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; Piegu *et al.*, 2006; 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 in many years (Torriani *et al.*, 2015). *Z. tritici* emerged during the domestication of wheat in the Fertile Crescent where the species retained high levels of genetic variation (Zhan *et al.*, 2005; Stukenbrock *et al.*, 2011). The pathogen migrated to all temperate zones where wheat is currently grown and underwent multiple migration bottlenecks, in particular when colonizing Oceania and North America (Zhan *et al.*, 2005; Estep *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

We detected 4,753 TE copies, grouped into 30 families with highly variable copy numbers in the reference genome IPO323 (Figure 2 - figure supplement 1 and Figure 2 - figure supplement 2A). To establish a comprehensive picture of within-species TE dynamics, we analyzed 295 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 and removed 11 genomes with an average read depth below 15X (Figure 1B). We tested for evidence of TEs using read depth at target site duplications (Figure 1C) and scanned the genome for mapped reads indicating gaps at TE loci (Figure 1D). We found robust evidence for a total of 18,864 TE insertions grouping into 2,465 individual loci. Of these loci, 35.5% (*n* = 876) have singleton TEs (*i.e.*, this locus is only present in one isolate: Figure 2A, figure supplement 3). An overwhelming proportion of loci (2,345 loci or 95.1%) have a TE frequency below 1%. Singleton TE insertions in particular can be the product of spurious Illumina read mapping errors (Nakamura *et al.*, 2011). To assess the reliability of the detected singletons, we focused on seven isolates for which PacBio long-read data was available (Badet *et al.*, 2020). Aligned PacBio reads confirmed the exact location of 71% (22 out of 31 singleton insertions among seven isolates; see Methods for further details). We found no significant difference in read coverage between confirmed and unconfirmed singleton insertions (Figure 2 - figure supplement 2C,-B and Figure 2 - figure supplement 4).

The abundance of singleton TE insertions strongly supports the idea that TEs actively copy into new locations but also indicates that strong purifying selection maintains nearly all TEs at low frequency (Figure 2A). The density of TE loci on accessory chromosomes, which are not shared among all isolates of the species, is almost twice the density found on core chromosomes (102 *versus* 58 TEs per Mb; Figure 2B and Figure 2 - figure supplement 5A). This suggests relaxed selection against TE insertion on the functionally dispensable and gene-poor accessory chromosomes. We found no difference in TE allele frequency distribution between recombination hotspots and the rest of the genome (Figure 2 - figure supplement 5B). Similarly, the TE density and the number of insertions did not vary between recombination hotspots and the genomic background (Figure 2 - figure supplement 5C).

TEs grouped into 23 families and 11 superfamilies, with 88.2% of all copies belonging to class I/retrotransposons (*n* = 2175; Figure 2C and Figure 2 - figure supplements 6A-B). RLG/*Gypsy* (*n* = 1,483) and RLC/*Copia* (*n* = 623) elements constitute the largest long terminal repeats (LTR) superfamilies. Class II/DNA transposons are dominated by DHH/*Helitron* (*n* = 249). As expected, TE families shared among fewer isolates tend to show also lower global copy numbers (*i.e.*, all isolates combined), while TE families that are present in all isolates generally have high global copy numbers (Figure 2D).

We detected 153 loci with TEs inserted into genes with most of the insertions being singletons (44.7%; *n* = 68) or of very low frequency (Figure 2E). 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 2F). 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 800-1,000 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-B and Figure 3 - figure supplement 1). To further analyze evidence for TE-mediated adaptive evolution, we screened a genome-wide SNP dataset for evidence of selective sweeps using selection scans. We found 16.5 % of all TE loci located in regions of selective sweep. Given our population sampling of two population pairs, we tested for adaptive TE insertions in selective sweep regions 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 7 candidate loci possibly underlying local adaptation (6 in North America, one in Europe; Figure 4A and Figure 4 - figure supplement 1). All loci carry inserted retrotransposons with 4 RLG_Luna, one RLG_Mercurius and one RLG_Deimos.

One TE insertion is 3,815 bp downstream of a gene encoding an RTA1-like protein, which can function as transporters with a transmembrane domain and have been associated with resistance against several antifungal compounds (Soustre *et al.*, 1996). The insertion is also 5785 bp upstream of a gene encoding a protein kinase domain (Figure 4B). The TE insertion was not detected in the Middle East or the two European populations, and was at low frequencies in the Australian (3.7%) and North American 1990 (1.7%) populations, but increased to 53% of all isolates in the North American 2015 population (fixation index F_{ST} = 0.42; Figure 4 - figure supplement 1). Isolates that carry the insertion show a significantly higher resistance to azole antifungal compounds (Figure 4C). The TE is in the subtelomeric region of chromosome 12, with a moderate GC content, a low TE and a high gene density (Figure 4D). The TE belongs to the family RLG_Luna, which shows a substantial burst across different chromosomes within the species (Figures 4E-F). We found no association between the phylogenetic relationships among isolates based on the two closest genes and the presence or absence of the TE insertion (Figure 4G). A second candidate adaptive TE insertion belongs to the RLG_Mercurius family and is located between two genes of unknown function (Figure 4 - figure supplement 2). A third potentially adaptive TE insertion of a RLC_Deimos is 229 bp upstream of a gene encoding a SNARE domain protein and 286 bp upstream of a gene encoding a flavin amine oxidoreductase. Furthermore, the TE is inserted in a selective sweep region (Figure 4 - figure supplement 2). SNARE domains play a role in vesicular transport and membrane fusion (Bonifacino & Glick, 2004). An additional four candidates for adaptive TE insertions belong to RLG_Luna and were located distantly to genes (Figure 4 - figure supplement 2). We experimentally tested whether the TE insertions in proximity to genes were associated with 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 two loci was positively associated with higher levels of fungicide resistance, suggesting that the adaptation was mediated by the TE (Figure 4C and Figure 4 - figure supplement 2).

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 (Figure 5A) and 900,193 genome-wide SNPs (Figure 5B). The population structure reflected the demographic history of the pathogen with clear continental differentiation and only minor within-site differentiation. To account for the lower number of TE loci, we performed an additional principal component analysis using a random SNP set of similar size to the number of TE loci. The reduced SNP set retained the geographic signal of the broader set of SNPs (Figure 5C). 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.

Unusual patterns in population differentiation at TE loci suggests that TE activity may substantially vary across populations (Figure 6, Figure 4 - figure supplement 1). 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 6A-C, and Figure 6 - figure supplement 1), 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 (Figure 2 - figure supplement 6C). 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 compared to the Middle East (Israel) was associated with an upwards shift in allele frequencies. This is consistent with transposition activity creating new copies in the genomes and stronger purifying selection in the Middle East. Similarly, the North American populations showed also signatures consistent with relaxation of selection against TEs (*i.e.*, fewer low frequency TEs). We found a significant difference (Two-sample Kolmogorov-Smirnov test, two-sided) in the curve shapes between the population from the Middle East and North America 2015 (Figure 6 - figure supplement 2). We analyzed variation in TE copy numbers across families and found that the expansions were mostly driven by RLG elements including the families Luna, Sol and Venus, the RLC family Deimos and the LINE family Lucy (Figure 6E and Figure 6 - figure supplement 3A). We also found a North American specific burst in DHH elements of the family Ada (increase from 4.6 to 6.1 copies on average per isolate), an increase specific to Swiss populations in LINE elements, and an increase in RLC elements in the Australian and the two North American populations. Analyses of complete *Z. tritici* reference-quality genomes that include isolates from the Israel, Australia, Switzerland (1999) and North American (1990) population revealed high TE contents in Australia and North America (Oregon 1990) (Badet *et al.*, 2020). The reference-quality genomes confirmed also that the increase in TEs was driven by LINE, RLG and RLC families in Australia and DHH, RLG and RLC families in North America (Badet *et al.*, 2020).

