MICROBIAL GENOMICS

RESEARCH ARTICLE

Pereira *et al.*, *Microbial Genomics* 2021;7:000540

DOI 10.1099/mgen.0.000540

MICROBIOLOGY SOCIETY

OPEN DATA OPEN ACCESS

Population genomics of transposable element activation in the highly repressive genome of an agricultural pathogen

Danilo Pereira[1]†, Ursula Oggenfuss[2], Bruce A. McDonald[1] and Daniel Croll[2,*]

Abstract

The activity of transposable elements (TEs) can be an important driver of genetic diversity with TE-mediated mutations having a wide range of fitness consequences. To avoid deleterious effects of TE activity, some fungi have evolved highly sophisticated genomic defences to reduce TE proliferation across the genome. Repeat-induced point mutation (RIP) is a fungal-specific TE defence mechanism efficiently targeting duplicated sequences. The rapid accumulation of RIPs is expected to deactivate TEs over the course of a few generations. The evolutionary dynamics of TEs at the population level in a species with highly repressive genome defences is poorly understood. Here, we analyse 366 whole-genome sequences of *Parastagonospora nodorum*, a fungal pathogen of wheat with efficient RIP. A global population genomics analysis revealed high levels of genetic diversity and signs of frequent sexual recombination. Contrary to expectations for a species with RIP, we identified recent TE activity in multiple populations. The TE composition and copy numbers showed little divergence among global populations regardless of the demographic history. Miniature inverted-repeat transposable elements (MITEs) and terminal repeat retrotransposons in miniature (TRIMs) were largely underlying recent intra-species TE expansions. We inferred RIP footprints in individual TE families and found that recently active, high-copy TEs have possibly evaded genomic defences. We find no evidence that recent positive selection acted on TE-mediated mutations rather that purifying selection maintained new TE insertions at low insertion frequencies in populations. Our findings highlight the complex evolutionary equilibria established by the joint action of TE activity, selection and genomic repression.

DATA SUMMARY

All Illumina sequence data is available from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) under BioProject numbers PRJNA606320, PRJNA398070 and PRJNA476481 (https://www.ncbi.nlm.nih.gov/bioproject). The Methods and Figs S1–S11, Tables S1–S6 (available with the online version of this article) provide all information on strain locations and outcomes of genome analyses.

INTRODUCTION

Genetic diversity in natural populations largely determines the evolutionary potential of populations. Polymorphism and diversity in haplotypes can arise from various processes, including single base mutations [1], reshuffling of alleles through recombination [2], chromosomal rearrangements [3] and the action of selfish DNA sequences [4]. Transposable elements (TEs) are ubiquitous selfish elements capable of proliferating throughout the genomic landscape [5]. Transposition activity can increase with senescence [6] or under environmental stress conditions increasing the risk for genetic modifications [7]. TE-mediated genetic changes can impact the fitness of the host. During transposition, TEs can create genetic variation by altering coding sequences, gene regulation or triggering chromosomal rearrangements through non-homologous recombination [8, 9]. The impact of TEs is in most cases negative or neutral, but can rarely also be positive by contributing to adaptation in humans and other organisms [10, 11]. In plant pathogens, TEs are important drivers of genome evolution and adaptation to the host [12–14]. For example, TE activity caused gene deletions and sequence reshuffling in the fungal

Received 12 November 2020; Accepted 03 February 2021; Published 23 August 2021

Author affiliations: [1]Plant Pathology, Institute of Integrative Biology, ETH Zürich, Zürich, Switzerland; [2]Laboratory of Evolutionary Genetics, Institute of Biology, University of Neuchâtel, Neuchâtel, Switzerland.

***Correspondence:** Daniel Croll, daniel.croll@unine.ch

Keywords: adaptation; genetic diversity; pathogen; repeat-induced point mutations; selfish elements; transposable elements.

Abbreviations: GO, gene ontology; LD, linkage disequilibrium; LTR, long terminal repeat; MITE, miniature inverted-repeat transposable element; ML, maximum likelihood; NCBI, National Center for Biotechnology Information; PCA, principal component analysis; RIP, repeat-induced point mutation; TRIM, terminal repeat retrotransposon in miniature.

†Present address: Max Planck Institute for Evolutionary Biology, August-Thienemann-Straße 2, D-24306 Plön, Germany.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Eleven supplementary figures and six supplementary tables are available with the online version of this article.

pathogen *Blumeria graminis* f. sp. *hordei* ultimately underpinning host specialization [15]. In *Zymoseptoria tritici*, a gain in virulence was observed after TE-mediated deletion of a single gene [16], and fungicide resistance emerged from overexpression and splicing alterations of target genes after upstream TE insertions [17, 18]. Despite the major impact on genome stability and the expression of phenotypic traits, ongoing TE activity at the intra-specific level is poorly understood in plant pathogens and other organisms. Major questions remain regarding the genome-wide distribution of active TEs and their impact on fitness.

A number of fungal genomes, including the genomes of various plant pathogens, contain regions enriched in TEs and are evolving at faster evolutionary rates largely through sequence rearrangements [12, 19–21]. Polymorphism in genes located within such regions is usually higher than in more conserved regions [21, 22]. Furthermore, key virulence factors tend to localize in the vicinity of repetitive regions [13, 23, 24]. Genes showing evidence for presence–absence polymorphism were shown to be closer to TEs than conserved genes [24–26]. Gene deletions in some pathogens are associated with gains in virulence if the affected proteins are recognized by the host [26]. Controlling the activity of TEs faces a costly trade-off between genomic defence mechanisms such as silencing and the bona fide expression of nearby genes [27]. Suppression of TE proliferation can be mediated by heterochromatin modifications, DNA methylation, meiotic silencing and post-transcriptional regulation [28, 29]. A powerful fungal-specific mechanism is to hypermutate duplicated sequences through repeat-induced point mutation (RIP), a mechanism first described in *Neurospora crassa* [30]. During sexual recombination, RIP acts via homology recognition changing C:G nucleotides to T:A nucleotides, leaving sequence hallmarks found in genomes of many plant pathogenic fungi [31–35]. Although genomic defences against TEs are frequent across fungi, substantial variation in TE abundance was found among genomes between species but also within species [36–39]. The population frequency of individual TE sequences is restrained by purifying selection [40]. A loss of control or neutral effects can permit copy number expansions [40, 41]. In rare cases, a TE insertion locus can undergo a selective sweep indicating that the insertion may have created adaptive genetic variation. Considering that genomic defences such as RIP were shown to vary between species [35, 42], the interplay of genomic defences, selection and demographics has the potential to drive or restrain TE proliferation among populations.

In the wheat pathogen *Parastagonospora nodorum*, TEs have caused significant genetic alterations and likely facilitated the transfer of a key virulence gene. *P. nodorum* is a pathogen with high evolutionary potential to adapt to local conditions [24, 43, 44] and a worldwide distribution causing significant wheat-yield losses [45]. The pathogen harbours a repertoire of various effector genes (named *Tox* genes) that confer host adaptation by causing cell death in susceptible plants. Interestingly, the effector gene *ToxA* was involved in a horizontal gene transfer (HGT) in the triad of the pathogens *P. nodorum*, *Pyrenophora tritici-repentis* and *Bipolaris sorokiniana* [46, 47]. The HGT was facilitated by the TE environment in which *ToxA* is embedded. Overall, repetitive elements in the genome of *P. nodorum* comprise less than 6.2% of the total genome size [24, 48]; proximity to TEs was correlated with gene presence–absence polymorphisms within populations [24]. Despite extensive evidence for RIP-mediated TE control in the reference genome [32, 49], the species display remarkable examples of TE-mediated phenotypic trait variation. Hence, TE activity may not be fully constrained by genomic defences and shape the evolutionary trajectory of the species.

Impact Statement

Selfish genetic elements or transposable elements (TEs) are ubiquitous in the genomes of many microbial eukaryotes. In fungal pathogens, TEs have reshaped genomes by copying and reinserting into new locations, potentially causing an expansion of the genome. The most intriguing effects of TEs occur over short evolutionary timescales by creating beneficial new variants in populations leading to drug resistance or increased virulence. Fungal genomes have also powerful defence mechanisms including repeat-induced point mutations (RIPs), which can deactivate TEs over a few generations. Hence, species are under complex selection regimes to modulate the activity of TEs. Yet, species-wide analyses of TE activity and deactivation are largely lacking. Here, we take a comprehensive view of TE dynamics based on 366 whole-genome sequenced strains representing the global diversity of the fungal wheat pathogen *Parastagonospora nodorum*. We are able to show that the species retained active TEs despite supposedly highly efficient genomic defences and the impact of purifying selection. Some recent TE activity may even have contributed to adaptive evolution. Our study establishes a microbial eukaryote model for species-wide investigations into mechanisms governing TE activity in genomes.

In this study, we screened whole-genome sequences of a global collection of 366 *P. nodorum* isolates for evidence of recent TE activity and genetic footprints of recent selection. TEs were exhaustively identified based on sequence similarity searches in three complete genomes of the species. We analysed the TE load among populations in relationship with the demographic history of the pathogen. Finally, we identified TE loci potentially contributing to local adaptation and inferred the effectiveness of genomic defence mechanisms shaping TE variation.

METHODS

Populations characterized using Illumina whole-genome sequencing

We analysed a total of 366 single spore isolates of *P. nodorum* sampled between 1991 and 2016. All isolates were collected from naturally infected wheat fields. The sampling locations included: Australia (2001 and 2010; n=23), Brazil (year unknown; n=1), Canada (2005; n=1), Finland (year unknown; n=1), Iran (2005 and 2010; n=20), South Africa (1995; n=23), Sweden (year unknown; n=2), Switzerland (1999A and

1999B; *n*=46), Arkansas (USA; 1995; *n*=6), Georgia (USA; 2008; *n*=5), Maryland (USA; 2008; *n*=3), Minnesota (USA; 2002, 2003 and 2005; *n*=14), New York (USA; 1991; *n*=21), North Carolina (USA; 2008; *n*=9), North Dakota (USA; 1998, 2003, 2005, 2007, 2008, 2010 and 2016; *n*=82), Ohio (USA; 2003; *n*=16), Oklahoma (USA; 2016; *n*=17), Oregon (USA; 1993 and 2011; *n*=28), South Carolina (USA; 2008; *n*=4), South Dakota (USA; 2016; *n*=8), Tennessee (USA; 2008; *n*=5), Texas (USA; 1992; *n*=27) and Virginia (USA; 2008; *n*=4). Earlier publications referred to the Switzerland 1999B population as being of Chinese origin [50–54]. A more recent study corrected the population origin [44]. Illumina whole-genome sequence data was generated for all 366 isolates with paired-end sequencing and a read length of 100–150 bp. Raw data was accessed from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) under BioProject ID numbers PRJNA606320, PRJNA398070 and PRJNA476481 [24, 37, 44].

