



HAL
open science

The black honey bee genome: insights on specific structural elements and a first step towards pan-genomes

Sonia Eynard, Christophe Klopp, Kamila Canale-Tabet, William Marande, Céline Vandecasteele, Céline Roques, Cécile Donnadiou, Quentin Boone, Bertrand Servin, Alain Vignal

► To cite this version:

Sonia Eynard, Christophe Klopp, Kamila Canale-Tabet, William Marande, Céline Vandecasteele, et al.. The black honey bee genome: insights on specific structural elements and a first step towards pan-genomes. 2024. hal-04473386v1

HAL Id: hal-04473386

<https://hal.inrae.fr/hal-04473386v1>

Preprint submitted on 23 Feb 2024 (v1), last revised 12 Jul 2024 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NoDerivatives 4.0 International License

1 **The black honey bee genome: insights on specific structural** 2 **elements and a first step towards pan-genomes**

3

4 Sonia E. Eynard¹, Christophe Klopp², Kamila Canale-Tabet¹, William Marande³, Céline
5 Vandecasteele⁴, Céline Roques⁴, Cécile Donnadiou⁴, Quentin Boone^{1,2}, Bertrand Servin¹ and Alain
6 Vignal^{1*}

7 ¹GenPhySE, Université de Toulouse, INRAE, INPT, INP-ENVT, Castanet Tolosan, France

8 ²Sigenae, MIAT, INRAE, Castanet Tolosan, France

9 ³CNRGV, INRAE, Castanet Tolosan, France

10 ⁴INRAE, US 1426, GeT-PlaGe, Genotoul, Castanet-Tolosan, France

11

12 *Corresponding author

13

14 E-mail addresses:

15 SE: sonia.eynard@inrae.fr

16 CK: christophe.klopp@inrae.fr

17 KC-T: kamila.tabet@inrae.fr

18 WM: william.marande@inrae.fr

19 CV: celine.vandecasteele@inrae.fr

20 CR: celine.lopez-roques@inrae.fr

21 QB: quentin.boone@inrae.fr

22 CD: cecile.donnadiou@inrae.fr

23 BS: bertrand.servin@inrae.fr

24 AV: alain.vignal@inrae.fr

25

26 **Abstract** (maximum 350 words).

27 **Background**

28 The actual honey bee reference genome, HAv3.1, was produced from a commercial line sample,
29 thought to have a largely dominant *Apis mellifera ligustica* genetic background. *Apis mellifera*
30 *mellifera*, often referred to as the black bee, has a separate evolutionary history and is the original
31 type in western and northern Europe. Growing interest in this subspecies for conservation and non-
32 professional apicultural practices, together with the necessity of deciphering genome backgrounds
33 in hybrids, triggered the necessity for a specific genome assembly. Moreover, having several high-
34 quality genomes is becoming key for taking structural variations into account in pan-genome
35 analyses.

36 **Results**

37 Pacific Bioscience technology long reads were produced from a single haploid black bee drone.
38 Scaffolding contigs into chromosomes was done using a high-density genetic map. This allowed for
39 a re-estimation of the honey recombination rate, over-estimated in some previous studies, due to
40 mis-assemblies resulting in spurious inversions in the older reference genomes. The sequence
41 continuity obtained is very high and the only limit towards continuous chromosome-wide sequences
42 seem to be due to tandem repeat arrays usually longer than 10 kb and belonging to two main
43 families, the 371 and 91 bp repeats, causing problems in the assembly process due to high internal
44 sequence similarity. Our assembly was used together with the reference genome, for genotyping
45 two structural variants by a pan-genome graph approach with GraphTyper2. Genotypes obtained
46 were either correct or missing, when compared to an approach based on sequencing depth analysis,
47 and genotyping rates were 89 and 76 % for the two variants respectively.

48 **Conclusions**

49 Our new assembly for the *Apis mellifera mellifera* honey bee subspecies demonstrates the utility of
50 multiple high-quality genomes for the genotyping of structural variants, with a test case on two
51 insertions and deletions. It will therefore be an invaluable resource for future studies, for instance
52 including structural variants in GWAS. Having used a single haploid drone for sequencing allowed
53 a refined analysis of very large tandem repeat arrays, raising the question of their function in the
54 genome. High quality genome assemblies for multiple subspecies such as presented here, are crucial
55 for emerging projects using pan-genomes.

56

57 **Background**

58 The honey bee *Apis mellifera* was originally found in Europe, Africa and the Middle East, with the
59 most eastern limit of its natural distribution situated in western Afghanistan until a new subspecies
60 was discovered in Kazakhstan [1]. The evolutionary origin of *Apis mellifera* is still unclear, with a
61 possible origin in Eastern Africa or the Middle East, followed by the colonization of Europe
62 through different routes, leading to high genetic differentiation between geographically close
63 populations or subspecies, namely *A. m. mellifera* (otherwise referred to as M-type) in western
64 Europe on one side and *A. m. ligustica* from Italy or *A. m. carnica* (known as C-type) from eastern
65 Europe on the other [2–5]. However, although *A. m. mellifera* is the original subspecies found in
66 western Europe, it has become commonplace amongst breeders, in order to increase production or
67 to facilitate the handling of colonies, to import other subspecies, mainly *A. m. ligustica* from Italy,
68 *A. m. carnica* from Slovenia and *A. m. caucasica* from Georgia, either to be bred as pure lines or as
69 hybrids generated by artificial or directed insemination [6,7]. As a consequence, these imported
70 subspecies and hybrid lines will mate naturally to local *A. m. mellifera* populations, threatening
71 them and prompting the establishment of conservation programmes [8]. However, although it has

72 been replaced in the majority of large professional beekeeper's facilities by imported honey bees, *A.*
73 *m. mellifera* is still used by dedicated breeders.

