Machine-learning predicts genomic determinants of meiosis-driven structural variation in a eukaryotic pathogen

Thomas Badet[1], Simone Fouché[1,2], Fanny E. Hartmann[3], Daniel Croll[1,*)]

[1] Laboratory of Evolutionary Genetics, Institute of Biology, University of Neuchâtel, Neuchâtel, Switzerland
[2] Plant Pathology, Institute of Integrative Biology, ETH Zurich, Zurich, Switzerland
[3] Ecologie Systématique Evolution, Bâtiment 360, Univ. Paris-Sud, AgroParisTech, CNRS, Université Paris-Saclay, 91400 Orsay, France

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

Keywords: structural variation, complete genomes, population genomics, machine-learning, meiotic crossovers

Running title: Determinants of meiosis-driven structural variation

Species harbour extensive structural variation underpinning recent adaptive evolution and major disease phenotypes. Most sequence rearrangements are generated non-randomly along the genome through non-allelic recombination and transposable element activity. However, the causality between genomic features and the induction of new rearrangements is poorly established. Here, we analyse a global set of telomere-to-telomere genome assemblies of a major fungal pathogen of wheat to establish a nucleotide-level map of structural variation. We show that the recent emergence of pesticide resistance has been disproportionally driven by rearrangements. We used machine-learning to train a model of indel and translocation events based on 30 chromosomal sequence features. We show that base composition and gene density are the major determinants of structural variation. Low-copy LINE retrotransposons and specific DNA transposons explain most indel and translocation dynamics, respectively. We retrain our model on *Arabidopsis thaliana* and show that our modelling approach can be extended to more complex genomes. Finally, we analysed complete genomes of haploid offspring in a four-generation pedigree. Meiotic crossover locations were enriched for newly generated structural variation consistent with crossovers being mutational hotspots. The model trained on species-wide structural variation predicted 79% and 99% of the translocations and indels in the pedigree, respectively. The predictive power highlights causality between specific sequence features and the induction of chromosomal rearrangements. Our work demonstrates that training sequence-derived models can accurately identify regions of intrinsic DNA instability in eukaryotic genomes.

Main text

Structural variation including duplications, inversions, insertions and deletions is often the most common form of genetic variation within species [1–3]. A wide range of phenotypes including human diseases, traits for crop improvements and pathogen virulence are associated with chromosomal rearrangements [4–7]. Breaks in genome collinearity can have important consequences for the faithful exchange of alleles during meiosis [8–10]. Hence, chromosomal rearrangements can also underpin adaptation stemming from co-adapted alleles including sex determination loci [11,12]. Structural variation arises either from non-allelic recombination between repetitive regions in the genome or the activity of transposable elements [13,14]. However, predicting the likelihood of sequence rearrangement events remains challenging. In most species, the repetitive sequence content is poorly resolved at the nucleotide-level leading to fuzzy predictions about fragile sites that are prone to rearrangements [15]. Furthermore, analysing segregating structural variants at the population level only partially reflects sequence rearrangement events because selection tends to remove deleterious variants over generations.

Most sequence rearrangements are generated through double-strand break repair [16]. Germline double-strand breaks are typically induced during meiotic recombination [1]. Microhomology, tandem repeats and large-scale duplicated sequences can be at the origin of fragile sites and cause chromosomal instability [17,18]. However, whether and how individual genetic and epigenetic sequence features attract higher rates of sequence rearrangements is poorly understood [16,19]. Pangenome studies across eukaryotes showed that structural variation segregating within species can exceed single nucleotide variation [3]. Species-wide structural variation is often heterogeneously distributed along chromosomes and shows strong associations with epigenetic marks and the activity of transposable elements [9,20–22]. Purifying selection in populations removes deleterious chromosomal rearrangements over generations, hence skewing the spectrum towards neutral or nearly neutral structural variation [8].

Here, we establish a nucleotide-level map of structural variation based on 19 chromosome-level assemblies of the fungal pathogen *Zymoseptoria tritici*, which is one of the most devastating pathogens of wheat causing over a billion euros in losses in Europe alone [23,24]. Based on genome-wide association mapping of 24 phenotypes characterizing the life cycle of the pathogen, we show that top SNPs tend to map near sequence rearrangement loci and that the recent emergence of pesticide resistance was disproportionally driven by structural variation. We applied machine-learning integrating 30 sequence metrics to train a species-wide model to predict causal factors underlying structural variation. We show that the model has high precision and accuracy to predict insertion/deletion (indel) events and we successfully apply the same approach to predict structural variation in the pangenome of *Arabidopsis thaliana*. To disentangle the effect of purifying selection from the mechanisms promoting structural variation, we generated chromosome-level assemblies of nine haploid progeny over four generations. We assess the power of the pangenome-derived model and quantify the impact of parental background on the generation of sequence rearrangements.

Pangenome-wide collinearity analyses reveal hotspots of structural variation

To build a model of how structural variation arises within a species, we analysed the high-quality pangenome of the fungal wheat pathogen *Z. tritici* (Figure 1A). The gapless chromosome-level assemblies are among the most complete pangenome of a eukaryotic species. Homologous chromosomes show substantial breaks in collinearity across the global species range driven both by indels and inversions (Figure 1B). Eight of the 21 chromosomes are accessory in the species with presence/absence polymorphism and can undergo drastic rearrangements such as the fusion of two chromosomes (Figure 1B-C).

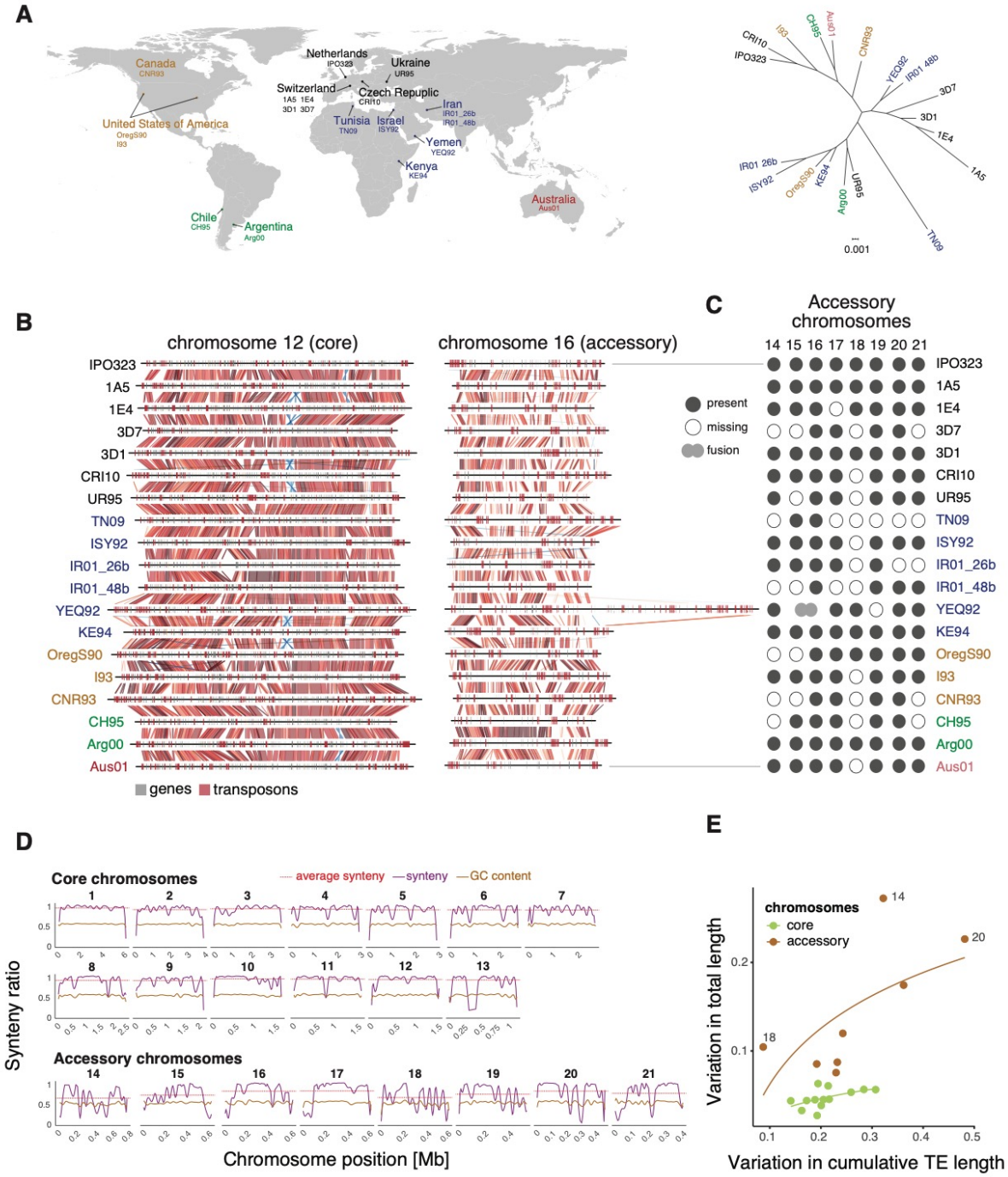


Figure 1: Species-wide genome structure of a highly polymorphic microbial eukaryote. A. Worldwide map showing the country of origin of the 19 *Zymoseptoria tritici* isolates and phylogenetic relationships based on 50 random orthologous genes. Colors indicate the continent of origin. **B.** Synteny plot of core chromosome 12 and accessory chromosome 16. The red and blue gradients represent the percentage of identity given BLAST alignments for syntenic and inverted segments, respectively. Darker colours show higher identity values. **C.** Presence/absence variation of accessory chromosomes. The isolate from Yemen carries a hybrid chromosome following a fusion between chromosomes 15 and 16. **D.** Genome-wide synteny ratios and GC content along chromosomes. **E.** Variation in the total transposable element content per chromosome. Length variation is expressed as the coefficient of variation. Curves show the best logarithmic fit.

We performed pairwise complete genome alignments against the reference genome to analyse whether the extent of structural variation within the species varied among chromosomal regions. Collinearity (or synteny) can be expressed as a ratio of syntenic segments over the total number of syntenic and non-syntenic segments across windows (Figure 1D). We averaged synteny scores across the pangenome and found distinct profiles for each chromosome. Telomeric regions were universally of low synteny consistent with higher densities of repetitive sequences and more opportunity for non-allelic recombination. Core chromosomes showed marked drops in synteny in regions of major sequence rearrangements including large inversions on chromosomes 3 and 12. The inversion on chromosome 12 present in the 3 isolates 1E4, CRI10 and KE94 encompasses near 3% of the total chromosome length and affects between 11 and 17 genes (Figure 1B). Accessory chromosomes showed greatly reduced synteny scores compared to core chromosomes with pronounced signatures of sequence degeneration. Overall, 40% of the species gene content is not fixed (*i.e* accessory genes). Chromosomal length variation was positively correlated with transposable element content (Figure 1E). Essential (core) chromosomes showed a cap at ~6% size variation despite significant variability in transposable elements. Accessory chromosomes showed significantly relaxed constraints on chromosome length and transposable element content.

Long-read based structural variant analyses

Local synteny scores do not resolve precise sequence rearrangement events. Hence, we analysed PacBio long reads mapping for 18 individuals against the reference genome (IPO323). Evaluating evidence from split read, mismatch and coverage data, we identified a total of 21’717 structural variants localized at base-pair resolution. Detected variants included duplications (0.85%), indels (56.6%), inversions (0.95%) and translocations (41.6%) (Table S1). Core chromosomes bear on average less structural variants compared to accessory chromosomes but showed a higher proportion of translocation events relative to all structural variant types, suggesting increased transposon activity (Figure 2A). Most indel variants (~70%) are short (<100 bp) and only 3% are larger than 10 kb (612 indels; Figure 2B). Similarly, most duplications (84%) are shorter than 100 bp. In contrast, 56% of inversions are longer than 10 kb. We found that translocations, duplications and inversions show a pronounced shift towards low allele frequencies compared to genome-wide SNPs consistent with purifying selection (Figure 2C). However, we also show that indels on accessory chromosomes are likely under relaxed selection.

