Draft Genome of the Wheat Rust Pathogen (*Puccinia triticina*) Unravels Genome-Wide Structural Variations during Evolution

Kanti Kiran[1], Hukam C. Rawal[1], Himanshu Dubey[1], Rajdeep Jaswal[1], B.N Devanna[1], Deepak Kumar Gupta[1], Subhash C. Bhardwaj[2], P. Prasad[2], Dharam Pal[3], Parveen Chhuneja[4], P. Balasubramanian[5], J. Kumar[6], M. Swami[7], Amolkumar U. Solanke[1], Kishor Gaikwad[1], Nagendra K. Singh[1] and Tilak Raj Sharma[1,*]

[1]ICAR-National Research Centre on Plant Biotechnology, New Delhi, India

[2]ICAR – Indian Institute of Wheat and Barley Research, Regional Station, Flowerdale, Shimla, India

[3]ICAR – Indian Agricultural Research Institute, Regional Station Tutikandi Centre, Shimla, India

[4]Punjab Agricultural University, Ludhiana, India

[5]Tamil Nadu Agricultural University, Coimbatore, Tamil Nadu, India

[6]ICAR – National Institute of Biotic Stress Management, Raipur, Chhattisgarh, India

[7]ICAR-Indian Agricultural Research Institute, Regional Station, Wellington, India

*Corresponding author: E-mail: trsharma1965@gmail.com; trsharma@nrcpb.org.

Accepted: August 6, 2016

Abstract

Leaf rust is one of the most important diseases of wheat and is caused by *Puccinia triticina*, a highly variable rust pathogen prevalent worldwide. Decoding the genome of this pathogen will help in unraveling the molecular basis of its evolution and in the identification of genes responsible for its various biological functions. We generated high quality draft genome sequences (approximately 100- 106 Mb) of two races of *P. triticina*; the variable and virulent Race77 and the old, avirulent Race106. The genomes of races 77 and 106 had 33X and 27X coverage, respectively. We predicted 27678 and 26384 genes, with average lengths of 1,129 and 1,086 bases in races 77 and 106, respectively and found that the genomes consisted of 37.49% and 39.99% repetitive sequences. Genome wide comparative analysis revealed that Race77 differs substantially from Race106 with regard to segmental duplication (SD), repeat element, and SNP/InDel characteristics. Comparative analyses showed that Race 77 is a recent, highly variable and adapted Race compared with Race106. Further sequence analyses of 13 additional pathotypes of Race77 clearly differentiated the recent, active and virulent, from the older pathotypes. Average densities of 2.4 SNPs and 0.32 InDels per kb were obtained for all *P. triticina* pathotypes. Secretome analysis demonstrated that Race77 has more virulence factors than Race 106, which may be responsible for the greater degree of adaptation of this pathogen. We also found that genes under greater selection pressure were conserved in the genomes of both races, and may affect functions crucial for the higher levels of virulence factors in Race77. This study provides insights into the genome structure, genome organization, molecular basis of variation, and pathogenicity of *P. triticina*. The genome sequence data generated in this study have been submitted to public domain databases and will be an important resource for comparative genomics studies of the more than 4000 existing *Puccinia* species.

Key words: *Puccinia triticina*, fungal pathogen, genome sequencing, molecular evolution, comparative genomics, genome variations.

Introduction

Wheat, a staple food of more than 50% of the world's population, is severely affected by three types of rust diseases, leaf or brown rust (*Puccinia triticina* Eriks); stem or black rust (*Puccinia graminis* Pers. f. sp. *tritici*); and stripe or yellow rust (*Puccinia striiformis* Westend f. sp. *tritici*). In India, leaf rust is

the most important and prevalent of these three diseases. The temporal and spatial patterns of occurrence of all of these rust diseases are variable and differ from one another (Joshi et al. 1986). Stem and stripe rusts are restricted to certain parts of India, whereas leaf rust is prevalent all over the country and appears in the months of February–March, when the wheat is in anthesis, or grain formation stage. Among the three rusts, leaf rust is comparatively more frequent worldwide and results in more wheat yield losses, due to reduced kernel weight and decreased numbers of kernels per head (German et al. 2007; Herrera-Foessel et al. 2011), while economic losses due to stem and stripe rusts are relatively low (Thind 1998). Draz et al. (2015) published a detailed survey of wheat yield losses in Egypt due to leaf rust and reported that it could reach up to 50%. Losses in kernel weight among wheat varieties due to leaf rust infection can range between 2.0% and 41% according to the resistance level of wheat varieties (Bajwa et al. 1986). The mean yield losses for susceptible, race-specific, and slow-rusting genotypes account for 51%, 5%, and 26%, respectively, in normally sown crops and 71%, 11%, and 44% in late sown crops (Herrera-Foessel et al. 2011). Among other approaches, leaf rust disease can be effectively managed using resistance (*R*) genes; however, the most challenging task for breeding programs is to incorporate durable leaf rust resistance genes into high yielding cultivars. The frequent appearance of new variants of *P. triticina* and shifts in virulence patterns are major problems for leaf rust management.

Puccinia triticina, an obligate biotroph, is a basidiomycete fungus with an estimated genome size of 100–120 Mb (http://www.broadinstitute.org/annotation/genome/puccinia_group/Info.html, last accessed 20 August 2016). This fungus has heteroecious life cycle in which the sexual phase is restricted to an alternate host; however, the major phase of the life cycle is completed on graminaceous hosts (Savile 2004).The availability of rust fungus genomic resources has facilitated extensive studies of the epidemiological, genetic, and molecular characteristics of leaf rust (Ayliffe and Lagudah 2004; Duplessis et al. 2011; Xu et al. 2011). Genome sequence data from a Canadian race of *P. triticina* (Race1), *P. graminis*, and *P. striiformis* are now available in the public domain (http://www.broadinstitute.org/annotation/genome/puccinia_group/Info.html; Cantu et al. 2011; Duplessis et al. 2011).

In India, efforts to identify new pathotypes of *P. triticina* and screening of wheat germplasm for resistance to rusts have been ongoing since 1930 (Jain et al. 2004; Bhardwaj et al. 2006). Initially, the enormous variants of each *formae speciale* of rust species were called as races and are further divided into pathotypes, which are differentiated on the basis of infection types on a set of differentials (Bhardwaj 2012). To separate minor differences within races, additional supplementary differentials were added to characterize pathotypes within a race and an international system for the identification of races of brown rust was given by Johnston and Mains (1932). Different countries have different systems of identification of races and various other research institutions have developed their own systems of analysis and race designation (Park 1996; Huerta-Espino et al. 2011). Kolmer (1997) preferred to call rust variants as "isolates" or "virulence phenotypes". Based on published literature on race and pathotype designation of Indian population of *P. triticina* (Bhardwaj et al. 2006; Bhardwaj 2012; Manjunatha et al. 2015), the terms pathotype has been used in the present study to distinguish the variants present within Race 77.

New pathotypes of *P. triticina* originate on wheat through mutation, parasexuality, and introduction from other, often unknown, areas. The pathotypes detected to date are maintained as live cultures and also cryopreserved at the Regional Station, Indian Institute of Wheat and Barley Research, Flowerdale, Shimla, India. Indian pathotypes are different from those found in the United States of America and adjoining countries (Bhardwaj et al. 2006). Of many races of *P. triticina* present in India, Race77, which was first detected in 1954, has been found to have 13 pathotypes/variants, whereas Race106, detected in 1934, is stable and not a single pathotype has been reported to date (Bhardwaj 2013). Race77 and its 13 pathotypes are highly virulent and are predominant under Indian conditions (Nayar et al. 1996; Bhardwaj 2011). Therefore, it is very pertinent to understand the molecular mechanisms of virulence and variation within Race77, and to unravel the molecular basis of the rapid evolution of Race77, compared with Race106. The objectives of the present study were (i) to generate high quality draft genome sequences of *P. triticina* races 106 and 77 (including its 13 pathotypes) using NGS methods, and (ii) to perform intra-species genome wide comparative analyses of the two distinct races, one of which is highly unstable and diversified, while the other has been absolutely uniform since its existence.

Results

Sequencing and Assembly of *P. triticina* Race77 and Race 106 Genomes

Two individual races of *P. triticina*, 77 and 106, were selected for *de novo* whole genome sequencing based on their prevalence and virulence, to provide a genetic platform for analysis of *P. triticina*, an obligate fungal parasite capable of producing infectious urediniospores on living leaf tissues (fig. 1*A*). A total of 3.41 and 2.91 Gb of high-quality sequence data were generated for races 77 and 106, respectively, using the 454 GSFLX platform (table 1). In addition, 1552 fosmid reads (594,676 bases), generated by Sanger sequencing, were used in an incremental assembly to improve the scaffold size. These data represent approximately 33.9× and 27.2× coverage of the genomes of races 77 and 106, respectively.

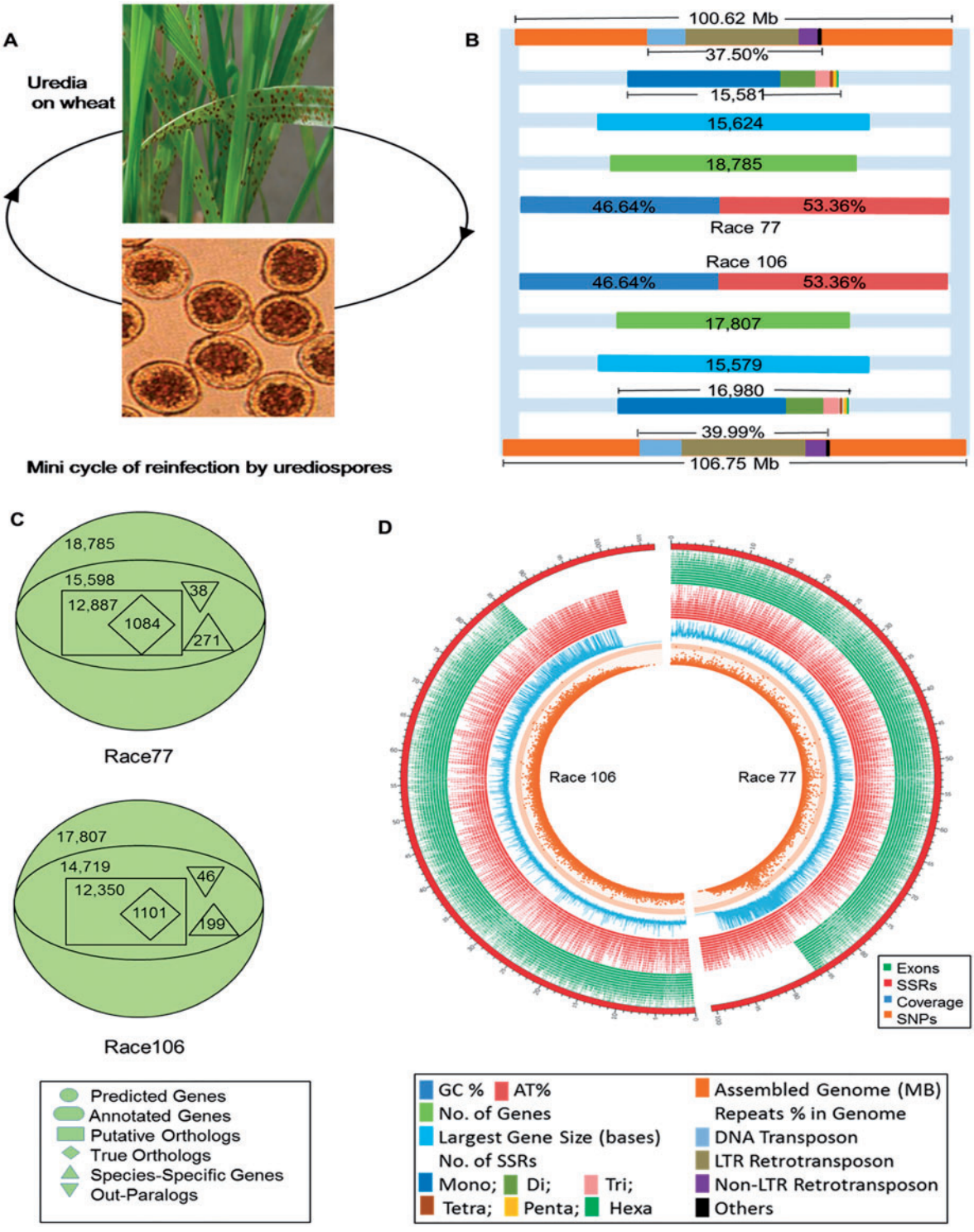

Fig. 1.—Genome structure of Race77 and Race106 based on assembly and repeat contents. (*A*) Asexual life cycle of *P. triticina*: Uredia on leaf containing single cells dikaryotic uredinospores originating from aeciospores or urediospores. This asexual uredinial stage is repeated on the wheat host as long as favorable conditions for infection occur. (*B*) Assembled genome size (orange), percentage of repeats in assembled genome (37.50% and 39.99% in Race77 and Race106, respectively), number of SSRs predicted (15,581 and 16,980 in Race77 and Race106, respectively), largest gene size (sky blue), number of predicted genes (light green), percentage of AT (red), and GC (blue) content in the assembled genome; (*C*) Distribution of predicted genes in Race77 and Race106 as functionally annotated genes, putative orthologs, true orthologs, species-specific genes and out-paralogs; (*D*) Genome wide distribution of exons, SSRs, coverage and SNPs in Race77 and Race106.