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. Consequently, mean genome sizes in populations should increase over generations. We estimated the cumulative length of TE insertions based on the length of the corresponding TE consensus sequences and found a strong increase in the total TE length in populations outside the Middle East center of origin, and a second increase between the two North American populations (Figure 1 - figure supplement 1). 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 reference-quality 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 core genome sizes (Figure 7C). We found a notable increase in core genome size in both the more recent European and North American populations. The increase in core genome size is positively correlated with the count and cumulative length of all inserted TEs (Figure 7D, 7E and 7G) and negatively correlated with the genome-wide GC content (Figure 7F and 7G). Hence, core genome size shows substantial variation within the species matching the recent expansion in TEs across continents. We found the most variable genome sizes in the more recent North American population (Figure 7 - figure supplement 1B). Finally, we contrasted variation in genome size with the detected TE insertion dynamics. For this, we assessed the variable genome segment as the difference between the smallest and largest analyzed core genome. To reflect TE dynamics, we calculated the cumulative length of all detected TE insertions in any given genome. We found that the cumulative length of inserted TEs represents between 4.8 and 184 % of the variable genome segment defined for the species or 0.2-2.6% of the estimated genome size per isolate (Figure 7 - figure supplement 1C-D).

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. However, our short read based analyses are possibly skewed towards over-counting singletons as indicated by independent long-read mapping evaluations. Nevertheless, the skew towards low frequency TE insertions 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. Some TE superfamilies have preferred insertion sites in coding regions and transcription start sites (Miyao *et al.*, 2003; Fu *et al.*, 2013; Gilly *et al.*, 2014; Quadrana *et al.*, 2016). Hence, some heterogeneity in the observed insertion site distribution across the genome is likely due to insertion preferences of individual TEs. We also found evidence for positive selection acting on TEs with the strongest candidate locus being a TE insertion on chromosome 12. This locus showed a frequency increase only in the more recent North

American population, 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). The nearest gene encodes a RTA1-like protein, a transmembrane exporter which is associated with resistance towards different stressors, including antifungal compounds, and shows strong copy number variation in several fungi (Soustre *et al.*, 1996; Rogers & Barker, 2003; Sirisattha *et al.*, 2004; Ali *et al.*, 2014; Yew *et al.*, 2016; Liang *et al.*, 2018). Hence, the TE insertion may have positively modulated RTA1 expression to resist antifungals.

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 reference-quality genomes from the same populations confirmed that genome expansions were primarily driven by the same TE families belonging to the RLG, RLC and DHH 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. By establishing large population genetic datasets, such as those possible for crop pathogens, analyses of genome size evolution become tractable at the population level.

TEs might not only contribute to genome expansion directly by adding length through additional copies, but also by increasing the rate of chromosomal rearrangements and ectopic recombination (Bourque *et al.*, 2018; Blommaert, 2020). However, TEs are not the only repetitive elements that can lead to a genome size expansion. In *Arabidopsis thaliana* genomes, the 45S rDNA has been shown to have the strongest impact on genome size variation, followed by 5S rDNA variation, and contributions by centromeric repeats and TEs (Long *et al.*, 2013). In conjunction, recent work demonstrates how repetitive sequences are drivers of genome size evolution over short evolutionary timescales.

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 (Figure 5 - figure supplement 1), 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 Eastern, Australian, European 1999 and North American 1990 populations were obtained from the NCBI Sequence Read Archive (SRA) 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 at the NCBI SRA 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; Figure 1 - figure supplement 5) 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 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 target site duplication (Figure 1C). We tested whether the number of reads with a mapped end around the target site duplication significantly deviated if the mapping ended exactly at the boundary. A mapped read ending exactly at the target site duplication boundary is indicative of a split read mapping to a TE sequence absent in the reference genome. To test for the deviation in the number of read mappings around the target site duplication, 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 target site duplication had a *p*-value < 0.001 (Figure 5 - figure supplement 1; Figure 1 - figure supplement 1B and Figure 1 - figure supplement 2).

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, indicating that the isolate had no inserted TE in this region. 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 Figure 1 - figure supplement 1C).

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 (Figure 1 - figure supplement 3). 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) and Singh et al (2020). 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 200000 to obtain an approximately equivalent number of SNPs as TE loci.

EVALUATION OF SINGLETON INSERTIONS

To evaluate the reliability of singleton TE insertion loci, we analyzed singleton loci in isolates for which we had both Illumina datasets and complete reference-quality genomes (Badet *et al.*, 2020). From a set of 19 long-read PacBio reference genomes spanning the global distribution of *Z. tritici*, one isolate each from Australia, Israel, North America (1990) and four isolates from Europe (1999) were also included in the TE insertion screening. To assess the reliability of singleton TE insertions, we first investigated structural variation analyses among the reference genomes (Badet *et al.*, 2021, Supplementary Data 1 and 2). The structural variation was called both based on split read mapping of PacBio reads and pairwise whole-genome alignments. Using bedtools intersect, we recovered for the 31 singleton TE loci in the 7 analyzed genomes a total of 17 loci showing either an indel, translocation, copy number polymorphism, duplication, inverted duplication, inversion, or inverted translocation at the same location. We visually inspected the PacBio read alignment bam files against the IPO323 reference genome using IGV version 2.4.16 (Robinson *et al.*, 2011), and found a typical coverage increase at the target site duplication, with most read mappings interrupted at the target site duplication as expected for an inserted TE. For the 14 remaining TE loci, we extracted the region of the predicted insertion and padded the sequence on both ends with an additional 500 bp using samtools faidx. We used blast to identify a homologous region in the assembled reference-quality genomes. Matching regions were inspected based on blastn for the presence of a TE sequence matching the TE family originally detected at the locus. With this second approach, we confirmed an additional five singletons to be true insertions. Both methods combined produced supportive evidence for 22 out of 31 singleton insertions (71%). We calculated the read coverage after mapping to the reference genome IPO323 with bedtools genomecov for each PacBio long-read dataset and calculated mean coverage for 500 bp regions around singleton TE insertions.

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 version 6.6.0 (Rice *et al.*, 2000) 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 calculated the population-specific allele frequency for TE loci and estimated the exponential decay curve with a self-starting Nls asymptomatic regression model nls(p_loci ~ SSasymp(p_round, Asym, R0, lrc) in R.

We analyzed signatures of selective sweeps based on genome-wide SNPs using the extended haplotype homozygosity (EHH) tests implemented in the R package *REHH* (Sabeti *et al.*, 2007; 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 signatures based on the XP-EHH statistic for 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 in the population pairs were within 10 kb of a region identified as a selective sweep by XP-EHH using the function intersect from bedtools.

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. 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 extracted the number of TEs present in 1 kb windows around each annotated core gene in the reference genome IPO323, using the function window in bedtools. We calculated the relative distances between each gene and the closest TE with the function bedtools closest. For the TEs inserted into genes, we used the function intersect in bedtools to distinguish intron and exon insertions with the parameters -wo and -v, respectively. TEs that overlap more than one exon were only counted once. For each 100 bp segment in the 1 kb windows as well as for introns and exons, we calculated the mean number of observed TE insertions per base pair. We calculated the mean number of TEs per window and calculated the log2 of the observed number of TE insertions divided by the expected value. We extracted information about recombination hotspots from Croll *et al.* (2015). This dataset is based on two experimental crosses initiated from isolates included in our analyses (1A5x1E4, 3D1x3D7). The recombination rates were assessed based on the reference genome IPO323 and analyzed with the *R/qtl* package in R. We used bedtools intersect to compare both TE density in IPO323 and TE insertion polymorphism with predicted recombination hotspots.

