SyRI: identification of syntenic and rearranged regions from whole-genome assemblies

Manish Goel[1], Hequan Sun[1], Wen-Biao Jiao[1], Korbinian Schneeberger[1*]

[1] Max Planck Institute for Plant Breeding Research, D-50829 Cologne, Germany

[*] To whom correspondence should be addressed

Email addresses:

Manish Goel: goel@mpipz.mpg.de

Hequan Sun: sun@mpipz.mpg.de

Wen-Biao Jiao: jiao@mpipz.mpg.de

Korbinian Schneeberger: schneeberger@mpipz.mpg.de

Running head: Synteny and rearrangement identification (SyRI)

Abstract

We present SyRI, an efficient tool for genome-wide identification of structural rearrangements (SR) from genome graphs, which are built up from pair-wise whole-genome alignments. Instead of searching for differences, SyRI starts by finding all co-linear regions between the genomes. As all remaining regions are SRs by definition, they can be classified as inversions, translocations, or duplications based on their positions in convoluted networks of repetitive alignments. Finally, SyRI reports local variations like SNPs and indels within syntenic and rearranged regions. We show SyRI's broad applicability to multiple species and genetically validate the presence of ~100 translocations identified in Arabidopsis.

Keywords

Structural rearrangements, structural variations, variant calling, genome alignments, genetics, genome assembly

Background

Different haploid genomes of the same species typically show high similarity in their genome structures including extensive syntenic regions. However, these syntenic regions can be interrupted by structural rearrangements (SRs), which are genomic regions characterised by different orientation and/or different location in the different genomes. SRs can be classified into inversions, translocations, and duplications. Despite being severe genomic differences, SRs were found to be implicated in various traits like disease phenotypes [1], reproductive strategies [2–4], or life cycle differences [5].

Many of the current SR prediction methods utilize short or long read alignments against reference sequences. Though local differences (like SNPs and small indels) can be found within such alignments with high accuracy, accurate prediction of complex SRs remains challenging using read alignments alone. In contrast, the comparison of high-quality genome assemblies is more powerful for accurate SR identification as the assembled contigs or scaffolds are typically much longer and of higher quality as compared to raw sequence reads [6]. However, despite significant technological improvements supporting the generation of *de-novo* whole-genome assemblies within the recent years [7], there are only a few tools which use whole-genome alignments (WGA) as the basis for the identification of genomic differences [8]. Available tools, for example, include AsmVar, which compares individual scaffolds of a *de novo* assembly against a reference sequence and analyses alignment breakpoints to identify inversions and translocations [9] and Assemblytics, which utilizes uniquely aligned regions within contig alignments to a reference sequence to identify various types of genomic differences including large indels or differences in local repeats [10].

Here, we introduce SyRI (Synteny and Rearrangement Identifier), a method to identify all structural genomic rearrangements between two related genomes (usually from the same species) using genome graphs generated from pair-wise WGA. SyRI starts by identifying syntenic regions between the pairs of homologous chromosomes of the two genomes. All non-syntenic regions are SRs by definition or otherwise, they would be part of the syntenic regions. This transforms the problem of SR *identification* to SR *classification*. SyRI classifies these non-syntenic regions into inversions, translocations, and duplications. For this, SyRI performs a simultaneous analysis of all rearranged regions to optimise the annotation of genomic differences (e.g. SyRI maximizes the size of the syntenic regions).

After the identification of syntenic and SR regions, SyRI identifies *local variation* (like SNPs and indels) within syntenic as well as rearranged regions. This introduces a hierarchy of variation and, for example, allows distinguishing between SNPs in co-linear regions as compared to SNPs in rearranged regions. This distinction is important, as rearranged regions (and the local variation in them) will not follow Mendelian segregation patterns in the offspring of the respective organisms, but can lead to changes in their copy number (Figure 1).

We show SyRI's performance in extensive simulations and comparison with existing tools for SR identification and analyse the effect of assembly contiguity. We also applied SyRI for the comparison of divergent genomes of five model species, including two *A. thaliana* strains, for which we genetically validated the existence of SR within the genomes of 50 recombinant off-spring genomes.

Results

Identifying genomic differences using SyRI. SyRI generates *genome graphs* from pairwise WGAs (here we used MUMmer3 toolbox to perform WGA [11,12], but other alignments tools like minimap2 [13] can be used as well) to identify genomic differences between any two haploid genomes (Additional File 1: Note 1). For an overview of the analysis workflow see Figure 2, and for algorithmic details see Figure 3.

Ideally, both of the haploid assemblies are at chromosome-level. If one or both of the assemblies are at scaffold-level, pseudo-chromosomes can be generated using homology between the assemblies themselves or homology to a reference sequence, which can be performed with tools like RaGOO (Additional File 1: Note 2, preprint [14]). Even though many of the local alignments between two genomes are redundant as they result from repeats, the genomes do not need to be repeat-masked. SyRI automatically distinguishes between repetitive and original alignments. Once done, SyRI outputs tab-separated and standard variant call format (VCF) files including all information on SR and local variation.

Even though translocations and transpositions are commonly distinguished, SyRI annotates both types as *translocations*. Moreover, translocations and duplications will be collectively referred to as *TDs*. Further, we distinguish between synteny and co-linearity. Regions are co-linear if their physical connection is conserved between two genomes. When considering the broad chromosomal structure, co-linear regions are also syntenic when they reside at the same, "orthologous" location in the two genomes.

Syntenic path identification. SyRI selects all forward alignments between a pair of homologous chromosomes and generates a genome graph in the form of a directed acyclic graph (DAG) (Figure 3a). For each of the alignments, a node is created with a score equal to the alignment score. Any pair of nodes is then connected by an edge if the two underlying alignments are co-linear and if no other co-linear alignment is between them.

After all nodes and edges are added to the genome graph, two additional 0-score nodes, called S (start) and E (end), are added to the genome graph. Edges are added from node S to all other nodes without any in-edge; similarly, edges are added from all nodes without any out-edge to node E. We identify the maximal syntenic path (i.e. the optimal set of non-conflicting, co-linear regions) by selecting the highest scoring path between node S and E using dynamic programming. This approach resembles an algorithm for the identification of conserved regions as seeds during whole-genome alignments as implemented in MUMmer [15,16]. This process is repeated for each pair of homologous chromosomes.

Inversion identification. Here, we define an inversion as a set of inverted alignments corresponding to one structural rearrangement event and which is in-between two syntenic alignments (Additional File 1: Figure S1). To identify them, SyRI selects all inverted alignments between a pair of corresponding chromosomes and then reverse complements one of the chromosomes. Then, analogous to the syntenic path identification, SyRI again builds up a genome graph now using this set of alignments based on reverse complemented regions. Start (S) and end (E) nodes are added and connected to all other nodes in the graph (Figure 3b (i)). In this graph, every path from S to E corresponds to a candidate inversion. For each candidate, SyRI

125 calculates an *inversion score* which is defined as the difference between the sum of the node
126 scores in the candidate inversion and the node scores of all syntenic alignments that would need
127 to be removed in case the inversion is selected (note: syntenic regions cannot reside within
128 inversions), since some inversions might cross syntenic alignments (see Additional File 1: Figure
129 S1 for example). Candidate inversions with a negative score are removed.
130 The remaining candidate inversions can overlap and/or contradict with each other (Figure 3b (ii)).
131 To resolve such conflicts, SyRI generates a second DAG where each node is a candidate inversion
132 with the inversion score as node-score and each edge connects inversions which do not conflict
133 (i.e. do not overlap or intersect). Similar to the syntenic path identification, additional start (S)
134 and end (E) nodes and corresponding edges are added, and dynamic programming is used to
135 select the highest-scoring path from S to E. Thus, SyRI identifies the best set of non-conflicting
136 inversions which maximises the total alignment score from inversions and the syntenic path
137 simultaneously.
138
139 **Translocation and duplication (TD) identification.** Since syntenic and inverted regions are
140 already annotated, all remaining alignments are either involved in TDs or are of repetitive nature
141 and need to be filtered out. For this, SyRI again starts by generating a third genome graph, this
142 time using only the so-far unannotated alignments, where again each node corresponds to an
143 alignment and an edge implies co-linearity between the respective alignments (Additional File 1:
144 Figure S2). Nodes corresponding to alignments which are separated by any other annotated
145 region (syntenic or inverted) are not connected by an edge (Additional File 1: Figure S3).
146 Additional start (S) and end (E) nodes and edges from these nodes to all other nodes are added.

In this graph, each path from S to E corresponds to a candidate TD. Each candidate TD is given a score based on its alignment length and gap length between consecutive alignments (Methods). Low scoring candidates and those that are overlapping with syntenic or inverted regions are filtered out.