Table 1
Assembly and gene prediction statistics of *P. triticina* Race77 and Race106 genomes

Assembly Parameters	Race77	Race106
Input reads	9,396,376 (3.41 Gb)	7,453,304 (2.91 Gb)
Total contigs	44,586 (100.62 Mb)	67,044 (106.75 Mb)
N50 (contigs) (bp)	6,040	4,341
Average contig length (bp)	2,256	1,592
Largest contig (bp)	69,400	45,977
Contigs ≥ 2K	13,235 (81.22 Mb)	13,780 (75.69 Mb)
Contigs ≥ 200 bases	37,447 (99.59 Mb)	55,512 (105.06 Mb)
Average contig length (>2k bp)	6,136	5,492
N50 (>2k contigs) (bp)	7,763	6,680
Total scaffolds	2,651 (102.20 Mb)	7,448 (93.99 Mb)
Average scaffold size (bp)	38,552	12,619
Largest scaffold (bp)	613,912	193,730
N50 (scaffolds)	102,413	20,667
Depth coverage	33×	27×
Number of genes predicted	27,678	26,384
Mean gene length	1129 bp	1086 bp
Total Number of Exons	124,443	117,168
Mean number of exons per gene	4.49	4.44
Largest Gene Length (bp)	15,624	15,579
Genes (>150 bases)	25,742	24, 446
Genes (>450 bases)	18,785	17,807
Average gene length (bp) (>450 bases genes)	1,545	1,488
Number of exons (>450 bases genes)	100,914	94,656

Genomes were initially assembled using three different software packages to determine the most suitable method (supplementary fig. S1 and table S1, supplementary material online). Final assembly was performed using GS Assembler (Newbler 2.5.3) to generate a draft genome of 100.62 Mb spanning 44,586 contigs for Race 77 and 106.75 Mb with 67,044 contigs for Race 106 (fig. 1*B*, Table 1). The N50 lengths of the assemblies were 6.04 and 4.34 kb for races 77 and 106, respectively. These contigs were then individually assembled into 2,651 and 7,448 scaffolds for races 77 and 106, respectively. Considering the heterozygous nature of the *P. triticina* genome, in order to assess the genetic variation between two independent nuclei of the dikaryotic spores, reads from races 77 and 106 were aligned to their respective assembled contigs to predict intra-race specific SNPs. On an average 3.82 ±1.50 SNPs/kb were identified within the two races. When the reads of Race 77 were aligned to the assembled contigs of Race 106 and vice-versa to predict inter-racial variations, we found that the number of heterokaryotic SNPs (2.57 ± 0.92 SNPs/kb) was similar to that of homokaryotic SNPs (2.35 ± 0.70 SNPs/kb) (supplementary table S2, Supplementary Material online).

Validation of Sequence Assembly

Assemblies generated by the three different assemblers, Newbler 2.5.3, SeqManNGen 4.0.1 (in DNA STAR), and CLC Genomics Workbench 6, were compared and validated using a Core Eukaryotic Genes Mapping Approach (CEGMA) and Assemblathon 2 scripts. To ensure the consistency of comparisons, we considered the estimated genome size of *P. triticina* Race 1 of 110 Mb (Puccinia Group Sequencing Project, Broad Institute of Harvard and MIT (http://www.broadinstitute.org, last accessed 20 August 2016) as a reference and specified the minimum contig length as 100 bp, with default parameters. An Assemblathon 2 script was used to calculate NG (X) values, (X = 1–99). The main aspects considered in comparison of the three methods were the number of contigs, assembled genome contig sequences (≥25 kb), largest contig size, and (more importantly) N50 and NG50 values. Validation results showed that Newbler assembler 2.5.3 was more suitable for our data, with higher NG (X) values for X = 1–56 (supplementary fig. S1, Supplementary Material online). By CEGMA the overall percentage of 248 ultra-conserved CEGs present in the assembled genomes was found to be optimal with assembly using Newbler (91.53%) (supplementary table S3, Supplementary Material online). Hence, the probability of identifying genes in the assembled genome was higher using the assembly generated with Newbler.

Gene Prediction and Functional Annotation

A total of 27,678 and 26,384 genes were predicted in the genomes of races 77 and 106, respectively. For Race 77, the average length of the predicted genes was 1,129 bases, with a mean of 4.49 exons per gene, whereas for Race 106, the average length was 1,086 bases, with 4.44 exons per gene (table 1). These genes comprised 31% and 26.8% of the genome sequences of races 77 and 106, respectively, with average gene densities of one gene per 3.63 kb and one per 4.04 kb. Assessment of the quality of gene prediction was performed by comparing the length distribution of the genes, CDS, exons, and introns, and the distribution of exon numbers and GC ratios of CDS per gene in both the genomes. All of the major parameters analyzed showed similar patterns in both genomes (supplementary fig.S2, Supplementary Material online). Further, genes longer than 450 bp (150 aa) from Race 77 (18,785) and Race 106 (17,807) were subjected to BLAST search against the NCBI expressed sequence tag (EST) database, resulting in 27% of gene sequences identified as expressed in both the races (supplementary table S4A, Supplementary Material online). All predicted genes (18,785 and 17,807) from both the races were then individually BLAST searched against NCBI-nr database and their coding regions comprised of 45.10% (Race77) and 39.05% (Race106) of

their assembled genome (supplementary table S4B, Supplementary Material online).

For functional annotation, predicted genes >450 bp from both the races were considered further. Functions were assigned only to those genes, which showed significant hits (15,598 and 14,719 in races 77 and 106, respectively) with *E*-values $\leq e^{-10}$. No significant hits in the nr database were found for 3,187 (16.96%) and 3,088 (17.34%) genes in races 77 and 106, respectively. Approximately 44% of the genes (excluding those in hypothetical, unclassified, and unknown categories) were associated with known functions (supplementary table S5, Supplementary Material online). Comparative analyses of the distribution of the genes in different functional classes revealed that genes belonging to the class "Mobile and extrachromosomal elements" showed some differences between the two races, with 1.2% Increased genes in this category in Race 77 (supplementary fig. S3A, Supplementary Material online). Further, to understand the interplay between ortholog, paralog, and speciation events, annotated genes of both races were BLAST searched against one another. In this analysis, 82.61% of annotated genes from Race 77 were found to be most similar to annotated genes of Race 106 with the same function, and hence predicted as orthologs to Race 106. Similarly, 83.90% of annotated genes of Race 106 were predicted as putative orthologs to Race 77 genes (fig. 1*C*, supplementary fig. S3B, Supplementary Material online). Race 106 was found to have 1101 true orthologs (putative orthologs with 100% sequence identity) as compared to1084 in Race 77. A high percentage of the true ortholog genes identified in Race 77 belonged to the class "Transport and Binding Proteins", whereas in Race 106 a greater percentage were annotated as "Energy Metabolism" (supplementary fig. S3C, Supplementary Material online). We found 271 genes specific to Race 77, whereas 199 genes were found to be Race 106 specific (fig. 1*C*). Categorization of these genes into functional classes (excluding hypothetical genes) revealed unique differences in various classes (supplementary table S6 and supplementary fig. S3D, Supplementary Material online). In spite of its lower number of predicted, annotated, and race-specific genes, Race 106 was found to contain higher percentages of true orthologs and out-paralogs (supplementary table S4B, Supplementary Material online). A comprehensive circular plot was generated to illustrate the total genome coverage after assembly, along with the distributional pattern of identified exons, SNPs, and simple sequence repeats (SSRs, microsatellites)in both races (fig. 1*D*).

Repetitive Elements in Race 77 and Race 106

De novo identification of dispersed repeats (Transposable Elements, TE) having at least 80% homology against the Fungi REPBASE library using the MapRep module of MolQuest software revealed the presence of 37.73 and 42.69 Mb of repeats in the genomes of Race 77 and Race 106, respectively, corresponding to 37.49% and 39.99% of their respective assembled genomes (fig 1*B*). A total of 36 different repeat categories were identified and these were grouped into three majo classes; LTR retrotransposons, non-LTR retrotransposons, and DNA transposon elements. The majority of repetitive sequences were retrotransposons (76.82% in Race 77 and 77.08% in Race 106). Individually LTR retrotransposons were predominant among repeat elements (66.7% and 66.5% in races 77 and 106, respectively), compared with non-LTR and DNA transposons, which made up approximately10% and 22% of repeat elements in both races. A very small proportion (approximately 1%) of elements did not fall into any of these major groups (although they may have important and specific roles in genome organization) and thus were grouped as "Others". These repeats consisted of paralogs, tDNAs, and retro-pseudogenes (supplementary fig. S4A, Supplementary Material online). The most abundant repeats were LTR retrotransposon *Gypsy*-type elements (approximately 40%) whereas *Copia*-type elements corresponded to approximately 23% in both races (supplementary fig. S4B, Supplementary Material online). In order to identify functional protein conserved domains within *Gypsy* and *Copia* elements, we screened them against the NCBI Conserved Domain Database (CDD). Of 34 domains identified among *Copia* elements, four were specific to Race 106 and one to Race 77. Screening with *Gypsy* elements identified two specific protein domains for Race 106 and four for Race 77. Some of the domains present in both genomes had specificity between *Gypsy* and *Copia* elements (supplementary fig. S4C, Supplementary Material online). The LTR retrotransposons in both races showed a marked difference in their occurrence (supplementary notes 1, Supplementary Material online).

Tandem (satellite) repeats were identified in genic (5′UTR, CDS, and 3′UTR) and non-genic (introns and intergenic) regions using MISA and Tandem Repeats Finder 4.07b software. The frequency of SSR repeats was similar between genic (18.6%) and non-genic (17.1%) regions in Race 77; however, a variation of approximately 7% was identified between genic (19.6%) and non-genic (12.3%) regions in Race 106 (supplementary notes 1, Supplementary Material online). In both races, minisatellites (>16 bp) made up approximately 33.2% and 38.4% of genic and non-genic regions, respectively (supplementary fig. S4D, Supplementary Material online). Megasatellite (>100bp) content was lower in Race 77 (4.1%) compared with Race 106 (5.7%) in genic regions; however, it was 5.8% in non-genic regions in both races (supplementary fig. S4E, Supplementary Material online). Of 18 genes identified in megasatellite repeats, 12 were present in Race 77 and 6 in Race 106. These genes mainly code for proteins involved in cellular processes, TE-related proteins, and hypothetical categories (supplementary fig. S5, Supplementary Material online).

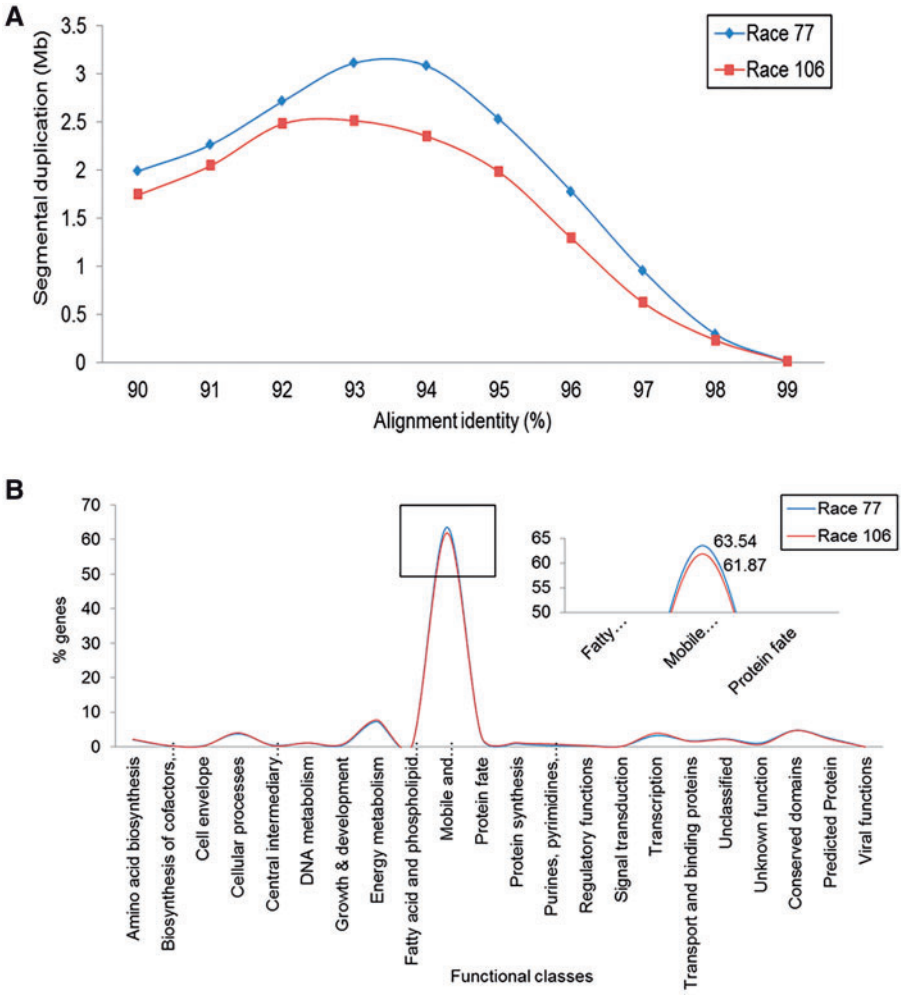

Fig. 2.—Segmental duplications within Race77 and Race106 genomes. (*A*) Comparison of Segmental duplications (SDs) among genomes of Race77 and Race106. Both the genomes showed alignments at all levels of identity, from 90 to 99%; however, the Race77 genome has greater amount of aligned bases relative to the Race106, for each of the identity level. (*B*) Functional annotation (excluding hypothetical genes) of genes predicted in SD regions showed a difference of approximately 2% genes falling under category 'Mobile and extra-chromosomal elements' in Race77 as compared to Race106.