CORE 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 (Figure 1 - figure supplement 4). 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 1 kb (Mikheenko *et al.*, 2018).

Fungicide resistance assay

To quantify susceptibility towards propiconazole we used a previously published microtiter plate assay dataset with 3 replicates performed for each isolate and concentration. Optical density was used to estimate growth rates under different fungicide concentrations (0, 0.00006, 0.00017, 0.0051, 0.0086, 0.015, 0.025, 0.042, 0.072, 0.20, 0.55, 1.5 mgL^{-1}) (Hartmann *et al.*, 2020). We calculated dose-response curves and estimated the half-maximal lethal concentration EC$_{50}$ with a 4-parameter logistics curve in the R package *drc* (Ritz & Streibig, 2005).

Data availability

Sequence data is deposited at the NCBI SRA 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. We also thank the three reviewers for very helpful suggestions. We thank C. Sarai Reyes-Avila for advice on statistical analyses. DC is supported by the Swiss National Science (grants 31003A_173265) and the Fondation Pierre Mercier pour la Science.

Competing interests

We declare to have no competing interests.

REFERENCES

Ali SS, Khan M, Mullins E, Doohan FM. **2014**. Identification of Fusarium oxysporum Genes Associated with Lignocellulose Bioconversion Competency. *Bioenergy Research* **7**: 110–119.

Badet T, Fouché S, Hartmann FE, Zala M, Croll D. **2021**. Machine-learning predicts genomic determinants of meiosis-driven structural variation in a eukaryotic pathogen. *Nature Communications* **12**.

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, Str□mberg 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.

Blommaert J. **2020**. Genome size evolution: towards new model systems for old questions. *Proceedings. Biological sciences* **287**: 20201441.

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.

Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvák Z, Levin HL, Macfarlan TS, *et al.* **2018**. Ten things you should know about transposable elements. *Genome Biology* **19**: 199.

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.

Fu Y, Kawabe A, Etcheverry M, Ito T, Toyoda A, Fujiyama A, Colot V, Tarutani Y, Kakutani T. **2013**. Mobilization of a plant transposon by expression of the transposon-encoded anti-silencing factor. *EMBO Journal* **32**: 2407–2417.

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.

Gilly A, Etcheverry M, Madoui MA, Guy J, Quadrana L, Alberti A, Martin A, Heitkam T, Engelen S, Labadie K, *et al.* 2014. TE-Tracker: systematic identification of transposition events through whole-genome resequencing. *Bmc Bioinformatics* **15**.

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.

Hartmann FE, Vonlanthen T, Singh NK, McDonald MC, Milgate A, Croll D. **2020**. The complex genomic basis of rapid convergent adaptation to pesticides across continents in a fungal plant pathogen. *Molecular Ecology*.

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.

Liang X, Wang B, Dong Q, Li L, Rollins JA, Zhang R, Sun G. **2018**. Pathogenic adaptations of Colletotrichum fungi revealed by genome wide gene family evolutionary analyses. *PLoS ONE* **13**: 1–25.

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.

Long Q, Rabanal FA, Meng D, Huber CD, Farlow A, Platzer A, Zhang Q, Vilhjálmsson BJ, Korte A, Nizhynska V, *et al.* **2013**. Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. *Nature Genetics* **45**: 884–890.

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.

Miyao A, Tanaka K, Murata K, Sawaki H, Takeda S, Abe K, Shinozuka Y, Onosato K, Hirochika H. **2003**. Target site specificity of the Tos17 retrotransposon shows a preference for insertion within genes and against insertion in retrotransposon-rich regions of the genome. *Plant Cell* **15**: 1771–1780.

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

Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H, *et al.* **2011**. Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Research* **39**.

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.

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.

Piegu B, Guyot R, Picault N, Roulin A, Sanyal A, Kim H, Collura K, Brar DS, Jackson S, Wing RA, *et al.* **2006**. Doubling genome size without polyploidization: Dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice. *Genome Research* **21**: 1201.

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

Quadrana L, Silveira AB, Mayhew GF, LeBlanc C, Martienssen RA, Jeddeloh JA, Colot V. **2016**. The Arabidopsis thaliana mobilome and its impact at the species level. *Elife* **5**.

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.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. **2011**. Integrative Genome Viewer. *Nature Biotechnology* **29**: 24–6.

Rogers PD, Barker KS. **2003**. Genome-wide expression profile analysis reveals coordinately regulated genes associated with stepwise acquisition of azole resistance in Candida albicans clinical isolates. *Antimicrobial Agents and Chemotherapy* **47**: 1220–1227.

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.

Singh NK, Chanclud E, Croll D. **2020**. Population-level deep sequencing reveals the interplay of clonal and sexual reproduction in the fungal wheat pathogen Zymoseptoria tritici.

Sirisattha S, Momose Y, Kitagawa E, Iwahashi H. **2004**. Toxicity of anionic detergents determined by Saccharomyces cerevisiae microarray analysis. *Water Research* **38**: 61–70.

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

Soustre I, Letourneux Y, Karst F. **1996**. Characterization of the Saccharomyces cerevisiae RTA1 gene involved in 7-aminocholesterol resistance. *Current Genetics* **30**: 121–125.

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.

Stukenbrock EH, Bataillon T, Dutheil JY, Hansen TT, Li RQ, Zala M, McDonald BA, Wang J, Schierup MH. **2011**. The making of a new pathogen: Insights from comparative population genomics of the domesticated wheat pathogen Mycosphaerella graminicola and its wild sister species. *Genome Research* **21**: 2157–2166.

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.

Yew SM, Chan CL, Kuan CS, Toh YF, Ngeow YF, Na SL, Lee KW, Hoh CC, Yee WY, Ng KP. **2016**. The genome of newly classified Ochroconis mirabilis: Insights into fungal adaptation to different living conditions. *BMC Genomics* **17**: 1–17.

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.

Figure Legends

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 absent 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) Decision map if a TE should be kept as a true insertion or rejected as a false positive. Only predicted TE insertions that overlap evidence of split reads were kept as TE insertions in downstream analyses. (E) Singleton validation using long-read PacBio sequencing. (i) Analysis if TE insertions overlap with a detected insertion/deletion locus (Badet *et al*, 2021). (ii) Homology search of the TE insertion flanking sequences based on the reference genome against PacBio reads. In addition, the consensus sequence of the inserted TE was used for matches between the flanks.

Figure supplement 1. Validation of transposable element (TE) insertion predictions. (A) TEs not present in the reference genome: distribution of additional TE hits found per locus after the outlier test. Color indicates superfamilies. (B) TEs not present in the reference genome: distribution of additional TE hits found per population after the outlier test. Colors indicate populations. (C) TEs present in the reference genome: distribution of missing data per locus after the validation with spliced junction reads. Missing data indicates that the TE was not predicted with ngs_te_mapper and that there was no indication of spliced reads. The red line (=20 %) indicates the threshold for missing data. TE loci with an amount of missing data > 20 % were completely excluded from further analyses. Color indicates superfamily. (D) TEs present in the reference genome: detection of strong outlier isolates with a high number of split reads. Color indicates the population.

Figure supplement 2. TE insertion validations for non-reference copies (Table).

Figure supplement 3. Establishment of transposable element (TE) loci with differing start and end positions in the isolates. Distribution of length of distance for start position, end position and both start and end combined after the correction.

Figure supplement 4. Bias for reads with a GC content lower than 30 % per population. Red lines indicate the mean.

Figure supplement 5. TE consensus sequences (Table).

