SyRI: finding genomic rearrangements and local sequence differences 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

Genomic differences range from single nucleotide differences to complex structural variation. Current methods typically annotate only differences in sequence (ranging from SNPs to large indels) but do not consider the full complexity of structural rearrangements (including inversions, translocations and duplications), where highly similar sequence changes in location, orientation, or copy-number. Here we present SyRI, a pairwise whole-genome comparison tool for chromosome-level assemblies. SyRI starts by finding rearranged regions and then searches for differences in the sequences, which are distinguished for residing in syntenic or rearranged regions. This distinction is important, as rearranged regions are inherited differently compared to syntenic regions.

Keywords

Genome comparison, structural rearrangements, structural variations, variant calling, genome alignments, genetics, genome assembly

Background

Genomic differences form the basis for heritable phenotypic variations and their identification can allow us to decipher evolutionary past and gene function. Differences in genomes can range from single nucleotides to highly complex genomic rearrangements and are usually described as local sequence differences in comparison to a reference sequence. But even though the annotation of all sequence differences would be sufficient to reconstruct the actual sequence of a genome, local annotation of differences does not describe the full complex dependencies between the rearrangements. For example, a translocation is a genomic rearrangement where a specific sequence has moved from one region in the genome to another region. Although such a translocation could be described as a deletion at one region in the genome and an insertion at the other regions, this annotation would miss out on the fact that the deleted/inserted sequence is the same and that the sequence is not actually deleted but moved to a different region of the other genome. Like translocations, inversions and duplications also introduce such "differences in the structure" of a genome by introducing differences in location, orientation and/or copy number of the same sequence in the genome. But even though this information is usually not considered when analysing whole-genome sequencing data, such "differences in the structure" are relevant as they can be the basis for diseases phenotypes [1], reproductive strategies [2–4], and survival strategies [5].

Many of the state-of-the-art methods used to predict genomic differences utilize short or long read alignments against reference sequences [6]. Even though such alignments allow to find local sequence differences (like SNPs, indels and structural variation) with high accuracy, accurate prediction of differences in the structure remains challenging using read alignments

alone. In contrast, whole-genome assemblies enable the identification of complex rearrangements as the assembled contigs are typically much longer and of higher quality as compared to raw sequence reads [7]. However, despite recent technological improvements to simplify the generation of whole-genome *de-novo* assemblies [8], there are so-far only a few tools which use whole-genome assemblies as the basis for the identification of genomic differences [9]. Available tools include AsmVar, which compares individual contigs of an assembly against a reference sequence and analyses alignment breakpoints to identify inversions and translocations [10], 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 [11] and Smartie-sv, which compares individual alignments between assembly and reference sequences [12].

Here, we introduce SyRI (Synteny and Rearrangement Identifier), a method to identify differences in structures as well as differences in the sequences (including small and complex local changes) of two whole-genome assemblies. SyRI expects whole-genome alignments (WGA) as input and starts by searching for differences in the structures of the genomes. Afterwards, SyRI identifies local sequence differences within both the rearranged and the non-rearranged (syntenic) regions. The distinction of local sequence differences in syntenic regions versus local sequence differences in rearranged regions is important as rearranged regions (and the variation in them) can show non-mendelian segregation patterns in offspring genomes.

Finally, we validate SyRI's performance with simulations and in comparison with existing tools used for the identification of genomic differences and apply SyRI to divergent genomes of five

72 model species, including two *Arabidopsis thaliana* strains, for which we experimentally validate
73 over 100 predicted translocations.

74 Results

75 The hierarchy in genomic differences

76 Genomes can differ in structure and sequence. *Differences in structure* occur if highly similar
77 regions have different locations/orientations in different genomes (here we will refer to these
78 regions as rearranged or non-syntenic regions). *Differences in sequence* are (simple to complex)
79 differences in the nucleotide sequence within otherwise conserved regions. Importantly,
80 differences in sequence can occur in both, rearranged as well as non-rearranged regions (Figure
81 1a). This introduces a hierarchy into the variation in genomes where some sequence
82 differences reside in the conserved (syntenic) regions whereas other reside in rearranged
83 regions of the genomes (e.g. a SNP within a translocated region).
84 Even though reference-based resequencing analyses typically do not allow to distinguish
85 between variation in syntenic versus non-syntenic regions, this distinction is important as non-
86 syntenic regions (and the local variation in them) do not follow Mendelian segregation patterns
87 in the offspring. Instead, due to the different locations in a genome, the inheritance of
88 rearrangements can lead to changes in copy number or loss of the rearranged regions (Figure
89 1b).

90 How SyRI works

91 SyRI is a whole-genome comparison tool that annotates all differences (including differences in
92 structure and sequence) between two whole-genome assemblies (Figure 2). It starts by
93 identifying all syntenic regions between the two genomes (Figure 2: Step 1). This at the same
94 time also reveals all rearranged regions, as all regions which are not annotated as syntenic are
95 rearranged by definition. In a second step, SyRI analyzes the location and orientation of the
96 rearranged regions and groups them into inversions, translocations, and duplications (Figure 2:
97 Step 2). As the last step, SyRI identifies sequence differences (like SNPs, indels, tandem
98 duplications, copy-number changes etc) within both rearranged and syntenic regions (Figure 2:
99 Step 3).
100 To perform these three steps, SyRI generates different *genome graphs* which are generated
101 from the local alignments from a pairwise whole-genome alignment (WGA). Here we used the
102 MUMmer3 toolbox to perform WGA [13,14], but other alignments tools like minimap2 [15] can
103 be used as well (Additional File 1: Note 1). In the following, we describe the individual steps of
104 SyRI in more detail.
105 **Syntenic region identification.** SyRI identifies syntenic regions by selecting the longest, non-
106 contradicting subset of aligned regions which are syntenic to each other in the two genomes
107 (Fig. 2: Step 1). For this, it selects all forward alignments between a pair of homologous
108 chromosomes and generates a genome graph in the form of a directed acyclic graph (DAG)
109 (Additional File 1: Note 2, Figure S1). SyRI then uses dynamic programming to identify the
110 highest scoring path from the nodes that represent one end of a chromosome to the nodes that
111 represent the other end (using similar algorithms as implemented in MUMmer [16,17]). This
112 process is repeated for each pair of homologous chromosomes and after all syntenic regions

113 are found, SyRI starts classifying the remaining non-syntenic regions into inversions,
114 translocations, and duplications (Figure 2: Step 2).

115 **Inversion identification.** An inversion is defined as a set of inverted alignments in-between two
116 syntenic alignments (Additional File 1: Figure S2). Hence, if one genome would be reverse
117 complemented then the alignments of an inversion would appear similar to syntenic
118 alignments. Following this idea, SyRI selects all inverted alignments between a pair of
119 corresponding chromosomes and reverse complements one of the chromosomes (Additional
120 File 1: Figure S3, Note 3). Then, analogous to the syntenic path identification, SyRI again builds
121 up a genome graph using these new forward alignments. From this graph, SyRI infers all
122 possible candidate inversions between the two genomes (Additional File 1: Figure S3a).
123 However, as candidate inversions can overlap and result in conflicting annotations (Additional
124 File 1: Figure S3b), SyRI compares the annotations of all candidate inversions simultaneously
125 and selects the best set of non-overlapping non-conflicting inversions while maximising the
126 overall alignment score of the two genomes.

