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

Here, we present SyRI, an efficient tool for genome-wide structural rearrangement identification from genome graphs, which are build 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 just have to 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

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

Many of the existing SR prediction methods utilize short or long read alignments against reference sequences. Though, local differences (like SNPs and small indels) can be found 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 of 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 the reference sequence and analyses alignment breakpoints to identify inversions and translocations [9] and Assemblytics, which utilizes contigs with uniquely aligned regions to a reference sequence and uses them to identify various different genomic differences including large indels or differences in local repeats [10].

62 Here, we introduce SyRI (Synteny and Rearrangement Identifier), a method to identify structural
63 genomic between two related genomes (usually from the same species) using genome graphs
64 generated from pair-wise WGA. SyRI starts by identifying all syntenic regions between
65 homologous chromosomes of the two genomes. Since all other regions are SRs by definition (or
66 otherwise they would be part of the syntenic regions), this transforms the problem of SR
67 identification to SR classification. SyRI classifies non-syntenic regions into inversions,
68 translocations and duplications. SyRI performs concurrent analysis of rearranged regions across
69 the entire genomes and annotates SRs globally to optimise the annotation of genomic
70 differences. Even though, it is common to distinguish between translocations and transpositions,
71 we will refer to both types as translocations. Additionally, translocations and duplications will be
72 collectively referred to as TDs.
73 Finally, SyRI identifies *local variation* across the entire genome including rearranged as well as
74 non-rearranged regions. It is important to note that local variations and structural
75 rearrangements are not distinguished by their size or complexity, as local variations include can
76 also include large structural variations like long deletions or insertions. Instead, local variation
77 can be found within syntenic as well as in structurally rearranged regions. This introduce a
78 hierarchy the variation and – for example – allows to distinguish between SNPs in co-linear
79 regions as compared to SNPs in rearranged regions. This distinction is important, as rearranged
80 regions (and the local variation in them) will not follow Mendelian segregation patterns in the
81 offspring of the respective organisms, but can even lead to changes in their copy number (Figure
82 1).

83 We used SyRI to analyse divergent genomes of five model species and genetically validated
84 translocations found in two *A. thaliana* strains by analysing Illumina whole-genome sequencing
85 data of 50 F_2 recombinant genomes.

86 ## Results

87 **Identifying genomic difference using SyRI.** SyRI generates *genome graphs* from pairwise WGAs
88 (here we used MUMmer3 toolbox to perform WGA [11,12]) to identify genomic differences
89 between any two haploid genomes (Additional File 1: Note 1). For an overview of the analysis
90 workflow see Figure 2, and for algorithmic details see Figure 3.
91 Ideally, both of the haploid assemblies are at chromosome-level. If one or both of the assemblies
92 are at scaffold-level, SyRi can generate pseudo-chromosomes using homology between
93 assemblies themselves or homology to a reference sequence (Additional File 1: Note 2, preprint
94 [13]). Even though many of the local alignments between two genomes are redundant as they
95 result from repeats, the genomes do not need to be repeat-masked as SyRI can distinguish
96 between repetitive and original alignments. Once done, SyRI outputs tab-separated and standard
97 variant call format (VCF) files including all information on SR and local variation as well as the
98 clustering of local variation in SRs.
99
100 **Syntenic path identification.** SyRI selects all forward alignments between a pair of homologous
101 chromosomes and generates a genome graph in the form of a directed acyclic graph (DAG)
102 (Figure 3a). For each of the alignments a node is created with a score equal to the alignment
103 score. Any pair of nodes is then connected by an edge if the two underlying alignments are co-

linear. Alignments are defined as colinear if the underlying regions are non-rearranged to each other 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 algorithms for the identification of conserved regions as seeds during whole-genome alignments as implemented in MUMmer [14,15]. 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 calculates an *inversion score* which is defined as the difference between the sum of the node scores in the candidate inversion and the node scores of all syntenic alignments that would need to be removed in case the inversion is selected, since some inversions might cross syntenic

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

147 Low scoring candidates and those that are overlapping with syntenic or inverted regions are
148 filtered out.
149 Like inversions, selected candidate TDs can overlap with each other. To resolve these conflicts,
150 SyRI first clusters overlapping candidate TDs, generating a network of co-aligned regions
151 (Additional File 1: Note 3). These networks are processed using *progressive elimination* which
152 involves iterative selection of candidates which align uniquely to a region (Figure 3c). A unique
153 alignment implies that these candidate TDs are necessary for annotating these genomic regions
154 as there are no alternative alignments which could annotate these regions. In contrast,
155 candidates which align to annotated regions within both of the genomes are removed as they
156 are redundant alignments. This procedure, however, can result in deadlocks, which refers to
157 cases when it is not clear which candidate should be selected next during *progressive elimination*.
158 To break such deadlocks, SyRI uses progressive elimination in conjunction with brute-force (for
159 small networks) and randomized-greedy algorithms (for large networks). In the brute-force
160 approach, all possible sets of non-conflicting candidate TDs within a network are generated. The
161 score of each set (sum of candidate TD scores) is calculated and the set with the highest score is
162 selected. In the randomized-greedy algorithm, deadlocks are resolved by selecting one of the
163 highest scoring candidates at random with its probability weighed to its alignment score followed
164 again by progressive elimination. This is repeated until all candidates are either annotated as TD
165 or removed as redundant. After SyRI identifies all intra-chromosomal TDs, inter-chromosomal
166 TDs are annotated in the same way.
167