Segmental Duplications in Race 77 and Race 106

Segmental duplications (SDs) with at least 90% identity and ≥1kb alignment length were identified using a BLAST-based whole-genome assembly comparison (WGAC). The results revealed that 18.60% of the genome of Race 77 was composed of SDs compared with 14.30% of the Race 106 genome. Recent segmental duplications (SDs with ≥95% identity) were more frequent at 5.52% of the genome in Race 77 compared with 3.86% in Race 106 (supplementary table S7, Supplementary Material online). By comparison of SDs at different levels of identity (90%–99%), we observed that the Race 77 genome contained a larger proportion of SD regions than Race 106, at each level of identity (fig. 2*A*).

In order to show that no haplotype sequences were identified within SD regions further analysis was performed. As GS Assembler utilizes a read as a whole or a portion of the read, only once during the entire process of de-novo genome assembly to form a contig and the output files contained only non-redundant contigs, the possibility that the SDs identified within contigs in this study contained any haplotype sequences were minimal. To confirm this, the nucleotide sequences of an identified SD region in each of the two contigs of a randomly selected SD pair was mapped back to total raw reads of its respective genome using CLC Genomics Workbench 7.5.1, and the mapped reads were aligned to check the conservation of reads across the aligned region. The results revealed that none of the reads showed 100% conservation at a particular position of alignment with maximum conservation levels of 34.70% for Race 77 and 43% for Race 106 (supplementary fig. S6, Supplementary Material

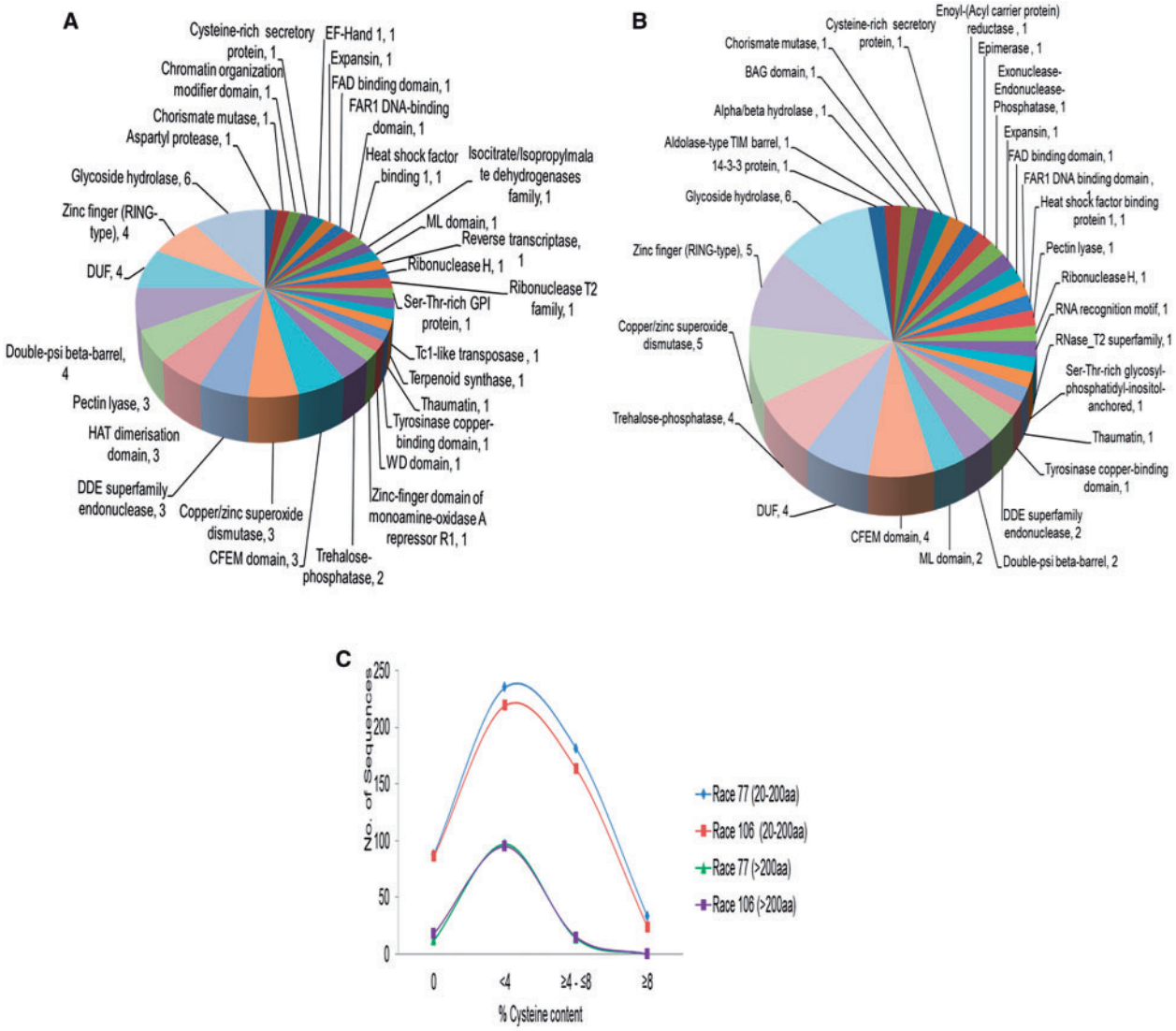

Fig. 3.—Secretome analysis in the whole genome of Race 77 and Race 106. Functional annotation of the secretory proteins with conserved domains in Race 77 (*A*) and Race 106 (*B*). (*C*) Secretory proteins of both the races featuring higher number of sequences containing cysteine mature peptide content of length 20 to ≤ 200 aa in Race77 as compared to Race106.

online). Moreover, no aligned stretches >20 bp were identified among the mapped reads, nullifying the possibility of any true copy of haplotype sequences accounting for SD regions. Within SD regions, a large number of genes were present in the "extra chromosomal and repeat elements" category in both genomes and there were approximately 2% greater genes of this category present in Race 77 compared with Race 106 (fig. 2*B*).

Secretome Analysis of the Race 77 and 106 Genomes

The degree of variation in the secretomes of both races was investigated using SignalP, TargetP, TMHMM, and ProtComp. We identified 660 secretory proteins in Race 77 and 620 in Race 106 (supplementary table S8, Supplementary Material online). The percentage of secretory genes producing significant hits was 3.4% greater in Race 77 than in Race 106 when these proteins were BLAST searched against at the NCBI EST database (supplementary fig. S7, Supplementary Material online). Functional annotation of these secretory genes revealed the presence of carbohydrate active enzymes, antioxidant enzymes, and proteins involved in host cell wall loosening (fig. 3*A* and *B*). Analysis of the cysteine content and size distribution within secretomes demonstrated that Race 77 possessed 416 peptides with >4–8% cysteine content among peptides with mature lengths of 20–200 amino

acids (aa), as compared to 384 peptides in Race 106, whereas proteins with >200 aa produced similar results in both races (fig. 3*C* and supplementary tables S9–S12, Supplementary Material online).

Pathogenicity Related Genes in Race 77 and Race 106

Virulence factors are the most important class of proteins in pathogens, as they can counteract the defense mechanisms of the host and enhance the spread of the pathogen, as well as facilitate the release of nutrients and water from host cells (Toth et al. 2003). Currently, the number of reported fungal virulence factors is limited, and the majority of data are only available in the literature (Idnurm and Howlett 2001; de Wit et al. 2009; Stergiopoulos and de Wit 2009; Van de et al. 2011; Gonzalez-Fernandez and Jorrin-Novo 2012). For the identification of pathogenicity related genes, we BLAST searched the predicted proteome data against PHIbase (www.phibase.org [last accessed 20 August 2016], Winnenburg et al 2006) and found that 6% (1,665 genes) of Race 77 genes and 5.7% (1,530 genes) of Race 106 genes account for pathogenicity related functions. A difference of 2% in the number of genes with reduced virulence between Race 77 and Race 106 was observed (fig. 4*A*). Since, the phenotypic and gene function information within the PHI-base, has been obtained by manual curation of the peer-reviewed literature with citations in more than 122 published articles including genetics, genomics, and bioinformatics research and review articles (Urban et al. 2015). Thus, indicating, that the potential pathogenicity genes identified in both the races may play important roles in the infection and development of the rust pathogen.

The majority of these genes were functionally annotated as "Energy metabolism" and "Mobile and Extra chromosomal elements" in both races. However, Race 106 possessed high percentage of genes categorized under "Energy metabolism", while Race 77 possessed a higher percentage of genes classified as "Extra-chromosomal elements" (fig. 4*B*).

Genome Wide Analysis of SNP and InDels in Race 77 and Race 106

Whole genome SNP and InDel analysis was performed to identify mutation events in both races. Analysis was carried out at a depth of 7X, with 136,082 and 228,738 SNPs identified in races 77 and 106, respectively. A total of 21,156 and 27,495 InDels were identified in races 77 and 106, respectively. The InDels were dispersed throughout the genomes and there were 11,505 and 11,269 insertions and 9,651 and 16,226 deletions in races 77 and 106, respectively. To further study the distribution patterns of SNPs in various genomic regions, we performed analysis using "snpEff" software. Large numbers of SNPs were identified within the exonic region in both races with 44.6% in Race 77 and 40% in Race 106, whereas there was a difference of 7% SNPs in the upstream regions of the genes between both the races with Race 77 possessing 23.7% SNPs and Rcae 106 possessing 30.5% SNPs. The lowest percentages of SNPs were identified in intronic and splice site regions of both races (fig. 5*A*). Breakdown of exonic SNPs into missense, non-sense, and silent variants revealed that the proportion of non-synonymous mutations was 29.2% in Race 77 and 26.3% in Race 106, while the proportions of synonymous SNPs were 15.4% in Race 77 and 13.7% in Race 106 (fig. 5*B*). These differences were relative to the reference genome (Race 1) used in the study.

Re-Sequencing of 13 Pathotypes of Race 77

In order to explore the intra-specific race evolution of *P. triticina*, 13 additional pathotypes of Race 77, collected from different geographical locations in India, were re-sequenced using Illumina platform Hiseq 1000 (supplementary tables S13 and S14, Supplementary Material online). For each genome, an average of 7Gb of raw sequence reads were generated and 70–93% of reads were mapped to the 97Mb Race 77 assembled reference genome with 66–94X depth coverage (table 2). The total numbers of genes predicted in all pathotypes ranged from 32,180 to 32,894. Phylogenetic analysis of these pathotypes, using the whole genome data, revealed their emergence from Race 77, with the recent, additionally active and virulent pathotypes, 77-9, 77-10, and 77-11 in one clade and the remaining pathotypes assigned to another clade (fig. 6*A*). An average of 273,845 SNPs and 35,945 InDels were identified in each pathotype, relative to the Race 77 reference genome. SNPs in coding regions accounted for approximately 21% of all SNPs, of which a large proportion (68%) were non-synonymous mutations. The average densities of SNPs and InDels in all pathotypes were 2.4 and 0.32 per kb, respectively. A slightly higher number of SNPs were identified in pathotypes 77-3, 77-9, 77-10, and 77-11, whereas the number of InDels varied randomly among all pathotypes (supplementary table S15, Supplementary Material online). To visualize the results, a comprehensive circular plot of SNP coverage was generated for all the 13 pathotypes (fig. 6*B*). To determine the density distribution of SNPs in all pathotypes along with Race 77, contig containing the maximum number of SNPs (average lengths approximately 16,000–22,000 bp) were split into 500 bp blocks. Density plots of SNPs in the genic and intergenic regions of pathotypes 77-1 to 77-7, 77A, and 77A-1 showed similar distributions. A detailed whole genome SNP analysis of the pathotypes is provided in supplementary figure S8, Supplementary Material online. However, higher densities of SNPs were identified in genic, than in intergenic, regions in pathotypes 77-8, 77-9, 77-10, and 77-11 (fig. 6*C*). Analysis of microsatellites showed that apart from the mono-nucleotide repeats, di-nucleotide repeats were the most abundant,

