A population-level invasion by transposable elements triggers genome expansion in a fungal pathogen

Ursula Oggenfuss[1], Thomas Badet[1], Thomas Wicker[2], Fanny E. Hartmann[3,4], Nikhil K. Singh[1], Leen N. Abraham[1], Petteri Karisto[4,6], Tiziana Vonlanthen[4], Christopher C. Mundt[5], Bruce A. McDonald[4], Daniel Croll[1,*]

[1] Laboratory of Evolutionary Genetics, Institute of Biology, University of Neuchâtel, 2000 Neuchâtel, Switzerland
[2] Institute for Plant and Microbial Biology, University of Zurich, Zurich, Switzerland
[3] Ecologie Systématique Evolution, Bâtiment 360, Univ. Paris-Sud, AgroParisTech, CNRS, Université Paris-Saclay, 91400 Orsay, France
[4] Plant Pathology, Institute of Integrative Biology, ETH Zurich, Zurich, Switzerland
[5] Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR 97331-2902, USA
[6] Department of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland

[*] Author for correspondence: daniel.croll@unine.ch

Running title: Transposable element invasion triggers genome expansion

ABSTRACT

Genome evolution is driven by the activity of transposable elements (TEs). The spread of TEs can have deleterious effects including the destabilization of genome integrity and expansions. However, the precise triggers of genome expansions remain poorly understood because genome size evolution is typically investigated only among deeply divergent lineages. Here, we use a large population genomics dataset of 284 individuals from populations across the globe of *Zymoseptoria tritici*, a major fungal wheat pathogen. We built a robust map of genome-wide TE insertions and deletions to track a total of 2'456 polymorphic loci within the species. We show that purifying selection substantially depressed TE frequencies in most populations but some rare TEs have recently risen in frequency and likely confer benefits. We found that specific TE families have undergone a substantial genome-wide expansion from the pathogen's center of origin to more recently founded populations. The most dramatic increase in TE insertions occurred between a pair of North American populations collected in the same field at an interval of 25 years. We find that both genome-wide counts of TE insertions and genome size have increased with colonization bottlenecks. Hence, the demographic history likely played a major role in shaping genome evolution within the species. We show that both the activation of specific TEs and relaxed purifying selection underpin this incipient expansion of the genome. Our study establishes a model to recapitulate TE-driven genome evolution over deeper evolutionary timescales.

INTRODUCTION

Transposable elements (TEs) are mobile repetitive DNA sequences with the ability to independently insert into new regions of the genome. TEs are major drivers of genome instability and epigenetic change (Eichler & Sankoff, 2003). Insertion of TEs can disrupt coding sequences, trigger chromosomal rearrangements, or alter expression profiles of adjacent genes (Lim, 1988; Petrov *et al.*, 2003; Slotkin & Martienssen, 2007; Hollister & Gaut, 2009; Oliver *et al.*, 2013). Hence, TE activity can have phenotypic consequences and impact host fitness. While TE insertion dynamics are driven by the selfish interest for proliferation, the impact on the host can range from beneficial to highly deleterious. The most dramatic examples of TE insertions underpinned rapid adaptation of populations or species (Feschotte, 2008; Chuong *et al.*, 2017), particularly following environmental change or colonization events. Beneficial TE insertions are expected to experience strong positive selection and rapid fixation in populations. However, most TE insertions have neutral or deleterious effects upon insertions. Purifying selection is expected to rapidly eliminate deleterious insertions from populations unless constrained by genetic drift (Walser *et al.*, 2006; Baucom *et al.*, 2008; Cridland *et al.*, 2013; Stuart *et al.*, 2016; Lai *et al.*, 2017; Stritt *et al.*, 2017). Additionally, genomic defense mechanisms can disable transposition activity. Across eukaryotes, epigenetic silencing is a shared defense mechanism against TEs (Slotkin & Martienssen, 2007). Fungi evolved an additional and highly specific defense system introducing repeat-induced point (RIP) mutations into any nearly identical set of sequences. The relative importance of demography, selection and genomic defenses determining the fate of TEs in populations remain poorly understood.

A crucial property predicting the invasion success of TEs in a genome is the transposition rate. TEs tend to expand through family-specific bursts of transposition followed by prolonged phases of transposition inactivity. Bursts of insertions of different retrotransposon families were observed across eukaryotic lineages including *Homo sapiens*, *Zea mays*, *Oryza sativa* and *Blumeria graminis* (Shen *et al.*, 1991; SanMiguel *et al.*, 1998; Eichler & Sankoff, 2003; Lu *et al.*, 2017; Frantzeskakis *et al.*, 2018). Prolonged bursts without effective counter-selection are thought to underpin genome expansions. In the symbiotic fungus *Cenococcum geophilum*, the burst of TEs resulted in a dramatically expanded genome compared to closely related species (Peter *et al.*, 2016). Similarly, a burst of a TE family in brown hydras led to an approximately three-fold increase of the genome size compared to related hydras (Wong *et al.*, 2019). Across the tree of life, genome sizes vary by orders of magnitude and enlarged genomes invariably show hallmarks of historic TE invasions (Kidwell, 2002). Population size variation is among the few correlates of genome size across major groups, suggesting that the efficacy of selection plays an important role in controlling TE activity (Lynch, 2007). Reduced selection efficacy against deleterious TE insertions is expected to lead to a ratchet-like increase in genome size. In fungi, TE-rich genomes often show an isochore structure alternating gene-rich and TE-rich compartments (Rouxel *et al.*, 2011). TE-rich compartments often harbor rapidly evolving genes such as effector genes in pathogens or resistance genes in plants (Raffaele & Kamoun, 2012; Jiao & Schneeberger, 2019). Taken together, incipient genome expansions are likely driven by population-level TE insertion dynamics.

The fungal wheat pathogen *Zymoseptoria tritici* is one of the most important pathogens on crops causing high yield losses (Torriani *et al.*, 2015). The genome is completely assembled and shows size variation between individuals sampled across the global distribution range (Feurtey *et al.*, 2020; Badet *et al.*, 2020) (Goodwin *et al.*, 2011). The TE content of the genome shows a striking variation of 17-24% variation among individuals (Badet *et al.*, 2020). *Z. tritici* recently gained major TE-mediated adaptations to colonize host plants and tolerate environmental stress (Omrane *et al.*, 2015, 2017; Krishnan *et al.*, 2018; Meile *et al.*, 2018). Clusters of TEs are often associated with genes encoding important pathogenicity functions (*i.e.* effectors), recent gene gains or losses (Hartmann & Croll, 2017), and major chromosomal rearrangements (Croll *et al.*, 2013; Plissonneau *et al.*, 2016). Transposition activity of TEs also had a genome-wide impact on gene expression profiles during infection (Fouché *et al.*, 2019). The well-characterized demographic history of the pathogen and evidence for recent TE-mediated adaptations make *Z. tritici* an ideal model to recapitulate the process of TE insertion dynamics, adaptive evolution and changes in genome size at the population level.

Here, we retrace the population-level context of TE insertion dynamics and genome size changes across the species range by analyzing populations sampled on four continents for a total of 284 genomes. We developed a robust pipeline to detect newly inserted TEs using short read sequencing datasets. Combining analyses of selection and knowledge of the colonization history of the pathogen, we tested whether population bottlenecks were associated with substantial changes in the TE content and the size of genomes.

RESULTS

A DYNAMIC TE LANDSCAPE SHAPED BY STRONG PURIFYING SELECTION

To establish a comprehensive picture of within-species TE dynamics, we analyzed 284 genomes from a worldwide set of six populations spanning the distribution range of the wheat pathogen *Z. tritici*. To ascertain the presence or absence of TEs across the genome, we developed a robust pipeline (Figure 1A). In summary, we called TE insertions by identifying reads mapping both to a TE sequence and a specific location in the reference genome. Then, we assessed the minimum sequencing coverage to reliably recover TE insertions, tested for evidence of TEs using read depth at target site duplications, and scanned the genome for mapped reads indicating gaps at TE loci. We found robust evidence for a total of 18’864 TE insertions grouping into 2’465 individual loci. More than 30% of these loci have singleton TEs (*i.e.* this locus is only present in one isolate; Figure 2B, Supplementary Table S3). An overwhelming proportion of loci (2’345 loci or 95.1%) have a TE frequency below 1%. This pattern strongly supports the hypothesis that TEs actively copy into new locations but also indicates that strong purifying selection maintains nearly all TEs at low frequency (Figure 2B). We found a higher density of TE loci on accessory chromosomes, which are not shared among all isolates of the species, compared to core chromosomes (Figure 2C). This suggests relaxed selection against TE insertion on the functionally dispensable accessory chromosomes.

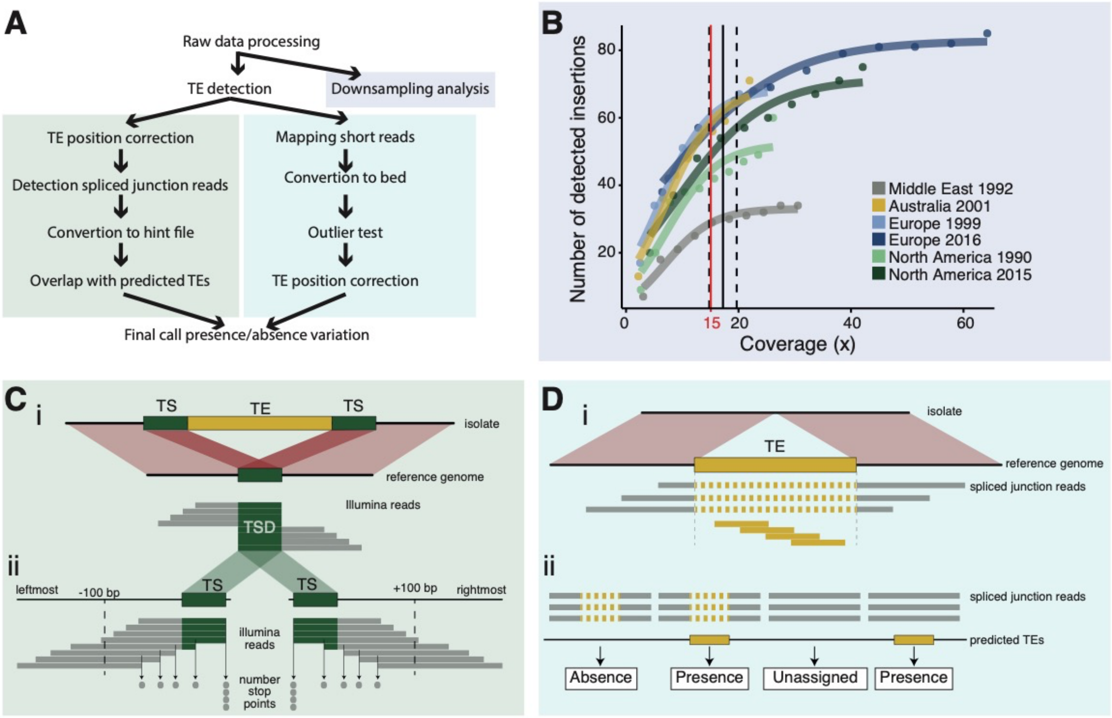


Figure 1: Robust discovery and validation of transposable element (TE) insertions: (A) General analysis pipeline. (B) Read depth down-sampling analysis for one isolate per population with an average coverage of the population. The vertical black line indicates the coverage at which on average 90% of the maximally detectable variants were recovered. Dashed black lines indicate the standard error. The threshold for a minimal mean coverage was set at 15X (red line). (C) Validation of insertions not present in the reference genome. (i) TE insertions that are not present in the reference genome show a duplication of the target site and the part of the reads that covers the TE will not be mapped against the reference genome. We thus expect reads to map to the TE surrounding region and the target site duplication but not the TE itself. At the target site, a local duplication of read depth is expected. (ii) We selected all reads in an interval of 100 bp up- and downstream including the target site duplication to detect deviations in the number of reads terminating near the target site duplication. (D) Validation of insertions present in the reference genome. (i) Analyses read coverage at target site duplications. (ii) Synthesis of evidence from ngs_te_mappr and split read mapping to determine TE presence or absence.

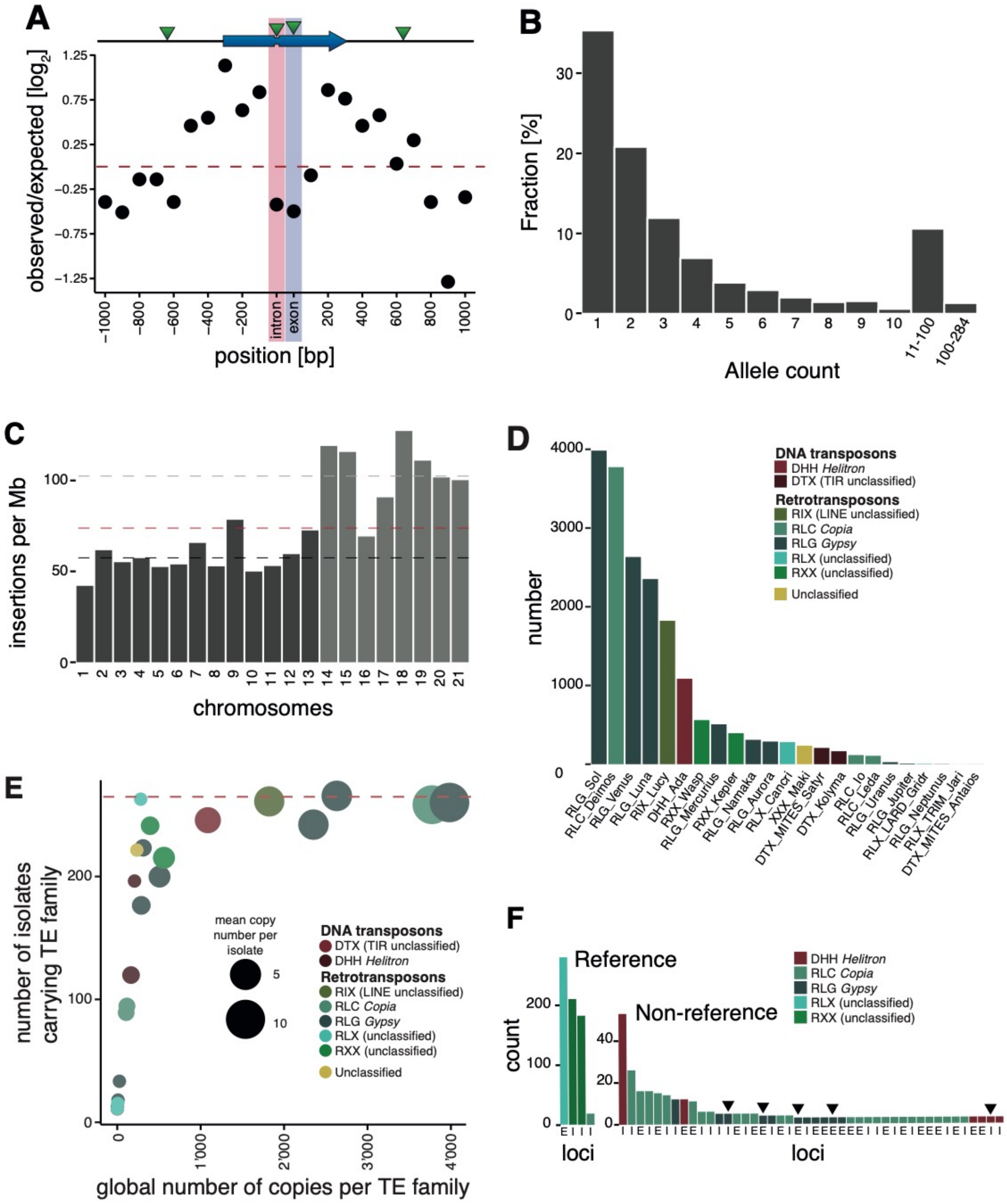


Figure 2: Transposable element (TE) landscape across populations. (A) Number of TE insertions 1 kb up- and downstream of genes on core chromosomes including introns and exons (100 bp windows). (B) Allele frequencies of the TE insertions across all isolates. (C) TE insertions per Mb on core chromosomes (dark) and accessory chromosomes (light). Dashed lines represent mean values. Red: global mean of 75.65 insertions/Mb, dark: core chromosome mean of 58 TEs/Mb, light: accessory chromosome mean of 102.24 insertions/Mb). (D) Number of TE insertions per family. (E) TE frequencies among isolates and copy numbers across the genome. The red line indicates the maximum number of isolates (n = 284). (F) TE insertions into introns and exons that are present in the reference genome and TEs absent from the reference genome but present in more than two copies in the populations. A hexagon indicates that the insertion was found in only one population, all other insertions were found in at least two populations. I = intron insertion, E = exon insertion.