168 **Grouping of alignments to generate annotation block.** SyRI combines annotated alignments to
169 generate annotation blocks where a *block* corresponds to genomic region (alignments and the
170 unaligned space within it) which constitute a structural event. For example, a syntenic block
171 would contain all consecutive uninterrupted co-linear alignments and the unaligned regions
172 between these alignments. Inversion (or TD) blocks would include all alignments which together
173 form the extent of one inversion (or one TD).
174
175 **Local variation identification.** Local, small variations (like SNPs and small indels) are parsed out
176 from individual alignments using MUMmer or corresponding CIGAR strings [12]. Local, large
177 variations (structural variations like long indels) lead to alignment discontinuity. SyRI compares
178 the overlaps and gaps present between all consecutive alignments within each of the annotation
179 blocks and annotates indels, highly divergent regions (HDRs) and CNVs/tandem repeats
180 (Additional File 1: Figure S4) similar to the structural variation predictions of Assemblytics [10].
181 Finally, SyRI reports all *un-aligned regions* which are not part of any annotation block. These
182 regions reside between neighboured annotation blocks and are not classified further.

183 Validating SyRI using genetics and a simulation study

184 We used SyRI to predict genomic differences between two of the best described *A. thaliana*
185 accessions, Col-0 and L*er*. Col-0 is the standard lab strain, which was used for the generation of a
186 high-quality reference sequence [16]. The L*er* genome was recently assembled using a
187 combination of different genomics technologies [17]. Both genomes are highly inbred and are
188 considered as fully homozygous. Here, we generated a new chromosome-level *de novo* assembly

189 of the homozygous L*er* genome using PacBio reads and homology to the reference sequence.
190 CN50 and CL50 (which are chromosome-number corrected metrics for the commonly used C50
191 and L50 values) were 12.6Mb and 1, respectively [18]. (Additional File 2: Table S1, Methods).
192 Based on a MUMmer WGA of these two assemblies, SyRI predicted 1,979 syntenic blocks
193 (104.5Mb, 88.5%), 4,630 SR blocks (9.7Mb, 8.2%) as well as 2,237 non-aligned regions (3.9Mb,
194 3.3%) (values are with respect to the L*er* genome; similar values for Col-0; Figure 4a). Syntenic
195 blocks were much larger (median length of ~22 kb) as compared to SR blocks with a median
196 length of 1.2 kb (Figure 4b). SRs were not randomly distributed in the genome, but clustered in
197 specific regions. An example for a highly rearranged region affected the location of several genes,
198 and how SyRI annotates the individual SR in this region can be found in Figure 4c. More than half
199 of the predicted SRs reside within the peri-centromeres, which are known to be structurally
200 diverse, repeat rich and low in gene density [16]. Overall, 4,095 (88.5%) of the SR blocks were at
201 least partially overlapping with transposable elements, while 2,238 (48%) of them were found
202 entirely within transposable elements. A summary of the local variations can be found in
203 Additional File 1: Figure S5.
204 To validate some of the predicted rearrangements we used a genetic approach which was based
205 on the assumption that recombinant genomes should have different copy numbers of
206 translocated DNA as compared to the parental genomes (Figure 1; 5a): The copy number of
207 translocated DNA in a recombinant genome should rely on the genotypes at the two insertion
208 sites of the translocation. For example, translocated DNA should be duplicated if the two loci of
209 translocated DNA are combined into one recombinant chromosome. In contrast, non-SR regions
210 should not vary in copy-number, irrespective of any recombination in the genome.

211 We used available whole-genome sequencing data of a set of 50 F_2 recombinant plants, which
212 were generated by crossing Col-0 and L*er*, followed by self-pollination of the resulting F_1 hybrids
213 (preprint [19]). We aligned the short reads (~5x genome coverage/sample) to the Col-0 reference
214 sequence and used the genotypes at ~500k SNP [17] markers to reconstruct the parental
215 haplotypes using TIGER [20] (Figure 5b). As expected, almost all of the crossing-over sites (99.5%)
216 overlapped with syntenic regions.
217 For each predicted translocation, we estimated a copy number in each sample using this
218 haplotype information and used three different metrics to test for its existence. The first two
219 were based on the assumption that all reads from a translocated region align to the same loci in
220 the reference genome independent of the location of actual allele(s) in the sequenced sample
221 (Figure 5b, [21]). Accordingly, it should be possible to estimate the copy-number of a
222 translocation using changes in read coverage. For the first test, we analysed the absence of reads
223 in translocated regions in recombinant genomes, which were predicted to feature no copy of the
224 translocated region (Figure 5c) (using 0.2x read coverage as a cut-off to distinguish between
225 absence or presence of a translocations). For the second test, we assessed the goodness-of-fit
226 between expected copy-number and observed copy-number for a translocation across all
227 recombinants (as estimated from the normalized read counts in the translocation regions; Figure
228 5d; Methods). Third, we tested for allele count differences, where allele count refers to the
229 number of reads having the Col-0 (or L*er*) alleles at SNP markers within a translocated region and
230 that allow to distinguish the two alleles of the translocation. Depending on the copy number of
231 the different alleles of a translocation, the allele count should also vary. In consequence, samples
232 with the same genotypes at the two loci of a translocation should have similar allele counts.