Genome alignment, variant calling and quality filtering

Raw reads were trimmed for remaining Illumina adaptors and read quality was assessed using Trimmomatic version 0.36 [55] with the following parameters: illuminaclip=TruSeq3 PE.fa:2:30:10, leading=10, trailing=10, slidingwindow=5:10, minlen=50. Trimmed reads were aligned against the reference genome established for the isolate Sn2000 [56] using the short-read aligner Bowtie2 version 2.3.3 [57] with the –very-sensitive-local option. PCR duplicates were marked using the MarkDuplicates option in Picard tools version 2.17.2 (http://broadinstitute.github.io/picard). All sequence alignment (SAM) files were sorted and converted to binary (BAM) files using SAMtools version 1.2 [58]. SNP calling and variant filtration were performed using the Genome Analysis Toolkit (GATK) version 3.8-0 [59]. We used HaplotypeCaller on each alignment file individually with the --emit-ref-confidence GVCF and -ploidy one options. Joint variant calls were produced using GenotypeGVCFs with the flag -maxAltAlleles 2. Finally, SelectVariants and VariantFiltration were used for hard filtering SNPs failing the following cut-offs: QUAL >200; QD >10.0; MQ >20.0; –2<BaseQRankSum<2; – 2<MQRankSum<2; –2<ReadPosRankSum<2. We kept only bi-allelic SNPs and with a maximum genotype missingness of 10% using vcftools version 0.1.15 [60].

Phylogenomic and population structure analyses

We inferred the evolutionary history of all 366 *P. nodorum* isolates using two phylogenetic tree reconstruction methods. A maximum-likelihood (ML) tree was estimated for all isolates using the GTRCAT model in RAxML version 8.2.12 [61]. We performed 100 rapid bootstraps with 20 ML searches. The best ML tree was chosen based on support values. We analysed evidence for reticulation among *P. nodorum* genotypes by building a Neighbor-Net (NN) network with SplitsTree4 version 4.15.1 [62]. The ML tree and the NN network were visualized using the online tool iTOL version 4 [63]. VCF files were converted to RAxML input (PHYLIP) and SplitsTree4 input formats (Nexus) using PGDSpider version 2.1.1.5 [64].

Population structure was inferred based on a principal component analysis (PCA) in TASSEL version 5.2.56 and a model-based clustering implemented in STRUCTURE version 2.3.4 [65, 66]. We performed the PCA based on SNP data filtered for a minor allele frequency above 5% and based on TE frequency insertion sites filtered for a minor allele frequency above 1%. We visualized the two first principal components using the ggplot2 package in R [67, 68]. For STRUCTURE analyses, we used an admixture model independent of prior population information and with correlated allele frequencies. The algorithm ran with a burn-in length of 50000 and a simulation length of 100 000 Markov chain Monte Carlo (MCMC) repetitions. We explore a range of *K* between 1 and 10, with 10 repetitions per *K*. The most likely number of populations (*K*) was estimated using the Delta K (ΔK) method [69] implemented in the R package *pophelper* version 2.3.0 [70]. For all phylogenetic and STRUCTURE analyses, we randomly selected SNPs at a mean distance of 15 kb using the vcftools parameter --thin 15000 to reduce linkage disequilibrium (LD) among loci and computational demands.

Population genetics and selective sweeps

Allele frequencies and nucleotide diversity (p) per site were determined using the options --freq and --site-pi in vcftools [60]. Histograms for minor allele spectrum were generated in ggplot2. LD decay was estimated per population based on 50 kb windows on chromosome one using r^2 with the option --hap-r2 in vcftools [60]. We performed Tajima's D neutrality tests [71] using the vcftools --haploid flag and analysed non-overlapping bins of 1 kb across the entire genome [60]. Selective sweeps were identified using a likelihood-based detection method implemented in the program SweeD version 3.0 using the option -folded [72]. We ran SweeD 3.0 individually for each of the 23 chromosomes of the reference genome Sn2000 in grids of 1 kb. If LD decay was slow in a population (i.e. r^2 ≥0.2 at 10 kb), we merged together adjacent genomic regions under selection if these were separated by less than 20 kb. We analysed regions of interest by adding 10 kb on each side of the window identified by SweeD.

Analysis of regions with signatures of selection using gene ontology (GO)

The gene annotation for the reference genome Sn2000 was previously established [43]. Briefly, predicted protein sequences of *P. nodorum* Sn15 strain were aligned and queries with a at least 95% identity score were maintained. Gene identifiers were retained. Genomic regions showing signatures of selection were used as coordinates in BEDtools version 2.29.0 [73] to intersect annotated genes in the Sn2000 reference strain annotation [56]. Genes were annotated for protein family (PFAM) domain and GO terms using interproscan version 5.36-75.0 with default parameters and a local pre-calculated match lookup service [74]. Protein secretion signals were predicted using SignalP version 4.01 [75], Phobius [76] and TMHMM version 2.0 [77]. GO enrichment analyses were

performed using the packages GSEABase version 1.35.0 and GOstats version 2.38.1 in R [78]. We used the false discovery rate of 5% as a cut-off and minimum GO term size of 5 for the hypergeometric test.

TE consensus sequence identification and classification

To obtain consensus sequences for TE families, we performed individual runs of RepeatModeler version 1.0.8 on the three complete genomes Sn2000, Sn4 and Sn79-1087. Identified sequences were annotated based on GIRI Repbase using RepeatMasker version 4.0.7 [56, 79, 80]. Further classification and processing of consensus sequences was performed using WICKERsoft [81]. We screened the three complete genomes for copies of the above detected consensus sequences with BLASTN filtering for sequence identity of >80% over >80% of the length of the sequence [82]. We then added flanks of at least 300 bp both upstream and downstream of each sequence. We made multiple sequence alignments using ClustalW and defined TE boundaries by visual inspection before updating consensus sequences [83]. If possible, consensus sequences were classified according to the presence and type of terminal repeats, superfamily-specific start and end bases or target site duplications, as well as homology of encoded proteins using BLASTX in the NCBI nr database. We excluded duplicated consensus sequences using dotter for inspection [84]. We assigned consensus sequence names according to the three-letter naming system [85]. Two predicted TE families showed strong length polymorphism and only weak sequence similarity between all individual insertions. Using BLASTN on individual insertions, we found that the sequences matched one of three different regions of the entire consensus sequences. We visualized the consensus sequence alignment with genoPlotR version 0.8.9 in R [86] and used BLASTN to identify matching regions in the Sn2000, Sn4 and Sn79-1087 complete genomes [56, 87, 88]. To re-define consensus sequences, we proceeded as described above. In a second round of TE annotation, we focused on protein-encoding sequences matching previously identified fungal TE superfamilies. We screened the three complete genomes for matches of protein sequences representative of each superfamily from other fungi using TBLASTN. We filtered hits for a minimal alignment length of 80 bp and a sequence similarity >25%. Identified sequences were retrieved including 300 bp flanking sequences. Hits were analysed for their matching sequencing with dotter and grouped into families based on visual inspection. We made further multiple sequence alignments using ClustalW and defined TE boundaries by visual inspection. Finally, the TE family consensus sequences from the two methods above were used to annotate the reference genomes using RepeatMasker version 4.0.7 [56, 79, 80]. We estimated the guanine–cytosine (G+C) content as G+C/G+C+A+T on the TE families retained in the consensus sequence: the FASTA consensus file was converted to EMBOSS sequence query using EMBOSS seqret [89], and run on EMBOSS geecee (https://www.bioinformatics.nl/cgi-bin/emboss/geecee). TE presence–absence annotation in the global collection of isolates was performed with the method proposed by Linheiro and Bergman [90], with ngs_te_mapper version 1 (https://github.com/bergmanlab/ngs_te_mapper/commit/fb23590200666fe66f1c417c5d5934385cb77ab9) implemented in R. Dependences to ngs_te_mapper pipeline were BWA version 0.7.17-r1188 [91], to map Illumina reads, and SAMtools version 1.2 [58], to perform output file conversion.

Depth of coverage can impact robust TE discovery; therefore, we set a minimum depth of 10×, which allows recovery of ≥90% of identified TEs across populations (Fig. S1). As a consequence, we removed 18 isolates and retained a total of 348 isolates for downstream TE polymorphism analyses. We clustered nearby TE insertions into a single locus if the insertions (i) belonged to the same TE family and (ii) were located within 100 bp distance. Clustering of TE insertion sites was performed using the package GenomicRanges version 1.38.0 implemented in R [92].