74 The honey bee reference genome, whose first version was obtained as soon as 2006 [9], was
75 updated twice: a first time in 2014 [10] and a second time in 2019, using long-read sequencing
76 together with Hi-C chromatin interaction and BioNano Optical maps for a chromosome-scale
77 assembly [11]. The sample used for this reference genome is from a commercial line (DH4), which
78 is not precisely genetically defined, but is thought to be mainly of *A. m. ligustica* descent [9]. As a
79 consequence, the genome of the genetically distinct *A. m. mellifera* may not be accurately
80 represented and future pangenome approaches, that were shown in other species to expand the
81 number of genomic regions available for analysis [12,13], would benefit from a high-quality
82 assembly for this important subspecies.

83 To ensure a faithful representation of the *A. m. mellifera* subspecies genetic background, an individual from the
84 black bee conservatory "Association Conservatoire de l'Abeille Noire Bretonne" in the island of Ouessant,
85 France was selected for sequencing. This very small island (15.5 km²) is located 20 km off the coast of Brittany,
86 the conservation population was set up starting in 1987, and further imports of other honey bees banned since
87 1991. Mitochondrial DNA analyses have shown a low haplotype diversity and the presence of only the M-type
88 in this population [14]. As expected from such a small population, microsatellite analysis has shown a low di-
89 versity [15].

90 Until the latest update [11], the current honey bee genome sequence, Amel4.5 [10] suffered from imperfections,
91 having numerous gaps in the assembly and possible sequence inversions. In order to construct a new *A. m. mel-*
92 *lifera* genome assembly with improved continuity, we used the Pacific Biosciences long-read technology and
93 produced all sequence reads from a single haploid drone to avoid assembly problems due to polymorphism. To

94 order and orient our contigs along the chromosomes, we used published sequencing reads from drones originat-
95 ing from three colonies that had previously been used to map meiotic crossovers and non-crossovers in the
96 honey bee [16], allowing also for the production of an updated genetic map and a re-estimation of the honey
97 bee recombination rate.
98 Our analyses of the assembly allowed the detection of a major family of tandem repeats, running in some in-
99 stances over more than 10 kb and found at the ends of most sequence contigs. Our assembly allows for the
100 first-time to perform detailed analyses of structural rearrangements, including at the population level, between
101 the genomes of *A. m. ligustica* and other C-type honey bees used by the majority of beekeepers and that of the
102 M-type subspecies *A. m. mellifera* black bee.

103 **Methods**

104 *Sampling, DNA extraction and PacBio long-read sequencing.*

105 Candidate drones for sequencing were sampled at the larval or pupae stage from the black bee
106 conservatory on the island of Ouessant, Brittany, France and extractions were performed from
107 several samples, to select the best DNA quality in terms of molecular weight and quantity. Each
108 sample was ground using a potter (see Additional file 1: Fig. S1) and DNA extraction performed
109 using the QIAGEN Genomic-tips 100/G kit (Cat No./ID: 10243), following the tissue protocol
110 extraction (see supplementary methods). DNA for sequencing was obtained from a single drone
111 OUE7B (see Additional file 1: Fig. S2). Library preparation and sequencing were performed at the
112 GeT-PlaGe core facility, INRAE Toulouse, following the manufacturer's instructions for "Shared
113 protocol-20kb Template Preparation Using BluePippin Size Selection system (15kb size Cutoff)".
114 At each step, DNA was quantified using the Qubit dsDNA HS Assay Kit (Life Technologies). DNA
115 purity was tested using the nanodrop (Thermofisher) and size distribution and degradation assessed
116 using the Fragment analyzer (AATI) High Sensitivity Large Fragment 50kb Analysis Kit.

117 Purification steps were performed using 0.45X AMPure PB beads (PacBio). Thirty μg of DNA was
118 purified to perform 3 libraries. Using SMRTBell template Prep Kit 1.0 (PacBio), a DNA and END
119 damage repair step was performed on 15 μg of unshared sample. Then blunt hairpin adapters were
120 ligated to the libraries. The libraries were treated with an exonuclease cocktail to digest unligated
121 DNA fragments. A size selection step using a 7kb (Library 1) or 9kb (libraries 2 and 3) cutoff was
122 performed on the BluePippin Size Selection system (Sage Science) with the 0.75% agarose
123 cassettes, Marker S1 high Pass 15-20kb. Conditioned Sequencing Primer V2 was annealed to the
124 size-selected SMRTbells. The annealed libraries were then bound to the P6-C4 polymerase using a
125 ratio of polymerase to SMRTbell at 10:1. Then after a magnetic bead-loading step (OCPW),
126 SMRTbell libraries were sequenced on 36 SMRTcells on a RSII instrument from 0.05 to 0.2 nM
127 with a 360 min movie.

128 *Assembly into contigs and alignment to Amel4.5 for chromosome assignments.*

129 Raw reads were assembled with Canu 1.3 [17] using standard parameters and a first polishing of the
130 assembly was done with quiver (version SMRT_Link v4.0.0) using standard parameters. The
131 contigs obtained after the assembly step were aligned to the Amel4.5 reference genome using LAST
132 v956 [18].

133 *Alignment of Illumina sequencing reads and SNP calling for crossing over analysis.*

134 All the Illumina paired-end sequences from Liu et al. (2015) [16] were downloaded from the NCBI
135 SRA project SRP043350 (see Additional file 2: Table S1). The reads were aligned to the assembled
136 contigs with BWA MEM v0.7.15 [19], duplicate reads removed with Picard (v2.1.1;
137 <http://picard.sourceforge.net>), and local realignment and base quality score recalibration (BQSR)
138 performed using GATKv3.7 [19]. SNPs were called in each drone independently with GATK
139 HaplotypeCaller and consolidated into a single set of master sites, from which all individuals were

140 genotyped with GATK GenotypeGVCFs (see scripts in supplementary material). Any SNP with
141 missing genotypes were filtered out. Further quality controls were applied and for each colony,
142 SNPs falling into any of the following categories were discarded: i) non-polymorphic SNPs in the
143 colony, ii) homozygous SNPs in the queen, iii) heterozygous SNPs in drones, iv) SNPs that
144 appeared inconsistent with the observations in the two other colonies and v) SNPs showing
145 inconsistent allelic versions between queen and drone genotypes.

