from the closest BLAST hit (see Materials and Methods). The TE order is color coded in each Class I and Class II TE plot.

Figure 2. A subset of transposable element (TE) superfamilies have driven the genome expansion of *A. psidii*. The blue line indicates the mean TE family percentage identity distribution relative to the consensus sequence of TE families as a proxy of TE age. Individual points indicate the relative frequency of a specific TE family plotted at their mean percentage identity relative to the consensus sequence. Points are color coded according to the TE superfamily. Only highly abundant TE families are included in the plot.

The high percent of TEs and the overall low GC content prompted us to investigate the presence of AT-rich region in the *A. psidii* genome. We identified AT-rich regions using OcculterCut [36], a program specifically designed to identify AT-isochore. The major peak identified is around 33% GC content and a second peak around 41% GC content (Figure 3A). The former largely overlaps with the GC content profile of TEs. The second peak overlaps with the GC content of genes in the *A. psidii* genome. This specific AT-isochore structure with two separate genome compartments appears to be specific to *A. psidii*, as the closely related *Puccinia striiformis* f. sp. *tritici* displays only a single peak at around 45% GC content (Figure 3B). We further investigated the GC-content for two scaffolds, APSI_P025 (8 512 731 bp) and APSI_P027 (3 316 254 bp). Plotting of gene density and repeat density shows a lack of repeat- or gene-rich islands, but rather accumulation of TEs between genes (Figure 3C).

Figure 3 (A) *Austropuccinia psidii* displays two distinct genome compartments in relation to GC content. Relative GC content of genome regions identified by genes, transposable elements (TEs), 1 kb sliding windows, or as identified by OcculterCut. (B) AT-isochores are specific to *A. psidii*. Relative GC content of genome regions identified by OcculterCut in *A. psidii* and *Puccinia striiformis f. sp. tritici*. (C) Karyoplots of scaffolds APSI_025 and APSI_P027. Gene and repeat density are shown at 20,000 bp windows. Mean GC-content of 33.8% is shown with a red line. OcculterCut regions of GC-content segmentation are shown as black lines.

Extended telomere repeats identified

Prior to scaffolding the assembled contigs, we identified 29 telomeric regions at the start or end of primary contigs (and 23 in the secondary contigs), based on the hexamer TTAGGG(n) [37]. We used these numbers to guide scaffolding and checked the telomere locations, post scaffolding. Three primary scaffolds; APSI_P012(37 965 047 bp), APSI_P025 (8 512 731 bp), APSI_P027 (3 316 254 bp), have telomeric regions at both the start and end indicating potential complete chromosomes. The final hexamer repeat numbers and lengths (Table 4) indicate extraordinarily long telomeres compared to other studied rust fungi [37,38]. Previous studies indicate that telomere localisation in the nucleus show affinity to the nuclear membrane [39], therefore cross-linking during our Hi-C library preparation may have proximity-ligated adjacent telomeres leading to errors in scaffolding. Manual curation before and after scaffolding corrected for this problem by identifying and breaking at embedded telomeres. With no current cytological evidence for the number of *A. psidii* haploid chromosomes to guide scaffolding, we made predictions for scaffold grouping based on telomeres and evidence from a related rust fungus, *Puccinia graminis* f.sp. *tritici*, (n = 18) [40].

Table 4. Telomere repeat regions identified for the *Austropuccinia psidii* primary assembly before and after scaffolding.

n(TTAGGG)	>1000	800-999	500-799	90-499	<25	Total
Av. Length (nt)	8 159	5 223	3 754	1 409	87	
Pre-scaffold	6	6	4	8	5	29
Post-scaffold	10	3	2	9	5	29

Numbers and average length in nucleotides (nt) of contigs/scaffolds with n(TTAGGG) sequences. n = number of hexamers, for example >1000 means more than 1000 x (TTAGGG) or more than 6000 nt.

RNA used for genome annotation

In order to annotate the genome, we extracted total RNA from infected young leaves of the susceptible host, *Syzygium jambos,* and harvested at 6, 12, 24, 48 hours (h), 5, and 7 days (d) post-inoculation with *A. psidii*. RNA-seq data were generated for the six time points after inoculation and the assembled transcripts were merged for a total of 80,804 representative genes and 108,659 alternative forms. Of the total transcripts, 8% and 9% were classified as fungi in the main and alternative gene sets respectively (Additional Figure 1.HTML).

RNA-seq alignments showed that the majority of reads were from the host plant and a minor percentage of data were from the fungal isolate. Mapping the RNA reads to the assembled *A. psidii* genome, the lowest overall mapping rate was observed in sample 6 h (0.77%) and the highest in 7 d (8.58%). Mapping to the *Metrosideros polymorpha* (ōhi□a) plant genome, the most closely related publicly available genome to the inoculation host plant, the overall mapping rate was stable and greater than 76.7% in all samples except 7 d (which was 71.55%).

Gene prediction and functional annotation

We used Braker2 [41] and RNA-seq data to computationally predict 18,875 (15,684) protein coding genes within the primary and secondary assemblies respectively (Table 5). We then used SignalP (v.4.1f)[42], followed by EffectorP (v.2.0) [20] to identify putative effector genes within the primary and secondary assemblies at 367 and 304 respectively. These putative effectors did not appear to be compartmentalized and had similar intergenic distances when compared to BUSCOs or all genes (Figure 4). We also ran the predicted proteins through ApoplastP [43] to identify sequences that are likely to be located to the apoplast within host plants. Predicted effectors from the primary assembly were visualized on scaffolds with MapChart v.2.3.2 [44] (Additional Figure 2) and expression validation, based on mapping reads to the primary assembly with Star (2.7.2b), was found for 10,613 sequences, including 210 predicted effectors. Of the predicted effectors in the primary and secondary assemblies, exact amino acid duplicates were found for 97 sequences. Annotation files and multi-sequence fasta files for predicted coding sequences, amino acid sequences and effectors for the combined primary and secondary assemblies are available at DOI:10.5281/zenodo.3567172.

Table 5. Structural and functional annotation of the primary and secondary *A. psidii* assemblies based on Braker2 (Hoff et al., 2016), SignalP [42]**, ApoplastP, EffectorP** [20]**. Numbers in brackets have expression evidence from mapped RNA-seq data using Star (2.7.2b).**

Genome	protein coding genes	Secreted	Apoplastic	Predicted effectors	Non-apoplastic effectors
Primary	18 875 (10 613)	1 157	374	367 (210)	194
Secondary	15 684	925	254	304	183

Figure 4. *Austropuccinia psidii* putative effectors are not found in gene sparse regions. Nearest-neighbor gene distance density hexplots for three gene categories including all genes, BUSCOs, and candidate effectors. Each subplot represents a distance density hexplot with the log10 3′-flanking and 5′-flanking distance to the nearest-neighboring gene plotted along the x axis and y axis, respectively.

Pucciniales orthologues and comparative statistics

We did a comparative analyses of fungi within the order Pucciniales with predicted proteins using Orthofinder (v2.2.7) [45]. We included protein sequences from two tree rust species, *Cronartium. quercuum* (white pine blister rust) [46] and *Melampsora larici-populina* (poplar rust) [47], to test whether protein orthologues may relate to life history of the host. We visualized the evolutionary phylogeny, using the species rooted tree based on orthologs, with Dendroscope (v3) [48]. Based on a species rooted phylogenetic tree from the multiple sequence alignment of single copy orthologues, *A. psidii* was placed closer to the cereal rusts than to tree rust species (Figure 5A). This finding is in accordance with previous analyses using ribosomal DNA and cytochrome c oxidase subunit 3 (CO3) of mitochondrial DNA to determine the type specimen for *A. psidii* [1]. A graphic presentation of the orthologues shows shared orthologroups between Pucciniales species (Figure 5B) indicating proteome closeness. Our orthologue analysis with other Puccinales species placed the *A. psidii* proteins closer to the cereal rusts phylogenetically (Figure 5A) however a group of 101 orthologues (Figure 5B, right) present in the tree rusts; *M. larici populina*, *C. quercuum* and *A. psidii* appear to be absent from the cereal rust species, perhaps indicating a common biological strategy for perennial host infection.

Figure 5. (A) Protein-protein species rooted tree based on multiple species alignment of orthogroups identified with Orthofinder [45]**. Scale represents substitutions per site and internal node values are species tree inference from all genes (STAG) supports** [49]**. (B) Protein-protein comparisons across rust fungal species with each concentric ring indicating a different species. Numbers external to the rings represent counts of orthologue groups and numbers within each concentric ring represents the number of orthologue genes in that species per section. The figure shows every possible combination of species included in this proteome ortholog analysis, using concentric circles graphically presents an overview of 'closeness' between the genomes. The species color code is consistent for Figures 5 and 6.**

Comparative statistics for the Pucciniales genomes show that the scaffolded assembly for *A. psidii* is more contiguous, more repetitive and larger than other assembled genomes (Table 6 and Additional Table 1.xlsx). A comparative analysis using annotation files downloaded from Mycocosm [50] and NCBI [50], reveal dramatically large expansions of intergenic lengths in *A. psidii*, while gene lengths, incorporating untranslated regions (UTRs), are similar across six species of the Pucciniales (Figure 6).

Table 6. Comparative assembly data across several available Pucciniales genomes. Common names indicated; myrtle rust (*Austropuccinia psidii* pandemic strain from the current study in bold), wheat stripe rust (*Puccinia striiformis* f. sp. *tritici*), wheat leaf rust (*P. triticina*), wheat stem rust (*P. graminis*), oat crown rust (*P. coronata*), maize common rust (*P. sorghi*), poplar rust (*Melampsora larici-populina*). *Predicted coding genes. nd=no data found.

Common name (rust)	**myrtle**	wheat stripe	wheat leaf	wheat stem	oat crown	maize	poplar
NCBI/ENA project	**PRJEB34084**	PRJNA396589	PRJNA36323	PRJNA18535	PRJNA398546	PRJNA277993	PRJNA242542
Genome (Mbp)	**1,018**	83	135	89	150	100	101
Scaffolds	**66**	156	14,818	393	1,636	15,715	462
N50	**56 243 252**	1 304 018	10 369	39 493	163 229	19 078	112 314
L50	**7**	57	2 866	557	241	1 530	265
Repeats (%)	**89**	54	40	nd	nd	nd	45

Coding genes*	**18 875**	15 928	15 685	15 979	26 323	21 078	16 372
GC (%)	**34**	44	37	44	45	45	41

Figure 6. Structural annotation comparisons of gene (A) and intergenic (B) length, incorporating untranslated regions, across six species of Pucciniales reveals dramatically large intergenic expansions in *A. psidii*.