233 Clustering the samples based on their allele count values should therefore result in a clustering
234 of the samples according to their genotypes at the loci associated with translocations (Figure 5e).
235 In contrast, the allele counts of falsely predicted translocations would be independent of the
236 genotype and as a result, samples with the same genotypes should not cluster together
237 (Methods).
238 We selected 135 translocations (intra- and inter- chromosomal), which were larger than 1kb and
239 were not part of the peri-centromeres for validation. We tested all translocations with at least
240 one of the tests described above, of which 92% (124) could be confirmed by at least one test,
241 while 39% (52/135) were even confirmed by two or three tests (Figure 5f). We manually checked
242 the read alignments in the regions of the eleven translocations that could not be confirmed and
243 found support for the existence of each of the translocations, which however had not been strong
244 enough to be identified by any of the three automated genetic tests.
245 In additiona, we performed a simulation analysis where we randomly introduced non-
246 overlapping SRs in the Col-0 reference sequence (Methods) and compared these modified
247 genomes against the original assembly using SyRI. For this, we simulated 100 genomes each of
248 the four classes: inversions, translocations, duplications, and translocations with duplications.
249 SyRI was able to consistently identify correct SRs and obtained >95% of sensitivity and precision
250 values across all samples (Additional File 1: Figure S6). False results were usually a consequence
251 of SRs in repeat regions where alternate annotations were equally likely.

252 **Comparing human, yeast, fruit fly, and maize genomes**

253 To test and demonstrate SyRI's usefulness and scalability regarding genome size and
254 repetitiveness, we searched for intra-species genomic differences in four different model
255 organisms: human, yeast, fruit fly, and maize (Additional File 2: Table S1). For its application to
256 human genomes, we compared two assemblies, which included the extensively studied sample
257 NA12878 [22] and an adult Yoruban female genome (NA19240, [23]) against the reference
258 genome GRCh38.p12 [24]. For yeast, we compared the *de novo* assembly of strain YJM1447 [25]
259 against the reference genome from strain S288C [26]. For fruit fly (*D. melanogaster*), the *de novo*
260 assembly of strain A4 [27] was compared to the reference genome [28]. For maize, we compared
261 the *de novo* assembly of PH207 [29] against the B73 reference genome [30]. To limit
262 computational requirements, we masked the highly repetitive maize genome while all other
263 genomes were analysed without masking [31].

264 SyRI was used to predict syntenic regions, SRs, and local variations in all genomes (Table 1,
265 Additional File 1: Figure S7–S11). In each comparison, including human, at least 5% of the
266 assembled genomes were found to be structurally different. Runtime and memory usage were
267 dependent on genome complexity (including repetitiveness). The CPU runtime for the smaller
268 and simpler yeast genomes was 43 seconds, whereas for the two human genomes SyRI took ~10
269 minutes, while memory usage was less than 1 GB for each of the comparisons (Table 1) (without
270 considering SNPs and small indel parsing). The exception was the comparison of the repetitive
271 maize genomes, which took ~1hr of CPU time and ~7GB of RAM. Since SyRI considers all
272 alignment combinations, the runtime and memory usage can be high in such repetitive genomes
273 (see Additional File 1: Note 4 and Figure S12 for more information), however, the number of

274 alignments can be drastically reduced by reducing the WGA sensitivity (i.e. omitting small, 10-
275 100s bp alignments), which in turn decreases runtime and memory consumption of SyRI.

276 Discussion

277 We introduced SyRI, a tool that identifies genomic differences using pair-wise comparisons of
278 whole genome assemblies. However, instead of identifying differences directly, SyRI starts by
279 identifying all syntenic (co-linear) regions between the genomes. As all non-syntenic regions must
280 result from rearrangements, identifying syntenic regions implies identification of rearranged
281 regions as well. This transforms the challenge of SR identification to the comparatively easier
282 challenge of SR classification.
283 SyRI is based on a genome graph and uses optimisation methods to identify the syntenic path
284 between the corresponding chromosomes and select alignments (nodes) which represent
285 structural rearrangements between all pairs of chromosomes while maximising the total
286 alignment score. This allows SyRI to annotate even highly complex and repeat rich regions.
287 In addition, SyRI identifies local genomic variations residing either in syntenic or rearranged
288 regions. The identification of local variations in rearranged regions introduces a hierarchy of
289 genomic variations (e.g., SNPs in translocated regions). It is important to distinguish between the
290 different types of SNPs, as SNPs in rearranged regions are differently inherited as compared to
291 SNPs in syntenic regions. The genotypes at rearranged SNPs can confound the interpretation of
292 genomic patterns during selection screens, genome-wide association or recombination analysis
293 as they can show unexpected genotypes [32,33]. SyRI now offers a straight-forward solution to
294 find SNPs in syntenic regions assuming whole-genome assemblies are available.

295 Finally, though implemented in a genome graph that is build up from local alignments generated
296 by a WGA, our algorithm can be easily adapted for SR identification in other types of genome
297 graphs as well [34,35].

298 Conclusions

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

308 Methods

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

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

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

316 **Assembly generation**

317 Raw reads were filtered to remove small and low-quality reads (length<50bp and QV<80),
318 corrected and *de novo* assembled using Falcon [36], followed by polishing with Arrow in the
319 SMRTLink5 package, and finally corrected using Illumina short read alignments with reads from
320 an earlier project [17]. The contigs from organellar DNA sequences were removed, all others
321 were anchored into pseudo-chromosome based on homology with the reference sequence.
322 Adjacent contigs were connected with a stretch of 500 "N" characters.

323 **Whole genome alignments**

324 All assemblies used in this work were filtered to select only chromosome-representing contigs
325 (unplaced scaffolds were removed) (Additional File 2: S2). We used the *nucmer* alignment tool
326 from the MUMmer toolbox [12] to perform WGAs. Nucmer was run with --maxmatch to get all
327 alignments between two genomes including -c, -b, and -l parameters which were selected to
328 balance alignment resolution and runtime based on genome size and number of repeat regions
329 (full commands are available in Additional File 2: Table S2). Alignments were filtered using the
330 *delta-filter* tool and the filtered delta files were converted to the tab-delimited files using the
331 *show-coords* command. Before whole-genome alignments, both maize genomes were masked
332 using RepeatMasker v4.0.6 [37].