RESULTS

Global population structure and phylogenomics

We analysed genome sequencing data for 366 *P. nodorum* isolates and identified a total of 487477 high-confidence SNP markers with a minor allele frequency of 5%. We identified three main groups by performing PCAs (Figs 1a and S2). Isolates from Oklahoma (USA) formed a distinct group flanked by the two largest clusters. Unsupervised Bayesian clustering analyses revealed a similar pattern of high admixture among genotypes from different continents with the optimal number of clusters being K=2 (Figs 1b and S3). At K=3, isolates from Iran and the northern United States (North Dakota, South Dakota and Minnesota) grouped as genetic cluster two, while isolates from the southern United States shared membership with cluster one. Genotypes from Australia and South Africa were assigned mainly to cluster three. Genotypes from Swiss populations and Oklahoma (USA) showed significant admixture. It is important to note that North America is overrepresented in our genomic sampling compared to other regions and the high local genotypic diversity is at least partially explained by the high number of samples. We used genome-wide SNPs to infer a ML tree and identified four well-supported major clades largely independent of sampling origin (Fig. 1c). We found evidence of reticulation indicative of recombination based on a SplitsTree network (Fig. 1d). The network also showed that multiple populations harboured groups of highly similar genotypes indicative of recent ancestry. Both the ML tree and the SplitsTree network identified four major genetic groups including a group of isolates from Australia and South Africa, a group of isolates from Brazil, Finland, Sweden, Switzerland and the southern United States, as well as a group of isolates from Canada, Iran and the northern United States. The fourth group is composed of isolates solely from Oklahoma (USA). The emergence of the genetic groups could stem from historic gene flow and may be shaped by environmental similarities. Our results indicate significant levels of gene flow among populations from the same continent, as well as among a set

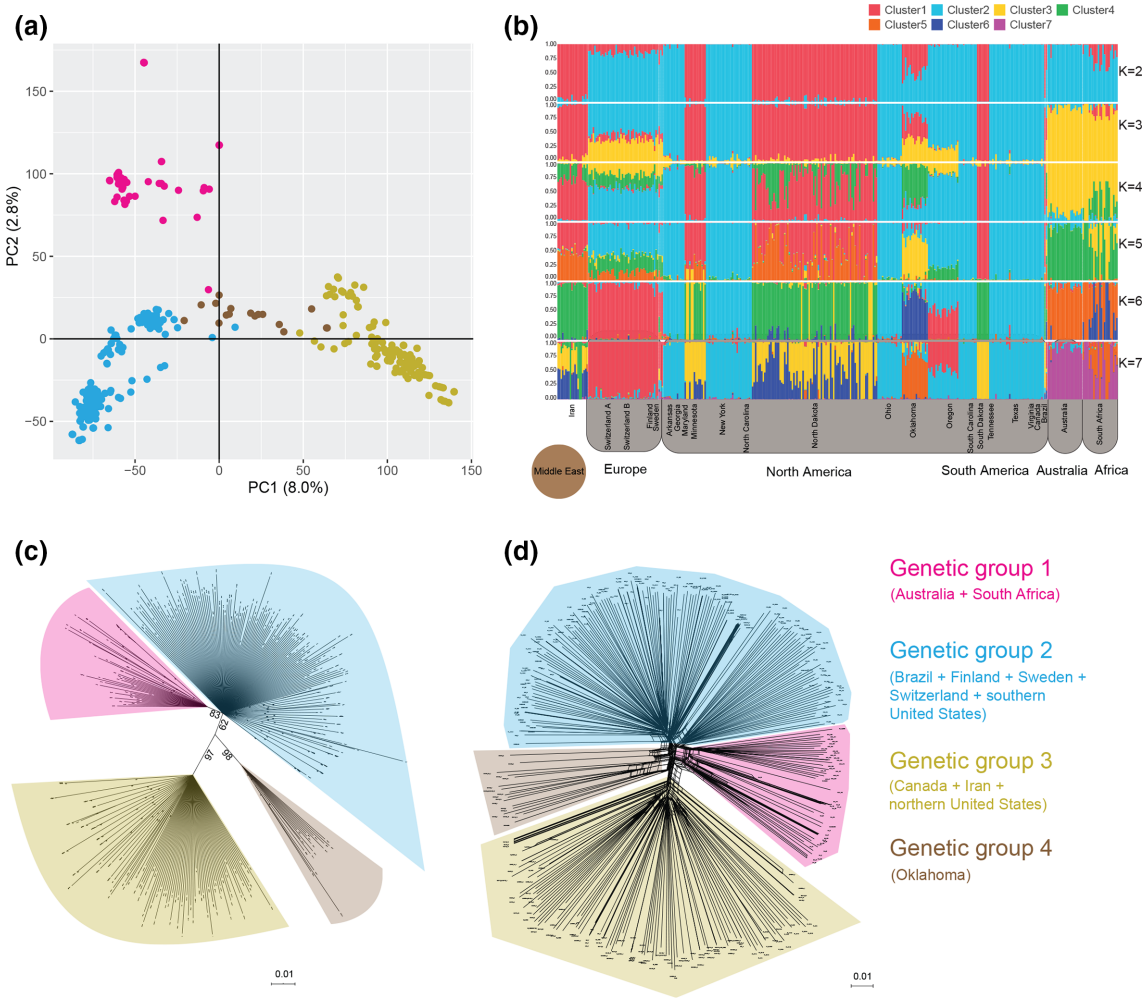

Fig. 1. Global population structure and phylogenomics of *P. nodorum*. (a) PCA with dots representing individual isolates coloured according to the genetic group. PC, Principal component. (b) Estimated population genetic structure. Each vertical bar represents one individual, coloured according to cluster membership values. The cluster designation is represented by *K* numbers. Iran (Middle East) is highlighted as matching the region of origin of *P. nodorum*. (c) ML phylogenetic tree. (d) Phylogenetic network reconstruction. Scale bars represents the mean number of nucleotide substitutions per site and the split support for the edges, respectively. Population genetic structure and phylogenetic reconstructions were based on 2278 genome-wide SNP markers spaced evenly at 15 kb. The PCA was based on the complete SNP dataset.

of populations from different continents. Populations constituting genetic group two were mostly sampled in regions classified as temperate climates (Cfa and Cfb; Table S1), while in genetic group three, samples from continental climates are predominant (Dfa and Dfb; Table S1). Although climate characteristics are not identical among sampling locations (e.g. recent populations of Australia and South Africa both group in genetic group one), similarities in climate regime could represent an important factor affecting global gene flow patterns in *P. nodorum*.

Population-level diversity and demographics

We used nucleotide diversity and minor allele frequency spectra to quantify variation in genetic diversity across major genetic groups and local populations (Fig. 2). Genetic group three (northern USA and Iran) had the highest nucleotide diversity (4.39×10^{-3}) but also the most considerable variation among sites (Fig. 2a). North Dakota (USA) and Iran were highly diverse in contrast to South Dakota (USA). Genetic groups two (mainly southern USA) and 4 (Oklahoma, USA) had intermediate levels of nucleotide diversity (3.96×10^{-3} and 3.80×10^{-3}, respectively). Genetic group one had the lowest nucleotide diversity (3.63×10^{-3}). We analysed minor allele frequency spectra across populations to detect evidence for past demographic events (e.g. bottlenecks and expansions). Except for South Dakota (USA), populations showed low-frequency alleles being more abundant than high-frequency alleles, suggesting random mixing and low

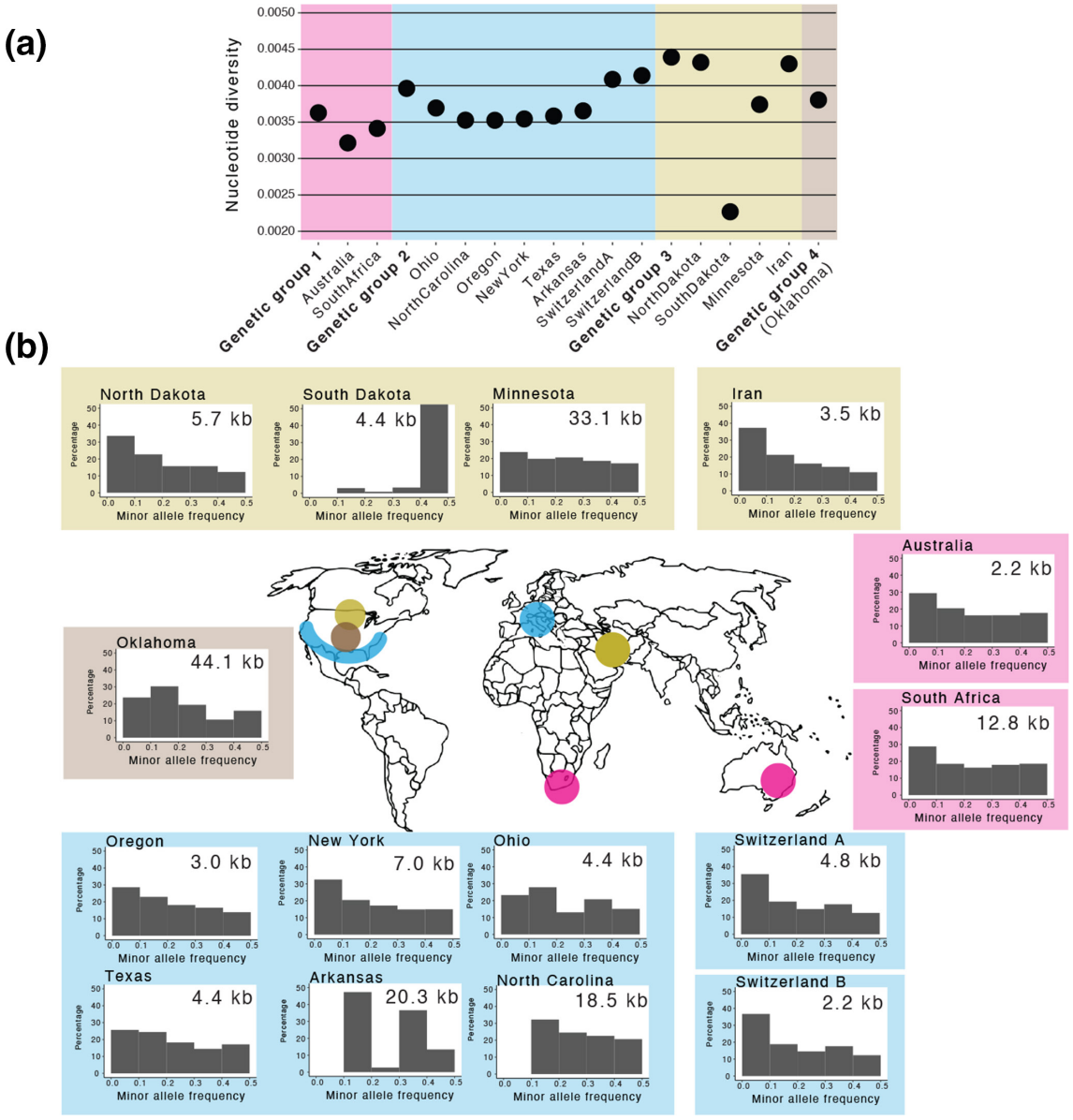

Fig. 2. Worldwide analyses of demographics and population diversity. (a) Nucleotide diversity for genetic groups and sampling locations. Colours represent different genetic groups. (b) Minor allele frequency spectrum for individual sampling locations. Sampling locations are represented on the map, highlighting northern United States and Iran in yellow; Oklahoma (USA) in brown; the southern United States and Switzerland in blue, and Australia and South Africa in pink. The number in the top right corner of histograms represents the distance for LD to decay below r^2 <0.2. All estimates are based on SNPs on chromosome one.

degrees of recent admixture (Fig. 2b). Next, we estimated LD decay and found that the distance for r^2 <0.2 varied between ~2.2 kb in Switzerland 1999B and Australia to 44.1 kb in Oklahoma (USA) (Fig. 2b). Most populations (10 out of 16) showed a fast LD decay (r^2 <0.2 within 10 kb), indicating high levels of diversity and ongoing recombination. The slow LD decay in South Africa and Oklahoma (USA) may indicate recent admixture or more dominant roles of asexual reproduction.