Pandemic isolate variation identified in Australia

Whole genome DNA resequencing on the Illumina sequencing platform (paired-end 150 bp reads) was obtained from six other Australian isolates, one isolate from Hawaii believed to be from the pandemic strain, and a Brazilian isolate stored at the at the University of Sydney, Plant Breeding Institute (PBI). The DNA from these isolates was previously tested using simple sequence repeat markers and all Australian isolates were determined to be genetically similar [6]. When comparing pathogen isolates, we found 44,375 SNPs in total. On average one SNP per 23 Kbp. The maximum likelihood phylogenetic tree, in newick format, was visualized in Dendroscope v3 [48] (Figure 7) and indicates variations present in the pandemic strain of the pathogen. These isolates were collected from diverse hosts and locations in Australia dating back to 2011 (Table 7, Methods). Isolate Au 7, collected from the host plant *Melaleuca quinquenervia* in 2012, Queenscliff, NSW 2096, could not be separated from the Brazilian isolate with strong bootstrap values. Nevertheless, the clustering away from other Australian isolates is concerning and perhaps indicates that a variant is already present in Australia.

Figure 7. Phylogenetic evolutionary relationship based on 44,375 single nucleotide polymorphisms within pathogen isolates collected in Australia (Au), Hawaii (Hw) and outgroup from Brazil (Bz from *Eucalyptus grandis*) inferred using the Maximum Likelihood method with Iqtree v1.6.7 and visualized with Dendrosope v3. The tree is UNROOTED although outgroup taxon Bz is drawn at root. Numbers are standard bootstrap support (%). Au 7 and Bz could not be separated with strong bootstrap values. All other isolates cluster with Hw but bootstrap values between sub-branches are weak.

Discussion

Austropuccinia psidii, the causal agent of the pandemic myrtle rust, is a dangerous and rapidly spreading plant pathogen. It infects perennial hosts in the Myrtaceae family, including iconic trees such as eucalypts, and has an expanding global host list [15]. Here we present the first highly contiguous genome assembly for *A. psidii* using DNA from an Australian isolate of the pandemic strain previously assembled with short-read sequence data [21]. We assembled the genome using long read sequence data to produce the largest assembled fungal genome to date at 1.02 Gbp haploid size.

Why is the *Austropuccinia psidii* genome so big?

Data based on flow cytometry determined the largest fungal genome at ~3.6 Gbp [51] and the largest known rust-type (Basidiomycota, Pucciniales) fungal genomes; *Gymnosporangium confusum* and *Uromyces bidentis*, at 893.2 Mbp and 2,489 Mbp respectively [52]. Rust genomes appear to be larger on average than other fungal genomes (which are around 44.2 Mbp) [25] and rusts infecting Poaceae (monocotyledon) are significantly smaller (average size 170.6 Mbp) than those infecting Fabaceae (eudicotyledon) (average size 556.6 Mbp) hosts. Interestingly, genome size does not appear to correlate with the perennial nature of host plants with the assembled genome of two important tree rust pathogens, *Melampsora larici-populina* (101 Mbp) [47] and

Cronartium quercuum (22 Mbp) having modest-sized genomes. While ploidy has been postulated as a possible reason for large fungal genome sizes [51], the evidence to date is that genome expansion in rust species is due to repetitive elements, notably a proliferation of transposable elements (TEs) [47,53].

We annotated 18,875 predicted protein coding genes, comparable to other cereal rusts (Table 6 and Additional Table 2.xlsx). We determined that the extraordinary size of the *A. psidii* genome is not a product of gene expansion but is due to extensive repetitive DNA (> 90%) and very large intergenic regions (an order of magnitude greater when compared to six other Puccinales species) (Figure 6). Despite these large intergenic regions, we see no difference in the gene lengths for these species, further explaining that the increased genome size is related to non-genic regions that are largely repetitive. Our investigation into the nature of the repeat regions identified Gypsy superfamilies belonging to the LTR order as the major contributor to overall genomic expansion covering 73% of the genome (744,220,388 bp).

Active telomerase and longer telomeres

Telomeres are structural 5′-TTAGGG-3′ repeats that cap the physical ends of chromosomes and are important in maintenance of DNA integrity [37]. In human cell lines, where the telomerase holo-enzyme is not expressed beyond embryo stage, short telomeres are linked to ageing and telomerase activity is linked to cancer [54]. While telomeres can be from 5-15 Kbp in humans, fungal telomeres are generally around 240 bp in length [37,38]. We identified very long telomeres in the *A. psidii* assembly (more than 6 Kb for 10 of the scaffolds) and early infection transcripts (12 hours post inoculation) showed the presence of the predicted telomerase enzyme, indicating a putative role in maintaining telomeres during active growth. Annotations for telomerase reverse transcriptase (RNA-dependent DNA polymerase) and telomerase

ribonucleoprotein complex - RNA binding domains (PF12009 and PF00078) within a predicted *A. psidii* protein were found to be expressed in in both the primary and secondary assemblies (protein ID: APSI_P017.12297.t3 and APSI_H002.12793.t2).

A study in 2000 [55] identified and located the *Magnaporthe grisea* (rice blast fungus) *AVR-Pita* effector gene directly adjacent to a telomeric region on chromosome 3. The gene encodes a secreted effector with a zinc metalloprotease domain and mutations occur spontaneously, due to telomeric proximity, leading to loss of avirulence. Effectors were not predicted within subtelomeric regions within the *A. psidii* assembly but there are several uncharacterised predicted genes within 10kb on scaffold 27, a putative complete chromosome (Additional Figure 2). High expression levels were identified for one of these predicted genes (APSI_P027.780) and might warrant further investigation during pathogenesis. Also of interest, several subtelomeric regions were annotated, perhaps due to roles as RNA templates for telomeric DNA synthesis [56]. The presence of transcriptional activity of telomerase within infecting spores provides interesting mechanistic evidence for potential genome maintenance or even telomere expansion.

Isolate variation: a concerning threat to native flora

A preliminary analysis of SNP data across Australian *A. psidii* isolates, using a Hawaiian pandemic strain and a Brazilian isolate for comparison, has raised some concerns due to isolate variation within Australia. These results require further confirmatory studies with much deeper sequencing but, if validated, indicate that new incursions may have already occurred and/or that genetic variants are occurring in Australia. Changes in *A. psidii* virulence has been observed on species previously considered resistant or highly tolerant including *Acmena ingens* (syn. *Syzygium ingens*) and *Waterhousea floribunda*, with any symptoms observed

restricted to limited leaf spots [57]. More recent surveys (G Pegg, personal communication, 25 November 2019) have identified significant levels of infection on both these species resulting in severe foliage blighting and dieback caused by *A. psidii* infection. In the case of *W. floribunda*, specific trees that have not previously shown infection, despite other species in the area recorded as infected back in 2012, have now succumbed to infection. Similarly, mature *Syzygium corynanthum* trees that were considered to have no or low levels of infection in 2016 [16] now present extensive levels of dieback and are in severe decline. Several Australian plants that were previously wide-spread are now listed as threatened or critically endangered since the arrival of the pathogen in 2010 [19,58,59] and similar concerning impacts have been noted within New Caledonia [8]. Further studies and deeper coverage sequencing will clarify these results and enable proper containment of newly introduced and more virulent isolates.

Potential implications of the research

We have generated a highly contiguous reference genome assembly for the pandemic strain of *A. psidii* using long-read sequencing and chromosomal confirmation capture sequencing technology. The resource of a genome assembly from long reads has enabled the spanning of highly repetitive sequence regions, previously problematic to assemble with short read sequence data. We used *in vivo* expression data for genome annotation and have identified genome features such as extended telomeres and very extensive repetitive regions dominated by transposable elements. At 1.02 Gbp this is the largest fungal genome yet to be assembled and provides a foundation for future molecular research as well as providing a base-line resource to monitor the spread of the pathogen. It is expected that this genome resource will enable improved biosecurity management by monitoring pathogen populations to detect new incursions and

populations shifts and permit a deeper understanding of the molecular interaction between the pathogen and the plant host.

Methods

***Austropuccinia psidii* spores**

Austropuccinia psidii urediniospores were initially collected from the host plant *Agonis flexuosa* (willow myrtle) in 2011, Leonay, NSW, and and a single pustule isolate was increased on the susceptible host plant *Syzygium jambos* (rose-apple) [6]. Resulting spores (isolate ID: Au_3, culture no. 622 and accession 115012) were maintained in liquid nitrogen at PBI and used for all subsequent workflows including flow cytometry, Hi-C libraries and high molecular weight DNA extraction.

Nuclei size determination using flow cytometry

We estimated nuclear DNA content of *A. psidii* cells using flow cytometry. The fungal nuclei fluorescence intensity were compared with that of *Eucalyptus grandis*, as an internal standard, 2C=1.33pg, 640 Mbp [27,60]. Nuclei were extracted from cells using methods previously described [25] but with slight modifications. Approximately 50 mg of leaf material from *E. grandis* (internal standard) were chopped in a petri dish with a sterile blade in 1 mL of chilled, modified Wood Plant Buffer (WPB) [25] consisting of 0.2M Tris-HCl, 4 mM $MgCl_2$, 0.1% Triton X-100, 2 mM Na_2EDTA, 86 mM NaCl, 20 mM $Na_2S_2O_5$, 1% PVP-40, pH 7.5. New leaves were taken from pathogen free *E. grandis* plants and *E. grandis* plants with *A. psidii* infection symptoms one-week post-inoculation. Inoculated plants were treated using previously described methods [61]. Hence, the inoculated leaves were used to determine the *A. psidii* fluorescence peaks in relation to the un-inoculated leaves as the reference standard. RNase A (50

µg/mL) was added and samples incubated on ice for 15 minutes prior to centrifugation through 25 µm nylon mesh. Two DNA stains were added to the samples; 4,6-diamidino-2-phenylindole (5 µM) and propidium iodide (50 µg/mL) (DAPI and PI; Sigma-Aldrich, St. Louis, USA) and incubated at room temperature for 30-60 minutes. Fluorescence intensity was measured for 10, 000 events per sample, threshold rate of 370 events per second and flow rate medium, with BD FACSVerse™ using laser settings for PI and DAPI. A minimum of three replicates were run for the sample and the reference standard. Calculations were made using the reference standard (un-inoculated *E. grandis* leaves) compared to the inoculated sample. Histograms, using linear scale, were gated to reduce background fluorescence for each replicate and descriptive statistics obtained for the peaks using Flowing Software (v2.5.1)[62]. Medians with maximum coefficient of variation < 8.3 were determined. Calculations to determine genome size were based on taking the ratio of the median fungal fluorescence intensity (FI) divided by that of the internal standard, in this case uninfected leaves of *Eucalyptus grandis* [63]. A minimum of three replicates per sample were used. It was assumed that the dominant cell cycle stage (G1) FI peak for the rust fungus was the 1C content, due to the dikaryotic nuclei released from cells, whereas the dominant peak for the reference was 2C (Tavares et al. 2014). The nuclear DNA content in picograms (pg) was then multiplied by 978 Mbp [63] to estimate the genome size (1C) for the rust. Flow cytometry results indicated the 1C nuclei at 1.42 pg (1,388.91 Mbp).