333 **TD candidates score**

334 Each selected candidate TD was given a score based on the length of alignments in it and the gap

335 length between the alignments; TD score $= \min\left(\frac{genA_aligned_length - genA_gap_length}{genA_aligned_length},\right.$

336 $\left.\frac{genB_aligned_length - genB_gap_length}{genB_aligned_length}\right)$.

337 **Data extraction and transformation of the 50 recombinant genomes**

338 We used whole-genome sequencing data of 50 F_2 recombinant plants that were generated in the

339 course of a different project (preprint [19]) for validating SR predictions. We extracted allele

340 count information from consensus call files generated by SHORE [38]. For each predicted

341 translocation, we estimated its copy number as the ratio between average read-coverage for the

342 translocated region and the average read-coverage across the entire genome of the respective

343 sample. Translocations in the centromeric regions and for which more than 25% of the

344 translocated sequence had at least 10% reads with Ns were filtered out. For allele count analysis,

345 we selected high-confidence (25bp conserved in both directions) SNPs in translocated regions as

346 markers. Translocated regions (and corresponding SNP markers) were classified according to the

347 genotypes (as predicted by TIGER) in individual samples at the two associated loci.

348 **Validation of translocation: Absence of reads (Test 1)**

349 We selected F2 samples which, according to predicted genotypes, should have lost the

350 translocated DNA and thus should not give rise to any reads from the translocated region. Only

351 translocations for which at least two samples that had lost the translocated regions were tested.

352 And only those translocations for which all tested samples had no reads were considered

353 validated.

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

355 For each translocation, we selected samples which had different genotypes at the two associated
356 loci for the translocation. This removes some of the samples with two copies, and helps to
357 remove a bias towards genomes with a copy number of two, which can affect this test. We further
358 only selected translocations for which we found samples with at least three different copy-
359 number values predicted. A linear model was fit using the *lm* function in *R*. *P*-values for the
360 model-fit were adjusted for multiple testing using the *BH*-method [39], and translocations for
361 which adjusted *p*-values were less than 10^{-6} and slope more than 0.75 were considered as valid.

362 **Validation of translocation: Genotype clustering (Test 3)**

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

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

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

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

380 **Simulation analysis**

381 We simulated SRs in the *A. thaliana* reference genome using *R* package *RSVSim* [40]. We
382 simulated 40, 436, 1241, and 836 events for inversions, translocations, duplications, and
383 translocations with duplications respectively. The number of rearrangements and their
384 corresponding sizes were sampled after real differences found between the Col-0 vs L*er*
385 genomes. The simulated genomes were aligned back to the unmodified reference genome and
386 SRs were identified using SyRI. A predicted SR was considered to match an original SR if the end
387 points for the predicted SR were within ±150bp of the end points of the original SR.

388 Declarations

389 **Availability of data and materials**

390 The assembly of the L*er* genome has been submitted to the European Nucleotide Archive
391 (http://www.ebi.ac.uk) and will be made publicly available upon publication under the accession
392 number GCA_900660825. All other assemblies are publicly available at NCBI
393 (https://www.ncbi.nlm.nih.gov/), and their accession numbers are GCA_000001735.3 [16],
394 GCA_000001405.27 [24], GCA_002077035.3 [22], GCA_001524155.4 [23], GCA_000146045.2 [26],
395 GCA_000977955.2 [25], GCA_000001215.4 [28], GCA_002300595.1 [27], GCA_000005005.6 [30],
396 GCA_002237485.1 [29]. Further details about the assemblies are in Additional File 2: Table S1. BAM
397 files for the 50 F_2 recombinant genomes are available at European Nucleotide Archive under the

project ID PRJEB29265 (preprint [19]). SyRI is freely available under MIT license and is available online [41]. 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

Project was conceived by KS and WBJ. MG and KS developed the algorithms and methods. MG implemented SyRI and performed analysis. HS processed recombinant genome sequencing data and identified crossing-over sites. WBJ generated the L*er* assembly. 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 for his helpful comments on the manuscript. We would also like to thank the researchers at the 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.

418 2. Tuttle EM, Bergland AO, Korody ML, Brewer MS, Newhouse DJ, Minx P, et al. Divergence
419 and Functional Degradation of a Sex Chromosome-like Supergene. Curr Biol. 2016;
420 26(3):344–50.
421 3. Küpper C, Stocks M, Risse JE, dos Remedios N, Farrell LL, McRae SB, et al. A supergene
422 determines highly divergent male reproductive morphs in the ruff. Nat Genet. 2016;
423 48(1):79–83.
424 4. Lamichhaney S, Fan G, Widemo F, Gunnarsson U, Thalmann DS, Hoeppner MP, et al.
425 Structural genomic changes underlie alternative reproductive strategies in the ruff
426 (Philomachus pugnax). Nat Genet. 2016; 48(1):84–8.
427 5. Lowry DB, Willis JH. A Widespread Chromosomal Inversion Polymorphism Contributes to
428 a Major Life-History Transition, Local Adaptation, and Reproductive Isolation. Barton NH,
429 editor. PLoS Biol. 2010; 8(9):e1000500.
430 6. Simpson JT, Pop M. The Theory and Practice of Genome Sequence Assembly. Annu Rev
431 Genomics Hum Genet. 2015; 16(1):153–72.
432 7. Jiao W-B. The impact of third generation genomic technologies on plant genome assembly.
433 Curr Opin Plant Biol. 2017; 36:64–70.
434 8. Sedlazeck FJ, Lee H, Darby CA, Schatz MC. Piercing the dark matter: Bioinformatics of long-
435 range sequencing and mapping. Nat Rev Genet. 2018; 19(6):329–46.
436 9. Liu S, Huang S, Rao J, Ye W, Krogh A, Wang J. Discovery, genotyping and characterization
437 of structural variation and novel sequence at single nucleotide resolution from de novo
438 genome assemblies on a population scale. Gigascience. 2015; 4(1).
439 10. Nattestad M, Schatz MC. Assemblytics: a web analytics tool for the detection of variants

440 from an assembly. Bioinformatics. 2016; 32(19):3021–3.