TE landscape and insertion dynamics among populations

Adaptation to local conditions may emerge from genetic variation produced by TE-mediated sequence rearrangements. Here, we performed *de novo* TE annotation using three complete genomes of *P. nodorum* [56] and combined this information into a single set of TE consensus sequences. Our consensus annotation identified 25 TE families in *P. nodorum* and revealed that the genomes Sn2000, Sn4 and Sn79-1087 are

composed of 4.40, 4.23 and 1.57% TEs, respectively (Fig. 3, Table S2). The TE density among the three reference genomes was heterogeneous (Fig. 3a). The reference isolate Sn79-1087 showed an overrepresentation of TEs mostly in subtelomeric regions. TEs in Sn2000 and Sn4 showed TE blocks along chromosomes with a particularly high density on chromosome 10 (Fig. 3a). Among the TE classes, class I elements comprised 3.76, 3.37 and 0.82% of the genome, and class II elements comprised 0.50, 0.72 and 0.42% of the genome in Sn2000, Sn4 and Sn79-1087, respectively. These findings are consistent with the previous estimates [48, 56]. Moreover, we show that complete genomes vary in the content and density of TEs highlighting the interest to screen TE dynamics at the population level.

We inferred population-level variation in TE activity by analysing the 348 isolates with genome sequences for evidence of newly inserted or deleted TEs. We identified 3850 TE insertions across all isolates with the insertions clustered into 167 unique loci in the genome (Fig. S4a). Recently inserted TEs were mainly DNA transposons (class II; *n*=2470) and included fewer retrotransposons (class I; *n*=1380). At the order level, we found only terminal inverted repeats (TIRs) and long terminal repeats (LTRs), respectively (Fig. 4a). Superfamilies were composed mainly of elements classified as class II *Tc1-Mariner* (*n*=1742), followed by class II *hAT* (*n*=182) and class I *Copia* (*n*=104; Fig. 4a). The remaining 1229 LTRs and 546 TIRs were not conserved enough to assign a superfamily. We found a stark difference between autonomous (*n*=602) and non-autonomous elements (*n*=3248; Figs 4a and S4b). Non-autonomous elements included 2019 miniature inverted-repeat transposable elements (MITEs; 62.1%) and 1229 terminal repeat retrotransposons in miniature (TRIMs; 37.9%). The total count of recent TE insertions in a population represents the load generated through TE activity. TEs were highly homogeneous between populations and genetic groups, with the highest mean numbers of new insertions in isolates of genetic group one (South Africa and Australia; mean=13.1 insertions) and the lowest in genetic group four (Oklahoma; mean=8.5 insertions; Figs 4c and S5). The TE superfamily composition is highly similar among genetic groups with a predominance in *Tc1-Mariner* and an unknown LTR (Fig. 4d). Recently inserted *Gypsy* elements were found only in genetic groups two and three. At the family level, insertions of DTX_MITE_Sirius were mostly found at a single locus on chromosome 19 (99%), and insertions of DTA_Mimas were mostly found at a single locus on chromosome 23 (67%). Chromosome 23 was identified as an accessory chromosome in *P. nodorum* [56, 93]. Similarly, TE families DTX_MITE_Ceti and DTX_MITE_Galatea were found only at a single locus on chromosomes 12 and 19, respectively. Most variation in TE loci numbers was driven by DTT_Tarvos (269 copies in total among isolates distributed across 59 loci; Fig. 3a). The RLC_Phobus showed a high degree of singletons with different 44 loci (Fig. 3a). Finally, TE copy expansion was driven by non-autonomous elements DTT_MITE_Geminga (821 copies), RLX_TRIM_Sinope (793 copies) and DTT_MITE_Eridani (652 copies). The species-wide screens of recently inserted TEs show that populations harbour fairly balanced TE loads with only a minor divergence in TE composition.

Differential activity among TEs and evidence for purifying selection

P. nodorum has an active genome defence called RIP, which induces C:G to T:A mutations within and proximal to duplicated sequences. The preferential AT-mutation bias of RIP will gradually lower the G+C content in RIP-affected genomic regions [32, 94]. We estimated G+C content as a proxy for RIP activity at the level of TE families. Active TEs not affected by RIP are expected to have a G+C content comparable to the genome-wide mean of coding sequences. We considered the population-level TE activity as the total number of copies per TE family across the 348 isolates. Interestingly, we found that TE families with higher copy numbers across the populations have higher G+C content than low-copy TE families (Fig. 3b). The three most-expanded TE families (DTT_MITE_Geminga, RLX_TRIM_Sinope and DTT_MITE_Eridani) showed the highest G+C content (mean G+C of 41.8, 43.7 and 48.2mol%, respectively). In contrast, TE families with copies divided across more loci were lower in G+C content, while those in fewer unique loci showed higher G+C content (Fig. 3c). The majority of TEs with high G+C content in both analyses were non-autonomous MITEs and TRIMs. We also found consistent variation in G+C content within TE families, with TEs of short length usually being of higher G+C content (e.g. MITEs and TRIMs) (Fig. 3d). Overall, recently active TEs in the *P. nodorum* genome are mostly of short length and constituted of non-autonomous elements.

TE frequencies across the genome are impacted by the joint actions of TE activity, selection on the insertion and demographics. We analysed whether the TE insertion frequency spectrum reflected neutral processes such as drift and migration, or selection. We performed a PCA using TE presence at insertion loci as a genetic marker. In contrast to genome-wide SNPs, we found no geographical signal for TE insertion loci (Fig. S6a–c). The majority of all TEs were present at <10% frequency in the global set of isolates, and about 48.5% of all TEs were singletons (Fig. S7). We found a particularly pronounced shift towards singletons in genetic group two, suggesting strong purifying selection acting against the TEs in this group (Figs 4b and S8).

Genomic signatures of recent selective sweeps

We analysed genetic groups for footprints of recent selective sweeps. We calculated the composite likelihood ratio implemented in the software SweeD for genetic groups one, two and three separately (group four composed of Oklahoma was omitted given the high admixture signatures). We detected a total of 46 genomic regions with significant signatures of recent sweeps across all groups (Fig. S9, Table S3). Selective sweeps were detected on all chromosomes except for chromosomes 1, 14 and 16. The size of the regions showing signatures of selective sweeps ranged from 20 to 40.7 kb. Interestingly, no overlaps in selective sweep regions were found among the