Kmer genome size estimation

As an additional approach to predict the genome size of *A. psidii*, raw Illumina data (SRR1873509) [21] was downloaded from the National Centre for Biotechnology Information (NCBI). Data was run through Jellyfish (v.2.2.6) software [64] with kmer size of 16 and 21 and the following parameters: count -t 8 -C -m (16) 21 -s 5G. Histogram outputs (Jellyfish v.2.2.6)

were visualised and coverage determined at 3X. The diploid genome size was calculated at 2,100 Mbp by dividing the total number of kmers (21mer) by the mean coverage. From these results a haploid genome of around 1 Gbp was predicted.

HMW DNA extraction and PacBio sequencing

For long read sequencing a modified CTAB high molecular weight DNA extraction procedure [65] was used with the modification of using RNase A (QIAGEN, Australia) instead of RNase T1. Frozen *A. psidii* spores (isolate ID: Au_3, culture no. 622 and accession 115012) were ground in liquid nitrogen and then all methods were run up to step 33 of the protocol [66]. The DNA solution was then further purified to separate high molecular weight DNA from other impurities and low molecular weight DNA as described by Dong [67]. DNA concentration and purity were measured with a Qubit 3 (Invitrogen) and Nanodrop ND-1000 (Thermo Fisher Scientific). If the ratio of concentration obtained from the Qubit to that of Nanodrop was smaller than 0.5, AMPure beads (Beckman, Coulter Inc.) were used to purify the DNA at the ratio of 0.45 beads to DNA (vol/vol) following the manufacturer's protocol. DNA integrity was checked by pulsed-field gel electrophoresis. The DNA was first sequenced on PacBio RSII (18 SMRT® cells) at Ramaciotti Centre for Genomics, University of NSW, Australia. Later, samples were sequenced on PacBio Sequel (total 11 SMRT® cells). The sequencing library was prepared using SMRT cell template prep Kit 1.0-SPv3 with BluePippin Size-selection with 15-20 kb cutoff, and then sequenced using Sequel Sequencing Kit 2.0 at the Plant Genomics Centre, School of Agriculture, Food and Wine, the University of Adelaide, and the Ramaciotti Centre for Genomics, University of NSW, Australia. Altogether, a total of 126.1 (bax.h5) and 110.5 (subreads.bam) Gigabytes (GB) of raw data was obtained. The data for this study are deposited

at the European Nucleotide Archive (ENA) at EMBL-EBI under study accession: PRJEB34084 and sample accession ERS3953144.

RNA isolation and sequencing

Syzygium jambos young leaves were infected with myrtle rust. Infected leaves were harvested at 6, 12, 24, 48 hours, 5, and 7 days post-inoculation and stored at -80 °C. Total RNA was extracted from about 100 mg tissue, using the Spectrum Total RNA Kit (Sigma-Aldrich, St. Louis, MO), protocol A. The total RNA was then treated with RNase-free DNase I (New England BioLabs Inc.), and column purified using ISOLATE II RNA Mini Kit (Bioline, Australia) according to the manufacturer's instruction. The quantity and quality of the total RNA were examined by Nanodrop (Thermo Scientific) and Agilent 2100 Bioanalyzer (Agilent Technologies). Library was constructed using a TruSeq Stranded mRNA-seq preparation kit (Illumina Inc. San Diego, CA, USA) and sequenced on NextSeq 500 (PE 150bp) in one MID flowcell at Ramaciotti Centre for Genomics, University of NSW, Australia.

Genome assembly

We optimized the sequencing based on our predicted genome size of ~1 Gbp and sequenced 29 SMRT® cells (18 X RSII plus 11 X Sequel). After converting RSII data to subreads.bam we obtained 162 GB of data. Details on the genome assembly are available on the github repository [29]. In brief, RSII bax.h5 files were converted to subreads.bam before fasta files were extracted from all datasets. After extracting fasta files for assembly we had approximately 72 times raw sequence coverage (7.24E+10 bases and 5.80E+06 reads). The reads were then mapped to an in-house *A. psidii* mitochondrial sequence to retain only genomic DNA sequence data (APSI_mitochondria.fa available on the github repository) [29]. Canu version 1.6 [31] long read

assembly software was used to assemble the genome with correctedErrorRate=0.040, batOptions="-dg 3 -db 3 -dr 1 -ca 500 -cp 50" and predicted genome size of 1 Gb. The assembly ran for over four months continually on the University of Sydney High Performance Compute (HPC) cluster. Approximately 359,878 CPU hours were used for the assembly stage. Grid options were specified, and job scheduling was managed with PBS-Pro v13.1.0.160576, thereby limiting the number of jobs running at any one time to around 400. Each job required 4 CPUs, around 16 GB memory allocation and progressive allocations of walltime to a maximum of 70 hours.

Deduplication of assembly and polish

To deduplicate the assembly into primary contigs and secondary (haplotigs) we used the Purge Haplotigs (v1.0) pipeline [32] with alignment coverage of 65 percent. We polished the primary assembly using the genomic consensus algorithm *arrow* within Anaconda2 (v.5.0.1) (PacBio), after aligning with Minimap2 (v2.3) [68] and processing the data for compatibility, as described in the github repository [29]. This process was repeated twice and genome completeness for the primary and secondary genomes was assessed with BUSCO (v3.1.0) [30] using genome mode and the Basidiomycota database (basidiomycota_odb9 downloaded 24/02/2017) of 1335 conserved genes. Where conserved single copy orthologues were present on the secondary contigs and absent from the primary contigs, these contigs were detected and moved to the primary assembly using custom python scripts.

Telomere identification for scaffolding

Prior to scaffolding we first checked for telomeric regions using blast+ v.2.3.0 (blastn evalue 1e-5) [69] based on the hexamer TTAGGG(n) [37] with the aim of guiding the scaffold numbers to

putative chromosomes. We tested n=20, 40, 80, 120. From n=40 (length=240 bp) the contigs and numbers remained consistent with 100% match and e-value=0 for 24 contigs. A further 5 contigs had e-values less than 1.00E-07. Further analysis with seqkit (v0.10.1) [70] operation *locate* and a minimum TTAGGG repeat of 8 (and then 6) confirmed the BLAST findings. The hexamer repeat locations were at the start and/or end of contigs and were visualized using Integrated Genomics Viewer (IGV v2.6.3) [71]. We reasoned that the repeats were likely to represent telomeric regions and used these numbers to guide Hi-C scaffolding.

Transposable elements (TE) analysis

Repeat regions were annotated as described previously [53] using the REPET pipeline (v2.5) for repeat annotation in combination with Repbase (v21.05). For de novo identification, we predicted repeat regions with TEdenovo in the initial assembly by subsampling 300 Mb of sequence space at random to reduce compute time in the risk of losing low abundant TEs. We called the resulting repeat library MR_P2A_300Mb. We annotated the primary assembly with three repeat libraries (repbase2005_aaSeq, repbase2005_ntSeq, and MR_P2A_300Mb) using TEanno from the REPET pipeline. Detailed description of the repeat annotation and analysis can be found in the associated github repository. We identified AT-isochomers using OcculterCut [36]. We calculated GC content for specific genome features or sliding windows (window size 1kb, slide 200bp) using bedtools[72]. Karyplots of scaffolds were plotted with karyplotR. [73]

Hi-C libraries and scaffolding

Hi-C libraries were prepared using the Phase Genomics, Inc. (www.phasegenomics.com) Proximo Hi-C Kit (Microbe) to the manufacturer's specifications. Illumina paired-end reads (125 bp on HiSeq) generated from Hi-C libraries were used to scaffold contigs. Reads were firstly trimmed with adapters and low-quality bases, and then mapped to the primary and secondary

assemblies independently following the mapping workflow from Arima Genomics [74]. Three Hi-C scaffolding programs were tested (LACHESIS, SALSA and ALLHIC) before settling with ALLHIC (v.0.8.11)) [33]. We tested the grouping, ordering and orienting of the contigs with both MluCl and Sau3AI, (SALSA) as enzyme cutting sites, both together and separately (LACHESIS and ALLHIC). Final scaffolds were generated from ALLHIC with the use of MluCl ("AATT") as the cutting site.

RNA data analysis

RNA-seq data were generated for six time points. Each set of data for different time point was processed in parallel in a consistent way. The quality of raw RNA-seq data was checked using FastQC (v.0.11.7) and an overall summary for all samples was created with MultiQC (v.1.5). Based on the QC reports, the data were cleaned using fastp (v.0.19.4) [75] trimming 15bp and 10bp from the 5' and 3' respectively, together with '--trim_poly_x --trim_poly_g' trimming. The cleaned data for each time point were assembled independently using Trinityrnaseq (v2.6.5) [76]. Six transcriptome assemblies were merged to create the final transcripts using Evigene (v.18-01-2018) [77].

Taxonomy classification based on final assemblies, as well as representative and alternative forms of genes within transcriptomes, was subsequently carried out with Kraken2 [78] against NCBI/RefSeq Fungi database (dated 20190905) and visualized with Krona (v.2.7) [79] (Additional Figure 1.HTML). A large percentage of 'no hit' were recorded as plant genomes were not included in the database (we used NCBI/RefSeq Fungi). In addition to transcriptome assemblies, the cleaned data were mapped to both the *A. psidii* genome and the *Metrosideros polymorpha* ('ōhi'a) genome, where the latter is the most closely related to the inoculation host plant among the publicly available ones. The *M. polymorpha* genome was

downloaded from http://getentry.ddbj.nig.ac.jp and mapping was done with bowtie2 (v.2.3.4) [80] in sensitive local mode.

We mapped the combined, untrimmed RNA-seq reads to the primary assembly with STAR (v.2.7.2b) with the following parameters –sjdbGTFfile --sjdbGTFtagExonParentTranscript Parent --sjdbGTFtagExonParentGene ID --quantMode GeneCounts --outSAMtype BAM Unsorted --outSAMprimaryFlag AllBestScore. The output APSI_ReadsPerGene file was sorted by counts in descending order, per predicted coding regions, to identify sequences with expression. Mapped reads were visualized using IGV (v.2.6.3). We also used transcriptome translations, made with TransDecoder (v.5.5.0) [74], to query predicted proteins, including effectors, from the primary and secondary assemblies with blast+ (v.2.3.0) (blastp evalue 1e-7) [69] for evidence of expression in the inoculated *Syzygium jambos*. The assembled transcriptome fasta file is available from DOI:10.5281/zenodo.3567172 and data (APSI_primary/secondary_v1.xlsx) are accessible from the github repository [29].