441 11. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: A fast and
442 versatile genome alignment system. Darling AE, editor. PLOS Comput Biol. 2018;
443 14(1):e1005944.

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

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

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

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

452 16. Initiative TAG. Analysis of the genome sequence of the flowering plant Arabidopsis
453 thaliana. Nature. 2000; 408(6814):796–815.

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

457 18. Jiao W-B, Accinelli GG, Hartwig B, Kiefer C, Baker D, Severing E, et al. Improving and
458 correcting the contiguity of long-read genome assemblies of three plant species using
459 optical mapping and chromosome conformation capture data. Genome Res. 2017;
460 27(5):778–86.

461 19. Sun H, Rowan BA, Flood PJ, Brandt R, Fuss J, Hancock AM, et al. Linked-read sequencing of

462 gametes allows efficient genome-wide analysis of meiotic recombination. bioRxiv. 2018;
463 :484022.
464 20. Rowan BA, Patel V, Weigel D, Schneeberger K. Rapid and Inexpensive Whole-Genome
465 Genotyping-by-Sequencing for Crossover Localization and Fine-Scale Genetic Mapping. G3
466 Genes, Genomes, Genet. 2015; 5(3):385–98.
467 21. Imprialou M, Kahles A, Steffen JG, Osborne EJ, Gan X, Lempe J, et al. Genomic
468 Rearrangements inArabidopsisConsidered as Quantitative Traits. Genetics. 2017;
469 205(4):1425–41.
470 22. The Genome Institute at Washington University School of Medicine. NA12878_prelim_3.0.
471 https://www.ncbi.nlm.nih.gov/assembly/GCA_00207703. 2018.
472 23. The Genome Institute at Washington University School of Medicine. NA19240_prelim_3.0.
473 https://www.ncbi.nlm.nih.gov/assembly/GCA_001524155.4. 2017.
474 24. Human Genome Sequencing Consortium I. Finishing the euchromatic sequence of the
475 human genome. Nature. 2004; 431(7011):931–45.
476 25. Strope PK, Skelly DA, Kozmin SG, Mahadevan G, Stone EA, Magwene PM, et al. The 100-
477 genomes strains, an S. cerevisiae resource that illuminates its natural phenotypic and
478 genotypic variation and emergence as an opportunistic pathogen. Genome Res. 2015;
479 25(5):762–74.
480 26. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, et al. Life with 6000
481 Genes. Science (80-). 1996; 274(5287):546–67.
482 27. Chakraborty M, VanKuren NW, Zhao R, Zhang X, Kalsow S, Emerson JJ. Hidden genetic
483 variation shapes the structure of functional elements in Drosophila. Nat Genet. 2018;

484 50(1):20–5.

485 28. Hoskins RA, Carlson JW, Wan KH, Park S, Mendez I, Galle SE, et al. The Release 6 reference
486 sequence of the *Drosophila melanogaster* genome. Genome Res. 2015; 25(3):445–58.

487 29. Hirsch CN, Hirsch CD, Brohammer AB, Bowman MJ, Soifer I, Barad O, et al. Draft Assembly
488 of Elite Inbred Line PH207 Provides Insights into Genomic and Transcriptome Diversity in
489 Maize. Plant Cell. 2016; 28(11):2700–14.

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

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

494 32. Wijnker E, Velikkakam James G, Ding J, Becker F, Klasen JR, Rawat V, et al. The genomic
495 landscape of meiotic crossovers and gene conversions in *Arabidopsis thaliana*. Elife. 2013;
496 2.

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

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

502 35. Consortium TCP. Computational pan-genomics: status, promises and challenges. Brief
503 Bioinform. 2016; (August):bbw089.

504 36. Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al. Phased diploid
505 genome assembly with single-molecule real-time sequencing. Nat Methods. 2016;

506 13(12):1050–4.

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

508 38. Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, Weigel D. Sequencing of
509 natural strains of Arabidopsis thaliana with short reads. Genome Res. 2008; 18(12):2024–
510 33.

511 39. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful
512 approach to multiple testing. J R Stat Soc Ser B. 1995; 57(1):289–300.

513 40. Bartenhagen C, Dugas M. RSVSim: an R/Bioconductor package for the simulation of
514 structural variations. Bioinformatics. 2013; 29(13):1679–81.

515 41. Goel M, Schneeberger K. SyRI: identification of syntenic and rearranged regions from
516 whole-genome assemblies. https://schneebergerlab.github.io/syri/. 2019; .