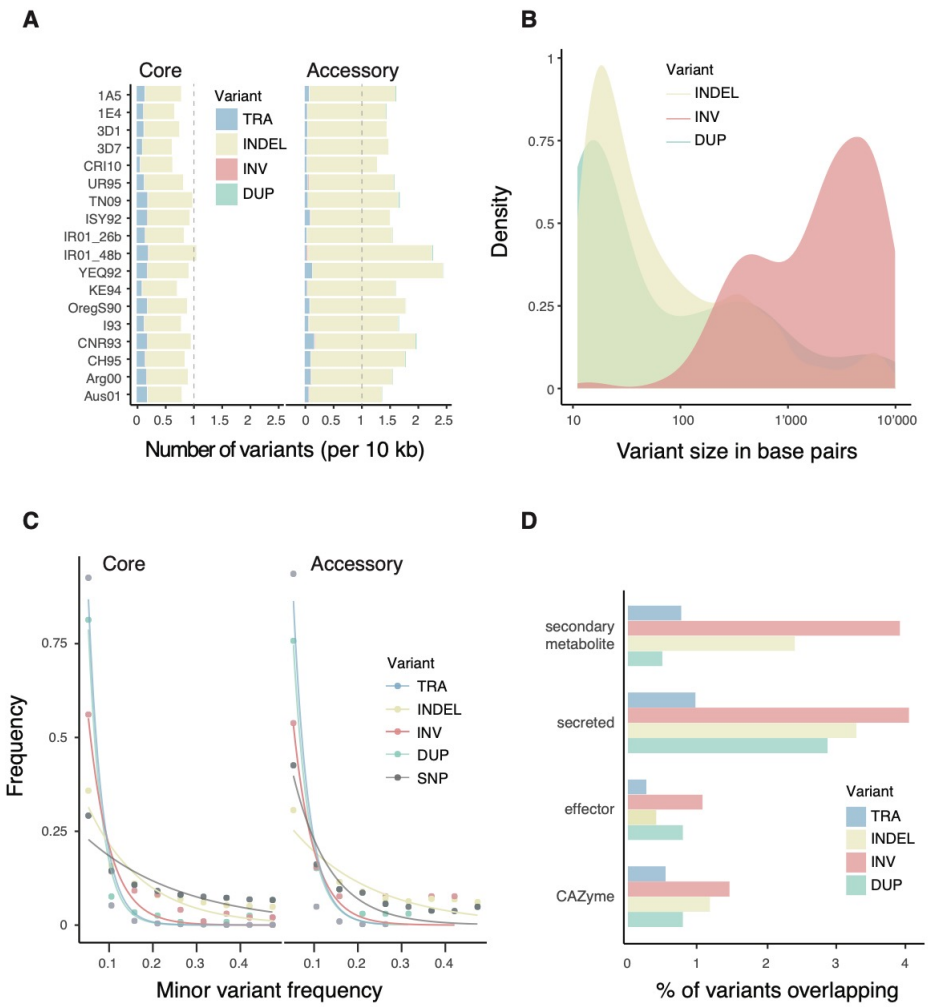


Figure 2: Structural variant heterogeneity across chromosomes and associations with genes. A. High-confidence structural variants among isolates. Translocations (TRA), indels (INDEL), inversion (INV) and duplications (DUP). **B.** Size distribution of the four structural variant types. Translocations were excluded as they have no size assigned. **C.** Minor allele frequency distribution. Curves show estimates of the exponential decay calculated for each variant type. **D.** Proportion of each type of variant overlapping with different gene category playing a role in fungal pathogenesis including secondary metabolite clusters, genes encoding for predicted secreted proteins, effectors and carbohydrate-active enzymes (CAZymes).

Structural variation underpinning fitness-related trait expression

Major phenotypic innovations originate from chromosomal rearrangements [10,25,26]. Duplications and deletions affecting coding sequences can promote phenotypic change. We show that more than 5% of the genome is affected by structural variation and that ~20% of the rearrangements overlap coding sequences (Figure S1). Inversions are balanced rearrangements that preserve sequence and thus gene content. We find that inversions are more likely to affect coding sequences (~79%; Figure S1). Genes encoding effector proteins with likely functions in host manipulation and genes encoding carbohydrate-degrading enzymes are overall less affected by structural variation (Figure 2D). Nevertheless, the telomeric effector gene *AvrStb6* known as the major genetic factor controlling disease development on a broad range of wheat cultivars is affected by a local inversion polymorphism (Figure

158 S2). The subtelomeric region harbouring *AvrStb6* shows length variation of 30-250 kb and is one of the
159 top 5% most rearranged chromosomal niches (Figure S2).
160
161 To quantify the contribution of structural variation to phenotypic trait variation, we performed genome-
162 wide association analyses on 24 traits representative of the life cycle of the pathogen. The phenotypes
163 included virulence on different wheat cultivars, temperature resilience and fungicide resistance. The
164 mapping population consisted of 106 isolates from all major wheat producing areas [27] (Figure 3A). We
165 identified 18 traits with significant SNPs located in regions affected by structural variation. For melanin
166 production, ~9% of the associated SNPs overlap regions with indel variation (Figure 3B-C). Additionally,
167 we identified 1'822 SNPs associated with tolerance to propiconazole, a widely used fungicide of the
168 demethylation inhibiting (DMI) family. A total of 7% of fungicide tolerance SNPs overlap regions with
169 indel variation. Significantly associated SNPs were over-represented near translocations, indels and
170 duplications for propiconazole tolerance related traits (odds ratio > 2; Figure 3D). Two fungicide
171 tolerance SNPs were near major facilitator superfamily (MFS) genes. One of the MFS genes was also
172 affected by a TE insertion and a loss of expression (Figure 3F). Multiple additional insertions near the
173 two MFS genes are associated with transcriptional variation (Figure 3F-G). Taken together, regions
174 underlying major phenotypic trait variation undergo high rates of structural variation.
175

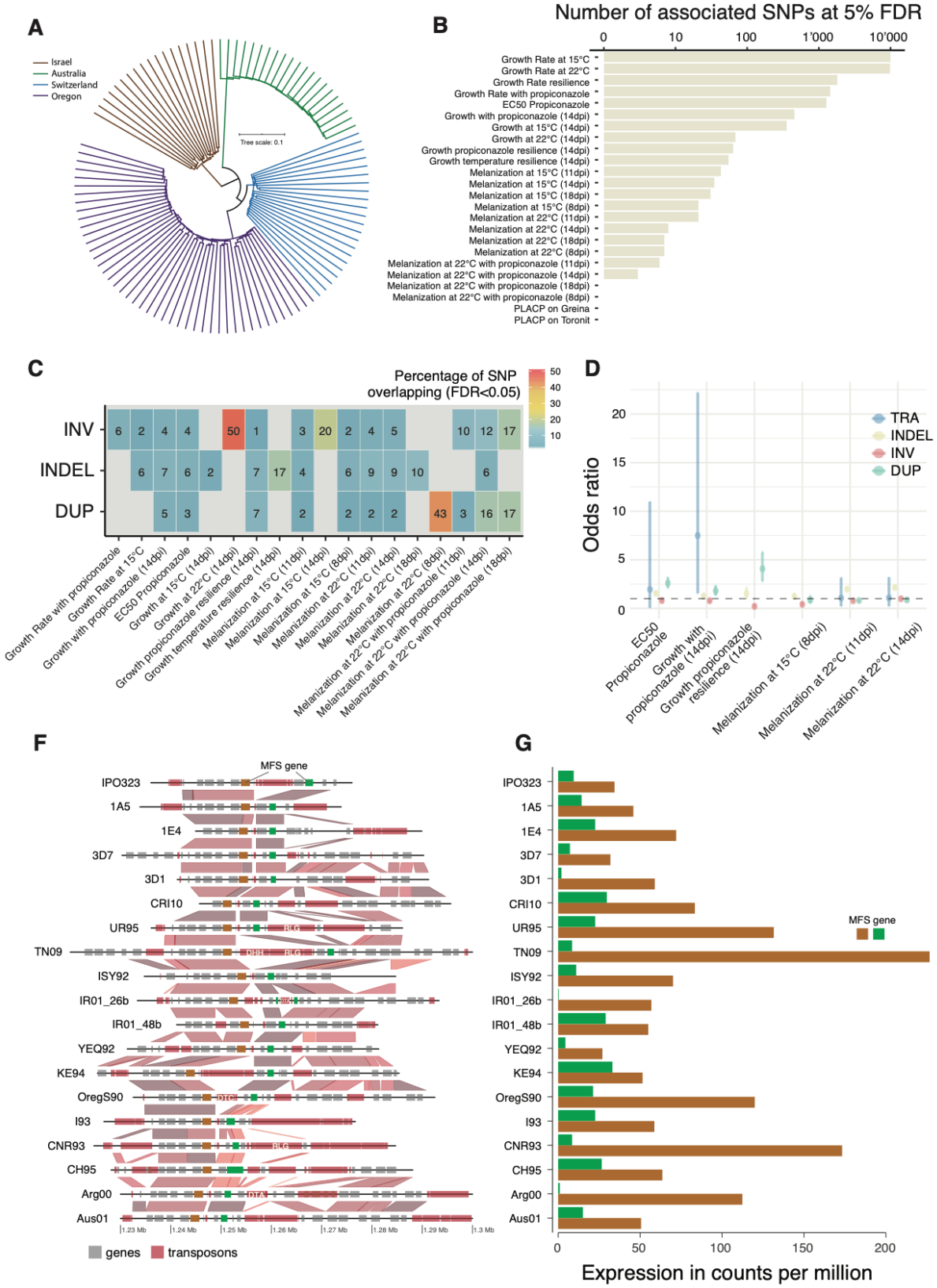


Figure 3: Structural variation drives phenotypic trait evolution. A. Unrooted phylogeny of 106 isolates used for genome-wide association mapping of 24 traits. **B.** Number of significantly associated SNPs across phenotypic traits (5% false discovery rate; FDR). **C.** Percentage of significantly associated SNPs overlapping with either indels, inversions or duplications across different phenotypes.

Translocations are not shown as they have no length assigned. **D.** Odds ratio of associated SNPs overlapping structural variants. Ratios are shown for six phenotypic traits associated with fungicide resistance including colony melanisation. **F.** Synteny plot of a locus associated with fungicide resistance on chromosome 13. Two genes encoding major facilitator superfamily (MFS genes) are coloured in brown and green. Nearby transposable elements are annotated according to their superfamily (shown in white). **G.** Relative expression of the two MFS genes, coloured as in (**F**).

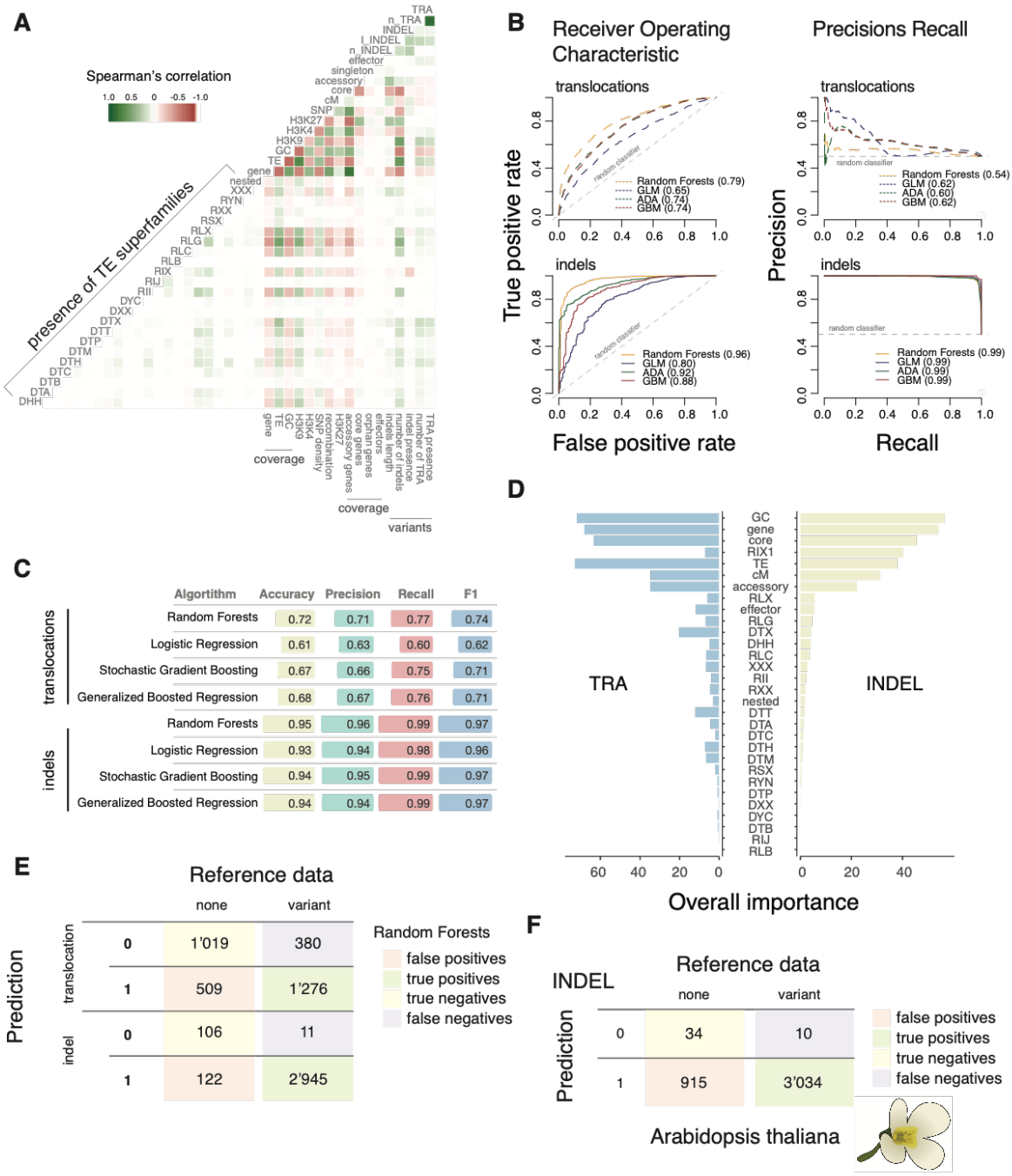
Chromosomal niches of structural variation hotspots

Structural variation shows a highly heterogeneous distribution along chromosomes. Hence, specific sequence features may contribute disproportionally to the likelihood of rearrangements. In order to systematically identify associations across the genome, we analysed variation in gene and transposable element densities, locations of core and accessory genes, isochores (*i.e.* AT-rich regions), recombination rates and histone methylation marks for a total of 35 different sequence metrics. Gene and transposon coverage are strongly negatively correlated and underlie strong genome compartmentalization (Figure 4A, Table S2). The highly abundant transposable element superfamily *Gypsy* shows the strongest correlation with TE regions. Transposable elements co-localize both with repressive and facultative heterochromatin marks ($H3K9_{met3}$ and $H3K27_{met3}$, respectively). We find that all four types of structural variations are over-represented near TEs with the strongest overrepresentation for translocations (Figure S3). This shows that transposable elements are drivers of structural variation in specific compartments of the genome. Indels showed no preferential associations across the genome but tended to be more abundant near accessory and depleted near core genes. Repressive heterochromatin marks and transposable element density (*Gypsy* elements in particular) were strongly predictive of indel length. Overall, chromosomal rearrangements hotspots are associated with histone methylation marks, specific transposable elements and accessory gene content.

