

1 Inferring synteny between genome assemblies: a systematic
2 evaluation

3

4

5 Dang Liu^{1,2}, Martin Hunt^{3,4} and Isheng. J. Tsai^{1,2*}

6

7

8

9 ¹Genome and Systems Biology Degree Program, National Taiwan University and
10 Academia Sinica, Taipei, Taiwan

11 ²Biodiversity Research Center, Academia Sinica, Taipei, Taiwan

12 ³Nuffield Department of Clinical Medicine, Experimental Medicine Division, John
13 Radcliffe Hospital, University of Oxford, Oxford, OX1 1NF, UK

14 ⁴European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus,
15 Hinxton, Cambridge CB10 1SD, UK

16

17

18

19

20 *Corresponding author

21 Email: ijtsai@gate.sinica.edu.tw

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41 **Abstract**

42 Identification of synteny between genomes of closely related species is an important
43 aspect of comparative genomics. However, it is unknown to what extent draft assemblies
44 lead to errors in such analysis. To investigate this, we fragmented genome assemblies of
45 model nematodes to various extents and conducted synteny identification and
46 downstream analysis. We first show that synteny between species can be underestimated
47 up to 40% and find disagreements between popular tools that infer synteny blocks. This
48 inconsistency and further demonstration of erroneous gene ontology enrichment tests
49 throws into question the robustness of previous synteny analysis when gold standard
50 genome sequences remain limited. In addition, determining the true evolutionary
51 relationship is compromised by assembly improvement using a reference guided
52 approach with a closely related species. Annotation quality, however, has minimal effect
53 on synteny if the assembled genome is highly contiguous. Our results highlight the need
54 for gold standard genome assemblies for synteny identification and accurate downstream
55 analysis.

56

57

58 **Author summary**

59

60 Genome assemblies across all domains of life are currently produced routinely.
61 Initial analysis of any new genome usually includes annotation and comparative
62 genomics. Synteny provides a framework in which conservation of homologous genes
63 and gene order is identified between genomes of different species. The availability of
64 human and mouse genomes paved the way for algorithm development in large-scale
65 synteny mapping, which eventually became an integral part of comparative genomics.
66 Synteny analysis is regularly performed on assembled sequences that are fragmented,
67 neglecting the fact that most methods were developed using complete genomes. Here, we
68 systematically evaluate this interplay by inferring synteny in genome assemblies with
69 different degrees of contiguation. As expected, our investigation reveals that assembly
70 quality can drastically affect synteny analysis, from the initial synteny identification to
71 downstream analysis. Importantly, we found that improving a fragmented assembly using
72 synteny with the genome of a related species can be dangerous, as this *a priori* assumes a
73 potentially false evolutionary relationship between the species. The results presented here
74 re-emphasize the importance of gold standard genomes to the science community, and
75 should be achieved given the current progress in sequencing technology.

76

77

78

79 Introduction

80

81 The essence of comparative genomics lies in how we compare genomes to reveal
82 species' evolutionary relationships. Advances in sequencing technologies have enabled
83 the generation and exploration of many new genomes across all domains of life [1–8].
84 Unfortunately, in most instances correctly aligning even just two genomes at base-pair
85 resolution can be challenging. A genome usually contains millions or billions of
86 nucleotides and is different from the genome of a closely related species as a result of
87 evolutionary processes such as mutations, chromosomal rearrangements, and gene family
88 expansion or loss. There are computational complexities when trying to align and assign
89 multiple copies of DNA, such as transposable elements that are identical to each other
90 [9–12]. In addition, it has been shown that popular alignment methods disagree with each
91 other [9].

92

93 An alternative and arguably more practical approach relies on the identification of
94 synteny blocks [13,14], first described as homologous genetic loci presenting on the same
95 chromosome [15,16]. Currently it is more formally defined as regions of chromosomes
96 between genomes that share a common order of homologous genes derived from a
97 common ancestor [17,18]. Alternative names such as conserved synteny or collinearity
98 have been used interchangeably [13,19–22]. Comparisons of genome synteny between
99 and within species have provided an opportunity to study evolutionary processes that lead
100 to diversity of chromosome number and structure in many lineages across the tree of life
101 [23,24]; early discoveries using such approaches include chromosomal conserved regions
102 in nematodes and yeast [25–27], evolutionary history and phenotypic traits of extremely
103 conserved Hox gene clusters across animals and MADS-box gene family in plants
104 [28,29], and karyotype evolution in mammals [30] and plants [31]. Analysis of gene
105 synteny against closely related species is now the norm in every new published genome.
106 However, assembly quality comes into question as it has been demonstrated to affect
107 subsequent analysis such as annotation or rate of lateral transfer [32,33].

108

109 In general, synteny identification is a filtering and organizing process of all local
110 similarities between genome sequences into a coherent global picture [34]. Orthologous
111 relationships of protein-coding genes are used as anchors to position statistically
112 significant local alignments. Approaches include the use of a directed acyclic graph
113 [35,36], a gene homology matrix (GHM) [37], and an algorithm using reciprocal best hits
114 (RBH) [38]. All of these methods generally agree on long synteny blocks, but have
115 differences in handling local shuffles as well as the resolution of synteny identification
116 [34,38]. Better resolution of micro-rearrangements in synteny patterns has been shown
117 when using an improved draft genome of *Caenorhabditis briggsae* versus *Caenorhabditis*
118 *elegans* [26,39]. Hence, synteny analysis depends highly on assembly quality. For

119 example, missing sequences in assembly lead to missing gene annotations and
120 subsequently missing orthologous relationship [40]. With respect to assembly
121 contiguation, it still remains unclear whether assembly fragmentation affects the
122 homology assignments for deciding anchors, features of genes for examining order and
123 gaps, or other factors in synteny analysis.

124

125 In this study, we focus on how assembly quality affects the identification of genome
126 synteny. We investigate the correlation between error rate (%) in detecting synteny and
127 the level of assembly contiguation using four popular software packages (DAGchainer
128 [35], i-ADHoRe [37], MCSanX [36], and SynChro [38]) on four nematodes:
129 *Caenorhabditis elegans*, *Caenorhabditis briggsae*, *Strongyloides ratti* and *Strongyloides*
130 *stercoralis*. We also carried out and explored the effects of three scenarios associated
131 with synteny analysis: gene ontology (GO) enrichment, reference-guided assembly
132 improvement, and annotation quality. Our findings show that assembly quality does
133 matter in synteny analysis, and fragmented assemblies ultimately lead to erroneous
134 findings. In addition, the true evolutionary relationship may be lost if a fragmented
135 assembly is improved using a reference-guided approach. Our main aim here is to
136 determine a minimum contiguation of assembly for subsequent synteny analysis to be
137 trustworthy, which should be possible using the latest sequencing technologies [41].

138

139

140 **Results**

141

142 **Definition of synteny block, break and coverage**

143

144 We begin with some terminology that is used throughout this study. As shown in Fig
145 1, a synteny block is defined as a region of genome sequence spanning a number of genes
146 that are orthologous and co-arranged with another genome. Orientation is not considered
147 (Fig 1, Block a and b). The minimum number of co-arranged orthologs which are said to
148 be the anchors can be set and vary between different studies. A larger number of
149 minimum anchors may result in fewer false positives, but also a more conservative
150 estimate of synteny blocks (S1 Fig). In some programs, some degrees of gaps are
151 tolerated (Fig 1, Block c), which are defined as the number of skipped genes or the length
152 of unaligned nucleotides. A score is usually calculated, and synteny breaks are therefore
153 regions that do not satisfy a certain score threshold. Possible scenarios that lead to
154 synteny breaks include a lack of anchors in the first place (Fig 1, break a), a break in
155 anchor order (Fig 1, break b), or gaps (Fig 1, break c). Genome insertions and
156 duplications are potential causes of over-sized gaps. An example is Break c of Fig 1,
157 which is due to either a large unaligned region (Fig 1, P¹-Q¹ and Q²-R²) or a high number

158 of skipped genes (Fig 1, $S^2-T^2-X^2$ within Q^2-R^2). Alternatively, an inversion (Fig 1,
159 orthologs K and L), deletion or translocation (Fig 1, ortholog X) may cause a loss of
160 anchors (Fig 1, gene D in species 1) or a break in the arrangement of anchors. Typically,
161 synteny coverage is commonly used as a summary metric, which is obtained by dividing
162 summed length of blocks by genome size. Note that synteny coverage is asymmetrical, as
163 demonstrated by the difference of block length in Block c (Fig 1).