TEs grouped into 23 families and 11 superfamilies, with most TEs belonging to class I/retrotransposons (*n* = 2175; Supplementary Figure S4A; Figure 2D). *Gypsy* (*n* = 1'483) and *Copia* (*n* = 623) elements constitute the largest long terminal repeats (LTR) superfamilies. Class II/DNA transposons are dominated by *Helitron* (*n* = 249). TE families shared among less isolates tend to show also lower copy numbers as expected (Figure 2E).

We detected 153 TE insertions into genes with most insertions being singletons (*n* = 68) or at very low frequency (Figure 2F). Overall, TE insertions into exonic sequences were less frequent than expected compared to insertions into up- and downstream regions, which is consistent with effective purifying selection (Figure 2A). Insertions into introns were also strongly under-represented, likely due to the small size of most fungal introns (~ 50-100 bp) and the high probability of disrupting splicing or adjacent coding sequences. We also found that insertions 800-1000 bp away from coding sequences of a focal gene were under-represented. Given the high gene density, with an average spacing between genes of 1,744 kb, TE insertions within 800-1000 bp of a coding gene tend to be near adjacent genes already. Taken together, TEs in the species show a high degree of transposition activity and are subject to strong purifying selection.

DETECTION OF CANDIDATE TE LOCI UNDERLYING RECENT ADAPTATION

The TE transposition activity can generate adaptive genetic variation. To identify the most likely candidate loci, we analyzed insertion frequency variation among populations as an indicator for recent selection. Across all populations, the insertion frequencies differed only weakly with a strong skew towards extremely low F_{ST} values (mean = 0.0163; Figure 3A, 3C). High F_{ST} loci tend to have high TE frequencies in either the North American population from 2015 or the Australian population. Given our population sampling, we tested for the emergence of adaptive TE insertions either in the North American or European population pairs. Hence, we selected loci having low TE insertion frequencies (< 5%) in all populations except either the recent North American or European population (> 20%) (Figure 3B). Based on these criteria, we obtained 26 candidate loci possibly underlying local adaptation in the North American populations with 22 loci showing retrotransposon insertions, three

Helitron, and one DNA TIR transposon. In parallel, we found six loci of retrotransposons possibly underlying local adaptation in the European populations (Figure 4A and Supplementary Table S4). To further analyze evidence for TE-mediated adaptive evolution, we screened the whole-genome sequencing datasets for evidence of selective sweeps using selection scans. Out of all 32 loci showing signatures of local adaptation in North American or European populations, we found five loci overlapping selective sweep regions. All TEs inserted in regions of selective sweeps are retrotransposons including *Copia* and *Gypsy* elements.

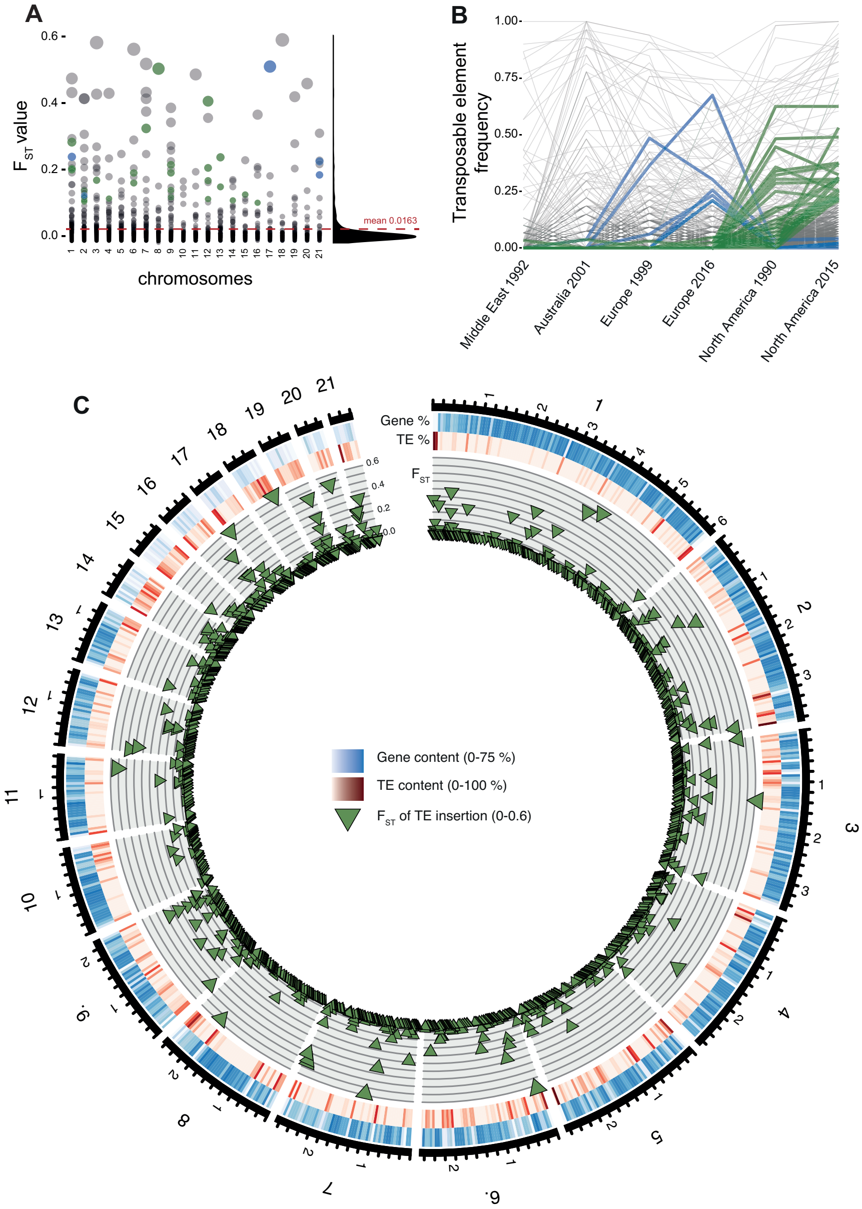


Figure 3: Differentiation in transposable element insertions frequencies across the genome. (A) Global pairwise F_{ST} distributions shown across the 21 chromosomes. The red horizontal line indicates the mean F_{ST} (=

0.0163). TEs with a strong local short-term frequency difference among populations are highlighted (blue: increase in Europe; green: increase in North America). (B) Allele frequency changes between the populations. The same TE loci as in panel A are highlighted. (C) Circos plot describing from the outside to the inside: The black line indicates chromosomal position in Mb. Blue bars indicate the gene density in windows of 100 kb with darker blue representing higher gene density. Red bars indicate the TE density in windows of 100 kb with a darker red representing higher TE density. Green triangles indicate positions of TE insertions with among population F_{ST} value shown on the y-axis.

We focused on five TE insertion loci in proximity to genes with a function likely associated with fungicide resistance or host adaptation. A TE insertion is 105 bp downstream of a major facilitator superfamily (MFS) transporter gene and 644 bp upstream of a TATA box (Figure 4B). MFS transporters can contribute to the detoxification of antifungal compounds in the species (Omrane *et al.*, 2017). The inserted *Helitron* TE was only found in North American populations (Figure 4G). The TE insertion occurred in a gene-rich, TE-poor region and the $F_{ST} = 0.51$ was one of the highest values of all TE loci (Figure 4C). Generally, the *Helitron* increased strongly in copy number from the Israel to the North American populations (Figure 4D, 4F). The phylogeny of the gene encoding the MFS showed a high degree of similarity for all isolates carrying the *Helitron* insertion compared to the isolates lacking the *Helitron* (Figure 4E). This is consistent with a rapid rise in frequency of the haplotype carrying the *Helitron* driven by positive selection. A second TE insertion that was only found in the two North American populations also contains a *Helitron* of the family Ada. The TE was inserted into an intron of a Phox domain-encoding gene (Supplementary Figure S8). Phox homologous domain proteins contribute to sorting membrane trafficking (Odorizzi *et al.*, 2000). A third potentially adaptive insertion of a *Copia* Deimos TE was 229 bp upstream of a gene encoding a SNARE domain protein and 286 bp upstream of a gene encoding a flavin amine oxidoreductase and located in a region of selective sweep (Supplementary Figure S9). SNARE domains play a role in vesicular transport and membrane fusion (Bonifacino & Glick, 2004). Additional strong candidates for adaptive TE insertions affected genes encoding a second MFS transporter and an effector candidate (Supplementary Figures 9 and 10). We experimentally tested whether the TE insertions in proximity to genes could contribute to higher levels of fungicide resistance. For this, we measured growth rates of the fungal isolates in the presence or absence of an azole fungicide widely deployed against the pathogen. We found that the insertion of TEs at three loci was positively associated with higher levels of fungicide resistance suggesting TE-mediated adaptations (Supplementary Figure S12).

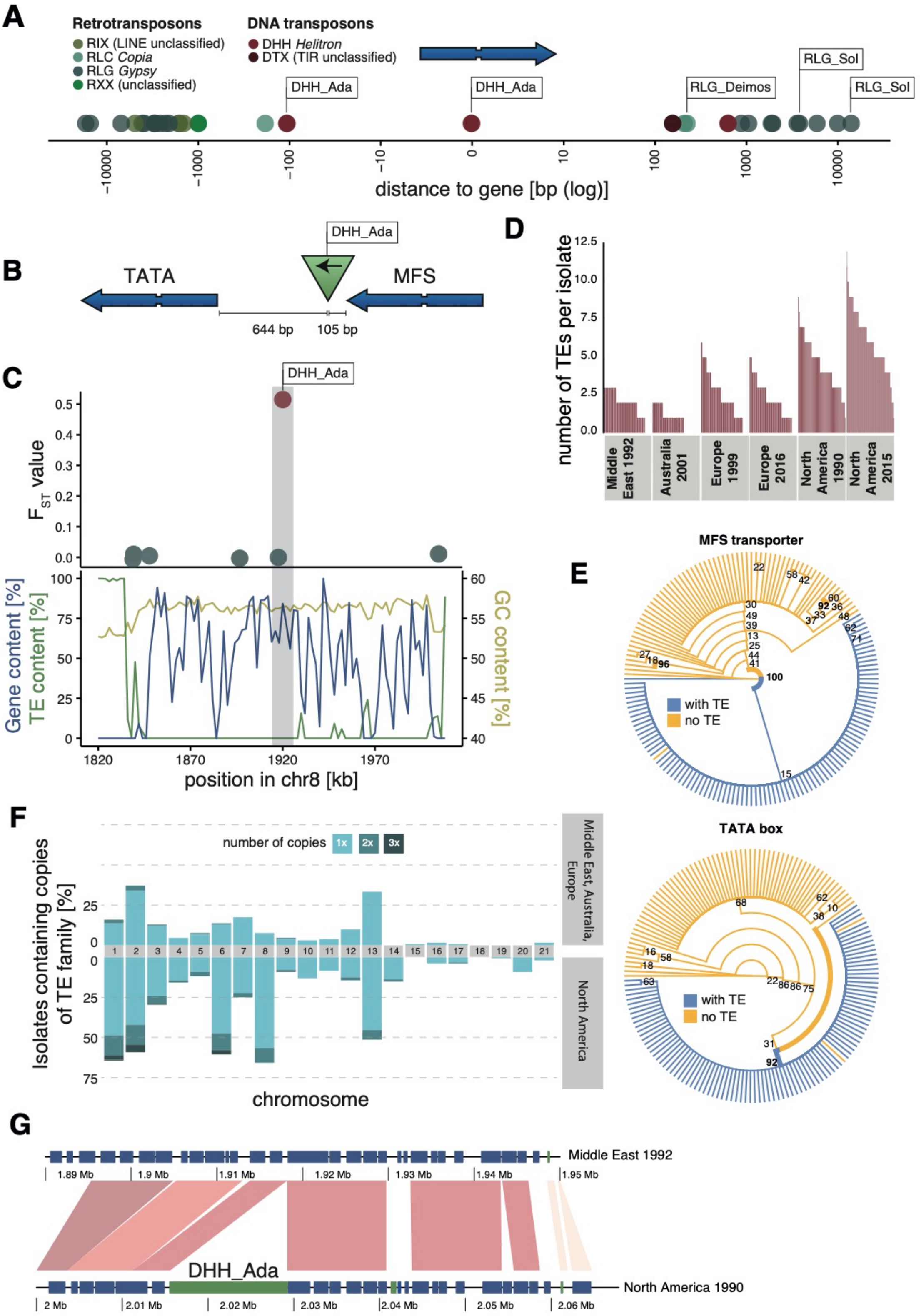


Figure 4: Candidate adaptive transposable element (TE) insertions. (A) Distribution of all extremely differentiated TEs and their distance to the closest gene. Color indicates the superfamily. TE sites potentially under selection according to F_{ST} are flagged. (B) Location of the *Helitron* Ada TE insertion on chromosome 8 corresponding to its two closest genes. (C) Genomic niche of the *Helitron* Ada TE insertion on chromosome 8: F_{ST} values for each TE insertion, gene content (blue), TE content (green) and GC content (yellow). The grey section highlights TE loci with extremely differentiated population frequencies. (D) Number of Ada copies per isolate and population. (E) Phylogenetic trees of the coding sequences of each the MFS transporter upstream and the TATA box downstream of the TE insertion. Isolates of the two North American populations and an additional 11 isolates from other populations not carrying the insertion are shown. Blue color indicates TE presence, yellow indicates TE absence. (F) Frequency changes of the TE family Ada between the two North American populations compared to the other populations. Colors indicate the number of copies per chromosome. (G) Synteny plot of the Ada insertion locus on chromosome 8 between two complete genomes from the Middle East (TE missing) and North America (TE present). Figures S8-S11 show additional candidate regions.