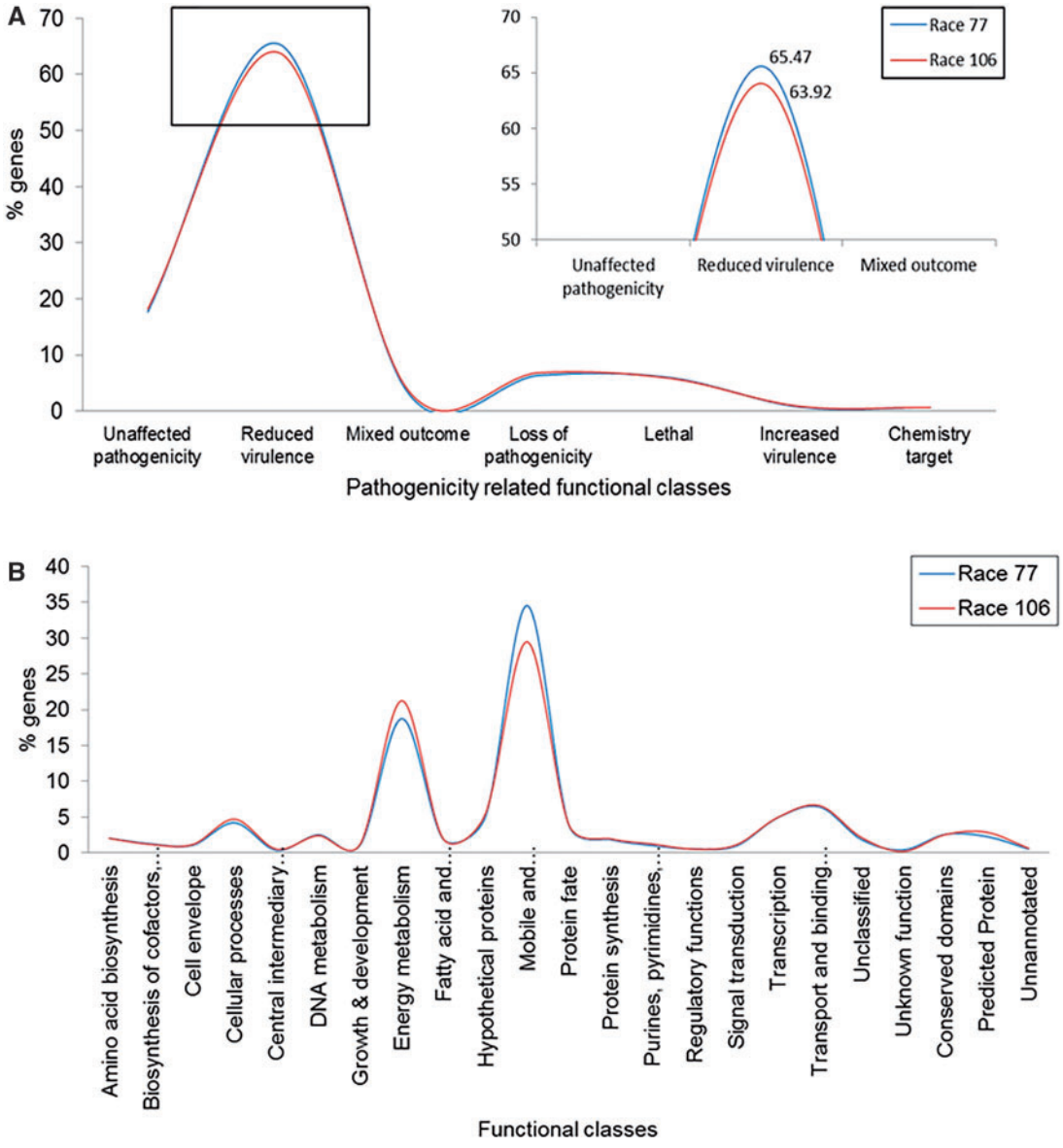

Fig. 4.—Pathogenicity genes identified in Race 77 and Race 106 (*A*) Pathogenicity genes having enhanced percentage of genes with "reduced virulence" in Race77. (*B*) Functional annotation (excluding hypothetical genes) of the genes showed a higher percentage of genes falling under category "Energy metabolism" and "Mobile and extra-chromosomal elements" in Race77 as compared to Race106.

followed by tri-nucleotide repeats; in addition, penta-nucleotide repeats were over represented among other repeat types (fig. 6*D*).

Transcriptome Sequencing of Race 106 and Pathotype 77-5

Race 106 is a highly avirulent and uniform Race of *P. triticina*, whereas pathotype 77-5 is one of the most virulent and a predominant pathotype of Race 77. These are, therefore, exceptional and diverse in terms of their wheat infection characteristics. Therefore, transcriptome sequencing of these two genomes was performed to discover, characterize, and compare the genes expressed by both genomes. For transcriptome profiling, mRNA from Un-germinated (UG) and 16 h germinated (G) urediniospores of Race 106 and pathotype 77-5 were used to construct four cDNA libraries; *P.t* 77-5 UG, *P.t* 77-5 G, *P.t* 106 UG, and *P.t* 106 G. The sequence data were generated on an Illumina HiSeq 2000 platform. After removing adaptor and low quality reads, approximately 71 and 72 million high quality reads were obtained for the two libraries of pathotype 77-5, and 67 and 77 million high quality were obtained for the race 106 libraries. The final unigene data set obtained from all four libraries was used for further analysis (supplementary table S16, Supplementary Material

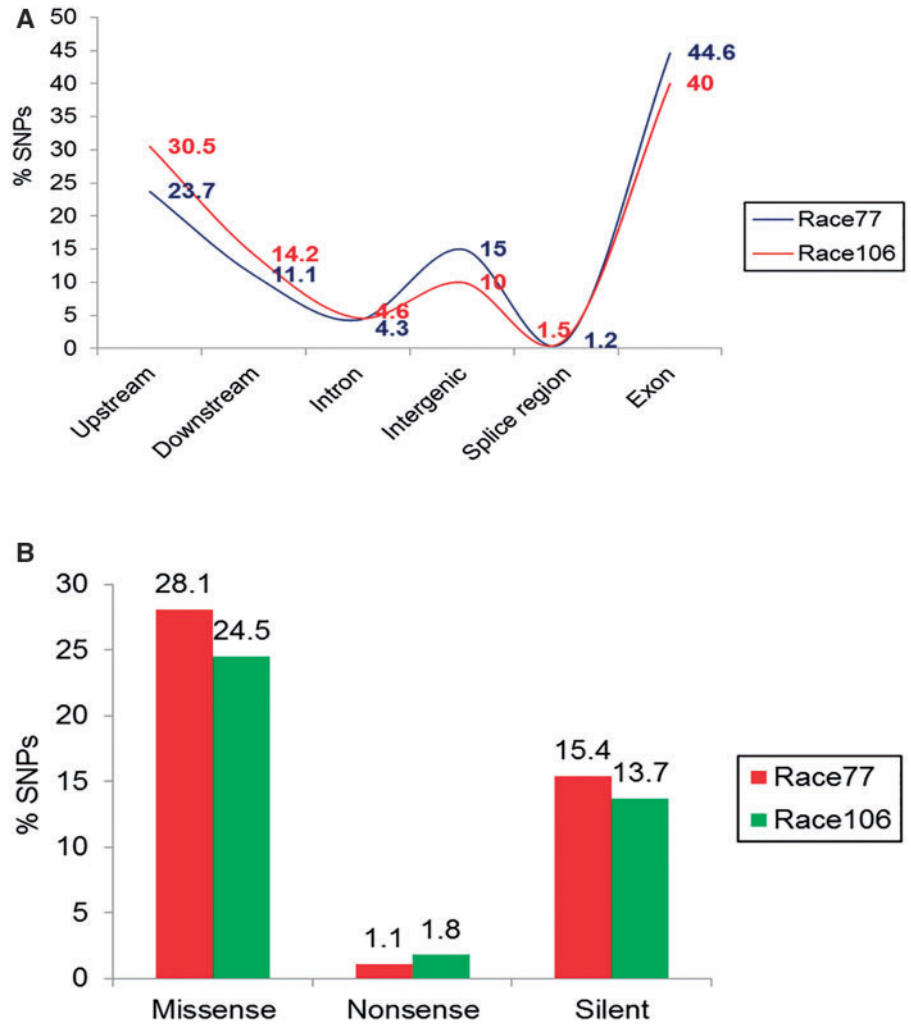

Fig. 5.—Genome wide SNP analysis in Race 77 and Race 106 (*A*) Single Nucleotide polymorphism (SNP) investigated in various regions of both genomes resulted higher percentages (30.5 and 14.2, respectively) in upstream and downstream non-coding regulatory regions of Race106. The intergenic region in Race77 has 15% SNPs as compared to 10% in Race106. SNPs in the intronic and splice regions were almost similar in both races, but in the exonic regions Race77 showed 4% increase as compared to Race106. (*B*) Distribution of types of SNPs (synonymous and non-synonymous) in the coding region of both the races.

online). Reference mapping of the assembled reads was performed using the software tools Bowtie 2 V 2.1.0, Samtools V 0.10.1, Bcftools V0.1.17, and FastQC V 0.10.1, generating 81.5% and 83.1% mate-mapped reads out of the total mapped reads for the pathotypes 77-5 libraries, and 83.77% and 82.1% for the Race 106 libraries from UG and G spore stages, respectively (supplementary table S17, Supplementary Material online). As unigene sequences are derived from the expressed genome, the markers developed from these resources can be assayed as gene-based functional markers for diversity assessment and gene mapping. We also characterized the unigene sequences with SSRs, SNPs, and InDels identified in the respective genomes (supplementary notes 2, Supplementary Material online).

Identification of Novel Genes in Race 106 and Pathotype 77-5

Transcripts from ungerminated spores and 16 h germinated spores were individually compared with the whole genome sequence data of both genomes. A total of 114 and 124 novel genes were identified in the geminated spores from pathotype 77-5 and Race 106, respectively, whereas 90 and 103 novel genes were identified in the resting spores (fig. 7). Significant BLAST hits (100% identity over full length) of the identified novel genes against the predicted genes of their respective genomes was achieved for 68 genes out of 90 and 116 genes out of 124 in the resting spores (UG) and 91genes out of 114 and 93 genes out of 103 in the germinated spores (G) in pathotype 77-5 and Race 106 respectively.

Table 2

Assembly and gene prediction analysis of 13 pathotypes of *P. triticina*Race77 genomes

Race 77 pathotypes	77-1	77-2	77-3	77-4	77-5	77-6	77-7	77-8	77-9	77-10	77-11	77-A	77-A1
Input reads	81,299,142 (7.91 Gb)	79,582,258 (7.78 Gb)	91,809,084 (8.93 Gb)	75,302,524 (7.34 Gb)	78,415,262 (7.64 Gb)	74,488,738 (7.30 Gb)	93,219,647 (9.15 Gb)	83,879,510 (8.22 Gb)	82,333,846 (8.05 Gb)	82,184,126 (8.06 Gb)	70,017,488 (6.86 Gb)	65,938,280 (6.45 Gb)	81,768,946 (7.98 Gb)
Mapped reads	90.04%	88.30%	88.73%	93.67%	91.26%	92.08%	90.43%	91.16%	89.63%	90.22%	88.84%	70.94%	91.54%
Mapped bases	91.35%	89.22%	89.96%	94.89%	92.51%	92.99%	91.23%	91.95%	90.55%	90.72%	89.65%	71.64%	92.62%
Total contigs	64,684 (97.26 Mb)	64,740 (97.25 Mb)	64,499 (97.33 Mb)	64,739 (97.22 Mb)	64,735 (97.25 Mb)	64,853 (97.22 Mb)	64,534 (97.26 Mb)	65,595 (96.41 Mb)	67,081 (96.14 Mb)	66,054 (95.98 Mb)	67,096 (95.94 Mb)	65,649 (97.02 Mb)	64,580 (97.25 Mb)
Avg. contig length	1503	1502	1509	1501	1502	1499	1507	1469	1433	1453	1430	1477	1506
Largest contig (bp)	44,512	44,512	44,513	44,507	44,513	44,509	44,513	44,513	44,512	44,515	44,501	44,503	44,509
Contigs (≥ 2K)	12,821 (69.78 Mb)	12,823 (69.76 Mb)	12,832 (69.92 Mb)	12,814 (69.69 Mb)	12,810 (69.74 Mb)	12,824 (69.70 Mb)	12,817 (69.86 Mb)	12,683 (68.88 Mb)	12,594 (68.47 Mb)	12,559 (68.41 Mb)	12,558 (68.08 Mb)	12,804 (69.15 Mb)	12,808 (69.81 Mb)
Contigs (≥ 500 bases)	31,985 (89.30 Mb)	31,979 (89.27 Mb)	32,005 (89.43 Mb)	32,021 (89.26 Mb)	31,983 (89.28 Mb)	32,034 (89.24 Mb)	31,927 (89.32 Mb)	31,686 (88.24 Mb)	31,559 (87.75 Mb)	31,560 (87.74 Mb)	31,674 (87.52 Mb)	32,217 (88.88 Mb)	31,940 (89.30 Mb)
Avg. contig length (≥ 2k)	5,442	5,440	5,449	5,438	5,444	5,435	5,451	5,430	5,436	5,447	5,421	5,400	5,450
N50 (contig)	6,528	6,493	6,515	6,479	6,513	6,493	6,522	6,484	6,501	6,530	6,460	6,431	6,529
Genes predicted (≥150 bases)	32,824	32,769	32,894	32,745	32,851	32,822	32,757	32,366	32,211	32,180	32,203	32,772	32,747
Genes predicted (≥450 bases)	21,606	21,550	21,624	21,523	21,606	21,609	21,597	21,269	21,126	21,098	21,090	21,491	21,577
Depth coverage	81x	80x	91x	75x	78x	75x	94x	85x	83x	83x	71x	66x	82x

These genes were also found to have matches in the NCBI (*Puccinia triticina)* ESTs database. The presence of these genes was confirmed by performing BLAST searches against the assembled contigs of their respective genomes and hits with100% identity were identified for all novel genes. BLAST analysis of the predicted genes (from assembled genome) was preformed individually for both, Race106 and pathotype 77-5 against their respective RNAseq data, 54.87% genes of Race 106 and 44.18% genes of Pathotype 77-5, showed significant hits (*E*-value $\leq e^{-20}$ and bit score ≥ 100).