164

165 **Fig 1. Definition of synteny block and synteny break**

166 Genes located on chromosomes of two species are denoted in letters. Each gene is
167 associated with a number representing the species they belong to (species 1 or 2).
168 Orthologous genes are connected by dashed lines and genes without an orthologous
169 relationship are treated as gaps in synteny programs. Under the criteria of at least three
170 orthologous genes (anchors): a synteny block can be orthologs with the same order
171 (Block a), reverse order (Block b), or allowing some gaps (Block c). In contrast, cases of
172 causing a synteny break can be lack of orthologs (Break a), gene order (Break b) or gaps
173 (Break c).

174

175

176 **Evaluation of synteny identification programs in fragmented assemblies**

177

178 There are several programs developed to identify synteny blocks, which can produce
179 quite different results [14]. Our first aim is to systematically assess the synteny
180 identification of four popular tools: DAGchainer [35], i-ADHoRe [37], MCScanX [36],
181 and SynChro [38]. As whole genome alignments between bacteria, which have relatively
182 small genomes, is becoming common practice [42], we conduct this study on species
183 with larger genome sizes. We chose *Caenorhabditis elegans*, a model eukaryote with a
184 100 megabase (Mb) reference genome. Our first question was whether these programs
185 would produce 100% synteny coverage if the *C. elegans* genome was compared to itself.
186 As expected, all tools accurately achieved almost 100% synteny coverage (Fig 2 and S2
187 Fig).

188

189 We then fragmented the *C. elegans* genome into fixed intervals of either 100kb,
190 200kb, 500kb or 1Mb to evaluate the performance of different programs when using
191 self-comparisons (Methods). Synteny coverages of the fragmented assembly (query)
192 against the original assembly (reference) were calculated for both query and reference
193 sequences. As expected, synteny coverage decreased as the assembly was broken into
194 smaller pieces. For example, an average of 16% decrease in synteny coverage was
195 obtained using the assembly with fixed fragment size of 100kb (S2 Table). Sites of
196 fragmentation are highly correlated with synteny breaks (Fig 2). One explanation is that
197 fragmented assembly introduced breaks within genes resulting in loss of anchors (Fig 1,
198 Break a), which can be common in real assemblies if introns contain hard to assemble

199 sequences [32]. Another explanation is that the breaks between genes lead to the number
200 anchors not reaching the required minimum (Fig 1, Break a). To assess our fragmented
201 approach to real data, we obtained a recent publicly available genome of *C. elegans* using
202 long reads data (Methods). The assembly has an N50 of ~1.6Mb and we annotated this
203 assembly *de novo*. A synteny coverage of 98.9% was obtained which is very similar to
204 our 1Mb fragmented assemblies of 98.4% (S2 Fig and S1 Table) suggesting robustness in
205 our fragmentation approach.

206

207 More fragmented assemblies led to greater differences in synteny coverage predicted
208 between the four tools (Fig 2 and S2-4 Figs). We carefully examined regions where
209 synteny was predicted in some programs but not the other (Figs 2 and 3). Fig 3
210 demonstrates such a case of disagreement. DAGchainer and i-ADHoRe produced the
211 same results, whilst MCSanX and SynChro detected less and more synteny, respectively
212 (Fig 3). MCSanX's gap scoring scheme used a stricter synteny definition, and more
213 synteny blocks can be identified when the gap threshold was lowered (Fig 3, situation a;
214 also see S5 Fig). SynChro has its own dedicated orthology assignment approach that
215 assigns more homologous anchors (Fig 3, situation b). Additionally, SynChro uses only 2
216 genes as anchors to form a synteny block, while the default is at least five gene anchors in
217 other three tools (Fig 3, situation b). Together, these results suggest that the default
218 parameters set by different tools will lead to differences in synteny identification and
219 need to be tuned before undertaking subsequent analysis.

220

221 **Fig 2. Synteny blocks identification between *C. elegans* chromosome IV.** The original
222 sequence is used as the reference and coloured in black. Established synteny regions
223 (outer number stands for synteny coverage) of the 4 different program packages:
224 DAGchainer (red), i-ADHoRe (yellow), MCSanX (green) and SynChro (light blue) are
225 joined to query sequences with different levels of fragmentation (un-fragmented, 1Mb
226 and 100kb fragmented). Chromosome positions are labeled in megabases (Mb). For plots
227 of other chromosomes see S1-3 Figs.

228 **Fig 3. A zoomed-in 600kb region of synteny identification between the reference *C.***
229 ***elegans* genome and a 100kb fragmented assembly.** Synteny blocks defined by the four
230 detection programs DAGchainer (red), i-ADHoRe (yellow), MCSanX (green) and
231 SynChro (light blue) are drawn as rectangles. Fragmented sites are labeled by vertical red
232 dashed lines. Genes are shown as burgundy rectangles, with gene starts marked using
233 dark blue lines. Two scenarios are marked: a) synteny block was not identified by
234 MCSanX, and b) synteny blocks only detected by SynChro.

235

236

237 **Contribution of assembly contiguation and intrinsic species effect to synteny**
238 **analysis**

239

240 To quantify the effect of assembly contiguation in synteny analysis, we used four
241 nematode genomes: *Caenorhabditis elegans*, *Caenorhabditis briggsae*, *Strongyloides*
242 *ratti* and *Strongyloides stercoralis*. Nematodes are useful models in synteny analysis as 1)
243 extensive chromosomal rearrangement are hallmarks of their genome evolution
244 [7,25,26,43,44]; 2) the genome sequences are highly contiguous and assembled into
245 chromosomes [7,25,26,43]. Also, these two genera were chosen to investigate the
246 intrinsic species effect as they differ in their density of genes (Table 1). Our
247 fragmentation approach was first used to break the *C. elegans* and *S. ratti* genomes into
248 fixed sequence size of either 100kb, 200kb, 500kb, or 1Mb. Here, we define the error rate
249 as the difference between the original synteny coverage (almost at 100%) and when
250 assembly is fragmented. For each fixed length, the fragmentation was repeated 100 times
251 so assemblies are broken at different places to obtain a distribution. As expected, there is
252 a positive correlation between error rate and level of fragmentation. Specifically, the
253 median error rate can be as high as 18% at assemblies with sequence length of 100kb (S2
254 Table). Amongst the four tools, fragmented assemblies have largest effect in MCScanX
255 and least in SynChro (Fig 4 and S2 Table).

256
257 A common analysis when reporting a new genome is inferring synteny against a
258 closely related species. Hence, we reanalyzed synteny between *C. elegans* and *C.*
259 *briggsae*, *S. ratti* and *S. stercoralis*. On average, the four tools found 77% and 83%
260 synteny between *C. elegans*-*C. briggsae* and *S. ratti*-*S. stercoralis* respectively (S2
261 Table). In contrast to broad agreement on within-species self-comparisons, the tools
262 varied considerably on these inter-species (i.e. more diverged) comparisons (S6 Fig and
263 S2 Table). For example, in the *C. elegans*-*C. briggsae* comparisons, a difference of 25%
264 in synteny coverage was found between the results of i-ADHoRe and SynChro (S6 Fig
265 and S2 Table), while this tool variation was interestingly much lower in *S. ratti*-*S.*
266 *stercoralis* with only 9% difference (S2 Table). To increase the complexity, we
267 fragmented *C. briggsae* and *S. stercoralis* into fixed sequence sizes using the same
268 approach as previously mentioned and compared them with the genome of *C. elegans* and
269 *S. ratti*, respectively. We found that MCScanX still underestimated synteny the most as
270 the scaffold size decreased from 1Mb to 100 kb. Strikingly, the median error rate was
271 high as 40% in *C. elegans*-*C. briggsae* but only 12% in *S. ratti*-*S. stercoralis* comparisons
272 (Fig 4). This observation suggests that higher gene density leads to more robustness in
273 synteny detection in fragmented assemblies as more anchors (genes) are available in a
274 given sequence (Table 1 and S2 Table).

