A 19-isolate reference-quality global pangenome for the fungal wheat pathogen *Zymoseptoria tritici*

Thomas Badet[1], Ursula Oggenfuss[1], Leen Abraham[1], Bruce A. McDonald[2], Daniel Croll[1, *]

[1] Laboratory of Evolutionary Genetics, Institute of Biology, University of Neuchâtel, Neuchâtel, Switzerland

[2] Plant Pathology, Institute of Integrative Biology, ETH Zürich, Zürich, Switzerland

[*] Corresponding author: daniel.croll@unine.ch

Abstract

Background: The gene content of a species largely governs its ecological interactions and adaptive potential. A species is therefore defined by both core genes shared between all individuals and accessory genes segregating presence-absence variation. There is growing evidence that eukaryotes, similar to bacteria, show intra-specific variability in gene content. However, it remains largely unknown how functionally relevant such a pangenome structure is for eukaryotes and what mechanisms underlie the emergence of highly polymorphic genome structures.

Results: Here, we establish a reference-quality pangenome of a fungal pathogen of wheat based on 19 complete genomes from isolates sampled across six continents. *Zymoseptoria tritici* causes substantial worldwide losses to wheat production due to rapidly evolved tolerance to fungicides and evasion of host resistance. We performed transcriptome-assisted annotations of each genome to construct a global pangenome. Major chromosomal rearrangements are segregating within the species and underlie extensive gene presence-absence variation. Conserved orthogroups account for only ~60% of the species pangenome. Investigating gene functions, we find that the accessory genome is enriched for pathogenesis-related functions and encodes genes involved in metabolite production, host tissue degradation and manipulation of the immune system. *De novo* transposon annotation of the 19 complete genomes shows that the highly diverse chromosomal structure is tightly associated with transposable elements content. Furthermore, transposable element expansions likely underlie recent genome expansions within the species.

Conclusions: Taken together, our work establishes a highly complex eukaryotic pangenome providing an unprecedented toolbox to study how pangenome structure impacts crop-pathogen interactions.

Background

Microbial species harbor substantial functional diversity at the level of gene presence-absence variation (Tettelin *et al.* 2008). Genes not fixed within a species (*i.e.* accessory genes) can account for a large fraction of the full gene repertoire (*i.e.* the pangenome). In bacteria, the proportion of core genes in the pangenome can range from 5-98% and challenge taxonomic classifications (Ramasamy *et al.* 2014; Rouli *et al.* 2015). The wide spectrum of pangenome sizes across species can be associated with the species distribution and lifestyle (McInerney *et al.* 2017). Species showing a wide geographical distribution and large population sizes characterized by frequent genetic exchange tend to have expansive, open pangenomes (Lefébure *et al.* 2010). In microbial pathogens, accessory genes play a major role in virulence and environmental adaptation (Jackson *et al.* 2011; Sánchez-Vallet *et al.* 2018a; Wu *et al.* 2018). The notion of a pangenome led to the discovery that major elements of intra-specific variation are often ignored in studies relying on a single reference genome. Large pangenomes also can challenge association studies aiming to identify the genetic basis of phenotypic traits because mapping is often performed against a single reference genome, making potentially relevant genetic variation inaccessible (Marschall *et al.* 2016; Sánchez-Vallet *et al.* 2018b). Despite their importance for unravelling the genetic basis of adaptive evolution, only a very limited number of eukaryotic species have well established pangenomes.

Copy number variation including gene deletion generates intraspecific gene content variation in nearly all species (Schrider & Hahn 2010). This variation can create extreme variance in fitness and promote adaptive evolution (Araki *et al.* 2006; Brynildsrud *et al.* 2016; Plissonneau *et al.* 2016a; Hartmann *et al.* 2018). In plant pathogens, the ability to infect a host often relies on the secretion of effector proteins that interfere with the host cell machinery (Wit De *et al.* 2009; Lo Presti *et al.* 2015; Toruño *et al.* 2016). Host plants evolved cognate resistance proteins that are able to recognize effector proteins and trigger immunity (Jones & Dangl 2006). Gains and losses of effector genes can therefore have a major impact on the outcome of host-pathogen interactions and challenge food security.

Recent studies on fungal pathogens highlighted that genes showing presence-absence variation are enriched for predicted effectors (Yoshida *et al.* 2016; Hartmann & Croll 2017; Hartmann *et al.* 2018). Effectors and transposable elements (TEs) are often tightly associated with fast-evolving compartments of the genome (Sperschneider *et al.* 2015; Faino *et al.* 2016), also known as the "two-speed" genome architecture (Dong *et al.* 2015). However, how TEs impact the birth and death of effectors in fast-evolving compartments remains largely unclear (Fouché *et al.* 2018a; Sánchez-Vallet *et al.* 2018a). The construction of pathogen pangenomes enabled crucial insights into functional diversity and the evolutionary trajectories of host adaptation. Recent pangenome analyses of four fungal species including opportunistic pathogens revealed that between ~9-19% of the pangenome is accessory. Accessory gene localization was preferentially in subtelomeric regions, suggesting both a mechanistic link to repeat-rich regions and relaxation of selective constraints (McCarthy & Fitzpatrick 2019). The wheat pathogen *Zymoseptoria tritici* was found to have one of the largest eukaryotic pangenomes with an estimate of at least 42% of all genes being accessory (Plissonneau *et al.* 2018). However, eukaryotic pangenomes remain shallow and are often based on not fully resolved chromosomal sequences.

Fungal plant pathogens such as *Z. tritici* show extreme cases of genome plasticity. The reference genome of *Z. tritici* has 21 chromosomes, of which eight are accessory and segregate presence-absence variation in populations (Goodwin *et al.* 2011). The pathogen rapidly evolved virulence on resistant wheat cultivars and has overcome all current fungicides (Cools & Fraaije 2008; Lucas *et al.* 2015; Blake *et al.* 2018). Host adaptation was driven among other factors by the rapid deletion of an effector gene and structural rearrangements (Hartmann *et al.* 2017; Krishnan *et al.* 2018; Meile *et al.* 2018). Pathogen populations are highly diverse with high rates of recombination (Croll *et al.* 2015; Stukenbrock & Dutheil 2018; Grandaubert *et al.* 2019). Meiosis can trigger large chromosomal rearrangements and lead to aneuploid chromosomes in the species (Croll *et al.* 2013; Fouché *et al.* 2018b). A pangenome constructed for five *Z. tritici* isolates revealed that chromosome length variation segregating within populations was mainly due to the presence-absence variation of large TE clusters (Plissonneau *et al.* 2016b, 2018). Furthermore, accessory genes tended to form clusters

dispersed along chromosomes. Accessory genes also tended to be in closer proximity to TEs than core genes and were therefore more likely to be affected by epigenetic silencing (Plissonneau *et al.* 2018). However, the constructed pangenome was very likely incomplete given the fact that four of the genomes originated from isolates collected in the same year from two nearby fields. Furthermore, accessory genes were enriched for pathogenesis-related functions but the pangenome size did not reach saturation. Given the global impact of the pathogen and the importance of accessory genes for adaptive evolution, a comprehensive pangenome capturing worldwide genetic diversity is essential.

In this study, we constructed the pangenome of *Z. tritici* by including 19 isolates sampled from six different continents and covering the global distribution of the pathogen. We find major chromosomal rearrangements segregating within the species together with extensive presence-absence variation for a range of pathogenicity-related gene functions. We also found major shifts in the TE content across the species. Our data represent the largest eukaryotic pangenome to date based on complete genome assemblies, providing an unprecedent toolbox for the analysis of adaptive evolution of pathogens.

Results

Major chromosomal rearrangements segregating within the species

We constructed a global pangenome of *Z. tritici* based on 19 isolates sampled from six continents and 13 different countries (Figure 1A). The isolates included the previously described reference isolate IPO323 sampled in the Netherlands and four isolates that were isolated from two nearby fields in Switzerland (Goodwin *et al.* 2011; Plissonneau *et al.* 2016b, 2018). The geographic regions of origin of the 19 isolates recapitulate a significant environmental gradient in mean annual temperature and humidity and span the distribution range of the species. The sampling period ranges from 1984 (IPO323) to 2010 (CRI10). Fungicide applications against *Z. tritici* became widespread in the 90s and early 2000s, hence the sampling covers both pre- and post-fungicide treatment regimes. We

sequenced long-read PacBio SMRTbell libraries to a depth of 40-110X and ~20 kb read coverage in order to generate chromosome-level assemblies. Assembly sizes ranged from 37.13 Mb (IR01_48b) to 41.76 Mb (Aus01) (Figure 1A). We recovered all eight known accessory chromosomes of the species but no additional chromosome. The accessory chromosome 18 is most often missing. Accessory chromosomes display an average of ~37% size variation among isolates and a maximum of 60% for chromosome 14 (Figure 1B). For core chromosomes, the average size variation accounts for 16% of chromosome length going up to 23% for chromosome 7.

A

isolate name	origin [country]	genome size (bp)	GC content [%]	TE content [%]	accessory chromosomes	gene count	BUSCO score [%]	accessory genome [%]
1A5	Switzerland	39698711	51.6	18.7		12092	98.3	23.2
1E4	Switzerland	38628401	51.8	17.7		12033	98.1	22.8
3D1	Switzerland	40657500	51.5	19.9		12006	98.5	22.5
3D7	Switzerland	37861420	52.1	18.4		11737	96.8	20.7
Arg00	Argentina	41149037	51.4	20.7		12347	98.4	24.8
Aus01	Australia	41760570	51.6	24.0		12787	98.1	27.6
CH95	Chile	39253131	51.7	19.6		12198	98.7	24.2
CNR93	Canada	38169704	52.1	19.2		12193	98.5	24.2
CRI10	Czech Republic	39697126	51.8	19.6		12265	98.1	24.4
I93	USA (Indiana)	39405809	51.5	19.1		12263	98.4	24.6
IPO323	Netherlands	39686251	51.7	19.1		11839	97.9	22.0
IR01_26b	Iran	38551850	51.5	16.5		11940	98.6	22.6
IR01_48b	Iran	37133562	52.0	18.1		11865	98.5	22.1
ISY92	Israel	39679942	51.1	19.1		12060	98.3	23.4
KE94	Kenya	38152782	51.9	17.3		11959	98.5	22.8
OregS90	USA (Oregon)	39591316	52.1	20.3		12483	98.4	26.0
TN09	Tunisia	37907319	51.4	18.8		11657	98.5	20.6
UR95	Ukraine	38369786	51.7	17.7		12241	98.5	24.4
YEQ92	Yemen	39871609	51.0	19.8		11734	98.3	21.3

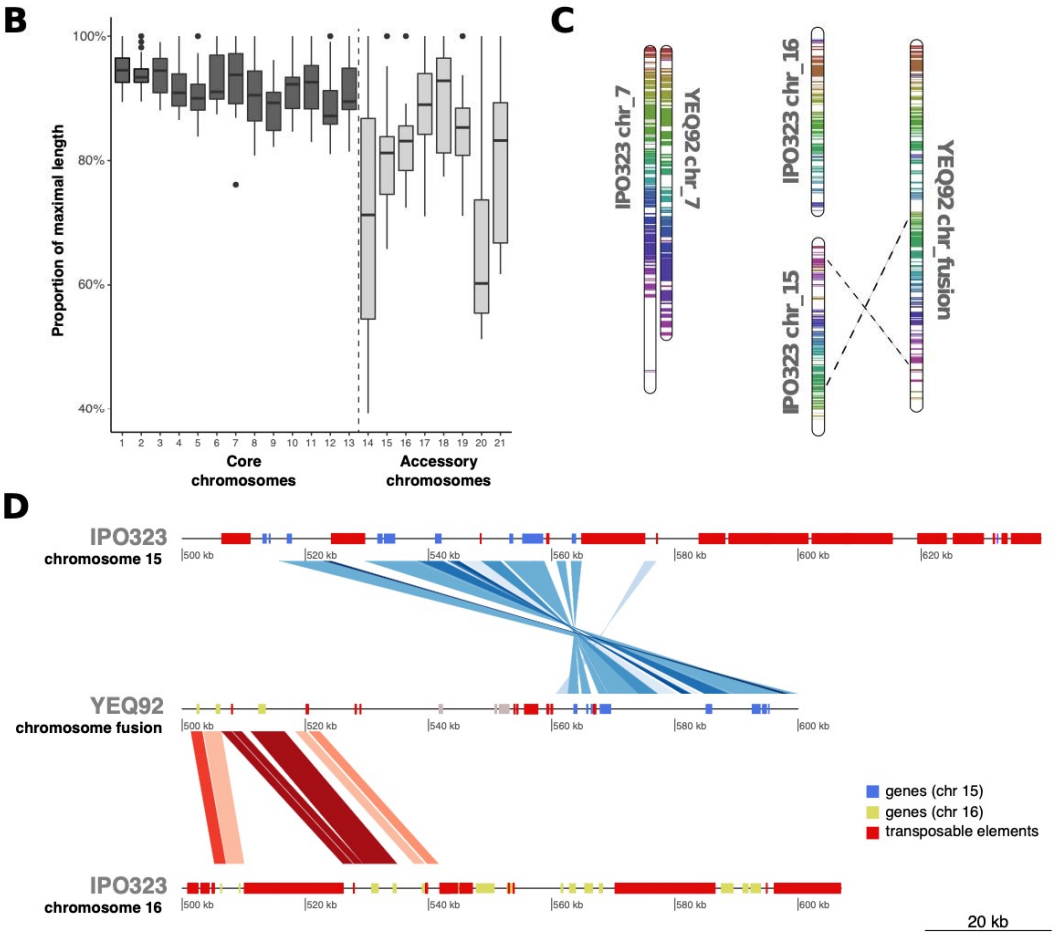


Figure 1: Large segregating chromosomal rearrangements. A. Summary of genome assembly characteristics for the 19 isolates. The bars represent the range of minimum to maximum values. Accessory chromosomes are shown from chromosome 14-21 with green dots for present and red dots