146 *Phasing and detection of recombination events.*

147 For each colony and informative SNP, genotyping results were used to define genotype vectors
148 across all drones for the colony. Identical genotype vectors following one another within a same
149 contig define a segment with no observed crossing over in the drones of the colony and were
150 grouped into bins. Not having access to grand-parental genotypes, genotype phase between two
151 successive bins within a contig was determined by finding which out of the two possible inverse
152 vectors minimised the number of recombination events. Non-crossing over gene conversion events,
153 which can be misinterpreted as double recombination events, occurring usually on short DNA
154 fragment often considered shorter than a few kb, [16] were removed to avoid inflating the size of
155 the genetic map. In our study, non-crossing over gene conversion events were identified as: i) bins
156 of length shorter than 2 kb, occurring between two identical bins, or ii) bins of length shorter than 2
157 kb for which the number of recombination events happening within this bin is higher than the
158 number of recombination events needed to go from the bin before to the bin after it. Bins detected
159 as non-crossing over gene conversions were merged with their two identical surrounding bins. Both
160 phasing and putative non-crossing over identification were performed iteratively from one bin to the
161 next and independently for each colony. As a consequence, a set of phased vectors minimising
162 recombination events was obtained for each contig in each colony.

163 *Scaffolding contigs into chromosomes.*

164 Using the *a priori* assignment of contigs to chromosomes by alignment to Amel4.5 as a starting
165 point, contigs were ordered and oriented iteratively in order to minimise the number of
166 recombination events between the genotype vectors defined at their extremities. The contig
167 scaffolding was first performed using the data for each colony separately and was thereafter
168 confirmed using markers informative across all three colonies.

169 *Correction of the assembly with Illumina reads.*

170 Genomic DNA from the same individual used for the PacBio sequencing was sequenced with an
171 Illumina NovaSeq6000 instrument, producing over 28 000 000 reads (estimated raw sequencing
172 depth = 37 X), NCBI SRA accession SRR15173860. These were aligned on the assembled genome
173 with BWA MEM version 0.7.12-r1039 [20] using standard parameters. Variant detection was done
174 with freebayes version 1.1.0 [21] and filtered to retain only those with a minimum quality score of
175 20 and '1/1' genotype or '0/1' with no read supporting the reference allele. Finally, corrections to the
176 genome assembly were done when alternative alleles were found in the VCF file using vcf-
177 consensus from the vcftools package (version 0.1.12a) [22] with standard parameters.

178 *Comparison with Amel 4.5 and HAv3.1 assemblies.*

179 Estimation of recombination rate and positioning recombination events along the Amel4.5 and
180 AMelMel1.1 assemblies was done following the same procedure as for the de-novo assembly. GC
181 content and sequence coverage for the queens' genotypes in AMelMel1.1 were measured in 0.5Mb
182 windows and the recombination rates were estimated using a script from Petit et al. (2017) [23] over
183 1Mb windows. Completeness of the assemblies was estimated with BUSCO 3.0.2 [24] using
184 OrthoDB v9.1 single-copy orthologs [25], from the Metazoa (n=978) and Hymenoptera (n=4415)

185 BUSCO core set. Alignments of AMelMel1.1 to Amel4.5 and to HAv3.1 were done using LAST
186 v956 [18]. Standard output psl files were produced to keep all alignments related to repeat elements,
187 together with psl files from split alignments [18], corresponding to one-to-one alignments. Dotplot
188 visualisation of alignments were produced with custom scripts. Inversions between the two genome
189 assemblies were detected in the split alignment psl file. Liftovers of the HAv3.1 gtf and gff
190 annotation to produce files with AMelMel1.1 annotation coordinates were done using CrossMap
191 [26] and the chained alignment format output from the AMelMel1.1 to HAv3.1 LAST alignments.

192 *Analysis of repeat elements.*

193 Analysis of tandem repeats was done with Tandem Repeat Finder v4.09 (TRF) [27], setting the
194 maximum period size to 2000 bp. The two major classes of repeat sizes: the 91 bp repeat and the
195 371 bp repeat were analysed by aligning all repeats within a class size with MAFFT v7.313 [28].
196 Sequences reported by TRF from different parts of the genome start at different positions of the
197 repeated element detected and as a consequence, the multifasta alignments produced by MAFFT
198 were processed with a custom script, to determine an identical arbitrary start point for all sequences,
199 before performing a second alignment with MAFFT. Phylogenetic trees were calculated in Jalview
200 v2.11.2 [29] with the average distance option. Consensus sequences from all sequences selected
201 within the groups defined based on the phylogenetic trees were used for a BLAST search in the
202 AMelMel1.1 assembly and hits following one another at distances shorter than the repeat period
203 size were grouped together.

204 The previously described monomer consensus sequences: accession X57427.1 for *AluI* and
205 X89530.1 for *AvaI* were used to detect their presence in the assembly by BLAST.