275

276 **Table 1. Genomic features of *Caenorhabditis* and *Strongyloides* species.**

	<i>C. elegans</i>	<i>C. briggsae</i>	<i>S. ratti</i>	<i>S. stercoralis</i>
Genome size (Mb)	100.3	108.4	43.2	42.7

Number of genes	20,247	21,814	12,453	13,098
Gene density (genes/Mb)	201.9	201.3	288.6	307.0
Gene coverage (%)	63.1	59.7	50.9	51.8
Median gene length (bp)	1,972	1,964	1,281	1,195
Mean gene length (bp)	3,124.2	2,967.5	1,763.8	1,687.8
Median intergenic length	925	1,183	923	808
Mean intergenic length	2,209.5	2,394.6	1,712.0	1,513.8

277 Features that may play a key role in synteny detection are highlighted in yellow.

278

279 **Fig 4. Error rate (%) of synteny identification in fragmented assemblies.** The error
280 rate is defined as the difference between the synteny coverage calculated with original
281 genome (almost 100%) and that in fragmented assemblies, where in both cases the
282 original assembly was used as the reference. 5 % and 2 % error rate positions are marked
283 by grey solid and dashed lines, respectively. Different pairs of synteny identification are
284 separated in different panels. The upper panels are self-comparisons, while the bottom are
285 comparisons between closely related species. Note that for a clear visualization of pattern
286 changes, the scales of error rate are different between upper and bottom panels. Colors
287 represent different types of synteny detection programs.

288

289 **Erroneous findings using fragmented assemblies in synteny analysis**

290

291 Functional enrichment of genes of interest are usually investigated after the
292 establishment of orthology and synteny [26,45–48]. Synteny breaks contain
293 rearrangements, novel genes, or genes that are too diverged to establish an orthologous
294 relationship or have undergone expansion or loss. Functions of these genes are often of
295 interest in comparative genomics analyses. To further estimate the effect of poor
296 assembly contiguation on synteny analysis, GO enrichment was performed for genes
297 present in synteny breaks in the original assembly of *C. elegans* versus fragmented
298 assemblies of *C. briggsae*. This approach was then repeated 100 times each with
299 assemblies fragmented randomly. We found that fragmented assemblies lead to GO terms
300 originally not in the top 100 ranks then consistently appearing in the top 10 during the
301 100 replicates (Fig 5 and S3-6 Table). Furthermore, the orders of the original top 10 GO
302 terms shifted in fragmented assemblies (Fig 5 and S3-6 Table). In addition, the 10th top
303 GO term failed to appear in the top 10 in 100 replicates of 100kb and 200kb fragmented
304 assemblies (Fig 5 and S3-6 table). These results suggest that an underestimation of
305 synteny relationship due to poor assembly contiguation can lead to a number of erroneous
306 findings in subsequent analysis.

307

308 **Fig 5. Comparison of gene ontology (GO) enriched terms in *C. briggsae* synteny**
309 **break between *C. elegans* vs. *C. briggsae* and 100 replicates of *C. elegans* vs. 100kb**

310 **fragmented *C. briggsae*.** Top ranks of GO terms in the original comparison are shown in
311 the Y axis. For original top ranking GO terms, only those that appeared more than 10
312 times in top 10 ranks of after-fragmentation comparison replicates were displayed (see S6
313 Table for more details). The X axis shows top 10 ranks and rank “out of top 10” in the
314 comparison when assemblies were fragmented. The darkness of color is proportional to
315 the occurrence of the GO term in that rank within 100 replicates. Regions in red are
316 indications of occurred errors.

317

318 **True synteny is compromised by reference-guided assembly methods**

319

320 Although assembly quality plays an important role in synteny analysis, it has been
321 demonstrated that poor assembly contiguity of one species can be scaffolded by
322 establishing synteny with a more contiguous assembly of a closely related species [40]
323 [49–51]. However, we hypothesised that the true synteny relationship between two
324 species may be incorrectly inferred when an assembly of one species is scaffolded based
325 on another closely related species, by assuming the two genomes are syntenic. In order to
326 investigate this, ALLMAPS [51] was used to order and orient sequences of 100kb
327 fragmented *C. briggsae* based on *C. elegans* as well as 100kb fragmented *S. stercoralis*
328 assembly based on *S. ratti*. ALLMAPS improved both fragmented assemblies
329 impressively, increasing the N50 from 100kb to 19Mb and 15Mb in *C. briggsae* and *S.*
330 *stercoralis*, respectively (S7 Table). Synteny coverage from these improved assemblies
331 was even higher than the original fragmented 100kb sequences in MCSanX, much lower
332 in i-ADHoRe, and similar in DAGchainer and SynChro (Fig 6). In addition, despite
333 synteny coverage close to that of the original assemblies in DAGchainer and SynChro,
334 further investigation of synteny block linkages in *C. elegans-C. briggsae* using
335 identification from DAGchainer revealed frequent false ordering and joining of synteny
336 blocks. Intra-chromosomal rearrangements are common between *C. elegans* and *C.*
337 *briggsae*, but ALLPMAPS improved assembly have shown a false largely collinear
338 relationship in the chromosomes between the two species (Fig 7). Hence, reference
339 guided assembly improvement produces pseudo-high quality assemblies that may have
340 ordering biased towards the reference genome and may not reflect the true evolutionary
341 scenario.

342

343 **Fig 6. Synteny coverage (%) between *C. elegans* and *S. ratti* assemblies against**
344 **original or ALLMAPS improved 100kb fragmented *C. briggsae* and *S. stercoralis*.**

345 **Fig 7. Synteny linkage between *C. elegans* vs. original *C. briggsae* assemblies and *C.***
346 ***elegans* vs. ALLMAPS *C. briggsae* assembly.** ALLMAPS assembly with L90 = 1,063
347 from 100kb fragmented *C. briggsae* assembly with L90 = 6 (top), original *C. elegans*
348 assembly with L90 = 6 (middle) and original *C. briggsae* assembly with L90 = 6 (bottom)
349 are shown in different horizontal lines. Vertical lines on chromosome lines show the
350 start/end positions of the first/last gene in a synteny block. Each panel shows a separate

351 chromosome. Block linkages in the same orientation are labeled in red, while those in
 352 inverted orientation are labeled in blue.

353

354

355 **Annotation quality has little effect on synteny analysis**

356

357 Genome annotation is a bridging step between genome assembly and synteny
 358 analysis. An incomplete annotation may lead to lack of homology information in synteny
 359 analysis. We compared synteny coverage in three datasets of *C. elegans* that differ in
 360 quality of annotation: 1) manually curated WormBase [52] *C. elegans* annotation, 2)
 361 optimized Augustus [53] annotation with its built-in *Caenorhabditis* species training set,
 362 and 3) semi-automated Augustus annotation with the BUSCO [54] nematoda species
 363 training set. In either case, we found that synteny coverage varies at most 0.02% in
 364 reference genome (Table 2). As a result, with a well-assembled genome and proper
 365 species training set, the quality of annotation has little effect on synteny analysis,
 366 compared to assembly quality.

367

368 **Table 2. Statistics of *C. elegans* annotations used for synteny analysis.**

	1	2	3
Species	<i>C. elegans</i>		
Assembly source	WormBase	WormBase	WormBase
Annotation info.	WormBase	Augustus + caenorhabditis	Augustus + BUSCO (nematoda)
Genome size (Mb)	100.3	100.3	100.3
Number of genes	20,247	22,930	17,074
Gene coverage (%)	63.0	64.4	55.9
Median gene length (bp)	1,972	1,999	2,149
Median intergenic length (bp)	925	640	1,139
Number of CDS	123,707	126,680	114,640
Median CDS length (bp)	146	146	147
Median CDS sum per gene (bp)	993	882	1041
Number of introns	103,460	103,750	97,566
Median intron length (bp)	63	65	72
Median intron sum per gene (bp)	704	660	967
Syteny coverage of 2 vs. 1 (%)	99.97	99.95	NA
Syteny coverage of 3 vs. 1 (%)	99.95	NA	99.95

369 The statistics that relate to variation of annotation that may play a key role in synteny
 370 detection are highlighted in yellow. The result of synteny detection by DAGchainer is
 371 highlighted in grey.