127 **Translocation and duplication (TD) identification.** After synteny and inversion identification, all
128 remaining alignments are either footprints of TDs or are redundant (repetitive) alignments. SyRI
129 analyzes these alignments to find TDs while removing redundant alignments (Additional File 1:
130 Note 4, Figure S4-S6). For this, SyRI first groups the alignments such that each group represents
131 all alignments of a putatively rearranged region (candidate TD). Each candidate TD is given a
132 score based on its alignment length and gap length between consecutive alignments. Low
133 scoring candidates and those that are overlapping with syntenic or inverted regions are filtered
134 out.

135 As a result of repeats, rearranged regions can have different candidate TDs aligning to different
136 copies of the same repeat region. Therefore, overlapping candidate TDs often result in
137 conflicting annotations. SyRI resolves these overlapping candidate TDs by selecting the non-
138 conflicting subset of candidate TDs with the highest alignment score (Additional File 1: Note 4,
139 Figure S5, S7).

140 **Grouping of alignments to generate annotation blocks.** As a final step of the analysis of
141 differences in structure, SyRI combines all neighbouring alignments of the same type to form
142 annotation blocks. For example, a syntenic block would contain all consecutive uninterrupted
143 syntenic alignments. Likewise, inversion or TD blocks include all alignments which together
144 form the extent of an inversion or of a TD.

145 **Identification of sequence differences.** Finally, SyRI reports differences in the sequence. This
146 includes small variation (like SNPs and small indels) which are found in the local alignments
147 generated by the whole-genome alignment algorithm as well as larger structural variations (like
148 indels or CNVs), which are not part of the local alignments (Figure 2: Step 3). To find these
149 structural variations, SyRI analyses the gaps between all consecutive alignments in annotation
150 blocks and identifies indels, highly divergent regions (HDRs) and CNVs/tandem repeats
151 (Additional File 1: Figure S8) similar to the SV identification of Assemblytics [11]. Finally, SyRI
152 reports all *un-aligned regions* which are not part of any annotation block.

153 Performance evaluation using simulated genomes

154 We validated SyRI using a simulation study with 400 simulated genomes which were generated
155 by randomly inserting inversions, translocations, duplications, and indels into the reference

156 sequence of *A. thaliana* (Methods). Comparing these genomes against the unaltered reference
157 sequence, SyRI showed sensitivity and precision values of more than 95% for each of the
158 different types of rearrangements (Figure 3a, b). False results were usually a consequence of
159 rearranged regions in repeat regions where alternative annotations were equally likely.
160 We compared this to the performance of six other tools: AsmVar, smartie-sv, assemblytics,
161 sniffles, picky, LUMPY [10–12,18–20]. AsmVar, smartie-sv, and assemblytics identify genomic
162 differences using whole-genome assemblies, sniffles and picky are based on long read
163 alignments, and LUMPY uses short read alignments against a reference sequence. For the tools
164 that require sequencing read data, we simulated reads from the whole genome assemblies and
165 aligned them to the reference sequence (Methods). However, unlike SyRI, these tools are not
166 designed to find all genomic differences (Additional File 2: Table S1). Generally, for variant
167 callers "translocation" identification is limited to identification of chromosome-end
168 rearrangements whereas "duplication" identification is limited to identification of tandem
169 duplications, in contrast, SyRI identifies all genome-wide rearrangements.
170 In our simulations, SyRI outperformed all other tools by consistently identifying all genomic
171 differences (Figure 3a, b). All tools identified indels with high sensitivity and precision, however,
172 only SyRI, AsmVar and Assemblytics identified indel breakpoints with high accuracy. For
173 inversions, only LUMPY and sniffles performed comparable to SyRI, whereas, for translocations,
174 SyRI and AsmVar showed the best performances.

175 **Performance evaluation using real genomes**

176 To test SyRI's performance with real data, we applied it to identify the variation in the human
177 genome NA19240 for which gold standard variation data was recently published (Additional
178 File 1: Figure S9, Methods) [21]. These gold standard variation data include differences which
179 were predicted based on whole-genome shotgun read alignments against the reference
180 sequence and therefore, include variation from both haplotypes of this genome. In addition, a
181 whole-genome assembly, which includes only one of the two haplotypes, is available as well
182 [22].
183 Using this whole genome assembly in comparison to the reference sequence, SyRI identified
184 55.2% (9685 out of 17545) of the gold standard insertions, 54.5% (9494 out of 17391) of the
185 deletions, and 49.7% (81 out of 163) of the inversions (Additional File 1: Figure S10, Methods),
186 which is consistent with the presence of only one of the haplotypes in the assembly. In
187 comparison to the other tools tested here, SyRI identified a higher proportion of all different
188 types of genomic variation of the gold standard variation (Additional File 1: Figure S10).
189 For a second comparison, we generated a chromosome-level assembly of the (homozygous) *A.*
190 *thaliana* L*er* genome using long PacBio reads. The assembly CN50 and CL50 values
191 (chromosome number normalized N50 and L50 values) were 12.6 Mb and 1 respectively
192 (Additional File 2: Table S2, Methods, Additional File 1: Figure S11) [23].
193 We again applied the other tools to identify differences between the Col-0 and L*er* genomes
194 (Additional File 1: Figure S12, Methods). SyRI and AsmVar found similar numbers of indels,
195 while Assemblytics found a much higher number of indels. However, all of these assembly-
196 based methods identified more indels than the read based methods. More inversions were

197 identified in the read based methods compared to SyRI and AsmVar. A similar number of
198 translocations were identified by SyRI and AsmVar, whereas only SyRI identified duplications.
199 The performance of the read based methods was limited by severe overcalling of genomic
200 differences. For every read based method, the cumulative length of deletions was more than
201 150Mbp, with multiple large deletions (>1,000,000, Additional File 2: Table S3) while the
202 genome size of A. *thaliana* is ~135Mbp [24]. Moreover, two of the tools annotated more than
203 70Mbp and 14Mbp of inversions, respectively. These results are highly improbable and most
204 likely resulted from the incorrect classification of structural rearrangements.

205 **Effect of genome contiguity**

206 SyRI requires whole-genome alignments from chromosome-level assemblies as input. If one or
207 both of the assemblies is/are incomplete (i.e. not on chromosome level), pseudo-chromosomes
208 can be generated using homology between the assemblies themselves or using homology to a
209 chromosome-level reference sequence, which can be performed with tools like RaGOO
210 (Additional File 1: Note 5, preprint [25]). To analyze the effect of the contiguity of the original
211 assembly on SyRI's performance, we performed a simulation analysis where we first generated
212 multiple incomplete assemblies from the chromosome-level assembly of *A. thaliana* L*er* by
213 randomly breaking the chromosome-level scaffolds in unconnected pieces (Methods). These
214 scattered assemblies were then reassembled with RaGOO using their homology to the *A.*
215 *thaliana* Col-0 reference genome (preprint [25]).
216 We then identified rearranged regions in each of these re-assemblies by comparing them to the
217 reference sequence using SyRI. These resulting rearrangements were then compared with the