206 *Analysis of indels in populations.*

207 Indels were detected by aligning the two genomes HAv3.1 and AMelMA11.1 to one another with
208 minimap2 [30], followed by variant calling with SVIM-asm [31]. Two nuclear mitochondrial DNA
209 (NUMT) were then selected for genotyping in a set of 80 haploid males representing the three major
210 European bee subspecies: *A. m. mellifera* ($n=35$), *A. m. ligustica* ($n=30$) and *A. m. caucasica*
211 ($n=15$) (see Additional file 2: Table S2). All 80 samples were aligned to both assemblies as
212 described in Wragg et al. (2022) [6] and sequencing depth was estimated using SAMtools [32].
213 Individual genotypes in the samples sequencing data was determined for the two indels by two
214 methods. One method consisted in using GraphTyper2 [33], that will detect breakpoints due to
215 insertions, deletions or inversions in the pangenome graph built with SVIM-asm using the two
216 assemblies HAv3.1 and AMelMel1.1. The other method consisted in using sequencing depths as an
217 indication of presence or absence of Indels. For a given Indel and for each sample, the sequencing
218 depth for the alignments on the genome in which the Indel is present was calculated and compared
219 to the sequencing depth of the sequences flanking the Indel on both sides. Normalisation was done
220 by calculating the ratio between sequencing depth in the indel and in the flanking sequences.
221 Genotype presence or absence was then done by K-means clustering with $K=2$.

222 **Results**

223 *PacBio long-read sequencing and assembly into contigs.*

224 All long-read sequence data comes from a single haploid drone selected amongst several tested, as
225 having the highest DNA concentration and a peak of DNA fragment length at 35 kb (see Additional
226 file 1: Fig. S2). A high proportion of reads exceeds 10 kb and a few reads are longer than 70 kb.
227 Their size distribution is shown in Additional file 1: Fig. S3 and S4. After assembly, a total of 200
228 contigs (gap-free sequence tracts) was obtained. The longest contig is 11.6 Mb and the N50 contig
229 size is 5.1 Mb (see Additional file 2: Table S3 and Additional file 1: Fig. S5). These results are a

230 major improvement in comparison to the 46 kb N50 contig of Amel4.5 and quite similar to the N50
231 contig of 5.4 Mb observed in the HAv3.1 assembly [11]. Analysis with BUSCO showed that
232 overall, AMelMel1.1 had a slightly larger gene content than both Amel 4.5 and the most recently
233 published reference assembly AmelHAv3.1 [34] (see Additional file 2: Table S4).

234 *Chromosomal assignation and ordering contigs with crossing-over data.*

235 A priori chromosomal assignment of contigs was done by alignment to the Amel4.5 assembly using
236 LAST v956 [18]. Out of the 200 contigs, 110 aligned successfully. Crossing-over data for
237 confirming assignation and ordering contigs along chromosomes was obtained by using the reads
238 from the sequencing of 43 drones from three colonies, initially used to estimate recombination rate
239 in honey bee [16]. Briefly, this data set contains sequence for three queens and their drone offspring
240 (15 to 13 depending on the colony). Three of the drones of colony 1 are sequenced in duplicate and
241 are used for the quality control of SNP calling. Aligning these reads to our contigs allowed the
242 detection of 2,103,924 SNPs, on 176 contigs before any quality control. Out of these, approximately
243 64.5% were discarded due to an absence of polymorphism within the three colonies analysed, 1%
244 for being homozygous in the queens and 1% for being heterozygous in the drones. Furthermore,
245 0.2% of the SNPs were discarded for being inconsistent between the three drone replicates and
246 finally 0.4% were discarded for having allelic inconsistencies between queen and drones of the
247 same colony. After all the quality controls and for each of the three colonies, 687,699; 698,123 and
248 672,728 reliable SNPs (approximately 32% of the initial SNPs), were detected in each of the three
249 colonies on 114, 112 and 113 contigs respectively (see Additional file 1: Fig. S6). In total 120
250 contigs were at least partially informative across the colonies, with 104 contigs informative in the
251 three colonies and 16 for only one or two. A total of 114,754 polymorphic SNPs was present overall
252 in the 104 contigs informative across all three colonies (see Additional file 1: Fig. S6). Genotype

253 vectors for each SNP across colony drones were then defined, allowing for within-contig crossing-
254 over detection (see Additional file 1: Fig. S7). Genotype vectors from the ends of contigs were then
255 used to join contig ends together by finding for each contig end, the best corresponding end from
256 another contig having either the same genotype vector or a genotype vector presenting a minimal
257 number of crossing-overs (see Additional file 1: Fig. S7). To minimize the number of comparisons,
258 the *a priori* chromosomal assignment by alignment to Amel4.5 (see above) was used.

259 One hundred and two contigs out of the 110 with chromosome assignment by sequence similarity to
260 Amel4.5, had SNP genotype data and were thus informative for crossing-over detection. At least
261 one crossing-over event, as evidenced by the presence of at least 2 genotype vector bins, could be
262 detected within 86 of these contigs, thus allowing for their orientation. The remaining 16 contigs
263 were oriented based on the alignment to Amel4.5. All these contigs were small, except one contig
264 on chromosome 7. Despite its large size, close to 2.4 Mb, it was indeed difficult to orientate using
265 the genetic map, as no crossing-over could be detected due to an unusually low number of SNPs
266 and a very low local recombination rate. Moreover, its orientation could not be deduced from
267 Amel4.5 or even from the more recent assembly HAv3.1, as both possible orientations induced
268 large inversions when compared to these other two assemblies. Contigs assigned to chromosomes
269 by alignment only (8 contigs) or by crossing-over data alone (16 contigs), were assigned to their
270 chromosomes, but at an unknown (unlocalised) position. All remaining 72 contigs were considered
271 unplaced (see Additional file 1: Fig. S6).