Population-level expansions in TE content

If TE insertion dynamics are largely neutral across populations, TE frequencies across loci should reflect neutral population structure. To test this, we performed a principal component analysis based on a set of six populations on four continents that represent the global genetic diversity of the pathogen and 900’193 genome-wide SNPs (Figure 5A-B). The population structure reflected the demographic history of the pathogen with clear continental differentiation and only minor within-site differentiation. In stark contrast, TE frequencies across loci showed only weak clustering by geographic origin with the Australian population being the most distinct (Figure 5D). We found a surprisingly strong differentiation of the two North American populations sampled at a 25-year interval in the same field in Oregon. To account for the lower number of TE loci, we performed an additional principal component analysis using a comparably sized SNP set to number of TE loci. Genome-wide SNPs retained the geographic signal of the broader set of SNPs (Figure 5C).

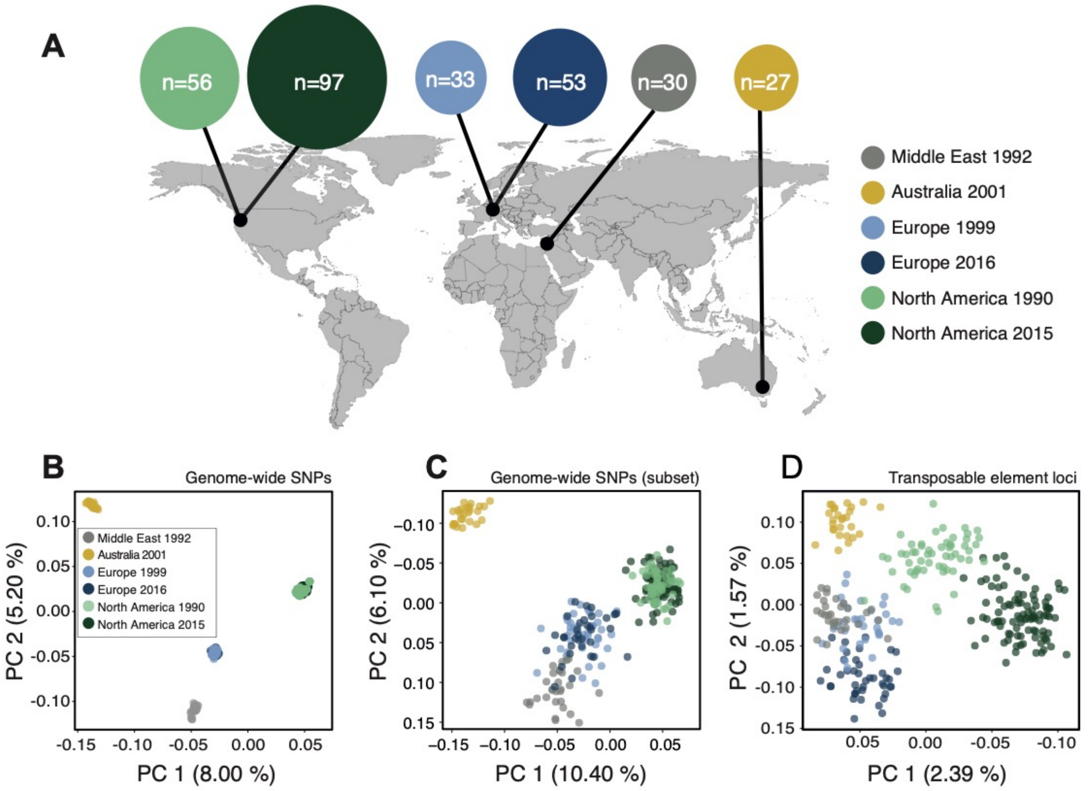


Figure 5: Population differentiation at transposable element (TE) and genome-wide SNP loci. (A) Sampling locations of the six populations. Middle East represents the region of origin of the pathogen. In North America, the two populations were collected at an interval of 25 years in the same field in Oregon. In Europe, two populations were collected at an interval of 17 years from two fields in Switzerland <20 km apart. Dark arrows indicate the historic colonization routes of the pathogen. (B) Principal component analysis (PCA) of 284 *Zymoseptoria tritici* isolates, based on 900'193 genome-wide SNPs. (C) PCA of a reduced SNP data set with randomly selected 203 SNPs matching approximately the number of analyzed TE loci. (D) PCA based on 193 TE insertion loci. Loci with allele frequency < 5% are excluded.

Unusual patterns in population differentiation at TE loci suggests that TE activity may substantially vary across populations (Figure 6). To analyze this, we first identified the total TE content across all loci per isolate. We found generally lower TE numbers in the Middle Eastern population from Israel (Figure 6B), which is close to the pathogen's center of origin (Stukenbrock *et al.*, 2007). Populations that underwent at least one migration bottleneck showed a substantial burst of TEs across all major superfamilies. These populations included the two populations from Europe collected in 1999 and 2016 and the North American population from 1990, as well as the Australian population. We found a second stark increase in TE content in the North American population sampled in 2015 at the same site as the population from 1990. Strikingly, the isolate with the lowest number of analyzed TEs collected in 2015 was comparable to the isolate with the highest number of TEs at the same site in

1990. We tested whether sequencing coverage could explain variation in the detected TEs across isolates, but we found no meaningful association (Supplementary Figure S4B). We analyzed variation in TE copy numbers across families and found that the expansions were mostly driven by *Gypsy* elements including the families Luna, Sol and Venus, the *Copia* family Deimos and the LINE family Lucy (Figure 6C; Supplementary Figures S5-6). We also found a North American specific burst in *Helitron* elements (Ada), an increase specific to Swiss populations in LINE elements, and an increase in *Copia* elements in the Australian and the two North American populations. Analyses of complete *Z. tritici* genomes from the same populations revealed high TE contents in Australia and North America (Oregon 1990) (Badet *et al.*, 2020). The complete genomes confirmed also that the increase in TEs was driven by LINE, *Gypsy* and *Copia* families in Australia and *Helitron*, *Gypsy* and *Copia* families in North America (Badet *et al.*, 2020).

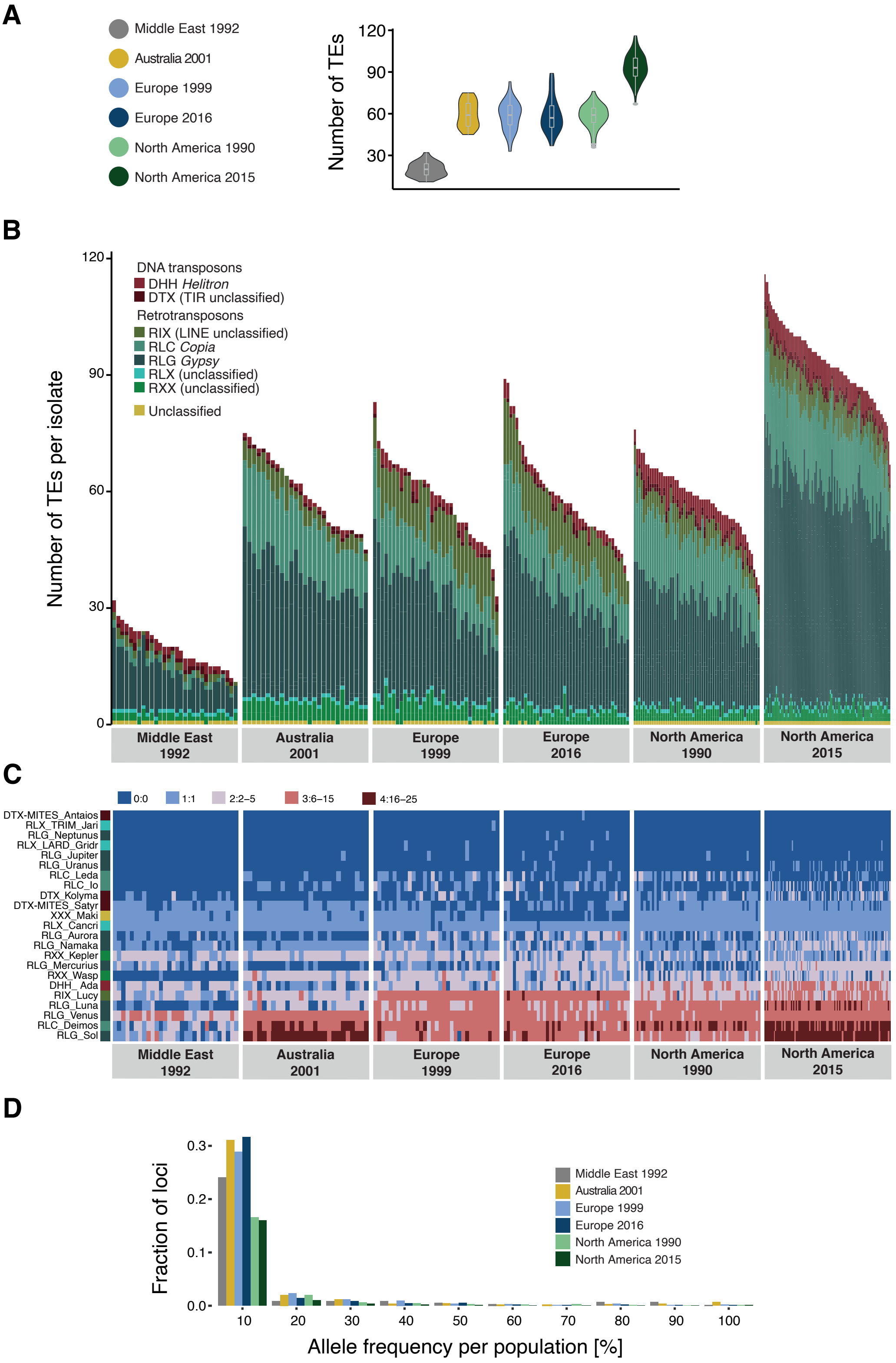


Figure 6: Global population structure of transposable element (TE) insertion polymorphism. (A) The number of transposable elements (TEs) per population. (B) Total TE copies per isolate. Colors identify TE superfamilies. (C) TE family copy numbers per isolate. (D) TE insertion frequency spectrum per population.

Finally, we analyzed whether the population-specific expansions were correlated with shifts in the frequency spectrum of TEs in the populations (Figure 6D). We found that the first step of expansions observed in Europe was associated with a downwards shift in allele frequencies. This is consistent with transposition activity creating new copies in the genomes and stronger purifying selection. In contrast, the North American populations showed an upwards shift in allele frequencies indicating relaxation of selection against TEs.

TE-MEDIATED GENOME SIZE EXPANSIONS

The combined effects of actively copying TE families and relaxed purifying selection leads to an accumulation of new TE insertions in populations. As a consequence, mean genome sizes in populations should increase over generations. To test for incipient genome expansions within the species, we first assembled genomes of all 284 isolates included in the study. Given the limitations of short-read assemblies, we implemented corrective measures to compensate for potential variation in assembly qualities. We corrected for variation in the GC content of different sequencing datasets by downsampling reads to generate balanced sequencing read sets prior to assembly (see Methods). We also excluded all reads mapping to accessory chromosomes because different isolates are known to differ in the number of these chromosomes. Genome assemblies were checked for completeness by retrieving the phylogenetically conserved BUSCO genes (Figure 7A). Genome assemblies across different populations carry generally >99% complete BUSCO gene sets, matching the completeness of fully finished genomes of the same species (Badet *et al.*, 2020). The completeness of the assemblies showed no correlation with either TE or GC content of the genomes. GC content was inversely correlated with genome size consistent with the expansion of repetitive regions having generally low GC content (Figure 7B). We found that the core genome size varied substantially among populations with the Middle East, Australia as well as the two older European and North American populations having the smallest genomes (Figure 7C and 7D). We found a notable increase in genome size in both the more recent European and North American populations. The increase in genome size is positively correlated with the count of TE insertions (Figure 7E and G) and negatively correlated with the genome-wide GC content (Figure 7F and G). Hence, genome size shows substantial variation within the species matching the recent expansion in TEs across continents.

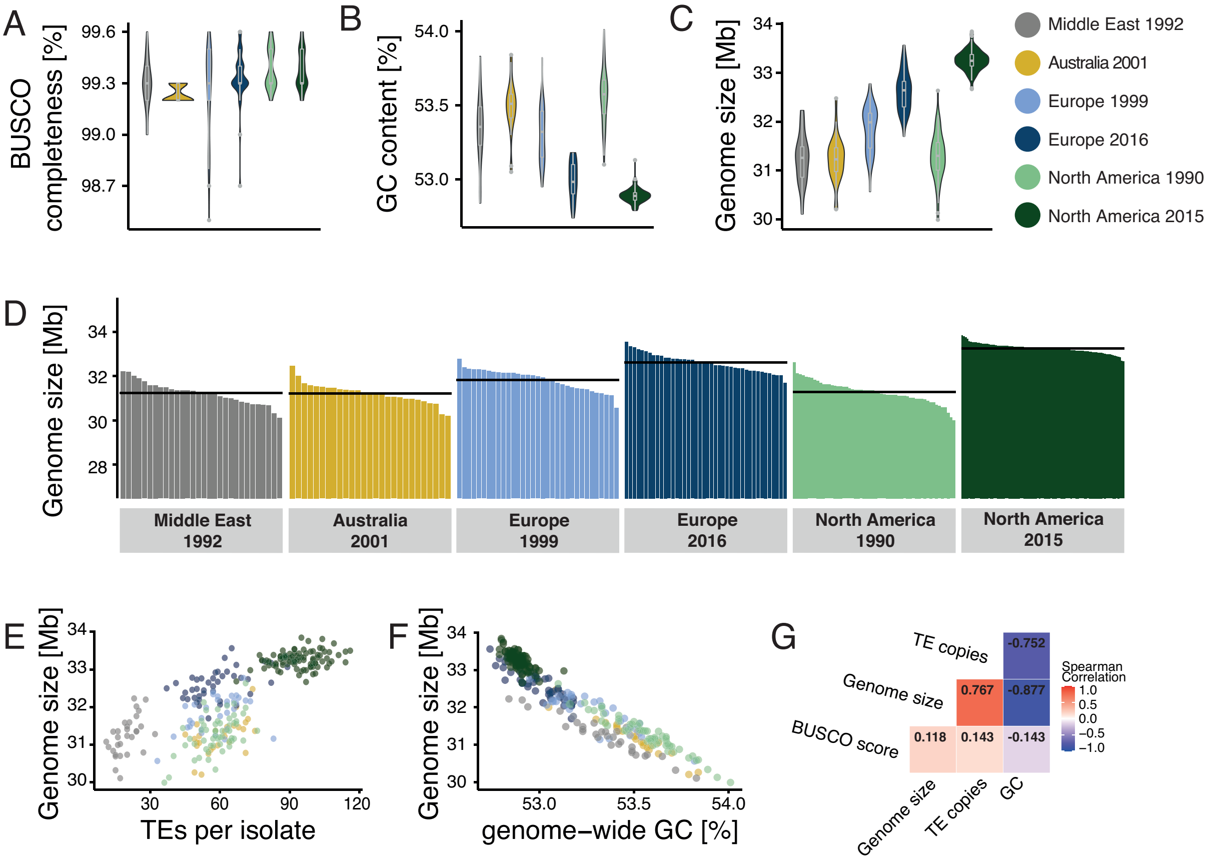


Figure 7: Genome size and transposable element (TE) evolution across populations. (A) BUSCO completeness variation among genome assemblies. Black lines indicate the mean genome size per population. (B) Genome-wide GC content variation. (C) Core genome sizes (excluding accessory chromosomes). (D) Genome size variation among population. (E) Correlation of core genome size and number of detected TEs. (F) Correlation of core genome size and genome-wide GC content. (G) Spearman correlation matrix of BUSCO completeness, core genome size, number of detected TEs and genome-wide GC content.