Figure 2: Transposable element (TE) landscape across populations. (A) Allele frequencies of the TE insertions across all isolates. (B) TE insertions per Mb on core chromosomes (dark) and accessory chromosomes (light). Dashed lines represent mean values. Blue: global mean of 75.65 insertions/Mb, dark: core chromosome mean of 58 TEs/Mb, light: accessory chromosome mean of 102.24 insertions/Mb). (C) Number of TE insertions per family. (D) TE frequencies among isolates and copy numbers across the genome. The blue line indicates the maximum number of isolates (*n* = 284). (E) Allele frequency distribution of TE insertions into introns and exons. (F) Number of TE insertions within 1 kb up- and downstream of genes on core chromosomes including introns and exons (100 bp windows). The blue arrow indicates a gene schematic with exons and an intron, the green triangles indicate TE insertions. The dotted blue line indicates no deviation from the expected value (*i.e.*, mean number of TEs per window).

Figure supplement 1. TEs in reference (Table).

Figure supplement 2. Validation of singleton insertions detected by mapped Illumina reads using PacBio read alignments for confirmation. (A) Comparison of TE family copy numbers per isolate to the number of copies found in the reference genome (IPO323). The color is indicating superfamilies. This figure includes only TE families that were detected in any of the isolates used for validation. (B) Confirmation of singleton TE insertions detected in the isolates CH99_SW5, CH99_SW39, CH99_3D7, CH99_3D1, ISR92_Ar_4f, AUS01_1H8 and ORE90Ste_4A10 using aligned PacBio reads. Confirmed/not confirmed TE insertions are shown by TE family. (C) PacBio read coverage (in 500 bp window) at singleton loci.

Figure supplement 3. Presence absence matrix TE loci (Table).

Figure supplement 4. Singletons (Table).

Figure supplement 5. TE insertion loci characteristics. (A) Number of TE insertions and density (insertions per Mb) in accessory and core genes. (B) Allele frequencies of TEs genome-wide and restricted to recombination hotspots. (C) TE insertion density and TE copy numbers within and outside of recombination hotspots.

Figure supplement 6. Hierarchy superfamilies. (A) Number of transposable element (TE) insertions per superfamily. Colors indicate the superfamily. (B) Number of TE loci and classification hierarchy. (C) Comparison of mean genome sequencing coverage and the number of detected TEs with ngs_te_mapper in isolates of the Middle East population. Dots indicate the coverage and colors indicate the superfamily.

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.

Figure supplement 1. Global pairwise FST distributions shown separately for the 21 chromosomes. The red horizontal line indicates the mean FST = 0.0163. Colors are according to the three main superfamilies (RLG, RLC, DHH).

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. The stars indicate TE insertions not found in the reference genome. (B) Location of the RLG_Luna TE insertion on chromosome 12 corresponding to its two closest genes. (C) Resistance against azole fungicides among isolates as a function of TE presence or absence. (D) Genomic niche of the RLG_Luna TE insertion on chromosome 12: F_{ST} values for each TE insertion, gene content (blue), TE content (green) and GC content (yellow). The grey section highlights the insertion site. (E) Number of RLG_Luna copies per isolate and population. (F) Frequency changes of RLG_Luna between the two North American populations compared to the other populations. Colors indicate the number of copies per chromosome. (G) Phylogenetic trees of the coding sequences of either the gene encoding the RTA1-like protein or the protein kinase domain. 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.

Figure supplement 1. Top loci information (Table).

Figure supplement 2. Additional top loci. Six additional candidate adaptive transposable element (TE) insertions. Each row corresponds to a candidate, with the first five being candidates detected in the North American populations and the last one in the European populations. For each candidate, the direction of the TE and the direction, function and distance of the closest two genes are indicated. The middle column indicates the location of the TE in the genomic niche, with TE content, gene content and GC content for the surrounding windows. The third column indicates resistance levels towards azole antifungals for isolates with and without the TE insertion.

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.

Figure supplement 1. Isolates (Table).

Figure 6: Global population structure of transposable element (TE) insertion polymorphism. (A) Total TE copies per isolate. Colors identify TE superfamilies. (B) TE copies per family and (C) superfamily. (D) TE insertion frequency spectrum per population. The curve fitting was performed with a self-starting Nls asymptomatic regression model (E). TE family copy numbers per isolate.

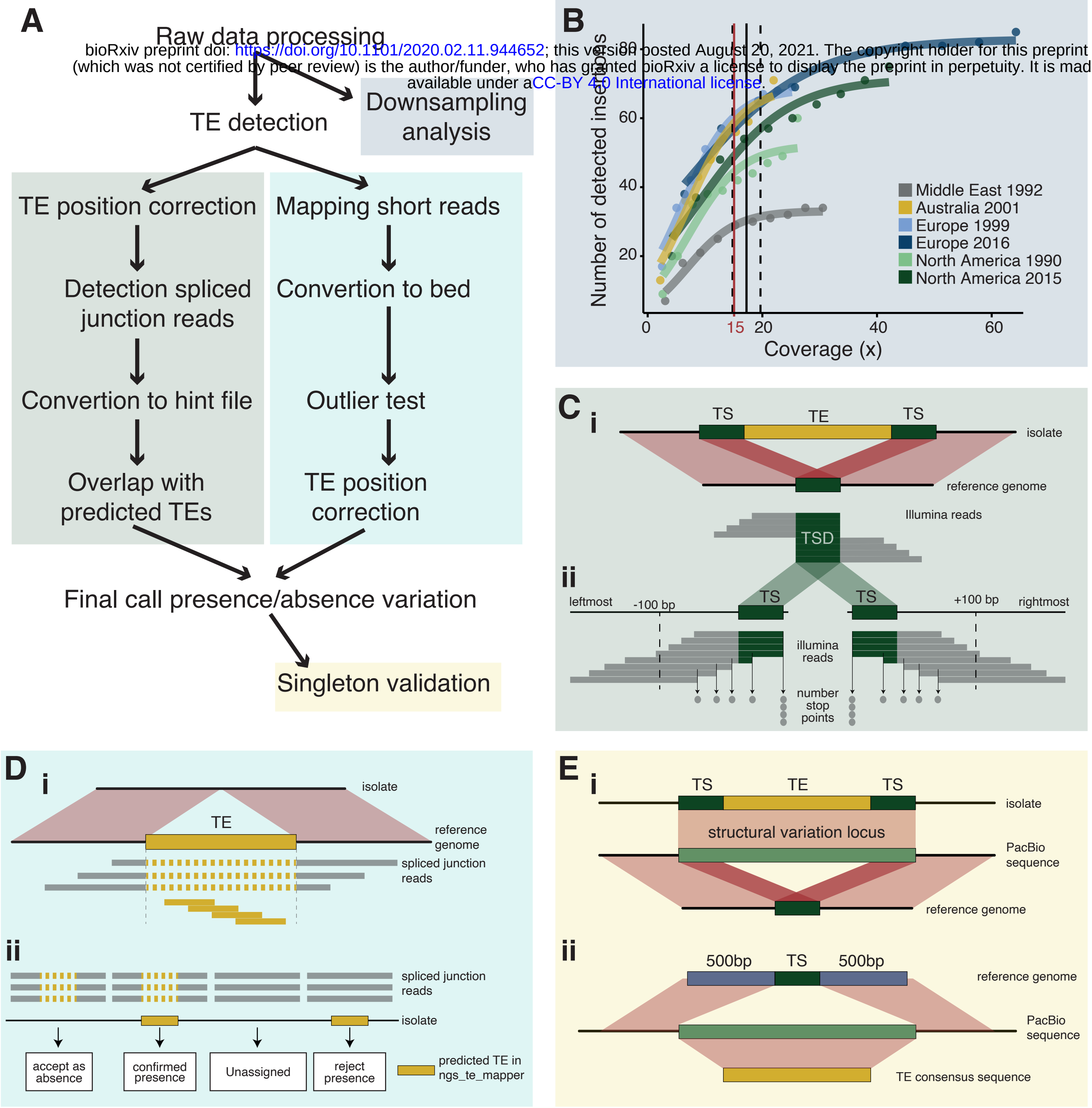
Figure supplement 1. Population changes additional. Variation in transposable element (TE) content per isolate across populations. (A) Total TE copies per superfamily (colored) and per isolate only including LTR (long terminal repeat) TEs *Copia* and *Gypsy*. Color indicates the family. (B) Total TE copies per superfamily (colored) and per isolate only on the core chromosomes. (C) Total TE copies per superfamily (colored) and per isolate only on the accessory chromosomes.