218 results SyRI generated when comparing the original chromosome-level assembly of L*er* against
219 the reference sequence.
220 More than 90% of the assemblies with N50 of more than 470 Kb (before the homology based
221 reassembly) had a sensitivity of more than 0.9 (Figure 3c). Similarly, more than 90% of the
222 assemblies with N50 more than 674Kb had a precision of more than 0.9. The shortest
223 assemblies we generated had N50 values in the range of 470-500 Kb and the predictions based
224 on these assemblies still had average sensitivity and precision values of 0.92 and 0.90
225 respectively.
226 We then evaluated SyRI's efficiency in identifying rearranged regions when both genomes are
227 at scaffold level. For this, we generated scattered assemblies from both the Col-0 reference
228 sequence and the L*er* assembly. Since current pseudo-chromosome generation tools only
229 concatenate scaffolds of one assembly using homology with another assembly, we developed a
230 heuristic script to generate homology-based pseudo-chromosomes using two incomplete
231 assemblies (Additional File 1: Note 5). As before, we identified rearranged regions from these
232 pseudo-genomes and compared them to the rearranged regions identified between the full-
233 length assemblies. For assemblies with N50 of more than 868Kb and 721Kb, sensitivity and
234 precision values of more than 0.7 in more than 70% of the cases (Additional File 1: Figure S13).
235 For assemblies with lower contiguity (N50: 470-500 Kb), sensitivity and precision were 0.56 and
236 0.65, respectively.
237 Together this shows that the prediction of genomic rearrangements is nearly complete even if
238 one of the genomes is not on chromosome-level, but has assembly contiguity of N50 > 500 kb.

239 If both assemblies are not on chromosome-level the quality of the predictions is reduced,
240 however, it is still possible to get useful insights on a subset of the rearrangements.

241 **Runtime estimation when comparing human, yeast, fruit fly, and maize genomes**

242 To analyze SyRI's runtime performance, we searched for intra-species genomic differences in
243 four different model organisms: human, yeast, fruit fly, and maize (Additional File 2: Table S2).
244 For its application to human genomes, we compared whole-genome assemblies of NA12878
245 and NA19240 against the reference genome GRCh38.p12 [22,26,27]. For yeast, we compared
246 the *de novo* assembly of strain YJM1447 against the reference genome from strain S288C
247 [28,29]. For fruit fly (*Drosophila melanogaster*), the *de novo* assembly of strain A4 was
248 compared to the reference genome [30,31]. For maize, we compared the *de novo* assembly of
249 PH207 against the B73 reference genome [32,33]. To limit computational requirements, we
250 masked the highly repetitive maize genome while all other genomes were analysed without
251 masking [34].

252 SyRI was used to predict syntenic and rearranged regions as well as local sequence differences
253 in all genomes (Table 1, Additional File 1: Figure S14–S18). In each comparison, including
254 human, at least 5% of the assembled genomes were found to be non-syntenic. The CPU
255 runtime for the smaller and simpler yeast genomes was 34.5 seconds, whereas for the two
256 human genomes SyRI took ~10 minutes, while memory usage was less than 1 GB for each of the
257 comparisons (Table 1) (without considering SNPs and small indels parsing). The exception was
258 the comparison of the repetitive maize genomes, which took ~1hr of CPU time and ~6GB of
259 RAM. Since SyRI considers all alignment combinations, the runtime and memory usage can be

260 high in repetitive genomes (Additional File 1: Note 6 and Figure S19). However, the number of
261 alignments can be drastically reduced by decreasing the WGA sensitivity (i.e. omitting small, 10-
262 100s bp alignments), which in turn decreases runtime and memory consumption of SyRI.

263 **Experimental validation**

264 To validate some of the predicted translocations in the genome of *A. thaliana* L*er*, we used a
265 genetic approach which was based on the observation that recombinant offspring genomes
266 feature different copy numbers of translocated DNA (Figure 1b; 4a), while non-translocated
267 regions always occur with the same copy number. The actual copy number of translocated DNA
268 in a recombinant genome relies on the genotypes at the two insertion sites of the translocation.
269 For example, translocated DNA is duplicated if the two insertion sites of a translocation are
270 combined into one recombinant haplotype.

271 We used available whole-genome sequencing data of a set of 50 F_2 recombinant plants, which
272 were generated by crossing Col-0 and L*er*, followed by self-pollination of the resulting F_1 hybrids
273 (preprint [35]). We aligned the short reads (~5x genome coverage/sample) to the Col-0
274 reference sequence and used the genotypes at ~500k SNP markers to reconstruct the parental
275 haplotypes using TIGER (Figure 4b) [36,37].

276 Based on this haplotype information, we estimated the expected copy number for 117
277 translocations, which were larger than 1kb, in each of the samples. The expected copy-number
278 was then compared to the outcome of three different tests. The first two tests were based on
279 the assumption that all reads from a translocated region align to the same loci in the reference
280 genome independent of the actual location of the rearranged region in the sequenced sample

281 (Figure 4b) [38]. This allows estimating copy-number of a translocation using read coverage in
282 the respective region of the reference. For the first test, we analysed the absence of reads in
283 translocated regions in recombinant genomes, which were predicted to feature no copy of the
284 translocated region (Figure 4c) (using 0.2x read coverage as a cut-off to distinguish between
285 absence or presence of a translocation). For the second test, we assessed the goodness-of-fit
286 between expected copy-number and observed copy-number for a translocation across all
287 recombinants (as estimated from the normalized read counts in the translocation regions;
288 Figure 4d; Methods). The third test was based on the sequence differences between the
289 different alleles of a translocation. For this, we tested differences in the read counts supporting
290 either the Col-0 (or L*er*) alleles of a translocation. Depending on the copy number of the
291 different alleles of a translocation, the allele count should also vary. In consequence, samples
292 with the same genotypes at the two loci of a translocation should have similar allele counts,
293 whereas samples with different genotypes should also show different allele counts (Figure 4e;
294 Methods).
295 Out of 117 translocations, 108 (92.3%) could be confirmed by at least one test (Figure 4f). We
296 manually checked the read alignments of the nine translocations that could not be confirmed
297 and found support for the existence of each of the translocations, which however had not been
298 strong enough to be identified by any of the three test criteria. In summary, this supports that a
299 large majority of the translocations predicted by SyRI are real.

300 ## Discussion

301 We introduced SyRI, a tool that identifies genomic differences between two whole-genome
302 assemblies. The genomic differences include structural differences as well as differences in
303 sequences. However, instead of identifying differences directly, SyRI starts by identifying all
304 syntenic regions between the genomes, as all other (non-syntenic) regions are rearranged by
305 definition.

306 Once the structural differences are found, SyRI identifies local sequence differences in both
307 syntenic and rearranged regions. The identification of local sequence differences in rearranged
308 regions introduces a hierarchy of genomic variations (e.g., SNPs in translocated regions). This
309 distinction is important as rearranged regions are differently inherited as compared to syntenic
310 regions. If this is not accounted for genotypes in rearranged SNPs can confound the
311 interpretation of genomic patterns during selection screens, genome-wide association or
312 recombination analysis [39,40]. SyRI now offers a straight-forward solution to filter SNPs in
313 rearranged regions assuming whole-genome assemblies are available.

314 SyRI provides comprehensive and accurate genomic difference annotation using chromosome-
315 level assemblies. Even though generating *de novo* assembly is still not trivial, consistent
316 improvements are being made. Finally, though SyRI is based on a genome graph that is build up
317 from the local alignments of a WGA, this algorithm can be easily adapted for rearrangement
318 identification in other types of genome graphs as well [41,42].