Discussion

TEs play a crucial role in generating adaptive genetic variation within species but are also drivers of deleterious genome expansions. We analyzed the interplay of TEs with selective and neutral processes including population differentiation and incipient genome expansions. TEs have substantial transposition activity in the genome but are strongly counter-selected and are maintained at low frequency. TE dynamics showed distinct trajectories across populations with more recently established populations having higher TE content and a concurrent expansion of the genome.

RECENT SELECTION ACTING ON TE INSERTIONS

TE frequencies in the species show a strong skew towards singleton insertions across populations. This indicates both that TEs are undergoing transposition and that purifying selection maintains frequencies at a low level. Similar effects of selection on active TEs were observed across plants and animals, including *Drosophila melanogaster* and *Brachypodium distachyon* (Cridland *et al.*, 2013; Stritt *et al.*, 2017; Luo *et al.*, 2020). TE insertions were under-represented in or near coding regions, showing a stronger purifying selection against TEs inserting into genes. Coding sequences in the *Z. tritici* genome are densely packed with an average distance of only ~1 kb (Goodwin *et al.*, 2011). Consistent with this high gene density, TE insertions were most frequent at a distance of 200-400 bp away from coding sequences. A rapid decay in linkage disequilibrium in the *Z. tritici* populations (Croll *et al.*, 2015; Hartmann *et al.*, 2018) likely contributed to the efficiency of removing deleterious insertions. We also found evidence for positive selection acting on TEs with the strongest candidate loci being two TE insertions near genes encoding MFS transporters. Both loci showed a frequency increase only in the North American populations, which experienced the first systematic fungicide applications and subsequent emergence of fungicide resistance in the decade prior to the last sampling (Estep *et al.*, 2015). TE-mediated overexpression of a MFS1 transporter is a known resistance mechanism of *Z. tritici* and acts by increasing efflux of fungicides out of the cell (Omrane *et al.*, 2017). TE-mediated fungicide resistance adaptation in the North American population is further supported by a significant association of levels of fungicide resistance in the population and the presence of the *Gypsy* insertion near the MFS gene. Furthermore, the locus experienced a selective sweep following the insertion of the TE.

Transposition activity in a genome and counter-acting purifying selection are expected to establish an equilibrium over evolutionary time (Charlesworth & Charlesworth, 1983). However, temporal bursts of TE families and changes in population size due to bottlenecks or founder events are likely to shift the equilibrium. Despite purifying selection, we were able to detect signatures of positive selection by scanning for short-term population frequency shifts. Population genomic datasets can be used to identify the most likely candidate loci underlying recent adaptation. The shallow genome-wide differentiation of *Z. tritici* populations provides a powerful background to test for outlier loci (Hartmann *et al.*, 2018). We found the same TE families to have experienced genome-wide copy number expansions, suggesting that the availability of adaptive TE insertions may be a by-product of TE bursts in individual populations.

Population-level TE invasions and relaxed selection

Across the surveyed populations from four continents, we identified substantial variation in TE counts per genome. The increase in TEs matches the global colonization history of the pathogen with an increase in TE copies in more recently established populations (Zhan *et al.*, 2003; Stukenbrock *et al.*, 2007). Compared to the Israeli population located nearest the center of origin in the Middle East, the European populations showed a three-fold increase in TE counts. The Australian and North American populations established from European descendants retained high TE counts. We identified a second increase at the North American site where TE counts nearly doubled again over a 25-year period. Compared to the broader increase in TEs from the Middle East, the second expansion at the North American site was driven by a small subset of TE families alone. Analyses of completely assembled genomes from the same populations confirmed that genome expansions were primarily driven by the same TE families belonging to *Gypsy*, *Copia* and *Helitron* superfamilies (Badet *et al.*, 2020). Consistent with the contributions from individual TEs, we found that the first expansion in Europe led to an increase in low-frequency variants, suggesting higher transposition activity of many TEs in conjunction with strong purifying selection. The second expansion at the North American site shifted TE frequencies upwards, suggesting relaxed selection against TEs. The population-level context of TEs in *Z. tritici* shows how heterogeneity in TE control interacts with demography to determine extant levels of TE content and, ultimately, genome size.

TE invasion dynamics underpins genome size expansions

The number of detected TEs was closely correlated with core genome size, hence genome size expansions were at least partly caused by the very recent proliferation of TEs. Genome assemblies of large eukaryotic genomes based on short read sequencing are often fragmented and contain chimeric sequences (Nagarajan & Pop, 2013). Focusing on the less repetitive core chromosomes in the genome of *Z. tritici* reduces such artefacts substantially. Because genome assemblies are the least complete in the most repetitive regions, any underrepresented sequences may rather underestimate than overestimate within-species variation in genome size. Hence, we consider the assembly sizes to be a robust correlate of total genome size. The core genome size differences observed across the species range match genome size variation typically observed among closely related species. Among primates, genome size varies by ~70% with ~10% between humans and chimpanzees (Rogers & Gibbs, 2014; Miga *et al.*, 2020). In fungi, genome size varies by several orders of magnitude within phyla but is often highly similar among closely related species (Raffaele & Kamoun, 2012). Interestingly, drastic changes in genome size have been observed in the *Blumeria* and *Pseudocercospora* genera where genome size changed by 35-130% between the closest known species (González-Sayer *et al.*; Frantzeskakis *et al.*, 2018). Beyond analyses of TE content variation correlating with genome size evolution, proximate mechanisms driving genome expansions are poorly understood. Establishing large population genetic datasets such as it is possible for crop pathogens, genome size evolution becomes tractable at the population level.

The activity of TEs is controlled by complex selection regimes within species. Actively transposing elements may accelerate genome evolution and underpin expansions. Hence, genomic defenses should evolve to efficiently target recently active TEs. Here, we show that TE activity and counteracting genomic defenses have established a tenuous equilibrium across the species range. We show that population subdivisions are at the origin of highly differentiated TE content within a species matching genome size changes emerging over the span of only decades and centuries. In conclusion, population-level analyses of genome size can recapitulate genome expansions typically observed across much deeper time scales providing fundamentally new insights into genome evolution.

METHODS

FUNGAL ISOLATE COLLECTION AND SEQUENCING

We analyzed 295 *Z. tritici* isolates covering six populations originating from four geographic locations and four continents (Supplementary Table S1), including: Middle East 1992 (*n* = 30 isolates, Nahal Oz, Israel), Australia 2001 (*n* = 27, Wagga Wagga), Europe 1999 (*n* = 33, Berg am Irchel, Switzerland), Europe 2016 (*n* = 52, Eschikon, ca. 15km from Berg am Irchel, Switzerland), North America 1990 and 2015 (*n* = 56 and *n* = 97, Willamette Valley, Oregon, United States) (McDonald *et al.*, 1996; Linde *et al.*, 2002; Zhan *et al.*, 2002, 2003, 2005). Illumina short read data from the Middle East, Australia, European 1999 and North American 1990 populations were obtained from the NCBI Short Read Archive under the BioProject PRJNA327615 (Hartmann *et al.*, 2017). For, the Switzerland 2016 and Oregon 2015 populations, asexual spores were harvested from infected wheat leaves from naturally infected fields and grown in YSB liquid media including 50 mgL^{-1} kanamycin and stored in silica gel at −80°C. High-quality genomic DNA was extracted from liquid cultures using the DNeasy Plant Mini Kit from Qiagen (Venlo, Netherlands). The isolates were sequenced on an Illumina HiSeq in paired-end mode and raw reads were deposited on the NCBI Short Read Archive under the BioProject PRJNA596434.

TE INSERTION DETECTION

The quality of Illumina short reads was determined with FastQC version 0.11.5 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) (Figure 1A). To remove spuriously sequenced Illumina adaptors and low quality reads, we trimmed the sequences with Trimmomatic version 0.36, using the following filter parameters: illuminaclip:TruSeq3-PE-2.fa:2:30:10 leading:10 trailing:10 slidingwindow:5:10 minlen:50 (Bolger *et al.*, 2014). We created repeat consensus sequences for TE families (sequences are available on https://github.com/crolllab/datasets) in the complete reference genome IPO323 (Goodwin *et al.*, 2011) with RepeatModeler version open-4.0.7 (http://www.repeatmasker.org/RepeatModeler/) based on the RepBase Sequence Database and de novo (Bao *et al.*, 2015). TE classification into superfamilies and families was based on an approach combining detection of conserved protein sequences and tools to detect non-autonomous TEs (Badet *et al.*, 2020). To detect TE insertions, we used the R-based tool ngs_te_mapper version 79ef861f1d52cdd08eb2d51f145223fad0b2363c integrated into the McClintock pipeline version 20cb912497394fabddcdaa175402adacf5130bd1, using bwa version 0.7.4-r385 to map Illumina short reads, samtools version 0.1.19 to convert alignment file formats and R version 3.2.3 (Li & Durbin, 2009; Li *et al.*, 2009; Linheiro & Bergman, 2012; R Core Team, 2017; Nelson *et al.*, 2017).

Down-sampling analysis

We performed a down-sampling analysis to estimate the sensitivity of the TE detection with ngs_te_mapper based on variation in read depth. We selected one isolate per population matching the average coverage of the population. We extracted the per-base pair read depth with the genomecov function of bedtools version 2.27.1 and calculated the genome-wide mean read depth (Quinlan & Hall, 2010). The number of reads in the original fastq file was reduced in steps of 10% to simulate the impact of reduced coverage. We analyzed each of the obtained reduced read subsets with ngs_te_mapper using the same parameters as described above. The correlation between the number of detected insertions and the read depth was visualized using the function nls with model SSlogis in R and visualized with ggplot2 (Wickham, 2016). The number of detected TEs increased with the number of reads until reaching a plateau indicating saturation (Figure 1B). Saturation was reached at a coverage of approximately 15X, hence we retained only isolates with an average read depth above 15X for further analyses. We thus excluded one isolate from the Oregon 2015 population and ten isolates from the Switzerland 2016 population.

Validation procedure for predicted TE insertions

ngs_te_mapper detects the presence but not the absence of a TE at any given locus. We devised additional validation steps to ascertain both the presence as well as the absence of a TE across all loci in all individuals. TEs absent in the reference genome were validated by re-analyzing mapped Illumina reads. Reads spanning both parts of a TE sequence and an adjacent chromosomal sequence should only map to the reference genome sequence and cover the target site duplication (TSD) of the TE (Figure 1C). We used bowtie2 version 2.3.0 with the parameter --very-sensitive-local to map Illumina short reads of each isolate on the reference genome IPO323 (Langmead & Salzberg, 2012). Mapped Illumina short reads were then sorted and indexed with samtools and the resulting bam file was converted to a bed file with the function bamtobed in bedtools. We extracted all mapped reads with an end point located within 100 bp of the TSD (Figure 1C). We tested whether the number of reads with a mapped end around the TSD significantly deviated if the mapping ended exactly at the boundary. A mapped read ending exactly at the TSD boundary is indicative of a split read mapping to a TE sequence not present in the reference genome. To test for the deviation in the number of read mappings around the TSD, we used a Poisson distribution and the *ppois* function in R version 3.5.1 (Figure 1C). We identified a TE as present in an isolate if tests on either side of the TSD had a *p*-value < 0.001 (Supplementary Table S1, S2, Figure S1B).

For TEs present in the reference genome, we analyzed evidence for spliced junction reads spanning the region containing the TE. Spliced reads are indicative of a discontinuous sequence and, hence, absence of the TE in a particular isolate (Figure 1D). We used STAR version 2.5.3a to detect spliced junction reads with the following set of parameters: --runThreadN 1 --outFilterMultimapNmax 100 --winAnchorMultimapNmax 200 --outSAMmultNmax 100 --outSAMtype BAM Unsorted --outFilterMismatchNmax 5 --alignIntronMin 150 --alignIntronMax 15000 (Dobin *et al.*, 2012). We then sorted and indexed the resulting bam file with samtools and converted split junction reads with the function bam2hints in bamtools version 2.5.1 (Barnett *et al.*, 2011). We selected loci without overlapping spliced junction reads using the function intersect in bedtools with the parameter -loj -v. We considered a TE as truly absent in an isolate if ngs_te_mapper did not detect a TE and evidence for spliced junction reads were found. If the absence of a TE could not be confirmed by spliced junction reads, we labelled the genotype as missing. Finally, we excluded TE loci with more than 20% missing data from further investigations (Figure 1D and Supplementary Figure S1C).