Like inversions, selected candidate TDs can overlap with each other. To resolve these conflicts, SyRI first clusters overlapping candidate TDs, generating a network of co-aligned regions (Additional File 1: Note 3). These networks are processed using *progressive elimination* which involves iterative selection of alignments which do not overlap with other alignments (Figure 3c). Such *unique alignments* imply that the respective candidate TDs are necessary for annotating these genomic regions as there are no alternative alignments which could annotate these regions. In contrast, alignments which correspond to annotated regions within both of the genomes are removed as they are redundant (repetitive) alignments. This procedure can result in deadlocks, which refers to cases when it is unclear which candidate should be selected next during *progressive elimination*. To resolve such deadlocks, SyRI uses progressive elimination in conjunction with brute-force (for small networks) and randomized-greedy algorithms (for large networks). In the brute-force approach, all possible sets of non-conflicting candidate TDs within a network are generated. The score of each set (sum of candidate TD scores) is calculated and the set with the highest score is selected. In the randomized-greedy algorithm, deadlocks are resolved by selecting one of the highest scoring candidates at random with its probability weighed to its alignment score followed again by progressive elimination. This is repeated until all candidates are either annotated as TD or removed as redundant. Afterwards, SyRI identifies all intra-chromosomal TDs, inter-chromosomal TDs are annotated in the same way.

Grouping of alignments to generate annotation blocks. SyRI combines annotated alignments to generate annotation blocks where a *block* corresponds to a genomic region (alignments and the unaligned space within it) which constitute a structural event. For example, a syntenic block would contain all consecutive uninterrupted co-linear alignments and the unaligned regions between these alignments. Inversion (or TD) blocks would include all alignments which together form the extent of one inversion (or one TD).

Local variation identification. Local, small variations (like SNPs and small indels) are parsed out from individual alignments using MUMmer or corresponding CIGAR strings [12]. Local, large variations (structural variations like long indels) lead to alignment discontinuity. SyRI compares the overlaps and gaps present between all consecutive alignments within each of the annotation blocks and annotates indels, highly divergent regions (HDRs) and CNVs/tandem repeats (Additional File 1: Figure S4) similar to the structural variation predictions of Assemblytics [10]. Finally, SyRI reports all *un-aligned regions* which are not part of any annotation block. These regions reside between neighboured annotation blocks and are not classified further.

Validating SyRI using genetics

We used SyRI to predict genomic differences between two of the best described *A. thaliana* accessions, Col-0 and L*er*. Col-0 is the standard lab strain, which was used for the generation of a high-quality reference sequence [17]. The L*er* genome was recently assembled using a combination of different genomics technologies [18]. Both genomes are highly inbred and are

considered as fully homozygous. Here, we generated a new chromosome-level *de novo* assembly of the homozygous L*er* genome using PacBio reads. CN50 and CL50 (which are chromosome-number corrected metrics for C50 and L50 values) were 12.6Mb and 1, respectively [19] (see Additional File 2: Table S1, Methods for further information).

Based on a WGA between the two assemblies, SyRI predicted 1,979 syntenic blocks (104.5Mb, 88.5%), 4,630 SR blocks (9.7Mb, 8.2%), and 2,237 non-aligned regions (3.9Mb, 3.3%) (values are with respect to the L*er* genome; similar values for Col-0; Figure 4a). Syntenic blocks were much larger (median length: 22 kb) as compared to SR blocks (median length: 1.2 kb) (Figure 4b). SRs were not randomly distributed in the genome, but clustered in specific regions. An example of a highly rearranged region affecting the location of several genes, and how SyRI annotates the individual SR in this region can be found in Figure 4c. More than half of the predicted SRs reside within the peri-centromeres, which are known to be structurally diverse, repeat rich, and low in gene density [17]. Overall, 4,095 (88.5%) of the SR blocks were at least partially overlapping with transposable elements, while 2,238 (48%) of them were found entirely within transposable elements. A summary of the local variations can be found in Additional File 1: Figure S5.

To validate some of the predicted rearrangements we used a genetic approach which was based on the observation that recombinant genomes can have different copy numbers of translocated DNA (Figure 1; 5a): The copy number of translocated DNA in a recombinant genome relies on the genotypes at the two insertion sites of the translocation. For example, translocated DNA is duplicated if the two insertion sites of translocated DNA are combined into one recombinant chromosome. In contrast, regions with are not translocated do not vary in copy-number - irrespective of any recombination in the genome.

We used available whole-genome sequencing data of a set of 50 F_2 recombinant plants, which were generated by crossing Col-0 and L*er*, followed by self-pollination of the resulting F_1 hybrids (preprint [20]). We aligned the short reads (~5x genome coverage/sample) to the Col-0 reference sequence and used the genotypes at ~500k SNP [18] markers to reconstruct the parental haplotypes using TIGER [21] (Figure 5b). As expected, almost all of the crossing-over sites (99.5%) overlapped with syntenic regions.

For each predicted translocation, we estimated a copy number in each sample using this haplotype information and used three different metrics to test for its existence. The first two were based on the assumption that all reads from a translocated region align to the same loci in the reference genome independent of the location of the actual allele(s) in the sequenced sample (Figure 5b, [22]). Accordingly, it should be possible to estimate the copy-number of a translocation using changes in read coverage. For the first test, we analysed the absence of reads in translocated regions in recombinant genomes, which were predicted to feature no copy of the translocated region (Figure 5c) (using 0.2x read coverage as a cut-off to distinguish between absence or presence of a translocation). For the second test, we assessed the goodness-of-fit between expected copy-number and observed copy-number for a translocation across all recombinants (as estimated from the normalized read counts in the translocation regions; Figure 5d; Methods). Third, we tested for allele count differences, where allele count refers to the number of reads having the Col-0 (or L*er*) alleles at SNP markers within a translocated region and that allow to distinguish the two alleles of the translocation. Depending on the copy number of the different alleles of a translocation, the allele count should also vary. In consequence, samples with the same genotypes at the two loci of a translocation should have similar allele counts.

Clustering the samples based on their allele count values should, therefore, result in a clustering of the samples according to their genotypes at the loci associated with translocations (Figure 5e). In contrast, the allele counts of falsely predicted translocations would be independent of the genotype and as a result, samples with the same genotypes should not cluster together (Methods).

We selected 135 translocations (intra- and inter- chromosomal), which were larger than 1kb and were not part of the peri-centromeres for validation. We tested all translocations with at least one of the tests described above, of which 92% (124) could be confirmed by at least one test, while 39% (52/135) were even confirmed by two or three tests (Figure 5f). We manually checked the read alignments in the regions of the eleven translocations that could not be confirmed and found support for the existence of each of the translocations, which however had not been strong enough to be identified by any of the three automated genetic tests.

Performance evaluation

We further validated SyRI by performing a simulation analysis where we randomly introduced SRs in the *A. thaliana* reference sequence (Methods) and compared these modified genomes against the original assembly using SyRI. For this, we simulated 100 genomes each for the three types of SRs: inversions, translocations, and duplications (Methods). SyRI was able to consistently identify correct SRs and obtained >95% of sensitivity and precision values across all samples (Additional File 1: Figure S6). False results were usually a consequence of SRs in repeat regions where alternative annotations were equally likely. For comparison, we also run AsmVar with these simulated genomes [9]. By design, AsmVar can also find translocations and inversions, but

will not identify duplications. The performance quality of AsmVar for the identification of translocations was very high and similar to SyRI (>95% sensitivity and precision), however, inversions were only found in only 40 samples (sensitivity <10%), albeit with high precision (Additional File 1: Figure S6).

To evaluate the prediction of local variation, we compared the indels identified by SyRI, Assemblytics [10], and AsmVar [9] when applied to the comparison of the *A. thaliana* Col-0 and L*er* assemblies. This allowed assessing their differences in the context of actual genomic differences. SyRI, Assemblytics, and AsmVar identified 188559, 166161, and 145676 indels respectively. Overall, 97% (160568) and 78% (114063) of all indels identified by Assemblytics and AsmVar were identified by SyRI as well (Methods). This shows that there is significant agreement between SyRI and Assemblytics, while the overlap between SyRI and AsmVar is lower.

In addition, SyRI separated the 188559 indel predictions into those which reside in syntenic regions (175260) and those in rearranged regions (13299). The overlap of the predictions in syntenic regions was high for both tools (89% (155607) and 63% (110891) for Assemblytics and AsmVar), while the number and the overlap for the indels predicted in structurally rearranged regions was low for both tools (35% (4652) and 32% (4238) for Assemblytics and AsmVar). Together this suggested that SyRI's performance is powerful even when analyzing complex regions.