319 ## Conclusions

320 We have developed SyRI which, to our knowledge, is the first tool to identify all structural and
321 sequence differences between two genome assemblies. This novel approach is highly efficient
322 and includes a classification of sequence differences for being in syntenic or rearranged regions.
323 Using SyRI, we identified genomic rearrangements and sequence differences in humans, *A.*
324 *thaliana*, fruit fly, yeast, and maize genomes. Additionally, we validated the existence of more
325 than 100 predicted translocations. SyRI is available as an open source tool and is being actively
326 developed and improved.

327 ## Methods

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

329 *A. thaliana* L*er* plants were grown in the greenhouse at the Max Planck Institute for Plant
330 Breeding Research. DNA was extracted using the NucleoSpin® Plant II Maxi Kit from Macherey-
331 Nagel. We used the PacBio template prep kit > 20 kb for Sequel systems (SMRTbell Template
332 Prep Kit 1.0-SPv3) with damage repair (SMRTbell Damage Repair Kit -SPv3) and BluePippin size
333 selection for fragments > 9/10 kb. Sequencing of two SMRT cells was done with the Sequel
334 Sequencing Plate 1.2 and the Sequel Binding Kit 1.0. Movie Time 360 min.

335 **Assembly generation**

336 Raw reads were filtered to remove small and low-quality reads (length<50bp and QV<80),
337 corrected and *de novo* assembled using Falcon [43], followed by polishing with Arrow in the
338 SMRTLink5 package, and finally corrected using Illumina short read alignments with reads from

339 an earlier project [44]. The contigs from organellar DNA sequences were removed, all others
340 were anchored into pseudo-chromosome based on homology with the reference sequence.
341 Adjacent contigs were connected with a stretch of 500 "N" characters.

342 **Whole-genome alignments**

343 All assemblies used in this work were filtered to select only chromosome-representing contigs
344 (unplaced scaffolds were removed). We used the *nucmer* alignment tool from the MUMmer
345 toolbox [14] to perform WGAs. Nucmer was run with --maxmatch to get all alignments between
346 two genomes including -c, -b, and -l parameters which were selected to balance alignment
347 resolution and runtime based on genome size and number of repeat regions (full commands
348 are available in Additional File 2: Table S4). Alignments were filtered using the *delta-filter* tool
349 and the filtered delta files were converted to the tab-delimited files using the *show-coords*
350 command. Before whole-genome alignments, both maize genomes were masked using
351 RepeatMasker v4.0.6 [45].

352 **Simulating rearranged genomes**

353 We simulated structural rearrangements in the *A. thaliana* reference genome using the *R*
354 package *RSVSim* [46]. We simulated 40, 436, and 1241 events for inversions, translocations, and
355 duplications respectively and for each rearrangement 100 genomes were simulated. The
356 number of rearrangements and their corresponding sizes were sampled from real differences
357 found between the Col-0 and L*er* genomes. For simulating indels, we used SURVIVOR [47] to
358 simulate 100 genomes containing 1,000 indels in the range of 1-500 bps.

359 From these rearranged genomes, we simulated PacBio and Nanopore reads using SURVIVOR.
360 We used the *A. thaliana* long read data generated by Michael et al. (NCBI project accession:
361 PRJEB21270) to generate read profiles required by SURVIVOR and simulated reads to get a 30x
362 coverage [48]. Short reads were simulated using wgsim (parameters used: -e 0.001 -d 550 -N
363 12000000 -1 150 -2 150) to get 30x coverage [49]. All reads were aligned to the A. *thaliana*
364 reference genome using minimap2 and the alignments were converted from SAM to BAM
365 format and sorted using samtools [15,50].

366 **Running tools on simulated genomes**

367 SyRI: Genome assemblies were aligned using nucmer (Additional File 2: Table S4) and SyRI was
368 run with default parameters. Assemblytics: We used the same alignments generated by nucmer
369 as used for SyRI. The default value for unique sequence length was used and variants size was
370 set from 1 to 100,000bp. AsmVar: The tool was run based on the demo script provided with the
371 tool. For genome alignment, lastdb was run using the default parameters, whereas lastal and
372 last-split were run using the parameters provided in the above demo [51]. Similarly, variants
373 were detected using ASV_VariantDetector tool of AsmVar with the default parameters.
374 Smartie-sv: The pipeline was run using the default settings. However, the number of jobs to be
375 run in parallel and job wait time was adjusted to make it suitable for the computer resources
376 available. Sniffles: Sniffles was run separately for PacBio and Nanopore simulated reads using
377 the default parameters. Alignments were generated through minimap2 and converted to BAM
378 and sorted using samtools. Picky: Picky was run using the same methodology and parameters as
379 described by the authors for both PacBio and Nanopore reads. LUMPY: Reads were aligned by

380 minimap2, and the alignments were pre-processed using samblaster [52] and samtools as per
381 the instructions provided by the authors. While running LUMPY, paired-end read distribution
382 parameters were changed to match the simulated reads (mean:550, read_length:150,
383 min_non_overlap:150).
384 Structural rearrangements predicted by these tools were considered to match an original
385 simulated rearrangement if the end-points for the predicted rearrangement were within
386 ±150bp of the end-points of the original rearrangement. For indels, we compared the location
387 and size of the predicted indels with the simulated indels, allowing for error in both location
388 and size. Two different error limits were checked: 5 and 100 bp.

389 **Performance evaluation with real genome data**

390 For both the *A. thaliana* (L*er*) and the human (NA19240) genome, we used the same methods
391 as above to simulate sequencing reads from whole genome assemblies, to perform alignments
392 with the reference genomes, and to identify genomic differences. For human genomes, we
393 used the error profiles provided by SURVIVOR [47]. Number and sizes of the variations were
394 extracted from the output files using in-house scripts. For the AsmVar comparison of Col-0 vs
395 L*er*, we used the .svd output file instead of the .vcf output file as the former had better suited
396 annotations for our analysis. An indel was considered as identified if there was a simulated
397 indel of the same type (insertion or deletion) within 100bp of the location of the predicted indel
398 and the size difference between two indels was not more than 100bps.

399 **Comparison with the gold standard variation dataset**

400 Variant calls for the gold standard dataset were downloaded from the NCBI [21]. The variants
401 were generated with an older version human reference genome (GRCh38) and were therefore
402 re-mapped to the newer GRCh38.p12 version of the human reference genome using the NCBI
403 Genome Remapping Service. An indel from the gold standard dataset was considered to be
404 identified if a predicted indel of the corresponding type existed within the surrounding 100bp.
405 For inversion predictions, we checked the overlap between inversions from the gold dataset
406 and the inversions, inverted translocations, and inverted duplications as annotated by SyRI.

407 **Pseudo-chromosome generation and output comparison**

408 We generated 200 fragmented assemblies of the L*er* genome by introducing 10-400 random
409 breakpoints. Pseudo-genomes were generated for each of the fragmented assemblies using
410 RaGOO with default parameters. Additionally, we generated 100 fragmented assemblies each
411 of Col-0 and L*er* again by introducing 10-400 random breakpoints. These fragmented assemblies
412 were assembled by a heuristic script (Additional File 1: Note 5) to generate pseudo-molecules.
413 For 16 assemblies, pseudo-molecule generation failed and these samples were skipped from
414 further analysis. A genomic rearrangement identified from the pseudo-genomes was
415 considered to be correct if the same rearrangement type was present within 100bp up or
416 downstream.