Clustering of TE insertions into loci

We identified insertions across isolates as being the same locus if all detected TEs belonged to the same TE family and insertion sites differed by ≤100 bp (Supplementary Figure S2). We used the R package *GenomicRanges* version 1.28.6 with the functions makeGRangesFromDataFrame and findOverlaps and the R package *devtools* version 1.13.4 (Lawrence *et al.*, 2013; Wickham & Chang, 2016). We used the R package *dplyr* version 0.7.4 to summarize datasets (https://dplyr.tidyverse.org/). Population-specific frequencies of insertions were calculated with the function allele.count in the R package *hierfstat* version 0.4.22 (Goudet, 2005). We conducted a principal component analysis for TE insertion frequencies filtering for a minor allele frequency ≥ 5%. We also performed a principal component analysis for genome-wide single nucleotide polymorphism (SNP) data obtained from Hartmann et al (2017). As described previously, SNPs were hard-filtered with VariantFiltration and SelectVariants tools integrated in the Genome Analysis Toolkit (GATK) (McKenna *et al.*, 2010). SNPs were removed if any of the following filter conditions applied: QUAL<250; QD<20.0; MQ<30.0; -2 > BaseQRankSum > 2; -2 > MQRankSum > 2; -2 > ReadPosRankSum > 2; FS>0.1. SNPs were excluded with vcftools version 0.1.17 and plink version 1.9 requiring a genotyping rate >90% and a minor allele frequency >5% (https://www.cog-genomics.org/plink2, Chang et al., 2015). Finally, we converted tri-allelic SNPs to bi-allelic SNPs by recoding the least frequent allele as a missing genotype. Principal component analysis was performed using the *gdsfmt* and *SNPRelate* packages in R (Zheng *et al.*, 2012, 2017). For a second principal component analysis with a reduced set of random markers, we randomly selected SNPs with vcftools and the following set of parameters: --maf 0.05 –thin 200’000 to obtain an approximately equivalent number of SNPs as TE loci.

Genomic location of TE insertions

To characterize the genomic environment of TE insertion loci, we split the reference genome into non-overlapping windows of 10 kb using the function splitter from EMBOSS version 6.6.0 (Rice *et al.*, 2000). TEs were located in the reference genome using RepeatMasker providing consensus sequences from RepeatModeler (http://www.repeatmasker.org/). To analyze coding sequence, we retrieved the gene annotation for the reference genome (Grandaubert *et al.*, 2015). We estimated the percentage covered by genes or TEs per window using the function intersect in bedtools. Additionally, we calculated the GC content using the tool get_gc_content (https://github.com/spundhir/RNA-Seq/blob/master/get_gc_content.pl). We also extracted the number of TEs present in 1 kb windows up- and downstream of each annotated gene with the function window in bedtools with the parameters -l 1000 -r 1000 and calculated the relative distances with the closest function in bedtools. For the TEs inserted into genes, we used the intersect function in bedtools to distinguish intron and exon insertions with the parameters -wo and -v, respectively. For each 100 bp segment in the 1kb windows as well as for introns and exons, we calculated the mean number of observed TE insertions per base pair.

POPULATION DIFFERENTIATION IN TE FREQUENCIES

We calculated Nei's fixation index (F_{ST}) between pairs of populations using the R packages *hierfstat* and *adegenet* version 2.1.0 (Jombart, 2008; Jombart & Ahmed, 2011). To understand the chromosomal context of TE insertion loci across isolates, we analyzed draft genome assemblies. We generated *de novo* genome assemblies for all isolates using SPAdes version 3.5.0 with the parameter --careful and a kmer range of "21, 29, 37, 45, 53, 61, 79, 87" (Bankevich *et al.*, 2012). We used blastn to locate genes adjacent to TE insertion loci on genomic scaffolds of each isolate. We then extracted scaffold sequences surrounding 10 kb up- and downstream of the localized gene with the function faidx in samtools and reverse complemented the sequence if needed. Then, we performed multiple sequence alignments for each locus across all isolates with MAFFT version 7.407 with parameter --maxiterate 1000 (Katoh & Standley, 2013). We performed visual inspections to ensure correct alignments across isolates using Jalview version 2.10.5 (Waterhouse *et al.*, 2009). To generate phylogenetic trees of individual gene or TE loci, we extracted specific sections of the alignment using the function extractalign in EMBOSS and converted the multiple sequence alignment into PHYLIP format with jmodeltest version 2.1.10 using the -getPhylip parameter. We then estimated maximum likelihood phylogenetic trees with the software PhyML version 3.0, the K80 substitution model and 100 bootstraps on the ATGC South of France bioinformatics platform (Guindon & Gascuel, 2003; Guindon *et al.*, 2010; Darriba *et al.*, 2012). Bifurcations with a supporting value lower than 10% were collapsed in TreeGraph version 2.15.0-887 beta and trees were visualized as circular phylograms in Dendroscope version 2.7.4 (Huson *et al.*, 2007; Stöver & Müller, 2010). For loci showing complex rearrangements, we generated synteny plots using 19 completely sequenced genomes from the same species using the R package *genoplotR* version 0.8.9 (Guy *et al.*, 2010; Badet *et al.*, 2020).

We analyzed signatures of selective sweeps using the extended haplotype homozygosity (EHH) tests (Sabeti *et al.*, 2007) implemented in the R package REHH (Gautier & Vitalis, 2012). We analyzed within-population signatures based on the iHS statistic and chose a maximum gap distance of 20 kb. We also analyzed cross-population EHH (XP-EHH) signatures testing the following two population pairs: North America 1990 versus North America 2015, Europe 1999 versus Europe 2016. We defined significant selective sweeps as being among the 99.9th percentile outliers of the iHS and XP-EHH statistics. Significant SNPs at less than 5 kb were clustered into a single selective sweep region adding +/- 2.5 kb. Finally, we analyzed whether TE loci were within 10 kb of a region identified as a selective sweep using the function intersect from bedtools.

GENOME SIZE ESTIMATION

Accessory chromosomes show presence/absence variation within the species and length polymorphism (Goodwin *et al.*, 2011; Croll *et al.*, 2013) and thus impact genome size. We controlled for this effect by first mapping sequencing reads to the reference genome IPO323 using bowtie2 with --very-sensitive-local settings and retained only reads mapping to any of the 13 core chromosomes using seqtk subseq v1.3-r106 (https://github.com/lh3/seqtk/). Furthermore, we found that different sequencing runs showed minor variation in the distribution of the per read GC content. In particular, reads of a GC content lower than 30 % were underrepresented in the Australian (mean reads < 30 % of the total readset: 0.05 %), North American 1990 (0.07 %) and Middle East (0.1 %) populations, and higher in the Europe 1999 (1.3 %), North American 2015 (3.0 %) and Europe 2016 (4.02 %) populations (Supplementary Figure S3). Library preparation protocols and Illumina sequencer generations are known factors influencing the recovery of reads of varying GC content (Benjamini & Speed, 2012).

To control a potential bias stemming from this, we subsampled reads based on GC content to create homogeneous datasets. For this, we first retrieved the mean GC content for each read pair using geecee in EMBOSS and binned reads according to GC content. For the bins with a GC content <30%, we calculated the mean proportion of reads from the genome over all samples. We then used seqtk subseq to subsample reads of <30% to adjust the mean GC content among readsets. We generated *de novo* genome assemblies using the SPAdes assembler version with the parameters --careful and a kmer range of "21, 29, 37, 45, 53, 61, 79, 87". The SPAdes assembler is optimized for the assembly of relatively small eukaryotic genomes. We evaluated the completeness of the assemblies using BUSCO v4.1.1 with the fungi_odb10 gene test set (Simão *et al.*, 2015). We finally ran Quast v5.0.2 to retrieve assembly metrics including scaffolds of at least 1kb (Mikheenko *et al.*, 2018).

FUNGICIDE RESISTANCE ASSAY

To quantify susceptibility towards propiconazole we performed a microtiter plate assay. Isolates were grown on yeast malt sucrose agar for five days and spores were harvested. We then tested for growth inhibition by growing spores (2.5 x 10^4 spores/ml) in Sabouraud-dextrose liquid medium with differing concentrations of propiconazole (0.00006, 0.00017, 0.0051, 0.0086, 0.015, 0.025, 0.042, 0.072, 0.20, 0.55, 1.5 mg/L). We incubated the plates stationary in the dark at 21°C and 80% relative humidity for four days and measured optical density at 605 nm. We calculated EC_{50} with the R package *drc* (Ritz & Streibig, 2005).

Data availability

Sequence data is deposited on the NCBI Short Read Archive under the accession numbers PRJNA327615, PRJNA596434 and PRJNA178194. Transposable element consensus sequences are available from https://github.com/crolllab/datasets.

Author contributions

UO and DC conceived the study, UO, TW and DC designed analyses, UO, TB, TV and FEH performed analyses, FEH, NKS, LNA, PK, CCM and BAM provided samples/datasets, BAM and DC provided funding, UO and DC wrote the manuscript with input from co-authors. All authors reviewed the manuscript and agreed on submission.

Acknowledgments

We thank Andrea Sánchez Vallet, Anne C. Roulin, Luzia Stalder, Adam Taranto, Emilie Chanclud and Alice Feurtey for helpful discussions and comments on previous versions of the manuscript. DC is supported by the Swiss National Science (grants 31003A_173265 and IZCOZO_177052) and the Fondation Pierre Mercier pour la Science.

Competing interests

We declare to have no competing interests.

References

Badet T, Oggenfuss U, Abraham L, McDonald BA, Croll D. **2020**. A 19-isolate reference-quality global pangenome for the fungal wheat pathogen Zymoseptoria tritici. *BMC Biology* **18**: 12.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, *et al.* **2012**. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology : a journal of computational molecular cell biology* **19**: 455–77.

Bao W, Kojima KK, Kohany O. **2015**. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**: 4–9.

Barnett DW, Garrison EK, Quinlan AR, Strï̈mberg MP, Marth GT. **2011**. Bamtools: A C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* **27**: 1691–1692.

Baucom RS, Estill JC, Leebens-Mack J, Bennetzen JL. **2008**. Natural selection on gene function drives the evolution of LTR retrotransposon families in the rice genome. *Genome Research* **19**: 243–254.

Benjamini Y, Speed TP. **2012**. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research* **40**: 1–14.

Bolger AM, Lohse M, Usadel B. **2014**. Trimmomatic: a flexible trimmer for Illumina sequence data.

Bioinformatics **30**: 2114–2120.

Bonifacino JS, Glick BS. **2004**. The Mechanisms of Vesicle Budding and Fusion. *Cell* **116**: 153–166.

Charlesworth B, Charlesworth D. **1983**. The population dynamics of transposable elements. *Genetical Research* **42**: 1–27.

Chuong EB, Elde NC, Feschotte C. **2017**. Regulatory activities of transposable elements: from conflicts to benefits. *Nature Reviews Genetics* **18**: 71–86.

Cridland JM, Macdonald SJ, Long AD, Thornton KR. **2013**. Abundance and distribution of transposable elements in two drosophila QTL mapping resources. *Molecular Biology and Evolution* **30**: 2311–2327.

Croll D, Lendenmann MH, Stewart E, McDonald BA. **2015**. The Impact of Recombination Hotspots on Genome Evolution of a Fungal Plant Pathogen. *Genetics* **201**: 1213-U787.

Croll D, Zala M, McDonald BA. **2013**. Breakage-fusion-bridge Cycles and Large Insertions Contribute to the Rapid Evolution of Accessory Chromosomes in a Fungal Pathogen (J Heitman, Ed.). *PLOS Genetics* **9**: e1003567.

Darriba D, Taboada GL, Doallo R, Posada D. **2012**. jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods* **9**: 772.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Gingeras TR, Batut P, Chaisson M. **2012**. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21.

Eichler EE, Sankoff D. **2003**. Structural dynamics of eukaryotic chromosome evolution. *Science* **301**: 793–797.

Estep LK, Torriani SFF, Zala M, Anderson NP, Flowers MD, Mcdonald BA, Mundt CC, Brunner PC. **2015**. Emergence and early evolution of fungicide resistance in North American populations of Zymoseptoria tritici. *Plant Pathology* **64**: 961–971.

Feschotte C. **2008**. Transposable elements and the evolution of regulatory networks. *Nature Reviews Genetics* **9**: 397–405.

Feurtey A, Lorrain C, Croll D, Eschenbrenner C, Freitag M, Habig M, Haueisen J, Möller M, Schotanus K, Stukenbrock EH. **2020**. Genome compartmentalization predates species divergence in the plant pathogen genus Zymoseptoria. *BMC genomics* **21**: 588.

Fouché S, Badet T, Oggenfuss U, Plissonneau C, Francisco CS, Croll D. **2019**. Stress-driven transposable element de-repression dynamics in a fungal pathogen. *Molecular Biology and Evolution*.

Frantzeskakis L, Kracher B, Kusch S, Yoshikawa-Maekawa M, Bauer S, Pedersen C, Spanu PD, Maekawa T, Schulze-Lefert P, Panstruga R. **2018**. Signatures of host specialization and a recent transposable element burst in the dynamic one-speed genome of the fungal barley powdery mildew pathogen. *BMC Genomics* **19**: 1–23.

Gautier M, Vitalis R. **2012**. Rehh An R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics* **28**: 1176–1177.

González-Sayer S, Oggenfuss U, García I, Aristizabal F. High-quality genome assembly of Pseudocercospora ulei the main threat to natural rubber trees. : 0–1.

Goodwin SB, Ben M'Barek S, Dhillon B, Wittenberg AHJ, Crane CF, Hane JK, Foster AJ, Van der Lee TAJ, Grimwood J, Aerts A, *et al.* **2011**. Finished Genome of the Fungal Wheat Pathogen Mycosphaerella graminicola Reveals Dispensome Structure, Chromosome Plasticity, and Stealth Pathogenesis (HS Malik, Ed.). *PLOS Genetics* **7**: e1002070.

Goudet J. **2005**. Hierstat, a package for R to compute and test heirarchical F-statistics. *Molecular*

Ecology Notes **5**: 184–186.

Grandaubert J, Bhattacharyya A, Stukenbrock EH. **2015**. RNA-seq-Based Gene Annotation and Comparative Genomics of Four Fungal Grass Pathogens in the Genus Zymoseptoria Identify Novel Orphan Genes and Species-Specific Invasions of Transposable Elements. *G3-Genes Genomes Genetics* **5**: 1323–1333.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. **2010**. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology* **59**: 307–321.

Guindon S, Gascuel O. **2003**. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* **52**: 696–704.

Guy L, Kultima JR, Andersson SGE. **2010**. GenoPlotR: comparative gene and genome visualization in R. *Bioinformatics* **26**: 2334–2335.

Hartmann F, Croll D. **2017**. Distinct Trajectories of Massive Recent Gene Gains and Losses in Populations of a Microbial Eukaryotic Pathogen. *Molecular Biology and Evolution*.

Hartmann F, McDonald M, Croll D. **2018**. Genome-wide evidence for divergent selection between populations of a major agricultural pathogen. *Molecular Ecology* **27**: 2725–2741.

Hartmann FE, Sánchez-Vallet A, McDonald BA, Croll D. **2017**. A fungal wheat pathogen evolved host specialization by extensive chromosomal rearrangements. *The ISME Journal* **11**: 1189–1204.

Hollister JD, Gaut BS. **2009**. Epigenetic silencing of transposable elements: A trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Research* **19**: 1419–1428.

Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, Rupp R. **2007**. Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics* **8**: 1–6.