Gene prediction and functional annotation

We annotated the primary and secondary scaffolds independently using Braker2 [81] by first soft masking with RepeatModeler (v.1.0.8) [82] and RepeatMasker (v. 4.0.6) [83]. RNA-seq reads (29.4 GB) were trimmed with Fastp (v.0.19.6) [75], and mapped to the masked scaffolds with Hisat2 [84] for alignment files (RNA-seq hints). Structural annotation outputs from Braker2 were utilized in downstream functional annotation steps. Automated functional annotation was performed on the 18,875 (primary) and 15,684 (secondary) Braker2 predicted proteins for domain and motif searches with InterProScan (v.5.34-73.0) [85]. A BUSCO [30] analysis was run in protein mode against the combined predicted annotated genes with Basidiomycota database of 1335 conserved genes. Additionally, we tested for transmembrane regions and signal

peptide motifs using SignalP (v.4.1f) [42]. A reduced file of sequences conforming to these predictions was submitted to both EffectorP (v.2.0) [20] and ApoplastP (v.1.01) [43] to identify predicted effectors and apoplastic proteins from each scaffolded assembly. Predicted effector fasta files were submitted to the online HMMER [86], using phmmer and the reference proteomes database, for an alternative annotation output based on homology. In order to determine allelic counterparts within both the primary and secondary assemblies we ran blast+ (v.2.3.0) (blastp evalue 1e-5) [69] with the predicted protein fasta files and identified query and subject equal aligned lengths at 100 percent match.

Comparative analysis of Pucciniales proteins

We used Orthofinder v2.2.7 (Emms and Kelly, 2018) with NCBI [50] and Mycocosm [50] (`*` represents data downloaded from Mycocosm) downloaded protein fasta files; **Puccinia coronata* var. *avenae* f. sp. *avenae*, *Puccinia striiformis* 93-210, *Puccinia graminis* f. sp. *tritici* CRL 75-36-700-3, *Puccinia sorghi* RO10H11247, **Puccinia triticina* 1-1 BBBD Race1, *Melampsora larici-populina* 98AG31, *Cronartium quercuum* f. sp. *fusiforme* G11. A visualized summary of ortholog clusters between species was generated with in-house PERL and R scripts.

Additionally, we reviewed the literature to determine basic genome statistics for these rust pathogens and compared the datasets gene architecture based on annotations (gff/gff3 files). For the analysis of gene and intergenic length we used the more complete assembly for *Puccinia striiformis* f. sp. *tritici*_104 E137 A, previously excluded from orthologue analysis due to diploidy, but excluded the strain *Puccinia striiformis* 93-210. We used custom R scripts to determine gene length and intergenic length distribution across datasets to determine alternative explanations for genome expansion. Untranslated regions (UTR) were not annotated for every gene therefore the analysis includes UTRs.

Variant analysis of Australian isolates

DNA from six other Australian isolates, one Hawaiian and one Brazilian isolate stored at the PBI were sequenced on Illumina sequencing platform (Table 7). A total of 209 GB of (150 bp) paired-end sequence data was first trimmed and then Bowtie2 [80] was used for mapping reads to the primary genome assembly using the --end-to-end option for each sample independently. Each sample sam file was them converted into bam files using Samtools [88], sorted and indexed against the reference genome. Samtools mpileup and bcftools were used to generate compressed variant calling files (vcf.gz). Variants were filtered using vcftools to a minimum depth of 6X and insertion-deletions were removed. The vcf file was sorted and format converted with Tassel v5 [89] and the phylip alignment file was run with iqtree v1.6.7 [90] with parameters -bb 1000 -nt AUTO and the Brazilian isolate as outgroup. Phylogenetic evolutionary relationships based on 44,375 single nucleotide polymorphisms within pathogen isolates collected in Australia (Au), Hawaii (Hw) and outgroup from Brazil (Bz from *Eucalyptus grandis*) were visualized with Dendroscope (v.3) [48].

Table 7. Details of local and exotic isolates of *Austropuccinia psidii* maintained at the University of Sydney, Plant Breeding Institute (PBI), and used for preliminary variant analysis. All Au isolates were increased from single pustule isolate on *Syzygium jambos* [6] **.**

Isolate ID	Acc. No.	Original host	Location	Year of Collection
Au_1	115001	*Syzygium jambos*	Lismore, NSW	2011
Au_4	125004	*Eucalyptus pilularis*	Newry, NSW	2012
Au_7	125009	*Melaleuca quinquenervia*	Queesncliff, NSW	2012
Au_9	125014	*Chamelaucium uncinatum*	Toowoomba, QLD	2012
Au_13	135001	*Rhodamnia maideniana*	Mooball, NSW	2013
Au_14	135002	*Rhodamnia rubescens*	Nightcap N. P., NSW	2013

Hw_17	135007	*Syzygium jambos*	Hawaii, USA	2013
Bz_15	135005	*Eucalyptus grandis*	Vicosa, Brazil	2013
Hw_17	135007	*Syzygium jambos*	Hawaii	2013

Availability of supporting data

Genome assembly and annotation data files:

The primary genome assembly file, named APSI_primary.v1.fasta, has been deposited at the European Nucleotide Archive (ENA) at EMBL-EBI under the following ENA accession: ERZ1194194 (GCA_902702905). Raw data is deposited at ENA under project: PRJEB34084 (Study accession ERP116938). Locus tags are registered as APSI (for *Austropuccinia psidii*) and scaffolds identified as APSI_Pxxx (where x indicates scaffold number) for primary assembly. Annotation data files and the dihaploid assembly file (APSI_v1.fa) are available from DOI:10.5281/zenodo.3567172. The dihaploid assembly incorporates the 67 secondary assembly scaffolds (APSI_Hxxx for secondary). Assembly and annotation scripts as well as data files are available at https://github.com/peritob/Myrtle-rust-genome-assembly: including the draft mitochondrial sequence, APSI_primary_v1.xlsx and APSI_secondary_v1.xlsx (containing HMMER and BLAST against transcriptome results)

Additional files

Additional Figure 1.html

APSI_primary and secondary_v1 genome and transcriptome mapping to the NCBI fungal database and visualised with Krona/2.7.

Additional Figure 2.pdf

Predicted effectors and telomeric regions identified on primary assembly scaffolds and visualised with MapChart v.2.3.2. Only scaffolds with relevant data are graphically depicted. Scale = Mbp. Gene ID and strand based on Braker2 annotations.

Additional Table 1.docx

Repetitive regions annotated with RepeatMasker (v4.0.6) (Smit et al., 1996-2010) and RepeatModeler (v1.0.8) (Smit and Hubley, 2008-2015) for the primary and secondary *A. psidii* assemblies.

Additional Table 2.xlsx

Pucciniales comparative statistics and Interproscan (v.5) results from primary and secondary assemblies.

Additional Table 3.docx

Hidden Markov model (HMMER) functional annotation of the predicted effectors within the primary assembly based on a reference proteome database [86]**. Listed are the 77 predicted effectors that had a functional annotation match.**

Abbreviations

BUSCO: Benchmarking Universal Single-Copy Orthologs BUSCO; gDNA: genomic DNA; TE: transposable elements; PacBio: Pacific Biosciences; RNA-seq: RNA sequencing; SMRT: single-molecule real time (SMRT®); GB: Gigabyte; Gb: Gigabase

Competing interests

The authors declare no competing interests in the preparation of these research results.

Funding

We acknowledge the support of the New Zealand Department of Primary Industries via RFP 18608 **Error! No text of specified style in document.** and The New Zealand Institute for Plant

and Food Research Limited for access to the high performance computing facility PowerPlant. This work was supported by an Australian Research Council DECRA (DE150101897) and Future Fellowship (FT180100024) to BS. This work was supported by an Australian Research Council DECRA (DE190100066) to JS. The provision of financial support to the University of Sydney by Judith and David Coffey and family is gratefully acknowledged.

Author's contributions

R.P. conceived of, supported the study and provided the samples. **P.T.** and **B.S.** are shared first authors. **P.T.** did flow cytometry, assembled and annotated the genome, ran analyses and wrote much of the manuscript. **B. S.** made substantial efforts in directing the research, made Hi-C libraries, ran analyses including TEs and genome structure. **C.Do**. extracted and sequenced the gDNA, ran inoculations and RNA extractions. **C.De.** and **C.W.** ran Hi-C and other analyses. **A.J.** made Hi-C libraries. **J.S.** ran analyses on scaffolds and contributed to the manuscript. **J.T.** provided mitochondrial sequence data and helped with RNA sequencing. **G.S.** with **D.C**. (who also ran SNP analysis), secured financial support for the research and helped to co-ordinate the study with **D.C.** All authors read and approved the final manuscript.

Acknowledgements

We wish to thank Mark Powrie who generously helped with computational support throughout the assembly and annotation process. The authors acknowledge the Sydney Informatics Hub and the University of Sydney's high-performance compute cluster, Artemis, for providing the computing resources that have contributed to the research results reported within this paper. Thanks also to Dr Geoff Pegg, Queensland Department of Agriculture Fisheries and Forestry, for sharing recent field observations on the impacts of myrtle rust.

References

1. Beenken L. *Austropuccinia*: a new genus name for the myrtle rust *Puccinia psidii* placed within the redefined family Sphaerophragmiaceae (Pucciniales). Phytotaxa. 2017;297:53–61.

2. Winter G. Rabenhorstii fungi europaei et extraeuropaei exsiccati cura Dr. G. Winter, Centuria XXXI et XXXII. Hedwigia. 1884;23:164–172.

3. Stewart JE, Ross-Davis AL, Gra□a RN, Alfenas AC, Peever TL, Hanna JW, et al. Genetic diversity of the myrtle rust pathogen (Austropuccinia psidii) in the Americas and Hawaii: Global implications for invasive threat assessments. For Pathol. 2018;48:1–13.

4. Machado P, Alfenas AC, Alfenas RF, Mohammed CL, Glen M. Microsatellite analysis indicates that Puccinia psidii in Australia is mutating but not recombining. Australas Plant Pathol [Internet]. 2015; Available from: http://link.springer.com/10.1007/s13313-015-0364-5

5. du Plessis E, Granados GM, Barnes I, Ho WH, Alexander BJR, Roux J, et al. The pandemic strain of Austropuccinia psidii causes myrtle rust in New Zealand and Singapore. Australas Plant Pathol. 2019;48:253–6.

6. Sandhu KS, Karaoglu H, Zhang P, Park RF. Simple sequence repeat markers support the presence of a single genotype of *Puccinia psidii* in Australia. Plant Pathol. 2015;

7. Carnegie A, Pegg G. Lessons from the Incursion of Myrtle Rust in Australia. Annu Rev Phytopathol. 2018;56:1–22.

8. Soewarto J, Carriconde F, Hugot N, Bocs S, Hamelin C, Maggia L. Impact of Austropuccinia psidii in New Caledonia, a biodiversity hotspot. For Pathol. 2018;48:1–12.

9. Carnegie AJ, Kathuria A, Pegg GS, Entwistle P, Nagel M, Giblin FR. Impact of the invasive

rust *Puccinia psidii* (myrtle rust) on native Myrtaceae in natural ecosystems in Australia. Biol Invasions. Springer International Publishing; 2016;18:127–44.

10. Coutinho TA, Wingfield MJ, Alfenas AC, Crous PW. Special Report Eucalyptus Rust□: A Disease with the Potential for Serious International Implications. Plant Dis. 1998;82:819–25.