372

373 Discussion

374

375 Synteny analysis is a practical way to investigate the evolution of genome structure
376 [28–30,55]. In this study, we have revealed how genome assembly contiguity affects
377 synteny analysis. We present a simple scenario of breaking assembly into more
378 fragmented state, which only mimics part of the poor assembly problem. Our genome
379 fragmentation method randomly breaks sequences into pieces with the same size, which
380 gives rise to an almost even distribution of sequence length. This is a simplification of
381 real assemblies, which usually comprise few large sequences and many more tiny
382 sequences that are difficult to assemble because of their repetitive nature [25,26]. It is
383 probable that we overestimate error rate in regions that can be easily assembled but
384 underestimate error rate in regions that will be more fragmented. Note some of the
385 sequences in real assemblies may contain gaps (scaffolds) which will result in more
386 missing genes and will result in further underestimation of synteny. Our result is quite
387 similar when a *de novo* Pacbio *C. elegans* assembly was compared to the reference
388 genome (S1 Table). The use of long read technology is becoming the norm in *de novo*
389 assembly projects. Assemblies with lower contiguation were not compared here as we
390 emphasize the responsibility of research groups to produce assemblies that are of the
391 higher contiguity made possible by long reads [56].

392

393

394 Synteny identification from different programs (*i.e.*, DAGchainer [35], i-ADHoRe
395 [37], MCSanX [36], and SynChro [38]) performed across different species (*C. elegans*,
396 *C. briggsae*, *S. ratti* and *S. stercoralis*) have allowed us to examine the wide-ranging
397 effects of assembly contiguation on downstream synteny analysis. Although the four
398 programs tend to produce the same results when the original assembly is compared to
399 itself, this was no longer the case as assemblies become fragmented. It is interesting to
400 note that DAGchainer and MCSanX use the same scoring algorithm for determining
401 synteny regions, except that DAGchainer uses the number of genes without orthology
402 assignment as gaps while MCSanX uses unaligned nucleotides to do this. When
403 comparing closely related species, results of the four programs fluctuate even without
404 fragmentation in *Caenorhabditis* species, while the pattern remains similar to
405 self-comparison in *Strongyloides* species. Sensitivity of synteny identification drops
406 sharply as the genome assembly becomes fragmented, and thus genome assembly
407 contiguation must be considered when inferring synteny relationships between species.
408 Our fragmentation approach only affects N50, which mostly leads to loss of anchors in
409 synteny analysis. Other scenarios such as unknown sequences (NNNs) in the assembly,
410 or rearrangements causing a break in anchor ordering (Fig 1, break b), or

411 insertions/deletions (Fig 1, break c) were not addressed and may lead to greater
412 inaccuracies.

413

414

415 We have shown that genomic features such as gene density and length of intergenic
416 regions play an essential role during the process of synteny identification (Fig 4 and
417 Table 1; S2 Table). Synteny identification can be established more readily in species with
418 higher gene density or shorter intergenic space, which is related to the initial setting of
419 minimum anchors for synteny identification (Fig 1 and S1 Fig). Repetitiveness of
420 paralogs is another factor in how anchors were chosen from homology assignment. For
421 example, we found that synteny coverage is low along chromosomal arm regions of *C.*
422 *elegans* in both self-comparison and versus *C. briggsae*, which has been reported to have
423 expansion of G protein-coupled receptor gene families [25] (Fig 2 and S6 Fig).
424 Nevertheless, this case may be a result of a combination of repetitive paralogs and high
425 gene density.

426

427 Interestingly, synteny comparison with improved assemblies using ALLMAPS [51]
428 exhibited unexpected variation among programs. Unfortunately, we did not resolve the
429 reason behind sharp decrease of synteny coverage in i-ADHoRe (Fig 6). Nevertheless, we
430 have shown that it is dangerous to improve an assembly using a reference from closely
431 related species without *a priori* information about their synteny relationship. Subsequent
432 synteny identification would be misleading if the same reference was compared again.
433 We also considered the interplay between genome annotation, assembly and synteny
434 identification. Although it may be intuitive to assume lower annotation quality can lead to
435 errors in synteny analysis, we demonstrated that such influence was minimal if an initial
436 genome assembly of good contiguation is available (Table 2).

437

438 In conclusion, this study has demonstrated that a minimum quality of genome
439 assembly is essential to synteny analysis. As a recommendation, to keep the error rate
440 below 5% in synteny identification, we suggest an N50 of 200kb and 1Mb is required
441 when gene density of species of interest are 290 and 200 genes per Mb, respectively
442 (Table 1 and S2 Table). This is a minimum standard and a higher N50 may be required
443 for other species with lower gene density or highly expanded gene families.

444

445

446

447

448 **Materials and Methods**

449

450 **Data Preparation and Fragmentation Simulation**

451

452 Assemblies and annotations of *C. elegans* and *C. briggsae* (release WS255), *S. ratti*
453 and *S. stercoralis* (release WBPS8) were obtained from WormBase
454 (<http://www.wormbase.org/>) [24]. A new assembly of *C. elegans* using long reads was
455 obtained from a Pacific Biosciences Dataset
456 (<https://github.com/PacificBiosciences/DevNet>). Since some genes produce multiple
457 alternative splicing isoforms and all of these isoforms represent one gene (locus), we used
458 the longest isoform as a representative. Further, non-coding genes were also filtered out
459 from our analysis. To simulate the fragmented state of assemblies, a script was made to
460 randomly break scaffolds into fixed size of fragments
461 (<https://github.com/dangliu/Assembly-breaking.git>). Sequences shorter than the fixed
462 length were kept unchanged.

463

464 **Fig 8. Pseudocode of genome assembly fragmentation.**

465

466 **Identification of Synteny Blocks**

467

468 Four different programs were used to identify synteny blocks: DAGchainer [35],
469 i-ADHoRe [37] (v3.0), MCScanX [36] and SynChro [38]. Settings for each program
470 were modified to resemble each other. All of the programs use gene orthology to find
471 anchor points during process of synteny blocks detection. For DAGchainer, i-ADHoRe
472 and MCScanX, we obtained gene orthology from OrthoFinder [57] (v0.2.8). For SynChro,
473 it has an implemented program called OPSCAN to achieve scanning of gene orthology.
474 We arranged the parameters, DAGchainer (accessory script filter_repetitive_matches.pl
475 was run with option 5 before synteny identification as recommended by manual; options:
476 -Z 12 -D 10 -A 5 -g 1), i-ADHoRe (only top 1 hit of each gene in input blast file was
477 used as recommended; options: cluster_type=collinear, alignment_method=gg2,
478 max_gaps_in_alignment=10, tandem_gap=5, gap_size=10, cluster_gap=10, q_value=0.9,
479 prob_cutoff=0.001, anchor_points=5, level_2_only=false), MCScanX (only top 5 hits of
480 each gene in the input blast file was used as suggested; options: default) and SynChro
481 (options: 0 6; 0 for all pairwise, and 6 for delta of RBH genes), for each program. To
482 calculate synteny coverage, the total length of merged synteny blocks along scaffolds was
483 divided by total assembly size.

484

485 **Data analysis**

486

487 Visualization of synteny linkage was made by R (v 3.3.1) and circos [58] (v0.69-4).
488 Gene ontology enrichment analysis was performed using topGO [59] (v1.0) package in R
489 and only focused on Biological Process (options: nodeSize = 3, algorithm = "weight01",
490 statistic = "Fisher"). Gene ontology associations files for *C. elegans* and *C. briggsae*

491 were downloaded from WormBase WS255 [24]. Gene orthology was assigned by
492 OrthoFinder [57]. Then, assemblies were assembled using ALLMAPS [51] with a guided
493 scaffolding approach. *de novo* Annotations of *C. elegans* was predicted using either the
494 manually trained species parameter or from BUSCO [54] (v2.0).

495
496

497 **Acknowledgements**

498
499 We thank John Wang for commenting the manuscript.