Evolutionary Analysis of Different *Puccinia* Strains

We compared races 77, 106, and 1 (Canadian race) of *P. triticina* by calculating the conservation distance matrix from genome alignments using the Mauve software tool. This analysis demonstrated that Race 106 is the most conserved and ancient genome among these three races, while races 1 and 77 appear to have emerged at the same time. The guide tree was consistent with the years of detection of the different strains; i.e., Race 77 (India, 1954), Race1 (Canada, 1954) (Ordonez and Kolmer 2009), and Race 106 (India, 1934) (fig. 8*A* and *B*).

Genome Wide Comparative Analysis of *Puccinia* Species

In order to compare the structural organizations of other publically available wheat rust genomes with respect to their assembled genomes sizes, genes predicted, total repeat element content identified, and AT/GC percentage, genome sequences of *P. graminis* (Race SCCL), and *P. striiformis* (Race 78), and *P. triticina* (Race 1) were downloaded from publically available databases (Puccinia Group Sequencing Project, Broad Institute of Harvard and MIT (http://www.broadinstitute.org [last accessed 20 August 2016], Cuomo et al 2016) .Our results revealed similarity in the %GC (43–46) and %AT (53–56) contents of all genomes. Among *P. triticina* races, the lowest number of genes (13,445) was identified in Race1, followed by Race 106 (17,807), and Race 77 (18,785). The number of genes in *P. graminis* (14,236) was less than that in *P. striiformis* (16,884) (fig. 9*A*).

TE elements are directly related to genome size; therefore, to further investigate the genetic relationships within the *Puccinia* genus, the total percentage of TE content was determined for all five genomes. Interspecies difference with respect to TE element content were observed consistent with their genome sizes; the genomes *P. graminis* and *P. striiformis* had TE contents of only 25.2% and 31%, compared with contents of between 34% and 40% TEs in the genomes of *P. triticina* races (fig. 9*B*).

Discussion

The wheat leaf rust fungus *P. triticina*, is globally distributed, unlike other rust pathogens; therefore, the wheat crop losses

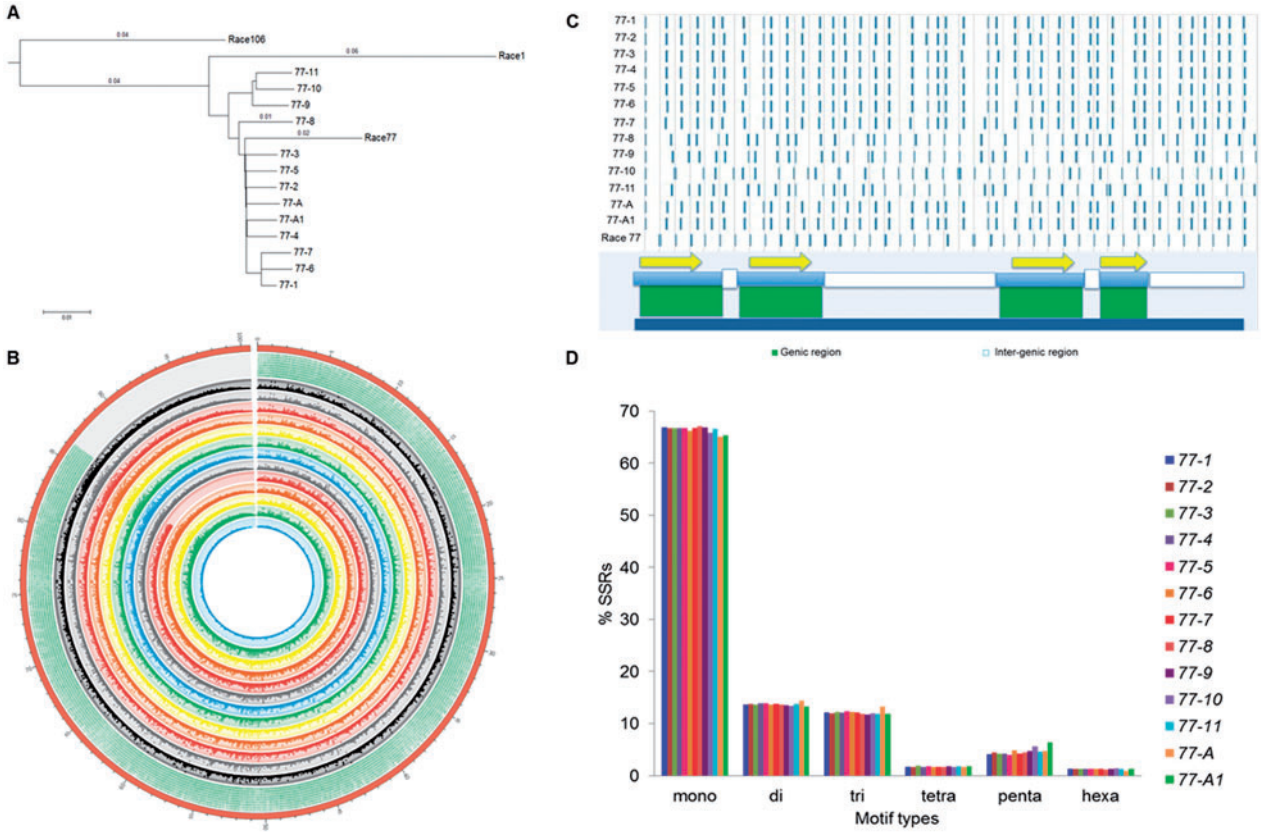

Fig. 6.—Resequencing and genome analysis of 13 pathotypes of Race77. (*A*) Phylogenetic tree of 13 pathotypes of Race77 including Race 106 and Race1, constructed on whole genome sequences featuring evolutionary relationships of the pathotypes with the recent, additionally active and virulent pathotypes, 77-9, 77-10, and 77-11 in one clade and rests are in the other clade (branch length ≥0.01 depicted in the figure). (*B*) Genome wide SNP distribution in all the 13 pathotypes of Race77, 77-1,77-2,77-3,77-4,77-5,77-6,77-7,77-8,77-9,77-10,77-11,77-A, and 77A1 from outer green layer towards inner most circle (*C*) SNP analysis performed on the pathotypes along with Race77 showing contigs richest in number of SNPs searched for their density distribution (marked with blue vertical lines) in 500 bp blocks in the genic and inter-genic regions. (*D*) Percentage occurrence of SSR motifs in 13 pathotypes of Race77.

incurred as a result of rusts are largely attributable to this species. Despite its economic significance, comparatively less genomic information is available for this pathogen. In the present study we generated high quality draft genome sequences of two Indian *P. triticina* races; Race 106, which is one of the oldest races with no known pathotypes to date, and Race 77, a highly virulent and variable race. We also generated genome sequences of 13 existing pathotypes of Race 77. The genome sequencing data from these fifteen pathotypes facilitated the systematic analysis of the *P. triticina* genome, with respect to evolutionary relationships, genome organization, genome duplication, SNP/InDel distribution, and pathogenicity.

Race 106 is relatively uniform, in terms of its prevalence and occurrence on wheat, and no pathotypes of this Race have been reported since its discovery in 1934 (Bhardwaj 2011). Race 106 is avirulent on the bread wheat cultivar "Thatcher", which possesses *Lr22b* gene (Mishra et al. 2001, 2005). In contrast, Race 77 is highly variable, with 13 pathotypes reported since 1954 when it was first detected in India. Previously, the major cause of the emergence of new virulent races of rust fungi was assumed to be spontaneous mutations (Kolmer 2005). *P. triticina* reproduces on wheat by the production of asexual dikaryotic urediniospores in the absence of suitable alternate hosts in India, and the majority of pathotypes of *Puccinia* species on wheat originate through mutation and parasexuality (Bhardwaj 2011). It has been hypothesized that new races of cereal rusts can develop through mutation followed by the selection for virulence against resistance genes deployed in wheat cultivars (Kolmer 1996), although experimental studies on *P. graminis* (Nelson et al. 1955), *P. triticina* (Rodenhiser and Hurd-Karrer 1947), *P. recondita* (Barr et al. 1964), and *P. striiformis* (Wright and Lennard 1980) have suggested the occurrence of somatic hybridization. Parasexual recombination in *P. triticina* is also proposed to occur through the process of germ tube anastomosis (Wang and McCallum,2009). Therefore, even if mutation is

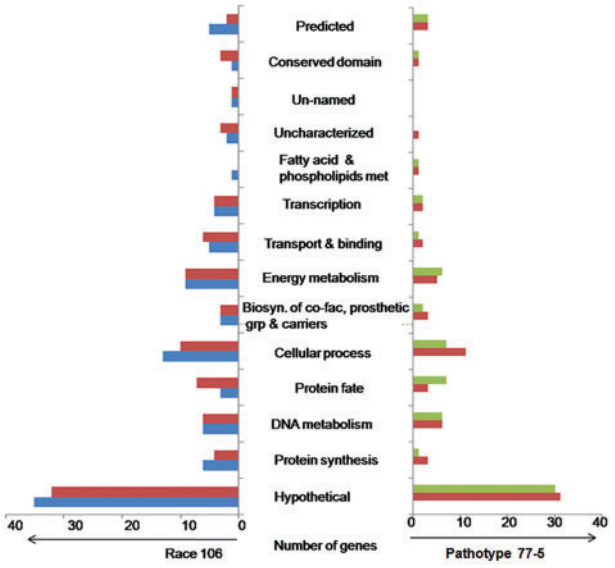

Fig. 7.—Identification of Novel genes in Race 106 and Pathotype 77-5. A set of novel genes were identified using the transcriptome data in the uniform Race 106 (blue bars germinated spores, red bars resting spores) and the most virulent pathotype 77-5 (red bars germinated spores, green bars resting spores) depicting the number and names of the genes post-annotation.

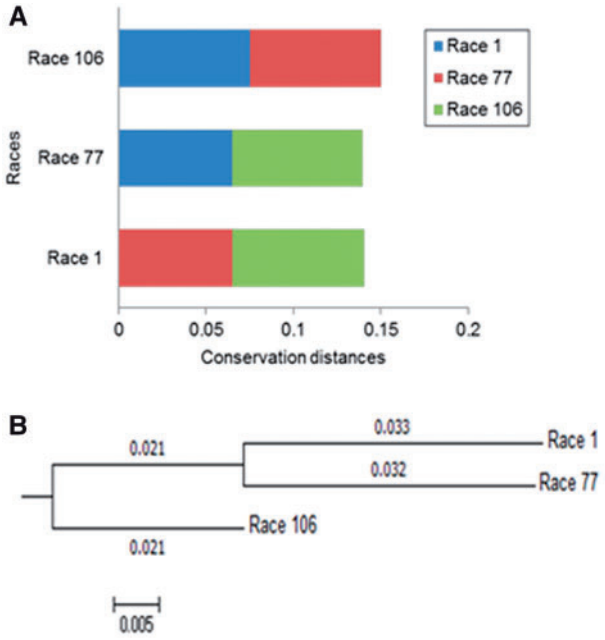

Fig. 8.—Evolutionary pattern of *Puccinia triticina spp*. (*A*) Genome conservation distance between three *Puccinia triticina* races. Race106 seems to be evolved as most conserved and ancient genome among these three, while Race1 and Race77 seemed to be evolved at the same time. (*B*) A guide tree obtained based on the whole genome alignment between three *Puccinia triticina*, i.e., Race1 (Canadian), Race77, and Race106.

believed to be the major source of variation affecting the virulence of cereal rusts, there are reports of virulence phenotypes arising from the same clonal lineage of *Puccinia* species, which often differ only in one or two virulence attributes (Kolmer 1996).

Our comparative genomic analyses identified Race 77 as a recent, highly variable, and adapted Race, compared with Race 106. Considering the heterozygous nature of the *P.triticina* genome, analysis was performed to assess the genetic variation between two independent nuclei of the dikaryotic spores as well as analysis also demonstrated that SDs identified within contigs of the two genomes in this study had a minimal possibility of any haplotype sequences assembled within The results further indicate that SDs constitute 18.60% of the Race 77 genome, a higher proportion than the 14.30% in the Race 106 genome. Moreover, a higher percentage (5.52%) of recent segmental duplication (SDs with ≥95% identity) was identified in the Race 77 genome, compared with 3.86% in the Race 106 genome. Comparative analysis also revealed that both the genomes show alignments at different percentages of identity ranging from 90 to 99% of the SD regions.

Analysis of the genome wide *de novo* assembly revealed that the recent duplication content of Race 77 is greater than that of Race 106. SDs have long been known to be involved in genome organization and evolution (Muller 1936; Ohno et al. 1968) and are believed to contribute to genomic instability (Koszul et al. 2006). Therefore, the relatively high number of SDs in Race 77 is likely to be a major factor in the evolution of large numbers of pathotypes of this Race. The identification of a higher number of true ortholog genes, compared with out-paralog genes, in both races is a further indication of the dominance of overall gene duplication events, over gene speciation events, in *P. triticina*.