272 *Tandem repeats at contig boundaries and Orientation of a large inversion on chromosome 7.*

273 With long read data, sequence contigs are large, but still don't cover the entire length of
274 chromosomes, with the exception of chromosome 16. When analysing the contig ends, we found
275 that almost all were composed of tandem repeats arrays usually longer than the read lengths, thus

276 preventing assembly. To orientate the large contig on chromosome 7, positioned as 5th in order
277 along the chromosome by the CO data, we took advantage of the fact that the repeat elements
278 detected by TRF and present at both extremities of the contig have different period sizes (258 and
279 1296 bp) and consensus sequences. These were compared to the proximal repeats of the 4th and the
280 6th contigs of chromosome 7. Interestingly, a tandem repeat element of 258 bp was detected at the
281 end of the 4th contig, and of 1296 bp at the end of the 6th contig, period sizes identical to the
282 extremities of the 5th contig, suggesting the correct orientation of the 5th contig. Correspondence
283 between these contig ends was further examined by pairwise alignment of the repeat sequences with
284 NCBI BLAST. The Identity was 100 % between the sequences of identical period sizes, whereas no
285 significant similarity could be found between the others (see Fig. 1 and Additional file 2: Table S9),
286 thus confirming the orientation of the contig. Dotplots comparing AMelMel1.1 and HAv3.1 are
287 shown in Additional file 3 and suggest a very small number of discrepancies, the major one residing
288 on chromosome 7.

289 *Telomeric and centromeric consensus sequences.*

290 The presence of telomeres is an indication of the completeness of the assembly. These were
291 analysed by searching for the accepted TTAGG consensus sequence for Hymenoptera [35] in TRF
292 analysis output, estimating their distance to the ends of chromosomes and comparing the results to
293 that of other 2-7 bp repeats, including non-TTAGG 5 bp repeats. Results (Fig. 2) show that TTAGG
294 are repeated with at least 842 copies when present at the extremities of chromosomes, whereas other
295 interstitial TTAGG repeats have only 117 repeats or less (mean = 21.3, median = 16.7), a size
296 distribution close to that of other pentanucleotide repeats (mean = 24.2, median = 14.4). See also
297 Additional file 1: Fig. S8 and Additional file 2: Table S6, S7 and S8 for data on other STR motifs.
298 In the AMelMel1.1 assembly, no TTAGG repeats were found on chromosomes 3, 7, 12 and 15 and

299 were found only at the beginning of chromosome 1, whereas in the HAv3.1 assembly, they could be
300 found at both extremities of this chromosome, but were absent from chromosomes 5 and 11 [11].

301 An AATAT repeat was found at the beginning of chromosome 15 in our assembly.

302 The *AluI* and *AvaI* repetitive sequences, previously described as being respectively telomeric and
303 centromeric were localised on the AMelMel1.1 assembly by BLAST search and the number of
304 copies per locus detected was counted (Fig 2). The *AluI* repeat was found at the start of
305 chromosomes 2 (6 repeats), 7 (3 repeats), 11 (46 repeats) and 12 (32 repeats). In addition, a single
306 *AluI* element was found around position 8 Mb on chromosome 15, at more than 1.5 Mb from the
307 distal end. Curiously, the *AluI* repeats found on chromosomes 2 and 11 were at the opposite end
308 from the TTAGG sequences we detected (Fig. 2). The *AvaI* repeat was found as arrays at single loci
309 on chromosomes 1, 2, 4, 9, 11 and 14. Only 4 copies in the array were found on chromosome 1, the
310 other arrays having between 10 and more than 30 copies. The *AvaI* repeats are at the start of
311 chromosomes 9 and 14, at the opposite end from the TTAGG repeats. On the other four
312 chromosomes, they are at least at 1.8 Mb from a chromosome end (Fig. 2).

313 *Recombination pattern*

314 Having used crossing-over detection and a genetic map for contig scaffolding, we could estimate
315 the total genetic map for AMelMel, which is approximately 50 Morgans long, giving an average
316 recombination rate in the genome of 23 cM/Mb, close to the first estimations based on the
317 microsatellite genetic map and to the most recent ones based on SNPs (Table 1). However, although
318 we used the same sequencing dataset as in Liu et al. (2015) [16], we found a drastic reduction in
319 recombination rate between our genetic map and the one they initially published, which is 37
320 cM/Mb (Table 1). A great difference is that the latter is based on alignments of the sequence reads
321 on Amel4.5. When aligning our assembly with Amel4.5, we find an agreement on the chromosomal

322 assignment of the contigs, but reveal many discrepancies in the orientation of large chromosome
323 segments. At most breakpoint positions between the two assemblies, recombination hotspots are
324 detected on Amel4.5 (Fig. 3 and Additional file 4), suggesting these assembly errors were
325 responsible for the overall higher recombination rate observed in Liu et al. (2015) [16]. This
326 reduction from 37 cM/Mb to 23 cM/Mb is explained by these artefactual recombination hotspots
327 detected on Amel4.5 at the breakpoint positions where the two assemblies disagree, that are absent
328 in AMelMel1.1 (i.e. for chromosome 3 shown in Fig. 3 and Additional file 4 for all the
329 chromosomes).

330 *High conservation of tandem repeat sequences across chromosomes.*

331 We used TRF to further localise and analyse the repeat arrays in the whole honey bee genome.
332 Interestingly, two major period size classes for tandem repeats could be found: one in the size range
333 of 91-93 bp, with a maximum number of 231 repeats, hereafter called the 91 bp repeat and the
334 second in the size range of 367-371 bp, with a maximum number of 100 repeats, called the 371 bp
335 repeat (see Additional file 1: Fig. S9). The 91 repeats are found on all chromosomes, whereas the
336 371 bp repeats are on all chromosomes except chromosome 16 (see Additional file 1: Fig. S10).
337 Interestingly, very long repeats whose length is within the range of the sequence reads, were often
338 found at the junction between two sequence contigs, confirming they could be responsible for the
339 impossibility to sequence and to assemble these regions properly (see Fig. 2, Additional file 1:
340 Fig11).