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

418 For validation, we used whole-genome sequencing data of 50 F_2 recombinant plants that we
419 generated recently (preprint [35]). We extracted allele count information from consensus call
420 files generated by SHORE [53]. For each predicted translocation, we estimated its copy number

421 as the ratio between average read-coverage for the translocated region and the average read-
422 coverage across the entire genome of the respective sample. Translocations in the centromeric
423 regions and for which more than 25% of the translocated sequence had at least 10% reads with
424 Ns were filtered out. For allele count analysis, we selected high-confidence (25bp conserved in
425 both directions) SNPs in translocated regions as markers.

426 **Validation of translocations: Absence of reads (Test 1)**

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

432 **Validation of translocations: Expected vs. observed copy number (Test 2)**

433 For each translocation, we selected samples which had different genotypes at the two
434 associated loci for the translocation. This removes some of the samples with two copies and
435 helps to remove a bias towards genomes with a copy number of two, which can affect this test.
436 We further selected translocations for which we found samples with at least three different
437 copy-number values predicted. A linear model was fit using the *lm* function in *R*. *P*-values for
438 the model-fit were adjusted for multiple testing using the *BH*-method [54], and translocations
439 for which adjusted *p*-values were less than 10^{-6} and slope more than 0.75 were considered as
440 valid.

441 **Validation of translocations: Genotype clustering (Test 3)**

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

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

451 *Simulating distributions:* Background distributions for the closeness_score were simulated by
452 generating random clusters. For each sample, allele counts (reference and alternate) were
453 sampled using a Poisson distribution. For true translocations, the closeness_score would be low
454 as samples from the same genotype would be much closer to each other, whereas samples
455 from different genotypes would be far. For each translocation, we calculated the lower-tail *p*-
456 value of retrieving the corresponding *closeness_score*. *P*-values were adjusted for multiple
457 testing using *BH*-method, and translocations with *p*-value < 0.05 were considered valid.

458 ## Declarations

459 **Availability of data and materials**

460 The assembly of the L*er* genome has been submitted to the European Nucleotide Archive
461 (http://www.ebi.ac.uk) and is publicly available under the accession number GCA_900660825.
462 All other assemblies are publicly available at NCBI (https://www.ncbi.nlm.nih.gov/), and their
463 accession numbers are GCA_000001735.3 [55], GCA_000001405.27 [26], GCA_002077035.3
464 [27], GCA_001524155.4 [22], GCA_000146045.2 [29], GCA_000977955.2 [28],
465 GCA_000001215.4 [31], GCA_002300595.1 [30], GCA_000005005.6 [33], GCA_002237485.1
466 [32]. Further details about the assemblies are in Additional File 2: Table S2. BAM files for the 50
467 F_2 recombinant genomes are available at European Nucleotide Archive under the project ID
468 PRJEB29265 (preprint [35]). SyRI is freely available under the MIT license and is available online
469 [56]. SyRI is developed using Python3 and is platform independent.

470 **Competing interests**

471 The authors declare that they have no competing interests.

472 **Funding**

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

475 **Authors' contributions**

476 The project was conceived by KS and WBJ. MG and KS developed the algorithms. MG
477 implemented SyRI and performed all analyses. HS processed recombinant genome sequencing
478 data and identified crossing-over sites. WBJ generated the L*er* assembly. The manuscript was

479 written by MG and KS with inputs from HS and WBJ. All authors read and approved the final
480 manuscript.

481 **Acknowledgements**

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

486 References

487 1. Weischenfeldt J, Symmons O, Spitz F, Korbel JO. Phenotypic impact of genomic structural
488 variation: insights from and for human disease. Nat Rev Genet. 2013; 14(2):125–38.
489 2. Tuttle EM, Bergland AO, Korody ML, Brewer MS, Newhouse DJ, Minx P, et al. Divergence
490 and Functional Degradation of a Sex Chromosome-like Supergene. Curr Biol. 2016;
491 26(3):344–50.
492 3. Küpper C, Stocks M, Risse JE, dos Remedios N, Farrell LL, McRae SB, et al. A supergene
493 determines highly divergent male reproductive morphs in the ruff. Nat Genet. 2016;
494 48(1):79–83.
495 4. Lamichhaney S, Fan G, Widemo F, Gunnarsson U, Thalmann DS, Hoeppner MP, et al.
496 Structural genomic changes underlie alternative reproductive strategies in the ruff
497 (Philomachus pugnax). Nat Genet. 2016; 48(1):84–8.
498 5. Lowry DB, Willis JH. A Widespread Chromosomal Inversion Polymorphism Contributes to

499 a Major Life-History Transition, Local Adaptation, and Reproductive Isolation. Barton NH,
500 editor. PLoS Biol. 2010; 8(9):e1000500.
501 6. Guan P, Sung W-K. Structural variation detection using next-generation sequencing data.
502 Methods. 2016; 102:36–49.
503 7. Simpson JT, Pop M. The Theory and Practice of Genome Sequence Assembly. Annu Rev
504 Genomics Hum Genet. 2015; 16(1):153–72.
505 8. Jiao W-B. The impact of third generation genomic technologies on plant genome
506 assembly. Curr Opin Plant Biol. 2017; 36:64–70.
507 9. Sedlazeck FJ, Lee H, Darby CA, Schatz MC. Piercing the dark matter: Bioinformatics of
508 long-range sequencing and mapping. Nat Rev Genet. 2018; 19(6):329–46.
509 10. Liu S, Huang S, Rao J, Ye W, Krogh A, Wang J. Discovery, genotyping and characterization
510 of structural variation and novel sequence at single nucleotide resolution from de novo
511 genome assemblies on a population scale. Gigascience. 2015; 4(1).
512 11. Nattestad M, Schatz MC. Assemblytics: a web analytics tool for the detection of variants
513 from an assembly. Bioinformatics. 2016; 32(19):3021–3.
514 12. Kronenberg ZN, Fiddes IT, Gordon D, Murali S, Cantsilieris S, Meyerson OS, et al. High-
515 resolution comparative analysis of great ape genomes. Science (80-). 2018;
516 360(6393):eaar6343.
517 13. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: A fast
518 and versatile genome alignment system. Darling AE, editor. PLOS Comput Biol. 2018;
519 14(1):e1005944.
520 14. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and

521 open software for comparing large genomes. Genome Biol. 2004; 5(2):R12.

522 15. Li H. Minimap2: pairwise alignment for nucleotide sequences. Birol I, editor.
523 Bioinformatics. 2018; 34(18):3094–100.

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

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

528 18. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, et al.
529 Accurate detection of complex structural variations using single-molecule sequencing.
530 Nat Methods. 2018; 15(6):461–8.

531 19. Gong L, Wong C-H, Cheng W-C, Tjong H, Menghi F, Ngan CY, et al. Picky comprehensively
532 detects high-resolution structural variants in nanopore long reads. Nat Methods. 2018;
533 15(6):455–60.

534 20. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural
535 variant discovery. Genome Biol. 2014; 15(6):R84.

536 21. Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, et al. Multi-
537 platform discovery of haplotype-resolved structural variation in human genomes. Nat
538 Commun. 2019; 10(1):1784.

539 22. The Genome Institute at Washington University School of Medicine.
540 NA19240_prelim_3.0. https://www.ncbi.nlm.nih.gov/assembly/GCA_001524155.4. 2017.