11. Hunt P. Cuticular penetration by germinating uredospores. Trans Br Mycol Soc [Internet]. British Mycological Society; 1968 [cited 2014 May 2];51:103-IN7. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0007153668801263

12. Beresford RM, Shuey LS, Pegg G. Symptom development and latent period of Austropuccinia psidii (myrtle rust) in relation to host species, temperature and ontogenic resistance. Plant Pathol. 2020;

13. Tobias PA, Park RF, Külheim C, Guest DI. Wild-sourced *Chamelaucium uncinatum* have no resistance to *Puccinia psidii* (myrtle rust). Australas Plant Dis Notes. 2015;10:1–3.

14. Carnegie AJ, Kathuria A, Pegg GS, Entwistle P, Nagel M, Giblin FR. Impact of the invasive rust Puccinia psidii (myrtle rust) on native Myrtaceae in natural ecosystems in Australia. Biol Invasions [Internet]. Springer International Publishing; 2015; Available from: http://link.springer.com/10.1007/s10530-015-0996-y

15. Soewarto J. GF and CAJ. *Austropuccinia psidii* (myrtle rust) global host list. Version 2. [Internet]. Aust. Netw. Plant Conserv. Canberra, ACT. 2019 [cited 2019 Dec 10]. Available from: http://www.anpc.asn.au/myrtle-rust

16. Pegg G, Taylor T, Entwistle P, Guymer G, Giblin F, Carnegie A. Impact of *Austropuccinia psidii*(myrtle rust) on Myrtaceae-rich wet sclerophyll forests in south east Queensland. PLoS One [Internet]. 2017; Available from: http://dx.doi.org/10.1371/journal.pone.0188058

17. Christenhusz MMJ, Byng JW. The number of known plants species in the world and its annual increase. Phytotaxa. 2016;261:201–17.

18. Berthon K, Esperon-Rodriguez M, Beaumont LJ, Carnegie AJ, Leishman MR. Assessment and prioritisation of plant species at risk from myrtle rust (Austropuccinia psidii) under current and future climates in Australia. Biol Conserv [Internet]. Elsevier; 2018;218:154–62. Available from: https://doi.org/10.1016/j.biocon.2017.11.035

19. Makinson RO. Myrtle Rust in Australia: A draft Action Plan [Internet]. Melbourne; 2018. Available from: apbsf.org.au

20. Sperschneider J, Dodds PN, Gardiner DM, Singh KB, Taylor JM. Improved prediction of fungal effector proteins from secretomes with EffectorP 2.0. Mol Plant Pathol. 2018;19:2094–110.

21. Tan M-K, Collins D, Chen Z, Englezou A, Wilkins MR. A brief overview of the size and composition of the myrtle rust genome and its taxonomic status. Mycol An Int J Fungal Biol [Internet]. 2014 [cited 2014 Jun 3];5:52–63. Available from: http://www.tandfonline.com/doi/abs/10.1080/21501203.2014.919967

22. 10XGenomics. 10XGenomics [Internet]. Available from: https://www.10xgenomics.com/

23. Mctaggart AR, Duong TA, Le VQ, Shuey LS, Smidt W, Naidoo S, et al. Chromium sequencing: the doors open for genomics of obligate plant pathogens. Biotechniques. 2018;65:1–4.

24. van Berkum NL, Lieberman-Aiden E, Williams L, Imakaev M, Gnirke A, Mirny LA, et al. Hi-C: A method to study the three-dimensional architecture of genomes. J Vis Exp. 2010;1–7.

25. Tavares S, Ramos AP, Pires AS, Azinheira HG, Caldeirinha P, Link T, et al. Genome size

analyses of Pucciniales reveal the largest fungal genomes. Front Plant Sci [Internet]. 2014;5:422. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4143883&tool=pmcentrez&rendertype=abstract

26. Thrimawithana AH, Jones D, Hilario E, Grierson E, Hanh M, Liachko I, et al. A whole genome assembly of Leptospermum scoparium (Myrtaceae) for mānuka research A whole genome assembly of Leptospermum scoparium (Myrtaceae) for m ā nuka research. New Zeal J Crop Hortic Sci [Internet]. Taylor & Francis; 2019;0:1–28. Available from: https://doi.org/10.1080/01140671.2019.1657911

27. Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J, et al. The genome of *Eucalyptus grandis*. Nature [Internet]. 2014 [cited 2014 Jun 11];510:356–362. Available from: http://www.nature.com/doifinder/10.1038/nature13308

28. Horns F, Petit E, Hood ME. Massive expansion of gypsy-like retrotransposons in microbotryum fungi. Genome Biol Evol. 2017;9:363–71.

29. Tobias PA, Deng CH, Wu C, Dong C, Jones A, Smith G, et al. Myrtle-rust-genome-assembly [Internet]. 2019. Available from: https://github.com/peritob/Myrtle-rust-genome-assembly

30. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2.

31. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu□: scalable and accurate long-read assembly via adaptive k--mer weighting and repeat separation. Genome Res. 2016;1–35.

32. Roach MJ, Schmidt SA, Borneman AR. Purge Haplotigs: Synteny Reduction for Third-gen Diploid Genome Assemblies. bioRxiv [Internet]. BMC Bioinformatics; 2018;286252. Available from: https://www.biorxiv.org/content/early/2018/03/22/286252

33. Zhang X, Zhang S, Zhao Q, Ming R, Tang H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. Nat Plants [Internet]. Springer US; 2019;5:833–45. Available from: http://dx.doi.org/10.1038/s41477-019-0487-8

34. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA [Internet]. Mobile DNA; 2015;6:4–9. Available from: http://dx.doi.org/10.1186/s13100-015-0041-9

35. Flutre T, Duprat E, Feuillet C, Quesneville H. Considering transposable element diversification in de novo annotation approaches. PLoS One. 2011;6.

36. Testa AC, Oliver RP, Hane JK. OcculterCut: A comprehensive survey of at-rich regions in fungal genomes. Genome Biol Evol. 2016;8:2044–64.

37. Lucía R, Gúmer P, Raúl C, Francisco S, G PA. Basidiomycetes Telomeres –. Bioinforma - Trends Methodol. 2010;19:393–424.

38. Schwessinger B, Chen Y-J, Tien R, Vogt JK, Sperschneider J, Nagar R, et al. Distinct life histories impact dikaryotic genome evolution in the rust fungus *Puccinia striiformis* causing stripe rust in wheat. bioRxiv [Internet]. 2020;859728. Available from: http://biorxiv.org/content/early/2020/02/18/859728.abstract

39. Laberthonnière C, Magdinier F, Robin JD. Bring It to an End: Does Telomeres Size Matter? Cells. 2019;8:30.

40. Boehm EWA, Wenstrom JC, McLaughlin DJ, Szabo LJ, Roelfs AP, Bushnell WR. An

ultrastructural pachytene karyotype for *Puccinia graminis* f.sp. *tritici*. Can J Bot. 1992;70:401–13.

41. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1: Unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics. 2016;32:767–9.

42. Nielsen H. Predicting Secretory Proteins with SignalP. In: Kihara D, editor. Protein Funct Predict Methods Mol Biol [Internet]. vol 1611. New York, NY: Humana Press; 2017. Available from: https://doi.org/10.1007/978-1-4939-7015-5_6

43. Sperschneider J, Dodds PN, Singh KB, Taylor JM. ApoplastP: prediction of effectors and plant proteins in the apoplast using machine learning. New Phytol. 2018;217:1764–78.

44. Voorrips RE. MapChart: software for the graphical presentation of linkage maps and QTLs. J Hered. 1994;93:77–8.

45. Emms DM, Kelly S. OrthoFinder2: fast and accurate phylogenomic orthology analysis from gene sequences. bioRxiv [Internet]. 2018; Available from: https://www.biorxiv.org/content/10.1101/466201v1.full

46. Pendleton AL, Smith KE, Feau N, Martin FM, Grigoriev I V., Hamelin R, et al. Duplications and losses in gene families of rust pathogens highlight putative effectors. Front Plant Sci. 2014;5:1–13.

47. Duplessis S, Cuomo CA, Lin YC, Aerts A, Tisserant E, Veneault-Fourrey C, et al. Obligate biotrophy features unraveled by the genomic analysis of rust fungi. Proc Natl Acad Sci U S A. 2011;108:9166–71.

48. Huson DH, Scornavacca C. Dendroscope 3: An interactive tool for rooted phylogenetic trees

and networks. Syst Biol. 2012;61:1061–7.

49. Emms DM, Kelly S. STAG: Species Tree Inference from All Genes. bioRxiv. 2018;1–29.

50. DOE Joint Genome Institute. Mycocosm [Internet]. [cited 2019 Sep 24]. Available from: https://mycocosm.jgi.doe.gov/mycocosm/home

51. Egertová Z, Sochor M. The largest fungal genome discovered in *Jafnea semitosta*. Plant Syst Evol. 2017;303:981–6.

52. Ramos AP, Tavares S, Tavares D, Silva MDC, Loureiro J, Talhinhas P. Flow cytometry reveals that the rust fungus, *Uromyces bidentis* (Pucciniales), possesses the largest fungal genome reported-2489Mbp. Mol Plant Pathol. 2015;16:1006–10.

53. Schwessinger B, Sperschneider J, Cuddy WS, Garnica DP, Miller ME, Taylor JM, et al. A near-complete haplotype-phased genome of the dikaryotic. MBio. 2018;9:e02275-17.

54. MacNeil DE, Bensoussan HJ, Autexier C. Telomerase regulation from beginning to the end. Genes (Basel). 2016;7.

55. Orbach MJ, Farrall L, Sweigard JA, Chumley FG, Orbach MJ, Farrall L, et al. A Telomeric Avirulence Gene Determines Efficacy for the Rice Blast Resistance Gene Pi-ta. Plant Cell. 2000;12:2019–32.

56. Kuprys P V., Davis SM, Hauer TM, Meltser M, Tzfati Y, Kirk KE. Identification of Telomerase RNAs from Filamentous Fungi Reveals Conservation with Vertebrates and Yeasts. PLoS One. 2013;8:1–11.

57. Pegg GS, Giblin FR, McTaggart AR, Guymer GP, Taylor H, Ireland KB, et al. *Puccinia psidii* in Queensland, Australia: Disease symptoms, distribution and impact. Plant Pathol. 2014;63:1005–21.

58. Berthon KA, Winzer LF, Carnegie AJ, Leishman MR. Endangered species face an extra threat□: susceptibility to the invasive pathogen *Austropuccinia psidii* (myrtle rust) in Australia. Australas Plant Pathol. Australasian Plant Pathology; 2019;

59. Fernandez Winzer L, Berthon KA, Carnegie AJ, Pegg GS, Leishman MR. Austropuccinia psidii on the move: survey based insights to its geographical distribution, host species, impacts and management in Australia. Biol Invasions [Internet]. Springer International Publishing; 2018;0. Available from: https://doi.org/10.1007/s10530-018-1891-0

60. Grattapaglia D, Bradshaw Jr HD. Nuclear DNA content of commercially important Eucalyptus species and hybrids. Can J For Res. 1994;24:1074–8.