A pangenome-informed model accurately predicts structural variation

The species-wide structural variation is the joint product of intrinsic instability of specific sequence features and selection acting on structural variants. To build a predictive model accounting for these different forces, we integrated 30 sequence features including the density and identity of transposable elements, the presence of virulence factors, local recombination rates, GC content as well as information on gene dispensability (see Table S3 for metric details). We evaluated four different machine-learning algorithms with three times 10-fold cross-validations. We selected the best models based on accuracy and maximized for the percentage of correctly classified instances overall. For both indels and translocations, we obtained the best results with the random forests method, yielding 0.95 and 0.72 accuracy, respectively (Figure 4B-C). The best predictors included GC content, recombination rate, the presence of core genes and transposable elements for both indels and translocations. Weaker predictors included the presence of retrotransposons for indels and DNA transposons for translocations (Figure 4D). The model based on random forests performed better for indels with a 0.99 recall and 0.96 area under the receiver operating characteristic (ROC) curve compared to translocations (Figure 4B-C). The model also yielded the lowest number of false negatives for indels (Figure 4E). The final models

222 for indels and translocations tended to call false positives variants rather than false negatives (122/11
223 and 509/380 for indels and translocations, respectively).
224



225

226 **Figure 4: A pangenome-informed model predicts the occurrence of structural variation. A.**
227 Correlation analyses of 35 genome-wide metrics in windows with the two most abundant structural
228 rearrangements (indels and translocations). **B.** Receiving operating characteristics and precision recall
229 curves of the final models trained using translocations (upper part) and indels (lower part). Four
230 algorithms were analysed including random forests, logistic regression (GLM), stochastic gradient
231 boosting (ADA) and generalised boosted regression (GBM). Numbers in parentheses indicate the area
232 under the curve for each of the four models. **C.** Summary statistics of the final models for translocations
233 and indels using the four algorithms. Accuracy describes the model capacity at only detecting true
234 variants. Precision measures the ability to predict true variants. Recall measures the ability to detect all
235 true variants and the F1 score considers the harmonic mean of both precision and recall. **D.** Relative

importance of each of the 30 sequence features used in the training of the random forests model. The importance of individual sequence features is shown as the unscaled averages of the class-specific scores. **E.** Confusion matrix resulting from indel and translocation predictions using the two respective models trained with the random forests algorithm. **F.** Confusion matrix of the indel predictions performed for the *Arabidopsis thaliana* pangenome.

We replicated our approach to predict structural variation in the model plant *Arabidopsis thaliana.* The pangenome shows hotspots of rearrangements associated with transposable elements and depleted for recombination [28]. We parametrized the random forests model based on Col-0 reference sequence metrics including transposable elements, core and accessory gene coverage, recombination rate and nucleotide-binding and leucine-rich repeats (NLRs) gene coverage [28,29]. Over the total of 3'993 sequence windows, we correctly predicted 99.8% of the indel polymorphism (n = 2'817) with an accuracy of 0.78 and a recall of 0.99 (Figure 4F). The successful application of machine-learning to predict structural variation suggests that eukaryotic genomes rearrange based on clearly identifiable mechanisms.

The predictive power of the species-wide model in an experimental pedigree

Structural variation within species are likely produced *de novo* during meiosis or accumulate in somatic tissue. Nearly all observed variants are expected to be neutral or nearly neutral because of purifying selection. To experimentally investigate mechanisms underlying the formation of new rearrangements, we produced a haploid pedigree of four generations of experimentally crossed isolates (Figure 5A). The parental isolates were collected from Swiss wheat fields and are included in the species pangenome analyses. We crossed the isolates using an established protocol to produce sexual progeny on inoculated wheat leaves. Sexual progenies were harvested and sub-cultured as single genotypes for further experiments. The haploid stage is dominant in *Z. tritici* meaning that any progeny has already undergone meiotic recombination of the two haploid parental genomes. Hence, chromosomal rearrangements triggered by sexual reproduction can be scored directly in an F1. We used progeny and parental strains to generate a total of five crosses with the first one being between the two parents and four subsequent backcrosses with one of the two parents (Figure 5A). The total of 9 progeny were separated each by one to four rounds of meiosis from the parents (see Methods). We produced gapless, chromosome-scale assemblies using PacBio sequencing. The genome assemblies revealed that the accessory chromosome 17 displayed significant instability undergoing large segmental duplications (Figure S4). Four accessory chromosomes were lost after the third and fourth round of meiosis likely due to non-disjunction events (Figure 5C). Progeny genomes showed large drops in synteny compared to parental isolate 1A5 consistent with the large inversions segregating among the parental genomes (Figure S5-S6). Regions of low synteny among progeny genomes were largely congruent with low synteny regions in the species pangenome, suggesting that same sequences promote recurring structural rearrangements. Following the procedures for structural variant calling to train the species-wide model, we identified rearrangements in progeny genomes by aligning PacBio long reads against one parental genome (*i.e.* 1A5). We filtered out any structural variant segregating between the two parental genomes. This ensured that we focused on structural variants that have originated within the pedigree. We identified a total 2'533 *de novo* rearrangements including 1'841 indels, 630 translocations,

30 inversions and 32 duplications matching the distribution and size observed in the species-wide genome collection (Figure 5B; Figure S7A; Figure S7B; Table S4). We found no significant difference in the rate or type of new rearrangements between the four meiotic rounds in the pedigree (two-way analysis of variance *p*-value = 0.899; interaction test for meiotic round and variant type). These observations are consistent with the patterns in the species-wide structural variant set, where genomes showed similar degrees of differentiation. Therefore, the few rounds of meiosis in the pedigree were sufficient to reproduce major aspects of the species-wide set of chromosomal rearrangements.

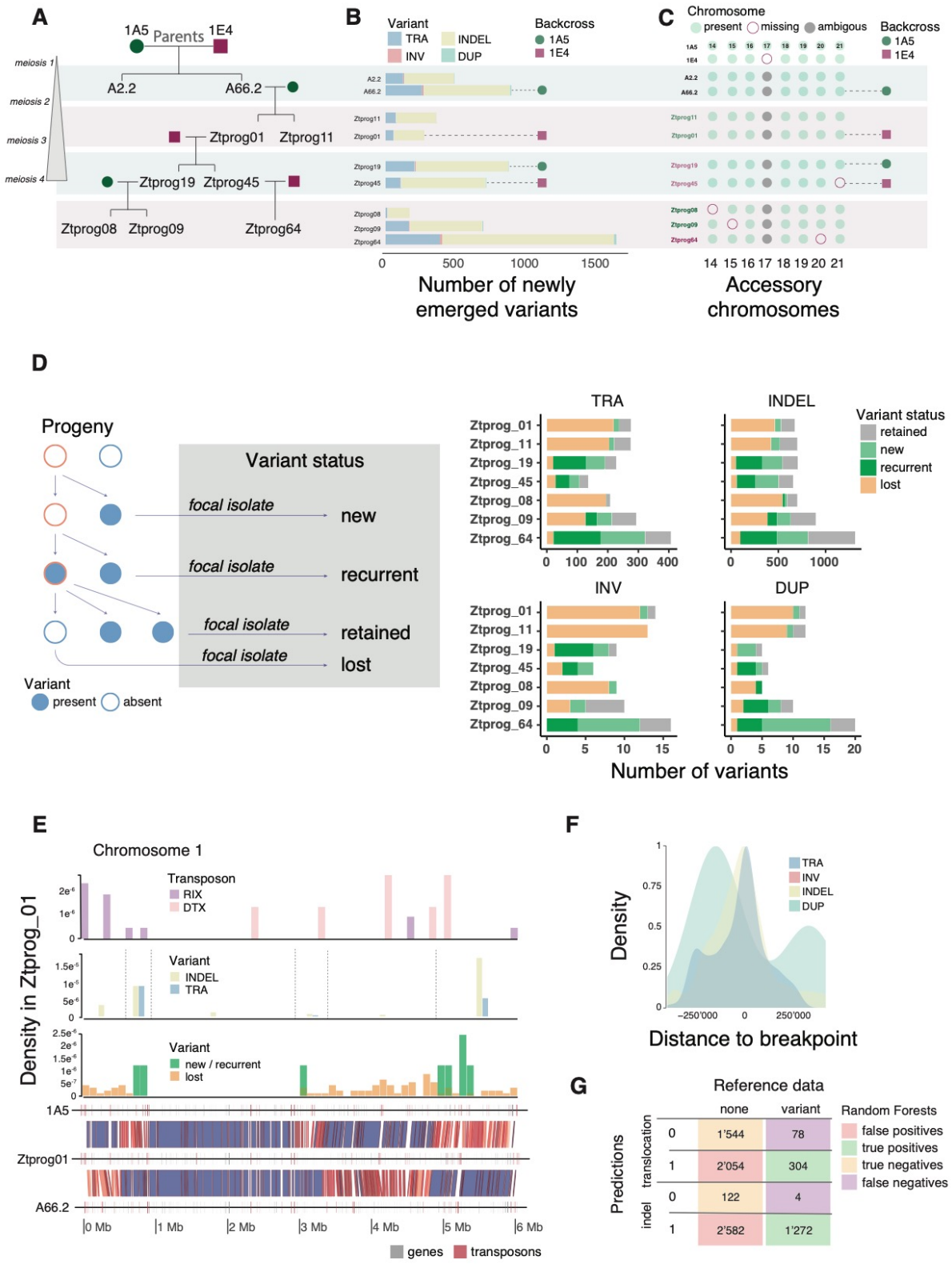


Figure 5: Predictions of structural variation generated by meiosis throughout a pedigree. A. Pedigree chart showing the relationship between the 9 progeny isolates retrieved across four rounds of meiosis. Coloured rounds and squares indicate the parental isolate used for backcrosses. **B.** Number of newly emerged high-confidence variants detected in each of the progeny isolates. Isolates are

ordered as in the pedigree chart and coloured rounds and squares indicate the parental isolate used for backcrosses. Translocations (TRA), indels (INDEL), inversions (INV) and duplications (DUP). **C.** Presence/absence of accessory chromosomes across the progeny set and the parental isolates. Chromosome 17 is absent in the 1E4 parent and shows complex segregation in the pedigree including truncations and duplications. **D.** Classification of variants according to their origin and persistence in the pedigree. Variants present in a focal progeny and the direct parents were defined as "retained" variants. Variants present in a focal progeny but in none of its ancestors in the pedigree were called "new" variants. Variants present in a progeny, absent in its direct parents but present already in an ancestor were defined as "recurrent" variants. Finally, variants absent from a progeny but present in both parents were called "lost" variants. **E.** Chromosome 1 synteny plot of the progeny Ztprog01 and its two parental isolates (lower panel). Regions of high identity are coloured in blue and highlight the three recombination events. Chromosome density of new and recurrent as opposed to lost structural variants (second panel from bottom). Number of new indels and translocations (second panel). Density of the two transposable elements contributing most to structural variation based on machine learning. RIX: unknown LINE retrotransposons, DTX: TIR DNA transposons. **F.** Density plot of new structural variants related to the closest crossover breakpoints on chromosome 1. **G.** Confusion matrix resulting from indel and translocation model predictions using the two respective models trained with the random forests algorithm.

Our design of crosses and backcrosses enables to retrace the emergence of structural variants through individual rounds of meiosis (Figure 5D). To dissect the history of new rearrangements along the crosses, we assigned each variant to one of the four categories. (1) New variants generated in the most recent round of meiosis. Overall, isolates resulting from a backcross with the 1E4 parent tend to show higher numbers of new variants (Figure 5B-D). The progeny isolate from the fourth meiosis backcrossed with the 1E4 parent shows the highest number of new structural variants (Ztprog64). (2) Lost variants compared to the parental genomes were most frequent as expected in the first two backcrosses. (3) Retained variants shared by all progeny of a cross are rare and constant in the pedigree. (4) We detected recurring mutations in nearly all progeny beyond the second round of meiosis (*i.e.* identical variants generated multiple times independently). New and lost variants tend to occur in distinct regions of the chromosome as expected from crossovers re-introducing parental sequences (Figure S8). We dissected the mechanism by which new and recurrent rearrangements are generated. For this, we identified crossovers from whole-genome alignments between each progeny and the two respective parental genomes (Figure 5E). As expected, segments inherited from the backcrossed parental isolate were depleted of structural variants (Figure 5E). The two transposable element superfamilies (DTX and RIX elements) driving structural variation in the across the species showed no clear association with crossover events and new rearrangements. Interestingly, we accurately predicted an indel and a translocation event that occurred in the progeny near the two MFS genes associated with fungicide resistance in our GWA study. New structural variants overall tend to be generated near crossover breakpoints (Figure 5E). To retrace how crossover generate new rearrangements, we used synteny to localize crossover breakpoints on chromosome 1 through the 4 meiotic cycles (n = 18 breakpoints). We show that indels and translocations are enriched in proximity to crossover breakpoints consistent with recombination-mediated double-strand breaks causing sequence rearrangements (Figure 5F).