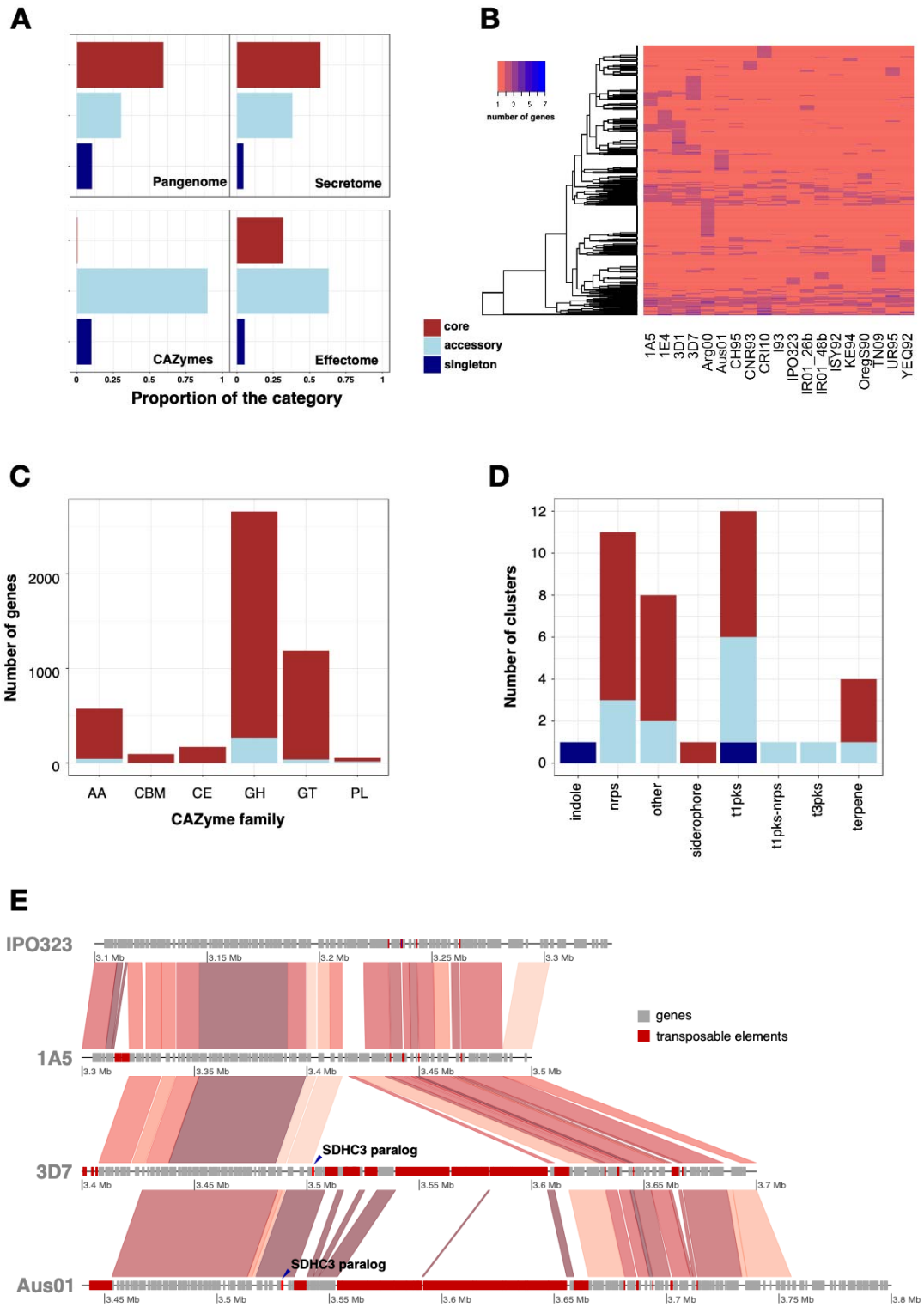
for missing chromosomes. The two linked dots for isolate YEQ92 represent a chromosomal fusion. **B**. Chromosome size variation. **C**. Two large chromosomal rearrangements present in YEQ92. Colors indicate macro-syntenic regions between YEQ92 and IPO323 chromosome homologs. The inversion of chromosome 15 is represented by crossing dotted lines. **D**. Chromosomal fusion of 15 and 16 identified in YEQ92 and compared to the reference genome IPO323. Transposons are shown in red, genes from chromosome 15 in purple, genes from chromosome 16 in green and genes specific to the fusion in grey boxes, respectively. Inversions are shown in blue.

We identified a major deletion spanning 406 kb and encompassing 107 genes on the right arm of core chromosome 7 of the Yemeni isolate (YEQ92; Figure 1C). The same isolate had chromosome 15 fused to the right arm of chromosome 16. The fusion event is supported by aligned PacBio reads spanning the region between the two chromosomal segments (Figure S1). The resulting chromosome is 1.20 Mb long and 49.5 kb shorter than the sum of the homologous chromosomes 15 and 16 of the IPO323 reference genome. Approximately 90% of the genes on the IPO323 chromosome 15 and 16 belong to accessory orthogroups, as they lack an ortholog in at least one of the other isolates. We find that the chromosomal fusion deleted about 150 kb affecting 1 and 12 genes on chromosomes 15 and 16, respectively (Figure 1D). We further assessed genome completeness using BUSCO analyses. All genomes exceed the completeness of the fully finished IPO323 reference genome (97.9%) with the exception of isolate 3D7 (96.8%; Figure 1A).

Substantial gene content variation across the pangenome

We generated RNAseq data to identify high-confidence gene models in all 14 newly assembled genomes based on a splice-site informed gene prediction pipeline. The total gene count varied between 11’657 and 12’787 gene models (Figure 1A). We assigned all genes to orthogroups using protein homology and constructed a pangenome of all 19 complete genomes. The pangenome consists of a total of 229’699 genes assigned to 15’474 orthogroups. The number of genes assigned per orthogroup varies among isolates (Figure 2B). Approximately 99.8% of all orthogroups (15’451) are single gene orthogroups and ~60% of all orthogroups are shared among all 19 isolates (9’193 core orthogroups). Around 96% of the core orthogroups (8’829 out of 9’193) have conserved gene copy numbers among isolates. Furthermore, we find that 30% of all orthogroups are shared between some

158 but not all genomes (4'690 accessory orthogroups) and 10% of the orthogroups are composed of
159 genes found in a single genome only (1'592 singletons; Figure 2A-B; Supplementary Table 1).



160

Figure 2: Pangenome diversification across gene categories. A. Pangenome proportions across all genes (upper-left), secreted (upper-right), carbohydrate-active enzymes (CAZymes; lower-left) and effectors (lower-right). **B**. Gene copy number variation in core orthogroups across the 19 genomes. **C**. Pangenome gene count across six CAZyme families. Families are divided into glycoside hydrolase (GH), glycosyl transferase (GT), auxiliary activity (AA), carbohydrate esterase (CE), carbohydrate-binding modules (CBM) and polysaccharide lyase activity (PL) categories. **D**. Pangenome categories of secondary metabolite gene clusters.

To infect wheat, *Z. tritici* relies on specific gene functions (Steinberg 2015; Palma-Guerrero *et al.* 2017). Effectors play a major role in establishing infection and exploiting host resources. Hence, we analysed how gene functions were structured across the pangenome components. Core orthogroups showing variation in gene-copy number among isolates include five encoding predicted effectors. Both accessory proteins and overall effector proteins are less conserved than core proteins at the amino acid level (Supplementary Figure S2). A total of 3.5% (691) of all orthogroups encode at least one predicted effector. Among orthogroups encoding at least one predicted effector, 31% were conserved among all isolates (219), 63% were accessory (436) and 5% were found in only one isolate (36 singletons). Notably, 99% of the predicted effector genes are located on core chromosomes. In addition to effectors, enzymes enabling access to nutrients are important pathogenicity components. We identified a total of 4’742 annotated carbohydrate-degrading enzymes (CAZymes) clustered into 263 orthogroups. Notably, 92% of the orthogroups encoding CAZymes were conserved among all isolates (Figure 2A). CAZymes grouped into 123 subfamilies. Glycoside hydrolases (GH) are the largest family and account for 57% of all annotated CAZymes (151 orthogroups for 2’717 genes). Glycosyl transferases (GT) are the second most abundant family with 1’188 genes and 66 orthogroups (25% of all CAZymes) (Figure 2C). We also identified 33 orthogroups encoding for auxiliary activities (AA), 9 for carbohydrate esterase activity (CE), 6 for carbohydrate-binding modules (CBM) and 3 for polysaccharide lyase activity (PL). The PL family includes 29% accessory genes. Across CAZyme families, 0-10% of the genes are accessory (Figure 2C). We found a singleton GH43 subfamily gene in the genome of the Australian isolate (Aus01).

The production of secondary metabolites contributes significantly to virulence and competitive abilities of fungal pathogens. We identified between 29 and 33 secondary metabolite gene clusters per genome depending on the isolate. A total of 70% of all genes predicted as components of a biosynthetic gene cluster are conserved between all isolates and 30% are accessory (Figure 2D, Figure S3). We identified 39 syntenic gene clusters in the pangenome classified into 12 type 1-polyketide synthase (PKS), 11 non-ribosomal peptide synthetase (NRPS), four terpene, one type 3-PKS, one siderophore, one indole and eight unclassified clusters. Sixteen (40%) of the identified syntenic clusters show presence-absence variation. In the CH95 isolate, a gene cluster on chromosome 7 was annotated as unclassified but annotated as a NRPS in 17 other isolates and absent from the IPO323 reference genome. The sole indole and type 1-PKS clusters located on chromosomes 5 and 10, respectively, were only found in isolate TN09. Two type 1-PKS and one NRPS cluster were missing in the isolates YEQ95, Aus01 and IPO323, respectively. Among the 39 identified syntenic gene clusters, 23 included a predicted effector and nine included a gene annotated as a cell-wall degrading enzyme.

The emergence of fungicide tolerance in *Z. tritici* is a major threat to wheat production. Succinate dehydrogenase (SDH) inhibitors are commonly used as control agents (Lucas *et al.* 2015; Steinhauer *et al.* 2019). We identified five SDH orthologs, of which three were conserved among all genomes (SDHB, SDHC and SDHD subunits). We find two distinct SDHC paralogs SDHC2 and SDHC3 in eleven and two isolates, respectively. The SDHC3 paralog conferring standing resistance to SDH inhibitors is located flanking a large cluster of TEs, suggesting that chromosomal rearrangements were underlying the paralog emergence (Figure 2E). Genes encoding major facilitator superfamily (MFS) transporters, which can confer multidrug resistance in *Z. tritici* (Omrane *et al.* 2017), grouped into 336 orthogroups for a total of 5’787 genes (Supplementary Table 2). We find that 39 (11%) of these orthogroups are part of a predicted secondary metabolite gene cluster and one is an annotated CAZyme from the GH78 family. Overall, the results reveal that gene families essential for pathogenicity and fungicide resistance show unexpectedly high levels of presence-absence variation in the *Z. tritici* pangenome.

Strong expression variation across major gene functions

Differential gene expression is a major driver of intraspecific phenotypic differences. We performed mRNA-sequencing of all 19 isolates grown on minimal media. Minimal media induces filamentous growth of *Z. tritici*, mimicking the morphology and nutrient starvation that occurs early during plant infection. We investigated isolate-specific gene expression by self-mapping RNAseq reads to each isolate's genome assembly. Overall, 91.3% of the genes show expression on minimal media and 68% have expression of more than 10 counts per million (CPM) (Figure 3A). Core genes have higher expression than accessory genes (Figure S4). Among the genes showing no expression on minimal media, 501 are predicted effector genes (8% of predicted effectors), 93 are predicted CAZymes (2% of CAZymes) and 838 are members of a predicted gene cluster (10% of all gene cluster genes). CAZymes are overall highly expressed on minimal media (~77% with CPM >10) when compared to effectors (~45% with CPM >10) and gene cluster genes (~60% with CPM >10) (Figure 3A). About 53% of core single copy orthogroups with non-zero expression have a coefficient of variation >50% (Figure 3B). Similarly, ~68% of CAZymes and ~60% of genes that are part of a secondary metabolite cluster have expression coefficient of variation > 50%. In contrast, about 90% of orthogroups encoding predicted effectors have a coefficient of variation >50%, together with ~81% of accessory orthogroups.