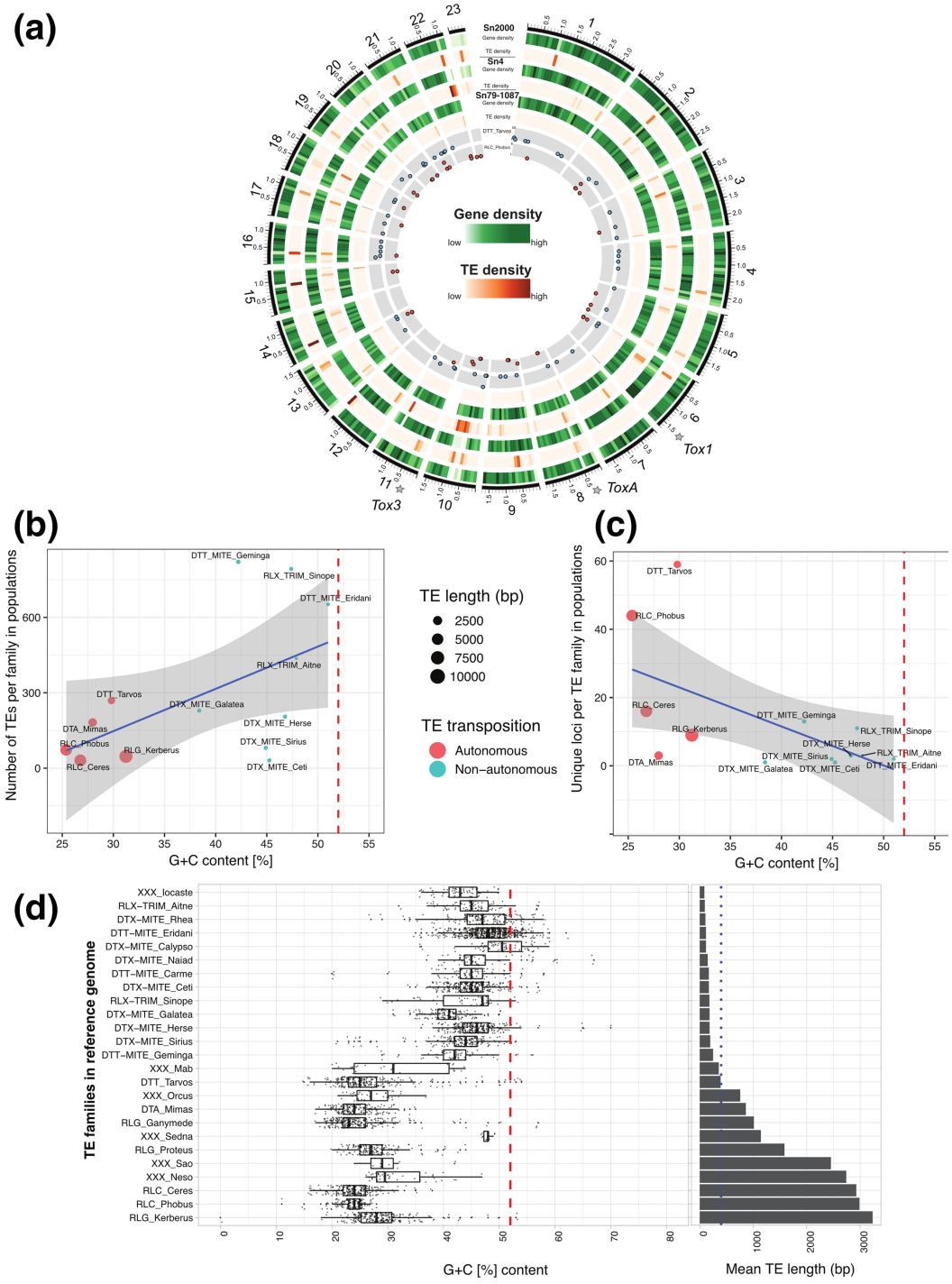

Fig. 3. Genome-wide detection of TEs. (a) Annotation of the completely assembled reference genomes. The outer black ring delimitates the 23 chromosomes and positions based on the Sn2000 reference genome (in Mb). Grey stars show approximate positions of the genes *ToxA*, *Tox1* and *Tox3* encoding effector proteins. The gene density is calculated for 100 kb windows (darker green for higher densities). The TE density was calculated for 100 kb windows (darker orange for higher densities). Inner grey rings highlight the occurrence of TE families with the highest number of loci in the genome (DTT_Tarvos with blue circles) and the highest degree of singleton (RLC_Phobus with red circles). The position on the *y*-axis shows the relative difference in the allele frequencies among populations. (b) Linear correlation plot of G+C content and TE copy number across populations. (c) Linear correlation plot of distinct loci per TE family across populations. (d) Analyses of TE copies in the three reference genomes (Sn2000, Sn4 and Sn79-1087) for G+C content and the mean TE length. Red vertical dashed lines indicate the genome-wide mean G+C content (52 mol%) and the dotted blue vertical line shows the 400 bp threshold.

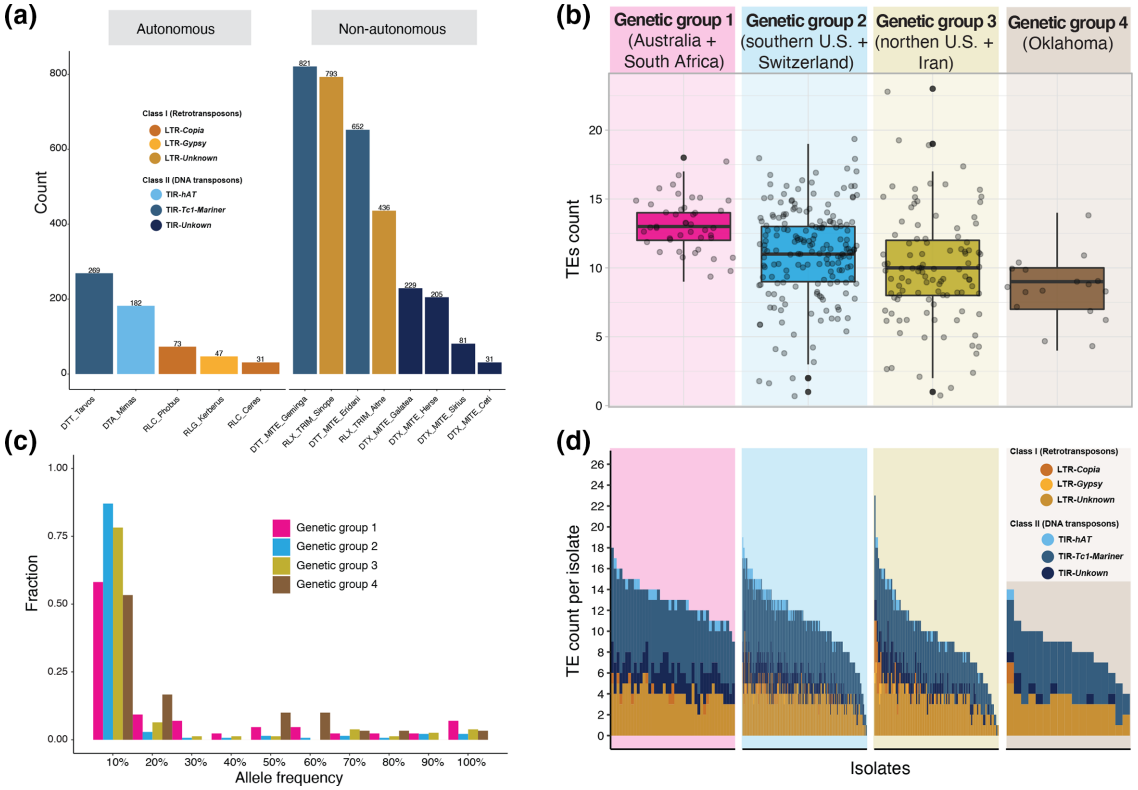

Fig. 4. Classification and differential load of TEs among genetic groups. (a) Bar plots of TE counts according to family and transposition mode. (b) TE insertion frequency distribution per genetic group. (c) The mean number of TEs per isolate within genetic groups. (d) TE counts per isolate and TE family across genetic groups. Each bar represents a single isolate and colours identify TE families.

genetic groups one, two and three, indicating strong heterogeneity in selection pressures and local adaptation. In genetic group one, a total of 16 genomic regions with sweep signatures were detected on chromosomes 2, 4–6, 8, 9, 11, 12, 15, 18–20 and 23 based on a 99.9% outlier threshold (Fig. 5a, Table S3). The mean length of the sweep regions was 21.3 kb spanning a total of 341.5 kb or 0.9% of the genome. The selective sweep region with the highest likelihood score (136) was located on chromosome two (Fig. 5b). We analysed Tajima's D values in the same region and found a positive but below-average value for the chromosome (Fig. 5c). LD decay was slower in the sweep region with a mean of r^2=0.48 over ~20.8 kb compared to a decay of r^2 <0.2 within 5.7 kb for genetic group one (Fig. 5e). The above-average LD is consistent with a recent selective sweep. In genetic groups two and three, we detected a total of 16 genomic regions with signatures of selection in total (Fig. S9, Table S3).

Genes potentially underlying local adaptation

Globally, we found 334 genes in selective sweep regions (genetic groups one, two and three; Table S3). The number of genes per sweep region ranged from 2 to 22. In genetic group one (South Africa and Australia), we identified a total of 99 genes (Table S4). The mean number of genes per sweep region was 6.2 ranging from 3 to 10 genes. We tested whether genes in sweep regions were overrepresented for genes encoding specific protein functions. We found 27 significantly overrepresented GO terms, including basal metabolic processes (e.g. cellular macromolecule biosynthetic/metabolic process), nutrient mobilization (e.g. nitrate assimilation) and transcriptional regulation (e.g. regulation of transcription, zinc ion binding; Table S5). The genomic region with the highest likelihood score in genetic group one contained two genes with roles in transcription, including *SNOG_02334* (a zinc-finger transcription factor) and *SNOG_30155* (a histidine phosphatase superfamily). We also identified genes encoding transporters, including *SNOG_02348* (amino acid permease) and *SNOG_02336* (cytochrome b5-like haem; Fig. 5d). Overall, genes in sweep regions were found to control metabolic functions and regulatory processes.

Among the 99 genes, 2 genes (*SNOG_07292*, *SNOG_30828*) were previously ranked as strong candidate effectors [37]. The genetic signal for a sweep found on chromosome eight (from 520242 to 540242 bp; coordinates on the reference genome Sn2000) was in close proximity to the locus containing the necrotrophic effector *ToxA*. A detailed analysis revealed a strong LD with the downstream region of *ToxA* and several TE insertion loci (Fig. S10). The presence of *Copia*, *Gypsy* and *Tc1-Mariner*-like elements were previously reported at the *ToxA* locus [47], here we identified additional MITEs and unclassified TEs (Fig. S10). In addition to TE insertions,