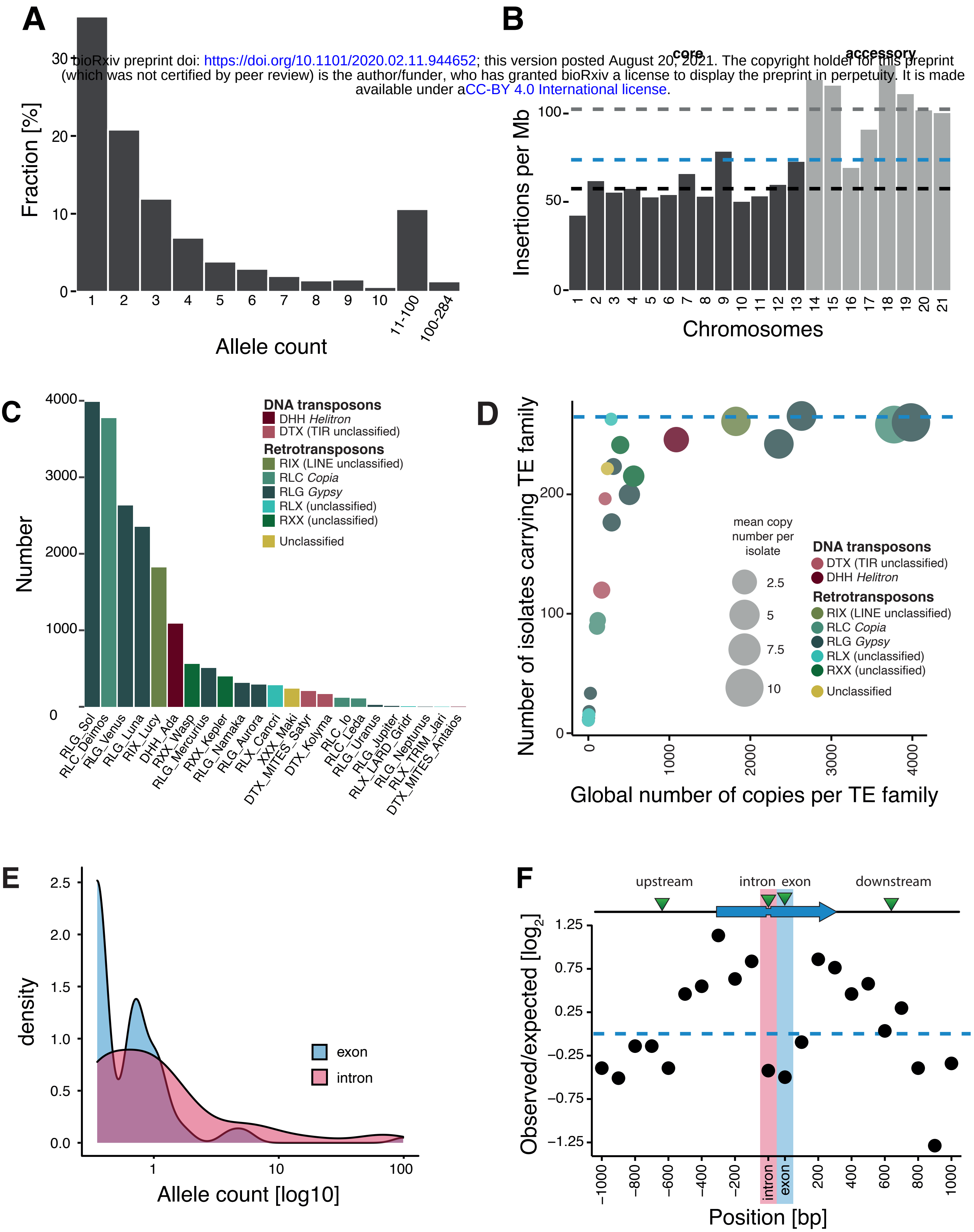
Figure supplement 2. Kolmogorof-Smirnov (Table).

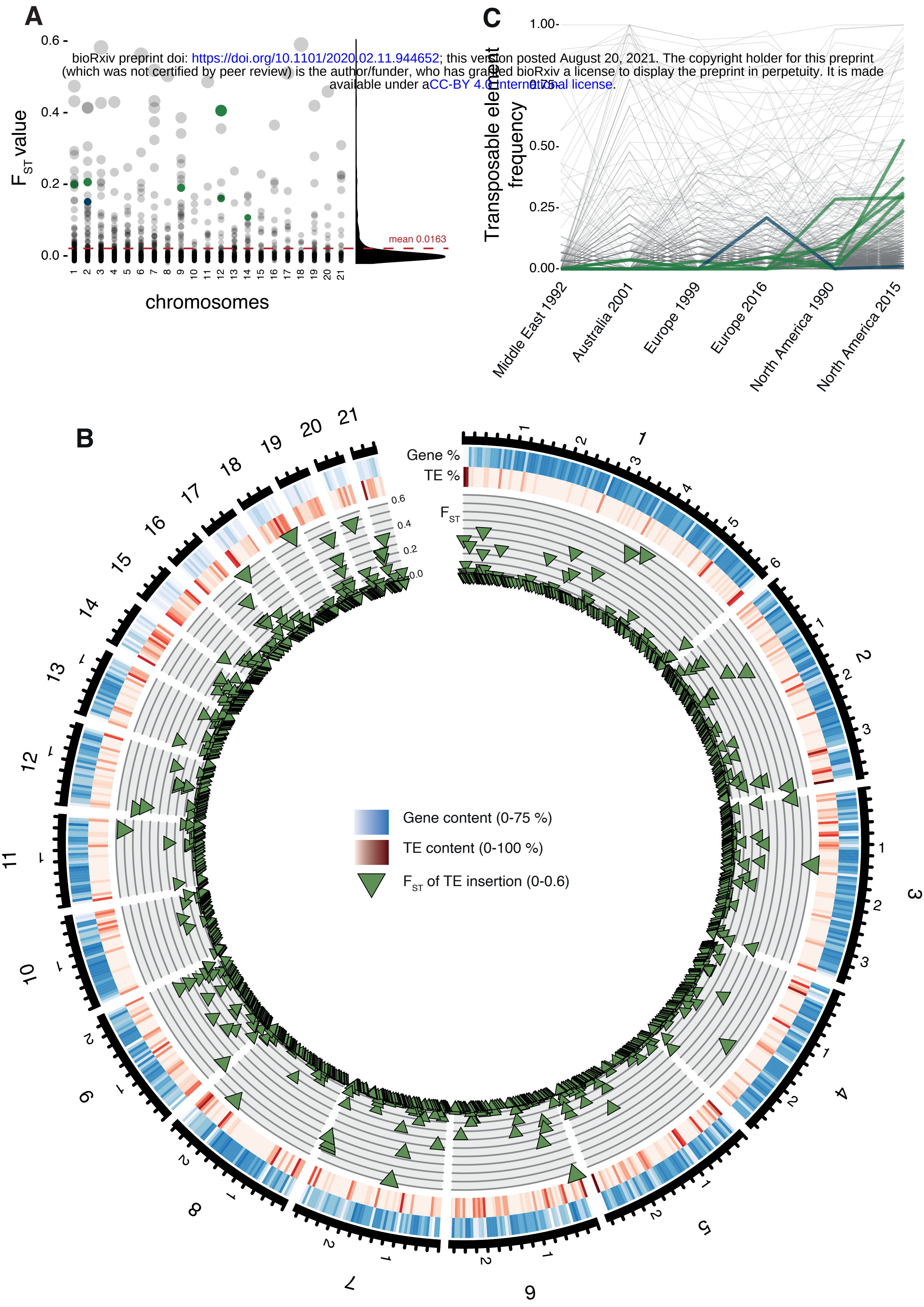
Figure supplement 3. Heatmap loci. (A) Presence (blue) and absence (yellow) matrix for all transposable element (TE) loci in all isolates per population. Colors on the left side indicate the superfamily. (B) Comparison of different genomic regions with and without TE insertions in IPO323.