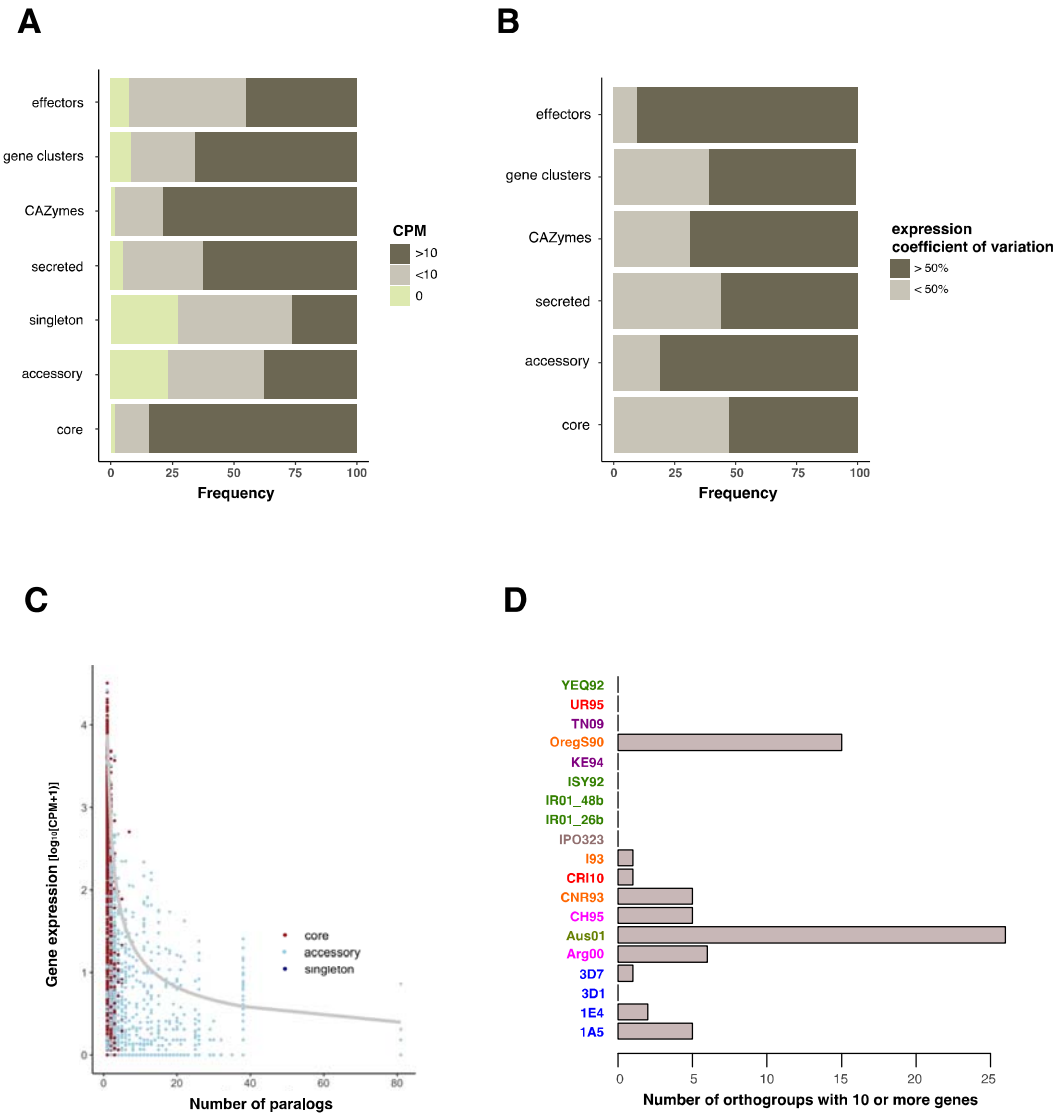


Figure 3: Expression polymorphism as a function of the pangenome. A. Proportion of genes showing expression >10 counts per million (CPM) across genes categories. **B**. Proportion of orthogroups for which the expression coefficient of variation is >50% [*cov* = *sd* (CPM) / *mean* (CPM)] among different gene categories. As is (A) we show results for orthogroups with annotated effectors, secondary metabolite cluster genes (gene cluster), carbohydrate-active enzymes (CAZymes), secreted genes, or annotated as singleton, accessory and core in the pangenome. **C**. Variation of single gene expression with the number of paralogs per genome. The grey line shows the logarithmic regression calculated as the linear model log_{10}(CPM+1) ~ log_{10}(number of paralogs). **D**. Number of orthogroups with 10 or more paralogs per genome. Isolates are colored by continent of origin.

To identify broad patterns in the pangenome expression landscape, we performed a clustering analysis of all core single gene orthogroups. We find that expression clustering does not reflect the

252 geographical origin or genetic distance with the exception of the four Swiss isolates (1A5, 1E4, 3D1
253 and 3D7; Figure S5). We also analysed the impact of copy-number variation on average expression
254 and find that single-copy orthologs are on average more highly expressed. In addition, we show that
255 gene expression rapidly decreases if an orthogroup includes 2-8 paralogs (Figure 3C).
256
257 **A highly variable transposable element content within the species**
258 TEs are drivers of pathogen evolution by generating adaptive genetic variation. To identify genes with
259 a potential role in the mobilisation of TEs, we analysed large homology groups. Among the
260 orthogroups with 10 or more paralogs, ~88% of the genes encode proteins without homology in
261 databases, ~7% of the genes encode nucleic acid binding functions (GO:0003676), ~2% of the genes
262 encode a retrotransposon nucleocapsid (GO:0000943) and ~1.5% of the genes encode a DNA
263 integration domain (GO:0015074). Orthogroups with 10 or more paralogs are all accessory. For
264 isolates sharing the same large orthogroups, we identified variability in the gene copy number within
265 those orthogroups. Indeed, the isolates Aus01 and OregS90 have 26 and 16 orthogroups, respectively,
266 with more than 10 assigned genes. The isolates I93 and Arg00 count between one and six orthogroups
267 and nine other isolates have no orthogroups larger than ten genes (Figure 3D). Altogether, these
268 results show that large orthogroups (>10 genes) essentially regroup genes that belong to TEs. Our data
269 also indicates regional TE-driven genome expansions given the enlarged genome sizes in Australian
270 and North American isolates.
271
272 To elucidate the role of transposition on generating genomic variation, we screened the 19 genomes
273 for TE content. For this, we jointly analysed all complete genomes to exhaustively identify repetitive
274 DNA sequences. We identified a total of 304 high-quality TE family consensus sequences grouped
275 into 22 TE superfamilies. The GC-content of the consensus sequences is highly variable, ranging from
276 23-77% (Figure S6). On average, TE superfamilies have a GC-content lower than 50%, except for
277 unclassified SINE families (RSX; GC% ~50.6). The genomic TE content ranges from 16.48%
278 (IR01_26b) to 23.96% (Aus01) and is positively correlated with genome size (cor = 0.78, $p < 0.001$;
279 Figure 4D). Genome size correlates with genome-wide TE proportions on both core and accessory

chromosomes but is negatively correlated with the proportion of coding sequences (Figure S7, S8). The average length of individual TEs ranges from 102 to 51'298 bp (Figure S9). The largest element is an unclassified LTR (RLX_LARD_Thrym) on chromosome 7, the size of which ranges from 6'282 bp in CNR93 to 59'390 bp in ISY92. This particular LTR is present at the locus only in 18 isolates including ISY92, which has a fragmented secondary copy on chromosome 3. The RLX_LARD_Thrym insertion on chromosome 7 overlaps with the ribosomal DNA locus and showed far above average mapped PacBio read coverage (~250X).

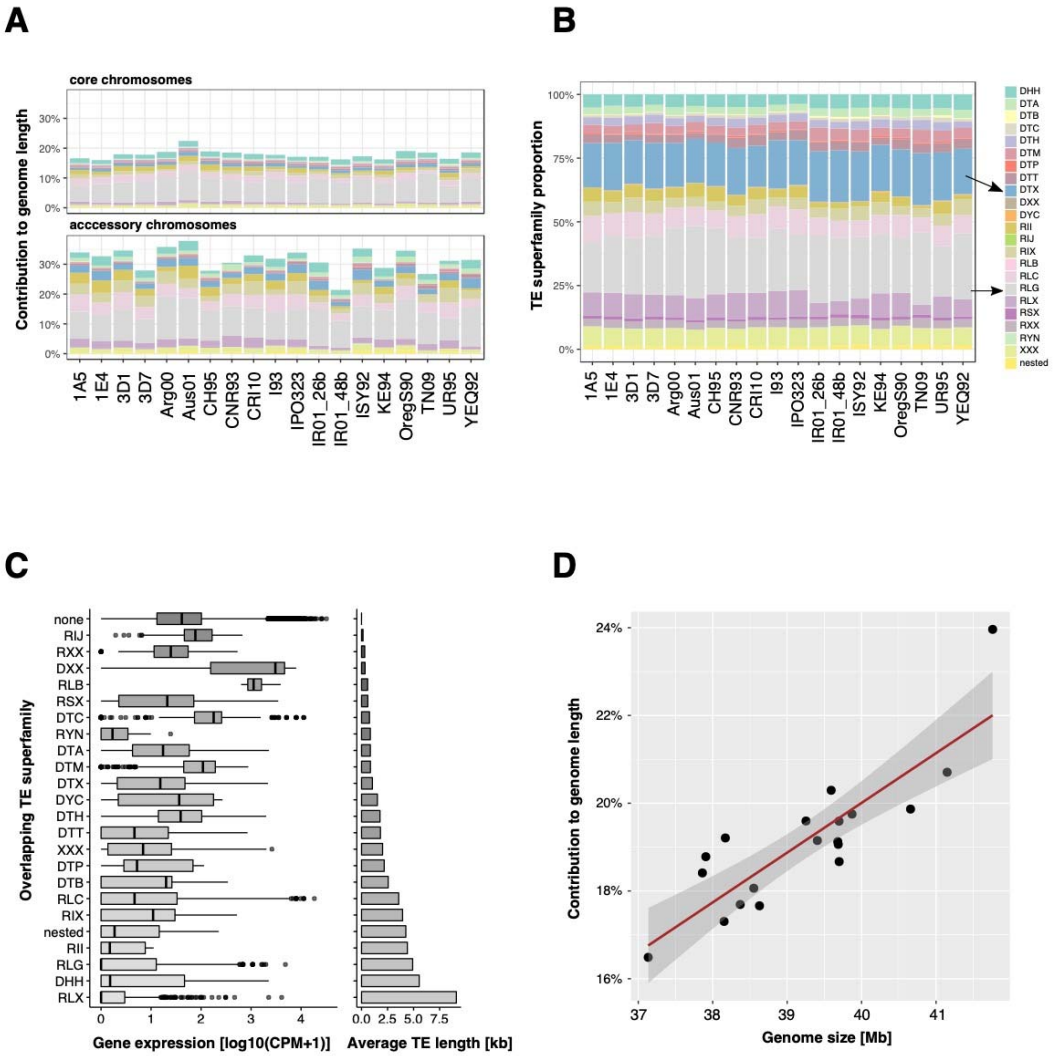


Figure 4: Transposable elements (TEs) contribute to *Z. tritici* genome size variation. A. Percentage of the core and accessory genomes covered by TEs across the 19 genomes. **B**. Relative frequency of the 23 TE superfamilies across all genomes. **C**. Relative impact of TE superfamilies on

genes affected by TE insertions (log_{10} CPM+1; left panel) and as a function of TE mean length (right panel). **D**. Correlation of genome-wide TE proportions with total genome size.

The genome-wide content of TEs shows substantial variation among the 19 isolates, however the relative abundance of different TE superfamilies is relatively conserved with LTR *Gypsy*, unclassified TIR and LTR *Copia* elements being the most frequent (Figure 4A-B). Accessory chromosomes contain consistently higher proportions of TEs compared to core chromosomes (26-41% versus 17-24%; Figure 4A). Aus01 and OregS90 isolates showed the highest TE content. Interestingly, the Aus01 genome shows LINE *I*, LTR *Gypsy* and LTR *Copia* family-specific expansion compared to other genomes. In contrast, the genome of OregS90 shows evidence for expansions of Helitron, LTR *Gypsy* and LTR *Copia* families. On average, 10% of all TEs overlap with genes. Overall, singleton and accessory genes tend to be closer to TEs and contain more often TE insertions than core genes (Figure S11-S12). The isolates Aus01 and OregS90 have 12.8% and 12.4% of all TEs overlapping with genes, respectively. In addition, Aus01 and OregS90 isolates have 7.4% and 5.4% of all genes that overlap with TEs, respectively (Figure S13). The composition of TEs inserted into genes reflects the overall TE composition in the genome, with more abundant TEs being more often inserted into genes (Figure S14). TEs can carry their own regulatory sequences and are often epigenetically silenced by the host. We found that orthogroups comprising a gene within 100 bp distance of a TE show stronger expression variation (~62% of orthogroups with a coefficient of variation >50%) compared to other orthogroups (~54% of orthogroups with a coefficient of variation >50%) (Figure S15). We also found that different TE superfamilies have contrasting effects on gene expression, with longer TEs having more drastic effects (Figure 4C). On average, genes with an inserted TE have lower expression levels (log10 CPM ~1.7-fold) and a higher coefficient of variation (log10 CPM ~2-fold) compared to genes without an inserted TE (Figure S16).