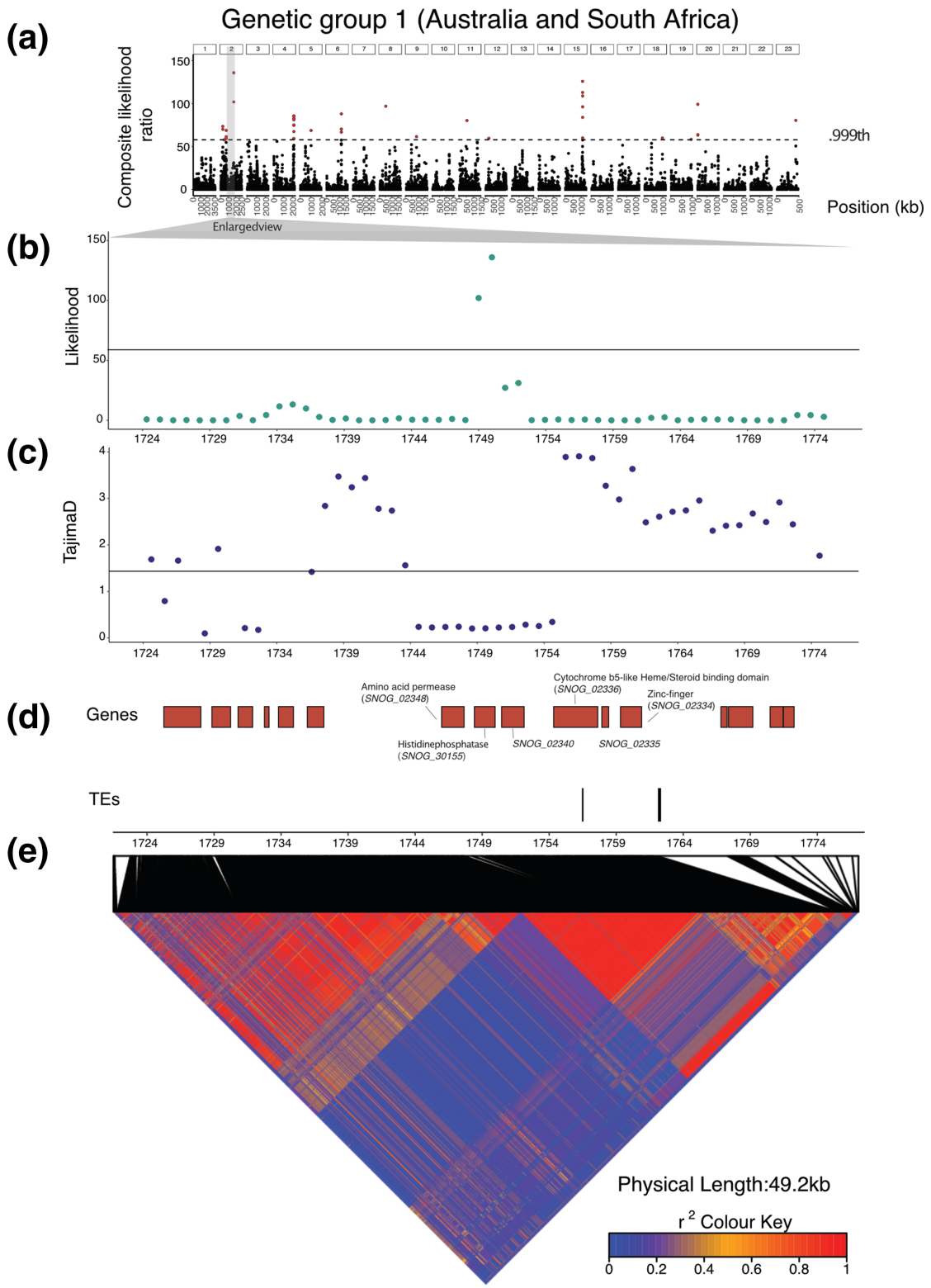

Fig. 5. Selective sweep loci in genetic group one. (a) Distribution of the composite likelihood scores in windows of 1 kb. The dashed horizontal line indicates the 99.9th percentile threshold and the grey highlighting shows the strongest sweep region. (b) Distribution of the composite likelihood scores within the strongest sweep region. (c) Tajima's D values calculated for 1 kb windows. (d) Schematic of genes highlighting predicted function and TEs. (e) Heatmap of pairwise LD r^2 within the sweep region. Gene annotation and chromosomal positions are based on the reference genome Sn2000.

the region covers 18 genes encoding among other proteins a kinase activator, a major facilitator superfamily transporter and signalling proteins (Table S6). We examined whether the 99 genes located in selective sweep regions from genetic group one were in close physical proximity of TEs, but we detected no overrepresentation of TEs in proximity of these genes compared to the genomic background (Fig. S11). Our findings show that the recent insertion dynamics of TEs in *P. nodorum* populations are unlikely to have driven recent selective sweeps. However, the insertion of individual TEs in proximity to known effector genes may well have had an impact on the evolution of virulence.

DISCUSSION

TE activity can be a major source of genetic variation in the fungal genome. Yet, we lack a comprehensive view of how evolutionary forces including genetic drift and selection act on TE dynamics in natural populations. Here, we analysed the intra-species population genetic diversity, genome-wide dynamics of TEs and signatures of selection in a worldwide collection of the plant pathogen *P. nodorum*. This species exhibits relatively low population subdivisions consistent with recent gene flow among continents. The TE compositions and copy numbers were very similar among populations. Despite strong evidence for highly active genomic defences [32, 49], we identified recent TE activity in the species pool. TEs do not appear to be the main drivers of recent adaptive evolution and are rather under purifying selection. Yet, we show that some high-copy TEs have sequence signatures consistent with an escape from genomic defence mechanisms.

For this work, we significantly expanded previous population genomic analyses of *P. nodorum* for both geographical coverage and sample size [24, 43, 44]. Our results corroborate evidence for genetic admixture among North America, Europe (Switzerland) and the Middle East (Iran) consistent with earlier analyses based on microsatellite loci [51]. High levels of gene flow among geographically widespread pathogen populations can lead to maladaptation and counteract local adaptation [95]. Yet, gene flow can be particularly advantageous for plant pathogens in the agro-ecosystem because agricultural practices and key susceptibility genes can be shared globally among crop growing areas [24, 95]. Indeed. we found high genetic diversity within and among field *P. nodorum* populations, with a rapid LD decay reflecting frequent sexual recombination. The slow LD decay found in more recently founded populations (e.g. South Africa) most likely reflects recent admixture. The historical expansion of wheat cultivation across the globe was likely accompanied by *P. nodorum* spreading from the Fertile Crescent to most continents. However, the global expansion of agriculture and pathogens likely imposed strong genetic bottlenecks in newly founded populations [45, 96]. Besides a reduction in genetic diversity, such demographic effects can influence genome-wide TE activity in some species [40, 97]. We found no striking differences in TE abundance among populations. Genetic group one, comprising the more recently founded populations of Australia and South Africa, showed similar TE composition as genetic group two (including Iranian isolates). Populations of the wheat pathogen *Z. tritici* showed striking expansion in most TE superfamilies [40, 98]. As *Z. tritici* and *P. nodorum* are thought to share common domestication origins and subsequent dispersal routes, differences in how TEs were impacted may be due to differences in genomic defences (e.g. cytosine DNA methylation polymorphism in *Z. tritici* [99]). Plant populations can similarly be shaped by bottleneck effects, but TEs did not show pronounced activation as found for example in the wild grass *Brachypodium distachyon* [97]. TE activity shaped by demographics highlights how genome-wide TE proliferation is complex and can be governed by stochastic effects.

We identified differences in the genome-wide TE density and distribution among the three completely assembled and re-annotated *P. nodorum* reference genomes. Polymorphism within the species in terms of TE content is indicative of TE activity and the existence of evolutionary hot spots. Such rapidly evolving regions in the genome were repeatedly associated with rapid evolution in plant pathogens. Rapid diversification of effector genes, which constitute a key virulence component, was thought as conferring strong advantages in the arms race with the plant host [12, 14]. However, loss of control over TE proliferation can have deleterious consequences to genome stability and the persistence of lineages [8, 100]. Compartmentalization of repetitive regions in the genome of *P. nodorum* was previously shown [56], and can mediate local adaptation by causing gene gains and losses [24]. The fact that the strain with the lowest TE content (Sn79-1087) is also avirulent on wheat may be an element of the evolutionary transitions in the lineage to become established on wheat in the agro-ecosystem [46, 56]. However, we found little evidence for TEs playing a major role in selective sweeps leading to local adaptation or sweep regions enriched in virulence factors in our populations. This is similar to other plant pathogens [101, 102], and the lack of enrichment in virulence-related genes may be explained by selection mostly acting on traits determined by small effect loci and not major gene-for-gene interaction loci (i.e. effectors).

The overall low number of repetitive sequences (<4.5% of genome length) and extensive signatures of RIP strongly suggest that the *P. nodorum* genome is largely devoid of TE activity [32, 48, 48]. This is consistent with the view that TE multiplication in a genome is determined by the balance between the TE's transposition potential and the effectiveness of genomic defences [103]. Here, we show that *P. nodorum* TE families exhibit a striking variation in G+C content, which is positively correlated with the TE copy numbers in the genome. The correlation suggests that an increase in TE copy numbers is associated with weak RIP activity on these TEs. RIP guards the genome against TE proliferation by targeting duplicated DNA (e.g. recent copies of TEs), but the RIP intensity can vary among TE families [104, 105]. The impact of RIP can also extend beyond repetitive regions, increasing mutation rates across the entire genome [106] and significantly impacting pathogenicity [22, 33]. In *N. crassa,* the action of

RIP is impaired in duplicated sequences shorter than ~400 bp [94, 107], and with an identity below 80% [108]. We found that high-copy TE families with high G+C content are mostly MITEs and TRIMs. These elements generally lack coding sequences [85] and are on average <400 bp in *P. nodorum*. Hence, the MITEs and TRIMs may have escaped RIP through their compactness. Both groups of TEs were found to mediate genome evolution and modulate gene expression across eukaryotes [41, 109–111]. In addition, MITEs are thought to increase in copy number despite genomic defences with possible benefits to DNA transposons [112, 113] and underlie the expression of pathogenicity [22, 114, 115]. MITEs and TRIMs are ubiquitous in the pangenome of the wheat pathogen *Z. tritici*, which has a similar genome size but more active TEs [39]. MITEs in *Z. tritici* become de-repressed upon encountering stressful conditions, such as during the colonization of the host plant [41]. The MITE activity in *P. nodorum* could also be a consequence of post-transcriptional regulation. In rice, short interfering RNA was found to preferentially suppress MITE expression [28, 116]. Genomic regions showing signatures of recent selective sweeps did not contain recently inserted TEs. Whether the activity of MITEs and TRIMs in *P. nodorum* is contingent on stressful conditions and whether transposition activity creates adaptive variation remains unknown.

In this study, we resolved the global population structure and retraced TE activity in the repressive genome of an important wheat pathogen. TEs can play a major role in plant pathogen adaptation by generating adaptive genetic variation to resist pesticides or to overcome host defence mechanisms. Key pathogenicity genes show tightly associated expression patterns with TEs in several plant pathogens. However, the selfish nature of TEs can also impose severe costs on the pathogen. Population genetic analyses informed by highly contiguous genome sequences are powerful tools to disentangle the evolutionary equilibrium between TE activation and repression.

Funding information
This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior Brasil (CAPES) finance code 001.

Acknowledgements
We thank Cécile Lorrain and Sandra Lorena Ament-Velásquez for comments and suggestions on a previous version of the manuscript. The Genetic Diversity Center (GDC) of ETH Zurich and the Functional Genomics Center in Zurich provided laboratory and sequencing facilities.

Conflicts of interest
The authors declare that there are no conflicts of interest.