61. Tobias PA, Guest DI, Külheim C, Park RF. De Novo Transcriptome Study Identifies Candidate Genes Involved in Resistance to *Austropuccinia psidii* (Myrtle Rust) in *Syzygium luehmannii* (Riberry). Phytopathology [Internet]. 2018;108:PHYTO-09-17-029. Available from: https://apsjournals.apsnet.org/doi/10.1094/PHYTO-09-17-0298-R

62. Terho P. Flowing Software [Internet]. Turku Centre for Biotechnology; 2013. Available from: http://flowingsoftware.btk.fi/index.php?page=1

63. Doležel J, Bartoš J. Plant DNA flow cytometry and estimation of nuclear genome size. Ann Bot. 2005;95:99–110.

64. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27:764–70.

65. Schwessinger B. High quality DNA from fungi for long read sequencing e.g. PacBio. protocols [Internet]. protocols.io:https://doi.org/10.17504/protocols.io.ewtbfen. 2016. Available from: https://doi.org/10.17504/protocols.io.ewtbfen

66. Dong C. Purification of HMW DNA from Fungi for long read sequencing [Internet]. protocols.io:https://doi.org/10.17504/protocols.io.ewtbfen. 2017. Available from: dx.doi.org/10.17504/protocols.io.hbvb2n6

67. Dong C. Purification of HMW DNA from Fungi for long read. 2017;3–4.

68. Li H. Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.

69. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol [Internet]. 1990;215:403–10. Available from: http://www.sciencedirect.com/science/article/pii/S0022283605803602

70. Shen W, Le S, Li Y, Hu F. SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS One. 2016;11:1–10.

71. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. Brief Bioinform. 2013;14:178–92.

72. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.

73. Gel B, Serra E. KaryoploteR: An R/Bioconductor package to plot customizable genomes displaying arbitrary data. Bioinformatics. 2017;33:3088–90.

74. Arima Genomics [Internet]. 2019. Available from: https://github.com/ArimaGenomics/mapping_pipeline

75. Chen S, Zhou Y, Chen Y, Gu J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34:i884–90.

76. 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 [Internet]. 2013 [cited 2014 Apr 28];8:1494–512. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3875132&tool=pmcentrez&rendertype=abstract

77. Gilbert D. EvigeneR [Internet]. 2013 [cited 2018 Jan 18]. Available from: http://arthropods.eugenes.org/EvidentialGene/

78. Wood DE, Salzberg SL. Kraken: Ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15.

79. Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics. 2011;385.

80. 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.

81. Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. Whole-Genome Annotation with BRAKER. In: M. K, editor. Methods Mol Biol [Internet]. New York, NY: Humana Press; 2019. p. 65–95. Available from: https://doi.org/10.1007/978-1-4939-9173-0_5

82. Smit A, Hubley R. RepeatModeler Open-1.0. [Internet]. Available from: http://www.repeatmasker.org

83. Smit A, Hubley R, Green P. RepeatMasker Open-4.0 [Internet]. Available from: http://www.repeatmasker.org

84. Kim D, Langmead B, Salzberg SL. HISAT: A fast spliced aligner with low memory

requirements. Nat Methods. 2015;12:357–60.

85. 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.

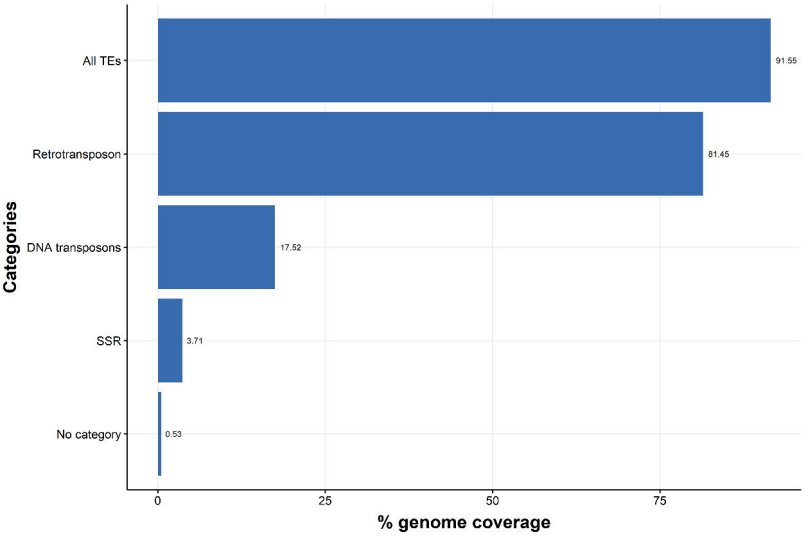
86. Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. HMMER web server: 2018 update. Nucleic Acids Res. 2018;46:W200–4.

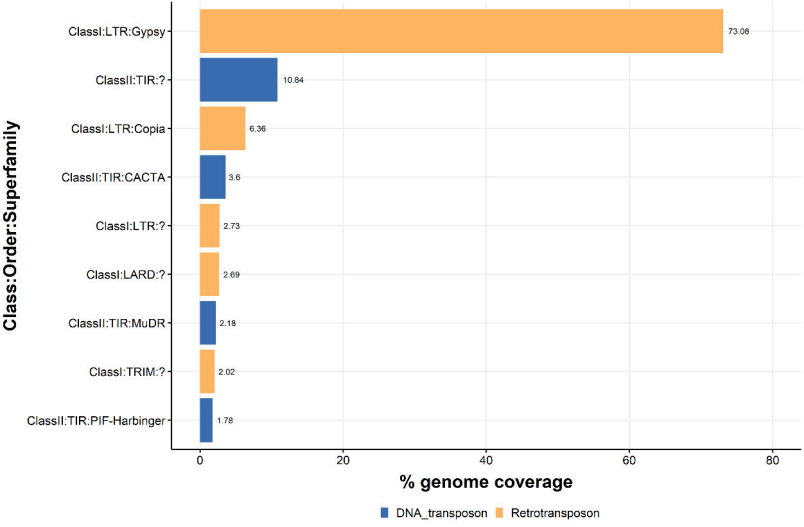
87. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol [Internet]. Genome Biology; 2015;16:1–14. Available from: http://dx.doi.org/10.1186/s13059-015-0721-2

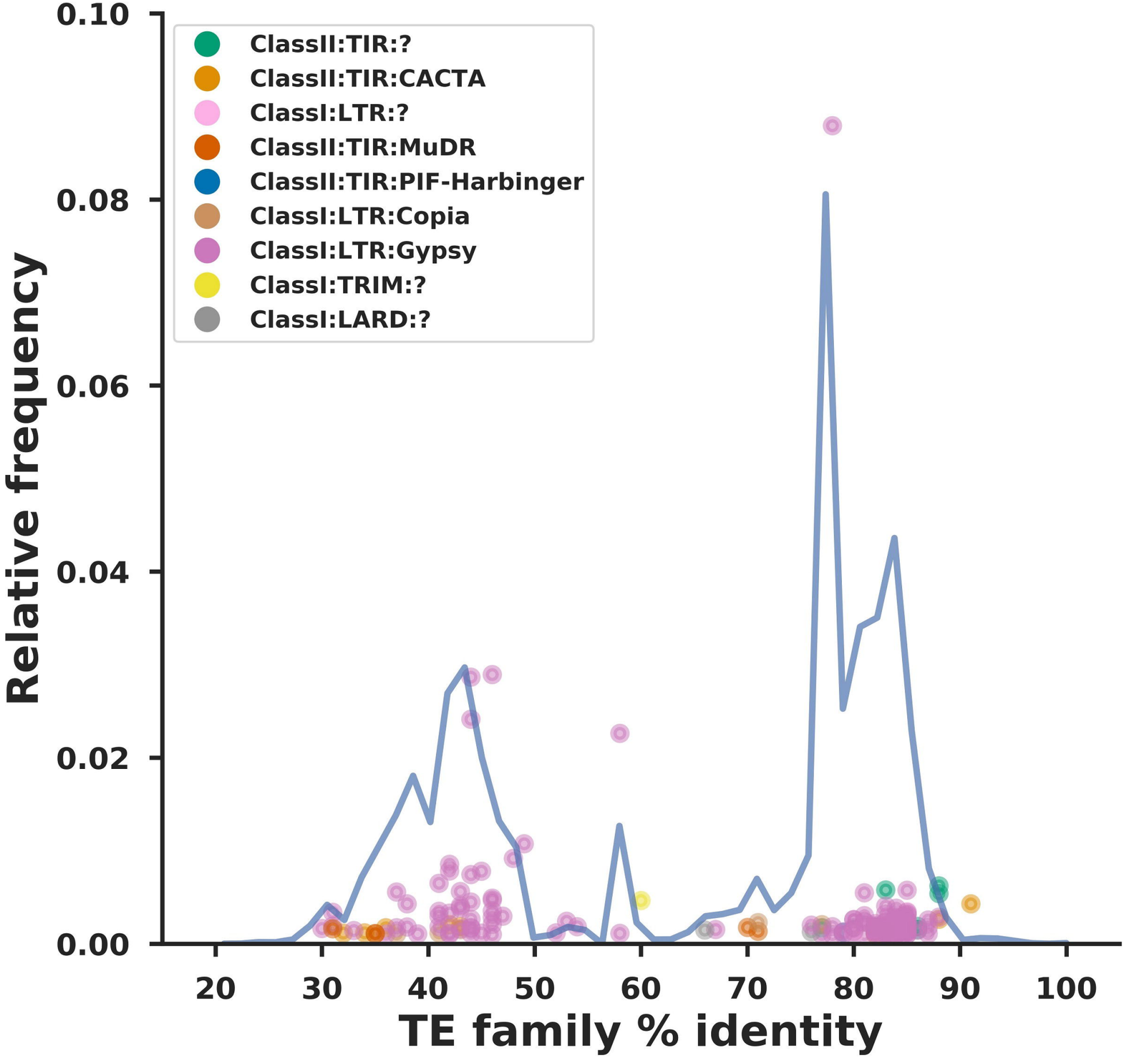
88. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