TE transcription correlates with relative frequency across isolates

Class I TEs replicate through an RNA intermediate and class II through a DNA intermediate. Nevertheless, class II TEs can also transcribe into RNA. To gain insights into the mechanisms of proliferation, we analysed the relative abundance of TE-derived transcripts across all genomes. The

highly repetitive nature of TEs typically prevents expression quantification at the individual copy level. Hence, we focused on normalized TE expression across all copies. Overall, more than 70% of the TE families have non-zero transcription levels. We find that the largest TE family, an unclassified LTR identified as RLX_LARD_Thrym, was the most transcribed with an average $\log_{10}$ CPM ~ 4.2 (Figure 5A). An unclassified DTX-MITE is the second most transcribed TE with an average $\log_{10}$ CPM ~ 3.6 followed by an unclassified TE (XXX_*Hermione* with an average $\log_{10}$ CPM ~ 3.4). At the superfamily level, LINEs have the highest expression overall followed by the aggregation of unclassified TEs (Figure 5B). Retroelements are more transcribed than DNA transposons (average $\log_{10}$ CPM ~2 and 1.2, respectively).

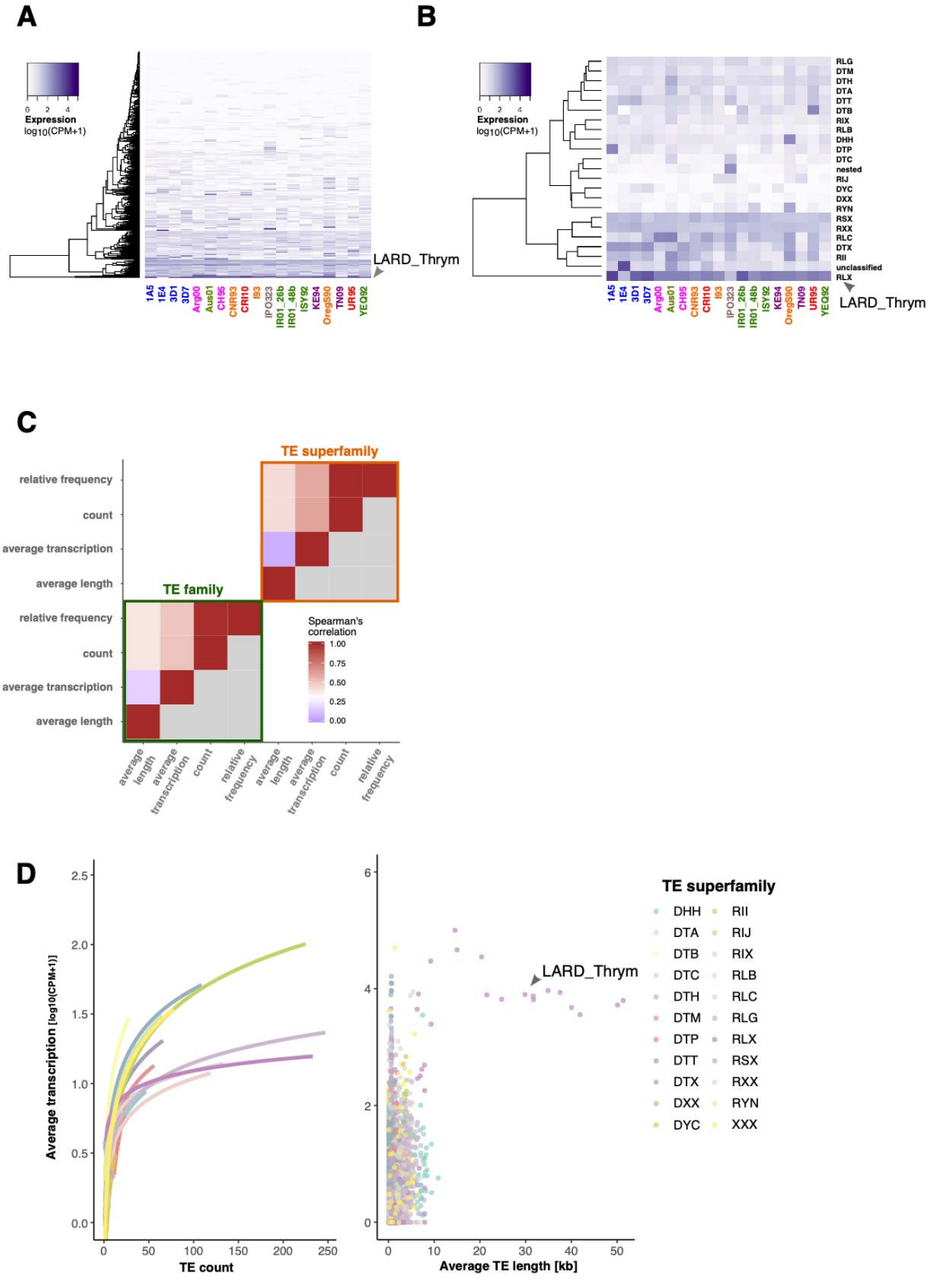


Figure 5: Transcriptional activity of transposable elements (TEs) and TE copy numbers. A. TE family transcription levels across all 19 genomes expressed as log_{10} (CPM +1). **B**. Average transcription levels of the 23 TE superfamilies across all genomes (average log_{10} (CPM +1). **C**.

Spearman's correlation matrix of four metrics (TE count, TE relative frequency, TE average length and transcription) at the TE family and superfamily level. **D**. Variation of TE transcription (average log_{10} (CPM +1)) as a function of TE count (left panel) or TE average length (right panel). Curves in the left panel show the logarithmic linear regression given by the linear model log_{10}(CPM+1) ~ log_{10} (TE count). The large highly expressed LARD_Thrym family (RLX) is highlighted by arrowheads in panels A, B, and D (right side).

To understand TE expression dynamics across the pangenome, we investigated associations between TE transcription, length and frequency (Figure 5C). We found TE transcription to be highly correlated with TE frequency in the genomes (Spearman's r = 0.49, p < 5e-307) and a weaker correlation at the TE superfamily level (Spearman's r = 0.1, p < 5e-13). Furthermore, TE transcription is negatively correlated with TE length at the superfamily level (Spearman's r =-0.4, p < 1e-187; Figure 5D). Interestingly, the average TE superfamily transcription levels are positively correlated with the frequency of the TE superfamily in the genome (Figure 5D). A notable exception is unclassified SINE retroelements. The correlation of TE transcription levels and TE frequency in the genome strongly suggests that transcriptional activity contributed to recent TE expansions in the genome.

Discussion

We established a global pangenome of a major fungal wheat pathogen based on the assembly and analysis of 19 high-quality genomes. *Z. tritici* segregates major chromosomal rearrangements affecting both the more conserved core chromosomes as well as the highly polymorphic accessory chromosomes. The gene content is highly variable among genomes with only 60% of all genes being conserved in the species. Accessory genes encode functions for a wide variety of interactions with both biotic and abiotic environments. An exhaustive map of TEs across all genomes pinpoints transposon-associated genome expansions across geographic regions.

We showed that the *Z. tritici* pangenome is expansive with ~40% accessory orthogroups. Compared to a previous construction of the *Z. tritici* pangenome based on genomes from a much narrower

363 geographic breadth (Plissonneau *et al.* 2018), we used more relaxed criteria to assign genes into
364 orthogroups. Based on the tendency to assign more divergent gene variants into the same orthogroup,
365 we recovered a total of 911 orthogroups with at least one paralog compared to only 76 identified
366 previously. The number of paralogs remains low compared to species with larger genomes that
367 retained more paralogs of gene duplication events (Goodwin *et al.* 2011). A likely constraint on gene
368 duplication is the genomic defence mechanism that introduces repeat-induced point (RIP) mutations
369 (Selker 2002). Although these defences evolved to suppress transpositional activity of TEs, they can
370 also affect genome evolution by targeting gene duplicates (Selker 2002; Galagan & Selker 2004).
371 Recent sequencing efforts oriented around important crop species reported impressively large
372 accessory genome proportions (Hirsch *et al.* 2014; Zhou *et al.* 2017; Zhao *et al.* 2018). However,
373 nearly all eukaryotic pangenomes are partially based on short-read assemblies that challenge the
374 resolution of segregating gene variants within a species. With the conservative estimate of ~24% non-
375 reference orthogroups, the *Z. tritici* accessory genome is the largest reported for a fungal species to
376 date (~40% of the pangenome). This falls outside the upper range of comparative analyses of human
377 fungal pathogens and *S. cerevisiae*, where estimates of the accessory genome ranged from 10-20%
378 (McCarthy & Fitzpatrick 2019). However, bacterial accessory genomes can range from 0 to 95% of
379 the total pangenome (Rouli *et al.* 2015). The effective population size of a species, its lifestyle, and
380 niche heterogeneity are main factors influencing bacterial pangenome sizes (McInerney *et al.* 2017).
381 Similar to bacteria, the effective population size is likely to be the major factor maintaining a large
382 accessory genome in *Z. tritici*. Previous studies identified *Z. tritici* as a highly polymorphic species
383 with a rapid decay in linkage disequilibrium, high SNP densities and high recombination rates (Croll
384 *et al.* 2015; Hartmann *et al.* 2017). As a consequence, the pathogen likely retains significant
385 functional variation within populations as long as the variation is nearly neutral.
386
387 Bacterial and fungal genomes show clear functional compartmentalization between core and
388 accessory genes (McInerney *et al.* 2017; McCarthy & Fitzpatrick 2019). In fungi, core orthogroups
389 are enriched for housekeeping functions in contrast to an enrichment for antimicrobial resistance and
390 pathogenicity factors among accessory genes (Plissonneau *et al.* 2018). Here we show that genes

encoding carbohydrate-active enzymes (CAZymes) are highly conserved within the species. CAZymes are involved in the degradation of the host cell wall and other storage compounds (Zerillo *et al.* 2013; Lyu *et al.* 2015). Strong conservation of the content in CAZymes may reflect a fundamental adaptation to wheat as a host plant. This contrasts with generalist pathogens, which often evolved larger CAZyme repertoires (Zhao *et al.* 2013). In contrast to CAZymes, secondary metabolite gene clusters show substantial presence-absence variation within the species. Fungi produce highly diverse secondary metabolites that play a role during various life cycle stages, but often have poorly understood functions (Calvo *et al.* 2002). Plant pathogens were also shown to depend on secondary metabolite production for full virulence (Pusztahelyi *et al.* 2015). Hence, variation in secondary metabolite production may underlie variation in virulence. Species from the genus *Aspergillus* produce a large diversity of secondary metabolites for which the gene clusters often segregate presence-absence (Kjærbølling *et al.* 2018; Raffa & Keller 2019). The *Z. tritici* pangenome was constructed from isolates coming from six different continents and a wide array of agricultural environments. Hence, differences in secondary metabolite production capacity may reflect local adaptation and trade-offs that balance the cost of metabolite production. Virulence of *Z. tritici* is thought to be largely governed by gene-for-gene interactions (Brown *et al.* 2015). In such interactions effector proteins either promote disease or are recognized by the host and trigger resistance (Jones & Dangl 2006). A gene encoding a recognized effector should therefore be rapidly eliminated from the species gene pool. *Z. tritici* populations responded rapidly to selection on effector gene loci by either mutating, deleting or silencing genes (Hartmann & Croll 2017; Krishnan *et al.* 2018; Meile *et al.* 2018). Our global pangenome analysis significantly expands our understanding of effector gene diversification. We identified 652 orthogroups encoding predicted effector functions of which 63% are accessory orthogroups. Accessory effector genes may be involved in arms races with strong selection driving the gain or loss of individual effector genes in populations. As a contrast, we identified 45 conserved and highly expressed effectors genes potentially encoding indispensable pathogenicity functions.

418 Ultimate mechanisms generating pangenomes may include large population sizes and niche
419 complexity, however the proximate mechanisms generating pangenome diversification are poorly
420 understood. TEs can be key drivers generating structural variation (Beck *et al.* 2011; Kim *et al.* 2019)
421 and *Z. tritici* readily undergoes TE-mediated chromosomal rearrangements during meiosis (Croll *et al.*
422 2013; Fouché *et al.* 2018b). Here we show that *Z. tritici* genomes contain 16-24% TEs, with the
423 overall proportion of TEs accounting for ~70% of the intraspecific genome size variation. Hence, TEs
424 are key drivers of genome evolution in this species. Among the most drastic chromosomal
425 rearrangements, we detected a significantly shorter chromosome 7 homolog. The longer homolog was
426 hypothesized to have originated from a fusion with an accessory chromosome based on evidence from
427 large scale epigenetic remodelling (Schotanus *et al.* 2015). Our analysis likely identified the ancestral
428 variant prior to the suspected chromosomal fusion event. Hence, the species retained two major
429 chromosomal variants of a core chromosome.
430
431 TEs are often implicated in gene copy number variation through duplication or pseudogenisation
432 events suggesting that TEs directly contribute to pangenome diversification. We show that specific
433 *Gypsy* and *Helitron* elements were integrated into genes generating highly paralogous orthogroups.
434 These orthogroups may underlie recent expansions of specific TEs in the genomes of Australian and
435 Oregon isolates. The *Helitron* element is among the most transcribed TEs in the Oregon isolate,
436 suggesting a high potential for new transpositions. In contrast, the *Gypsy* element is only weakly
437 transcribed in the Australian isolate, suggesting that this TE has become deactivated by genomic
438 defences. In addition to transpositional activity causing loss-of-function mutations in genes, TEs can
439 also contribute to genome expansions (Naville *et al.* 2019). We found a strong correlation of overall
440 TE content and genome size across the analysed genomes. Intra-specific variation in genome size is
441 unexpected. However, local population bottlenecks could lead to changes in mean genome sizes given
442 the large structural variation segregating among isolates. Hence, the population history might have
443 impacted genome size evolution. Populations in Australia underwent a significant founder event
444 during the recent colonization of the continent (Zhan *et al.* 2005). Hence, our observation of an
445 expanded Australian genome may be causally linked to this bottleneck. Genome expansions may also

446 be triggered by TE activation through mobilisation triggered by stressors such as host infections
447 (Fouché *et al.* 2019b). Taken together, TE dynamics and large effective population sizes likely
448 constitute the proximate and ultimate drivers of pangenome size evolution. Understanding the birth
449 and death cycles of gene functions in such evolving pangenomes will help address major questions
450 related to crop-pathogen co-evolution.