541 23. Jiao W-B, Accinelli GG, Hartwig B, Kiefer C, Baker D, Severing E, et al. Improving and
542 correcting the contiguity of long-read genome assemblies of three plant species using

543 optical mapping and chromosome conformation capture data. Genome Res. 2017;
544 27(5):778–86.
545 24. Sun H, Ding J, Piednoël M, Schneeberger K. FindGSE: estimating genome size variation
546 within human and Arabidopsis using k-mer frequencies. Bioinformatics. 2018; 34(4):550–
547 7.
548 25. Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, et al. Fast and
549 accurate reference-guided scaffolding of draft genomes. bioRxiv. 2019; :519637.
550 26. Human Genome Sequencing Consortium I. Finishing the euchromatic sequence of the
551 human genome. Nature. 2004; 431(7011):931–45.
552 27. The Genome Institute at Washington University School of Medicine.
553 NA12878_prelim_3.0. https://www.ncbi.nlm.nih.gov/assembly/GCA_00207703. 2018.
554 28. Strope PK, Skelly DA, Kozmin SG, Mahadevan G, Stone EA, Magwene PM, et al. The 100-
555 genomes strains, an S. cerevisiae resource that illuminates its natural phenotypic and
556 genotypic variation and emergence as an opportunistic pathogen. Genome Res. 2015;
557 25(5):762–74.
558 29. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, et al. Life with 6000
559 Genes. Science (80-). 1996; 274(5287):546–67.
560 30. Chakraborty M, VanKuren NW, Zhao R, Zhang X, Kalsow S, Emerson JJ. Hidden genetic
561 variation shapes the structure of functional elements in Drosophila. Nat Genet. 2018;
562 50(1):20–5.
563 31. Hoskins RA, Carlson JW, Wan KH, Park S, Mendez I, Galle SE, et al. The Release 6
564 reference sequence of the *Drosophila melanogaster* genome. Genome Res. 2015;

565 25(3):445–58.

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

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

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

573 35. Sun H, Rowan BA, Flood PJ, Brandt R, Fuss J, Hancock AM, et al. Linked-read sequencing
574 of gametes allows efficient genome-wide analysis of meiotic recombination. bioRxiv.
575 2018; :484022.

576 36. Zapata L, Ding J, Willing E-M, Hartwig B, Bezdan D, Jiao W-B, et al. Chromosome-level
577 assembly of Arabidopsis thaliana Ler reveals the extent of translocation and inversion
578 polymorphisms. Proc Natl Acad Sci U S A. 2016; 113(28):E4052-60.

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

582 38. Imprialou M, Kahles A, Steffen JG, Osborne EJ, Gan X, Lempe J, et al. Genomic
583 rearrangements in Arabidopsis considered as quantitative traits. Genetics. 2017;
584 205(4):1425–41.

585 39. Wijnker E, Velikkakam James G, Ding J, Becker F, Klasen JR, Rawat V, et al. The genomic
586 landscape of meiotic crossovers and gene conversions in *Arabidopsis thaliana*. Elife.

587 2013; 2.

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

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

593 42. Consortium TCP. Computational pan-genomics: status, promises and challenges. Brief
594 Bioinform. 2016; (August):bbw089.

595 43. Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al. Phased
596 diploid genome assembly with single-molecule real-time sequencing. Nat Methods.
597 2016; 13(12):1050–4.

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

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

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

604 47. Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, et al. Transient structural variations
605 have strong effects on quantitative traits and reproductive isolation in fission yeast. Nat
606 Commun. 2017; 8(1):14061.

607 48. Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, et al. High contiguity
608 Arabidopsis thaliana genome assembly with a single nanopore flow cell. Nat Commun.

2018; 9(1):541.

49. Li H. Wgsim: Reads simulator. https://github.com/lh3/wgsim.

50. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25(16):2078–9.

51. Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. Genome Res. 2011; 21(3):487–93.

52. Faust GG, Hall IM. SAMBLASTER: fast duplicate marking and structural variant read extraction. Bioinformatics. 2014; 30(17):2503–5.

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

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

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

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

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 the method and additional results
Additional File 2	.xlsx	Additional tables	Information about methodology and data used

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

629

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	542.71	581	size	2.8 Gb	7.0 Mb	9.0 Mb	21.8 Mb	8.6 Mb	224.1 Mb
					% genome	91.1	0.2	0.3	0.7	0.3	7.4
					number	1147	66	156	2420	1460	840
	NA19240	3.04 Gb	528.79	1003	size	2.8 Gb	3.8 Mb	10.3 Mb	20.6 Mb	8.0 Mb	208.8 Mb
					% genome	91.7	0.1	0.3	0.7	0.3	6.9
					number	1134	68	153	2288	1242	848
Yeast	YJM1447	12.1 Mb	34.51	5	size	11.2 Mb	1.8 Kb	52.1 Kb	603.2 Kb	166.2 Kb	87.3 Kb
					% genome	92.5	0.02	0.4	5.0	1.4	0.7
					number	222	3	19	92	313	164
Fruit Fly	A4	135.5 Mb	522.02	289	size	124.8 Mb	130.3 Kb	1.3 Mb	3.9 Mb	4.2 Mb	1.2 Mb
					% genome	92.1	0.1	1.0	2.9	3.1	0.8
					number	1947	15	390	2206	2427	1365
Maize	PH207	2.06 Gb	3342.62	5873	size	1.3 Gb	82.9 Mb	4.3 Mb	5.2 Mb	16.2 Mb	669.6 Mb
					% genome	62.2	4.0	0.2	0.3	0.8	32.5
					number	8779	195	1383	2688	9495	15166

630 ## Figure legends

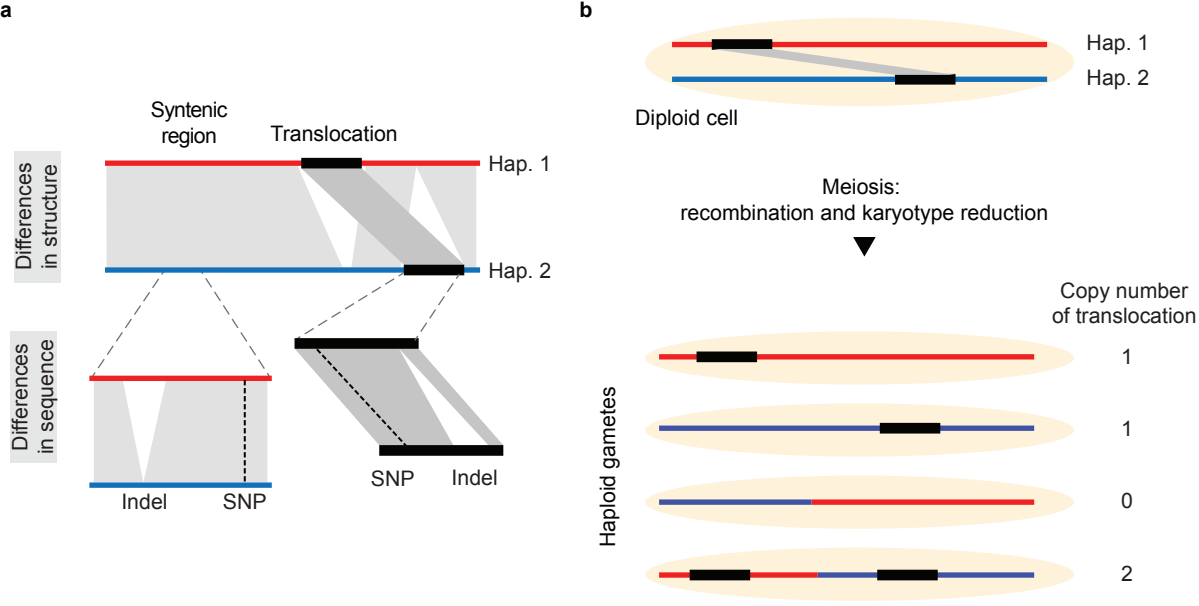
631 **Figure 1: Hierarchy of genomic differences and their propagation.** a) Genomic
632 differences include differences in the structure (like inversions, translocations or duplications)
633 as well as local sequence differences like SNPs/indels. Differences in sequence can occur in
634 syntenic regions as well as in rearranged regions. b) A diploid cell containing two haplotypes,
635 which can be distinguished by a translocation. Following meiosis and recombination, the
636 haploid gametes can feature different copy-number variation for the translocated regions (and
637 the sequence differences in it).