341 We investigated further the nature of the 91 and 371 repeats by analysing their potential
342 homogeneity in terms of sequence content. Summary statistics for the two classes show very
343 different distributions in terms of repeat copy numbers within tandem arrays (see Additional file 1:
344 Fig. S12 and Additional file 2: Table S9). There is a total of 345 arrays of the 91 bp repeat in the

345 genome and 131 arrays of the 371 bp repeats. However, these numbers drop to 43 and 74
346 respectively when only considering tandem arrays of more than 10 repeats, suggesting that most of
347 the 91 bp repeats have less than 10 elements (see Additional file 1: Fig. S12). To investigate
348 sequence homogeneity within each of the two repeat classes, we selected the repeat sequence
349 defined by TRF for repeats having strictly more than ten copies in tandem within an array. For the
350 91 bp repeat, we selected for $91 \leq \text{period size} \leq 93$ and for the 371 bp repeat $367 \leq \text{period size} \leq 371$,
351 as suggested by the graph shown in Additional file 1: Fig. S9. Then, for each repeat class, we
352 performed a multi-sequence alignment with MAFFT, and produced an average distance tree with
353 Jalview. Results show that out of the 74 sequences of the 371 bp repeat class, 72 were clearly
354 grouped together, having high similarity (Fig. 4), whereas the 43 sequences of the 91 bp repeat class
355 showed lower similarity. We therefore decided to subdivide the 91 bp repeat class into three groups
356 of 20, 10 and 3 sequences, based on the average distance tree (Fig. 4). The remaining ten 91 bp
357 repeat class sequences were singletons. A consensus sequence was made for each of the four group
358 of sequences, and was used for a BLAST search in the AMelMel1.1 assembly. The homogeneity of
359 the 371 bp consensus sequence was confirmed by the detection of a very high number of hits of
360 high similarity covering the overall length of the queries (see Additional file 1: Fig. S13). On the
361 contrary, for the three different consensus sequences used separately for the 91 bp repeat, alignment
362 length and sequence similarities were lower, confirming that it to correspond more to a class size,
363 rather than a specific repeat family based also on sequence composition (see Additional file 1: Fig.
364 S13).

365 We then searched for the possible existence of the 371 and 91 bp repeats in other organisms.
366 BLAST searches with each of the 371 bp repeat consensus sequences did not allow to find any
367 significant hit in the NCBI nucleic collection database. When searching with each of the 91 bp
368 consensus repeats, four hits were found: three consensus sequences from repeat arrays from

369 chromosome 11 and one consensus sequences from a repeat array from chromosome 12 showed
370 sequence similarity to fragments of predicted lncRNAs *LOC116185390*, *LOC105734921*,
371 *LOC116415009*, *LOC116185696*, from unknown scaffolds of the genome assemblies of *Apis*
372 *dorsata* and *Apis florea*. However, these lncRNAs are composed of two exons and span close to 1.5
373 kb in the genomes of *Apis dorsata* and *Apis florea*, suggesting that the 91 bp sequences correspond
374 to only a portion (one out of two exons) of these lncRNAs. To investigate further, we performed
375 BLAST searches with each of the consensus sequences directly on the refseq_genomes databases of
376 *Apis cerana*, *Apis dorsata* and *Apis florea*. A very high number of hits were found, suggesting the
377 371 bp and 91 bp repeats were also present in these three genomes, with an apparent slightly higher
378 percent identity for the 91 bp repeat (see Additional file 1: Fig. S14).

379 *Difference in the number of repeats of 5S ribosomal RNA genes.*

380 Genes that are repeated in tandem can often vary in numbers between individuals through unequal
381 crossing-over [36]. They are therefore good candidates to study functional variation related to large
382 rearrangements. A typical example of such genes is the 5S ribosomal RNA genes whose copy
383 number can vary greatly in the genome [37–39]. Alignment of a region from the AMelMel1.1 and
384 HAv3.1 assemblies in a region on chromosome 3 containing 5S ribosomal RNA genes, show a
385 variation in the number of these genes between the two genomes (Fig. 5.). The period size of one of
386 the repeat arrays of 5S genes is 357 bp, while that of the second is 373 bp. However, inclusion of
387 this sequence in the multiple sequence analysis of the 371 bp repeat shows that these two sequences
388 are different (see Additional file 1: Fig 15)

389 *Inversions between AMelMel1.1 and HAv3.1.*

390 One-to-one split alignments produced by aligning AMelMel1.1 on HAv3.1 with LAST were used to
391 detect inversions larger than 1000 bp between both genomes. The largest inversion detected is on

392 chromosome 7 and is larger than 1.6 Mb (see Additional file 3 and Additional file 5). It should be
393 noted, that a similar rearrangement on chromosome 7 was previously detected when comparing a
394 genome assembly of an *A. m. ligustica* samples with the HAv3.1 reference [40]. Although close to
395 one hundred other inversions could be detected, their visual inspection on dotplot graphs show that
396 53 are within complex repeat patterns present at the junction between contigs, 32 within other
397 complex repeat elements and only 12, are in the middle of the high-quality sequence contigs in both
398 assemblies, thus representing well supported inversions. Apart the large inversion on chromosome
399 7, the smallest is 1055 bp long and the largest 25608 bp long (see Additional file 2: Table S10 and
400 Additional file 5). Interestingly, some inversions will concern genes and can involve repeat
401 elements differentially found in both assemblies. In the example shown in Fig. 6, a local inverted
402 duplicated region seen in the HAv3.1 assembly, is absent in AMelMel1.1. This chromosomal
403 segment contains a portion of the gene model *LOC113218640*, which has no direct annotation in
404 the HAv3.1 assembly, but is described as coding for a *bric-a-brac 1-like* protein. *Bric-a-brac* was
405 shown to be involved in body pigmentation in drosophila [41]. Another interesting inversion is 11
406 kb long on chromosome 3, in an intron of *Rhomboid*, a gene involved in the formation of wing
407 veins in *Drosophila* [42]. A more complex rearrangement involves a gene labelled as a probable
408 nuclear hormone receptor *HR38*, involved in synchronizing the reproductive activity in *Agrotis*
409 *ippsilon* [43] and in the larval-pupal transition in *Leptinotarsa decemlineata* [44]. Other genes
410 involved in the inversions described are reported in Additional file 2: Table S10.