517

List of Additional Files:

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

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

521

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

522 ## Figure legends

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

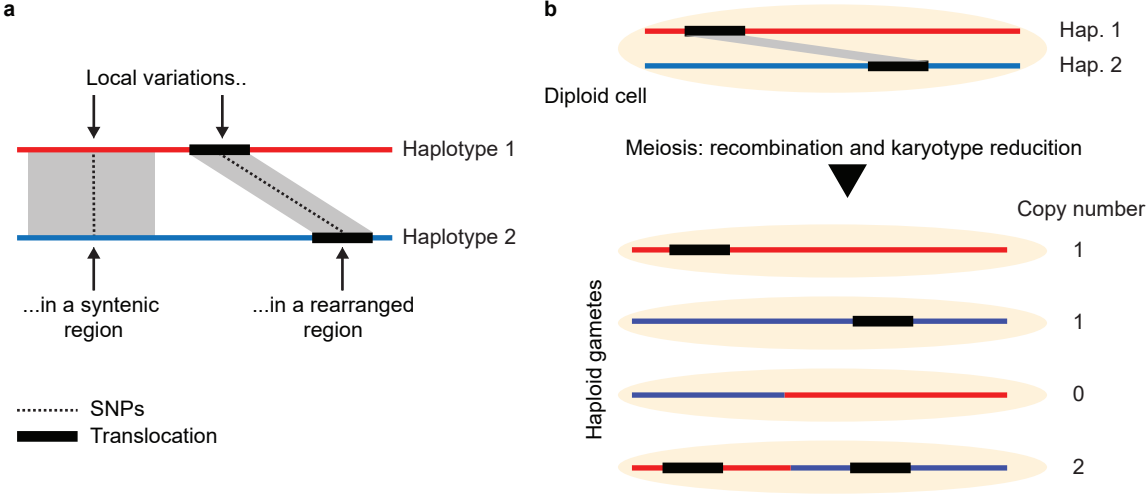
527

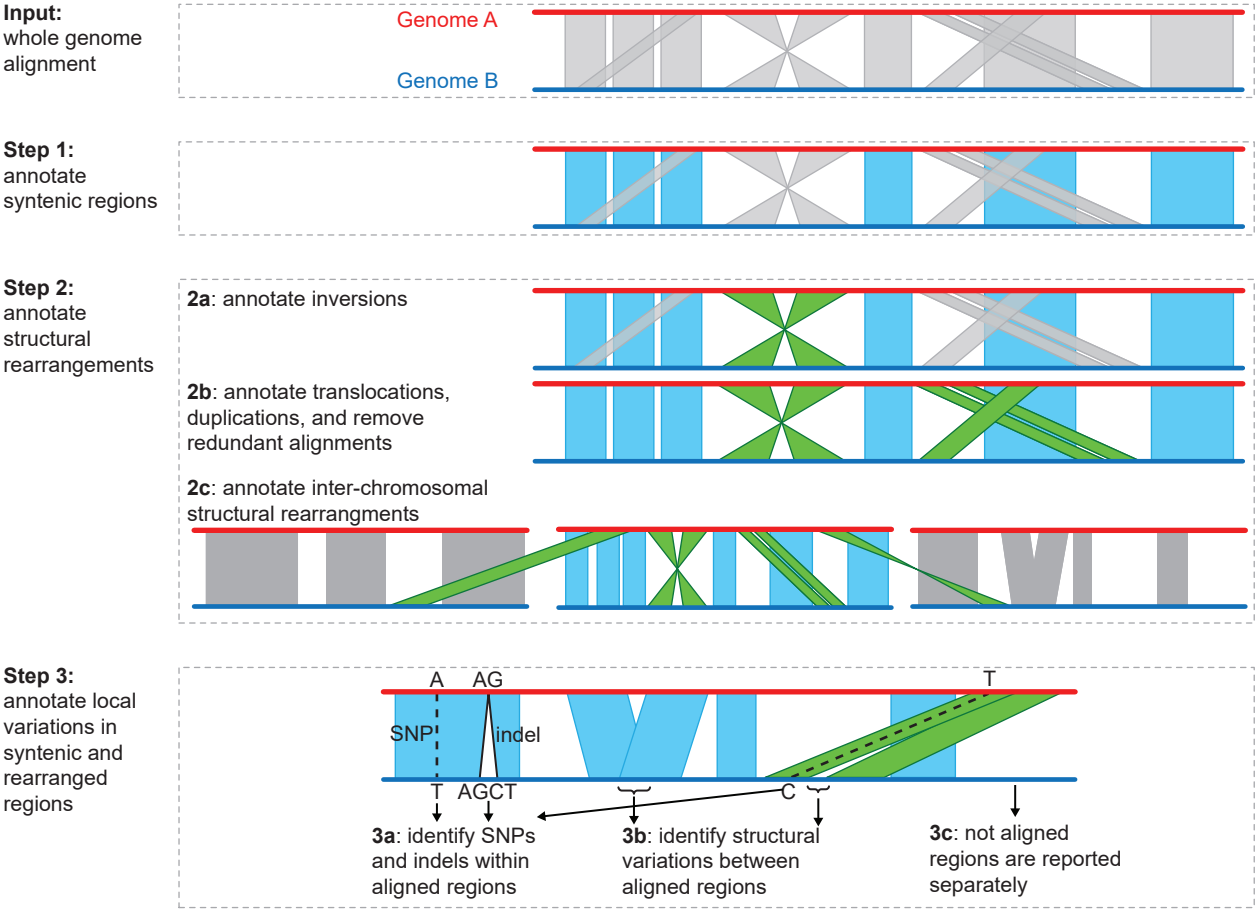
528 **Figure 2: Workflow for the identification of genomic differences.** SyRI takes WGA as input. A
529 WGA consists of local alignments, where each local alignment (grey polygon) connects a specific
530 region in one genome to a specific region in the other genome. Step 1: SyRI identifies the highest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