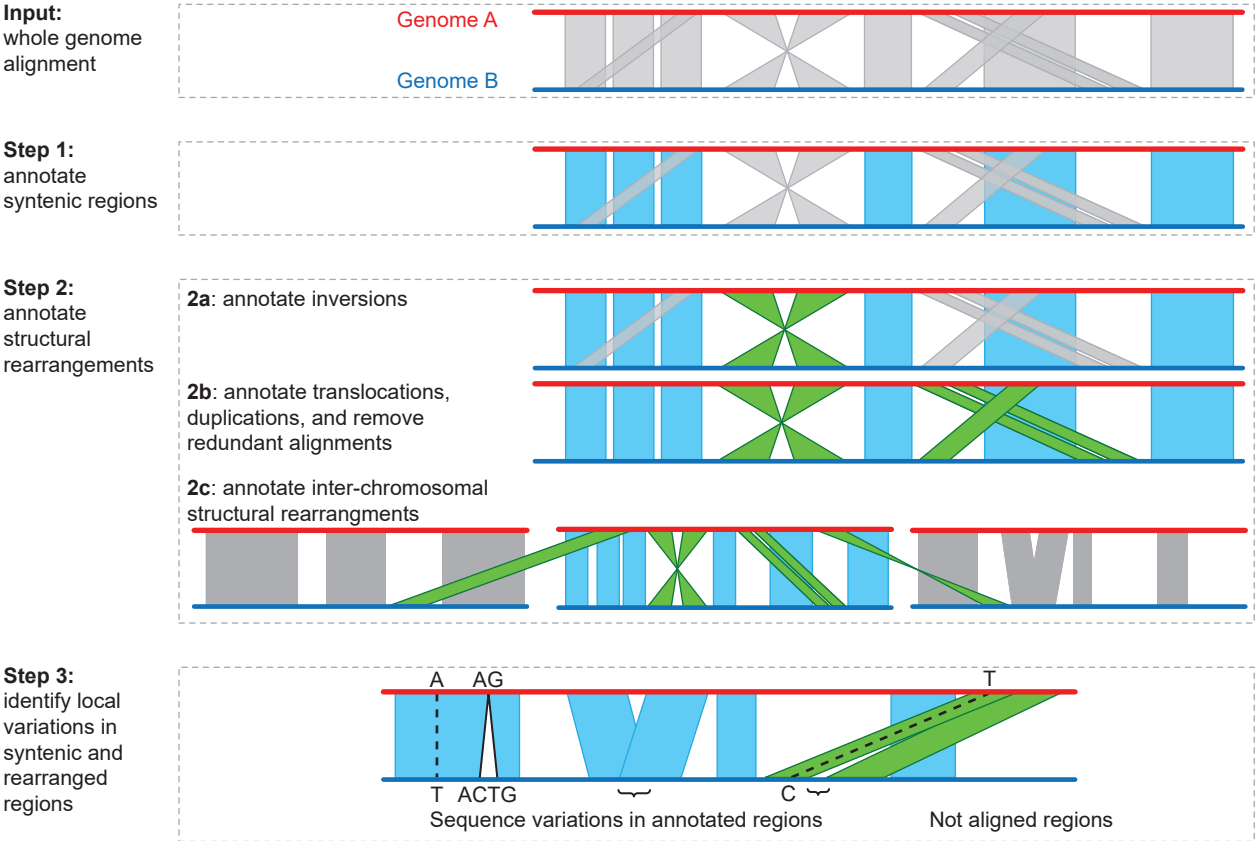
638

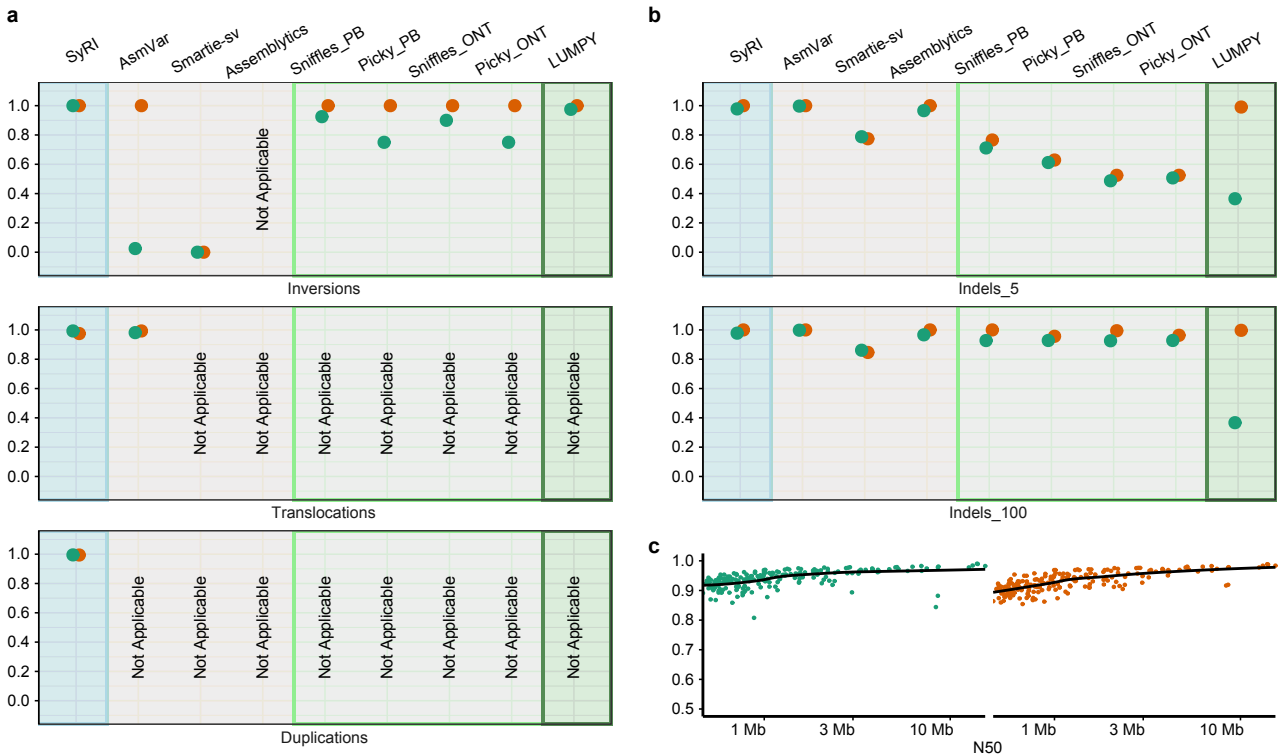
639 **Figure 2: Workflow for the identification of genomic differences.** SyRI uses whole-
640 genome alignments (WGA) as input. A WGA consists of a set of local alignments, where each
641 local alignment (grey polygon) connects a specific region in one genome to a specific region in
642 the other genome. Step 1: SyRI identifies the highest scoring syntenic path between the
643 corresponding genomes (blue alignments). The syntenic path represents the longest set of non-
644 rearranged regions between two genomes. Step 2 (a-c): The remaining alignments are
645 separated into structural rearrangements and redundant alignments. Structural
646 rearrangements (green alignments) are classified into inversions, intra-chromosomal
647 translocations and duplications, and finally inter-chromosomal rearrangements. Step 3: Local
648 differences in the sequences are identified in all syntenic and rearranged regions. SNPs and
649 small indels are parsed directly from the local alignments, whereas more complex sequence
650 variation (e.g. like large indels and CNVs) are identified in the overlaps and gaps between