Effect of genome contiguity

For the identification of SRs from incomplete assemblies (not at chromosome level), pseudo-chromosomes can be generated using homology to a reference sequence (or any other

assembly). To analyse the effect of the original assembly contiguity on SyRI's performance, we performed a simulation analysis where we first generated scattered assemblies of the L*er* genome (Methods) which were then reassembled to generate pseudo-genomes using their homology with Col-0 genome by RaGOO (preprint [13]). We then identified SRs between these pseudo-L*er* genomes and Col-0 (similarly to earlier) using SyRI and compared the predicted SRs with the SRs identified between the chromosome level assemblies of L*er* and Col-0.

Based on these simulations, assemblies with N50 of more than 571Kb had a more than 90% chance to have a sensitivity of more than 0.9 (Figure 6). Similarly, there was a 90% chance to have a precision of more than 0.9 for assemblies with N50 more than 674Kb. However, even for assemblies with N50 of ~470Kb, sensitivity and precision values were 0.89 and 0.86 respectively. We also evaluated SyRI's efficiency in identifying SRs when both genomes are at scaffold level. For this, we generated scattered assemblies from the Col-0 and L*er* genomes. Since current pseudo-chromosome generation tools only concatenate scaffolds of one assembly using homology with another assembly, we developed a heuristic script to generate homology-based pseudo-chromosomes using two incomplete assemblies (Additional File 1: Note 2). As earlier, we identified SRs from these pseudo-genomes and compared them to the SRs identified between the full-length assemblies of Col-0 and L*er*. For assemblies with N50 of more than 940Kb, there was a 70% chance to have sensitivity and precision values of more than 0.7 (Additional File 1: Figure S7). For assemblies with low contiguity (N50: ~470Kb), sensitivity and precision values of 0.66 and 0.67, respectively, could be obtained. The median sensitivity and precision values were 0.7. This shows that SyRI can be used to get useful insights about SRs even in situations when a chromosome-level reference genome is not available.

Comparing human, yeast, fruit fly, and maize genomes

To test and demonstrate SyRI's usefulness and scalability regarding genome size and repetitiveness, we searched for intra-species genomic differences in four different model organisms: human, yeast, fruit fly, and maize (Additional File 2: Table S1). For its application to human genomes, we compared two assemblies, which included the extensively studied sample NA12878 [23] and an adult Yoruban female genome (NA19240, [24]), against the reference genome GRCh38.p12 [25]. For yeast, we compared the *de novo* assembly of strain YJM1447 [26] against the reference genome from strain S288C [27]. For fruit fly (*D. melanogaster*), the *de novo* assembly of strain A4 [28] was compared to the reference genome [29]. For maize, we compared the *de novo* assembly of PH207 [30] against the B73 reference genome [31]. To limit computational requirements, we masked the highly repetitive maize genome while all other genomes were analysed without masking [32].

SyRI was used to predict syntenic regions, SRs, and local variations in all genomes (Table 1, Additional File 1: Figure S8–S12). In each comparison, including human, at least 5% of the assembled genomes were found to be structurally different. Runtime and memory usage were dependent on genome complexity (including repetitiveness). The CPU runtime for the smaller and simpler yeast genomes was 43 seconds, whereas for the two human genomes SyRI took ~10 minutes, while memory usage was less than 1 GB for each of the comparisons (Table 1) (without considering SNPs and small indel parsing). The exception was the comparison of the repetitive maize genomes, which took ~1hr of CPU time and ~7GB of RAM. Since SyRI considers all alignment combinations, the runtime and memory usage can be high in such repetitive genomes (see Additional File 1: Note 4 and Figure S12 for more information), however, the number of

320 alignments can be drastically reduced by reducing the WGA sensitivity (i.e. omitting small, 10-
321 100s bp alignments), which in turn decreases runtime and memory consumption of SyRI.

322 Discussion

323 We introduced SyRI, a tool that identifies any kind of genomic differences using pair-wise
324 comparisons of whole genome assemblies. However, instead of identifying differences directly,
325 SyRI starts by identifying all syntenic regions between the genomes. As all non-syntenic regions
326 must result from rearrangements, identifying syntenic regions implies the identification of
327 rearranged regions as well. This transforms the challenge of SR identification to the
328 comparatively easier challenge of SR classification.
329 SyRI is based on a genome graph and uses optimisation methods to identify the syntenic path
330 between the corresponding chromosomes and select alignments (nodes) which represent
331 structural rearrangements between all pairs of chromosomes while maximising the total
332 alignment score. This allows SyRI to annotate even highly complex and repeat rich regions.
333 In addition, SyRI identifies local genomic variations residing either in syntenic or rearranged
334 regions. The identification of local variations in rearranged regions introduces a hierarchy of
335 genomic variations (e.g., SNPs in translocated regions). It is important to distinguish between the
336 different types of SNPs, as SNPs in rearranged regions are differently inherited as compared to
337 SNPs in syntenic regions. The genotypes at rearranged SNPs can confound the interpretation of
338 genomic patterns during selection screens, genome-wide association or recombination analysis
339 as they can show unexpected genotypes [33,34]. SyRI now offers a straight-forward solution to
340 find SNPs in syntenic regions assuming whole-genome assemblies are available.

Finally, though implemented in a genome graph that is build up from local alignments generated by a WGA, our algorithm can be easily adapted for SR identification in other types of genome graphs as well [35,36].

Conclusions

We have developed SyRI which, to our knowledge, is the first tool to identify all classes of structural rearrangements between two genome assemblies following the identification of syntenic regions. This novel approach of SR identification is highly efficient and also identifies all local genomic difference in their genomic context (i.e. whether local variation resides in syntenic or rearranged regions). Using SyRI, we identified SRs and local variations in humans, *A. thaliana*, fruit fly, yeast, and maize genomes. Further, using the individual genomes of a population of recombinant individuals generated by crossing two *A. thaliana* accessions allowed us to validate the predicted translocations. SyRI is available as an open source tool and is being actively developed and improved.

Methods

Long read sequencing of the genome of *A. thaliana* L*er*

A. thaliana L*er* plants were grown in the greenhouse at the Max Planck Institute for Plant Breeding Research. DNA was extracted using the NucleoSpin® Plant II Maxi Kit from Macherey-Nagel. We used the PacBio template prep kit > 20 kb for Sequel systems (SMRTbell Template Prep Kit 1.0-SPv3) with damage repair (SMRTbell Damage Repair Kit -SPv3) and BluePippin size

selection for fragments > 9/10 kb. Sequencing of two SMRT cells was done with the Sequel Sequencing Plate 1.2 and the Sequel Binding Kit 1.0. Movie Time 360 min.

Assembly generation

Raw reads were filtered to remove small and low-quality reads (length<50bp and QV<80), corrected and *de novo* assembled using Falcon [37], followed by polishing with Arrow in the SMRTLink5 package, and finally corrected using Illumina short read alignments with reads from an earlier project [18]. The contigs from organellar DNA sequences were removed, all others were anchored into pseudo-chromosome based on homology with the reference sequence. Adjacent contigs were connected with a stretch of 500 “N” characters.

Whole genome alignments

All assemblies used in this work were filtered to select only chromosome-representing contigs (unplaced scaffolds were removed). We used the *nucmer* alignment tool from the MUMmer toolbox [12] to perform WGAs. Nucmer was run with --maxmatch to get all alignments between two genomes including -c, -b, and -l parameters which were selected to balance alignment resolution and runtime based on genome size and number of repeat regions (full commands are available in Additional File 2: Table S2). Alignments were filtered using the *delta-filter* tool and the filtered delta files were converted to the tab-delimited files using the *show-coords* command. Before whole-genome alignments, both maize genomes were masked using RepeatMasker v4.0.6 [38].

TD candidates score

Each selected candidate TD was given a score based on the length of alignments in it and the gap length between the alignments; TD score $= \min\left(\frac{genA\_aligned\_length - genA\_gap\_length}{genA\_aligned\_length}, \frac{genB\_aligned\_length - genB\_gap\_length}{genB\_aligned\_length}\right)$.

Data extraction and transformation of the 50 recombinant genomes

We used whole-genome sequencing data of 50 F_2 recombinant plants that were generated in the course of a different project (preprint [20]) for validating SR predictions. We extracted allele count information from consensus call files generated by SHORE [39]. For each predicted translocation, we estimated its copy number as the ratio between average read-coverage for the translocated region and the average read-coverage across the entire genome of the respective sample. Translocations in the centromeric regions and for which more than 25% of the translocated sequence had at least 10% reads with Ns were filtered out. For allele count analysis, we selected high-confidence (25bp conserved in both directions) SNPs in translocated regions as markers. Translocated regions (and corresponding SNP markers) were classified according to the genotypes (as predicted by TIGER) in individual samples at the two associated loci.