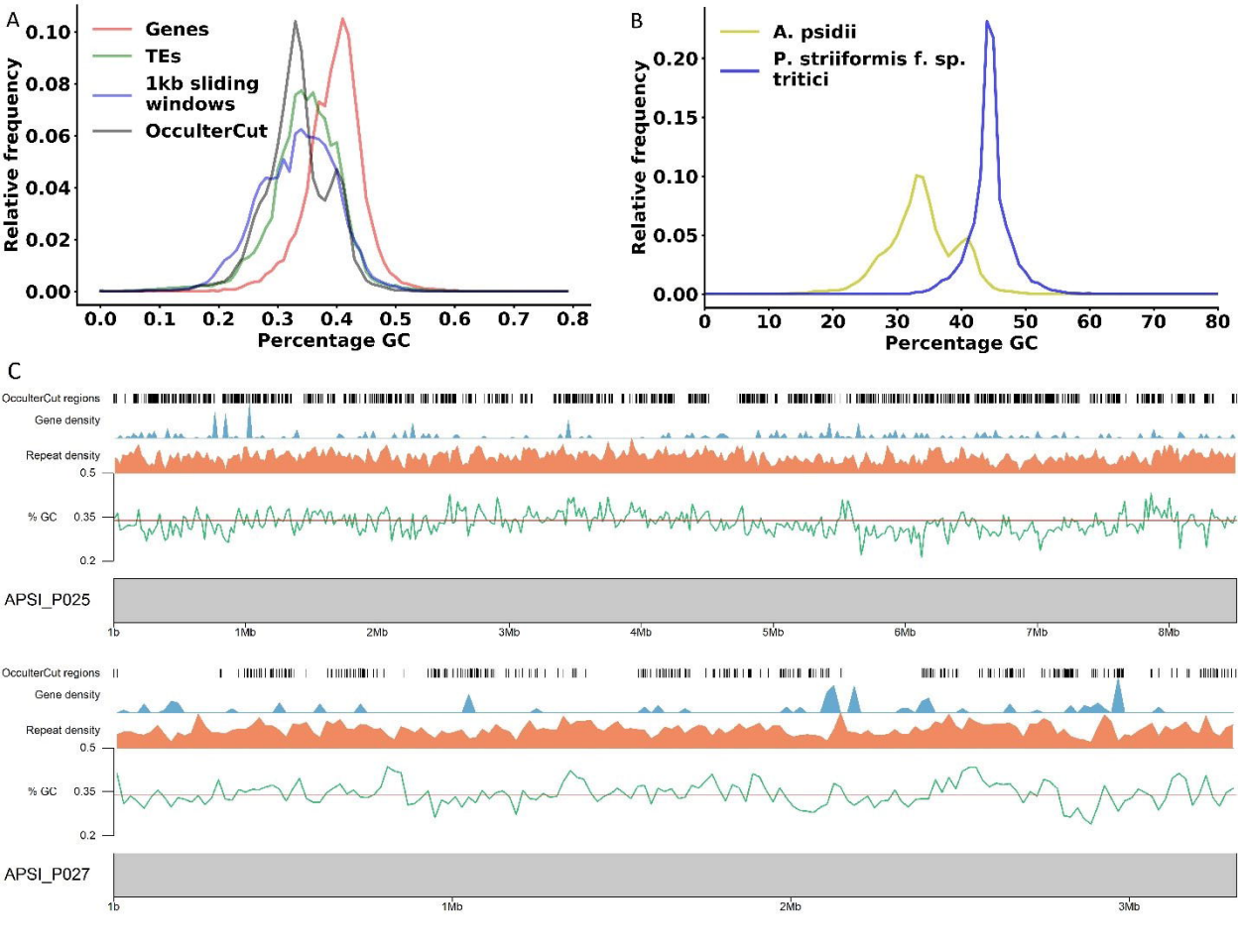
89. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics. 2007;23:2633–5.

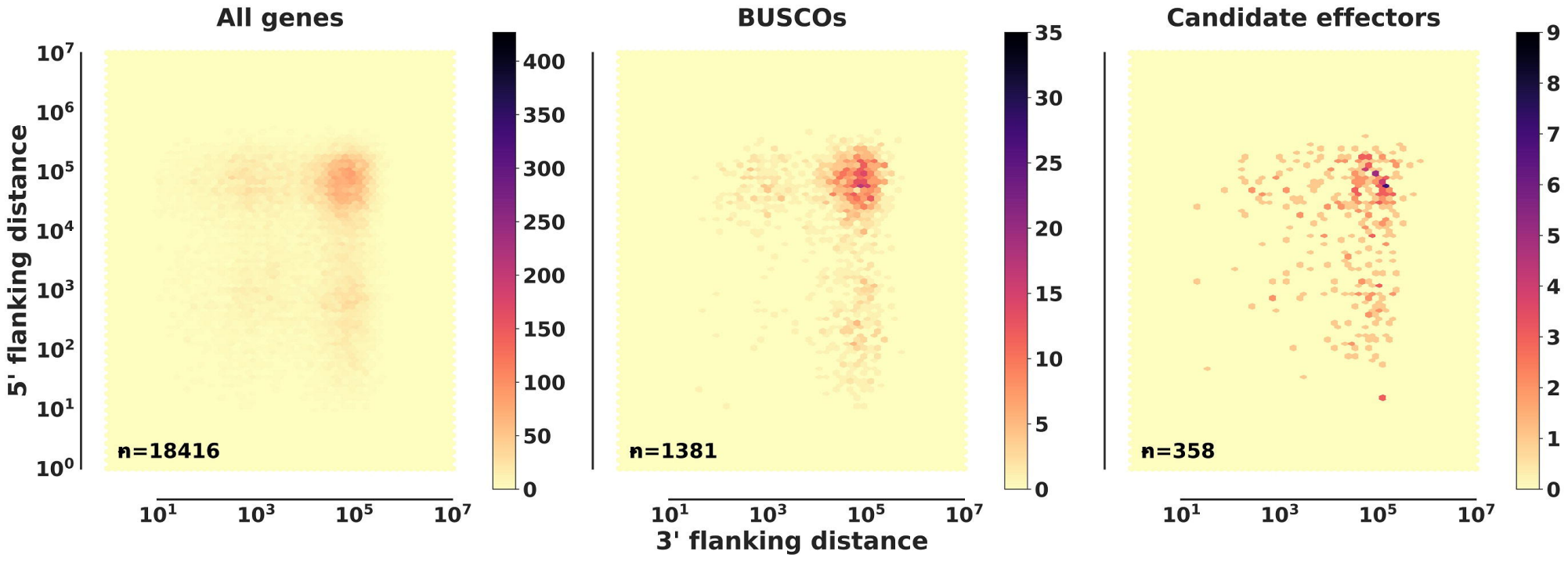
90. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32:268–74.


Categories
All TEs
Retrotransposon
DNA transposons
SSR
No category
91.55
81.45
17.52
3.71
0.53
0
25
50
75
% genome coverage

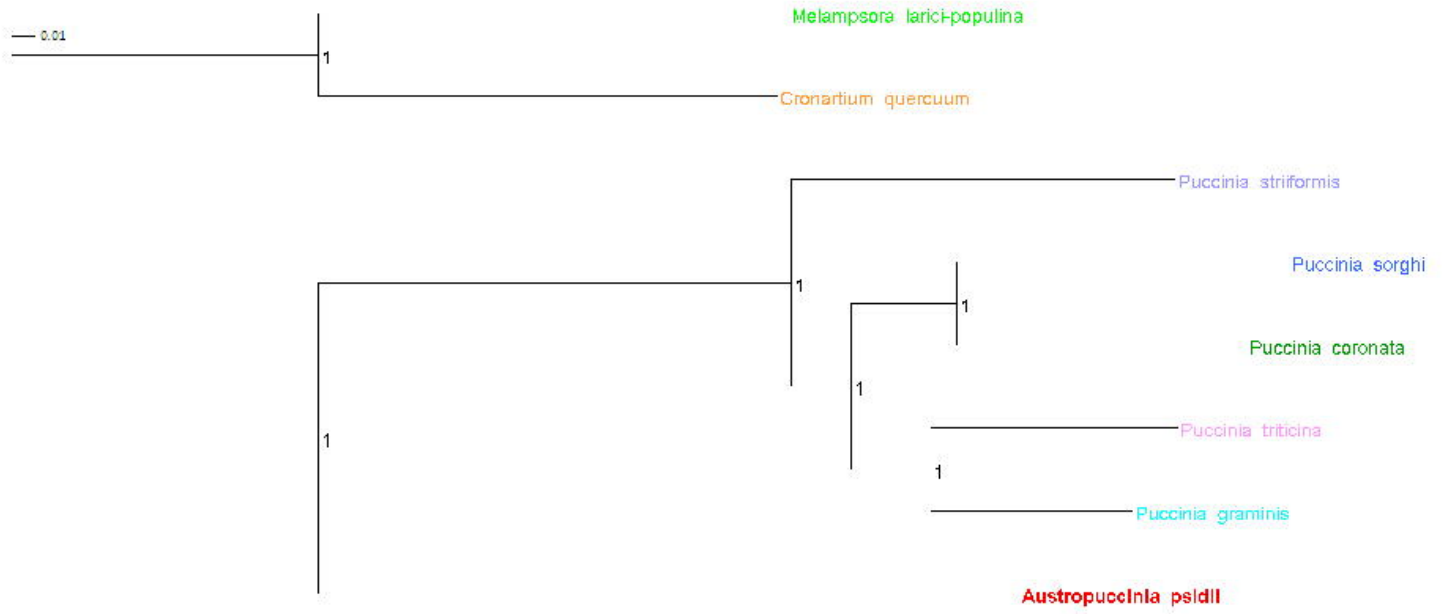

Class:Order:Superfamily
ClassI:LTR:Gypsy
ClassII:TIR:?
ClassI:LTR:Copia
ClassII:TIR:CACTA
ClassI:LTR:?
ClassI:LARD:?
ClassII:TIR:MuDR
ClassI:TRIM:?
ClassII:TIR:PIF-Harbinger
73.08
10.84
6.36
3.6
2.73
2.69
2.18
2.02
1.78
0
20
40
60
80
% genome coverage
DNA_transposon
Retrotransposon


ClassII:TIR:?
ClassII:TIR:CACTA
ClassI:LTR:?
ClassII:TIR:MuDR
ClassII:TIR:PIF-Harbinger
ClassI:LTR:Copia
ClassI:LTR:Gypsy
ClassI:TRIM:?
ClassI:LARD:?
Relative frequency
0.10
0.08
0.06
0.04
0.02
0.00
20
30
40
50
60
70
80
90
100
TE family % identity