500
501

502 **References**

503

- 504 1. Gordon D, Huddleston J, Chaisson MJ, Hill CM, Kronenberg ZN, Munson KM, et
505 al. Long-read sequence assembly of the gorilla genome. *Science*. 2016;352:
506 aae0344. doi:10.1126/science.aae0344
- 507 2. Lien S, Koop BF, Sandve SR, Miller JR, Matthew P, Leong JS, et al. The Atlantic
508 salmon genome provides insights into rediploidization. *Nature*. 2016;533: 200–205.
509 doi:10.1038/nature17164
- 510 3. Iorizzo M, Ellison S, Senalik D, Zeng P, Satapoomin P, Huang J, et al. A
511 high-quality carrot genome assembly provides new insights into carotenoid
512 accumulation and asterid genome evolution. *Nat Genet*. 2016;advance on: 657–666.
513 doi:10.1038/ng.3565
- 514 4. Jarvis DE, Ho YS, Lightfoot DJ, Schmöckel SM, Li B, Borm TJA, et al. The
515 genome of *Chenopodium quinoa*. *Nature*. 2017; 1–6. doi:10.1038/nature21370
- 516 5. Ma L, Chen Z, Huang DW, Kutty G, Ishihara M, Wang H, et al. Genome analysis of
517 three *Pneumocystis* species reveals adaptation mechanisms to life exclusively in
518 mammalian hosts. *Nat Commun*. Nature Publishing Group; 2016;7: 10740.
519 doi:10.1038/ncomms10740
- 520 6. de Man TJB, Stajich JE, Kubicek CP, Teiling C, Chenthamara K, Atanasova L, et al.
521 Small genome of the fungus *Escovopsis weberi*, a specialized disease agent of ant
522 agriculture. *Proc Natl Acad Sci*. 2016;113: 3567–3572.
523 doi:10.1073/pnas.1518501113
- 524 7. Hunt VL, Tsai IJ, Coghlan A, Reid AJ, Holroyd N, Foth BJ, et al. The genomic basis
525 of parasitism in the *Strongyloides* clade of nematodes. *Nat Genet*. Nature Publishing
526 Group; 2016;48: 1–11. doi:10.1038/ng.3495
- 527 8. Cotton JA, Bennuru S, Grote A, Harsha B, Tracey A, Beech R, et al. The genome of
528 *Onchocerca volvulus*, agent of river blindness. *Nat Microbiol*. 2016;2: 16216.
529 doi:10.1038/nmicrobiol.2016.216
- 530 9. Chen X, Tompa M. Comparative assessment of methods for aligning multiple
531 genome sequences. *Nat Biotechnol*. Nature Publishing Group; 2010;28: 567–572.
532 doi:10.1038/nbt.1637

- 533 10. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and
534 genotyping. *Nat Rev Genet.* Nature Publishing Group; 2011;12: 363–76.
535 doi:10.1038/nrg2958
- 536 11. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing:
537 computational challenges and solutions. *Nat Rev Genet.* 2012;46: 36–46.
538 doi:10.1038/nrg3164
- 539 12. Uricaru R, Michotey C, Chiapello H, Rivals E. YOC, A new strategy for pairwise
540 alignment of collinear genomes. *BMC Bioinformatics.* ???; 2015;16: 111.
541 doi:10.1186/s12859-015-0530-3
- 542 13. Ehrlich J, Sankoff D, Nadeau JH. Synteny conservation and chromosome
543 rearrangements during mammalian evolution. *Genetics.* 1997;147: 289–296.
544 doi:10.1159/000322358
- 545 14. Ghiurcuta CG, Moret BME. Evaluating synteny for improved comparative studies.
546 *Bioinformatics.* 2014;30: 9–18. doi:10.1093/bioinformatics/btu259
- 547 15. Renwick JH. The mapping of human chromosome. *Annu Rev Genet.* 1971;5:
548 81–120.
- 549 16. Nadeau JH. Maps of linkage and synteny homologies between mouse and man.
550 *Trends Genet.* 1989; 1–5.
- 551 17. Vergara IA, Chen N. Large synteny blocks revealed between *Caenorhabditis*
552 *elegans* and *Caenorhabditis briggsae* genomes using OrthoCluster. *BMC Genomics.*
553 2010;11: 516. doi:10.1186/1471-2164-11-516
- 554 18. Tang H, Lyons E, Pedersen B, Schnable JC, Paterson AH, Freeling M. Screening
555 synteny blocks in pairwise genome comparisons through integer programming.
556 2011; 1–11.
- 557 19. Schmidt R. Synteny - Recent Advances and Future Prospects. *Curr Opin Plant Biol.*
558 2000;3: 97–102.
- 559 20. Vandepoele K, Saeys Y, Simillion C, Raes J, Van de Peer Y. The automatic
560 detection of homologous regions (ADHoRe) and its application to microcolinearity
561 between *Arabidopsis* and rice. *Genome Res.* 2002;12: 1792–1801.
562 doi:10.1101/gr.400202
- 563 21. Coghlan A, Eichler EE, Oliver SG, Paterson AH, Stein L. Chromosome evolution in
564 eukaryotes: A multi-kingdom perspective. *Trends Genet.* 2005;21: 673–682.
565 doi:10.1016/j.tig.2005.09.009
- 566 22. Molinari NA, Petrov DA, Price HJ, Smith JD, Gold JR, Vassiliadis C, et al. Synteny
567 and Collinearity in Plant Genomes. *Science (80-).* 2008; 486–489. Available:
568 <http://www.sciencemag.org/content/320/5875/486.full.pdf>
- 569 23. Zhang G, Li B, Li C, Gilbert MTP, Jarvis ED, Wang J. Comparative genomic data of
570 the Avian Phylogenomics Project. *Gigascience.* 2014;3: 26.
571 doi:10.1186/2047-217X-3-26
- 572 24. Howe KL, Bolt BJ, Cain S, Chan J, Chen WJ, Davis P, et al. WormBase 2016:
573 Expanding to enable helminth genomic research. *Nucleic Acids Res.* 2016;44:
574 D774–D780. doi:10.1093/nar/gkv1217
- 575 25. Consortium TC *elegans* S. Genome Sequence of the Nematode *C. elegans*: A
576 Platform for Investigating Biology. *Science (80-).* 1998;282: 2012–2018.
577 doi:10.1126/science.282.5396.2012

- 578 26. Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, Chen N, et al. The genome
579 sequence of *Caenorhabditis briggsae*: A platform for comparative genomics. *PLoS*
580 *Biol.* 2003;1. doi:10.1371/journal.pbio.0000045
- 581 27. Wong S, Wolfe KH. Birth of a metabolic gene cluster in yeast by adaptive gene
582 relocation. *Nat Genet.* 2005;37: 777–782. doi:10.1038/ng1584
- 583 28. Lemons D, McGinnis W. Genomic evolution of Hox gene clusters. *Science* (80-).
584 2006/09/30. 2006;313: 1918–1922. doi:10.1126/science.1132040
- 585 29. Ruelens P, de Maagd RA, Proost S, Theißen G, Geuten K, Kaufmann K.
586 FLOWERING LOCUS C in monocots and the tandem origin of
587 angiosperm-specific MADS-box genes. *Nat Commun.* 2013;4: 2280.
588 doi:10.1038/ncomms3280
- 589 30. Kemkemer C, Kohn M, Cooper DN, Froenicke L, Högel J, Hameister H, et al. Gene
590 synteny comparisons between different vertebrates provide new insights into
591 breakage and fusion events during mammalian karyotype evolution. *BMC Evol Biol.*
592 2009;9: 84. doi:10.1186/1471-2148-9-84
- 593 31. Murat F, Armero A, Pont C, Klopp C, Salse J. Reconstructing the genome of the
594 most recent common ancestor of flowering plants. *Nat Genet.* Nature Publishing
595 Group; 2017;49: 490–496. doi:10.1038/ng.3813
- 596 32. Denton JF, Lugo-Martinez J, Tucker AE, Schridder DR, Warren WC, Hahn MW.
597 Extensive Error in the Number of Genes Inferred from Draft Genome Assemblies.
598 *PLoS Comput Biol.* 2014;10. doi:10.1371/journal.pcbi.1003998
- 599 33. Dupont P-Y, Cox MP. Genomic Data Quality Impacts Automated Detection of
600 Lateral Gene Transfer in Fungi. *G3 (Bethesda).* 2017;7: g3.116.038448.
601 doi:10.1534/g3.116.038448
- 602 34. Batzoglou S. The many faces of sequence alignment. *Brief Bioinform.* 2005;6: 6–22.
603 doi:10.1093/bib/6.1.6
- 604 35. Haas BJ, Delcher AL, Wortman JR, Salzberg SL. DAGchainer: A tool for mining
605 segmental genome duplications and synteny. *Bioinformatics.* 2004;20: 3643–3646.
606 doi:10.1093/bioinformatics/bth397
- 607 36. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, et al. MCScanX: A toolkit for
608 detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids*
609 *Res.* 2012;40: 1–14. doi:10.1093/nar/gkr1293
- 610 37. Proost S, Fostier J, De Witte D, Dhoedt B, Demeester P, Van De Peer Y, et al.
611 i-ADHoRe 3.0-fast and sensitive detection of genomic homology in extremely large
612 data sets. *Nucleic Acids Res.* 2012;40: 1–11. doi:10.1093/nar/gkr955
- 613 38. Drillon G, Carbone A, Fischer G. SynChro: A fast and easy tool to reconstruct and
614 visualize synteny blocks along eukaryotic chromosomes. *PLoS One.* 2014;9: 1–8.
615 doi:10.1371/journal.pone.0092621
- 616 39. Ross JA, Koboldt DC, Staisch JE, Chamberlin HM, Gupta BP, Miller RD, et al.
617 *Caenorhabditis briggsae* recombinant inbred line genotypes reveal inter-strain
618 incompatibility and the evolution of recombination. *PLoS Genet.* 2011;7.
619 doi:10.1371/journal.pgen.1002174
- 620 40. Bhutkar A, Russo S, Smith TF, Gelbart WM. Techniques for multi-genome synteny
621 analysis to overcome assembly limitations. *Genome Inform.* 2006;17: 152–161.
622 Available:
623 <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=175033>