Sequence rearrangements generated through meiosis can impact fitness and therefore challenge the analyses of structural variation in natural populations. Progeny from our pedigree were minimally cultured on nutrient-rich medium likely reducing the strength of selection on slightly deleterious

337 rearrangements. To test the predictive power of the species-wide model of structural variation, we re-
338 parametrized the input metrics for the one parental genome (*i.e.* 1A5) in the pedigree. Using the same
339 set of sequence features across windows, we analysed the occurrence of a total of 382 and 1'276
340 translocations and indels, respectively. In contrast to the *Z. tritici* species-wide model and the *A. thaliana*
341 model, the logistic regression method yielded the highest accuracy and the boosting implementation
342 yielded the best recall for translocations (0.52 and 0.81, respectively) (Figure S9). For indels, the
343 random forests achieved the highest accuracy and precision (0.35 and 0.33, respectively). Using the
344 random forests model, we were able to accurately predict 79% of the translocations and 99% of the
345 indels generated by meiosis in the pedigree (Figure 5F). The lower accuracy of the model applied to
346 the pedigree compared to the species-wide model stems from an over-prediction of structural
347 rearrangements (Figure 5G). We investigated the genomic context of the over-predictions using synteny
348 ratios between the parental genomes. For indels, we find that true negatives tend to be in regions of
349 low synteny (Figure S10). Furthermore, true positive translocations tend also to be in regions of low
350 synteny compared to false positives (Wilcoxon rank sum test *p*-value = 6.335e-8, average synteny of
351 0.89 and 0.94 respectively). We find no significant difference in synteny for false and true positives indel
352 calls. This suggests that model over-predictions are not caused by variation in synteny alone (Figure
353 S10). Indeed, false positive indels but not translocations are found in region of lower recombination rate
354 compared to true positives (Wilcoxon rank sum test, *p*-value < 0.0001). In agreement, we find that the
355 model predicts less false positive indels within 5 kb of crossover events. In addition, the ratio of false
356 and true discovery rates among progeny is independent of the round of meiosis in the pedigree (Figure
357 S11). In general, the performance of the models increases towards the last round of meiosis (*i.e.* with
358 an increase in detected indels and translocations in; see *e.g.* Ztprog64). Altogether, we show that
359 meiosis contributes significantly to the formation of structural variation in the species, and that machine
360 learning enables the precise prediction of indel events across rounds of meiosis.
361
362
363
364 We show that integrating sequence data into an optimized model enables to make powerful predictions
365 where structural variation is generated along chromosomes. Similar sequence features underpin the
366 likelihood of rearrangements both in the sets of species-wide genomes of *Z. tritici* and *A. thaliana*. The
367 major factors leading to higher rates of sequence rearrangements are low GC and gene content. Low
368 GC-content is associated with a variety of sequence features including short repeats and inactivated
369 transposable elements. The higher rate of sequence rearrangements triggered by these regions is likely
370 a combined effect of relaxed selection and higher degrees of homology along the genome. Indeed, AT-
371 rich regions are more likely to form secondary structures that disturb DNA replication [30]. Structural
372 variation occurring in gene dense regions is more likely to be deleterious. Here, we show that recent
373 adaptation to resist pesticide applications has been primarily driven by structural variation. We identified
374 rearrangements near two major facilitator superfamily (MFS) encoding transporters of xenobiotics. This
375 shows that strong positive selection on individual rearrangements can exceed purifying selection acting
376 on gene-rich, conserved regions.
377

Germline sequence rearrangements are mainly generated during meiosis. However, selection on deleterious rearrangements can potentially skew how population-level structural variation relates to underlying sequence features. The haploid fungus *Z. tritici* enables observing rearrangements generated by meiosis already at the F1 stage and growth on rich medium minimizes the breadth of selection compared to natural environments. Meiotically-induced structural variants are often the result of non-allelic homologous recombination [16]. Genome-wide structural variation breakpoints in the human genome revealed that such NAHR events are associated with higher recombination rates, but also higher GC content and open chromatin, suggesting conserved mechanisms across eukaryotes [18]. In *A. thaliana*, structural variation is negatively correlated with recombination rates likely due to structural variation being able to suppress crossover formation [9]. Here, we show that new variants are preferentially generated near crossovers showing that breakpoints can act as hotspots for creating structural variation including recurring variants.

Conclusions

Our species-wide model of structural variation enables powerful predictions about the occurrence of sequence rearrangements during meiosis. Our pedigree analyses additionally revealed a significant number of recurrent rearrangements at relatively short distances, which were generated independently multiple times. This shows that chromosomal instability can be mapped precisely in regard to chromosomal sequence features. Recurrent and nonrecurrent structural rearrangements in humans are thought to be generated by different mechanisms. NAHR between low-copy repeats is a major factor for the recurrent mutations underlying genetic disorders in humans [31]. Similarly, recurrent deletions and insertions influences developmental processes in sticklebacks [32] and cell aggregation in yeasts [33], respectively. The ability to predict the likelihood and nature of structural variants from sequence features alone provides a conceptual ground for the growing evidence of parallel evolution at the molecular level [34,35]. Here, we show that indels are most likely generated near low-copy retrotransposons. In contrast, translocations are generated primarily by high-copy DNA transposons. In both cases, the transposable elements showed no close homology to major repeat families indicating recent, lineage-specific repeat expansions driving sequence rearrangements. Genomic models trained on species-wide structural variation will have powerful applications in research on genetic disorders, assessing the persistence of genetic engineering in crop breeding and the development of evolutionarily stable antimicrobials.

409 **Methods**
410
411 **Experimental crossings**
412 We generated a four-generation pedigree started from a cross of the isolates 1A5 with 1E4
413 (PRJEB15648 and PRJEB20900 respectively) initially described in [36]. One haploid progeny was
414 selected for further sexual crosses. The isolate A66.2 was subjected to further crosses and backcrosses
415 with the parental isolates following the protocol described by [37]. The full pedigree is described in Figure
416 5A. Crosses were made by co-inoculating spores of the parental pairs on wheat plants and incubated
417 until the development of pseudothecia (sexual structures of the fungus). Ascospores ejected from
418 pseudothecia were isolated on water agar plates. Germinating ascospores were transferred for clonal
419 propagation and further DNA extraction.
420
421 **Species pangenome and pedigree analyses based on complete genomes**
422 The 19 isolate pangenome was first reported by Badet et al. [24]. Briefly, isolates were collected from six
423 different continents on multiple wheat cultivars and span the climatic gradient of the species. High-
424 molecular weight DNA was extracted from cultured spores using a phenol-chloroform-isoamyl alcohol
425 solution [38]. Libraries were prepared from at least 15 µg of DNA and the sequencing was performed
426 using P6/C4 chemistry on a PacBio RSII or Sequel instrument at the Functional Genomics Center,
427 Zurich, Switzerland. PacBio reads were assembled to chromosome-level using Canu v1.7.1 and Ragout
428 v2.1.1 and polished twice using Arrow v2.2.2 [39,40]. Genes were predicted using BRAKER v2.1 pipeline
429 with RNA-seq and protein data for intron calling [41–48]. Orthogroups were defined based on protein
430 homology using Orthofinder v2.1.2 [49,50]. Transposable elements consensus sequences were identified
431 *de novo* using RepeatModeler open-1.0.11 and annotated using RepeatMasker after manual curation
432 [51]. We used the consensus sequences from Badet et al, 2020 for repeat annotation. Briefly, repeats
433 were identified *de novo* with RepeatModeler software using the 19 pangenome isolates together with
434 the genome from the closest sister species *Zymoseptoria pseudotritici*. Consensus sequences were
435 manually curated and classified according to the GIRI Repbase [52]. Transposon superfamilies were
436 named based on the three-letter classification system [53]. The genome assemblies from the 19
437 pangenome isolates and the 9 progeny isolates were annotated using the curated consensus
438 sequences with RepeatMasker software. The cut-off value was set to 250 and simple repeats as well
439 as low complexity regions were ignored. We filtered out all the elements that were shorter than 100 bp.
440 Adjacent identical elements that were overlapping by more than 100 bp were merged. Elements that
441 were different and overlapping by more than 100 bp were considered as nested insertions and renamed
442 accordingly. Interrupted elements were grouped into a single element using minimal start and maximal
443 stop positions if separated by less than 200 bp.
444
445 The 9 progeny from the pedigree selected for genome sequencing were subjected to the same
446 procedure as describe above for DNA extraction and sequencing. Libraries were prepared from 15-31
447 µg of DNA after size selection with an 8 kb cutoff on a BluePippin system (Sage Science). The average
448 fragment length was ~15 kb and further sequencing was performed using either P4/C2 or P6/C4

449 chemistry on a PacBio RSII or Sequel instrument at the Functional Genomics Center, Zurich,
450 Switzerland. PacBio read assembly was performed using HGAP version 4 of the SMRTanalysis suite
451 (version 6, release 6.0.0.47841) [54]. HGAP was run with the default parameters, except for the minimum
452 seed read length, to initiate the self-correction. First, we produced assemblies for all chromosomes
453 except chromosome 17 using the cutoffs automatically chosen by HGAP. In order to improve the
454 contiguity of the assembly, we tested minimum seed read lengths of 8000, 10000, 12000, 15000, 20000.
455 Chromosomes which could not be assembled into a single contig were scaffolded into chromosomes
456 using Ragout (Reference-Assisted Genome Ordering UTility) version 2.2 with default parameters and
457 Sibelia for synteny block decomposition [40,55]. Unusual assemblies were analysed using Quiver
458 6.0.0.47835 with default settings as implemented in the SMRTanalysis suite and mapped reads were
459 inspected visually in IGV (2.4.10). The synteny of assembled progenies was compared to the parent
460 isolate 1A5 using Nucmer (mummer 3.23) with the parameter's delta-filter -i 80 -l 1000 -1 -q. A
461 mummerplot was generated with the parameters --filter --large --fat --layout-t postscript to visually
462 inspect assembly contiguity.
463
464 **Chromosome-wide synteny ratios**
465 To estimate the sequence collinearity between multiple chromosome-level assemblies, we adapted the
466 method by Jiao et al, 2020. Briefly, we performed whole-genome alignments of 18 pangenome isolates
467 against the IPO323 reference using nucmer v3.1 allowing for multiple matches (--maxmatch option).
468 We filtered the alignments for a minimal identity of 90% and a minimal single match length 100 bp using
469 the delta-filter in MUMmer v3.1. We then extracted the alignment coordinates using show-coords with
470 the -THrd options and mapped the position of syntenic and rearranged regions using the syri v1.0
471 software [56]. Regions absent from the reference genome where removed. We calculated the intersect
472 between the syntenic and rearranged regions given by syri and 10 kb windows along the reference
473 genome using bedtools intersect (v2.29.2). The synteny ratio was then calculated as the number of
474 syntenic blocks (named "SYN" and "SYNAL" in the syri output) over the number of total syntenic and
475 non-syntenic blocks in each 10 kb window. For accessory chromosomes not shared by all isolates, the
476 total number of chromosomes included in the synteny ratio was lower than for core chromosomes. For
477 the pedigree analyses, we applied the same method to calculate the synteny ratio of the 9 progeny
478 genomes against the 1A5 parental genome.
479
480 **Long-read structural variant analysis**
481 Structural variants were identified following the pipeline described in [57]. Briefly, PacBio reads from 18
482 isolates were aligned to the IPO323 reference genome using the NGMLR software v0.2.7 with default
483 parameters. Variants longer than 10 bp (-l 10 option) were called using Sniffles v1.0.10 allowing for a
484 maximum distance of 5 kb for grouping variants (-d 5000), a minimum mapping quality of 30 (-q 30) and
485 a minimum read support of 80% (-f 0.8). Sniffles adds a tag to each variant reporting the confidence of
486 the breakpoints mapping. We only retained precisely mapped variants that are smaller than 100 kb
487 (translocation events have unresolved length with the variant calling method). All 18 individual variant