The overall content of repeat elements in races 77 and 106 was similar; however, their distribution pattern in various genome regions exhibited some differences. Therefore, although a direct correlation was observed between evolutionary time and the increase in the overall percentage of genome repeats, there was no clear relationship observed between the elements and the respective genome sizes. Further, Klevytska et al. (2001), reported that the mechanisms responsible for tandem repeated sequences and their associated variations are potent and have tremendous potential for generating genetic diversity within protein-coding genes over very short evolutionary timescales. Also, different repeat motifs show tremendous variation in their densities in different genomic regions; however, the density of each class of repeat may be comparable across various genomic regions (Subramanian et al. 2003). Consistent with this, the remarkably higher percentage of pentameric repeats in the coding regions of Race 77 may be responsible for its variable nature, relative to Race 106. In this study, the relative abundance of

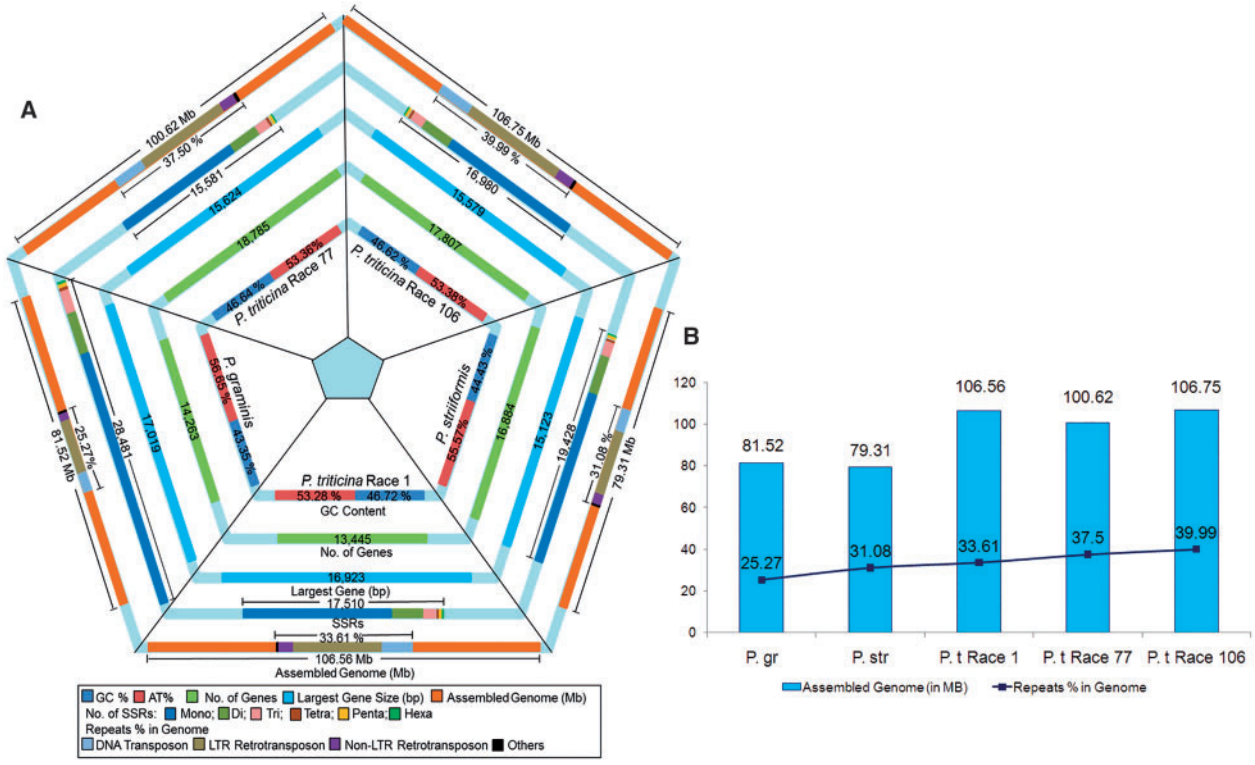

Fig. 9.—Comparative analyses of five genomes of *Puccinia* species. (*A*) Different values are plotted on five different rings as started from outer ring. (1) Assembled genome and repeats % in genome; (2) SSRs predicted and % of their types; (3) Largest gene size; (4) Number of genes identified; and (5) AT and GC % in the genome. Values on each ring are comparative and up to scale to each other's, but independent of other rings. (*C*) Correlation between genome size and percentage of transposable elements. A direct correlation was observed between percentages of transposable elements (TE%) of species at different progressive stages of the evolution, irrespective of their genome sizes.

SSRs showed comparatively similar distribution patterns in both races. Furthermore, TEs are likely candidates for participation in macro- and micro-rearrangements by chromosome breakage (Bennetzen 2005). To date, no linkage map is available for *P. triticina*, hence the effects of these elements on the fate of chromosome based rearrangements could not be commented The presence of TEs within genes in segmental duplicated regions and genes involved in the pathogenicity of Race 77 may have led them to have significant roles in shaping the additionally diverse and adaptive nature of Race 77.

SNP and InDel analysis supported that races 77 and 106, which are from different regions of India, possess a large amount of genetic variation at the genome level. One possible explanation for the higher percentages of non-synonymous and synonymous SNPs in the coding regions of Race 77 over Race 106 could be due to the presence of specific genes which are under strong selection pressure; for example, the polyproteins, which are essential for protein biogenesis and proteolytic processing. Our study also showed that SNPs tend to cluster to a greater extent in non-coding regulatory regions (upstream, and downstream) of Race 106 than in those of Race 77, raising the possibility that mutations in these regulatory regions have no deleterious effects on the functionality of the genes in Race 106.

Secretome analysis demonstrated that Race 77 has higher number of virulence factors than Race 106, which may equip the pathogen to be better adapted to various external environmental factors encountered while multiplying within its host. Some secreted protein families, such as cutinase, pectin esterase, endo1-4 β-D glucanase, and mannanase, showed gene expansion in *P. striiformis* and *P. graminis* (Zheng et al. 2013); however, this phenomenon was not observed in the genomes of *P. triticina* races 77 and 106.

Unraveling the complex molecular networks that regulate and control fungal pathogenicity is a substantial challenge, particularly in poorly annotated genomes, where the functional categorization of genes and gene families is limited. Genes with lost or reduced capacity to cause disease are vital pathogenic genes (Elliott and Howlett 2004). A detailed study on how individual genes or multigene families can have different roles and capacities to cause disease was reported by Tudzynski and Sharon (2003). Consistent with this, we identified genes with reduced virulence as a major category in

both races, with a 2% increase of these genes in Race 77. Biological validation is required to provide further support for the hypothesis that these genes are highly active in regulatory processes during the early stages of infection; however, several mutational studies of genes, which resulted in reduced virulence in their respective hosts in the early stages of infection, have been reported (Gale et al. 1998; Wang et al. 2001; Viaud et al. 2002). In addition, there were differences between the two races in respect of specific genes classified into different functional categories. Genes related to energy metabolism, transport and binding proteins were prominent in Race 106, whereas TEs and conserved domains related genes were more in Race 77. This is in contrast to other fungal rust pathogens, where genes in the functional category "DNA replication and repair" are more in number (Duplessis et al. 2011; Zheng et al. 2013). These differences clearly suggest divergence in evolutionary adaptation, not only between *P. triticina* and other rust pathogens, but also within the *P. triticina* species (i.e., races 77 and 106).

Phylogenetic analysis of the 13 pathotypes of Race 77, based on their whole genome sequences, demonstrated that the relationships among these strains are associated with their virulence features on differential wheat cultivars. The 13 pathotypes share many (~18) *Lr* virulence genes and have differentiated among themselves by overcoming only few *Lr* genes. For instances, pathotypes 77A, 77 A1 have similar virulence profiles and are virulent on 27 *Lr* genes and avirulent on 16 *Lr* genes, and differ to each other only with 77A-1 being avirulent to *Lr* 20. Also, pathotypes 77-6 and 77-7 are quite similar to each other and are virulent on 32 common *Lr* genes and avirulent 11 *Lr* genes and being different to each other only with genes *Lr* 20 and *Lr* 9 to which pathotype 77-7 is virulent (Bharawaj 2012; supplementary table S14, Supplementary Material online). Similarly, pathotypes 77-9, 77-10, and 77-11 have similar virulence profiles among themselves for being avirulent to *Lr* genes 2a, 2b, and 2c unlike the rest eight pathotypes which are virulent to these three *Lr* genes. Overall, the difference in the distributional pattern of SNPs among these pathotypes indicates a higher rate of polymorphism, with some specific mutational events undergone in certain pathotypes, namely 77-9, 77-10, 77-11, and 77-3. The most virulent and recent pathotypes, 77-9, 77-10, and 77-11 were clustered into same clade, and are the most closely related among the pathotypes (Shiwani and Saini 1993). This finding provides support for the idea that recent pathotypes are active and virulent to a greater extent than older pathotypes, such as 77-A, 77-A1, 77-1, and 77-2 which have variable virulence abilities and were active during the early 1990s (Bhardwaj et al. 2015). Our study clearly demonstrated the emergence of these pathotypes from Race 77, indicating rapid, ongoing adaptation within this Race, resulting in the emergence of novel pathotypes.

Conclusions

Genome analysis of the *P. triticina* draft sequences has dramatically improved our knowledge of the genomic variation and gene content in this species. Sequence data from multiple genomes have provided the largest sequence information resource to date for obligate pathogens, and facilitate comparative analysis within and between races. The in-depth genomic information generated in this study, along with other available genome sequences, serve as a starting point for a set capable of distinguishing the >4,000 existing *Puccinia* species. Considering the heterozygous nature of the *P. triticina* genome, this is a maiden report on decoding the genomes of one of the most important wheat pathogen, prevalent in the Indian subcontinent and many other countries worldwide. It will contribute to the wheat improvement program of the Indian subcontinent.

Materials and Methods

Source of Material

Spore mass samples of *P. triticina* races 106 and 77, along with 13 additional pathotypes of Race 77 (77-1, 77-2, 77-3, 77-4, 77-5, 77-6, 77-7, 77-8, 77-9, 77-10,77-11, 77A, and 77A-1), were obtained from the national collection, which is maintained at the Regional Station, Indian Institute of Wheat and Barley Research, Flowerdale, Shimla, India.

Genomic DNA Isolation

Genomic DNA was isolated from urediniospores of *P. triticina* Race 106 and Race 77 and its 13 pathotypes as described by Roose-Amsaleg et al. (2002) with minor modifications. Dried urediniospores (30 μg) were ground to fine powder in liquid nitrogen using pestle and mortar. Then, 550 μl of extraction buffer (100mM Tris-HCl, pH 8.0; 20 mM EDTA, pH 8.0; 1.4 mM NaCl; 2% cetyltrimethylammonium bromide) was added to the powdered mass of spores and transferred to a 1.5 ml microcentrifuge tube. Protinase K (Fermentas, USA) was added to a final concentration of 0.2 mg/ml and the tube incubated for 2 h at 65 °C. Denatured proteins were removed by extracting once with 600 μl Tris saturated phenol:chloroform:isoamylalcohol (25:24:1, v/v/v), followed by repeated extractions with 600 μl Tris saturated chloroform:isoamylalcohol (24:1, v/v). After centrifugation, the aqueous phase was removed and DNA precipitated with 1/10th volume of sodium acetate (3M, pH5.3) and two volumes of absolute alcohol. DNA was pelleted, dried, and resuspended in 40 μl of Tris–EDTA buffer (10 mMTris-HCl, pH 7.5; 1 mM EDTA). An aliquot of the extracted DNA was separated by electrophoresis on 1% agarose gels for visualization and quantification. Genomic DNA (approximately 65μg) was used for sequencing experiments.

Generation of High Quality Genome Sequences

The Race 77 genome was sequenced using a combination of Sanger sequencing and next generation pyrosequencing methods. Whole genome *de-novo* sequencing was performed using the 454 GS FLX platform (Roche, Applied Science, IN, USA). In order to cover the whole genome, Single runs of libraries with different sized inserts (8 and 3 kb) were generated for paired end sequencing, along with five runs of whole genome shotgun (WGS) sequencing libraries. In addition, a single fosmid library (nine 96-well plates) was sequenced using the Sanger sequencing method. DNA Baser v3.5.4 software tool was used to incorporate the reads for assembly using Newbler assembler (ver 2.5.3). Race 106 whole genome *de novo* sequencing was performed by single runs of paired-end sequencing of 3 and 8kb libraries, along with five and a half runs of WGS libraries.

For the preparation of the 8kb paired end library, 20 μg of sample genomic DNA (in Tris-HCl to a final volume of 150 μl) was sheared using HydroShear apparatus (DIGILAB, HydroShear DNA Shearing Device, Holliston, MA), following the manufacturer's instructions. DNA was subjected to 20 cycles at a calibrated speed setting of 16, with standard assembly. The next steps were specific to 454GS-FLXTitanium technology and were carried out strictly according to the manufacturer's instructions (Roche Applied Science). This entailed fragment end polishing, circularization adaptor ligation (A and B), library span size selection on agarose gel, fill-in reaction and DNA circularization, followed by nebulization by applying 45 psi (3.1 bar) of nitrogen for 2 min. WGS libraries do not require circularization therefore for each run, 500ng of sample DNA was directly subjected to nebulization by applying 30 psi (2.1 bar) of nitrogen for 1 min. Further steps for both types of libraries (paired end and WGS) were similar, and comprised fragment end repair and final adaptor ligation, library immobilization, amplification, and final library size selection using AMPure XP beads (Agencourt, Beckman Coulter). The quality of the DNA library (1μl) was checked by size range analysis using a High Sensitivity DNA chip on the 2100 Bioanalyzer (Agilent Technologies) and quantified by fluorescent measurement using the Quant-iT PicoGreenDNA assay (Invitrogen).