411 *Using both assemblies for the analysis of two medium-size InDels in honey bee subspecies.*

412 To demonstrate the utility of using two reference genomes for analysing structural variants, we
413 studied two indels corresponding to nuclear mitochondrial DNA (NUMT), that were detected by
414 using minimap2 [30] and SVIM-asm [31]. The first one, NUMT_Chr2, is 745 bp long, has 92.7 %

415 identity over 99 % of its length to HAv3.1 mitochondrial DNA, is present in the AMelMel1.1
416 assembly on chromosome 2 at positions 12,212,275 – 12,213,020 and absent from the HAv3.1
417 assembly. The second one, NUMT_Chr10, is 576 bp long, has 92.5 % identity over 94 % of its
418 length to HAv3.1 mitochondrial DNA, is present in the HAv3.1 assembly on chromosome 10 at
419 positions 670,675 – 671,251 and absent in the AMelMel1.1 assembly. The presence and absence of
420 these two NUMTs were tested in three honey bee subspecies: *A. m. mellifera* (n=35), *A. m. ligustica*
421 (n=30) and *A. m. caucasia* (n=15), for which Illumina sequencing data was aligned to both
422 reference genomes. Inspection of mean sequencing depth over all 80 samples in the regions of
423 NUMT_Chr2 and NUMT_Chr10 indicates a decrease of the mean depth and an increase of its
424 variance (see Additional file 1: Fig. 16), suggesting the existence of a presence / absence
425 polymorphism. When inspecting the sequencing depth per population, the *A. m. mellifera* samples
426 show a constant value over NUMT_Chr2 and have a depth close to zero over NUMT_Chr10,
427 whereas the *A. m. ligustica* show an inverse tendency (Fig. 7). The *A. m. caucasia* samples seem not
428 to have NUMT_Chr2 in their genomes, whereas a few may have NUMT_Chr10, as although there
429 is a drop of mean sequencing depth on HAv3.1 in the corresponding region, there is still some low
430 coverage (Fig. 7). To genotype our samples individually, we used two methods. The first was to
431 estimate individual sequencing depth in the chromosomal region delimiting the NUMTs, by using
432 AMelMel1.1 as reference genome for NUMT_Chr2 and HAv3.1 for NUMT_Chr10 (see methods).
433 All 80 samples could thereafter be called unambiguously assigned to one of two groups (presence
434 or absence) by K-means clustering (see Additional file 1: Fig. 17). The second method tested was to
435 use GraphTyper2 [33], allowing the genotyping of structural variation using pangenome graphs.
436 Our GraphTyper2 results, showed that the calling of samples was incomplete, with a high
437 proportion of no-calls, and that the fact of using individual bam files of alignments to one or to the
438 other reference genome can greatly influence the call rate (see Additional file 2: Table S11). Indeed,

439 for the detection of variants with minimap2 and SVIM-asm, a reference genome must be specified
440 and bam files of alignments to this specific reference genome must be used to perform the
441 individual genotyping. So, we first used HAv3.1 as reference and the results were a genotyping call
442 rate of 78.7 % for NUMT_Chr2 and null for NUMT_Chr10, as the line describing its potential
443 genotypes didn't appear in the output file from GraphTyper2 at all. To check if the reference
444 genome could influence the results, we also performed the analysis by using AMelMel1.1 as
445 reference and this time the call rate was 85.0 % for NUMT_Chr2, and 76.2 % for NUMT_Chr10.
446 When genotype calls were successfully obtained in both analyses, results were identical and were
447 also concordant with the analysis based on sequencing depth, showing that when genotyping was
448 possible with GraphTyper2, the results were consistent. Two samples were called as heterozygotes
449 for NUMT_Chr2, when using AMelMel1.1 as reference and were counted as “no calls”, given our
450 samples were haploid. Low sequencing depth could have been a possible explanation for the
451 absence of genotyping results with GraphTyper2 in some of the samples, but this does not seem to
452 be the case, as all samples that failed genotyping had at least 8X average sequencing depth in the
453 sequence flanking the NUMTs analysed, whereas successful genotyping could be obtained for
454 samples having as little as 2X sequencing depth (see Additional file: Fig. 18). Substantially, the
455 individual genotyping results confirm the overall impression that the presence or absence of the
456 NUMT insertions are specific to the subspecies analysed, with most, if not all samples having
457 identical within-population genotypes, except for NUMT_Chr10 in *A. m. caucasia*, for which four
458 out of eleven samples have a different allele. Interestingly, NUMT_Chr2 is present in all *A. m.*
459 *mellifera* and only two *A. m. ligustica* samples, and absent from all other samples, whereas
460 NUMT_Chr10 is absent from *A. m. mellifera* samples and present in all but one *A. m. ligustica*
461 samples and four *A. m. caucasia* samples (Fig. 7, Fig. 8).