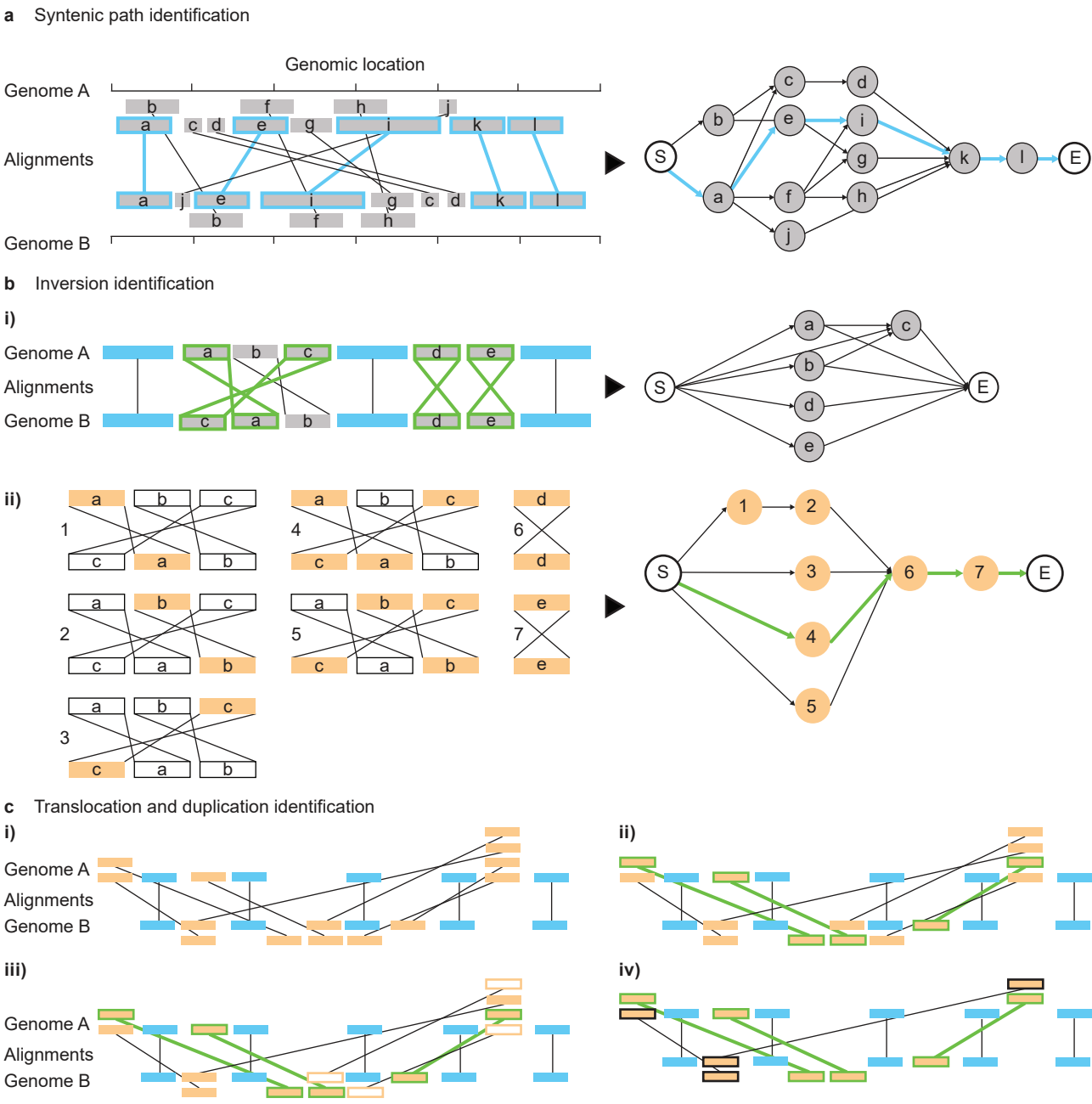
541

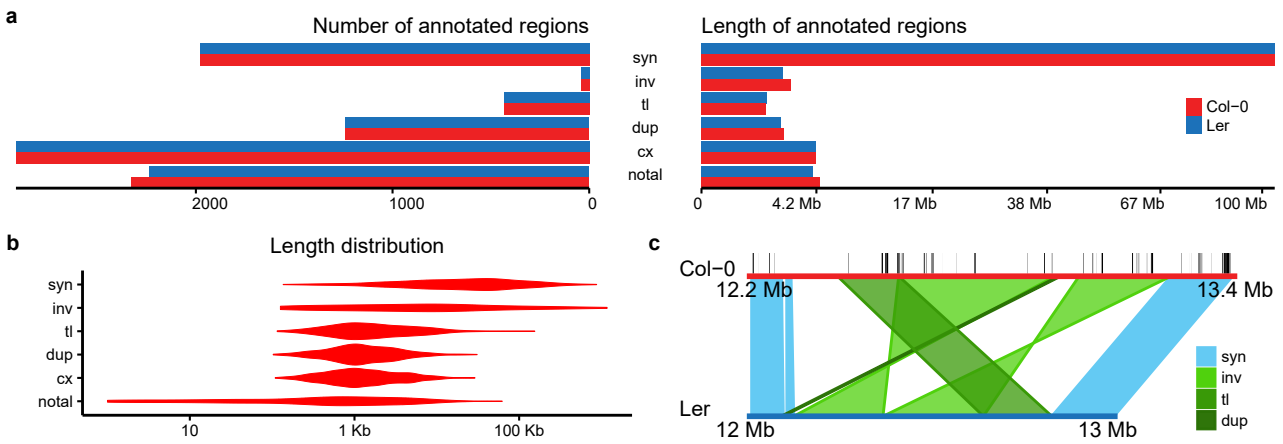
Figure 3: Overview of the different algorithms implemented in SyRI. (a) Syntenic region identification. A WGA between genomes A and B is converted to a genome graph where nodes represent local alignments and edges are introduced if the corresponding alignments are co-linear (like alignments 'a' and 'e'). 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 are 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.