References

1. Baranova MA, Logacheva MD, Penin AA, Seplyarskiy VB, Safonova YY *et al*. Extraordinary genetic diversity in a wood decay mushroom. *Mol Biol Evol* 2015;32:2775–2783.
2. Goddard MR, Godfray HCJ, Burt A. Sex increases the efficacy of natural selection in experimental yeast populations. *Nature* 2005;434:636–640.
3. Wolfe KH, Shields DC. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 1997;387:708–713.
4. Horns F, Petit E, Hood ME. Massive expansion of *Gypsy*-like retrotransposons in *Microbotryum* fungi. *Genome Biol Evol* 2017;9:363–371.
5. Kidwell MG. Transposable elements and the evolution of genome size in eukaryotes. *Genetica* 2002;115:49–63.
6. De Cecco M, Criscione SW, Peckham EJ, Hillenmeyer S, Hamm EA *et al*. Genomes of replicatively senescent cells undergo global epigenetic changes leading to gene silencing and activation of transposable elements. *Aging Cell* 2013;12:247–256.
7. Chen F, Everhart SE, Bryson PK, Luo C, Song X *et al*. Fungicide-induced transposon movement in *Monilinia fructicola*. *Fungal Genet Biol* 2015;85:38–44.
8. Hedges DJ, Deininger PL. Inviting instability: transposable elements, double-strand breaks, and the maintenance of genome integrity. *Mutat Res* 2007;616:46–59.
9. Burns KH, Boeke JD. Human transposon tectonics. *Cell* 2012;149:740–752.
10. Chou H-H, Hayakawa T, Diaz S, Krings M, Indriati E *et al*. Inactivation of CMP-N-acetylneuraminic acid hydroxylase occurred prior to brain expansion during human evolution. *Proc Natl Acad Sci USA* 2002;99:11736–11741.
11. Desalvo MK, Voolstra CR, Sunagawa S, Schwarz JA, Stillman JH *et al*. Differential gene expression during thermal stress and bleaching in the Caribbean coral *Montastraea faveolata*. *Mol Ecol* 2008;17:3952–3971.
12. Raffaele S, Kamoun S. Genome evolution in filamentous plant pathogens: why bigger can be better. *Nat Rev Microbiol* 2012;10:417–430.
13. Möller M, Stukenbrock EH. Evolution and genome architecture in fungal plant pathogens. *Nat Rev Microbiol* 2017;15:756–771.
14. Seidl MF, Thomma BPHJ. Transposable elements direct the coevolution between plants and microbes. *Trends Genet* 2017;33:842–851.
15. Spanu PD, Abbott JC, Amselem J, Burgis TA, Soanes DM *et al*. Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. *Science* 2010;330:1543–1546.
16. Hartmann FE, Sánchez-Vallet A, McDonald BA, Croll D. A fungal wheat pathogen evolved host specialization by extensive chromosomal rearrangements. *ISME J* 2017;11:1189–1204.
17. Omrane S, Sghyer H, Audéon C, Lanen C, Duplaix C *et al*. Fungicide efflux and the *MgMFS1* transporter contribute to the multidrug resistance phenotype in *Zymoseptoria tritici* field isolates. *Environ Microbiol* 2015;17:2805–2823.
18. Steinhauer D, Salat M, Frey R, Mosbach A, Luksch T *et al*. A dispensable paralog of succinate dehydrogenase subunit C mediates standing resistance towards a subclass of SDHI fungicides in *Zymoseptoria tritici*. *PLoS Pathog* 2019;15:e1007780.
19. Dutheil JY, Mannhaupt G, Schweizer G, M K Sieber C, Münsterkötter M *et al*. A tale of genome compartmentalization: the evolution of virulence clusters in smut fungi. *Genome Biol Evol* 2016;8:681–704.
20. Faino L, Seidl MF, Shi-Kunne X, Pauper M, van den Berg GCM *et al*. Transposons passively and actively contribute to evolution of the two-speed genome of a fungal pathogen. *Genome Res* 2016;26:1091–1100.
21. Wang Q, Jiang C, Wang C, Chen C, Xu J-R, *et al*. Characterization of the two-speed subgenomes of *Fusarium graminearum* reveals the fast-speed subgenome specialized for adaption and infection. *Front Plant Sci* 2017;8:140.
22. Rouxel T, Grandaubert J, Hane JK, Hoede C, van de Wouw AP *et al*. Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by repeat-induced point mutations. *Nat Commun* 2011;2:202.
23. Fouché S, Plissonneau C, Croll D. The birth and death of effectors in rapidly evolving filamentous pathogen genomes. *Curr Opin Microbiol* 2018;46:34–42.

24. Richards JK, Stukenbrock EH, Carpenter J, Liu Z, Cowger C *et al*. Local adaptation drives the diversification of effectors in the fungal wheat pathogen *Parastagonospora nodorum* in the United States. *PLoS Genet* 2019;15:e1008223.
25. Yoshida K, Saunders DGO, Mitsuoka C, Natsume S, Kosugi S *et al*. Host specialization of the blast fungus *Magnaporthe oryzae* is associated with dynamic gain and loss of genes linked to transposable elements. *BMC Genomics* 2016;17:370.
26. Hartmann FE, Croll D. Distinct trajectories of massive recent gene gains and losses in populations of a microbial eukaryotic pathogen. *Mol Biol Evol* 2017;34:2808–2822.
27. Hollister JD, Gaut BS. Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Res* 2009;19:1419–1428.
28. Nolan T. The post-transcriptional gene silencing machinery functions independently of DNA methylation to repress a LINE1-like retrotransposon in *Neurospora crassa*. *Nucleic Acids Res* 2005;33:1564–1573.
29. Gladyshev E. Repeat-induced point mutation and other genome defense mechanisms in fungi. *Microbiol Spectr* 2017;5:FUNK-0042-2017.
30. Selker EU. Premeiotic instability of repeated sequences in *Neurospora crassa*. *Annu Rev Genet* 1990;24:579–613.
31. Ikeda K, Nakayashiki H, Kataoka T, Tamba H, Hashimoto Y. Repeat-induced point mutation (RIP) in *Magnaporthe grisea*: implications for its sexual cycle in the natural field context: repeat-induced point mutation in *Magnaporthe grisea*. *Mol Microbiol* 2002;45:1355–1364.
32. Hane JK, Oliver RP. RIPCAL: a tool for alignment-based analysis of repeat-induced point mutations in fungal genomic sequences. *BMC Bioinform* 2008;9:478.
33. Fudal I, Ross S, Brun H, Besnard A-. L, Ermel M, *et al*. Repeat-induced point mutation (RIP) as an alternative mechanism of evolution toward vi rulence in *Leptosphaeria maculans*. *Mol Plant Microbe Interact* 2009;22:932–941.
34. Dhillon B, Gill N, Hamelin RC, Goodwin SB. The landscape of transposable elements in the finished genome of the fungal wheat pathogen *Mycosphaerella graminicola*. *BMC Genom* 2014;15:1132.
35. Van de Wouw AP, Elliott CE, Popa KM, Idnurm A. Analysis of repeat induced point (RIP) mutations in *Leptosphaeria maculans* indicates variability in the RIP process between fungal species. *Genetics* 2019;211:89–104.
36. Plissonneau C, Hartmann FE, Croll D. Pangenome analyses of the wheat pathogen *Zymoseptoria tritici* reveal the structural basis of a highly plastic eukaryotic genome. *BMC Biol* 2018;16:5.
37. Syme RA, Tan K-C, Rybak K, Friesen TL, McDonald BA *et al*. Pan-*Parastagonospora* comparative genome analysis – effector prediction and genome evolution. *Genome Biol Evol* 2018;10:2443–2457.
38. Wyatt NA, Richards JK, Brueggeman RS, Friesen TL. Reference assembly and annotation of the *Pyrenophora teres* f. *teres* isolate 0-1. *G3* 2018;8:g3.117.300196.
39. Badet T, Oggenfuss U, Abraham L, McDonald BA, Croll D. A 19-isolate reference-quality global pangenome for the fungal wheat pathogen *Zymoseptoria tritici*. *BMC Biol* 2020;18:12.
40. Oggenfuss U, Badet T, Wicker T, Hartmann FE, Singh NK, *et al*. A population-level invasion by transposable elements in a fungal pathogen. *bioRxiv* 2020:944652.
41. Fouché S, Badet T, Oggenfuss U, Plissonneau C, Francisco CS *et al*. Stress-driven transposable element de-repression dynamics and virulence evolution in a fungal pathogen. *Mol Biol Evol* 2020;37:221–239.
42. Smith KM, Phatale PA, Bredeweg EL, Connolly LR, Pomraning KR, *et al*. Epigenetics of filamentous fungi. In: Meyers R (eds). *Encyclopedia of Molecular Cell Biology and Molecular Medicine*. Weinheim: Wiley-VCH Verlag; 2012.
43. Pereira D, McDonald BA, Croll D. The genetic architecture of emerging fungicide resistance in populations of a global wheat pathogen. *Genome Biol Evol* 2020;12:2231–2244.
44. Pereira D, Croll D, Brunner PC, McDonald BA. Natural selection drives population divergence for local adaptation in a wheat pathogen. *Fungal Genet Biol* 2020;141:103398.
45. Oliver R, Friesen T, Faris J, Solomon P. *Stagonospora nodorum*: from pathology to genomics and host resistance. *Annu Rev Phytopathol* 2012;50:23–43.
46. Friesen TL, Stukenbrock EH, Liu Z, Meinhardt S, Ling H *et al*. Emergence of a new disease as a result of interspecific virulence gene transfer. *Nat Genet* 2006;38:953–956.
47. McDonald MC, Taranto AP, Hill E, Schwessinger B, Liu Z *et al*. Transposon-mediated horizontal transfer of the host-specific virulence protein ToxA between three fungal wheat pathogens. *mBio* 2019;10:e01515-19.
48. Hane JK, Lowe RGT, Solomon PS, Tan K-. C, Schoch CL, *et al*. Dothideomycete–plant interactions illuminated by genome sequencing and EST analysis of the wheat pa thogen *Stagonospora nodorum*. *Plant Cell* 2007;19:3347–3368.
49. Hane JK, Oliver RP. *In silico* reversal of repeat-induced point mutation (RIP) identifies the origins of repeat families and uncovers obscured duplicated genes. *BMC Genom* 2010;11:655.
50. Sommerhalder RJ, McDonald BA, Zhan J. The frequencies and spatial distribution of mating types in *Stagonospora nodorum* are consistent with recurring sexual reproduction. *Phytopathology* 2006;96:234–239.
51. Stukenbrock EH, Banke S, McDonald BA. Global migration patterns in the fungal wheat pathogen *Phaeosphaeria nodorum*. *Mol Ecol* 2006;15:2895–2904.
52. McDonald MC, Razavi M, Friesen TL, Brunner PC, McDonald BA. Phylogenetic and population genetic analyses of *Phaeosphaeria nodorum* and its close relatives indicate cryptic species and an origin in the Fertile Crescent. *Fungal Genet Biol* 2012;49:882–895.
53. McDonald MC, Oliver RP, Friesen TL, Brunner PC, McDonald BA. Global diversity and distribution of three necrotrophic effectors in *Phaeosphaeria nodorum* and related species. *New Phytol* 2013;199:241–251.
54. Pereira DA, McDonald BA, Brunner PC. Mutations in the *CYP51* gene reduce DMI sensitivity in *Parastagonospora nodorum* populations in Europe and China. *Pest Manag Sci* 2017;73:1503–1510.
55. Bolger A, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30:2114–2120.
56. Richards JK, Wyatt NA, Liu Z, Faris JD, Friesen TL. Reference quality genome assemblies of three *Parastagonospora nodorum* isolates differing in virulence on wheat. *G3* 2018;8:393–399.
57. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;9:357–359.
58. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25:2078–2079.
59. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K *et al*. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20:1297–1303.
60. Danecek P, Auton A, Abecasis G, Albers CA, Banks E *et al*. The variant call format and VCFtools. *Bioinformatics* 2011;27:2156–2158.
61. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014;30:1312–1313.
62. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* 2006;23:254–267.
63. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* 2019;47:W256–W259.
64. Lischer HEL, Excoffier L. PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* 2012;28:298–299.

65. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics* 2000;155:945–959.
66. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y *et al*. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 2007;23:2633–2635.
67. R Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2019. https://www.R-project.org
68. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*, 2nd edn. New York: Springer; 2009.
69. Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software structure: a simulation study. *Mol Ecol* 2005;14:2611–2620.
70. Francis RM. pophelper: an R package and web app to analyse and visualize population structure. *Mol Ecol Resour* 2017;17:27–32.
71. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 1989;123:585–595.
72. Pavlidis P, Živkovic D, Stamatakis A, Alachiotis N. SweeD: likelihood-based detection of selective sweeps in thousands of genomes. *Mol Biol Evol* 2013;30:2224–2234.
73. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;26:841–842.
74. Jones P, Binns D, Chang H-Y, Fraser M, Li W *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* 2014;30:1236–1240.
75. Nielsen H. Predicting secretory proteins with signalP. In: Kihara D (ed). *Protein Function Prediction*. New York: Springer; 2017. pp. 59–73.
76. Käll L, Krogh A, Sonnhammer ELL. A combined transmembrane topology and signal peptide prediction method. *J Mol Biol* 2004;338:1027–1036.
77. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* 2001;305:567–580.
78. Falcon S, Gentleman R. Using GOstats to test gene lists for GO term association. *Bioinformatics* 2007;23:257–258.
79. Bao W, Kojima KK, Kohany O. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* 2015;6:11.
80. Smit AFA, Hubley R, Green P. RepeatMasker Open-4.0 (http://www.repeatmasker.org); 2013.
81. Breen JM, Wicker T, Kong X, Zhang J, Ma W *et al*. A highly conserved gene island of three genes on chromosome 3B of hexaploid wheat: diverse gene function and genomic structure maintained in a tightly linked block. *BMC Plant Biol* 2010;10:98.
82. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25:3389–3402.
83. Higgins DG, Sharp PM. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* 1988;73:237–244.
84. Sonnhammer ELL, Durbin R. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* 1995;167:GC1–GC10.
85. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P *et al*. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* 2007;8:973–982.
86. Guy L, Kultima JR, Andersson SGE. genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* 2010;26:2334–2335.
87. Morgulis A, Coulouris G, Raytselis Y, Madden TL, Agarwala R *et al*. Database indexing for production MegaBLAST searches. *Bioinformatics* 2008;24:1757–1764.
88. Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. *J Comput Biol* 2000;7:203–214.
89. Madeira F, Park YM, Lee J, Buso N, Gur T *et al*. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res* 2019;47:W636–W641.
90. Linheiro RS, Bergman CM. Whole genome resequencing reveals natural target site preferences of transposable elements in *Drosophila melanogaster*. *PloS One* 2012;7:e30008.
91. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;25:1754–1760.
92. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M *et al*. Software for computing and annotating genomic ranges. *PLoS Comput Biol* 2013;9:e1003118.
93. Ohm RA, Feau N, Henrissat B, Schoch CL, Horwitz BA *et al*. Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen Dothideomycetes fungi. *PLoS Pathog* 2012;8:e1003037.
94. Watters MK, Randall TA, Margolin BS, Selker EU, Stadler DR. Action of repeat-induced point mutation on both strands of a duplex and on tandem duplications of various sizes in Neurospora. *Genetics* 1999;153:705–714.
95. Croll D, McDonald BA. The genetic basis of local adaptation for pathogenic fungi in agricultural ecosystems. *Mol Ecol* 2017;26:2027–2040.
96. McDonald BA, Linde C. Pathogen population genetics, evolutionary potential, and durable resistance. *Annu Rev Phytopathol* 2002;40:349–379.
97. Stritt C, Gordon SP, Wicker T, Vogel JP, Roulin AC. Recent activity in expanding populations and purifying selection have shaped transposable element landscapes across natural accessions of the Mediterranean grass *Brachypodium distachyon*. *Genome Biol Evol* 2018;10:304–318.
98. Stukenbrock EH, Banke S, Javan-Nikkhah M, McDonald BA. Origin and domestication of the fungal wheat pathogen *Mycosphaerella graminicola* via sympatric speciation. *Mol Biol Evol* 2007;24:398–411.
99. Möller M, Habig M, Lorrain C, Feurtey A, Haueisen J, *et al*. Recent loss of the Dim2 DNA methyltransferase decreases mutation rate in repeats and changes evolutionary trajectory in a fungal pathogen. *bioRxiv* 2020:012203.
100. Quadrana L, Bortolini Silveira A, Mayhew GF, LeBlanc C, Martienssen RA *et al*. The Arabidopsis thaliana mobilome and its impact at the species level. *eLife* 2016;5:e15716.
101. Badouin H, Gladieux P, Gouzy J, Siguenza S, Aguileta G *et al*. Widespread selective sweeps throughout the genome of model plant pathogenic fungi and identification of effector candidates. *Mol Ecol* 2017;26:2041–2062.
102. Mohd-Assaad N, McDonald BA, Croll D. Genome-Wide detection of genes under positive selection in worldwide populations of the barley scald pathogen. *Genome Biol Evol* 2018;10:1315–1332.
103. Muszewska A, Steczkiewicz K, Stepniewska-Dziubinska M, Ginalski K. Cut-and-paste transposons in fungi with diverse lifestyles. *Genome Biol Evol* 2017;9:3463–3477.
104. Santana MF, Silva JCF, Mizubuti ESG, Araújo EF, Queiroz MV. Analysis of Tc1-Mariner elements in *Sclerotinia sclerotiorum* suggests recent activity and flexible transposases. *BMC Microbiol* 2014;14:256.
105. Amselem J, Lebrun M-H, Quesneville H. Whole genome comparative analysis of transposable elements provides new insight into mechanisms of their inactivation in fungal genomes. *BMC Genom* 2015;16:141.
106. Wang L, Sun Y, Sun X, Yu L, Xue L *et al*. Repeat-induced point mutation in *Neurospora crassa* causes the highest known mutation rate and mutational burden of any cellular life. *Genome Biol* 2020;21:142.
107. Yeadon PJ, Catcheside DEA. Guest: a 98 bp inverted repeat transposable element in *Neurospora crassa*. *Mol Gen Genet* 1995;247:105–109.
108. Cambareri EB, Singer MJ, Selker EU. Recurrence of repeat-induced point mutation (RIP) in *Neurospora crassa*. *Genetics* 1991;127:699–710.

109. Witte C-P, Le QH, Bureau T, Kumar A. Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. *Proc Natl Acad Sci USA* 2001;98:13778–13783.
110. Santiago N, Herráiz C, Goñi JR, Messeguer X, Casacuberta JM. Genome-wide analysis of the emigrant family of MITEs of *Arabidopsis thaliana*. *Mol Biol Evol* 2002;19:2285–2293.
111. Naito K, Zhang F, Tsukiyama T, Saito H, Hancock CN *et al*. Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature* 2009;461:1130–1134.
112. Feschotte C, Swamy L, Wessler SR. Genome-wide analysis of mariner-like transposable elements in rice reveals complex relationships with stowaway miniature inverted repeat transposable elements (MITEs). *Genetics* 2003;163:747–758.
113. Feschotte C, Pritham EJ. DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet* 2007;41:331–368.
114. Kang S, Lebrun MH, Farrall L, Valent B. Gain of virulence caused by insertion of a Pot3 transposon in a *Magnaporthe grisea* avirulence gene. *Mol Plant Microbe Interact* 2001;14:671–674.
115. Wachter S, Raghavan R, Wachter J, Minnick MF. Identification of novel MITEs (miniature inverted-repeat transposable elements) in *Coxiella burnetii*: implications for protein and small RNA evolution. *BMC Genomics* 2018;19:247.
116. Lu C, Chen J, Zhang Y, Hu Q, Su W *et al*. Miniature inverted-repeat transposable elements (MITEs) have been accumulated through amplification bursts and play important roles in gene expression and species diversity in *Oryza sativa*. *Mol Biol Evol* 2012;29:1005–1017.