Methods

High molecular-weight DNA extraction and single molecule real-time (SMRT) sequencing

Origin and year of sampling of all the isolates are described in Figure 1. High-molecular-weight DNA was extracted from lyophilized spores following a modified version of a cetyltrimethylammonium bromide (CTAB) protocol developed for plant tissue described in (Plissonneau *et al.* 2016b). Briefly, ~100 mg of lyophilized spores were crushed with a mortar and transferred to a phenol-chloroform-isoamyl alcohol solution. The supernatant was centrifuged and the pellet resuspended twice in fresh phenol-chloroform-isoamyl alcohol. The resulting pellet was then washed three times and resuspended in 100 µl of sterile water. For each isolate, PacBio SMRTbell libraries were prepared using between 15 µg and 30 µg of high molecular-weight DNA. Sequencing was performed on a PacBio Sequel instrument at the Functional Genomics Center, Zürich, Switzerland.

Complete genome assemblies

We largely followed the pipeline described in (Yue & Liti 2018). In summary, raw PacBio sequencing reads were assembled using *Canu* v1.7.1 (Koren *et al.* 2017). All assemblies were performed with an estimated genomeSize of 39.678 Mb. Two correctedErrorRate (0.045 and 0.039) and minReadLength (500 and 5000) were tested and the most contiguous chromosome-level assemblies were retained for further analysis based on reference alignment. Reads were aligned to the obtained assemblies using *pbalign* v0.3.1 from Pacific Biosciences suite (https://github.com/PacificBiosciences/pbalign). The assemblies were polished twice using the software Arrow v2.2.2 from the Pacific Biosciences suite with default settings (https://github.com/PacificBiosciences/GenomicConsensus) and chromosome-level assemblies were performed using Ragout v2.1.1 (Kolmogorov *et al.* 2014).

RNA extraction, library preparation, sequencing and quantification

For isolates 1A5, 1E4, 3D1 and 3D7, RNA sequencing experiments on minimal media were performed by (Fouché *et al.* 2019a; Francisco *et al.* 2019). Raw reads were retrieved from the NCBI Short Read Archive accession number SRP077418. Similarly, the 15 additional fungal isolates (Table S1) were grown in YSB media (10g sucrose + 10g yeast extract per liter) and then 10e5 cells were inoculated on liquid minimal media without a carbon source (Metzenberg 2003) for 7-10 days prior to extraction. RNA was extracted using a NucleoSpin® RNA Plant kit following the manufacturer's instructions. Library preparation was carried out according to the Illumina TruSeq Stranded mRNA Library Prep protocol with unique indexes for each sample. Single-end 100-bp sequencing was performed on a HiSeq 4000 at the iGE3 platform in Geneva, Switzerland. RNA-seq reads were first filtered using Trimmomatic v0.38 (Bolger *et al.* 2014) using the following parameters: ILLUMINACLIP:TruSeq3-SE.fa: 2:30:10 LEADING:10 TRAILING:10 SLIDINGWINDOW:5:10 MINLEN: 50, and then aligned to the corresponding genome assembly using STAR v2.6.0a (Dobin *et al.* 2013) allowing for multiple read mapping (parameters set as --outFilterMultimapNmax 100 --winAnchorMultimapNmax 200 --outFilterMismatchNmax 3). We used HTSeq-count v0.11.2 (Anders *et al.* 2015) with -s reverse and –m union parameters to recover counts per feature (joint counting of reads in genes and TEs). We calculated normalized feature counts expressed as counts per million using the EdgeR package v3.24.3 (Robinson *et al.* 2010). We restricted our analyses to features with a count per million >1.

Gene prediction and genome annotation

We used the gene prediction pipeline BRAKER v2.1 to predict genes in the 14 newly assembled genomes (Altschul *et al.* 1990; Stanke *et al.* 2006, 2008; Camacho *et al.* 2009; Li *et al.* 2009; Barnett *et al.* 2011; Lomsadze *et al.* 2014; Hoff *et al.* 2016). BRAKER combines coding sequence and intron hints based on the mapping of conserved protein sequences and introns identified in RNA-seq data, respectively. The above described RNA-seq datasets were joined with predicted protein sequences from the reference isolate IPO323 (Goodwin *et al.* 2011) and used to predict gene features and guide splice site mapping. RNA alignment files were generated with HISAT2 v2.1.0 using the --rna-strandness R option (Kim *et al.* 2015). The resulting bam files were provided to BRAKER (--bam

option) together with mapped IPO323 reference proteins (--prot_seq option) to generate gene predictions for each assembled genome using the --alternatives-from-evidence=false --prg=gth --etpmode --fungus parameters. Orthologous genes were identified using protein sequences from all 19 isolates and Orthofinder v2.1.2 with default parameters (Emms & Kelly 2015, 2019).

TE consensus identification, classification and annotation

To obtain consensus sequences for TE families, individual runs of RepeatModeler were performed on the 19 complete genomes in addition to the genome of *Z. pseudotritici* (REF Stukenbrock et al 2010 PLOS Genetics). The classification was based on the GIRI Repbase using RepeatMasker (Bao *et al.* 2015; Smit, AFA, Hubley, R & Green 2015). In order to finalize the classification of TE consensus sequences, we used WICKERsoft (Breen *et al.* 2010). The 19 complete genomes were screened for copies of consensus sequences with blastn filtering for sequence identity of > 80% on >80% of the length of the sequence (Altschul *et al.* 1997). Flanks of 300 bp were added and new multiple sequence alignments were performed using ClustalW (G Higgins & M Sharp 1988). Boundaries were visually inspected and trimmed if necessary. Consensus sequences were classified according to the presence and type of terminal repeats and homology of encoded proteins using hints from blastx on NCBI. Consensus sequences were renamed according to a three-letter classification system (Wicker *et al.* 2007).

A second round of annotation was performed based on predicted protein sequences of TE superfamilies from other fungal species. Here again, the 19 complete genomes were screened for a protein sequence of each superfamily using tblastn. Blast hits were filtered for a minimal alignment size of 80 bp and sequence similarity >35%. Flanks of 3'000 bp or more both up and downstream of the sequence were then added. Hits were pairwise compared with dotplots using dotter and grouped into families based on visual inspection (Sonnhammer & Durbin 1995). Finally, multiple sequence alignments were performed with ClustalW to construct consensus sequences and the consensus sequences were renamed according to the three-letter system (Wicker *et al.* 2007).

A third round of annotation of the 19 complete genomes was done to identify four groups of short non-autonomous TEs. LTR-Finder was used to screen for LARDs (LArge Retrotransposon Derivates) and TRIMs (Terminal Repeat retrotransposons In Miniature) with the filters -d 2001 -D 6000 -l 30 -L 5000 and -d 30 -D 2000 -l 30 -L 500 respectively. MITE-Tracker was used to screen for MITEs (Miniature Inverted-repeat Transposable Elements) and SINE-Finder in Sine-Scan to screen for SINEs (Short Interspersed Nuclear Elements) (Xu & Wang 2007; Wenke *et al.* 2011; Ma *et al.* 2015; Gao *et al.* 2016; Mao & Wang 2017; Crescente *et al.* 2018). For each detected LARD, TRIM and SINE, consensus sequences were created as described above and duplicates excluded. All genome assemblies were then annotated with the curated consensus sequences using RepeatMasker with a cut-off value of 250 and ignored simple repeats as well as low complexity regions. Annotated elements shorter than 100 bp were filtered out, and adjacent identical TEs overlapping by more than 100 bp were merged. Different TE families overlapping by more than 100 bp were considered as *nested* insertions and were renamed accordingly. Identical elements separated by less than 200 bp indicative of putative interrupted elements were grouped into a single element using minimal start and maximal stop positions. TEs overlapping ≥1 bp with genes were recovered using the *bedtools* v2.27.1 suite and the *overlap* function (Quinlan & Hall 2010). Correlations were calculated in RStudio version 1.1.453 using Spearman's coefficient for pairwise complete observations and statistics were inferred with the *psych* package using the Holm correction method (Revelle 2017).

Functional annotation of predicted genes

Protein functions were predicted for all gene models using InterProScan v 5.31-70.0 (Jones *et al.* 2014) adding -goterms -iprlookup and -pathway information. Secretion peptides and transmembrane domains (TM) were identified using SignalP v 4.1 and Phobius (Käll *et al.* 2004; Petersen *et al.* 2011). The secretome was defined as the set of proteins with a signal peptide but no TM as predicted by either SignalP and Phobius. Putative effectors were identified with EffectorP v 2.0 with default parameters using the set of predicted secreted proteins (Sperschneider *et al.* 2016). Carbohydrate-active enzymes (CAZymes) were identified using dbCAN2 release 7.0 server (Lombard *et al.* 2014; Zhang *et al.* 2018) with the three tools HMMER, DIAMOND and Hotpep (Finn *et al.* 2011; Buchfink

et al. 2015; Busk *et al.* 2017). Proteins were classified as a CAZyme if predicted by each of the three tools. We searched for secondary metabolite gene clusters using the online version 4 of antiSMASH (Blin *et al.* 2017). Genes belonging to an identified cluster were annotated as "biosynthetic", "biosynthetic-additional", "transport", "regulatory" or "other". Gene clusters mapping at a conserved, orthologous locus shared by two or more isolate were considered as syntenic.

Declarations

Ethics approval and consent to participate: n/a

Consent for publication: n/a

Availability of data and materials: The genome assembly and annotation for new genome assemblies are available at the European Nucleotide Archive (http://www.ebi.ac.uk/ena) under accession number PRJEB33986. The RNA-sequencing raw sequencing data was deposited at the NCBI Short Read Archive under the accession number PRJNA559981.

Competing interests: none

Funding: BAM and DC received support from the Swiss National Science Foundation (grants 31003A_155955 and 31003A_173265, respectively). DC was also supported by a grant from the Fondation Pierre Mercier pour la Science for this work.

Authors' contributions: TB and DC conceived the study; TB and UO performed analyses; LA and BAM provided datasets and strains; BAM and DC provided funding; TB and DC wrote the manuscript.

Acknowledgements: We are grateful for helpful comments by Simone Fouché on a previous version of this manuscript. Data generated for this manuscript was obtained in collaboration with the Genetic Diversity Centre (GDC), ETH Zurich and the Functional Genomics Center Zurich (FGCZ).

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990). Basic local alignment search tool. *J. Mol. Biol.*, 215, 403–10.

589 Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., *et al.* (1997). Gapped
590 BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids*
591 *Res.*, 25, 3389–3402.
592 Anders, S., Pyl, P.T. & Huber, W. (2015). HTSeq--a Python framework to work with high-throughput
593 sequencing data. *Bioinformatics*, 31, 166–9.
594 Araki, H., Tian, D., Goss, E.M., Jakob, K., Halldorsdottir, S.S., Kreitman, M., *et al.* (2006).
595 Presence/absence polymorphism for alternative pathogenicity islands in Pseudomonas
596 viridiflava, a pathogen of Arabidopsis. *Pnas*, 103, 5887–5892.
597 Bao, W., Kojima, K.K. & Kohany, O. (2015). Repbase Update, a database of repetitive elements in
598 eukaryotic genomes. *Mob. DNA*, 6, 11.
599 Barnett, D.W., Garrison, E.K., Quinlan, A.R., Stromberg, M.P. & Marth, G.T. (2011). BamTools: a
600 C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics*, 27, 1691–1692.
601 Beck, C.R., Garcia-Perez, J.L., Badge, R.M. & Moran, J. V. (2011). LINE-1 Elements in Structural
602 Variation and Disease. *Annu. Rev. Genomics Hum. Genet.*, 12, 187–215.
603 Blake, J.J., Gosling, P., Fraaije, B.A., Burnett, F.J., Knight, S.M., Kildea, S., *et al.* (2018). Changes in
604 field dose-response curves for demethylation inhibitor (DMI) and quinone outside inhibitor
605 (QoI) fungicides against *Zymoseptoria tritici* , related to laboratory sensitivity phenotyping and
606 genotyping assays. *Pest Manag. Sci.*, 74, 302–313.
607 Blin, K., Wolf, T., Chevrette, M.G., Lu, X., Schwalen, C.J., Kautsar, S.A., *et al.* (2017). antiSMASH
608 4.0-improvements in chemistry prediction and gene cluster boundary identification. *Nucleic*
609 *Acids Res.*, 45, W36–W41.
610 Bolger, A.M., Lohse, M. & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina
611 sequence data. *Bioinformatics*, 30, 2114–20.
612 Breen, J., Wicker, T., Kong, X., Zhang, J., Ma, W., Paux, E., *et al.* (2010). A highly conserved gene
613 island of three genes on chromosome 3B of hexaploid wheat: diverse gene function and genomic
614 structure maintained in a tightly linked block. *BMC Plant Biol.*, 10, 98.
615 Brown, J.K.M., Chartrain, L., Lasserre-Zuber, P. & Saintenac, C. (2015). Genetics of resistance to
616 Zymoseptoria tritici and applications to wheat breeding. *Fungal Genet. Biol.*, 79, 33–41.