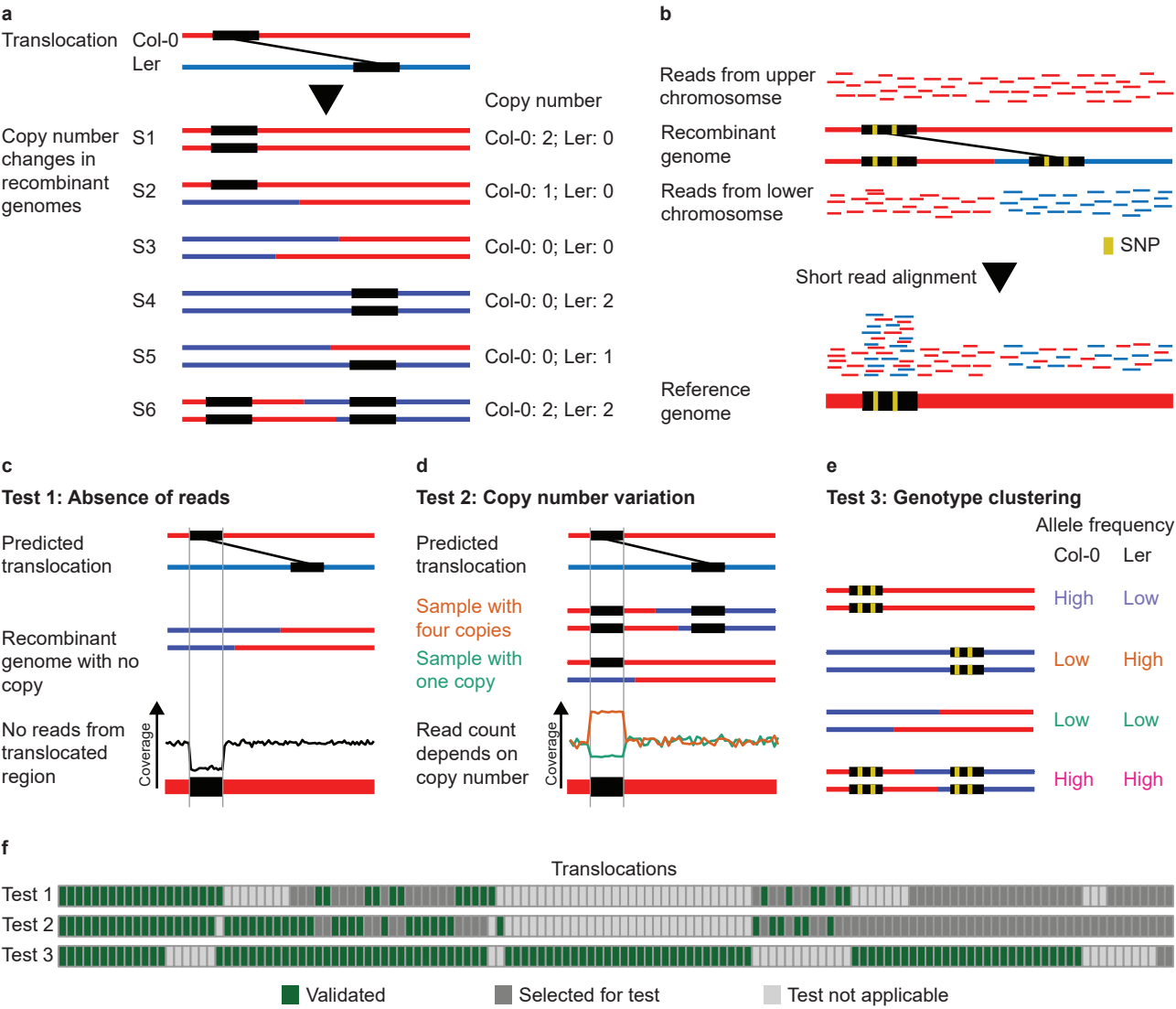
564 **Figure 4: Genomic differences between Col-0 and L*er* accessions of *A. thaliana.*** (a) Total length
565 and number of syntenic, SR, and un-aligned regions identified between the Col-0 and L*er*
566 genomes (syn: syntenic regions, inv: inversions, tl: translocations, dup: duplications, cx: cross-
567 chromosomal exchange, notal: un-aligned). (b) Length distribution of all annotation types. (c)
568 Example of a highly rearranged region at chromosome 5. Two large inversions, a translocation,
569 and a duplication occur together. Multiple genes (black lines) are present in this region. Smaller
570 SRs in this region have been filtered for cleaner representation.
571
572 **Figure 5: Recombination can cause copy-number variation of translocated regions.** (a)
573 Recombination can lead to differences in copy number. (b) This can be observed by aligning short-
574 read sequencing data from recombinant genomes between Col-0 and L*er* to the reference
575 genome. (c-e) Three different tests to assess the validity of the predicted translocations have
576 been applied. These included (c) testing for the absence of reads in samples with no copy of the
577 translocated DNA, (d) goodness-of-fit between expected copy number and observed copy-
578 number, and (e) clustering of samples with the same genotypes at the translocation. (f) In the
579 heatmap, columns correspond to individual translocations and rows correspond to the three
580 different tests, while the color of a cell represents whether a translocation was validated (green),
581 was selected but could not be validated (dark grey), or was filtered out as the test was not
582 applicable (grey).
583


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 reducition
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
AGCT
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
i)
ii)
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