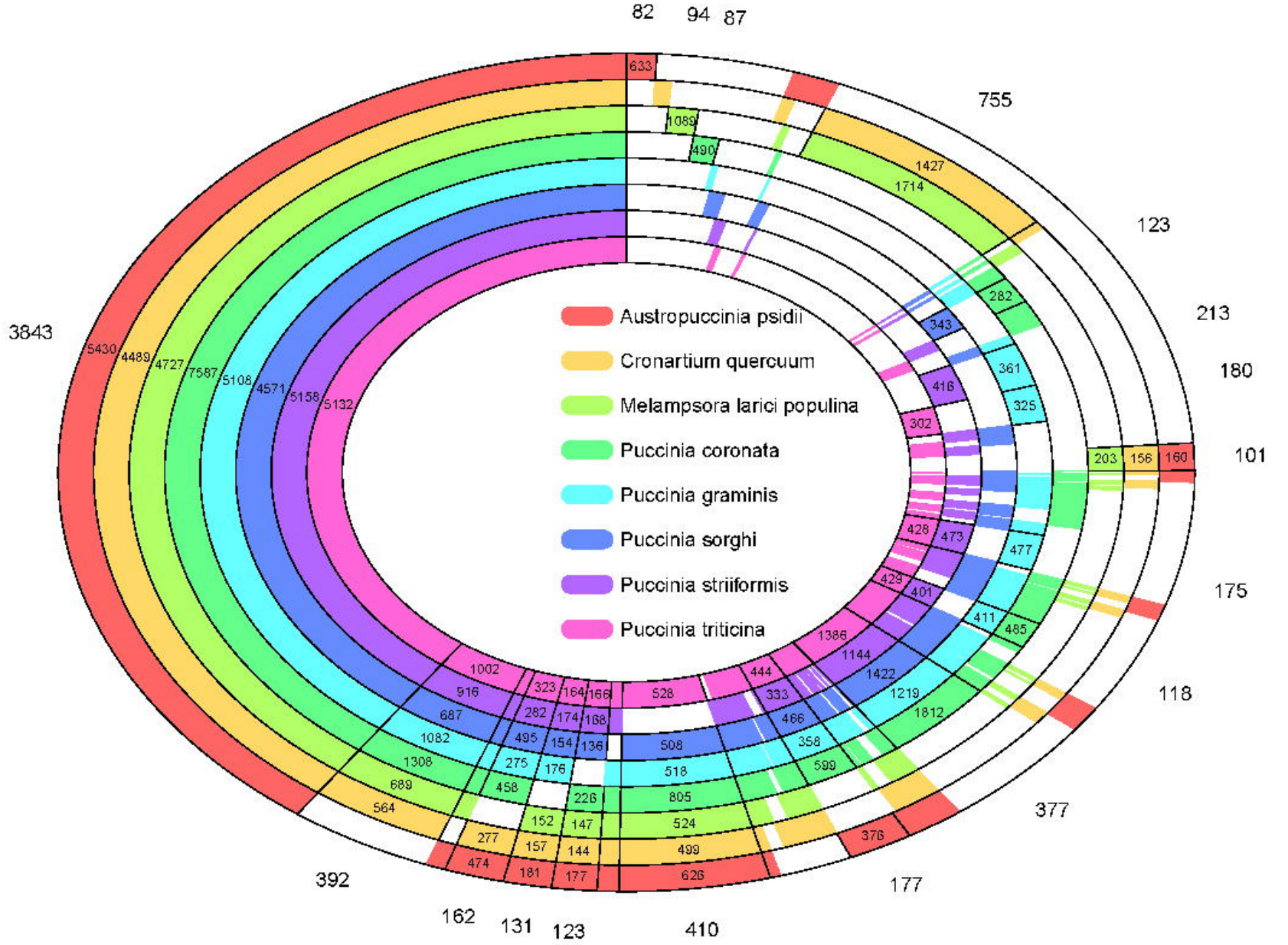

A
Genes
TEs
1kb sliding windows
OcculterCut
Relative frequency
0.10
0.08
0.06
0.04
0.02
0.00
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Percentage GC
B
A. psidii
P. striiformis f. sp. tritici
Relative frequency
0.20
0.15
0.10
0.05
0.00
0
10
20
30
40
50
60
70
80
Percentage GC
C
OcculterCut regions
Gene density
Repeat density
0.5
% GC
0.35
0.2
APSI_P025
1b
1Mb
2Mb
3Mb
4Mb
5Mb
6Mb
7Mb
8Mb
OcculterCut regions
Gene density
Repeat density
0.5
% GC
0.35
0.2
APSI_P027
1b
1Mb
2Mb
3Mb

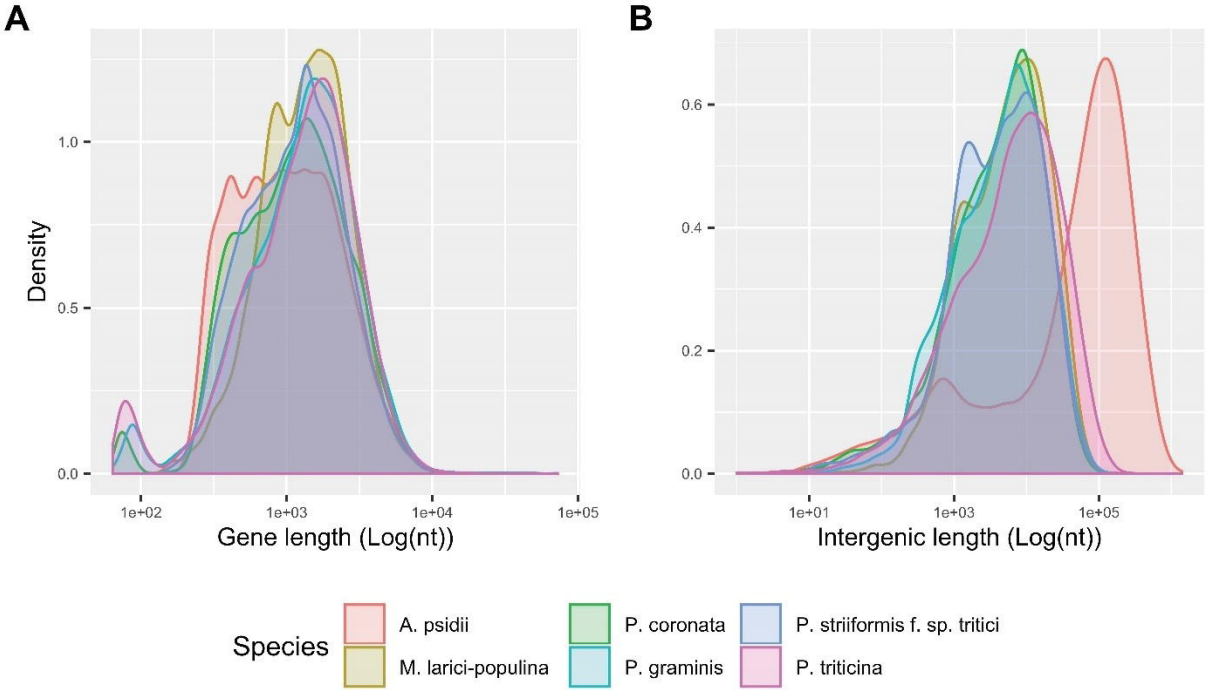

All genes
BUSCOs
Candidate effectors
5' flanking distance
3' flanking distance
n=18416
n=1381
n=358
10^0
10^1
10^2
10^3
10^4
10^5
10^6
10^7
0
50
100
150
200
250
300
350
400
5
10
15
20
25
30
35
1
2
3
4
5
6
7
8
9

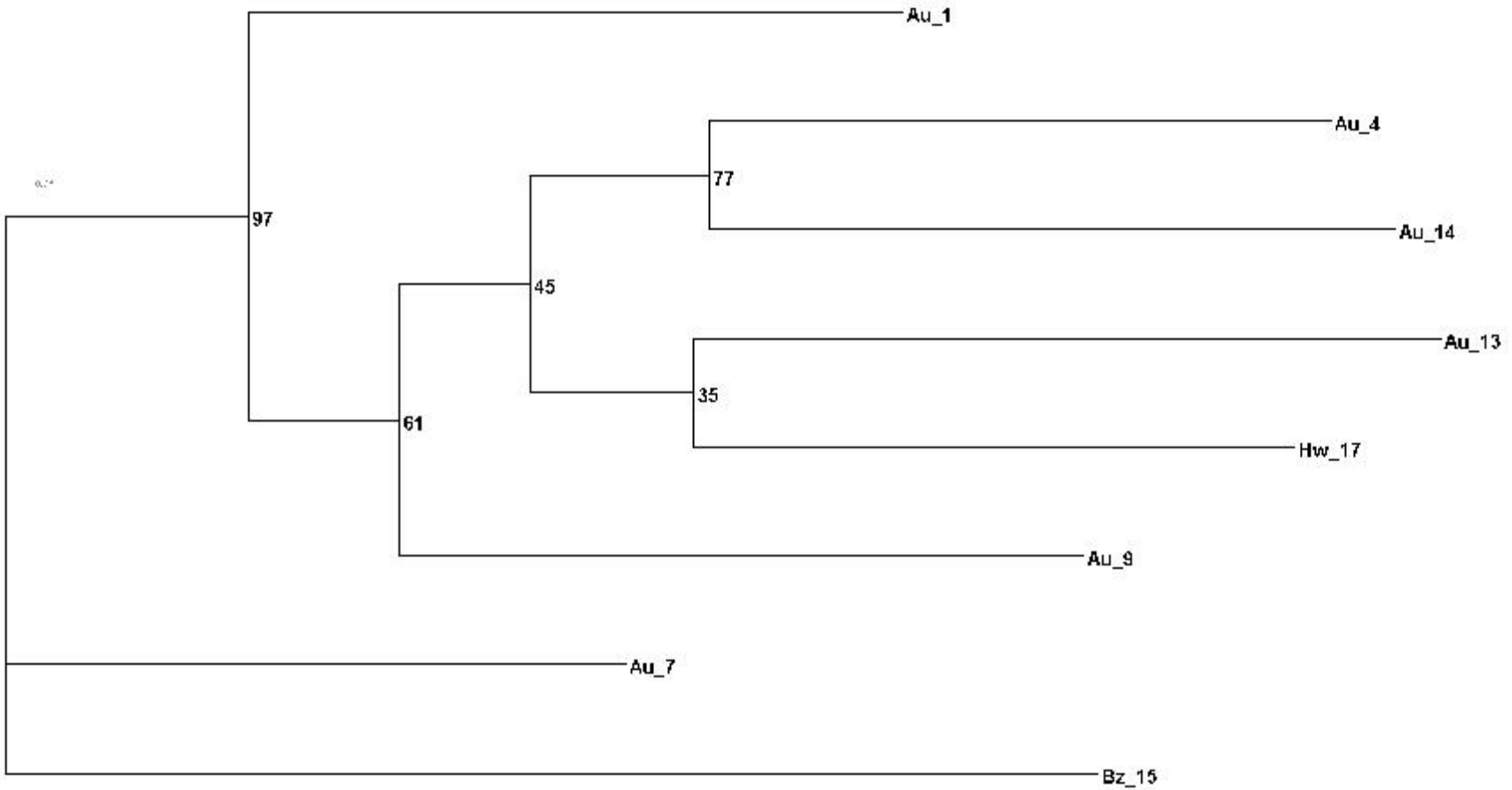
A


0.01
Melampsora larici-populina
Cronartium quercuum
Puccinia striiformis
Puccinia sorghi
Puccinia coronata
Puccinia triticina
Puccinia graminis
Austropuccinia psidii
1

B


Austropuccinia psidii
Cronartium quercuum
Melampsora larici populina
Puccinia coronata
Puccinia graminis
Puccinia sorghi
Puccinia striiformis
Puccinia triticina
82
94
87
755
123
213
180
101
175
118
377
177
410
123
131
162
392
3843


A
B
Density
1.0
0.5
0.0
1e+02
1e+03
1e+04
1e+05
Gene length (Log(nt))
0.6
0.4
0.2
0.0
1e+01
1e+03
1e+05
Intergenic length (Log(nt))
Species
A. psidii
M. larici-populina
P. coronata
P. graminis
P. striiformis f. sp. tritici
P. triticina


Au_1
Au_4
77
97
Au_14
45
Au_13
35
61
Hw_17
Au_9
Au_7
Bz_15