An appropriate working dilution of the library was made for use in small-scale emulsion PCR (emPCR) titrations and to determine the exact amount of library required for the final emulsion PCR. The DNA library was then immobilized onto DNA capture beads carrying oligonucleotides complementary to the adaptors. The resulting DNA-beads were added to a mixture of amplification mix and oil and shaken vigorously on a Tissue Lyser (Qiagen) to create "micro-reactors", containing both amplification mix and a single bead. The emulsion was then dispensed into a 96-well plate and the PCR amplification program was run according to the manufacturer's recommendations. Following amplification, the emulsion was chemically broken and the beads carrying the amplified DNA library was recovered and washed by filtration. Positive beads were purified using streptavidin-coated magnetic beads. The DNA library beads were then separated from the magnetic beads by melting the double-stranded amplification products, leaving a population of bead-bound single-stranded template DNA fragments. The sequencing primer was then annealed to the amplified single-stranded DNA. Finally, beads carrying amplified single-stranded DNA were counted using a Hemocytometer (Beckman Coulter) under a light microscope. DNA libraries were sequenced after the addition of DNA polymerase, sulfurylase, and luciferase using the 454 GS-FLX Titanium System (Roche Applied Science).

For preparation of the 40 kb fosmid library for Race 77, a Copy Control HTP fosmid library production kit was used according to the manufacturer's protocol (Epicenter, USA). High quality genomic DNA was isolated from ungerminated urediniospores. Genomic DNA was sheared, end repaired, and 40kb size fragments were separated by electrophoresis. Blunt-ended DNA was purified from LMP agarose and ligated to pCC2FOS vector. Ligated products were packaged into the empty lambda phage particles and used to infect EPI300-T1 *Escherichia coli* cells. Selection of transformed *E. coli* cells was carried out on LB agar-chloramphenicol plates. Individual clones from the plates were picked and grown in 96-well plates containing LB broth, chloramphenicol and Copy Control fosmid auto-induction solution. All cultures were grown overnight at 37 °C with shaking at 250 rpm. Plasmid clones containing inserts were selected and end-sequencing was performed using ABI 3730xl DNA Analyser (Life Technologies).

Sequence Assembly

De novo assembly of the signal processed data was performed using Roche GS Assembler (Newbler 2.5.3) with default parameters including, minimum read length = 20 bp, minimum overlap length = 40 bp, minimum overlap identity = 90%, alignment identity score = 2, and all contig threshold = 100. An incremental approach was followed for *de novo* assembly of WGS data, followed by incorporation of paired end data from the 3 and 8 kb libraries and fosmid (40kb) sequence data. We also compared the assembly obtained using Newbler2.5.3 with those assembled using SeqManNGen 4.0.1 (inDNASTAR) and CLC Genomics Workbench 6 for assembly verification with an estimated genome size of 110 Mb and a minimum contig length of 100 bp for comparative studies. Various aspects of the assembly, including the number of contigs in the assembled genome, useful contig sequences (≥ 25 kb), largest contig size, and (more importantly) N50 and NG50 values, were compared using Assemblathon2 scripts (supplementary fig. S1 and table S1, Supplementary Material online). The resulting assembled sequences were

validated using the Core Eukaryotic Genes Mapping Approach (CEGMA) as described in (supplementary table S3, Supplementary Material online).

The raw reads from races 77 and 106 were then mapped to their respective contigs using CLC Genomics Workbench 7.5.1.Intra-race SNPs were identified in the mapped sequences using the "Basic Variant Detection" module of CLC with default parameters (Ploidy=2; minimum coverage, 10x;Variant Frequency $\geq$ 35%). For inter-racial analysis, the reads of Race 77 were mapped against the assembled contigs from Race 106 and vice-versa followed by "Basic Variant Detection". These SNPs were classified as either "homozygous" or "heterozygous", when one, or more than one, variant was called at a position, respectively. If homozygous SNPs were found to be polymorphic between the two races, we classified them as homokaryotic SNPs, while the heterozygous SNPs were classified as heterokaryotic SNPs if they referred to a variant position that was homozygous in the other Race (Cantu et al. 2013).

Resequencing, Assembly, and Phylogenetic Tree Construction for 13 Pathotypes of Race 77

Paired end libraries with a 100bp insert size were separately prepared from genomic DNA of all 13 pathotypes of Race 77 and sequenced on the HiSeq 1000 (Illumina) platform by NXGenBio life sciences. High-quality reads were compared with the reference sequence of Race 77 using ABySS software to check the authenticity of data. Further, the raw reads were finally assembled to produce contigs using GS Reference Mapper 2.5.3. Whole genome sequence alignment of Race 77 and its pathotypes, including Race1 and Race 106, was performed using Mauve2.3.1 software to obtain a conservation distance matrix and a guide tree was constructed to depict the evolutionary relationships among these variants.

Gene Prediction and Annotations

Genes were predicted from large contigs ($\geq$2 kb) using FGENESH 3.1.2 (MolQuest 2.2) against *Puccinia* sp. with default parameters. In-built PERL scripts were used to parse the FGENESH output and sequences. Predicted genes were BLAST searched against the NCBI EST database for expression analysis. Genes (>450 bases) were BLAST searched against the NCBI non-redundant protein (nr) database for functional annotations. Genes with significant hits (E-value $\leq e^{-10}$) were assigned to specific functional categories. Annotated genes of races 77 and 106 were BLAST searched against each other to predict orthologs (genes with the same function in both races), paralogs (similar genes with different functions), and species-specific genes (with no significant hits in the other genome).

Identification of Putative Repeat Elements within *P. triticina* Genomes

Repeat elements belonging to various classes, including LTRs, non-LTRs, and DNA transposon elements were identified using the CENSOR software tool (Genetic Information Research Institute; http://www.girinst.org/, last accessed 20 August 2016). Significant hits in the BLAST output from comparison of the whole genome sequences against the repeat database were considered repeat elements. Nucleotide sequences > 200 bp were extracted from total TEs for two major repeat groups, *gypsy* and *copia*. Annotation of these elements was performed by BLAST search of the FASTA files against the publicly available fungi repeat database, REPBASE (http://www.girinst.org/repbase/update/). The output file was used to create a unified repeat library that could be compared to previously characterized elements. Annotated repeats of *P. triticina* Race1 published in REPBASE reports were then used for comparative analysis. The unified library lists of *gypsy* and *copia* elements sequences were converted into amino acid sequences in all six frames. Multiple sequence alignments were then performed using ClustalX (http://www.clustal.org/clustal2/, last accessed 20 August 2016) to identify full-length *gypsy* and *copia* elements. Tandem repeat sequences were detected using Tandem Repeats Finder 4 software with default parameters (Benson 1999), (https://tandem.bu.edu/trf/trf.html, last accessed 20 August 2016). Tandem repeats were classified, based on their monomer length, as either minisatellites (16–100 bp) or megasatellites (100–2000 bp). SSR identification was performed using MISA software and categorized using standard parameters (http://pgrc.ipk-gatersleben.de/misa/, last accessed 20 August 2016) for all 15 genomes.

Identification of Novel Genes from Transcriptome Data of Race 106 and Pathotype 77-5

Total RNA was extracted from germinated (G,16h) and ungerminated/resting (UG) spores of Race 106 and pathotype 77-5. The quality and quantity of mRNA was tested using the Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). RNA samples were processed using the RNA-Seq Sample Prep kit from Illumina (Illumina, Inc., CA, USA). Each library was loaded on one lane of an Illumina flowcell, clusters generated were sequenced on a HiSeq 2000 (Illumina, Inc., CA, USA); 100bp paired-end reads were obtained. Reads were initially preprocessed to remove possible contaminants and successively aligned to the Race 77 reference genome using GS Reference Mapper 2.5.3 with default parameters.

A total of 2,651 scaffolds of pathotype 77-5 and 7,448 scaffolds of Race 106 were used as input to search for putative novel genes. Identification of putative novel genes was performed using FGENESH with the complete

gene option enabled. A total of 26,263 and 27,292 candidate genes were found in pathotype 77-5 and Race 106, respectively. Using MEGABLAST, we mapped the predicted mRNAs to the public EST *Puccinia triticina* sp. database downloaded from NCBI using minimal-score = 100 and minimal-identity =90% as filters. This analysis identified 179 genes in pathotype 77-5 and 173 genes in Race 106 with supporting evidence among *P. triticina* ESTs. Next, we eliminated all genes with MEGABLAST matches, using the same parameters, with the transcripts from77-5 and 106 at two stages of spore germination (G16 h and UG). This process yielded 114 and 124 novel genes from ungerminated spores of both races and 90 and 103 novel genes from germinated spores. Functional annotation of the novel genes was performed using the NCBI nr database (significance threshold, e^{-10}).

SNP Analysis

Single nucleotide polymorphisms were detected using the Sequence Alignment/Map tools (SAMtools) software package at 7x coverage with a quality threshold value of Phred score ≥20. The sam file generated by BWA was converted to a bam file and processed using the mpileup utility of SAMtools to generate a pileup of read bases aligned to the sequence of Race 1 (The Broad Institute), which was used as reference sequence for the prediction of SNPs. Annotation of SNPs was performed using SnpEff software (Cingolani et al. 2012)

Identification of Whole Genome Segmental Duplications

The strategy used to detect segmental duplications (SDs) was based on the WGAC (whole genome assembly comparison) method. Self-BLAST searches were performed for all assembled contig sequences. Sequences with ≥90% identity over a ≥1000 bp alignment length were considered possible SDs. Self-hits, duplicate entries, and partial and reverse BLAST hits were removed to obtain the final list and the amount of SDs in the genome. Sequences of SDs were extracted from whole assemblies using PERL scripts and then subjected to FGENESH and MapRep in MolQuest 2.2 Software, for the prediction of genes and transposable elements, respectively. Predicted genes were self-BLAST searched and parsed to remove duplicates, and partial and redundant genes. Genes were BLAST searched against the NCBI nr database for functional annotation.

Identification of Secretory Proteins

A combination of software programs was used to define the secretome of *P. triticina*. Initially all proteins with a SignalP D-score = Y (SignalP v4.1; http://www.cbs.dtu.dk/services/SignalP, last accessed 20 August 2016) and a TargetP Loc = S (TargetP v1.1; www.cbs.dtu.dk/services/TargetP, last accessed 20 August 2016) were combined. These were then scanned for transmembrane spanning regions using TMHMM (TMHMM v2.0; http://www.cbs.dtu.dk/services/TMHMM, last accessed 20 August 2016). All peptides lacking transmembrane helices, along with peptides having a single transmembrane helix located in the first nine amino acids in the mature peptide, were analyzed further. Proteins with highly probable GPI-anchor sequences were predicted by predGPI (http://gpcr.biocomp.unibo.it/predgpi/pred.htm, last accessed 20 August 2016). The eventual locations of these proteins were predicted by integral prediction of protein location scores obtained using ProtComp v9.0 (http://linux1.softberry.com/berry.phtml/berry.phtml?topic=protcompan&group=programs&subgroup=proloc, last accessed 20 August 2016). Proteins demonstrating the integral prediction of protein location, "extracellular secreted", were retained in the final secretome dataset. WolfPSort analysis was performed using "runWolfPsortSummary fungi" to find peptides with a high probability of secretion using the WoLFPSORT v0.2 package (www.wolfpsort.org/WoLFPSORT_package/version0.2, last accessed 20 August 2016). BlastP was used for annotation of the predicted secretome. Conserved domains in the secretome were predicted using the NCBI conserved domain database (http://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi, last accessed 20 August 2016).

Comparative Analysis of Pathogenicity Genes Identified in *P. triticina* Races

Predicted genes from each genome were BLAST searched against 2647 protein sequences in PHI-base (Pathogen–host Interactions database, Version 3.6). Genes with significant hits ($\leq e^{-20}$ and bit score ≥100) against PHI were considered pathogenicity related genes. Annotation of pathogenicity genes was performed by one-to-one alignment against the whole genome annotated file of the predicted genes in both races.

Comparative Analysis of Races 77 and 106 with Publicly Available *Puccinia* Strains

Publicly available genome sequences of the Canadian *P. triticina* race, Race1, *P. graminis*, and *P. striiformis* were downloaded from public databases (Puccinia Group Sequencing Project, Broad Institute of Harvard and MIT (http://www.broadinstitute.org, last accessed 20 August 2016) and then compared based on the different features, including genome size, %AT-GC, % repeats, number of genes, and SSRs. Genome alignment of *P. triticina* races 77, 106, and 1 was performed using Mauve2.3.1 software to obtain a conservation distance matrix and a guide tree to depict the evolutionary relationships among these three races.