617 Brynildsrud, O., Gulla, S., Feil, E.J., Nørstebø, S.F. & Rhodes, L.D. (2016). Identifying copy number
618 variation of the dominant virulence factors msa and p22 within genomes of the fish pathogen
619 Renibacterium salmoninarum. *Microb. genomics*, 2, e000055.
620 Buchfink, B., Xie, C. & Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND.
621 *Nat. Methods*, 12, 59–60.
622 Busk, P.K., Pilgaard, B., Lezyk, M.J., Meyer, A.S. & Lange, L. (2017). Homology to peptide pattern
623 for annotation of carbohydrate-active enzymes and prediction of function. *BMC Bioinformatics*,
624 18, 214.
625 Calvo, A.M., Wilson, R.A., Bok, J.W. & Keller, N.P. (2002). Relationship between secondary
626 metabolism and fungal development. *Microbiol. Mol. Biol. Rev.*, 66, 447–59, table of contents.
627 Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., *et al.* (2009).
628 BLAST+: architecture and applications. *BMC Bioinformatics*, 10, 421.
629 Cools, H.J. & Fraaije, B.A. (2008). Are azole fungicides losing ground against Septoria wheat
630 disease? Resistance mechanisms inMycosphaerella graminicola. *Pest Manag. Sci.*, 64, 681–684.
631 Crescente, J.M., Zavallo, D., Helguera, M. & Vanzetti, L.S. (2018). MITE Tracker: an accurate
632 approach to identify miniature inverted-repeat transposable elements in large genomes. *BMC*
633 *Bioinformatics*, 19, 348.
634 Croll, D., Lendenmann, M.H., Stewart, E. & McDonald, B.A. (2015). The impact of recombination
635 hotspots on genome evolution of a fungal plant pathogen. *Genetics*, 201, 1213–1228.
636 Croll, D., Zala, M., McDonald, B.A., Smoot, M. & Shumway, M. (2013). Breakage-fusion-bridge
637 Cycles and Large Insertions Contribute to the Rapid Evolution of Accessory Chromosomes in a
638 Fungal Pathogen. *PLoS Genet.*, 9, e1003567.
639 Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., *et al.* (2013). STAR:
640 ultrafast universal RNA-seq aligner. *Bioinformatics*, 29, 15–21.
641 Dong, S., Raffaele, S. & Kamoun, S. (2015). The two-speed genomes of filamentous pathogens: waltz
642 with plants. *Curr. Opin. Genet. Dev.*, 35, 57–65.
643 Emms, D.M. & Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome
644 comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.*, 16, 157.

645 Emms, D.M. & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative
646 genomics. *bioRxiv*, 466201.
647 Faino, L., Seidl, M.F., Shi-Kunne, X., Pauper, M., Van Den Berg, G.C.M., Wittenberg, A.H.J., *et al.*
648 (2016). Transposons passively and actively contribute to evolution of the two-speed genome of a
649 fungal pathogen. *Genome Res.*, 26, 1091–1100.
650 Finn, R.D., Clements, J. & Eddy, S.R. (2011). HMMER web server: interactive sequence similarity
651 searching. *Nucleic Acids Res.*, 39, W29-37.
652 Fouché, S., Badet, T., Oggenfuss, U., Plissonneau, C., Francisco, C.S. & Croll, D. (2019a). Stress-
653 driven transposable element de-repression dynamics and virulence evolution in a fungal
654 pathogen. *Mol. Biol. Evol.*
655 Fouché, S., Badet, T., Oggenfuss, U., Plissonneau, C., Francisco, C.S. & Croll, D. (2019b). Stress-
656 driven transposable element de-repression dynamics in a fungal pathogen. *bioRxiv*, 633693.
657 Fouché, S., Plissonneau, C. & Croll, D. (2018a). The birth and death of effectors in rapidly evolving
658 filamentous pathogen genomes. *Curr. Opin. Microbiol.*, 46, 34–42.
659 Fouché, S., Plissonneau, C., McDonald, B.A. & Croll, D. (2018b). Meiosis Leads to Pervasive Copy-
660 Number Variation and Distorted Inheritance of Accessory Chromosomes of the Wheat Pathogen
661 Zymoseptoria tritici. *Genome Biol. Evol.*, 10, 1416–1429.
662 Francisco, C.S., Ma, X., Zwyssig, M.M., McDonald, B.A. & Palma-Guerrero, J. (2019).
663 Morphological changes in response to environmental stresses in the fungal plant pathogen
664 Zymoseptoria tritici. *Sci. Rep.*, 9, 9642.
665 G Higgins, D. & M Sharp, P. (1988). CLUSTAL: a package for performing multiple sequence
666 alignment on a microcomputer. *Gene*, 73, 237–244.
667 Galagan, J.E. & Selker, E.U. (2004). RIP: the evolutionary cost of genome defense. *Trends Genet.*,
668 20, 417–23.
669 Gao, D., Li, Y., Kim, K. Do, Abernathy, B. & Jackson, S.A. (2016). Landscape and evolutionary
670 dynamics of terminal repeat retrotransposons in miniature in plant genomes. *Genome Biol.*, 17,
671 7.
672 Goodwin, S.B., Ben M'Barek, S., Dhillon, B., Wittenberg, A.H.J., Crane, C.F., Hane, J.K., *et al.*

673 (2011). Finished Genome of the Fungal Wheat Pathogen Mycosphaerella graminicola Reveals
674 Dispensome Structure, Chromosome Plasticity, and Stealth Pathogenesis. *PLoS Genet.*, 7,
675 e1002070.
676 Grandaubert, J., Dutheil, J.Y. & Stukenbrock, E.H. (2019). The genomic determinants of adaptive
677 evolution in a fungal pathogen. *Evol. Lett.*, 3, 299–312.
678 Hartmann, F.E. & Croll, D. (2017). Distinct Trajectories of Massive Recent Gene Gains and Losses in
679 Populations of a Microbial Eukaryotic Pathogen. *Mol. Biol. Evol.*, 127, 1–18.
680 Hartmann, F.E., Rodríguez de la Vega, R.C., Brandenburg, J.-T., Carpentier, F. & Giraud, T. (2018).
681 Gene Presence–Absence Polymorphism in Castrating Anther-Smut Fungi: Recent Gene Gains
682 and Phylogeographic Structure. *Genome Biol. Evol.*, 10, 1298–1314.
683 Hartmann, F.E., Sánchez-Vallet, A., McDonald, B.A. & Croll, D. (2017). A fungal wheat pathogen
684 evolved host specialization by extensive chromosomal rearrangements. *ISME J.*, 11, 1189–1204.
685 Hirsch, C.N., Foerster, J.M., Johnson, J.M., Sekhon, R.S., Muttoni, G., Vaillancourt, B., *et al.* (2014).
686 Insights into the maize pan-genome and pan-transcriptome. *Plant Cell*, 26, 121–35.
687 Hoff, K.J., Lange, S., Lomsadze, A., Borodovsky, M. & Stanke, M. (2016). BRAKER1:
688 Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS:
689 Table 1. *Bioinformatics*, 32, 767–769.
690 Jackson, R.W., Vinatzer, B., Arnold, D.L., Dorus, S. & Murillo, J. (2011). The influence of the
691 accessory genome on bacterial pathogen evolution. *Mob. Genet. Elements*, 1, 55–65.
692 Jones, J.D.G. & Dangl, J.L. (2006). The plant immune system. *Nature*, 444, 323–9.
693 Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., *et al.* (2014). InterProScan 5:
694 genome-scale protein function classification. *Bioinformatics*, 30, 1236–40.
695 Käll, L., Krogh, A. & Sonnhammer, E.L.. (2004). A Combined Transmembrane Topology and Signal
696 Peptide Prediction Method. *J. Mol. Biol.*, 338, 1027–1036.
697 Kim, D., Langmead, B. & Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory
698 requirements. *Nat. Methods*, 12, 357–360.
699 Kim, S., Mun, S., Kim, T., Lee, K.-H., Kang, K., Cho, J.-Y., *et al.* (2019). Transposable element-
700 mediated structural variation analysis in dog breeds using whole-genome sequencing. *Mamm.*

Genome.

Kjærbølling, I., Vesth, T.C., Frisvad, J.C., Nybo, J.L., Theobald, S., Kuo, A., *et al.* (2018). Linking secondary metabolites to gene clusters through genome sequencing of six diverse Aspergillus species. *Proc. Natl. Acad. Sci. U. S. A.*, 115, E753–E761.

Kolmogorov, M., Raney, B., Paten, B. & Pham, S. (2014). Ragout-a reference-assisted assembly tool for bacterial genomes. *Bioinformatics*, 30, i302-9.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H. & Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.*, 27, 722–736.

Krishnan, P., Meile, L., Plissonneau, C., Ma, X., Hartmann, F.E., Croll, D., *et al.* (2018). Transposable element insertions shape gene regulation and melanin production in a fungal pathogen of wheat. *BMC Biol.*, 16, 78.

Lefébure, T., Pavinski Bitar, P.D., Suzuki, H. & Stanhope, M.J. (2010). Evolutionary Dynamics of Complete Campylobacter Pan-Genomes and the Bacterial Species Concept. *Genome Biol. Evol.*, 2, 646–655.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., *et al.* (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25, 2078–2079.

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P.M. & Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.*, 42, D490-5.

Lomsadze, A., Burns, P.D. & Borodovsky, M. (2014). Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res.*, 42, e119–e119.

Lucas, J.A., Hawkins, N.J. & Fraaije, B.A. (2015). The Evolution of Fungicide Resistance. In: *Advances in applied microbiology*. pp. 29–92.

Lyu, X., Shen, C., Fu, Y., Xie, J., Jiang, D., Li, G., *et al.* (2015). Comparative genomic and transcriptional analyses of the carbohydrate-active enzymes and secretomes of phytopathogenic fungi reveal their significant roles during infection and development. *Sci. Rep.*, 5, 15565.

Ma, B., Li, T., Xiang, Z. & He, N. (2015). MnTEdb, a collective resource for mulberry transposable elements. *Database*, 2015.

729 Mao, H. & Wang, H. (2017). SINE_scan: an efficient tool to discover short interspersed nuclear
730 elements (SINEs) in large-scale genomic datasets. *Bioinformatics*, 33, btw718.
731 Marschall, T., Marz, M., Abeel, T., Dijkstra, L., Dutilh, B.E., Ghaffaari, A., *et al.* (2016).
732 Computational pan-genomics: status, promises and challenges. *Brief. Bioinform.*, 19, bbw089.
733 McCarthy, C.G.P. & Fitzpatrick, D.A. (2019). Pan-genome analyses of model fungal species. *Microb.*
734 *genomics*, 5.
735 McInerney, J.O., McNally, A. & O'Connell, M.J. (2017). Why prokaryotes have pangenomes. *Nat.*
736 *Microbiol.*, 2, 17040.
737 Meile, L., Croll, D., Brunner, P.C., Plissonneau, C., Hartmann, F.E., McDonald, B.A., *et al.* (2018). A
738 fungal avirulence factor encoded in a highly plastic genomic region triggers partial resistance to
739 septoria tritici blotch. *New Phytol.*, 219, 1048–1061.
740 Metzenberg, R.L. (2003). Vogel's Medium N salts: avoiding the need for ammonium nitrate. *Fungal*
741 *Genet. Rep.*, 50, 14–14.
742 Naville, M., Henriet, S., Warren, I., Sumic, S., Reeve, M., Volff, J.-N., *et al.* (2019). Massive
743 Changes of Genome Size Driven by Expansions of Non-autonomous Transposable Elements.
744 *Curr. Biol.*, 29, 1161-1168.e6.
745 Omrane, S., Audéon, C., Ignace, A., Duplaix, C., Aouini, L., Kema, G., *et al.* (2017). Plasticity of the
746 MFS1 Promoter Leads to Multidrug Resistance in the Wheat Pathogen Zymoseptoria tritici.
747 *mSphere*, 2, e00393-17.
748 Palma-Guerrero, J., Ma, X., Torriani, S.F.F., Zala, M., Francisco, C.S., Hartmann, F.E., *et al.* (2017).
749 Comparative Transcriptome Analyses in *Zymoseptoria tritici* Reveal Significant Differences in
750 Gene Expression Among Strains During Plant Infection. *Mol. Plant-Microbe Interact.*, 30, 231–
751 244.
752 Petersen, T.N., Brunak, S., von Heijne, G. & Nielsen, H. (2011). SignalP 4.0: discriminating signal
753 peptides from transmembrane regions. *Nat. Methods*, 8, 785–786.
754 Plissonneau, C., Daverdin, G., Ollivier, B., Blaise, F., Degrave, A., Fudal, I., *et al.* (2016a). A game
755 of hide and seek between avirulence genes *AvrLm4-7* and *AvrLm3* in *Leptosphaeria maculans*.
756 *New Phytol.*, 209, 1613–1624.