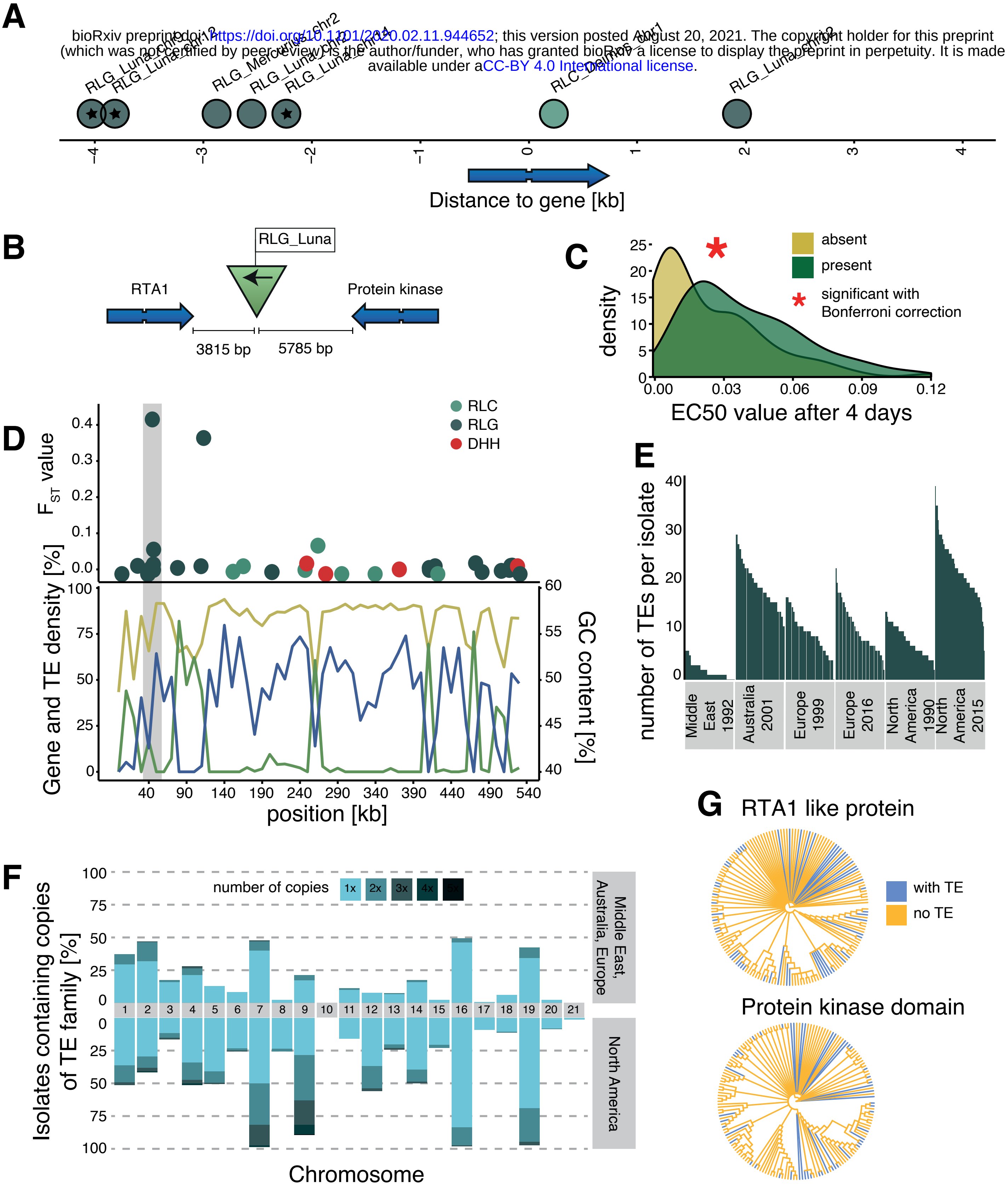
Figure 7: Core genome size and transposable element (TE) evolution across populations. (A) BUSCO completeness variation among genome assemblies. Black lines indicate the mean genome size per population. (B) Genome-wide GC content variation. (C) Core genome size variation among the isolates of the populations (excluding accessory chromosomes). (D) Correlation of core genome size and number of detected TEs. (E) Correlation of core genome size and the cumulative length of all TEs detected as inserted. (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.

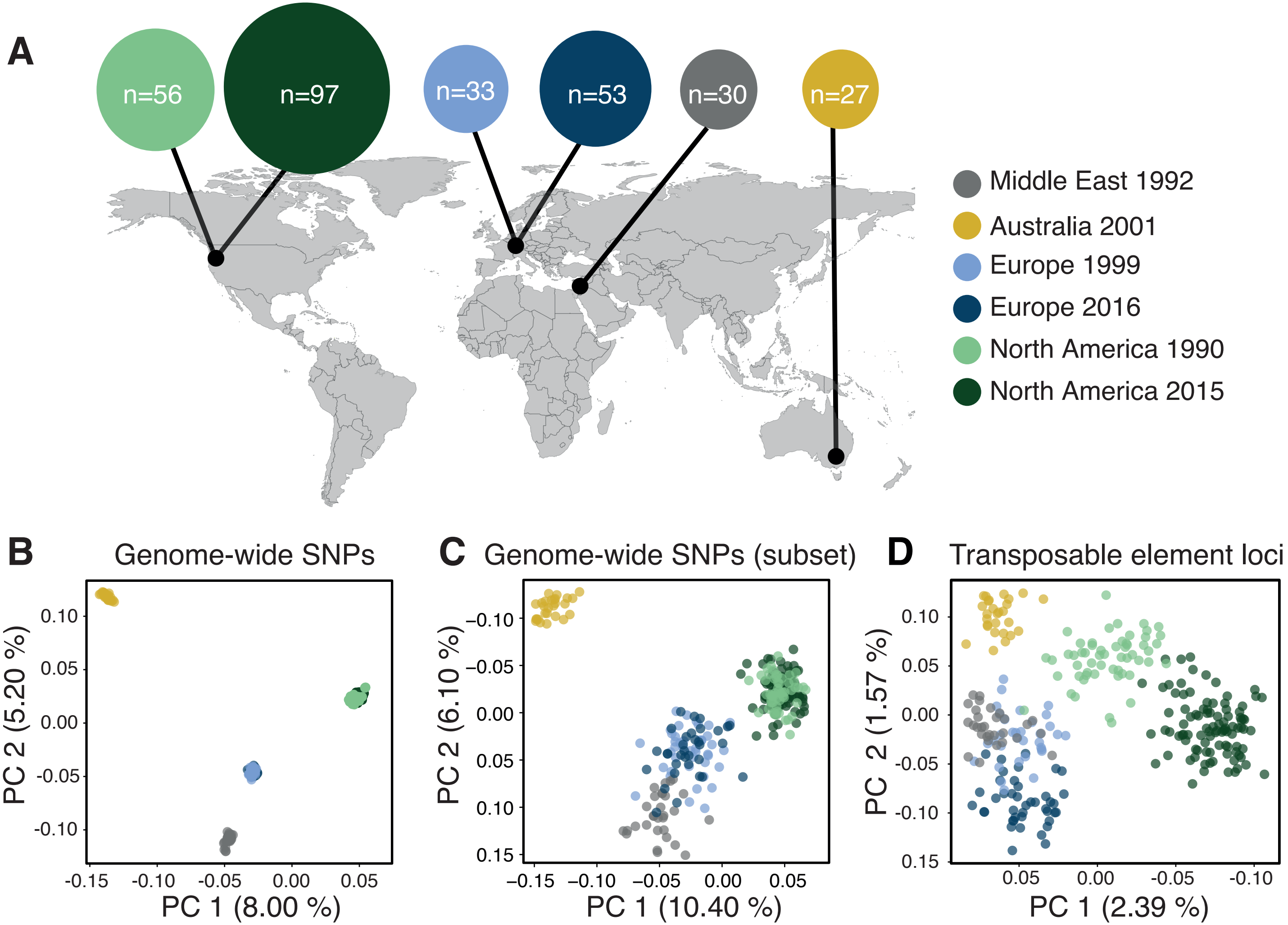
Figure supplement 1. Genome size expansion. (A) Estimated length of TE insertions per isolate and population. (B) Genome size variation per population. (C) Percentage of TEs content variation compared to the variation in genome size. (D) TE contributions to genome size variation compared to full genome size.