488 files were merged using the SURVIVOR v1.0.7 tool allowing a distance of 1kb to bin variants [58]. Variants
489 larger than 100'000 bp were filtered out for all further analyses.
490
491 **Genome-wide association mapping**
492 The genome-wide association study (GWAS) was performed on 106 *Z. tritici* isolates collected from
493 single wheat fields in Australia, Israel, Switzerland and USA (Oregon) as described in [27] The total SNP
494 dataset included 779'178 SNPs. We analysed phylogenetic relationships using the vk suite (https://vcf-
495 kit.readthedocs.io/en/latest/phylo/). We generated a phylogenetic tree using the vk phylo tree function
496 using the unweighted pair grouping method based on the arithmetic mean. The phylogenetic tree was
497 visualized using the iTOL web interface (https://itol.embl.de). We performed GWAS for a set of 24 traits
498 related to fungicide resistance, growth and host adaptation (for values see Table S5). Virulence on two
499 wheat cultivars was measured as the percentage of leaf area covered by pycnidia (PLACP) and
500 described in [27]. The other 22 traits were measured in vitro following methods from earlier studies [59–62].
501 Briefly, isolates were regenerated from -80°C culture stocks on yeast malt sucrose agar petri-dishes for
502 4-5 days at 18°C (4 gl^{-1} yeast extract, 4 gl^{-1} malt extract, 4 gl^{-1} sucrose, 50 mgl^{-1} kanamycin). For
503 plate inoculations, blastospores were diluted in sterile water to 200 spores/mL final concentration using
504 KOVA counting slides and 500 µl spore suspensions were then spread on potato dextrose agar petri
505 dishes (PDA, 4 g l^{-1} potato starch, 20 g l^{-1} dextrose, 15 g l^{-1} agar). Control treatments were performed
506 at 22°C and cold treatment at 15°C, both at 70 relative humidity. Plate pictures were taken at 8, 11and
507 14 days post inoculation using a digital camera for five technical replicates. Images were analysed using
508 ImageJ macros from [59] to measure colony area. For each isolate, estimates of colony area are based
509 on nine single spore colonies. Growth rates (mm/day) were estimated from average colony radius by
510 fitting a generalized linear model over the three time points. Melanisation was estimated from the same
511 images by calculating the average values of grey (colours ranging from 0/black to 255/white). For
512 fungicide treatments, 0.05 ppm of propiconazole was added to the PDA plates to address its impact on
513 growth at 15°C and 22°C. Further formal testing of propiconazole resistance was carried on microtiter
514 plates. For that spores were grown on 100 µl Sabouraud-dextrose liquid media (inoculated with 100 µl
515 of spore suspension at $2.5e^4$ spores/mL) with varying concentrations of propiconazole (0.00006,
516 0.00017, 0.0051, 0.0086, 0.015, 0.025, 0.042, 0.072, 0.20, 0.55, 1.5 mg/L including a control without
517 propiconazole). Microtiter plates were sealed and incubated at 22°C for four days at dark with 80%
518 relative humidity. Growth was estimated given the optical density at the 605 nm using an Elisa plate
519 reader (MR5000, Dynatech). Five technical replicates were performed, and dose-response curves were
520 used to estimate the concentration at which the growth is reduced by half (EC_{50}) using the drc v.3.0-1
521 R package [63]. Prior the association study we computed the genetic relatedness matrix (GRM) and
522 conduced a principal component analysis (PCA) to investigate the genetic structure in the dataset using
523 TASSEL v.20200220 [64]. The genome wide association studies (GWAS) were then performed by
524 applying a mixed linear model which included the GRM as random factors using the GAPIT v3.0 R
525 package [65]. We retrieved false discovery rate (FDR) thresholds for each trait using the p.adjust function
526 implemented in R. The overlap between 5% FDR SNPs and structural variants was assessed with
527 bedtools intersect [66]. We built a contingency table summarising for each trait the number of 5% FDR

528 SNPs overlapping with indels, translocations, inversions and duplications. Odd ratios were calculated
529 using two-sided Fisher tests in R.
530
531 **Genome-wide sequence feature correlations**
532 The IPO323 genome was divided into 10 kb bins for which we retrieved sequence based statistics using
533 the EMBOSS v6.6.0.0 suite and bedtools [66,67]. The dataset included a total of 35 sequence metrics and
534 5 additional including the presence, average length and total number of indels and translocations
535 identified across the 18 worldwide isolates (summarized in Table S3). Data on histone methylation
536 marks (H3K9, H3K4 and H3K27) was previously generated by [68] and included the average read count
537 for enriched domains as identified by RSEG [69]. Recombination rates were calculated using 214
538 progenies from the cross between isolates 1A5 and 1E4 [36] (both included in the pangenome analyses).
539 Pairwise Spearman's correlation values were computed in R using complete observations (*use* =
540 pairwise.complete.obs) and visualized using the *ggcorr* function from the GGally package.
541
542 **Species-wide model of structural variant occurrence**
543 We applied a supervised machine learning approach using 30 sequence features with the aim of
544 predicting the occurrence of translocations and indels. Full details on the analysed sequence features
545 are shown in Table S3. The training dataset was made of a random 80% subset of the 3'980 10 kb
546 windows. We trained the models for translocations and indels separately. To identify the best-
547 performing machine learning method, we explored four different algorithms based on regression and
548 classification [70–72] (Logistic regression, Stochastic Gradient Boosting, Random Forests and Boosted
549 Classification Tree methods). Regression and classification methods mostly differ by their procedure to
550 identify decision boundaries. Regression methods implement binary decisions while classification
551 methods (*e.g.* decision trees) allow for more complex non-linear boundaries [73]. Decision trees can
552 improve the assignment if the two classes follow a non-linear separation, but these methods tend to
553 overfit the training dataset when the two classes are not well separated. Here all models were trained
554 using the *caret* package in R (https://www.rdocumentation.org/packages/caret) with 10-fold cross-
555 validation repeated three times. For the logistic regression models, we used the *glm* method
556 implemented by the caret package using the binomial family definition. For the random forest models,
557 we implemented a random selection of 1 to 10 predictors with the *rf* method (including the mtry option).
558 For the boosted classification tree models, we applied the *ada* method using 100, 1'000 or 3'000
559 boosting iterations with a maximum tree depth of 1, 5 or 20 and a fixed learning rate of 0.01. For the
560 stochastic gradient boosting models, we used the *gbm* method with 50, 500 or 1'000 boosting iterations
561 with a maximum tree depth of 1, 5 or 10. The stochastic gradient boosting models were trained with
562 two values of shrinkage (0.001 or 0.01) and a terminal node minimal size (1 or 5). For all the models,
563 the accuracy measure was used for the selection of the best model during training (metric = "accuracy").
564 For the final random forest model, we estimated the relative importance of predictors using the *varImp*
565 function from the *caret* package with the scale option set to false.
566

567 The eight final models (the 4 different methods using indels and translocations as classifiers,
568 respectively) were tested on a 20% validation subset of the pangenome dataset using the *predict*
569 function. To address the final model performance, we investigated how the true and false positive rate
570 were related using the receiver operating characteristic (ROC) curve implemented in the ROCR R
571 package [74]. The area under the ROC curve (AUC) was calculated with the performance function from
572 ROCR package. In addition, we quantified the relationship between the model precision and recall using
573 the precision recall curves calculated with the *pr.curve* function from the PRROC package [75]. Confusion
574 matrices and the evluation metrics of each model were recovered using the *confusionMatrix* function in
575 the *caret* package.
576
577 ***Arabidopsis thaliana* pangenome sequence metrics and indel model**
578 For the *Arabidopsis thaliana* dataset, we used indels identified across 8 chromosome-level assemblies
579 [28]. Orthogroups identified by Jiao et al. (2020) were defined as core (orthogroup present in all 8
580 genomes), accessory (orthogroup present in 1 to 7 genomes) and singletons (orthogroup present in
581 one genome only). Transposable elements and gene annotations were retrieved from the TAIR10
582 database (https://www.arabidopsis.org/). The list of reference NLRs (Col-0) was retrieved from Weyer
583 *et al.* [29]. Transposons annotation form TAIR10 database were renamed according to the three-letters
584 nomenclature [53] (Table S6). Recombination rate in 30 kb windows was estimated from the position of
585 crossover events established by Rowan et al, 2019. We counted the number of crossovers overlapping
586 the sequence windows (Table S7). The recombination rate was then calculated as the number of
587 crossovers per 30 kb. GC coverage along the 30 kb windows was calculated using bedtools nuc
588 (v2.29.0). We trained the same four models as for the *Z. tritici* pangenome but focused only on indels
589 (3'993 30 kb windows) using the *predict* function in R. Receiver operating characteristic and precision
590 recall curves were also calculated identically as for the *Z. tritici* pangenome dataset using the ROCR
591 and PRROC R packages [74,75]. The confusion matrices and model evaluation metrics were recovered
592 from the *confusionMatrix* function implemented in the *caret* package.
593
594 **Identifying sequence rearrangements across the pedigree**
595 We applied a similar method to call variants in the progenies resulting from the cross between the
596 parental isolates 1A5x1E4 [60]. PacBio reads from the 9 progenies and the two parents (1A5 and 1E4)
597 were aligned to the 1A5 parental genome using NGMLR and variants were called using Sniffles as
598 described for the pangenome. We removed all variants, which were present in either one of the two
599 parents (1A5 or 1E4), thus retaining only the variants generated during the different rounds of meiosis.
600 To dissect the emergence and reshuffling of individual rearrangements at each meiotic cycle, we
601 defined four categories based on the variant history in the pedigree (Figure 5D). Variants present in a
602 focal progeny and the direct parents were defined as retained variants. Variants present in a focal
603 progeny but in none of its ancestors in the pedigree were called new variants. Variants present in a
604 progeny, absent in its direct parents but present already in an ancestor were defined as recurrent
605 variants. Finally, variants absent from a progeny but present in both parents were called lost variants.
606

607 **Identification of crossover breakpoints on chromosome 1**
608 To identify the crossover events on chromosome 1 over the 4 rounds of meiosis, we performed whole-
609 genome alignments of each progeny against its two immediate parents using blastn [41,44] (v2.8.1+). To
610 visualise crossovers, we plotted chromosome-wide synteny using the genoplotR package in R [76]. To
611 recover the crossover breakpoints, we manually inspected the blastn results (-outputfmt 6) and
612 recovered the approximate positions of shifts in sequence alignment identity between the two parental
613 genomes. We identified a total of 18 crossover breakpoints summarised in Table S8. To investigate the
614 relative distance of structural variants to crossover breakpoints and given the overall high synteny in
615 the progeny, we assigned the same breakpoint position to the 1A5 reference genome. We calculated
616 the distance between sequence rearrangements on chromosome 1 and crossovers using the *bedtools*
617 *closest* command [66,67].
618
619 **Evaluation of pangenome sequence rearrangement models for the pedigree**
620 We applied the four pangenome-trained models to predict indels and translocations to the nine progeny
621 dataset using the *predict* function in the R package *caret* (https://github.com/topepo/caret/). Receiver
622 operating characteristic and precision recall curves were calculated as for the pangenome dataset using
623 the ROCR and PRROC packages [74,75]. Similarly, the confusion matrices and model evaluation metrics
624 were recovered from the *confusionMatrix* function implemented in *caret* package.
625
626 **Data access**
627 Genome assemblies for the species-wide pangenome are available under the project PRJEB33986.
628 The progeny genomes are available under the project PRJNA645795. Variant calls for GWAS were
629 retrieved from the variant call format (VCF) file deposited in the European Nucleotide Archive (ENA)
630 under the accession numbers PRJEB15502/ERP017268 and the analysis number ERZ330467.
631 Analyses scripts are available on Github (https://github.com/crolllab/datasets).
632
633 **Author contributions**
634 TB and DC designed the study; TB performed analyses; SF and FH contributed datasets; TB and DC
635 wrote the manuscript with input from all co-authors.
636
637 **Competing interests**
638 We declare that we have no competing interests.
639
640 **Acknowledgements**
641 We are grateful to Leen Abraham, Nikhil Kumar Singh, Sylvain Raffaele, Sam Yeaman and Aurélien
642 Tellier for critical feedback on a previous version of this manuscript. We thank Marcello Zala for advice
643 and help with the crosses and the DNA extractions. The sequencing was performed at the Functional
644 Genomics Center Zurich. DC was supported by the Fondation Pierre Mercier pour la science and the
645 Swiss National Science Foundation (grant number 31003A_173265).