Plissonneau, C., Hartmann, F.E. & Croll, D. (2018). Pangenome analyses of the wheat pathogen *Zymoseptoria tritici* reveal the structural basis of a highly plastic eukaryotic genome. *BMC Biol.*, 16, 5.

Plissonneau, C., Stürchler, A. & Croll, D. (2016b). The Evolution of Orphan Regions in Genomes of a Fungal Pathogen of Wheat. *MBio*, 7, e01231-16.

Lo Presti, L., Lanver, D., Schweizer, G., Tanaka, S., Liang, L., Tollot, M., *et al.* (2015). Fungal Effectors and Plant Susceptibility. *Annu. Rev. Plant Biol.*, 66, 513–545.

Pusztahelyi, T., Holb, I.J. & Pócsi, I. (2015). Secondary metabolites in fungus-plant interactions. *Front. Plant Sci.*, 6, 573.

Quinlan, A.R. & Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26, 841–842.

Raffa, N. & Keller, N.P. (2019). A call to arms: Mustering secondary metabolites for success and survival of an opportunistic pathogen. *PLOS Pathog.*, 15, e1007606.

Ramasamy, D., Mishra, A.K., Lagier, J.-C., Padhmanabhan, R., Rossi, M., Sentausa, E., *et al.* (2014). A polyphasic strategy incorporating genomic data for the taxonomic description of novel bacterial species. *Int. J. Syst. Evol. Microbiol.*, 64, 384–391.

Revelle, W.R. (2017). psych: Procedures for Personality and Psychological Research.

Robinson, M.D., McCarthy, D.J. & Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26, 139.

Rouli, L., Merhej, V., Fournier, P.-E. & Raoult, D. (2015). The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New microbes new Infect.*, 7, 72–85.

Sánchez-Vallet, A., Fouché, S., Fudal, I., Hartmann, F.E., Soyer, J.L., Tellier, A., *et al.* (2018a). The Genome Biology of Effector Gene Evolution in Filamentous Plant Pathogens. *Annu. Rev. Phytopathol.*, 56, 21–40.

Sánchez-Vallet, A., Hartmann, F.E., Marcel, T.C. & Croll, D. (2018b). Nature's genetic screens: using genome-wide association studies for effector discovery. *Mol. Plant Pathol.*, 19, 3–6.

Schotanus, K., Soyer, J.L., Connolly, L.R., Grandaubert, J., Happel, P., Smith, K.M., *et al.* (2015). Histone modifications rather than the novel regional centromeres of Zymoseptoria tritici

distinguish core and accessory chromosomes. *Epigenetics Chromatin*, 8, 41.

Schrider, D.R. & Hahn, M.W. (2010). Gene copy-number polymorphism in nature. *Proc. R. Soc. B Biol. Sci.*, 277, 3213–3221.

Selker, E.U. (2002). Repeat-Induced Gene Silencing in Fungi. *Adv. Genet.*, 46, 439–450.

Smit, AFA, Hubley, R & Green, P. (2015). RepeatMasker Open-4.0.

Sonnhammer, E.L.L. & Durbin, R. (1995). A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene*, 167, Gc1–Gc10.

Sperschneider, J., Gardiner, D.M., Dodds, P.N., Tini, F., Covarelli, L., Singh, K.B., *et al.* (2016). Effector P: predicting fungal effector proteins from secretomes using machine learning. *New Phytol.*, 210, 743–761.

Sperschneider, J., Gardiner, D.M., Thatcher, L.F., Lyons, R., Singh, K.B., Manners, J.M., *et al.* (2015). Genome-Wide Analysis in Three Fusarium Pathogens Identifies Rapidly Evolving Chromosomes and Genes Associated with Pathogenicity. *Genome Biol. Evol.*, 7, 1613–27.

Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, 24, 637–644.

Stanke, M., Schöffmann, O., Morgenstern, B. & Waack, S. (2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*, 7, 62.

Steinberg, G. (2015). Cell biology of Zymoseptoria tritici: Pathogen cell organization and wheat infection. *Fungal Genet. Biol.*, 79, 17–23.

Steinhauer, D., Salat, M., Frey, R., Mosbach, A., Luksch, T., Balmer, D., *et al.* (2019). A dispensable paralog of succinate dehydrogenase subunit C mediates standing resistance towards a subclass of SDHI fungicides in Zymoseptoria tritici. *bioRxiv*, 616904.

Stukenbrock, E.H. & Dutheil, J.Y. (2018). Fine-Scale Recombination Maps of Fungal Plant Pathogens Reveal Dynamic Recombination Landscapes and Intragenic Hotspots. *Genetics*, 208, 1209–1229.

Tettelin, H., Riley, D., Cattuto, C. & Medini, D. (2008). Comparative genomics: the bacterial pan-genome. *Curr. Opin. Microbiol.*, 11, 472–7.

Toruño, T.Y., Stergiopoulos, I. & Coaker, G. (2016). Plant-Pathogen Effectors: Cellular Probes Interfering with Plant Defenses in Spatial and Temporal Manners. *Annu. Rev. Phytopathol.*, 54, 419–441.

Wenke, T., Dobel, T., Sorensen, T.R., Junghans, H., Weisshaar, B. & Schmidt, T. (2011). Targeted Identification of Short Interspersed Nuclear Element Families Shows Their Widespread Existence and Extreme Heterogeneity in Plant Genomes. *Plant Cell Online*, 23, 3117–3128.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., *et al.* (2007). A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.*, 8, 973–982.

Wit De, P.J.G.M., Mehrabi, R., Burg Van den, H.A. & Stergiopoulos, I. (2009). Fungal effector proteins: past, present and future. *Mol. Plant Pathol.*, 10, 735–47.

Wu, Y., Zaiden, N. & Cao, B. (2018). The Core- and Pan-Genomic Analyses of the Genus Comamonas: From Environmental Adaptation to Potential Virulence. *Front. Microbiol.*, 9, 3096.

Xu, Z. & Wang, H. (2007). LTR-FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.*, 35, 265–268.

Yoshida, K., Saunders, D.G.O., Mitsuoka, C., Natsume, S., Kosugi, S., Saitoh, H., *et al.* (2016). Host specialization of the blast fungus Magnaporthe oryzae is associated with dynamic gain and loss of genes linked to transposable elements. *BMC Genomics*, 17, 370.

Yue, J.-X. & Liti, G. (2018). Long-read sequencing data analysis for yeasts. *Nat. Protoc.*, 13, 1213–1231.

Zerillo, M.M., Adhikari, B.N., Hamilton, J.P., Buell, C.R., Lévesque, C.A. & Tisserat, N. (2013). Carbohydrate-Active Enzymes in Pythium and Their Role in Plant Cell Wall and Storage Polysaccharide Degradation. *PLoS One*, 8, e72572.

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

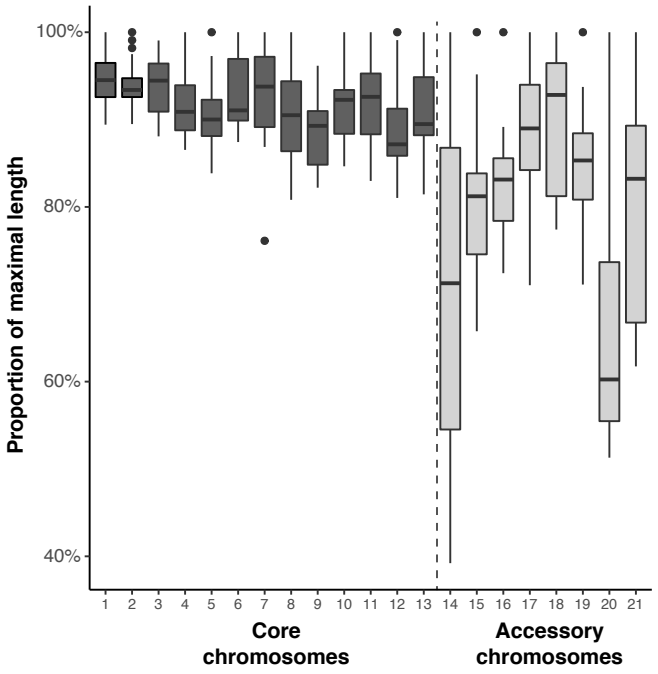
Zhang, H., Yohe, T., Huang, L., Entwistle, S., Wu, P., Yang, Z., *et al.* (2018). dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.*, 46, W95–W101.

841 Zhao, Q., Feng, Q., Lu, H., Li, Y., Wang, A., Tian, Q., *et al.* (2018). Pan-genome analysis highlights
842 the extent of genomic variation in cultivated and wild rice. *Nat. Genet.*, 50, 278–284.
843 Zhao, Z., Liu, H., Wang, C. & Xu, J.-R. (2013). Comparative analysis of fungal genomes reveals
844 different plant cell wall degrading capacity in fungi. *BMC Genomics*, 14, 274.
845 Zhou, P., Silverstein, K.A.T., Ramaraj, T., Guhlin, J., Denny, R., Liu, J., *et al.* (2017). Exploring
846 structural variation and gene family architecture with De Novo assemblies of 15 Medicago
847 genomes. *BMC Genomics*, 18, 261.
848

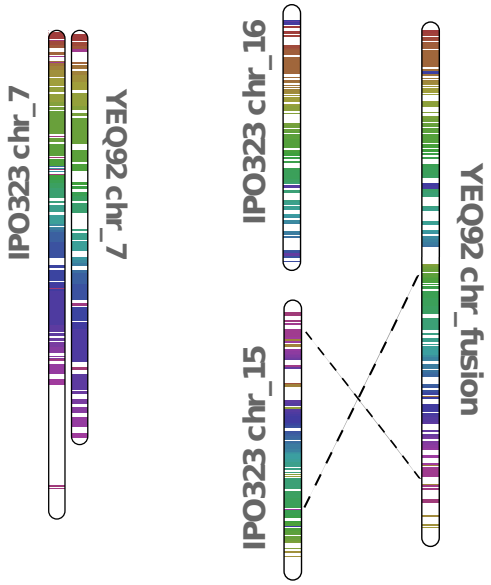
A

isolate name	origin [country]	genome size (bp)	GC content [%]	TE content [%]	accessory chromosomes	gene count	BUSCO score [%]	accessory genome [%]
1A5	Switzerland	39698711	51.6	18.7		12092	98.3	23.2
3D1	Switzerland	40657500	51.5	19.9		12006	98.5	22.5
3D7	Switzerland	37861420	52.1	18.4		11737	96.8	20.7
Arg00	Argentina	41149037	51.4	20.7		12347	98.4	24.8
Aus01	Australia	41760570	51.6	24.0		12787	98.1	27.6
CH95	Chile	39253131	51.7	19.6		12198	98.7	24.2
CNR93	Canada	38169704	52.1	19.2		12193	98.5	24.2
CRI10	Czech Republic	39697126	51.8	19.6		12265	98.1	24.4
I93	USA (Indiana)	39405809	51.5	19.1		12263	98.4	24.6
IPO323	Netherlands	39686251	51.7	19.1		11839	97.9	22.0
IR01_26b	Iran	38551850	51.5	16.5		11940	98.6	22.6
IR01_48b	Iran	37133562	52.0	18.1		11865	98.5	22.1
ISY92	Israel	39679942	51.1	19.1		12060	98.3	23.4
KE94	Kenya	38152782	51.9	17.3		11959	98.5	22.8
OregS90	USA (Oregon)	39591316	52.1	20.3		12483	98.4	26.0
TN09	Tunisia	37907319	51.4	18.8		11657	98.5	20.6
UR95	Ukraine	38369786	51.7	17.7		12241	98.5	24.4
YEQ92	Yemen	39871609	51.0	19.8		11734	98.3	21.3