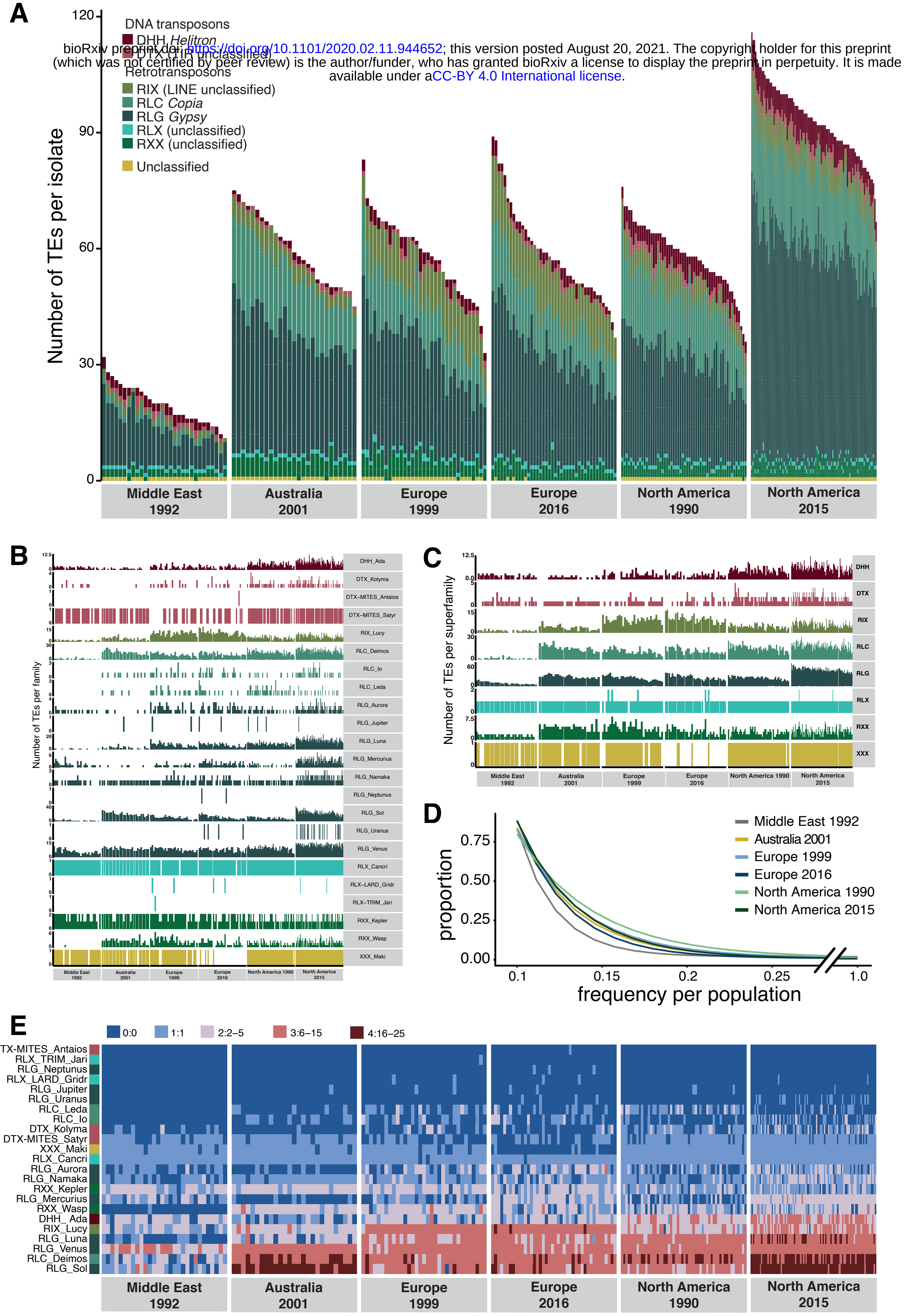

A
Raw data processing
TE detection
Downsampling analysis
TE position correction
Detection spliced junction reads
Convertion to hint file
Overlap with predicted TEs
Mapping short reads
Convertion to bed
Outlier test
TE position correction
Final call presence/absence variation
Singleton validation
B
Number 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
isolate
accept as absence
confirmed presence
Unassigned
reject presence
predicted TE in ngs_te_mapper
E
i
TS
TE
TS
isolate
structural variation locus
PacBio sequence
reference genome
ii
500bp
TS
500bp
reference genome
PacBio sequence
TE consensus sequence