646 **References**

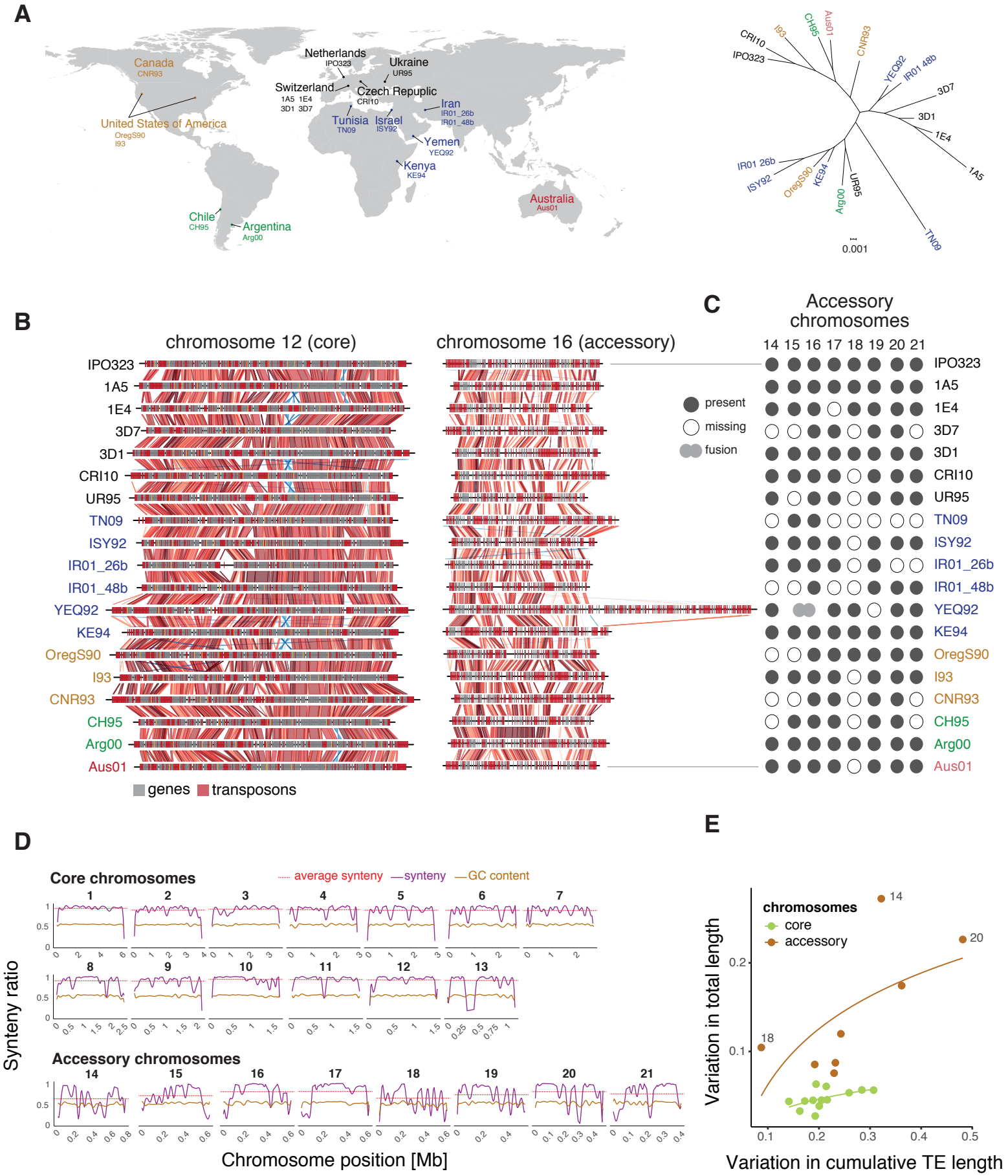
647

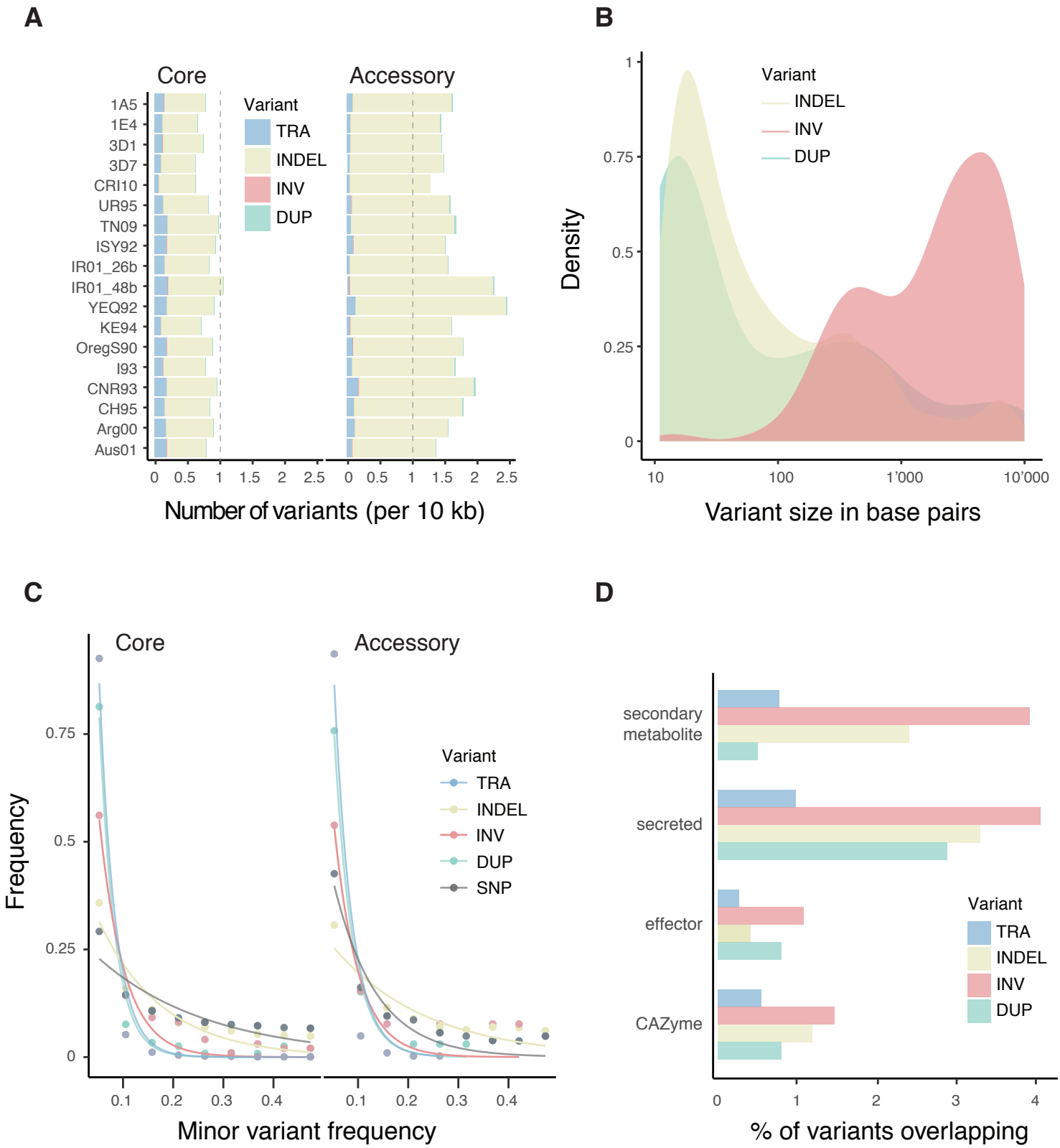
648 1. Quinlan, A. R. & Hall, I. M. Characterizing complex structural variation in germline and somatic
649 genomes. *Trends in Genetics* **28**, 43–53 (2012).
650 2. Fuentes, R. R. *et al.* Structural variants in 3000 rice genomes. *Genome Res.* **29**, 870–880
651 (2019).
652 3. Catanach, A. *et al.* The genomic pool of standing structural variation outnumbers single
653 nucleotide polymorphism by threefold in the marine teleost *Chrysophrys auratus*. *Mol. Ecol.*
654 **28**, 1210–1223 (2019).
655 4. Plaumann, P. L., Schmidpeter, J., Dahl, M., Taher, L. & Koch, C. A dispensable chromosome
656 is required for virulence in the hemibiotrophic plant pathogen Colletotrichum higginsianum.
657 *Front. Microbiol.* **9**, (2018).
658 5. Langner, T. *et al.* Genomic rearrangements generate hypervariable mini-chromosomes in
659 host-specific lineages of the blast fungus. *bioRxiv* 2020.01.10.901983 (2020).
660 doi:10.1101/2020.01.10.901983
661 6. Schiessl, S. V., Katche, E., Ihien, E., Chawla, H. S. & Mason, A. S. The role of genomic
662 structural variation in the genetic improvement of polyploid crops. *Crop Journal* **7**, 127–140
663 (2019).
664 7. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic
665 structural variation: Insights from and for human disease. *Nature Reviews Genetics* **14**, 125–
666 138 (2013).
667 8. Sanseverino, W. *et al.* Transposon Insertions, Structural Variations, and SNPs Contribute to
668 the Evolution of the Melon Genome. *Mol. Biol. Evol.* **32**, 2760–2774 (2015).
669 9. Rowan, B. A. *et al.* An ultra high-density arabidopsis thaliana crossover map that refines the
670 influences of structural variation and epigenetic features. *Genetics* **213**, 771–787 (2019).
671 10. Miles, A. *et al.* Indels, structural variation, and recombination drive genomic diversity in
672 Plasmodium falciparum. *Genome Res.* **26**, 1288–1299 (2016).
673 11. Natri, H. M., Merilä, J. & Shikano, T. The evolution of sex determination associated with a
674 chromosomal inversion. *Nat. Commun.* **10**, 1–13 (2019).
675 12. Thompson, M. J. & Jiggins, C. D. Supergenes and their role in evolution. *Heredity* **113**, 1–8
676 (2014).
677 13. McClintock, B. Mutable loci in maize. *Carnegie Inst. Wash. Year B.* **47**, 155–169 (1948).
678 14. Argueso, J. L. *et al.* Double-strand breaks associated with repetitive DNA can reshape the
679 genome. *Proc. Natl. Acad. Sci.* **105**, 11845–11850 (2008).
680 15. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping.
681 *Nature Reviews Genetics* **12**, 363–376 (2011).
682 16. Carvalho, C. M. B. & Lupski, J. R. Mechanisms underlying structural variant formation in
683 genomic disorders. *Nature Reviews Genetics* **17**, 224–238 (2016).
684 17. Kidd, J. M. *et al.* A human genome structural variation sequencing resource reveals insights
685 into mutational mechanisms. *Cell* **143**, 837–847 (2010).
686 18. Abyzov, A. *et al.* Analysis of deletion breakpoints from 1,092 humans reveals details of
687 mutation mechanisms. *Nat. Commun.* **6**, 7256 (2015).
688 19. Weckselblatt, B. & Rudd, M. K. Human Structural Variation: Mechanisms of Chromosome
689 Rearrangements. *Trends Genet.* **31**, 587–599 (2015).
690 20. Möller, M. *et al.* Destabilization of chromosome structure by histone H3 lysine 27 methylation.
691 *PLoS Genet.* **15**, (2019).
692 21. Fudenberg, G. & Pollard, K. S. Chromatin features constrain structural variation across
693 evolutionary timescales. *Proc. Natl. Acad. Sci.* **116**, 2175–2180 (2019).
694 22. Morgan, A. P. *et al.* Structural variation shapes the landscape of recombination in mouse.
695 *Genetics* **206**, 603–619 (2017).
696 23. Torriani, S. F. F. *et al.* Zymoseptoria tritici: A major threat to wheat production, integrated
697 approaches to control. *Fungal Genet. Biol.* **79**, 8–12 (2015).

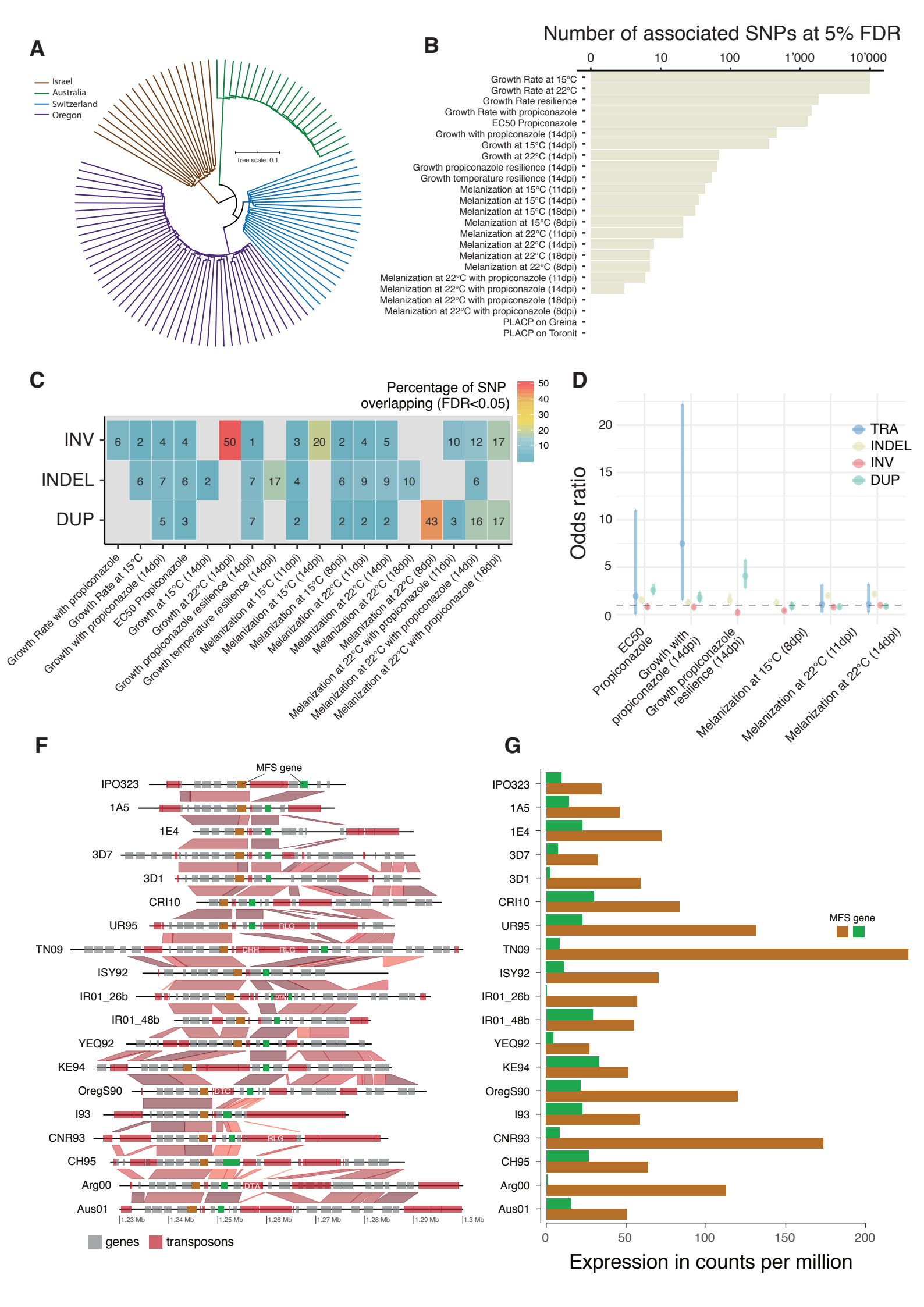
698 24. Badet, T., Oggenfuss, U., Abraham, L., McDonald, B. A. & Croll, D. A 19-isolate reference-
699 quality global pangenome for the fungal wheat pathogen Zymoseptoria tritici. *BMC Biol.* **18**, 1–
700 18 (2020).
701 25. Christmas, M. J. *et al.* Chromosomal inversions associated with environmental adaptation in
702 honeybees. *Mol. Ecol.* **28**, 1358–1374 (2019).
703 26. Fuller, Z. L., Koury, S. A., Phadnis, N. & Schaeffer, S. W. How chromosomal rearrangements
704 shape adaptation and speciation: Case studies in *Drosophila pseudoobscura* and its sibling
705 species *Drosophila persimilis*. *Mol. Ecol.* **28**, 1283–1301 (2019).
706 27. Hartmann, F. E., Sánchez-Vallet, A., McDonald, B. A. & Croll, D. A fungal wheat pathogen
707 evolved host specialization by extensive chromosomal rearrangements. *ISME J.* **11**, 1189–
708 1204 (2017).
709 28. Jiao, W. B. & Schneeberger, K. Chromosome-level assemblies of multiple Arabidopsis
710 genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nat.*
711 *Commun.* **11**, 1–10 (2020).
712 29. Van de Weyer, A.-L. *et al.* A Species-Wide Inventory of NLR Genes and Alleles in Arabidopsis
713 thaliana. *Cell* **178**, 1260-1272.e14 (2019).
714 30. Zlotorynski, E. *et al.* Molecular Basis for Expression of Common and Rare Fragile Sites. *Mol.*
715 *Cell. Biol.* **23**, 7143–7151 (2003).
716 31. Stankiewicz, P. & Lupski, J. R. Molecular-evolutionary mechanisms for genomic disorders.
717 *Current Opinion in Genetics and Development* **12**, 312–319 (2002).
718 32. Chan, Y. F. *et al.* Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of
719 a pitxl enhancer. *Science (80-.).* **327**, 302–305 (2010).
720 33. Hope, E. A. *et al.* Experimental evolution reveals favored adaptive routes to cell aggregation in
721 yeast. *Genetics* **206**, 1153–1167 (2017).
722 34. Sackton, T. B. & Clark, N. Convergent evolution in the genomics era: new insights and
723 directions. *Philos. Trans. R. Soc. B Biol. Sci.* **374**, 20190102 (2019).
724 35. Martin, A. & Orgogozo, V. The loci of repeated evolution: A catalog of genetic hotspots of
725 phenotypic variation. *Evolution (N. Y).* **67**, 1235–1250 (2013).
726 36. Croll, D., Lendenmann, M. H., Stewart, E. & McDonald, B. A. The impact of recombination
727 hotspots on genome evolution of a fungal plant pathogen. *Genetics* **201**, 1213–1228 (2015).
728 37. Kema, G. H. J., Verstappen, E. C. P., Todorova, M. & Waalwijk, C. Successful crosses and
729 molecular tetrad and progeny analyses demonstrate heterothallism in Mycosphaerella
730 graminicola. *Curr. Genet.* **30**, 251–258 (1996).
731 38. Plissonneau, C., Hartmann, F. E. & Croll, D. Pangenome analyses of the wheat pathogen
732 *Zymoseptoria tritici* reveal the structural basis of a highly plastic eukaryotic genome. *BMC Biol.*
733 **16**, 5 (2018).
734 39. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting
735 and repeat separation. *Genome Res.* **27**, 722–736 (2017).
736 40. Kolmogorov, M., Raney, B., Paten, B. & Pham, S. Ragout-a reference-assisted assembly tool
737 for bacterial genomes. *Bioinformatics* **30**, i302-9 (2014).
738 41. Altschul, S. F. *et al.* Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–10 (1990).
739 42. Stanke, M., Schöffmann, O., Morgenstern, B. & Waack, S. Gene prediction in eukaryotes with
740 a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*
741 **7**, 62 (2006).
742 43. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped
743 cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644 (2008).
744 44. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421
745 (2009).
746 45. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform.
747 *Bioinformatics* **25**, 1754–60 (2009).
748 46. Barnett, D. W., Garrison, E. K., Quinlan, A. R., Stromberg, M. P. & Marth, G. T. BamTools: a
749 C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* **27**, 1691–1692