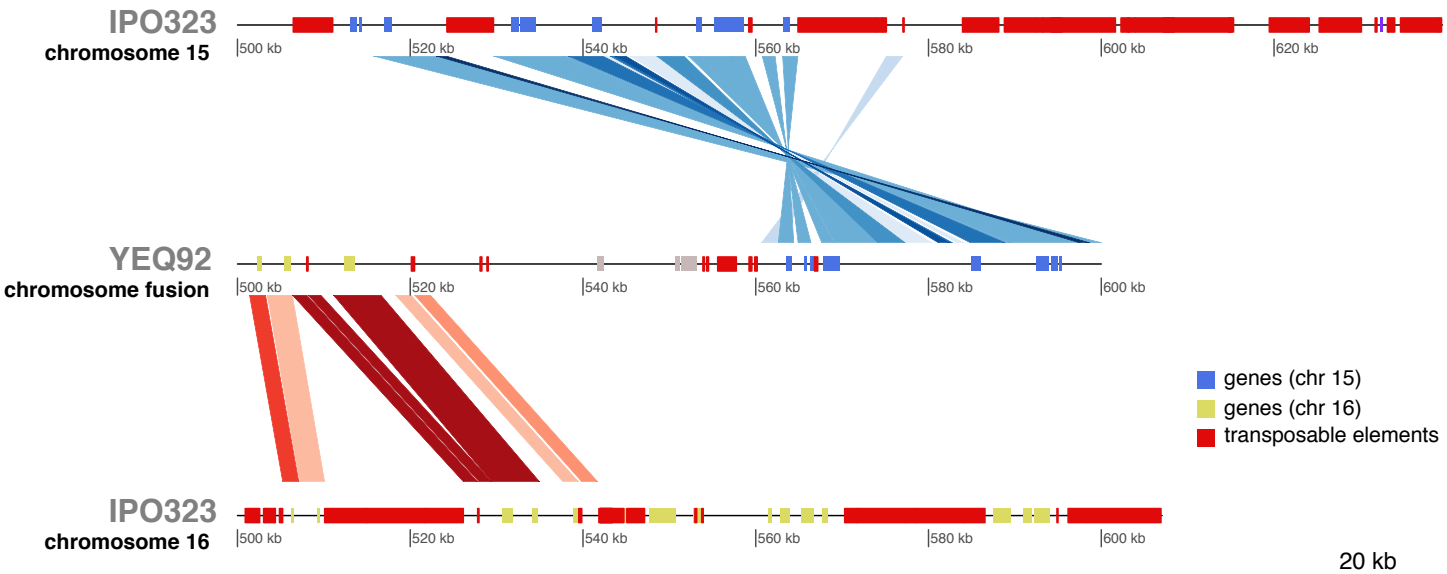
B

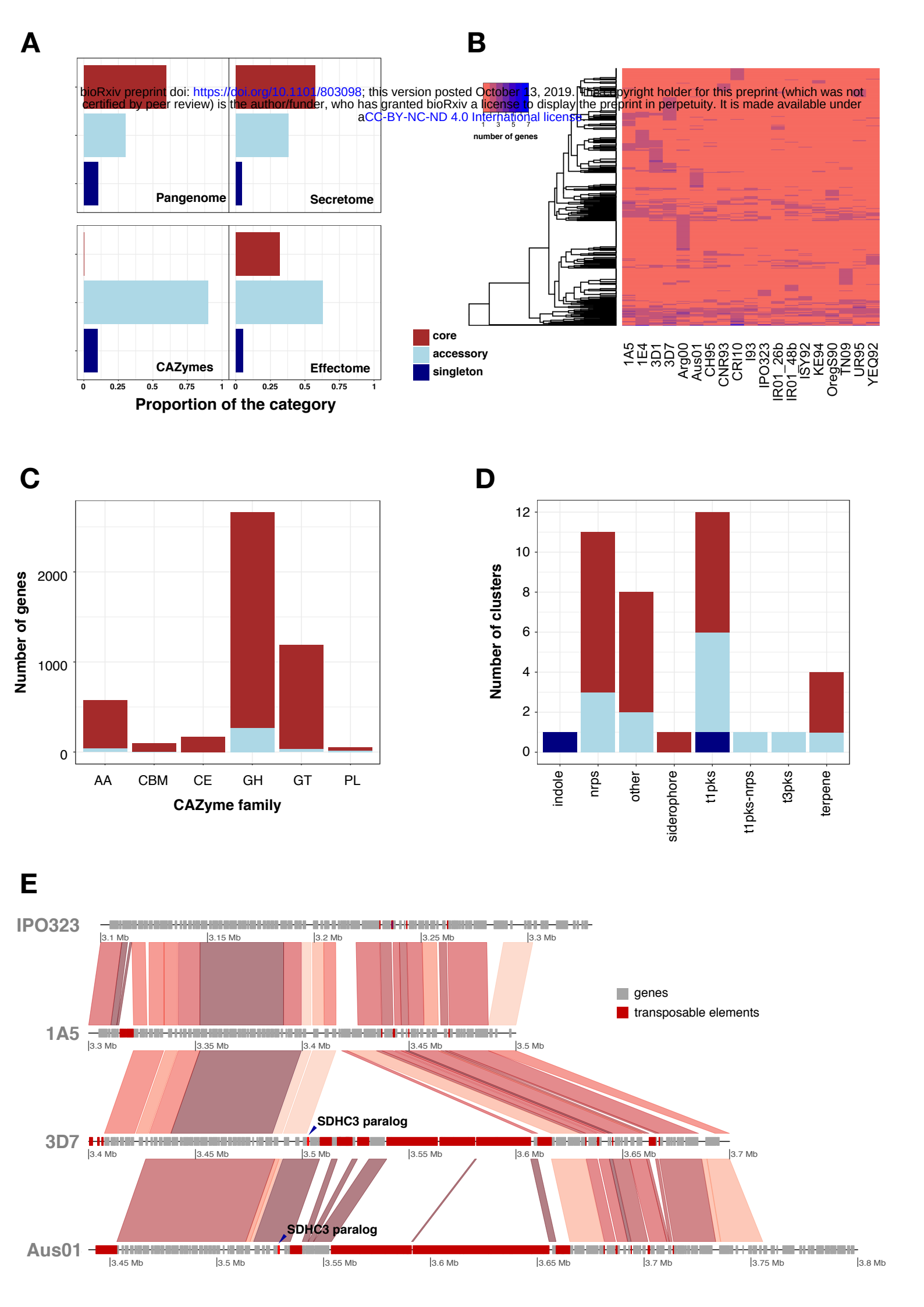


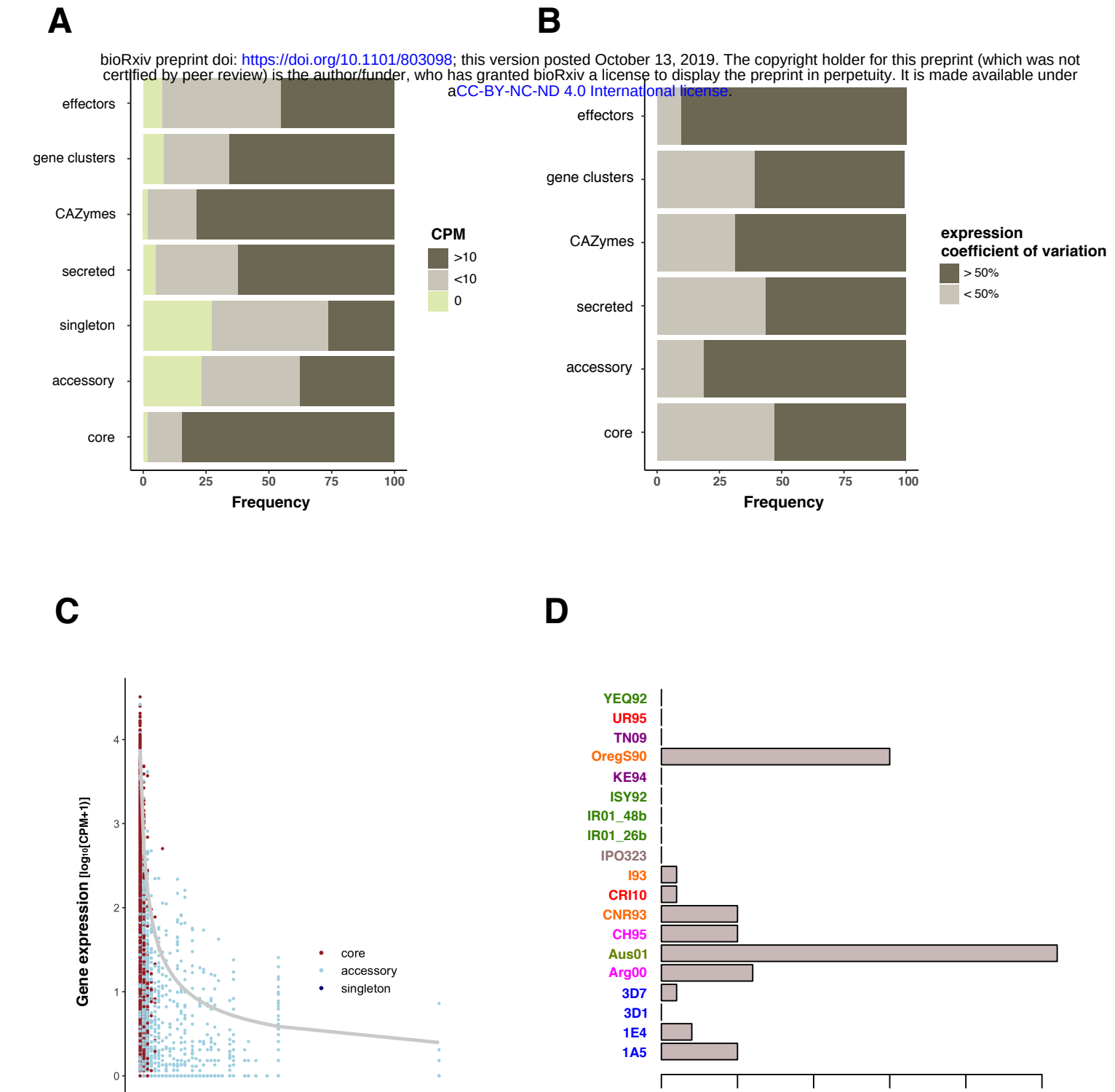
C



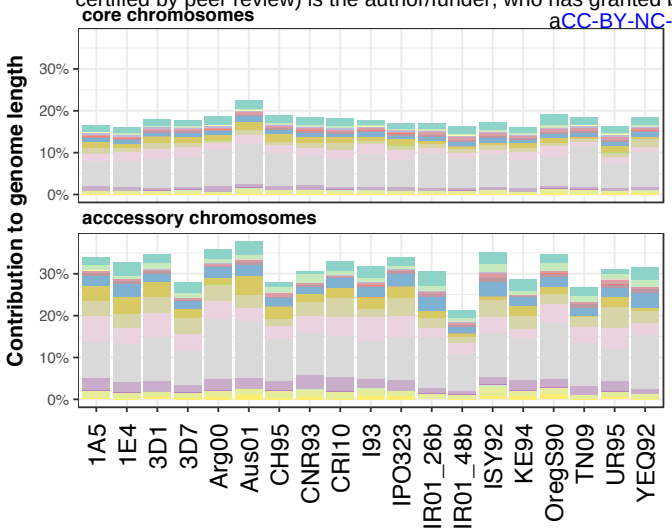
D



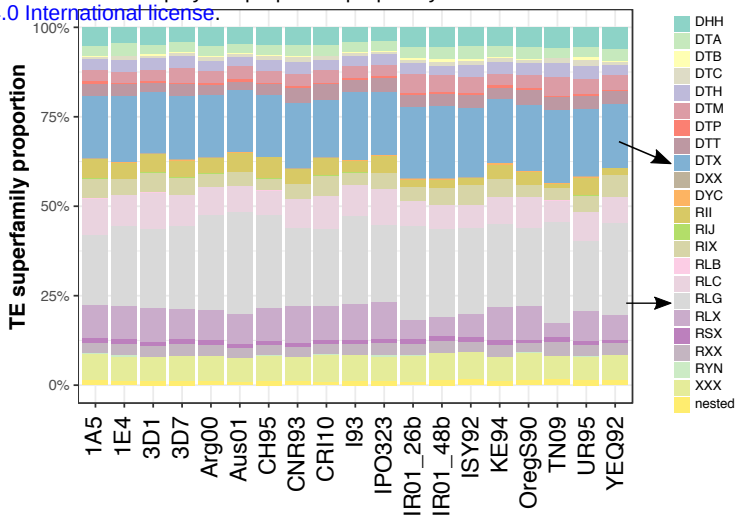




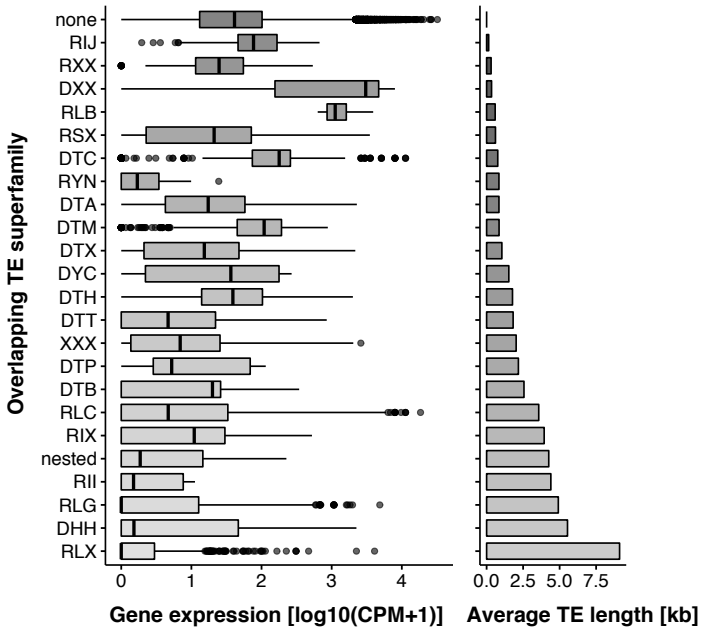
A



B



C



D