Additional Information

Accession codes: Assemblies were deposited at NCBI GenBank and their BioProject IDs and WGS accession numbers are listed in Table below:

Puccinia triticina Race/Pathotype name	BioProject ID	Accession numbers
Race77	PRJNA231547	AZRO00000000
Race106	PRJNA231548	AZRP00000000
Pathotype 77-1	PRJNA277090	LACV00000000
Pathotype 77-2	PRJNA277090	LACW00000000
Pathotype 77-3	PRJNA277090	LACX00000000
Pathotype 77-4	PRJNA277090	LACY00000000
Pathotype 77-5	PRJNA277090	LACZ00000000
Pathotype 77-6	PRJNA277090	LADA00000000
Pathotype 77-7	PRJNA277090	LADB00000000
Pathotype 77-8	PRJNA277090	LADC00000000
Pathotype 77-9	PRJNA277090	LADD00000000
Pathotype 77-10	PRJNA277090	LADE00000000
Pathotype 77-11	PRJNA277090	LADF00000000
Pathotype 77-A	PRJNA277090	LADG00000000
Pathotype 77-A-1	PRJNA277090	LADH00000000

Supplementary Material

Supplementary figures S1–S8, tables S1–S17, and notes 1–2 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was supported by Department of Biotechnology, Govt. of India (BT/PR-13631/AGR/02/717). Authors thank Indian Council of Agricultural Research, New Delhi for providing institutional infrastructure facilities to conduct experiments at National Research Centre on Plant Biotechnology (NRCPB), New Delhi. We sincerely thank "Puccinia Group Sequencing Project, Broad Institute of Harvard and MIT (http://www.broadinstitute.org)" for availability of *Puccinia triticina* Race1 genome data in public domain to be used for comparative studies. We thank Dr S. Nagarajan former Director, Indian Agricultural Research Institute, New Delhi for his valuable technical inputs to start this study. We thank Ashwani Mishra, NxGenBio Life Sciences, Delhi for support in preliminary data analysis. T.R.S. is thankful to the Department of Science and Technology, Govt. of India for JC Bose National Fellowship. K.K. generated sequence data, performed analyses of various experiments. H.R. worked on genome assembly and analyses. H.D. performed secretory protein analysis. S.C.B. provided biological material for sequencing. S.C.B., H.R. and D.B.N. helped in the writing of manuscript. A.U.S., D.K.G., P.P., D.P., P.C., P.B., J.K., M.S., K.G., and N.K.S. provided input during execution of the project. T.R.S. conceived and directed the project and planning of experiments. T.R.S. and K.K. wrote the manuscript. All authors read and approved the final manuscript.

Literature Cited

Ayliffe MA, Lagudah ES. 2004. Molecular genetics of disease resistance in cereals. Ann. Bot. 94:765–773.

Babayants O, et al. 2015. Physiologic specialization of *Puccinia triticina* Erikss. and effectiveness of *Lr*-genes in the south of Ukraine during 2013-2014. Chilean J. Agri. Res. 75:443–450.

Bajwa MA, Aqil KA, Khan NI. 1986. Effect of leaf rust on yield and kernel weight of spring wheat. Rachis 5:25–28.

Barr R, Caldwell RM, Amacher RH. 1964. An examination of vegetative recombination of urediniospore color and virulence in mixtures of certain races of *Puccinia recondita*. Phytopathology 54:104–109.

Bennetzen JL. 2005. Transposable elements, gene creation and genome rearrangement in flowering plants. Curr. Opin. Genet. Dev. 15:621–627.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27:573–580.

Bhardwaj SC, Gangwar OP, Pramod P, Khan H. 2015. ICAR. Indian Inst. of Wheat and Barley Research, Regional Station, Flowerdale, Shimla, India. Mehtaensis News Lett. 35:1–16.

Bhardwaj SC, Prashar M, Subodh K, Datta D. 2006. Virulence and diversity of *Puccinia triticina* on wheat in India during 2002–04. Indian J. Agric. Sci. 76:324–426.

Bhardwaj SC. 2013. *Puccinia–Triticum* interaction: an update. J Indian Phytopath. 66:14–19.

Bhardwaj SC. 2011. Strategic centres. 100 years of wheat research in India—A saga of distinguished achievements. Directorate of Wheat Research (ICAR) India 243–262.

Bhardwaj SC. 2012. Wheat rust pathotypes in Indian subcontinent then and now. In: Singh SS, Hanchinal RR, Singh G, Sharma RK, Tyagi BS, Sharan MS, Sharma I, editors. Wheat productivity enhancement under changing climate. New Delhi: Narosa Publishing House Pvt Ltd, p. 227–238.

Cantu D, et al. 2011. Next generation sequencing provides rapid access to the genome of *Puccinia striiformis* f. sp. *tritici*, the causal agent of wheat stripe rust. PLoS One 6:1–8.

Cantu D, et al. 2013. Genome analyses of the wheat yellow (stripe) rust pathogen *Puccinia striiformis* f. sp. *tritici* reveal polymorphic and haustorial expressed secreted proteins as candidate effectors. BMC Genomics 14:270.

Cingolani P, et al. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff, SNPs in the genome of *Drosophila melanogaster* strain w^{1118}; iso-2; iso-3. Fly (Austin) 6:80–92.

Cuomo CA, et al. 2016. Comparative analysis highlights variable genome content of wheat rusts and divergence of the mating loci. bioRxiv, Cold Spring Harbor Laboratory, doi:http://dx.doi.org/10.1101/060665.

de Wit PJ, Mehrabi R, Van den Burg HA, et al. 2009. Fungal effector proteins: past, present and future. Mol. Plant Pathol. 10:735–747.

Draz IS, Abou-Elseoud MS, Kamara A-EM, Alaa-Eldein OA-E, El-Bebany AF. 2015. Screening of wheat genotypes for leaf rust resistance along with grain yield. Ann. Agric. Sci. 60:29–39.

Duplessis S, et al. 2011. Obligate biotrophy features unraveled by the genomic analysis of rust fungi. Proc. Natl. Acad. Sci. U S A. 108:9166–9171.

Elliott CE, Howlett BJ. 2004. Approaches for identification of fungal genes essential for plant disease. In: Setlow JK, editor. Genetic engineering. New York: Springer, pp. 85–103.
Gale CA, et al. 1998. Linkage of adhesion, filamentous growth, and virulence in *Candida albicans* to a single gene, *INT1*. Science 279:1355–1358.
German S, et al. 2007. The situation of common wheat rusts in the Southern Cone of America and perspectives for control. Aust. J. Agric. Res. 58:620–630.
Gonzalez-Fernandez R, Jorrin-Novo JV. 2012. Contribution of proteomics to the study of plant pathogenic fungi. J. Proteome Res. 11:3–16.
Herrera-Foessel SA, et al. 2011. New slow-rusting leaf rust resistance genes Lr67 and Yr46 in wheat are pleiotropic or closely linked. Theor. Appl. Genet. 122:239–249.
Huerta-Espino J, et al. 2011. Global status of wheat leaf rust caused by *Puccinia triticina*. Euphytica 179:143–160.
Idnurm A, Howlett BJ. 2001. Pathogenicity genes of phytopathogenic fungi. Mol. Plant Pathol. 2:241–255.
Jain SK, Nayar SK, Prashar M, Bhardwaj SC, Subodh K. 2004. Pathotypes of *Puccinia recondita tritici* in India during 1999-2001. Indian J. Agric. Sci. 74:185–189.
Johnston CO, Mains EB. 1932. Studies on physiological specialization in *Puccinia triticina* U.S. Dep. Agric. Tech. Bull. 313:22.
Joshi LM, Singh DV, Srivastava KD. 1986. Wheat and wheat diseases in India. In: Joshi L M, Singh D V, Srivastava K D. editors. Problem and progress of wheat pathology in South Asia. New Delhi: Malhotra Publishing House.
Klevytska AM, et al. 2001. Identification and characterization of variable-number tandem repeats in the *Yersinia pestis* genome. J. Clin. Microbiol. 39:3179–3185.
Kolmer JA. 1996. Genetics of resistance to wheat leaf rust. Annu. Rev. Phytopathol. 34:435–455.
Kolmer JA. 1997. Virulence in *Puccinia recondita* f. sp. *tritici* isolates from Canada to genes for adult-plant resistance to wheat leaf rust. Plant Dis. 81:267–271.
Kolmer JA. 2005. Tracking wheat rust on a continental scale. Curr. Opin. Plant Biol. 8:441–449.
Koszul R, Dujon B, Fischer G. 2006. Stability of large segmental duplications in the yeast genome. Genetics 172:2211–2222.
Manjunatha C, Aggarwal R, Bhardwaj SC, Sharma S. 2015. Virulence analysis and molecular characterization of *Puccinia triticina* pathotypes causing wheat leaf rust in India. Res. J. Biotech. 10:98–107.
Mishra AN, Kaushal K, Jain SK, Pandey HN. 2001. 'Thatcher'-avirulent leaf rust pathotypes from India. Wheat Int. Serv. 93:32–34.
Mishra AN, Kaushal K, Shirsekar GS, Yadav SR, Brahma RN. 2005. Genetic basis of seedling-resistance to leaf rust iin bread wheat 'Thatcher'. Plant Breed. 124:514–516.
Muller HJ. 1936. Bar duplication. Science 83:528–530.
Nayar SK, Prashar M, Bhardwaj SC, Verma LR. 1996. Distribution pattern of *Puccinia recondite tritici* pathotypes in India during 1990-94. Indian J. Agric. Sci. 66:621–630.
Nelson RR, Wilcoxson RD, Christensen JJ. 1955. Heterokaryosis as a basis for variation in *Puccinia graminis* var. *tritici.* Phytopathology 45:639–643.
Ohno S, Wolf U, Atkin NB. 1968. Evolution from fish to mammals by gene duplication. Hereditas 59:169–187.
Ordonez ME, Kolmer JA. 2009. Differentiation of molecular genotypes and virulence phenotypes of *Puccinia triticina* from common wheat in North America. Phytopathology 99:750–758.
Park RF. 1996. Pathogenic specialisation of *Puccinia recondita* f. sp. *tritici* in Australia and New Zealand in 1990 and 1991. Aust. Plant Pathol. 25:12–17.
Rodenhiser HA, Hurd-Karrer AM. 1947. Evidence of fusion bodies from urediniospore germ tube of cereal rusts on nutrient solution agar. Phytopathology 46:744–756.
Roose-Amsaleg C, Vallavieille-Pope C, Brygoo Y, Levis C. 2002. Characterization of a length polymorphism in the two intergenic spacers of ribosomal RNA in *Puccinia striiformis* f. sp. *tritici*, the causal agent wheat yellow rust. Mycol. Res. 106:918–924.
Savile DBO. 2004. Taxonomy of the cereal rust fungi. In: Bushnell WR, Roelfs AP, editors. The cereal rusts, origins, specificity, structures, and physiology. Orlando, FL: Academic press, pp. 9–112.
Shiwani, Saini RG. 1993. Diversity for resistance to leaf rust in *Triticum aestivum.* Plant Dis. 77:359–363.
Stergiopoulos I, de Wit PJ. 2009. Fungal effector proteins. Annu. Rev. Phytopathol. 47:233–263.
Subramanian S, Mishra RK, Singh L. 2003. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. Genome Biol. 4:R13.
Thind TS. 1998. Diseases of field crops and their management. Ludhiana, India: National Agricultural Technology Information Centre.
Toth IK, Bell KS, Holeva MC, Birch PRJ. 2003. Soft rot erwiniae: from genes to genomes. Mol. Plant Pathol. 4:17–30.
Tudzynski P, Sharon A. 2003. Fungal pathogenicity genes. In: Arora DK, Khachatourians GG, editors. Fungal genomics. Applied mycology and biotechnology. Amsterdam: Elsevier, pp. 187–212.
Urban M, Irvine AG, Cuzick A, Hammond-Kosack KE. 2015. Using the pathogen-host interactions database (PHI-base) to investigate plant pathogen genomes and genes implicated in virulence. Front. Plant Sci. 6:605.
Van de WAP, Howlett BJ. 2011. Fungal pathogenicity genes in the age of 'omics'. Mol. Plant Pathol. 12:507–514.
Viaud MC, Balhadère PV, Talbot NJ. 2002. A *Magnaporthe grisea* cyclophilin acts as a virulence determinant during plant infection. Plant Cell 14:917–930.
Wang P, Cardenas ME, Cox GM, Perfect JR, Heitman J. 2001. Two cyclophilin A homologs with shared and distinct functions important for growth and virulence of *Cryptococcus neoformans.* EMBO Rep. 2:511–518.
Wang X, McCallum B. 2009. Fusion body formation, germ tube anastomosis, and nuclear migration during the germination of urediniospores of the wheat leaf rust fungus, *Puccinia triticina.* Phytopathology 99:1355–1364.
Winnenburg R, Baldwin TK, Urban M, et al. 2006. PHI-base: a new database for pathogen host interactions. Nucleic Acids Res. 34:459–464.
Wright RG, Lennard JH. 1980. Origin of new race of *Puccinia striiformis.* Trans. Br. Mycol. Soc. 74:283–287.
Xu J, et al. 2011. Gene discovery in EST sequences from the wheat leaf rust fungus *Puccinia triticina* sexual spores, asexual spores and haustoria, compared to other rust and corn smut fungi. BMC Genomics 12:161.
Zheng W, et al. 2013. High genome heterozygosity and endemic genetic recombination in the wheat stripe rust fungus. Nat. Commun. 4:2673.

Associate editor: B. Venkatesh