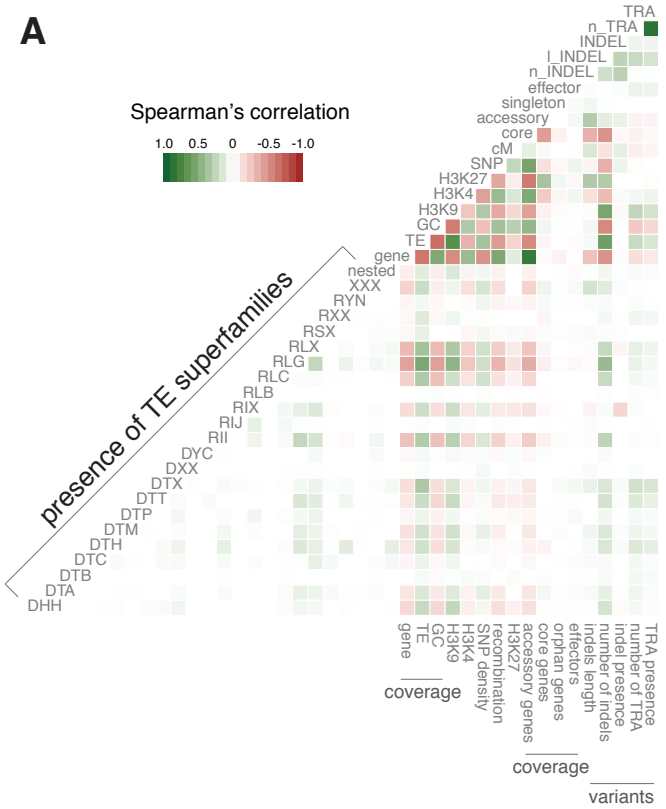
750 (2011).
751 47. Lomsadze, A., Burns, P. D. & Borodovsky, M. Integration of mapped RNA-Seq reads into
752 automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res.* **42**, e119–e119
753 (2014).
754 48. Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M. & Stanke, M. BRAKER1: Unsupervised
755 RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS: Table 1.
756 *Bioinformatics* **32**, 767–769 (2016).
757 49. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome
758 comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157
759 (2015).
760 50. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative
761 genomics. *bioRxiv* 466201 (2019). doi:10.1101/466201
762 51. Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0. (2015).
763 52. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in
764 eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).
765 53. Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. *Nat. Rev.*
766 *Genet.* **8**, 973–982 (2007).
767 54. Chin, C. S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT
768 sequencing data. *Nat. Methods* **10**, 563–569 (2013).
769 55. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using
770 repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
771 56. Goel, M., Sun, H., Jiao, W. B. & Schneeberger, K. SyRI: finding genomic rearrangements and
772 local sequence differences from whole-genome assemblies. *Genome Biol.* **20**, 277 (2019).
773 57. Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single-
774 molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
775 58. Jeffares, D. C. *et al.* Transient structural variations have strong effects on quantitative traits
776 and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 14061 (2017).
777 59. Lendenmann, M. H., Croll, D., Stewart, E. L. & McDonald, B. A. Quantitative trait locus
778 mapping of melanization in the plant pathogenic fungus Zymoseptoria tritici. *G3 Genes,*
779 *Genomes, Genet.* **4**, 2519–2533 (2014).
780 60. Lendenmann, M. H., Croll, D. & McDonald, B. A. QTL mapping of fungicide sensitivity reveals
781 novel genes and pleiotropy with melanization in the pathogen Zymoseptoria tritici. *Fungal*
782 *Genet. Biol.* **80**, 53–67 (2015).
783 61. Lendenmann, M. H., Croll, D., Palma-Guerrero, J., Stewart, E. L. & Mcdonald, B. A. QTL
784 mapping of temperature sensitivity reveals candidate genes for thermal adaptation and growth
785 morphology in the plant pathogenic fungus Zymoseptoria tritici. *Heredity (Edinb).* **116**, 384–
786 394 (2016).
787 62. Mohd-Assaad, N., McDonald, B. A. & Croll, D. Multilocus resistance evolution to azole
788 fungicides in fungal plant pathogen populations. *Mol. Ecol.* **25**, 6124–6142 (2016).
789 63. Ritz, C., Baty, F., Streibig, J. C. & Gerhard, D. Dose-Response Analysis Using R. *PLoS One*
790 **10**, e0146021 (2015).
791 64. Bradbury, P. J. *et al.* TASSEL: Software for association mapping of complex traits in diverse
792 samples. *Bioinformatics* **23**, 2633–2635 (2007).
793 65. Tang, Y. *et al.* GAPIT Version 2: An Enhanced Integrated Tool for Genomic Association and
794 Prediction. *Plant Genome* **9**, plantgenome2015.11.0120 (2016).
795 66. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic
796 features. *Bioinformatics* **26**, 841–842 (2010).
797 67. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software
798 Suite. *Trends Genet.* **16**, 276–7 (2000).
799 68. Schotanus, K. *et al.* Histone modifications rather than the novel regional centromeres of
800 Zymoseptoria tritici distinguish core and accessory chromosomes. **8**, 41 (2015).
801 69. Song, Q. & Smith, A. D. Identifying dispersed epigenomic domains from ChIP-Seq data.

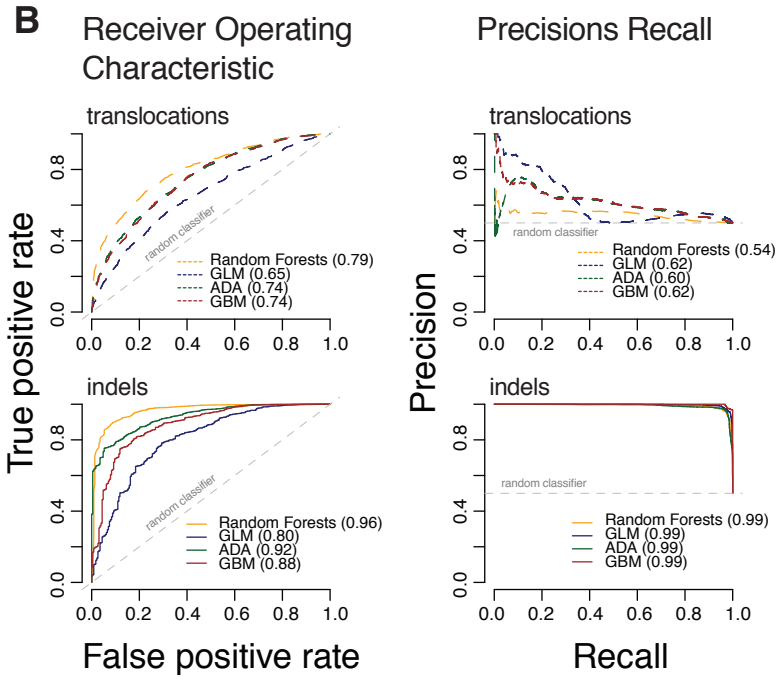
Bioinformatics **27**, 870–1 (2011).
70. Praveena, M. & Jaiganesh, V. A Literature Review on Supervised Machine Learning Algorithms and Boosting Process. *Int. J. Comput. Appl.* **169**, 32–35 (2017).
71. Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A. & Aljaaf, A. J. A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. in 3–21 (Springer, Cham, 2020). doi:10.1007/978-3-030-22475-2_1
72. Libbrecht, M. W. & Noble, W. S. Machine learning applications in genetics and genomics. *Nature Reviews Genetics* **16**, 321–332 (2015).
73. Malley, J. D., Kruppa, J., Dasgupta, A., Malley, K. G. & Ziegler, A. Probability Machines: Consistent probability estimation using nonparametric learning machines. *Methods Inf. Med.* **51**, 74–81 (2012).
74. Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. ROCR: Visualizing classifier performance in R. *Bioinformatics* **21**, 3940–3941 (2005).
75. Grau, J., Grosse, I. & Keilwagen, J. PRROC: Computing and visualizing Precision-recall and receiver operating characteristic curves in R. *Bioinformatics* **31**, 2595–2597 (2015).
76. Guy, L., Roat Kultima, J. & Andersson, S. G. E. genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* **26**, 2334–2335 (2010).