Jiao W-B, Schneeberger K. **2019**. Chromosome-level assemblies of multiple Arabidopsis thaliana accessions reveal hotspots of genomic rearrangements. *bioRxiv*: 738880.

Jombart T. **2008**. Adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**: 1403–1405.

Jombart T, Ahmed I. **2011**. adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**: 3070–3071.

Katoh K, Standley DM. **2013**. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution* **30**: 772–780.

Kidwell MG. **2002**. Transposable elements and the evolution of genome size in eukaryotes. *Genetica* **115**: 49–63.

Krishnan P, Meile L, Plissonneau C, Ma X, Hartmann FE, Croll D, McDonald BA, Sánchez-Vallet A. **2018**. Transposable element insertions shape gene regulation and melanin production in a fungal pathogen of wheat. *BMC Biology* **16**: 1–18.

Lai X, Schnable JC, Liao Z, Xu J, Zhang G, Li C, Hu E, Rong T, Xu Y, Lu Y. **2017**. Genome-wide characterization of non-reference transposable element insertion polymorphisms reveals genetic diversity in tropical and temperate maize. *BMC Genomics* **18**: 1–13.

Langmead B, Salzberg SL. **2012**. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**: 357–359.

Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. **2013**. Software for Computing and Annotating Genomic Ranges (A Prlic, Ed.). *PLOS Computational Biology* **9**: e1003118.

Li H, Durbin R. **2009**. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. **2009**. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Lim JK. **1988**. Intrachromosomal rearrangements mediated by hobo transposons in Drosophila melanogaster. *PNAS* **85**: 9153–9157.

Linde CC, Zhan J, McDonald BA. **2002**. Population Structure of *Mycosphaerella graminicola* : From Lesions to Continents. *Phytopathology* **92**: 946–955.

Linheiro RS, Bergman CM. **2012**. Whole Genome Resequencing Reveals Natural Target Site Preferences of Transposable Elements in Drosophila melanogaster (JE Stajich, Ed.). *PLOS ONE* **7**: e30008.

Lu L, Chen J, Robb SMC, Okumoto Y, Stajich JE, Wessler SR. **2017**. Tracking the genome-wide outcomes of a transposable element burst over decades of amplification. *Proceedings of the National Academy of Sciences*: 201716459.

Luo S, Zhang H, Duan Y, Yao X, Clark AG, Lu J. **2020**. The evolutionary arms race between transposable elements and piRNAs in Drosophila melanogaster. *BMC Evolutionary Biology* **20**: 14.

Lynch M. **2007**. *The Origins of Genome Architecture*. Sunderland MA: Sinauer Associates.

McDonald BA, Mundt CC, Chen R. **1996**. The role of selection on the genetic structure of pathogen populations : Evidence from field experiments with Mycosphaerella graminicola on wheat. *Euphytica* **92**: 73–80.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, *et al.* **2010**. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**: 1297–1303.

Meile L, Croll D, Brunner PC, Plissonneau C, Hartmann FE, McDonald BA, Sánchez-Vallet A. **2018**. A fungal avirulence factor encoded in a highly plastic genomic region triggers partial resistance to septoria tritici blotch. *New Phytologist* **219**: 1048–1061.

Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, Brooks S, Howe E, Porubsky D, Logsdon GA, *et al.* **2020**. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**: 79–84.

Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. **2018**. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* **34**: i142–i150.

Nagarajan N, Pop M. **2013**. Sequence assembly demystified. *Nature Reviews Genetics* **14**: 157–167.

Nelson MG, Linheiro RS, Bergman CM. **2017**. McClintock: An Integrated Pipeline for Detecting Transposable Element Insertions in Whole-Genome Shotgun Sequencing Data. *G3&amp;#58; Genes|Genomes|Genetics* **7**: 2763–2778.

Odorizzi G, Babst M, Emr SD. **2000**. Phosphoinositide signaling and the regulation of membrane trafficking in yeast. *Trends in Biochemical Sciences* **25**: 229–235.

Oliver KR, McComb JA, Greene WK. **2013**. Transposable elements: Powerful contributors to angiosperm evolution and diversity. *Genome Biology and Evolution* **5**: 1886–1901.

Omrane S, Audéon C, Ignace A, Duplaix C, Aouini L, Kema G, Walker A-S, Fillinger S. **2017**. Plasticity of the MFS1 promoter leads to multi drug resistance in the wheat pathogen Zymoseptoria tritici. *mSphere*: 1–42.

Omrane S, Sghyer H, Audeon C, Lanen C, Duplaix C, Walker AS, Fillinger S. 2015. Fungicide efflux and the MgMFS1 transporter contribute to the multidrug resistance phenotype in Zymoseptoria tritici field isolates. *Environmental Microbiology* **17**: 2805–2823.

Peter M, Kohler A, Ohm RA, Kuo A, Krützmann J, Morin E, Arend M, Barry KW, Binder M, Choi C, *et al.* 2016. Ectomycorrhizal ecology is imprinted in the genome of the dominant symbiotic fungus Cenococcum geophilum. *Nature Communications* **7**: 1–15.

Petrov DA, Aminetzach YT, Davis JC, Bensasson D, Hirsh AE. 2003. Size matters: Non-LTR retrotransposable elements and ectopic recombination in Drosophila. *Molecular Biology and Evolution* **20**: 880–892.

Plissonneau C, Stürchler A, Croll D. 2016. The Evolution of Orphan Regions in Genomes of a Fungal Pathogen of Wheat. *mBio* **7**: 1–13.

Quinlan AR, Hall IM. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

R Core Team. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Raffaele S, Kamoun S. 2012. Genome evolution in filamentous plant pathogens: why bigger can be better. *Nature Reviews Microbiology* **10**: 417–430.

Rice P, Longden L, Bleasby A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* **16**: 276–277.

Ritz C, Streibig JC. 2005. Bioassay analysis using R. *Journal of Statistical Software* **12**: 1–22.

Rogers J, Gibbs RA. 2014. Content and Dynamics. *Nature Reviews Genetics* **15**: 347–359.

Rouxel T, Grandaubert J, Hane JK, Hoede C, van de Wouw AP, Couloux A, Dominguez V, Anthouard V, Bally P, Bourras S, *et al.* 2011. Effector diversification within compartments of the Leptosphaeria maculans genome affected by Repeat-Induced Point mutations. *Nature communications* **2**: 202.

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, *et al.* 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**: 913–918.

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL. 1998. The paleontology of intergene retrotransposons of maize. *Nature Genetics* **20**: 43–45.

Shen RM, Batzer MA, Deininger PL. 1991. Evolution of the master Alu gene(s). *Journal of Molecular Evolution* **33**: 311–320.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**: 3210–3212.

Slotkin RK, Martienssen R. 2007. Transposable elements and the epigenetic regulation of the genome. *Nature Reviews Genetics* **8**: 272–285.

Stöver BC, Müller KF. 2010. TreeGraph 2: Combining and visualizing evidence from different phylogenetic analyses. *BMC Bioinformatics* **11**: 1–9.

Stritt C, Gordon SP, Wicker T, Vogel JP, Roulin AC. 2017. Recent activity in expanding populations and purifying selection have shaped transposable element landscapes across natural accessions of the Mediterranean grass Brachypodium distachyon. *Genome Biology and Evolution* **10**: 1–38.

Stuart T, Eichten SR, Cahn J, Karpievitch Y V, Borevitz JO, Lister R. 2016. Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. *eLife* **5**: 1–27.

Stukenbrock EH, Banke S, Javan-Nikkhah M, McDonald BA. **2007**. Origin and domestication of the fungal wheat pathogen Mycosphaerella graminicola via sympatric speciation. *Molecular Biology and Evolution* **24**: 398–411.

Torriani SFF, Melichar JPE, Mills C, Pain N, Sierotzki H, Courbot M. **2015**. Zymoseptoria tritici: A major threat to wheat production, integrated approaches to control. *Fungal Genetics and Biology* **79**: 8–12.

Walser J-C, Chen B, Feder ME. **2006**. Heat-Shock Promoters: Targets for Evolution by P Transposable Elements in Drosophila. *PLOS Genetics* **2**: e165.

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. **2009**. Jalview Version 2-A multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**: 1189–1191.

Wickham H. **2016**. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.

Wickham H, Chang W. **2016**. devtools: Tools to Make Developing R Packages Easier.

Wong WY, Simakov O, Bridge DM, Cartwright P, Bellantuono AJ, Kuhn A, Holstein TW, David CN, Steele RE, Martínez DE. **2019**. Expansion of a single transposable element family is associated with genome-size increase and radiation in the genus Hydra. *Proceedings of the National Academy of Sciences* **116**: 22915–22917.

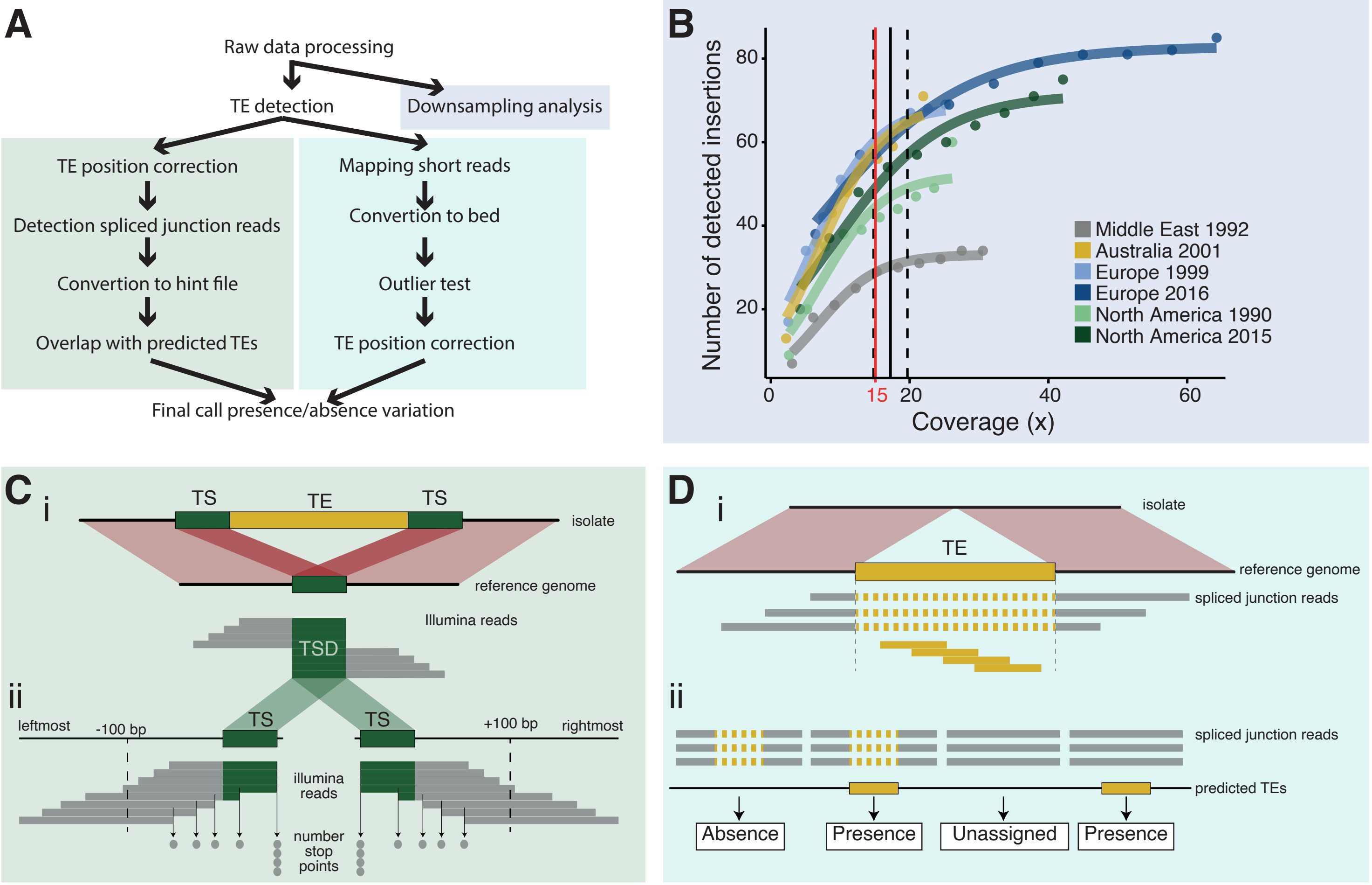
Zhan J, Kema GHJ, Waalwijk C, McDonald BA. **2002**. Distribution of mating type alleles in the wheat pathogen Mycosphaerella graminicola over spatial scales from lesions to continents. *Fungal Genetics and Biology* **36**: 128–136.

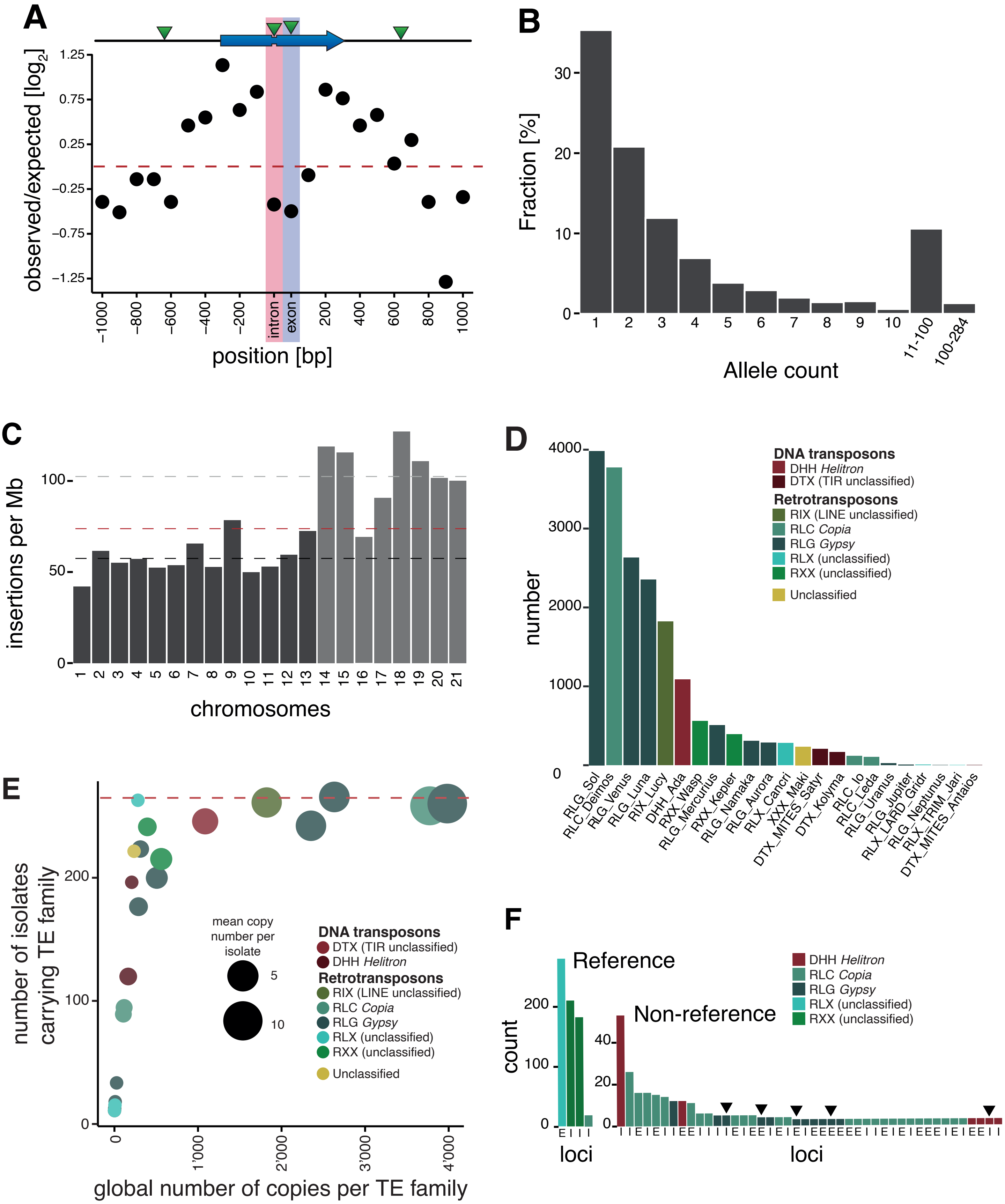
Zhan J, Linde CC, Jurgens T, Merz U, Steinebrunner F, McDonald BA. **2005**. Variation for neutral markers is correlated with variation for quantitative traits in the plant pathogenic fungus Mycosphaerella graminicola. *Mol Ecol* **14**: 2683–2693.