Validation of translocation: Absence of reads (Test 1)

We selected F2 samples which, according to predicted genotypes, should have lost the translocated DNA and thus should not give rise to any reads from the translocated region. Only translocations for which at least two samples that had lost the translocated regions were tested. And only those translocations for which all tested samples had no reads were considered validated.

Validation of translocation: Expected vs. observed copy number (Test 2)

For each translocation, we selected samples which had different genotypes at the two associated loci for the translocation. This removes some of the samples with two copies and helps to remove bias towards genomes with a copy number of two, which can affect this test. We further only selected translocations for which we found samples with at least three different copy-number values predicted. A linear model was fit using the *lm* function in *R*. *P*-values for the model-fit were adjusted for multiple testing using the *BH*-method [40], and translocations for which adjusted *p*-values were less than 10^{-6} and slope more than 0.75 were considered as valid.

Validation of translocation: Genotype clustering (Test 3)

Allele count values at the SNP markers were normalized and outliers (markers having very high allele counts) were removed. Translocations were tested only when they had at least two different classes of samples (genotypes) with each class having at least three samples, and at least three SNP markers in the translocated regions. Translocations for which alternate allele counts did not change across the samples (variance < 1) were also filtered out.

Cluster fit calculation: First, the distance between two samples was defined as the Euclidean distance between their reference allele counts and alternate allele counts. Then, the *closeness_score* was calculated as the sum of ratios of the average distance between the samples belonging to a genotype to the average distance to samples of other genotypes.

Simulating distributions: Background distributions for the closeness_score were simulated by generating random clusters. For each sample, allele counts (reference and alternate) were sampled using a Poisson distribution. For true translocations, the closeness_score would be low as samples from the same genotype would be much closer to each other (leading to smaller numerator), whereas samples from different genotypes would be far (leading to large

423 denominator). For each translocation, we calculated the lower-tail *p*-value of getting the
424 corresponding closeness_score. *P*-values were adjusted for multiple testing using *BH*-method,
425 and translocations with *p*-value < 0.05 were considered valid.

426 **Running AsmVar and Assemblytics**

427 AsmVar was run based on the demo script provided with the tool. For genome assembly
428 alignment, lastdb was run using the default parameters, whereas lastal and last-split were run
429 using the parameters provided in the above demo [41]. Similarly, variants were detected using
430 ASV_VariantDetector tool of AsmVar using the default parameters.
431 Assemblytics was run through the assemblytics web-page (http://assemblytics.com/). We used
432 the same filtered delta file which was used to perform Col-0 vs L*er* comparison. Assemblytics was
433 run with default parameters but minimum and maximum variant size were set to 1 and 100000
434 respectively.

435 **Comparing SyRI with AsmVar and Assemblytics**

436 For comparing SRs, we simulated SRs in the *A. thaliana* reference genome using the *R* package
437 *RSVSim* [42]. We simulated 40, 436, and 1241 events for inversions, translocations, and
438 duplications respectively. The number of rearrangements and their corresponding sizes were
439 sampled from real differences found between the Col-0 and L*er* genomes. The simulated
440 genomes were aligned back to the unmodified reference genome and SRs were identified using
441 SyRI and AsmVar. A predicted SR was considered to match an original SR if the end-points for the
442 predicted SR were within ±150bp of the end-points of the original SR.

For comparing SVs, we compared the *A. thaliana* reference sequence against the newly generated L*er* assembly. For running Assemblytics, we used the web-tool with delta file, same as used by SyRI, as input. For AsmVar, we used the pipeline as described above. An insertion (or deletion) identified by Assemblytics was considered to be predicted by SyRI if an insertion (or deletion) of the same length exists at the same coordinate in the reference genome. To accommodate for variation arising because of use of different aligners (MUMmer for SyRI, and LAST for AsmVar), an indel predicted by AsmVar was considered to be predicted by SyRI if its position in reference genome is within 50bp up/down-stream of an indel of the same type predicted by SyRI and both indels have the same length. Similarly, an indel identified by SyRI was considered to be present in Assemblytics output if an indel of same type and length exists at the same reference genome coordinate, whereas it was considered to be present in AsmVar output if an indel of same type and length exists with within 50bp margin loci at the reference genome.

Pseudo-chromosome generation and output comparison

We generated 200 fragmented assemblies of the L*er* genome by introducing 10-400 random break points in the corresponding chromosome-level assemblies. Pseudo-genomes were generated for each of the fragmented assemblies using RaGOO with default parameters. Additionally, we generated 100 fragmented assemblies each of Col-0 and L*er* genomes by introducing 10-400 random breakpoints. These fragmented assemblies were assembled by a heuristic script (Additional File 1: Note 2) to generate pseudo-molecules. For 16 assemblies, pseudo-molecule generation failed and these samples were skipped from further analysis.

An SR identified from the pseudo-genomes was considered to be correct if the same rearrangement type was present at the same reference genome loci (within a 100bp up/downstream margin).

Declarations

Availability of data and materials

The assembly of the L*er* genome has been submitted to the European Nucleotide Archive (http://www.ebi.ac.uk) and is publicly available under the accession number GCA_900660825. All other assemblies are publicly available at NCBI (https://www.ncbi.nlm.nih.gov/), and their accession numbers are GCA_000001735.3 [17], GCA_000001405.27 [25], GCA_002077035.3 [23], GCA_001524155.4 [24], GCA_000146045.2 [27], GCA_000977955.2 [26], GCA_000001215.4 [29], GCA_002300595.1 [28], GCA_000005005.6 [31], GCA_002237485.1 [30]. Further details about the assemblies are in Additional File 2: Table S1. BAM files for the 50 F_2 recombinant genomes are available at European Nucleotide Archive under the project ID PRJEB29265 (preprint [20]). SyRI is freely available under the MIT license and is available online [43]. SyRI is developed using Python3 and is platform independent.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the German Federal Ministry of Education and Research in the frame of RECONSTRUCT (FKZ 031B0200A-E).

Authors' contributions

The project was conceived by KS and WBJ. MG and KS developed the algorithms. MG implemented SyRI and performed all analyses. HS processed recombinant genome sequencing data and identified crossing-over sites. WBJ generated the L*er* assembly. The manuscript was written by MG and KS with inputs from HS and WBJ. All authors read and approved the final manuscript.

Acknowledgements

The authors would like to thank Ulrike Hümann for help with plant work and Detlef Weigel and Gunnar Klau for helpful comments on the manuscript. We would also like to thank the researchers at the Genome Institute at Washington University School of Medicine who shared the assemblies for the NA12878 and NA19240 genomes.

References

1. Weischenfeldt J, Symmons O, Spitz F, Korbel JO. Phenotypic impact of genomic structural variation: insights from and for human disease. Nat Rev Genet. 2013; 14(2):125–38.
2. Tuttle EM, Bergland AO, Korody ML, Brewer MS, Newhouse DJ, Minx P, et al. Divergence and Functional Degradation of a Sex Chromosome-like Supergene. Curr Biol. 2016; 26(3):344–50.
3. Küpper C, Stocks M, Risse JE, dos Remedios N, Farrell LL, McRae SB, et al. A supergene determines highly divergent male reproductive morphs in the ruff. Nat Genet. 2016; 48(1):79–83.
4. Lamichhaney S, Fan G, Widemo F, Gunnarsson U, Thalmann DS, Hoeppner MP, et al.

Structural genomic changes underlie alternative reproductive strategies in the ruff (Philomachus pugnax). Nat Genet. 2016; 48(1):84–8.

5. Lowry DB, Willis JH. A Widespread Chromosomal Inversion Polymorphism Contributes to a Major Life-History Transition, Local Adaptation, and Reproductive Isolation. Barton NH, editor. PLoS Biol. 2010; 8(9):e1000500.

6. Simpson JT, Pop M. The Theory and Practice of Genome Sequence Assembly. Annu Rev Genomics Hum Genet. 2015; 16(1):153–72.

7. Jiao W-B. The impact of third generation genomic technologies on plant genome assembly. Curr Opin Plant Biol. 2017; 36:64–70.

8. Sedlazeck FJ, Lee H, Darby CA, Schatz MC. Piercing the dark matter: Bioinformatics of long-range sequencing and mapping. Nat Rev Genet. 2018; 19(6):329–46.

9. Liu S, Huang S, Rao J, Ye W, Krogh A, Wang J. Discovery, genotyping and characterization of structural variation and novel sequence at single nucleotide resolution from de novo genome assemblies on a population scale. Gigascience. 2015; 4(1).