A
Canada
CNR93
United States of America
OregS90
I93
Netherlands
IPO323
Ukraine
UR95
Switzerland
1A5 1E4
3D1 3D7
Czech Republic
CRI10
Tunisia
TN09
Israel
ISY92
Iran
IR01_26b
IR01_48b
Yemen
YEQ92
Kenya
KE94
Australia
Aus01
Chile
CH95
Argentina
Arg00
CRI10
IPO323
I93
CH95
Aus01
CNR93
YEQ92
IR01 48b
3D7
3D1
1E4
1A5
IR01 26b
ISY92
OregS90
KE94
Arg00
UR95
TN09
0.001
B
chromosome 12 (core)
chromosome 16 (accessory)
IPO323
1A5
1E4
3D7
3D1
CRI10
UR95
TN09
ISY92
IR01_26b
IR01_48b
YEQ92
KE94
OregS90
I93
CNR93
CH95
Arg00
Aus01
genes
transposons
C
Accessory chromosomes
14 15 16 17 18 19 20 21
present
missing
fusion
D
Core chromosomes
average synteny
synteny
GC content
Synteny ratio
Accessory chromosomes
Chromosome position [Mb]
E
chromosomes
core
accessory
Variation in total length
Variation in cumulative TE length


A
Core
Accessory
Variant
TRA
INDEL
INV
DUP
1A5
1E4
3D1
3D7
CRI10
UR95
TN09
ISY92
IR01_26b
IR01_48b
YEQ92
KE94
OregS90
I93
CNR93
CH95
Arg00
Aus01
0 0.5 1 1.5 2 2.5
Number of variants (per 10 kb)
B
Density
Variant
INDEL
INV
DUP
0 0.25 0.5 0.75 1
10 100 1'000 10'000
Variant size in base pairs
C
Core
Accessory
Frequency
0 0.25 0.5 0.75
Variant
TRA
INDEL
INV
DUP
SNP
0.1 0.2 0.3 0.4
Minor variant frequency
D
secondary metabolite
secreted
effector
CAZyme
Variant
TRA
INDEL
INV
DUP
0 1 2 3 4
% of variants overlapping

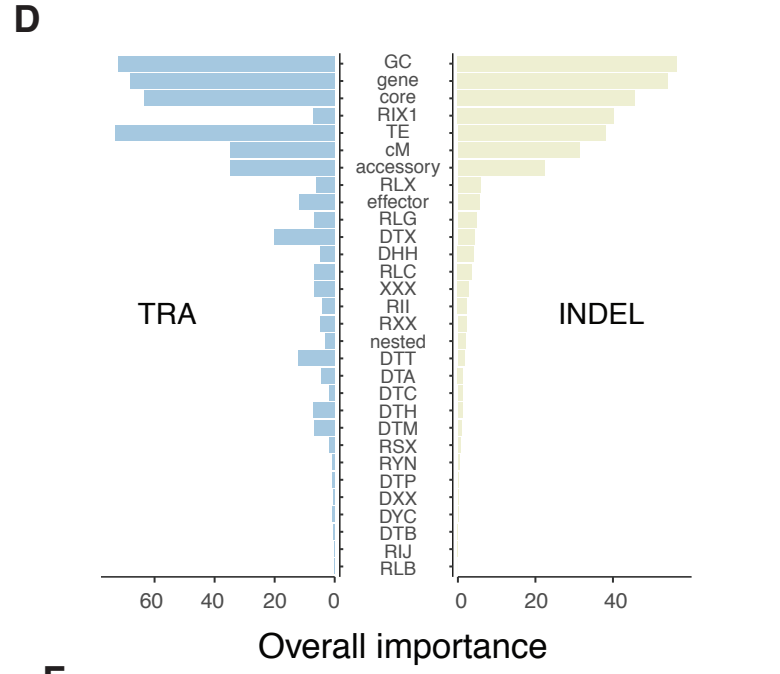

A
Israel
Australia
Switzerland
Oregon
Tree scale: 0.1
B
Number of associated SNPs at 5% FDR
0
10
100
1'000
10'000
Growth Rate at 15°C
Growth Rate at 22°C
Growth Rate resilience
Growth Rate with propiconazole
EC50 Propiconazole
Growth with propiconazole (14dpi)
Growth at 15°C (14dpi)
Growth at 22°C (14dpi)
Growth propiconazole resilience (14dpi)
Growth temperature resilience (14dpi)
Melanization at 15°C (11dpi)
Melanization at 15°C (14dpi)
Melanization at 15°C (18dpi)
Melanization at 15°C (8dpi)
Melanization at 22°C (11dpi)
Melanization at 22°C (14dpi)
Melanization at 22°C (18dpi)
Melanization at 22°C (8dpi)
Melanization at 22°C with propiconazole (11dpi)
Melanization at 22°C with propiconazole (14dpi)
Melanization at 22°C with propiconazole (18dpi)
Melanization at 22°C with propiconazole (8dpi)
PLACP on Greina
PLACP on Toronit
C
Percentage of SNP overlapping (FDR<0.05)
50
40
30
20
10
INV
INDEL
DUP
Growth Rate with propiconazole
Growth Rate at 15°C
Growth with propiconazole (14dpi)
EC50 Propiconazole
Growth at 15°C (14dpi)
Growth at 22°C (14dpi)
Growth propiconazole resilience (14dpi)
Growth temperature resilience (14dpi)
Melanization at 15°C (11dpi)
Melanization at 15°C (14dpi)
Melanization at 15°C (8dpi)
Melanization at 22°C (11dpi)
Melanization at 22°C (14dpi)
Melanization at 22°C (18dpi)
Melanization at 22°C (8dpi)
Melanization at 22°C with propiconazole (11dpi)
Melanization at 22°C with propiconazole (14dpi)
Melanization at 22°C with propiconazole (18dpi)
D
Odds ratio
TRA
INDEL
INV
DUP
EC50 Propiconazole
Growth with propiconazole (14dpi)
Growth propiconazole resilience (14dpi)
Melanization at 15°C (8dpi)
Melanization at 22°C (11dpi)
Melanization at 22°C (14dpi)
F
MFS gene
IPO323
1A5
1E4
3D7
3D1
CRI10
UR95
TN09
ISY92
IR01_26b
IR01_48b
YEQ92
KE94
OregS90
I93
CNR93
CH95
Arg00
Aus01
RLG
DHH
DTC
DTA
1.23 Mb
1.24 Mb
1.25 Mb
1.26 Mb
1.27 Mb
1.28 Mb
1.29 Mb
1.3 Mb
genes
transposons
G
MFS gene
0
50
100
150
200
Expression in counts per million





C

	Algorithm	Accuracy	Precision	Recall	F1
translocations	Random Forests	0.72	0.71	0.77	0.74
	Logistic Regression	0.61	0.63	0.60	0.62
	Stochastic Gradient Boosting	0.67	0.66	0.75	0.71
	Generalized Boosted Regression	0.68	0.67	0.76	0.71
indels	Random Forests	0.95	0.96	0.99	0.97
	Logistic Regression	0.93	0.94	0.98	0.96
	Stochastic Gradient Boosting	0.94	0.95	0.99	0.97
	Generalized Boosted Regression	0.94	0.94	0.99	0.97



E

Reference data

Prediction		none	variant
translocation	0	1'019	380
	1	509	1'276
indel	0	106	11
	1	122	2'945

Random Forests

- false positives
- true positives
- true negatives
- false negatives

F

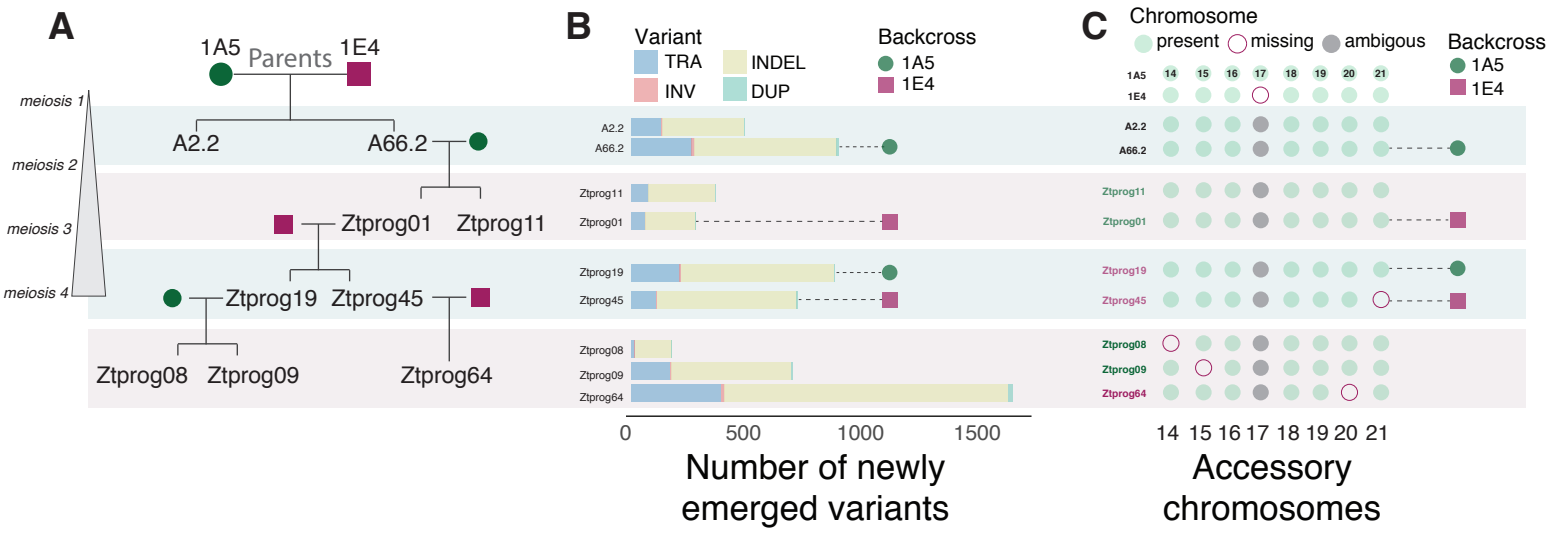
INDEL

Reference data

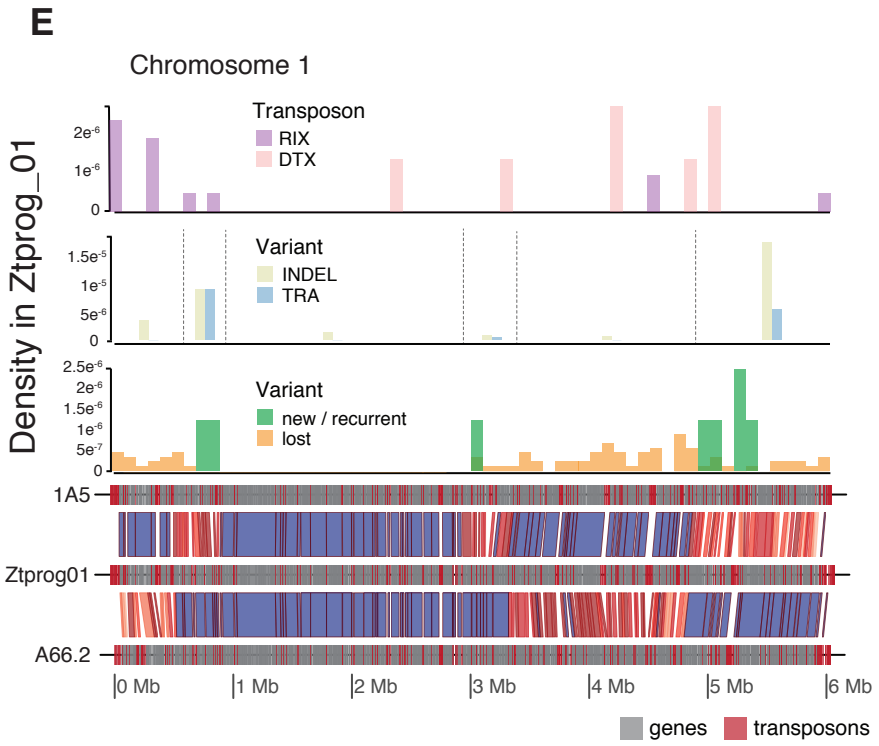
Prediction	none	variant
0	34	10
1	915	3'034

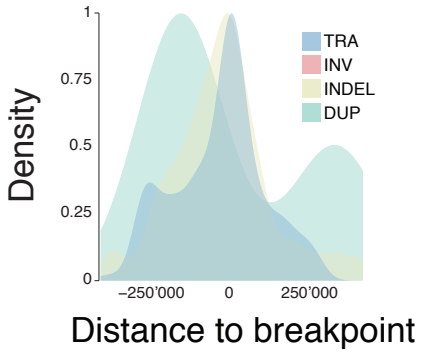
- false positives
- true positives
- true negatives
- false negatives

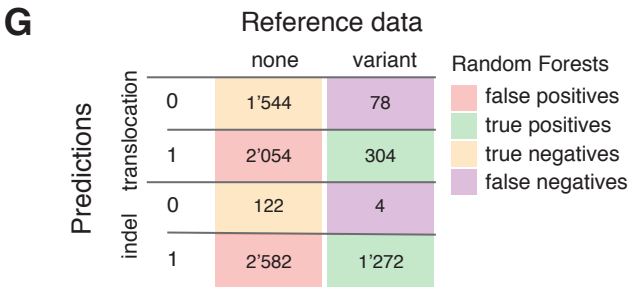
Arabidopsis thaliana


A
1A5
Parents
1E4
meiosis 1
meiosis 2
meiosis 3
meiosis 4
A2.2
A66.2
Ztprog01
Ztprog11
Ztprog19
Ztprog45
Ztprog08
Ztprog09
Ztprog64
B
Variant
TRA
INDEL
INV
DUP
Backcross
1A5
1E4
0
500
1000
1500
Number of newly emerged variants
C
Chromosome
present
missing
ambigous
Backcross
1A5
1E4
14 15 16 17 18 19 20 21
Accessory chromosomes

D
Progeny
Variant status
focal isolate
new
focal isolate
recurrent
focal isolate
retained
focal isolate
lost
Variant
present
absent
TRA
INDEL
INV
DUP
Ztprog_01
Ztprog_11
Ztprog_19
Ztprog_45
Ztprog_08
Ztprog_09
Ztprog_64
Variant status
retained
new
recurrent
lost
Number of variants


E
Chromosome 1
Density in Ztprog_01
Transposon
RIX
DTX
Variant
INDEL
TRA
Variant
new / recurrent
lost
1A5
Ztprog01
A66.2
0 Mb
1 Mb
2 Mb
3 Mb
4 Mb
5 Mb
6 Mb
genes
transposons


F
Density
TRA
INV
INDEL
DUP
−250'000
0
250'000
Distance to breakpoint


G
Reference data
none
variant
Predictions
translocation
indel
0
1
0
1
1'544
78
2'054
304
122
4
2'582
1'272
Random Forests
false positives
true positives
true negatives
false negatives