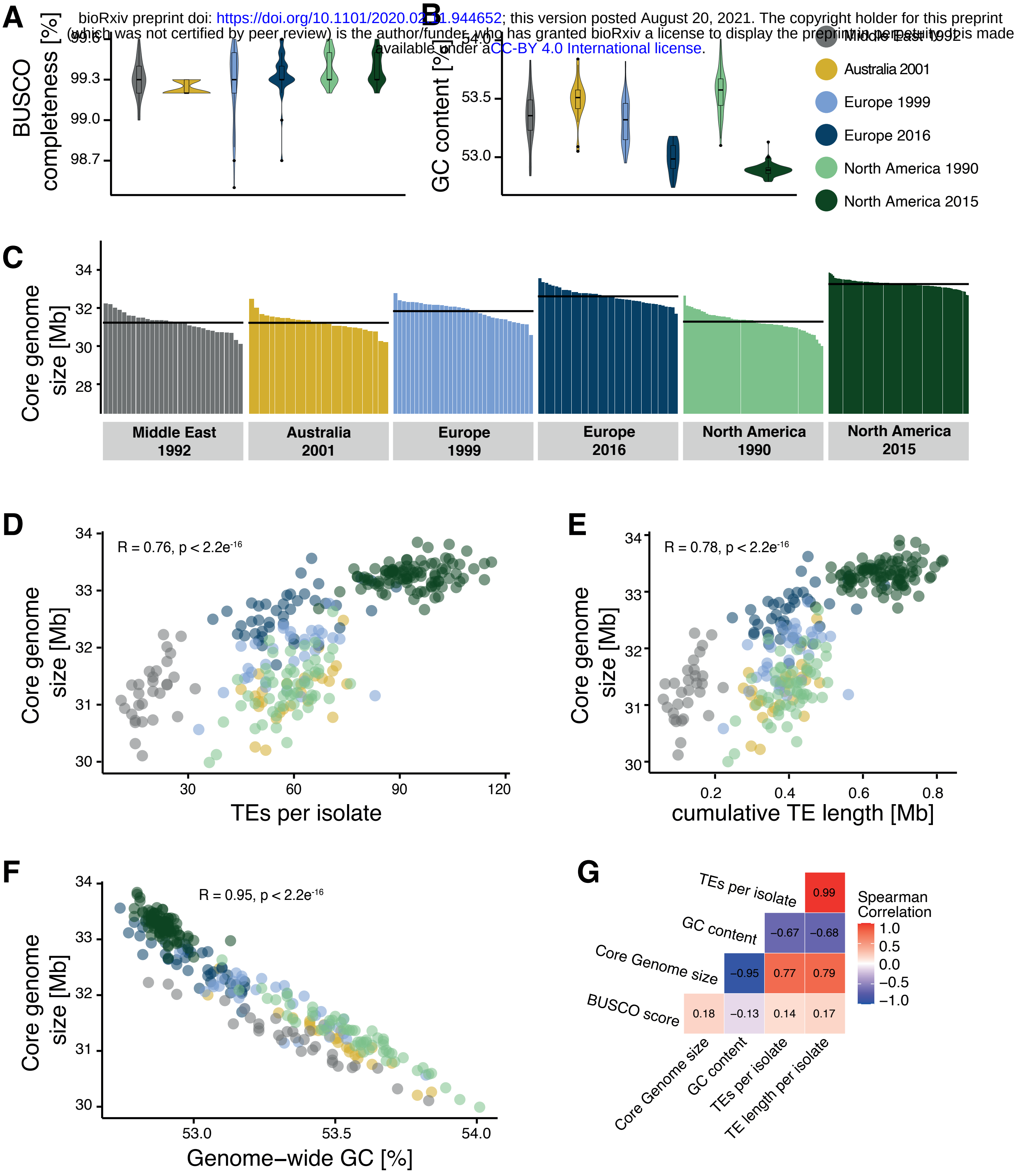

A
Fraction [%]
0
10
20
30
1
2
3
4
5
6
7
8
9
10
11-100
100-284
Allele count
B
core
accessory
Insertions per Mb
0
50
100
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
Chromosomes
C
Number
0
1000
2000
3000
4000
DNA transposons
DHH Helitron
DTX (TIR unclassified)
Retrotransposons
RIX (LINE unclassified)
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
D
Number of isolates carrying TE family
0
100
200
0
1000
2000
3000
4000
Global number of copies per TE family
mean copy number per isolate
2.5
5
7.5
10
DNA transposons
DTX (TIR unclassified)
DHH Helitron
Retrotransposons
RIX (LINE unclassified)
RLC Copia
RLG Gypsy
RLX (unclassified)
RXX (unclassified)
Unclassified
E
density
0.0
0.5
1.0
1.5
2.0
2.5
1
10
100
Allele count [log10]
exon
intron
F
upstream
intron
exon
downstream
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]


A
F_{ST} value
0.6
0.4
0.2
0.0
mean 0.0163
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
chromosomes
C
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
B
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
RLG_Luna_chr3
RLG_Luna_chr12
RLG_Mercurius_chr2
RLG_Luna_chr2
RLG_Luna_chr14
RLC_Deimos_chr1
RLG_Luna_chr12
−4
−3
−2
−1
0
1
2
3
4
Distance to gene [kb]
B
RLG_Luna
RTA1
Protein kinase
3815 bp
5785 bp
C
density
25
20
15
10
5
0
0.00
0.03
0.06
0.09
0.12
EC50 value after 4 days
absent
present
significant with Bonferroni correction
D
FST value
0.4
0.3
0.2
0.1
0.0
RLC
RLG
DHH
Gene and TE density [%]
100
75
50
25
0
GC content [%]
60
55
50
45
40
40
90
140
190
240
290
340
390
440
490
540
position [kb]
E
number of TEs per isolate
40
30
20
10
0
Middle East 1992
Australia 2001
Europe 1999
Europe 2016
North America 1990
North America 2015
F
Isolates containing copies of TE family [%]
100
75
50
25
0
number of copies
1x
2x
3x
4x
5x
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
RTA1 like protein
with TE
no TE
Protein kinase domain


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
PC 2 (5.20 %)
0.10
0.05
0.00
-0.05
-0.10
-0.15
-0.10
-0.05
0.00
0.05
PC 1 (8.00 %)
C
Genome-wide SNPs (subset)
PC 2 (6.10 %)
−0.10
−0.05
0.00
0.05
0.10
0.15
−0.15
−0.10
−0.05
0.00
0.05
PC 1 (10.40 %)
D
Transposable element loci
PC 2 (1.57 %)
0.10
0.05
0.00
-0.05
-0.10
0.15
0.05
0.00
-0.05
-0.10
PC 1 (2.39 %)


A
120
90
60
30
0
Number of TEs per isolate
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
B
Number of TEs per family
DHH_Ada
DTX_Kolyma
DTX–MITES_Antaios
DTX–MITES_Satyr
RIX_Lucy
RLC_Deimos
RLC_Io
RLC_Leda
RLG_Aurora
RLG_Jupiter
RLG_Luna
RLG_Mercurius
RLG_Namaka
RLG_Neptunus
RLG_Sol
RLG_Uranus
RLG_Venus
RLX_Cancri
RLX–LARD_Gridr
RLX–TRIM_Jari
RXX_Kepler
RXX_Wasp
XXX_Maki
Middle East 1992
Australia 2001
Europe 1999
Europe 2016
North America 1990
North America 2015
C
Number of TEs per superfamily
12.5
0.0
5
0
15
0
30
0
60
0
2
0
7.5
0
1
0
DHH
DTX
RIX
RLC
RLG
RLX
RXX
XXX
Middle East 1992
Australia 2001
Europe 1999
Europe 2016
North America 1990
North America 2015
D
proportion
0.75
0.50
0.25
0.00
0.1
0.15
0.2
0.25
1.0
frequency per population
Middle East 1992
Australia 2001
Europe 1999
Europe 2016
North America 1990
North America 2015
E
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


A
BUSCO completeness [%]
99.6
99.3
99.0
98.7
B
GC content [%]
54.0
53.5
53.0
Middle East 1992
Australia 2001
Europe 1999
Europe 2016
North America 1990
North America 2015
C
Core genome size [Mb]
34
32
30
28
Middle East 1992
Australia 2001
Europe 1999
Europe 2016
North America 1990
North America 2015
D
R = 0.76, p < 2.2e-16
Core genome size [Mb]
34
33
32
31
30
30
60
90
120
TEs per isolate
E
R = 0.78, p < 2.2e-16
Core genome size [Mb]
34
33
32
31
30
0.2
0.4
0.6
0.8
cumulative TE length [Mb]
F
R = 0.95, p < 2.2e-16
Core genome size [Mb]
34
33
32
31
30
53.0
53.5
54.0
Genome−wide GC [%]
G
TEs per isolate
GC content
Core Genome size
BUSCO score
0.99
−0.67
−0.68
−0.95
0.77
0.79
0.18
−0.13
0.14
0.17
Core Genome size
GC content
TEs per isolate
TE length per isolate
Spearman Correlation
1.0
0.5
0.0
−0.5
−1.0