10. Nattestad M, Schatz MC. Assemblytics: a web analytics tool for the detection of variants from an assembly. Bioinformatics. 2016; 32(19):3021–3.

11. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: A fast and versatile genome alignment system. Darling AE, editor. PLOS Comput Biol. 2018; 14(1):e1005944.

12. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004; 5(2):R12.

13. Li H. Minimap2: pairwise alignment for nucleotide sequences. Birol I, editor.

526 Bioinformatics. 2018; 34(18):3094–100.

527 14. Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, et al. Fast and
528 accurate reference-guided scaffolding of draft genomes. bioRxiv. 2019; :519637.

529 15. Delcher AL, Phillippy A, Carlton J, Salzberg SL. Fast algorithms for large-scale genome
530 alignment and comparison. Nucleic Acids Res. 2002; 30(11):2478–83.

531 16. Delcher AL, Kasif S, Fleischmann RD, Peterson J, White O, Salzberg SL. Alignment of whole
532 genomes. Nucleic Acids Res. 1999; 27(11):2369–76.

533 17. Initiative TAG. Analysis of the genome sequence of the flowering plant Arabidopsis
534 thaliana. Nature. 2000; 408(6814):796–815.

535 18. Zapata L, Ding J, Willing E-M, Hartwig B, Bezdan D, Jiao W-B, et al. Chromosome-level
536 assembly of *Arabidopsis thaliana* L *er* reveals the extent of translocation and inversion
537 polymorphisms. Proc Natl Acad Sci. 2016; 113(28):E4052–60.

538 19. Jiao W-B, Accinelli GG, Hartwig B, Kiefer C, Baker D, Severing E, et al. Improving and
539 correcting the contiguity of long-read genome assemblies of three plant species using
540 optical mapping and chromosome conformation capture data. Genome Res. 2017;
541 27(5):778–86.

542 20. Sun H, Rowan BA, Flood PJ, Brandt R, Fuss J, Hancock AM, et al. Linked-read sequencing of
543 gametes allows efficient genome-wide analysis of meiotic recombination. bioRxiv. 2018;
544 :484022.

545 21. Rowan BA, Patel V, Weigel D, Schneeberger K. Rapid and Inexpensive Whole-Genome
546 Genotyping-by-Sequencing for Crossover Localization and Fine-Scale Genetic Mapping. G3
547 Genes, Genomes, Genet. 2015; 5(3):385–98.

548 22. Imprialou M, Kahles A, Steffen JG, Osborne EJ, Gan X, Lempe J, et al. Genomic
549 Rearrangements inArabidopsisConsidered as Quantitative Traits. Genetics. 2017;
550 205(4):1425–41.
551 23. The Genome Institute at Washington University School of Medicine. NA12878_prelim_3.0.
552 https://www.ncbi.nlm.nih.gov/assembly/GCA_00207703. 2018.
553 24. The Genome Institute at Washington University School of Medicine. NA19240_prelim_3.0.
554 https://www.ncbi.nlm.nih.gov/assembly/GCA_001524155.4. 2017.
555 25. Human Genome Sequencing Consortium I. Finishing the euchromatic sequence of the
556 human genome. Nature. 2004; 431(7011):931–45.
557 26. Strope PK, Skelly DA, Kozmin SG, Mahadevan G, Stone EA, Magwene PM, et al. The 100-
558 genomes strains, an S. cerevisiae resource that illuminates its natural phenotypic and
559 genotypic variation and emergence as an opportunistic pathogen. Genome Res. 2015;
560 25(5):762–74.
561 27. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, et al. Life with 6000
562 Genes. Science (80-). 1996; 274(5287):546–67.
563 28. Chakraborty M, VanKuren NW, Zhao R, Zhang X, Kalsow S, Emerson JJ. Hidden genetic
564 variation shapes the structure of functional elements in Drosophila. Nat Genet. 2018;
565 50(1):20–5.
566 29. Hoskins RA, Carlson JW, Wan KH, Park S, Mendez I, Galle SE, et al. The Release 6 reference
567 sequence of the *Drosophila melanogaster* genome. Genome Res. 2015; 25(3):445–58.
568 30. Hirsch CN, Hirsch CD, Brohammer AB, Bowman MJ, Soifer I, Barad O, et al. Draft Assembly
569 of Elite Inbred Line PH207 Provides Insights into Genomic and Transcriptome Diversity in

570 Maize. Plant Cell. 2016; 28(11):2700–14.

571 31. Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, et al. Improved maize reference genome
572 with single-molecule technologies. Nature. 2017; 546(7659):524.

573 32. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 maize genome:
574 complexity, diversity, and dynamics. Science. 2009; 326(5956):1112–5.

575 33. Wijnker E, Velikkakam James G, Ding J, Becker F, Klasen JR, Rawat V, et al. The genomic
576 landscape of meiotic crossovers and gene conversions in *Arabidopsis thaliana*. Elife. 2013;
577 2.

578 34. Qi J, Chen Y, Copenhaver GP, Ma H. Detection of genomic variations and DNA
579 polymorphisms and impact on analysis of meiotic recombination and genetic mapping.
580 Proc Natl Acad Sci U S A. 2014; 111(27):10007–12.

581 35. Paten B, Novak AM, Eizenga JM, Garrison E. Genome graphs and the evolution of genome
582 inference. Genome Res. 2017; 27(5):665–76.

583 36. Consortium TCP. Computational pan-genomics: status, promises and challenges. Brief
584 Bioinform. 2016; (August):bbw089.

585 37. Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al. Phased diploid
586 genome assembly with single-molecule real-time sequencing. Nat Methods. 2016;
587 13(12):1050–4.

588 38. Smit, AFA, Hubley, R & Green P. RepeatMasker Open-4.0. http://www.repeatmasker.org.

589 39. Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, Weigel D. Sequencing of
590 natural strains of Arabidopsis thaliana with short reads. Genome Res. 2008; 18(12):2024–
591 33.

592 40. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful
593 approach to multiple testing. J R Stat Soc Ser B. 1995; 57(1):289–300.
594 41. Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence
595 comparison. Genome Res. 2011; 21(3):487–93.
596 42. Bartenhagen C, Dugas M. RSVSim: an R/Bioconductor package for the simulation of
597 structural variations. Bioinformatics. 2013; 29(13):1679–81.
598 43. Goel M, Schneeberger K. SyRI: identification of syntenic and rearranged regions from
599 whole-genome assemblies. https://schneebergerlab.github.io/syri/.
600

601 ## List of Additional Files:

602

File name	File format	Title of data	Description of data
Additional File 1	.pdf	Additional notes and figures	Additional notes describing the method and additional results
Additional File 2	.xlsx	Additional tables	Information about methodology and data used

603 **Table 1:** Structural differences identified by SyRI and corresponding computation resources

604

Species	Sample	Assembly Size	CPU runtime (in secs)	Memory Usage (in MB)		Syntenic Regions	Structural Rearrangements				Un-aligned
							Inversion	Translocation	Duplication	CX	
Human	NA12878	3.03 Gb	505.7	722	size	2.77 Gb	7.0 Mb	10.4 Mb	21.1 Mb	8.5 Mb	220.7 Mb
					% genome	91.1	0.2	0.3	0.7	0.3	7.3
					number	1066	67	154	2335	1463	748
	NA19240	3.04 Gb	278.3	528	size	2.79 Gb	3.8 Mb	11.9 Mb	19.2 Mb	8.0 Mb	205.6 Mb
					% genome	91.8	0.1	0.4	0.6	0.3	6.8
					number	1021	69	151	2221	1250	748
Yeast	YJM1447	12.1 Mb	43.38	3	size	11.2 Mb	1.8 Kb	52.1 Kb	602.2 Kb	166.2 Kb	86.6 Kb
					% genome	92.5	0.02	0.4	5.0	1.4	0.7
					number	198	3	19	91	317	144
Fruit Fly	A4	135.5 Mb	545.6	242	size	124.9 Mb	131.6 Kb	1.35 Mb	3.6 Mb	4.2 Mb	1.2 Mb
					% genome	92.2	0.1	1.0	2.7	3.1	0.9
					number	1693	16	393	2058	2478	1160
Maize	PH207	2.06 Gb	3400	6934	size	1.3 Gb	83 Mb	4.4 Mb	4.9 Mb	16.2 Mb	638 Mb
					% genome	63.7	4.0	0.2	0.2	0.8	31
					number	8439	195	1384	2548	9527	14319

Figure legends

Figure 1: Hierarchy of genomic differences and their propagation. a) Local variation like SNPs/indels can occur in syntenic regions as well as in regions which are structurally rearranged (translocated in this example). b) A diploid cell containing such haplotypes. Following meiosis, gametes can feature different copy-number variation for the translocated regions.