- 624 88&retmode=ref&cmd=prlinks%5Cnpapers3://publication/uuid/7717ABDA-5CC
625 B-48C2-9AF7-C51B12BDEAF8
- 626 41. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of
627 next-generation sequencing technologies. *Nat Rev Genet.* Nature Publishing Group;
628 2016;17: 333–351. doi:10.1038/nrg.2016.49
- 629 42. Treangen TJ, Ondov BD, Koren S, Phillippy AM. The Harvest suite for rapid
630 core-genome alignment and visualization of thousands of intraspecific microbial
631 genomes. *Genome Biol.* 2014;15: 524. doi:10.1186/s13059-014-0524-x
- 632 43. Viney ME. The biology and genomics of Strongyloides. *Med Microbiol Immunol.*
633 2006;195: 49–54. doi:10.1007/s00430-006-0013-2
- 634 44. Ward JD. Rendering the intractable more tractable: Tools from caenorhabditis
635 elegans ripe for import into parasitic nematodes. *Genetics.* 2015. pp. 1279–1294.
636 doi:10.1534/genetics.115.182717
- 637 45. Armengol L, Marquès-Bonet T, Cheung J, Khaja R, González JR, Scherer SW, et al.
638 Murine segmental duplications are hot spots for chromosome and gene evolution.
639 *Genomics.* 2005;86: 692–700. doi:10.1016/j.ygeno.2005.08.008
- 640 46. Davidson RM, Gowda M, Moghe G, Lin H, Vaillancourt B, Shiu SH, et al.
641 Comparative transcriptomics of three Poaceae species reveals patterns of gene
642 expression evolution. *Plant J.* 2012;71: 492–502.
643 doi:10.1111/j.1365-313X.2012.05005.x
- 644 47. Lovell P V, Wirthlin M, Wilhelm L, Minx P, Lazar NH, Carbone L, et al. Conserved
645 syntenic clusters of protein coding genes are missing in birds. *Genome Biol.* 2014;
646 1–27. doi:10.1186/s13059-014-0565-1
- 647 48. Baldauf J, Marcon C, Paschold A, Hochholdinger F. Nonsyntenic genes drive
648 tissue-specific dynamics of differential, nonadditive and allelic expression patterns
649 in maize hybrids. *Plant Physiol.* 2016;171: pp.00262.2016.
650 doi:10.1104/pp.16.00262
- 651 49. Assefa S, Keane TM, Otto TD, Newbold C, Berriman M. ABACAS:
652 Algorithm-based automatic contiguation of assembled sequences. *Bioinformatics.*
653 2009;25: 1968–1969. doi:10.1093/bioinformatics/btp347
- 654 50. Husemann P, Stoye J. r2cat: Synteny plots and comparative assembly.
655 *Bioinformatics.* 2009;26: 570–571. doi:10.1093/bioinformatics/btp690
- 656 51. Tang H, Zhang X, Miao C, Zhang J, Ming R, Schnable JC, et al. ALLMAPS: robust
657 scaffold ordering based on multiple maps. *Genome Biol.* 2015;16: 3.
658 doi:10.1186/s13059-014-0573-1
- 659 52. Howe KL, Bolt BJ, Cain S, Chan J, Chen WJ, Davis P, et al. WormBase 2016:
660 Expanding to enable helminth genomic research. *Nucleic Acids Res.* 2016;44:
661 D774–D780. doi:10.1093/nar/gkv1217
- 662 53. Stanke M, Steinkamp R, Waack S, Morgenstern B. AUGUSTUS: a web server for
663 gene finding in eukaryotes. *Nucleic Acids Res.* 2004;32: W309–W312.
664 doi:10.1093/nar/gkh379
- 665 54. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V. BUSCO: assessing
666 genome assembly and annotation completeness with single-copy orthologs.
667 *Genome Anal.* 2015;31: 9–10. doi:10.1093/bioinformatics/btv351
- 668 55. Hillier LW, Miller RD, Baird SE, Chinwalla A, Fulton LA, Koboldt DC, et al.
669 Comparison of *C. elegans* and *C. briggsae* genome sequences reveals extensive

- 670 conservation of chromosome organization and synteny. PLoS Biol. 2007;5:
671 1603–1616. doi:10.1371/journal.pbio.0050167
- 672 56. Chain PSG, Grafham D V, Fulton RS, Fitzgerald MG, Hostetler J, Muzny D, et al.
673 Genome Project Standards in a New Era of Sequencing. Science. 2009;326: 4–5.
674 doi:10.1126/science.1180614
- 675 57. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome
676 comparisons dramatically improves orthogroup inference accuracy. Genome Biol.
677 Genome Biology; 2015;16: 157. doi:10.1186/s13059-015-0721-2
- 678 58. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos:
679 an information esthetic for comparative genomics. Genome Res. 2009;19:
680 1639–1645. doi:10.1101/gr.092759.109
- 681 59. Alexa A and Rahnenfuhrer J. topGO: Enrichment Analysis for Gene Ontology. In:
682 R package version 2.26.0. [Internet]. 2016. Available:
683 <http://bioconductor.org/packages/release/bioc/html/topGO.html>
684
685
686

687 **Supporting information**

688

689 **S1 Fig. Synteny coverage in different number of minimum anchor using**
690 **DAGchainer.** The Y axis shows synteny coverage (%). The X axis is the number of
691 minimum anchor to identify a synteny block from 2 to 8. The 4 colors are 4 combinations
692 of synteny detection among species: *C. elegans* vs. *C. elegans* (CEvsCE, green), *C.*
693 *elegans* vs. *C. briggsae* (CEvsCBG, orange), *S. ratti* vs. *S. ratti* (SRvsSR, blue) and
694 *S.ratti* vs. *S. stercoralis* (SRvsSS, purple).

695

696 **S2 Fig. Synteny blocks in *C. elegans* versus *C. elegans*.** Chromosomes are separated
697 into panels with Roman number labels. The Y axis stands for categories of distribution.
698 Synteny blocks defined by four detection programs: DAGchainer (red), i-ADHoRe
699 (yellow), MCScanX (green) and SynChro (light blue) are drawn as rectangles.
700 Distribution of genes is the bottom smaller rectangles in burgundy. The X axis is the
701 position of chromosome.

702

703 **S3 Fig. Synteny blocks in *C. elegans* versus 1Mb fragmented *C. elegans*.**
704 Chromosomes are separated into panels with Roman number labels. The Y axis stands for
705 categories of distribution. Synteny blocks defined by four detection programs:
706 DAGchainer (red), i-ADHoRe (yellow), MCScanX (green) and SynChro (light blue) are
707 drawn as rectangles. Distribution of genes is the bottom smaller rectangles in burgundy.
708 The X axis is the position of chromosome.

709

710 **S4 Fig. Synteny blocks in *C. elegans* versus 100kb fragmented *C. elegans*.**

711 Chromosomes are separated into panels with Roman number labels. The Y axis stands for
712 categories of distribution. Synteny blocks defined by four detection programs:
713 DAGchainer (red), i-ADHoRe (yellow), MCScanX (green) and SynChro (light blue) are
714 drawn as rectangles. Distribution of genes is the bottom smaller rectangles in burgundy.
715 The X axis is the position of chromosome.