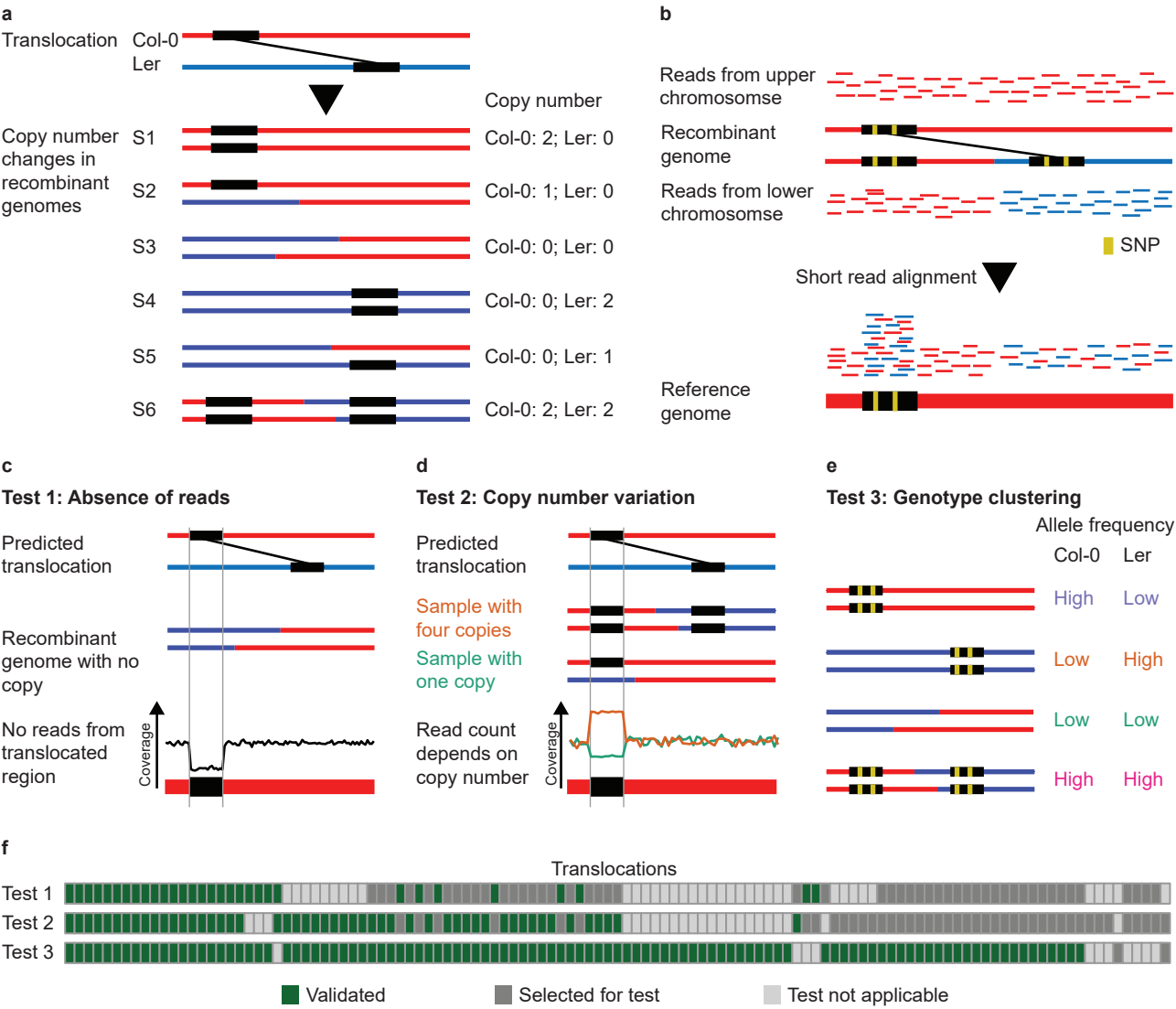
651 consecutive local alignments. Also, all non-aligned regions in between syntenic and rearranged
652 regions are reported for completeness.
653
654 **Figure 3: SyRI's performance compared to six other tools using simulated assemblies**
655 **and on incomplete assemblies.** Sensitivity (green points) and precision (orange points)
656 values are shown for the identification of (a) rearranged regions and (b) indels. Each point
657 describes the median sensitivity (and precision) values from the analysis of 100 simulated
658 genomes. Background colors represent the data type required by the respective tools: light blue
659 for chromosome-level *de novo* assembly, pale yellow for *de novo* assembly, pale green for long
660 sequencing reads (both PacBio (PB) and Oxford Nanopore (ONT) reads), green for short
661 sequencing reads.
662 (a) Performance values for inversions, translocations and duplications. (As AsmVar found
663 inversions in only 40 samples, performance values for AsmVar considered only these 40
664 samples.) (b) Performance values for indel identification using stringent (upper) or relaxed
665 (lower) thresholds for matching predicted vs. simulated indel positions. (LUMPY cannot identify
666 insertions, however, sensitivity calculation did not account for that.) (c) SyRI's performance for
667 the identification of rearranged regions from incomplete assemblies. Each point represents the
668 performance value generated with one of the simulated assemblies. The black line represents
669 the polynomial fit.
670
671 **Figure 4: Recombination introduces copy-number variation.** (a) Recombination between
672 two haplotypes with translocated regions can lead to copy-number differences in the

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


a
Syntenic region
Translocation
Differences in structure
Hap. 1
Hap. 2
Differences in sequence
Indel
SNP
SNP
Indel
b
Hap. 1
Hap. 2
Diploid cell
Meiosis:
recombination and karyotype reduction
Copy number of translocation
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:
identify local variations in syntenic and rearranged regions
A
AG
T
T
ACTG
C
Sequence variations in annotated regions
Not aligned regions


a
b
c
SyRI
AsmVar
Smartie-sv
Assemblytics
Sniffles_PB
Picky_PB
Sniffles_ONT
Picky_ONT
LUMPY
Inversions
Translocations
Duplications
Indels_5
Indels_100
Not Applicable
1.0
0.8
0.6
0.4
0.2
0.0
0.9
0.7
0.5
1 Mb
3 Mb
10 Mb
N50


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