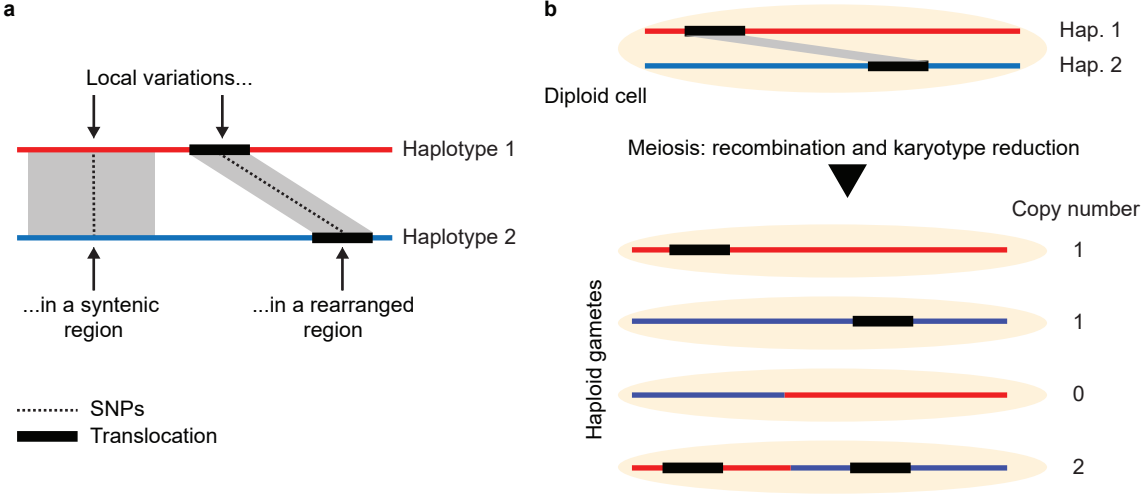
Figure 2: Workflow for the identification of genomic differences. SyRI takes WGA as input. A WGA consists of a set of local alignments, where each local alignment (grey polygon) connects a specific region in one genome to a specific region in the other genome. Step 1: SyRI identifies the highest scoring syntenic path between the corresponding genomes (blue alignments). The syntenic path represents the longest set of non-rearranged regions between two genomes. Step 2 (a-c): The remaining alignments are separated into structural rearrangements and redundant alignments. Structural rearrangements (green alignments) are further classified into inversions, intra-chromosomal translocations and duplications, and finally inter-chromosomal rearrangements. Step 3: Local variations are identified in all non-redundant regions. SNPs and small indels are parsed directly from the local alignments, whereas more complex variation (e.g. like large indels and CNVs) are identified in the overlaps and gaps between consecutive local alignments. Finally, all un-aligned regions in between syntenic and rearranged regions are reported for completeness.

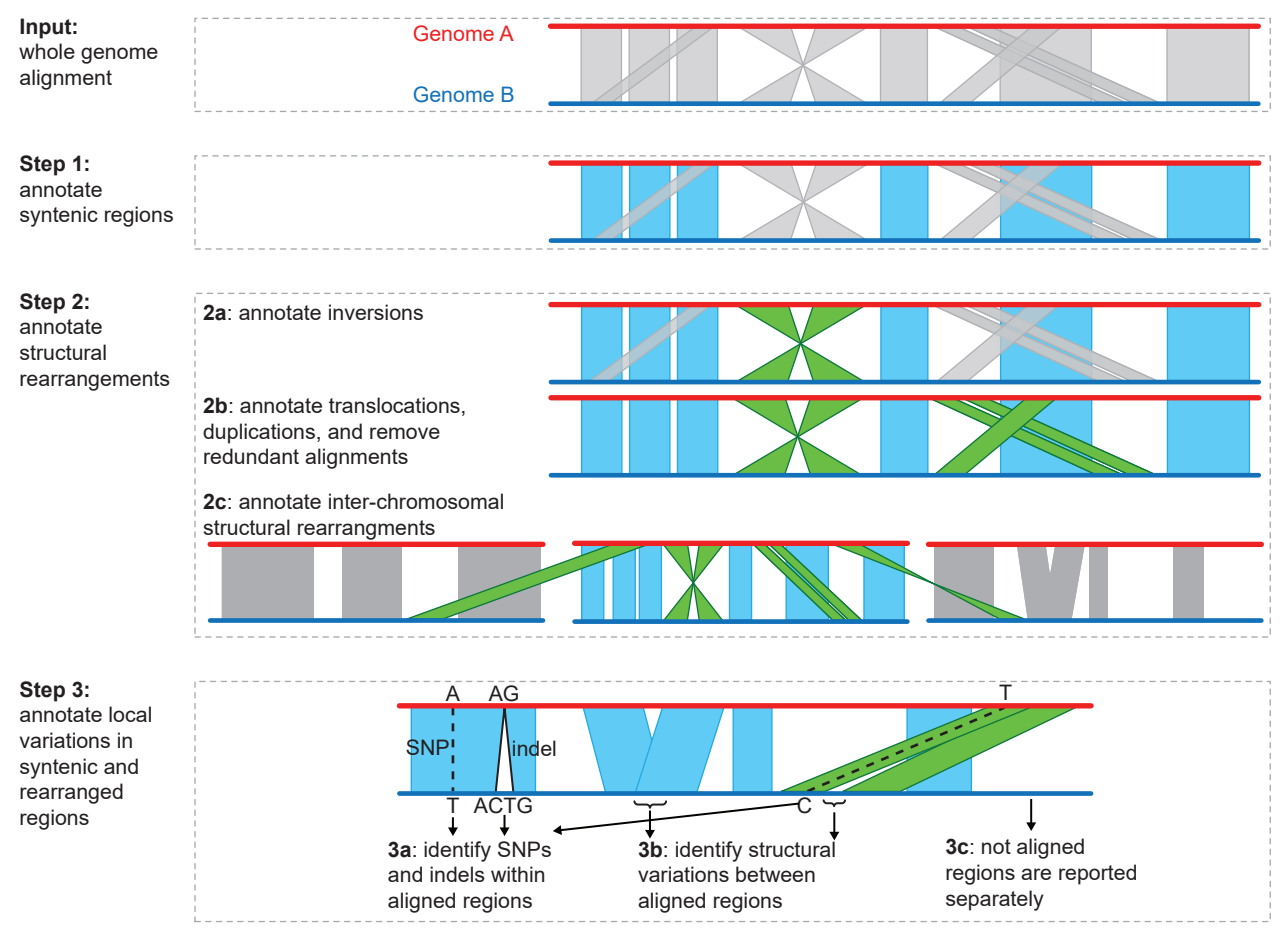
Figure 3: Overview of the different algorithms implemented in SyRI. (a) Syntenic region identification. A WGA between genomes A and B is converted to a genome graph where nodes represent local alignments and edges are introduced if the corresponding alignments are co-linear (like alignments 'a' and 'e'). Nodes corresponding to overlapping (like 'b' and 'e') or inter-crossing (like 'g' and 'i') alignments are not connected. Additional start and end (S/E, white) nodes are added and connected to the nodes without in-going (S) or out-going (E) edges. The longest path from S to E is identified (blue line) and the corresponding alignments represent syntenic path (blue outline). (b) Inversion identification is performed in two steps. First, the (i) alignments based on reverse complemented regions (grey) are converted to a genome graph, where each node is an alignment and an edge is added between two nodes if they can be part of the same inversion. Start (S) and end (E) nodes are added and edges are introduced to all other nodes. Each path from S to E corresponds to a candidate inversion. (ii) Here, seven inversions are possible. SyRI then generates a second graph (again including S and E nodes) where nodes represent candidate inversions and edges illustrate that these inversions can co-exist. The longest path from S to E (green line) corresponds to the highest scoring set of inversions that can co-exist (alignments with green outline). (c) Translocation and duplication identification. (i) Each set of overlapping alignments (orange alignments) are analyzed individually. Within such a network of alignments, (ii) SyRI selects necessary alignments (which are present in regions where no other alignments can be found, green outline), and (iii) removes redundant alignments (white). (iv) Deadlocks occur when no further alignments can be classified and are solved using brute-force and randomized-greedy methods.