716

717 **S5 Fig. A zoomed-in region of synteny identification with lower gap threshold in**
718 **MCScanX between original *C. elegans* and 100kb fragmented assembly.** The Y axis
719 stands for categories of distribution. Synteny blocks defined by four detection programs:
720 DAGchainer (red), i-ADHoRe (yellow), MCScanX (green) and SynChro (light blue) are
721 drawn as rectangles. Fragmented sites are labeled by vertical red dashed lines.
722 Distribution of genes in burgundy rectangles is marked with dark blue lines as gene starts.
723 The X axis is the position of chromosome. Scenario (a) is that synteny block was
724 identified after gap threshold tuned lower.

725

726 **S6 Fig. Synteny blocks in *C. elegans* versus *C. briggsae*.** Chromosomes are separated
727 into panels with Roman number labels. The Y axis stands for categories of distribution.
728 Synteny blocks defined by four detection programs: DAGchainer (red), i-ADHoRe
729 (yellow), MCScanX (green) and SynChro (light blue) are drawn as rectangles. The
730 bottom four categories are orthologs between the two species assigned by Opscan (OP;
731 burgundy) and OrthoFinder (OF; deep blue), and we further categorized orthologs into 1
732 to 1 orthology (1-1) or many to many orthology (N-N). The X axis is the position of
733 chromosome.

734

735 **S1 Table. Statistics of annotation and synteny coverage using WormBase *C. elegans***
736 **versus PacBio *C. elegans*.** Yellow highlights the statistics that relate to variation of
737 annotation that may play a key role in synteny detections. Grey highlights the result of
738 synteny detections by DAGchainer. Assembly source in column 2 is obtained from
739 Pacific Bioscienceces Dataset. Annotation information in column 2 is predicted by
740 Augustus using implanted caenorhabditis (*elegans*) species data set.

741 **S2 Table. Quantification of synteny coverage and error rate.**

742 **S3 Table. Gene ontology (GO) enrichment analysis of *C. briggsae* genes in synteny**
743 **break between *C. elegans* and 1Mb fragmented *C. briggsae* assemblies.** GO terms that
744 appeared in the top 10 ranks either in the original comparison or after when assemblies
745 were fragmented, are displayed. The original rank, median rank and number of
746 occurrences that reached top 10 in 100 replications are shown for each GO term. GO
747 terms not belonging to original assembly but reached top 10 after fragmentation are
748 shaded in green.

749 **S4 Table. Gene ontology (GO) enrichment analysis of *C. briggsae* genes in synteny**
750 **break between *C. elegans* and 500kb fragmented *C. briggsae* assemblies.** GO terms

751 that appeared in the top 10 ranks either in the original comparison or after when
752 assemblies were fragmented, are displayed. The original rank, median rank and number
753 of occurrences that reached top 10 in 100 replications are shown for each GO term. GO
754 terms not belonging to original assembly but reached top 10 after fragmentation are
755 shaded in green.

756 **S5 Table. Gene ontology (GO) enrichment analysis of *C. briggsae* genes in synteny**
757 **break between *C. elegans* and 200kb fragmented *C. briggsae* assemblies.** GO terms
758 that appeared in the top 10 ranks either in the original comparison or after when
759 assemblies were fragmented, are displayed. The original rank, median rank and number
760 of occurrences that reached top 10 in 100 replications are shown for each GO term. GO
761 terms not belonging to original assembly but reached top 10 after fragmentation are
762 shaded in green. GO:0043066 was in the original top 10 rank but failed to reach top 10 in
763 all of 100 replications.

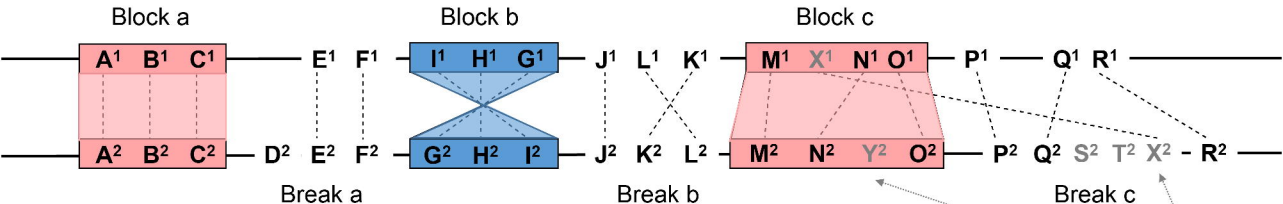
764 **S6 Table. Gene ontology (GO) enrichment analysis of *C. briggsae* genes in synteny**
765 **break between *C. elegans* and 100kb fragmented *C. briggsae* assemblies.** GO terms
766 that appeared in the top 10 ranks either in the original comparison or after when
767 assemblies were fragmented, are displayed. The original rank, median rank and number
768 of occurrences that reached top 10 in 100 replications are shown for each GO term. GO
769 terms not belonging to original assembly but reached top 10 after fragmentation are
770 shaded in green. GO:0043066 was in the original top 10 rank but failed to reach top 10 in
771 all of 100 replications.

772 **S7 Table. Assembly statistics among *Caenorhabditis* species and *Strongyloides***
773 **species including ALLMAPS results.**