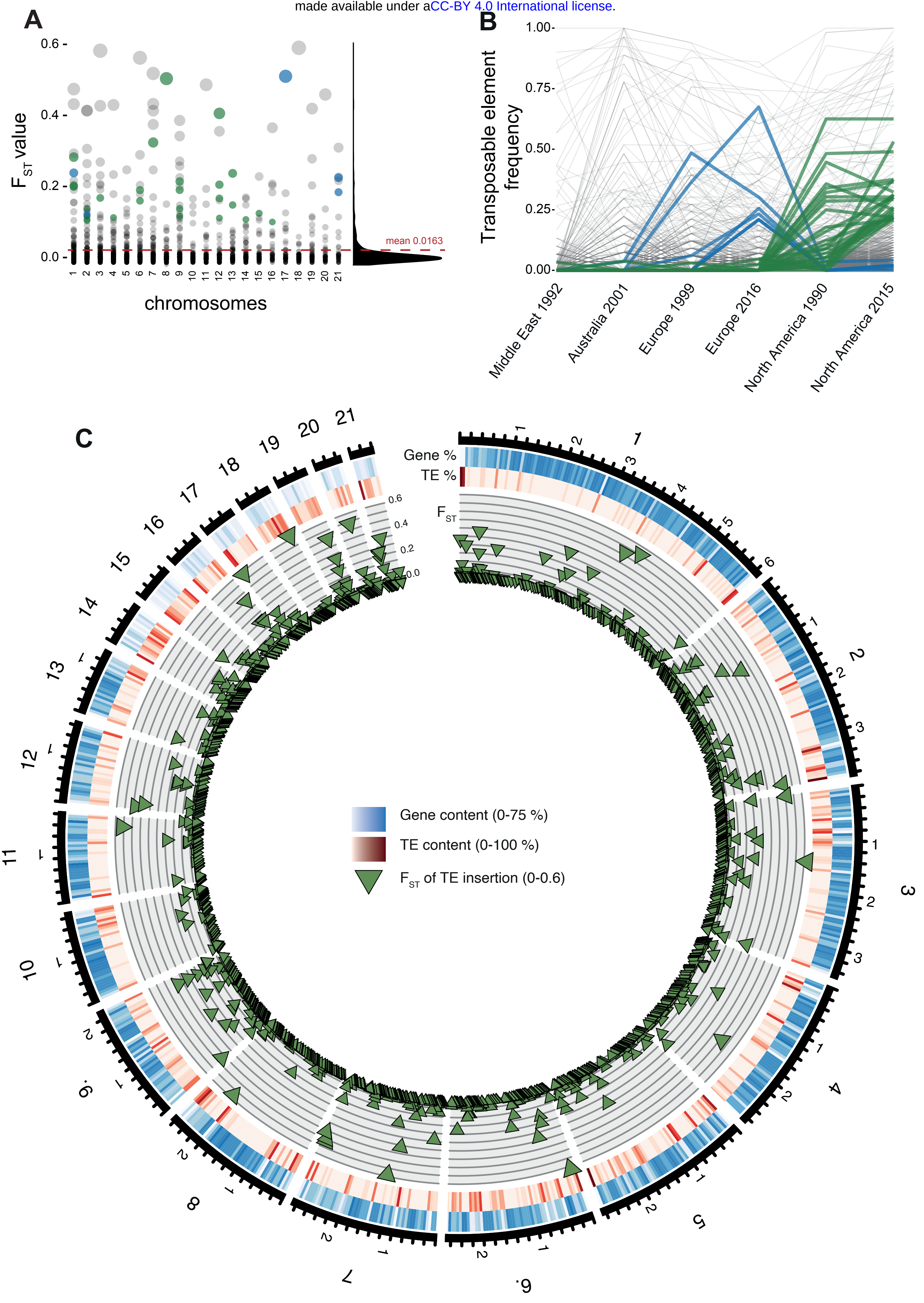
Zhan J, Pettway RE, McDonald BA. **2003**. The global genetic structure of the wheat pathogen Mycosphaerella graminicola is characterized by high nuclear diversity, low mitochondrial diversity, regular recombination, and gene flow. *Fungal Genetics and Biology* **38**: 286–297.

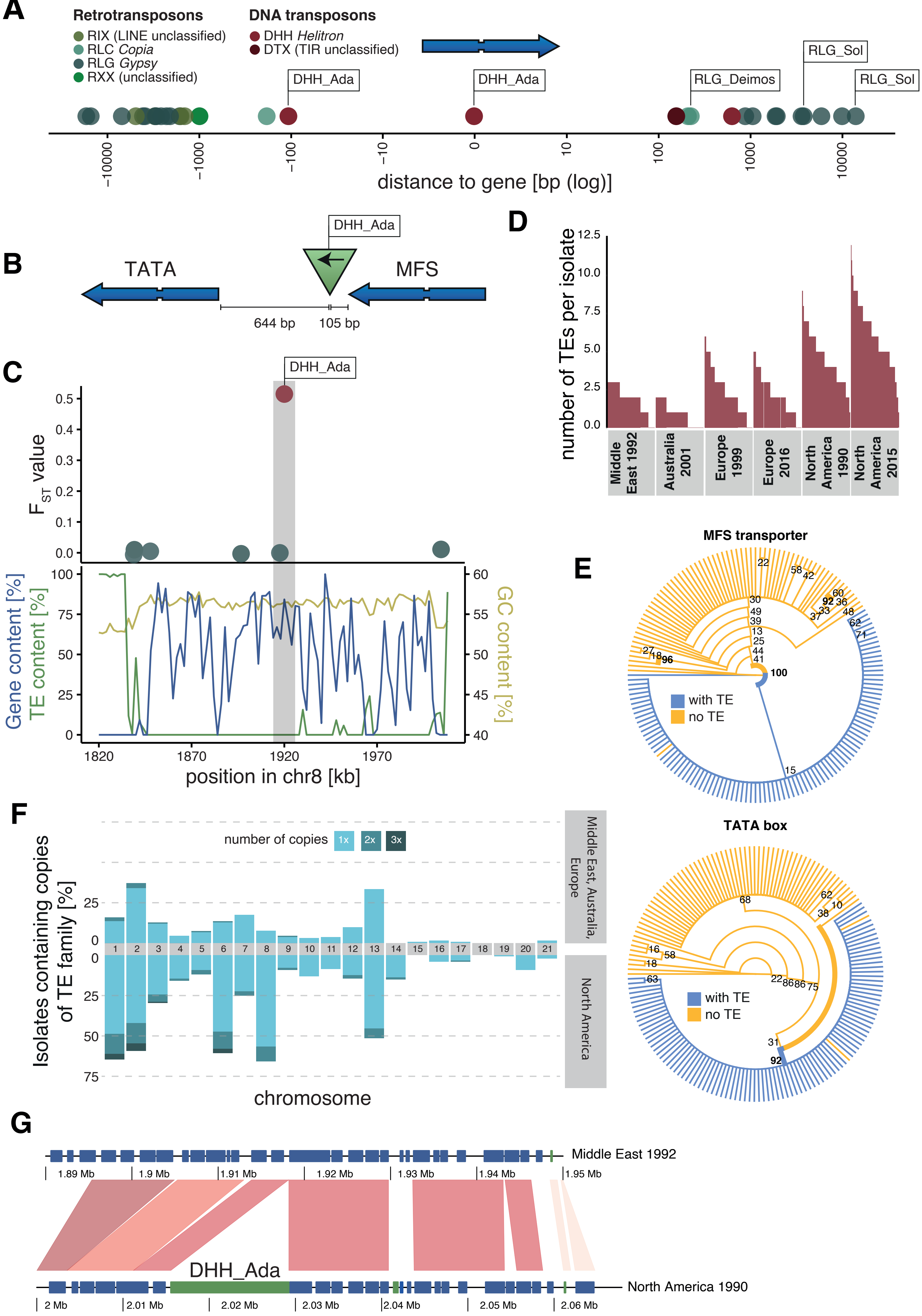
Zheng X, Gogarten SM, Lawrence M, Stilp A, Conomos MP, Weir BS, Laurie C, Levine D. **2017**. SeqArray-a storage-efficient high-performance data format for WGS variant calls. *Bioinformatics* **33**: 2251–2257.

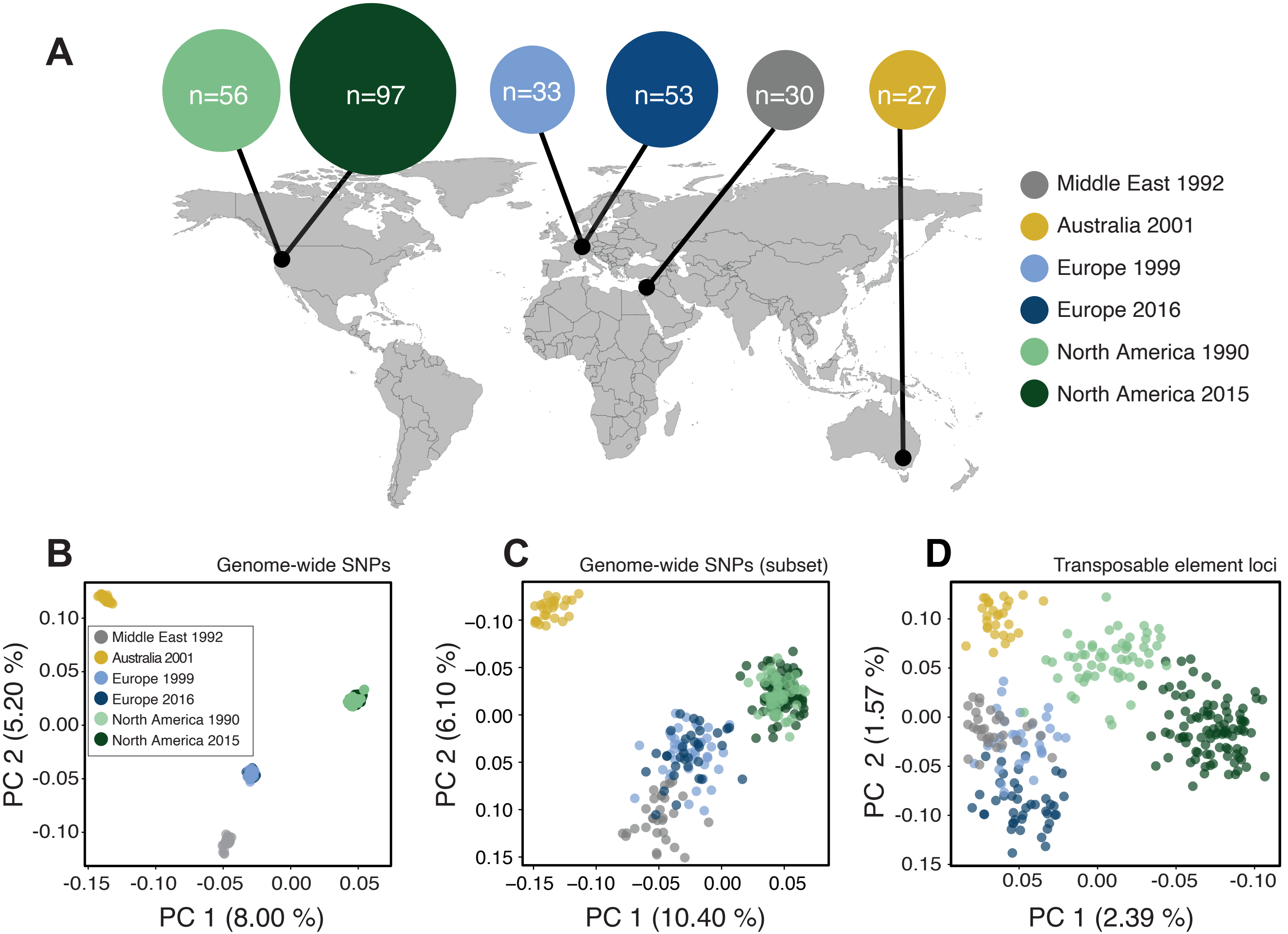
Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. **2012**. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**: 3326–3328.