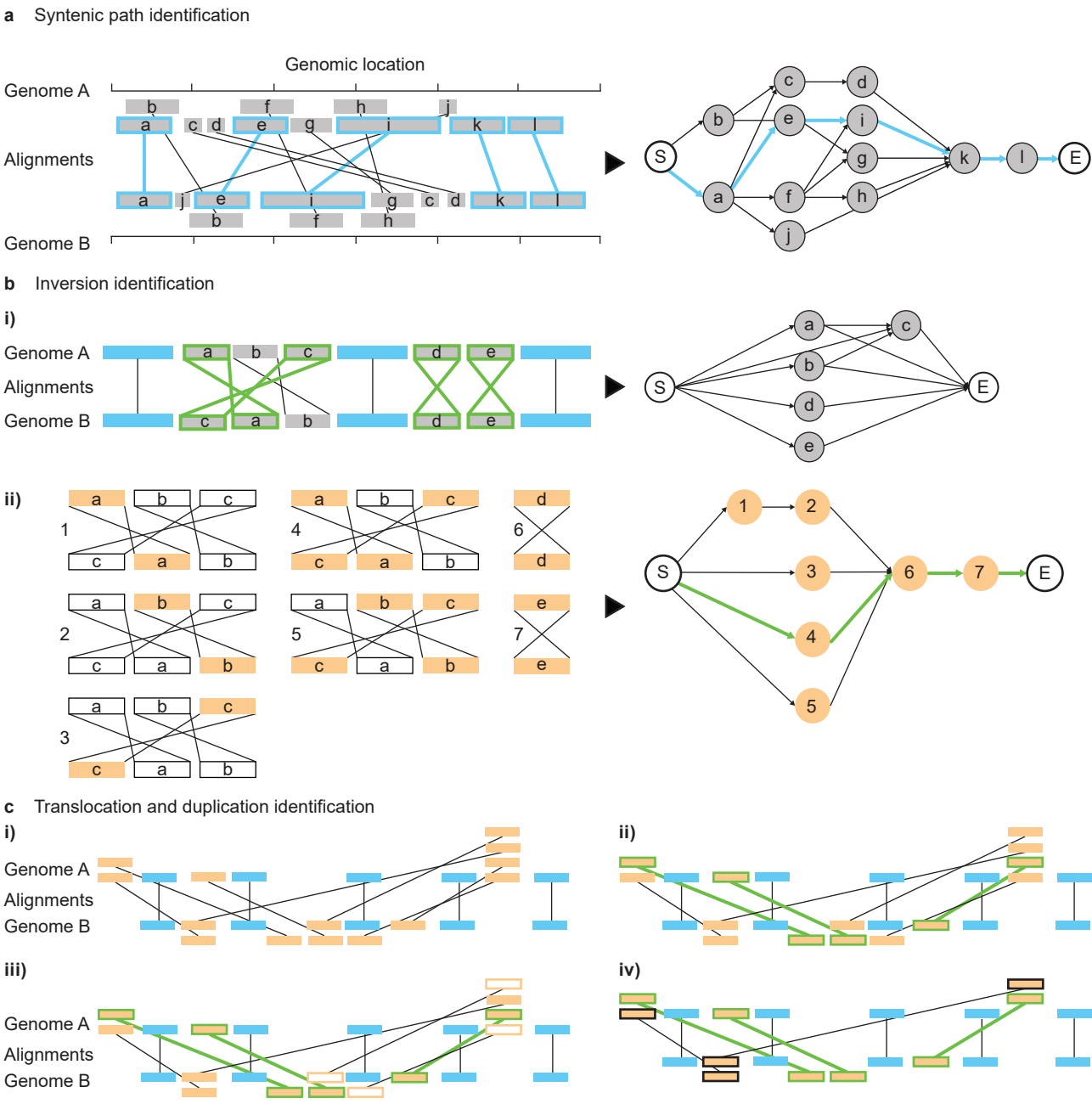
Figure 4: Genomic differences between *A. thaliana* Col-0 and L*er*. (a) Total length and number of syntenic, SR, and un-aligned regions identified between the Col-0 and L*er* genomes (syn: syntenic regions, inv: inversions, tl: translocations, dup: duplications, cx: cross-chromosomal exchange, notal: un-aligned). (b) Length distribution of all annotation types. (c) Example of a highly rearranged region at chromosome 5. Two large inversions, a translocation, and a duplication occur together. Multiple genes (black lines) are present in this region. Smaller SRs in this region have been filtered for illustrative reasons.

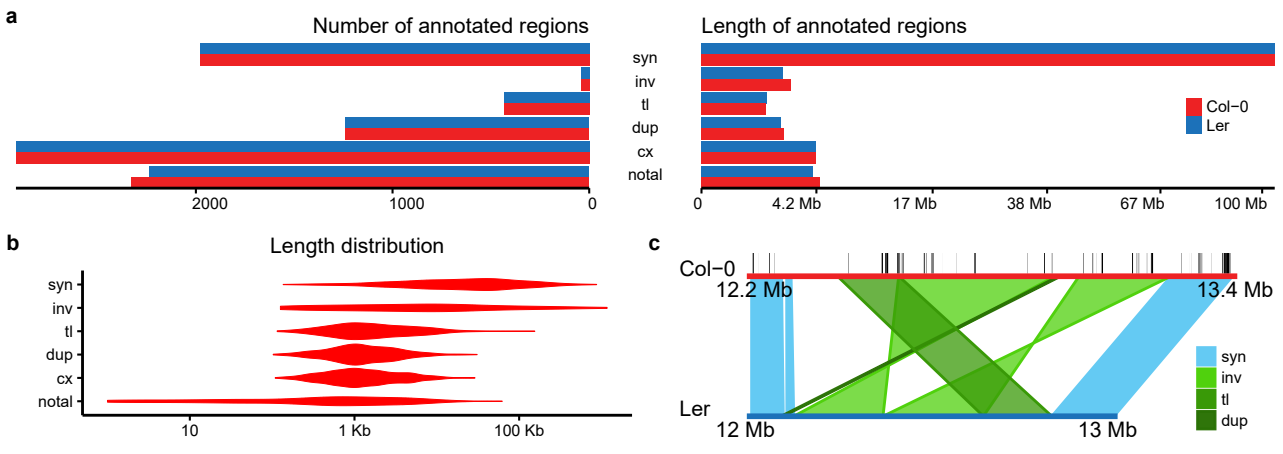
Figure 5: Recombination introduces copy-number variation. (a) Recombination can lead to differences in copy number. (b) This can be observed by aligning short-read sequencing data from recombinant genomes to the reference genome. (c-e) Three different tests to assess the validity of the predicted translocations have been applied. These included (c) testing for the absence of reads in samples with no copy of the translocated DNA, (d) goodness-of-fit between expected copy number and observed copy-number, and (e) clustering of samples with the same genotypes at the translocation. (f) In the heatmap, columns correspond to individual translocations and rows correspond to the three different tests, while the colour of a cell represents whether a translocation was validated (green), was selected but could not be validated (dark grey), or was filtered out as the test was not applicable (grey).