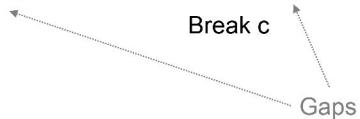
774

775

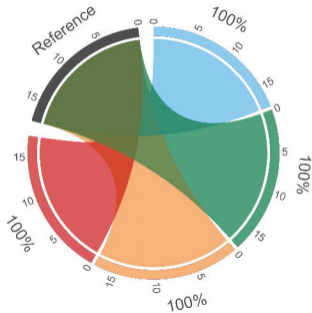
Species 1



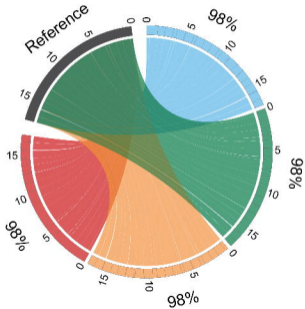
Species 2



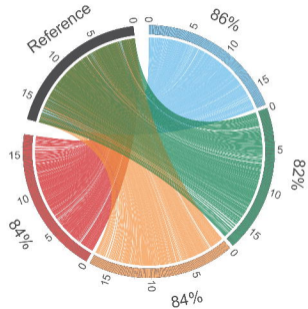
Un-fragmented

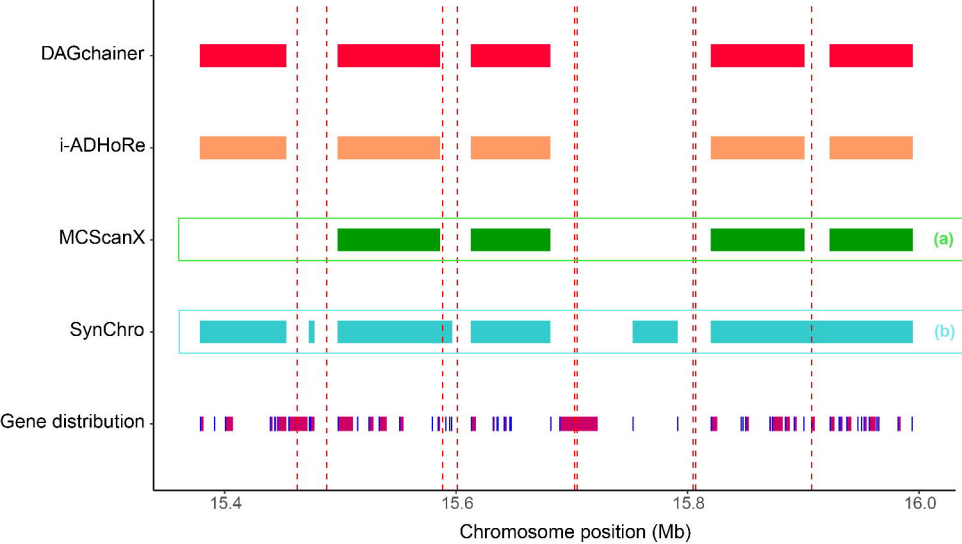


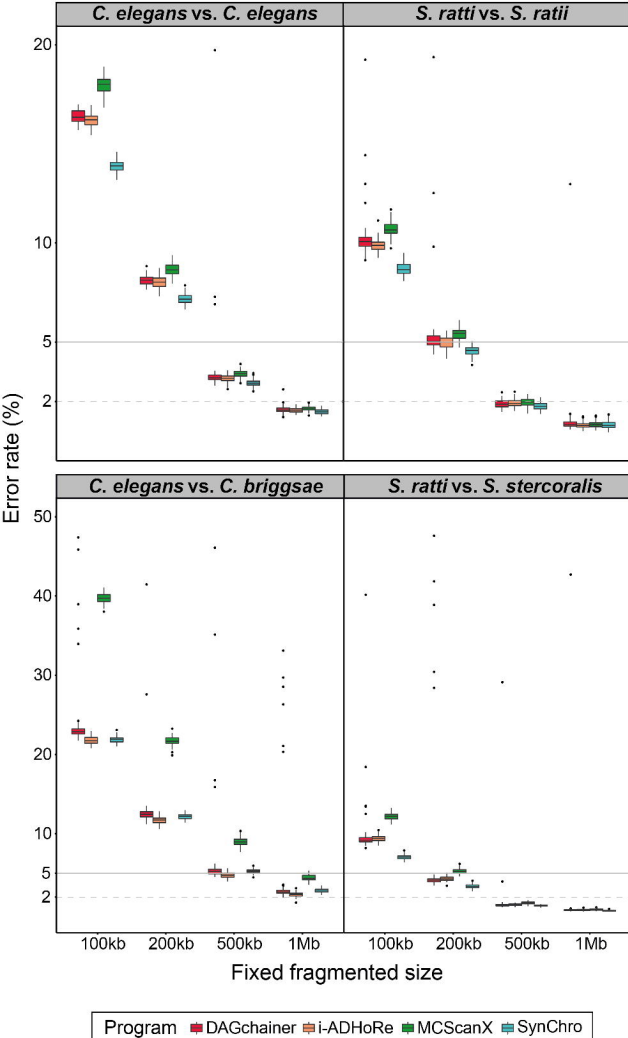
1Mb fragmented



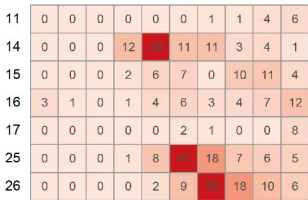
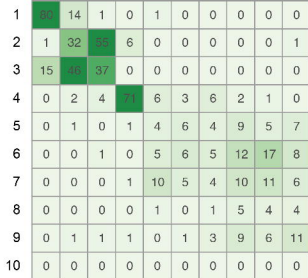
100kb fragmented





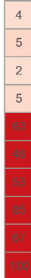


Rank before fragmentation



1 2 3 4 5 6 7 8 9 10 Out of top 10

Rank after fragmentation

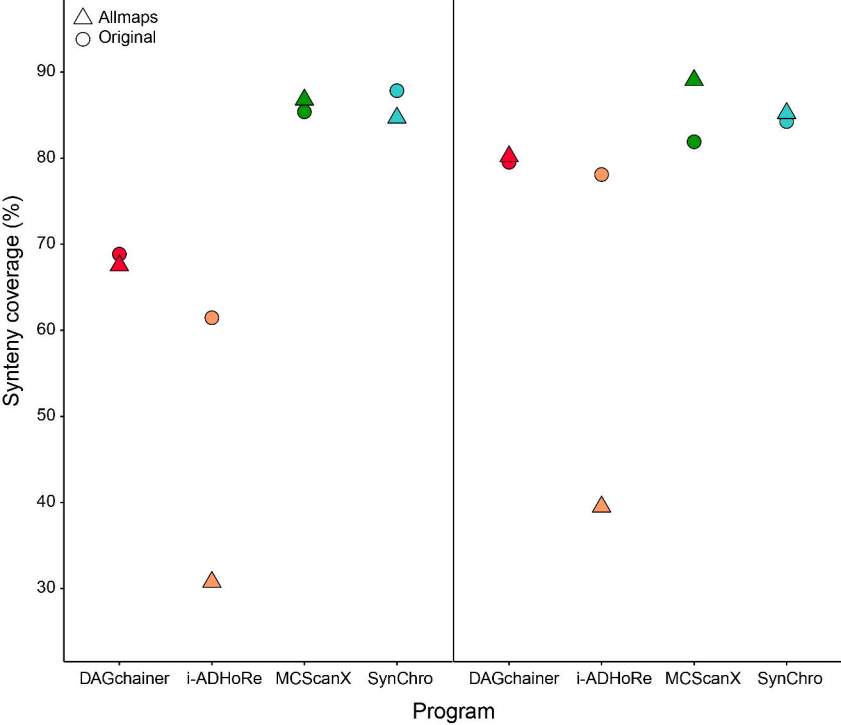


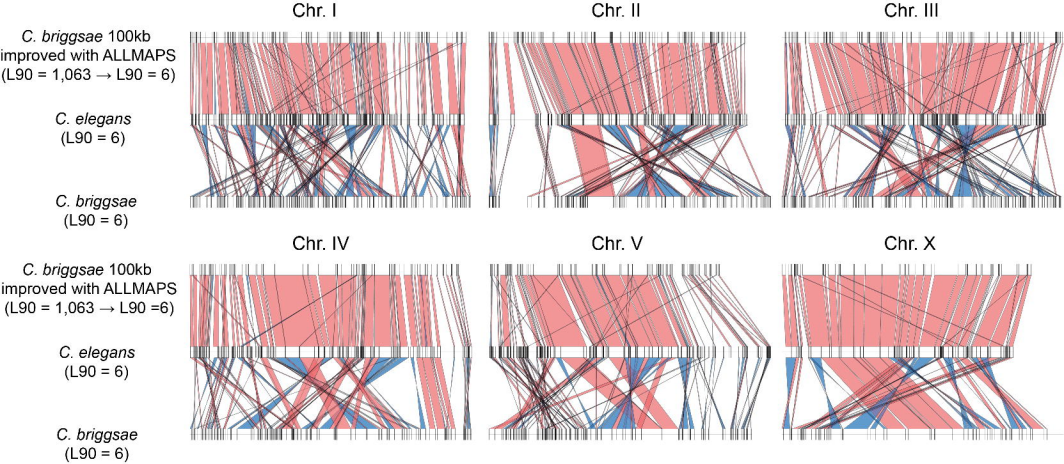
GO:0006355	Regulation of transcription, DNA-templated
GO:0043401	Steroid hormone mediated signaling pathway
GO:0015074	DNA integration
GO:0006032	Chitin catabolic process
GO:0007606	Sensory perception of chemical stimulus
GO:0016573	Histone acetylation
GO:0006468	Protein phosphorylation
GO:0005975	Carbohydrate metabolic process
GO:0006310	DNA recombination
GO:0043066	Negative regulation of apoptotic process

GO ID

GO term

GO:0034729	Histone H3-K79 methylation	86
GO:0000122	Negative regulation of transcription from RNA polymerase II promoter	10
GO:0008152	Metabolic process	60
GO:0006470	Protein dephosphorylation	39
GO:0007155	Cell adhesion	35
GO:0001709	Cell fate determination	14
GO:0045596	Negative regulation of cell differentiation	15

*C. elegans vs. C. briggsae**S. ratti vs. S. stercoralis*



Algorithm 1 Genome assembly fragmentation

original \leftarrow a dictionary containing original sequences of an assembly
n = fixed fragmentation size
result \leftarrow an empty list for collecting fragments information
while sequence **in** *original* **do**
 seq \leftarrow **random** sequence **from** *original*
 if $\text{len}(seq) < n$ **then**
 move *seq* **to** *result*
 else
 x \leftarrow **random** number **from** 1 to $\text{len}(seq) - n$
 newseq \leftarrow *seq* from *x* to *x* + *n*
 seq.left \leftarrow *seq* from 1 to *x* - 1
 seq.right \leftarrow *seq* from *x* + *n* + 1 to $\text{len}(seq)$
 move *newseq* **to** *result*
 move *seq.left* , *seq.right* **to** *original*
 delete *seq* **from** *original*
print *result*