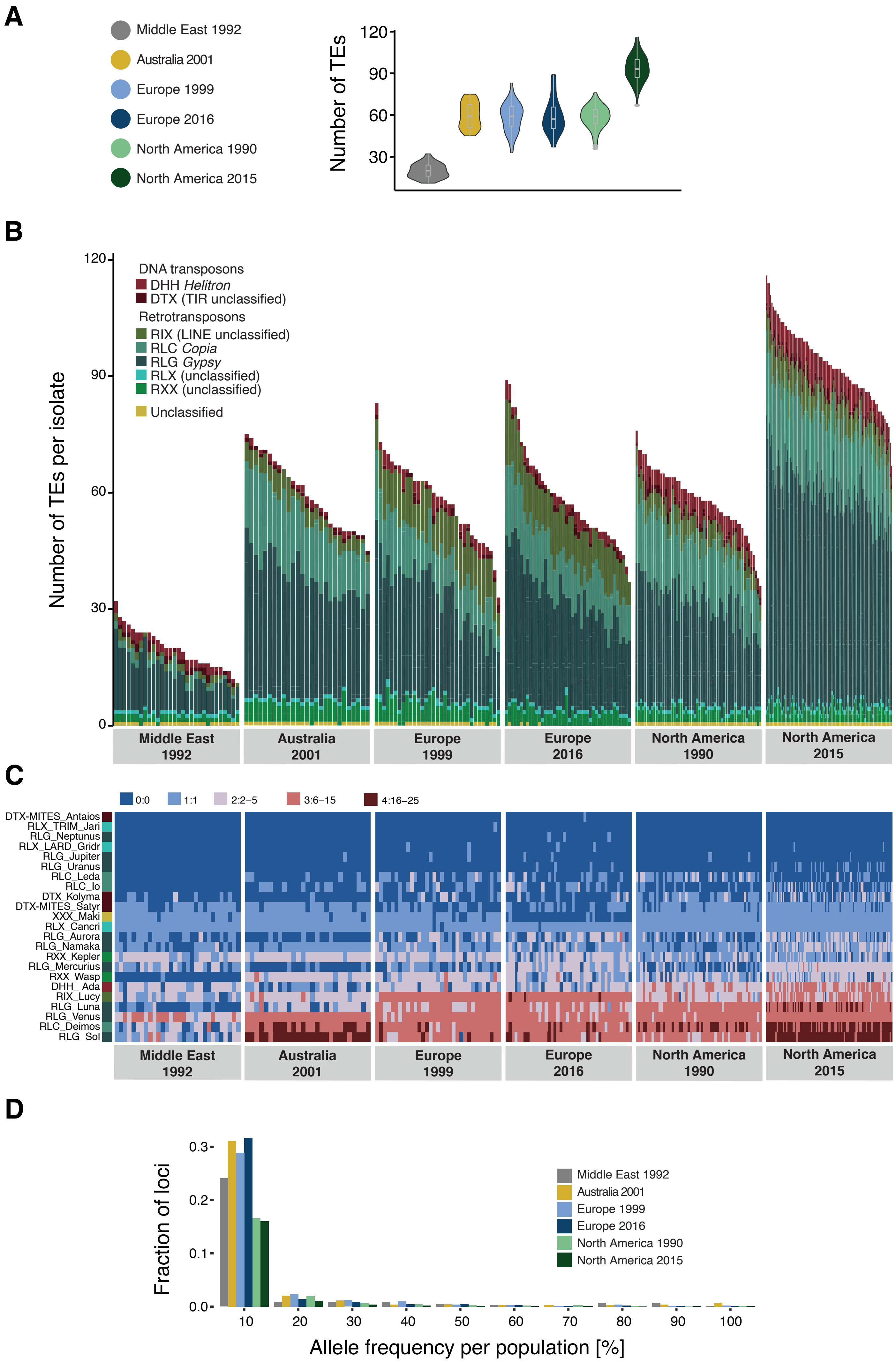

A
Raw data processing
TE detection
Downsampling analysis
TE position correction
Detection spliced junction reads
Convertion to hint file
Overlap with predicted TEs
Mapping short reads
Convertion to bed
Outlier test
TE position correction
Final call presence/absence variation
B
Number of detected insertions
80
60
40
20
0
15
20
40
60
Coverage (x)
Middle East 1992
Australia 2001
Europe 1999
Europe 2016
North America 1990
North America 2015
C
i
TS
TE
TS
isolate
reference genome
Illumina reads
TSD
ii
leftmost
-100 bp
TS
TS
+100 bp
rightmost
illumina reads
number stop points
D
i
isolate
TE
reference genome
spliced junction reads
ii
spliced junction reads
predicted TEs
Absence
Presence
Unassigned
Presence

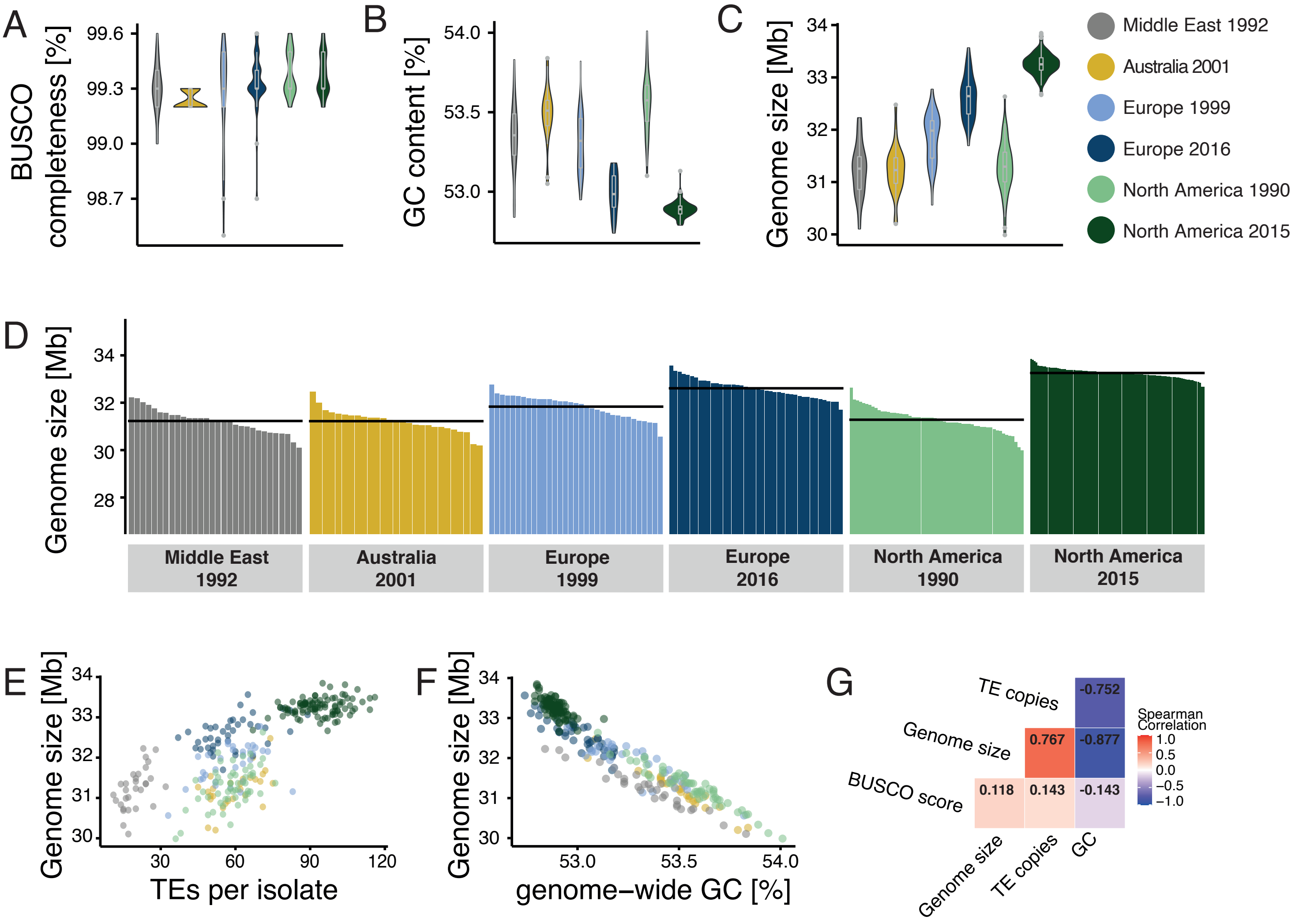

A
observed/expected [log2]
1.25
0.75
0.25
−0.25
−0.75
−1.25
−1000
−800
−600
−400
−200
intron
exon
200
400
600
800
1000
position [bp]
B
Fraction [%]
30
20
10
0
1
2
3
4
5
6
7
8
9
10
11-100
100-284
Allele count
C
insertions per Mb
100
50
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
chromosomes
D
number
4000
3000
2000
1000
0
DNA transposons
DHH Helitron
DTX (TIR unclassified)
Retrotransposons
RIX (LINE unclassified)
RLC Copia
RLG Gypsy
RLX (unclassified)
RXX (unclassified)
Unclassified
RLG_Sol
RLC_Deimos
RLG_Venus
RLG_Luna
RIX_Lucy
DHH_Ada
RXX_Wasp
RLG_Mercurius
RXX_Kepler
RLG_Namaka
RLG_Aurora
RLX_Cancri
XXX_Maki
DTX_MITES_Satyr
DTX_Kolyma
RLC_Io
RLC_Leda
RLG_Uranus
RLG_Jupiter
RLX_LARD_Gridr
RLG_Neptunus
RLX_TRIM_Jari
DTX_MITES_Antaios
E
number of isolates carrying TE family
200
100
0
0
1'000
2'000
3'000
4'000
global number of copies per TE family
mean copy number per isolate
5
10
DNA transposons
DTX (TIR unclassified)
DHH Helitron
Retrotransposons
RIX (LINE unclassified)
RLC Copia
RLG Gypsy
RLX (unclassified)
RXX (unclassified)
Unclassified
F
Reference
Non-reference
count
200
100
0
40
20
0
DHH Helitron
RLC Copia
RLG Gypsy
RLX (unclassified)
RXX (unclassified)
E I I I
loci
I I E I E I I E E I I I I E I E E I E I E I E E E E E E I I E I E I E E E I E I E E I I
loci


A
F_{ST} value
0.6
0.4
0.2
0.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
chromosomes
mean 0.0163
B
Transposable element frequency
1.00
0.75
0.50
0.25
0.00
Middle East 1992
Australia 2001
Europe 1999
Europe 2016
North America 1990
North America 2015
C
Gene %
TE %
F_{ST}
0.6
0.4
0.2
0.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
Gene content (0-75 %)
TE content (0-100 %)
F_{ST} of TE insertion (0-0.6)


A
Retrotransposons
RIX (LINE unclassified)
RLC Copia
RLG Gypsy
RXX (unclassified)
DNA transposons
DHH Helitron
DTX (TIR unclassified)
DHH_Ada
DHH_Ada
RLG_Deimos
RLG_Sol
RLG_Sol
−10000
−1000
−100
−10
0
10
100
1000
10000
distance to gene [bp (log)]
B
DHH_Ada
TATA
MFS
644 bp
105 bp
C
DHH_Ada
F_{ST} value
0.5
0.4
0.3
0.2
0.1
0.0
Gene content [%]
TE content [%]
GC content [%]
100
75
50
25
0
60
55
50
45
40
1820
1870
1920
1970
position in chr8 [kb]
D
number of TEs per isolate
12.5
10.0
7.5
5.0
2.5
0.0
Middle East 1992
Australia 2001
Europe 1999
Europe 2016
North America 1990
North America 2015
E
MFS transporter
22
58
42
60
92
36
48
37
33
62
71
30
49
39
13
25
44
41
27
18
96
100
15
with TE
no TE
TATA box
68
62
10
38
16
58
18
63
22
86
86
75
31
92
with TE
no TE
F
number of copies
1x
2x
3x
Isolates containing copies of TE family [%]
25
0
0
25
50
75
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
Middle East, Australia, Europe
North America
chromosome
G
Middle East 1992
1.89 Mb
1.9 Mb
1.91 Mb
1.92 Mb
1.93 Mb
1.94 Mb
1.95 Mb
DHH_Ada
North America 1990
2 Mb
2.01 Mb
2.02 Mb
2.03 Mb
2.04 Mb
2.05 Mb
2.06 Mb


A
n=56
n=97
n=33
n=53
n=30
n=27
● Middle East 1992
● Australia 2001
● Europe 1999
● Europe 2016
● North America 1990
● North America 2015
B
Genome-wide SNPs
● Middle East 1992
● Australia 2001
● Europe 1999
● Europe 2016
● North America 1990
● North America 2015
0.10
0.05
0.00
-0.05
-0.10
PC 2 (5.20 %)
-0.15
-0.10
-0.05
0.00
0.05
PC 1 (8.00 %)
C
Genome-wide SNPs (subset)
−0.10
−0.05
0.00
0.05
0.10
0.15
PC 2 (6.10 %)
−0.15
−0.10
−0.05
0.00
0.05
PC 1 (10.40 %)
D
Transposable element loci
0.10
0.05
0.00
-0.05
-0.10
0.15
PC 2 (1.57 %)
0.05
0.00
-0.05
-0.10
PC 1 (2.39 %)


A
Middle East 1992
Australia 2001
Europe 1999
Europe 2016
North America 1990
North America 2015
Number of TEs
120
90
60
30
B
Number of TEs per isolate
120
90
60
30
0
DNA transposons
DHH Helitron
DTX (TIR unclassified)
Retrotransposons
RIX (LINE unclassified)
RLC Copia
RLG Gypsy
RLX (unclassified)
RXX (unclassified)
Unclassified
Middle East 1992
Australia 2001
Europe 1999
Europe 2016
North America 1990
North America 2015
C
0:0
1:1
2:2–5
3:6–15
4:16–25
DTX-MITES_Antaios
RLX_TRIM_Jari
RLG_Neptunus
RLX_LARD_Gridr
RLG_Jupiter
RLG_Uranus
RLC_Leda
RLC_Io
DTX_Kolyma
DTX-MITES_Satyr
XXX_Maki
RLX_Cancri
RLG_Aurora
RLG_Namaka
RXX_Kepler
RLG_Mercurius
RXX_Wasp
DHH_ Ada
RIX_Lucy
RLG_Luna
RLG_Venus
RLC_Deimos
RLG_Sol
Middle East 1992
Australia 2001
Europe 1999
Europe 2016
North America 1990
North America 2015
D
Fraction of loci
0.3
0.2
0.1
0.0
10
20
30
40
50
60
70
80
90
100
Allele frequency per population [%]
Middle East 1992
Australia 2001
Europe 1999
Europe 2016
North America 1990
North America 2015


A
BUSCO completeness [%]
99.6
99.3
99.0
98.7
B
GC content [%]
54.0
53.5
53.0
C
Genome size [Mb]
34
33
32
31
30
● Middle East 1992
● Australia 2001
● Europe 1999
● Europe 2016
● North America 1990
● North America 2015
D
Genome size [Mb]
34
32
30
28
Middle East 1992
Australia 2001
Europe 1999
Europe 2016
North America 1990
North America 2015
E
Genome size [Mb]
34
33
32
31
30
30
60
90
120
TEs per isolate
F
Genome size [Mb]
34
33
32
31
30
53.0
53.5
54.0
genome–wide GC [%]
G
TE copies
Genome size
BUSCO score
-0.752
0.767
-0.877
0.118
0.143
-0.143
Genome size
TE copies
GC
Spearman Correlation
1.0
0.5
0.0
−0.5
−1.0