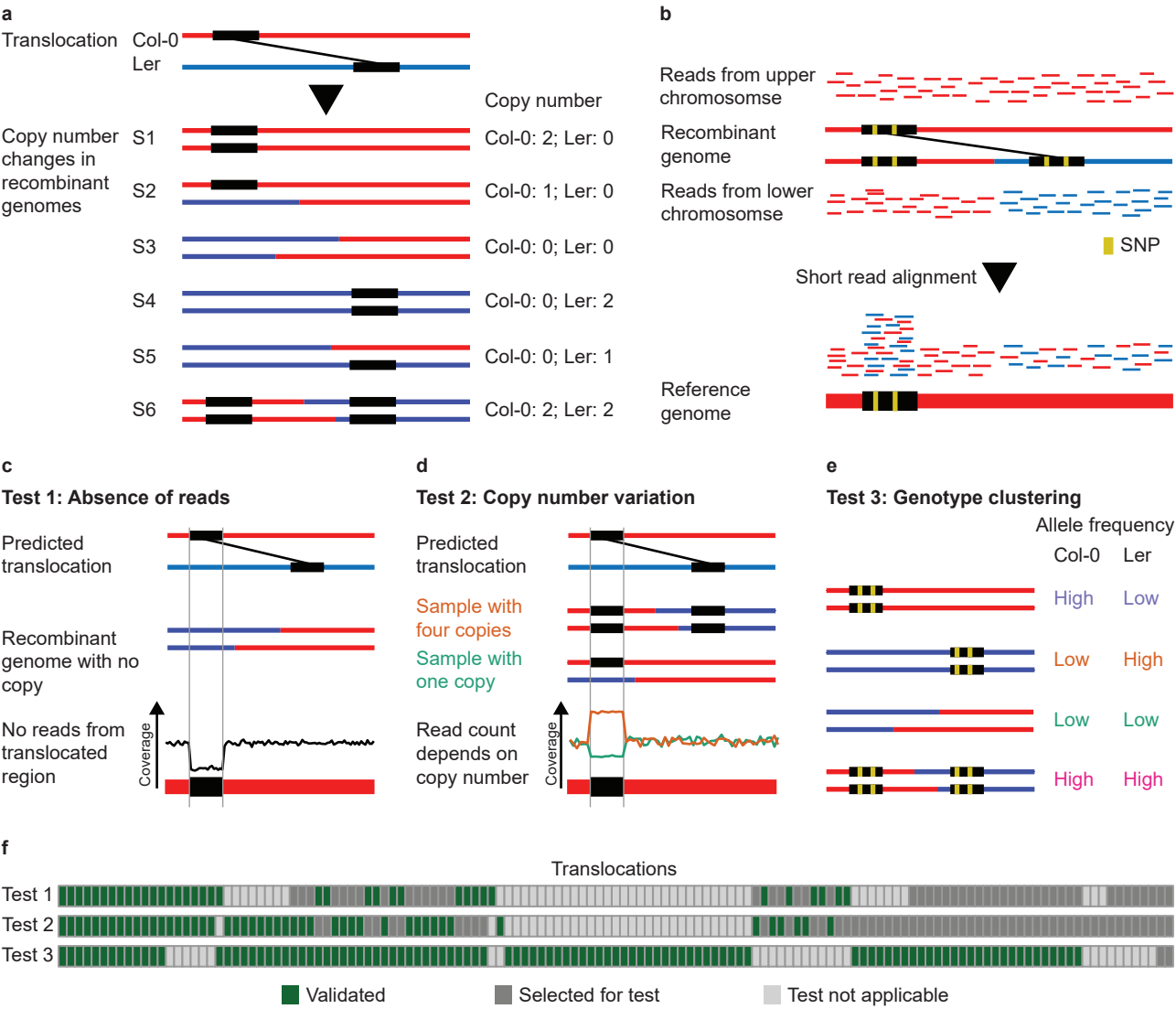
Figure 6: SyRI can identify SRs from incomplete assemblies. Sensitivity (green) and precision (brown) values for SR identification from incomplete assemblies as assessed with simulated data. Each point represents a fragmented assembly. The black line represents the polynomial fit.

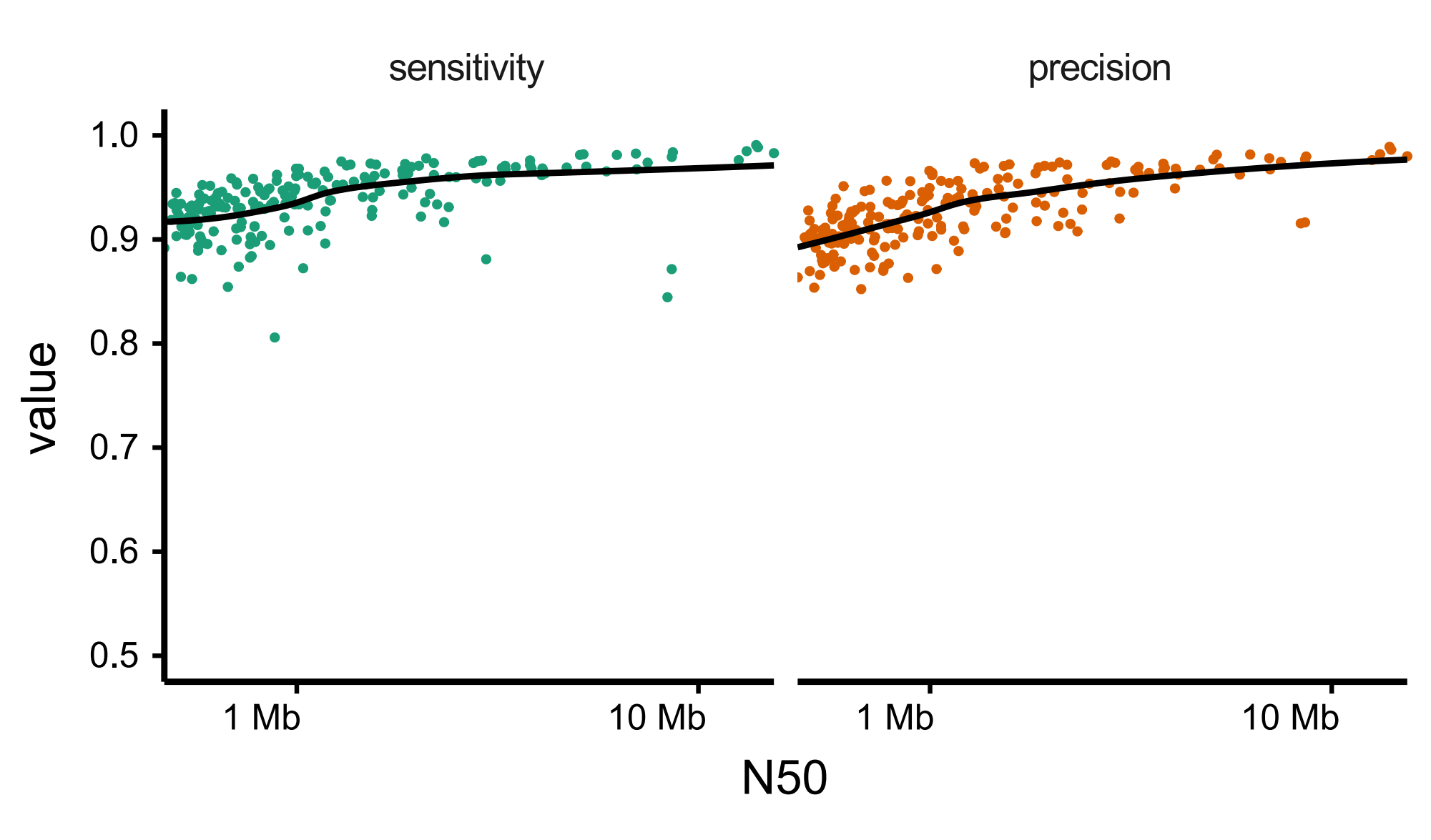

a
Local variations...
Haplotype 1
Haplotype 2
...in a syntenic region
...in a rearranged region
SNPs
Translocation
b
Hap. 1
Hap. 2
Diploid cell
Meiosis: recombination and karyotype reduction
Copy number
Haploid gametes
1
1
0
2


Input:
whole genome alignment
Genome A
Genome B
Step 1:
annotate syntenic regions
Step 2:
annotate structural rearrangements
2a: annotate inversions
2b: annotate translocations, duplications, and remove redundant alignments
2c: annotate inter-chromosomal structural rearrangments
Step 3:
annotate local variations in syntenic and rearranged regions
A
AG
T
SNP
indel
T
ACTG
C
3a: identify SNPs and indels within aligned regions
3b: identify structural variations between aligned regions
3c: not aligned regions are reported separately


a Syntenic path identification
Genomic location
Genome A
Alignments
Genome B
b Inversion identification
i)
ii)
c Translocation and duplication identification
iii)
iv)


a
Number of annotated regions
2000
1000
0
Length of annotated regions
syn
inv
tl
dup
cx
notal
Col−0
Ler
0
4.2 Mb
17 Mb
38 Mb
67 Mb
100 Mb
b
Length distribution
syn
inv
tl
dup
cx
notal
10
1 Kb
100 Kb
c
Col−0
12.2 Mb
13.4 Mb
Ler
12 Mb
13 Mb
syn
inv
tl
dup


a
Translocation
Col-0
Ler
Copy number
Copy number changes in recombinant genomes
S1
Col-0: 2; Ler: 0
S2
Col-0: 1; Ler: 0
S3
Col-0: 0; Ler: 0
S4
Col-0: 0; Ler: 2
S5
Col-0: 0; Ler: 1
S6
Col-0: 2; Ler: 2
b
Reads from upper chromosomse
Recombinant genome
Reads from lower chromosomse
SNP
Short read alignment
Reference genome
c
Test 1: Absence of reads
Predicted translocation
Recombinant genome with no copy
No reads from translocated region
Coverage
d
Test 2: Copy number variation
Predicted translocation
Sample with four copies
Sample with one copy
Read count depends on copy number
Coverage
e
Test 3: Genotype clustering
Allele frequency
Col-0
Ler
High
Low
Low
High
Low
Low
High
High
f
Translocations
Test 1
Test 2
Test 3
Validated
Selected for test
Test not applicable


sensitivity
precision
1.0
0.9
0.8
0.7
0.6
0.5
value
1 Mb
10 Mb
1 Mb
10 Mb
N50