462

463 **Discussion**

464 *AMelMel assembly quality and comparison to other honey bee assemblies*

465 Although five chromosome level genome assemblies for *Apis mellifera* are available [45] ours has
466 the originality of representing *Apis mellifera mellifera*. Indeed, this subspecies is genetically distinct
467 from *Apis mellifera ligustica*, *Apis mellifera carnica* and *Apis mellifera caucasia* [46] represented
468 by the four other assemblies. Another originality of our study, is that the contigs we obtained were
469 scaffolded into chromosomes using a genetic (recombination) map rather than the now more
470 common HiC chromatin conformation and Bionano optical maps methods. Compared to the current
471 HAv3.1 reference genome [11], our assembly is slightly longer (227 Mb versus 225 Mb), is built
472 from a lower number of contigs (200 versus 228) with very similar N50 contig values (5.1 Mb
473 versus 5.4 Mb). However, the overall final coverage was slightly smaller in our study (137X Pac
474 Bio and Illumina reads versus 192X in HAv3.1). BUSCO statistics are also very similar due to the
475 fact that contig building was based in both cases on PacBio reads with some correction using
476 Illumina reads. Assembly of contigs into chromosomes using the recombination data failed to
477 accurately order and orient in only one instance for a large contig on chromosome seven. Despite
478 this limitation, the orientation of this contig was possible thanks to a careful analysis of tandem
479 repeat elements at its boundaries. Sequencing data for both HAv3.1 and our assembly, AmelMel1.1,
480 are from a single haploid drone, which is a tremendous advantage for the resolution of regions
481 largely composed of repeat elements. This was recently demonstrated in the human Telomere-to-
482 Telomere project, for which a complete hydatidiform mole haploid cell line was used, helping to
483 solve complex structures such as centromeres [47]. Our results show however, that although the
484 sequencing of repeat elements and especially of challenging tandem repeats seems resolved by the
485 use of a single haploid sample and long reads, there are cases in which the total length of
486 monotonous repeats is larger than the reads lengths, preventing local assembly. As a result, for

487 almost all contig boundaries investigated, long stretches of tandem repeats were found (Fig. 2).
488 Interestingly, chromosome 16, which was obtained as a single contig, has no stretch of tandem
489 repeats exceeding 10 kb.

490 *Genetic maps and recombination rate in the honey bee*

491 Having used genetic recombination data to scaffold our contigs, we could build a new recombination
492 maps and give an estimation of 23 cM/Mbp for the overall recombination rate in the honey bee [16],
493 which is of the same magnitude as the latest values from [34] and also congruent with prior values
494 [48–50]. It is interesting to note, that the public sequencing dataset we used, representing 43 drone
495 genome offspring of three queens, gave a much higher estimate of 37 cM/Mb when previously used
496 for generating genotyping data by alignment on the Amel4.5 reference genome [16]. On closer
497 inspection, this higher overall recombination rate in Liu et al. (2015) [16], is due to very specific
498 false recombination hotspots that appear at contig junctions in Amel4.5, when at least one of them
499 is inverted as compared to AMelMel1.1 (Fig. 3 and Additional file 4). This illustrates the
500 importance of the quality of the reference genome for such studies. Errors in the local estimations of
501 recombination rate when using mis-assembled reference genome will in turn affect any analysis
502 based on recombination maps or including linkage disequilibrium.

503 *Tandem repeats and the current limits for obtaining chromosome-wide contigs*

504 We found a high occurrence of conserved tandem repeats in the honey bee genome, whose length
505 and sequence conservation caused problems for scaffolding contigs into chromosomes, the ultimate
506 goal being each chromosome covered by a single contig. Indeed, long stretches of such repeats were
507 found at the boundaries between contigs. Luckily, the only large contig in the assembly, that could
508 be placed on chromosome 7, but not oriented due to lack of sufficient genetic data, had different

509 tandem repeats at each of its extremities, allowing to decide on a correct orientation. However,
510 other regions may still be problematic, the most striking example being the region between 1 and 3
511 Mb on chromosome 10. In this region, the contigs are small (< 0.2 Mb) due to a high occurrence of
512 tandem repeats, leading to difficulties in their ordering along the chromosome and their orientation.
513 Moreover, these repeats appear mostly to belong to the highly conserved 371 bp family, preventing
514 their use for contig mapping. This portion of chromosome 10 has also been described as difficult to
515 assemble in other studies [51].

516 *General chromosome structure: telomeres, centromeres.*

517 Cytogenetic studies based on fluorescent *in situ* hybridization of *AluI* and *AvaI* probes suggest that
518 the honey bee genome is composed of one large metacentric and 15 acrocentric chromosomes [52].
519 This is to date still considered as the honey bee standard karyotype structure [34,35]. However,
520 other data could question this structure, for instance the suggested positions of the centromeres
521 based on sequence characteristics of the HAv3.1 genome assembly such as the (GC) % and the
522 presence of *AluI* and *AvaI* repeats on chromosomes 7, 8 and 11 in Wallberg et al. (2019) [34].

523 Regarding telomeres, we were not able to identify the TTAGG consensus sequences on all the 17
524 chromosome ends (two for the metacentric chromosome 1 and one for each of the other fifteen
525 acrocentric chromosomes) where they were expected: none were detected on the right arm of
526 chromosome 1 and on chromosomes 3, 12 and 15. Interestingly, some chromosomes also lacked
527 TTAGG repeats in the HAv3.1 assembly, but these were not the same (chromosomes 5 and 11).
528 These discrepancies can be due to problems in the assembly of these repeat regions, either due to
529 variations in the sequence quality between the two datasets or to local variations in repeat content,
530 rendering the assembly of varying difficulty due to biological reasons. It is interesting to note, that
531 in the older assemblies of the bee genome, based on the same DH4 strain as HAv3.1, extended

532 analyses of telomeric and subtelomeric repeats showed that some chromosomes were easier to
533 analyse than others and that no TTAGG repeats were identified for chromosome 5 [35]. Taken
534 altogether, although the current sequencing data supports the actual consensus karyotype structure,
535 we didn't find that the *AluI* repeat elements [52] could be considered as a marker of telomeres, as
536 when such repeats were detected at the extremity of a chromosome, this was at the opposite end
537 from the TTAGG repeats (see specifically chromosome 11 in Fig2).

538 The question of the exact position of the centromeres is a more complex one: the centromeres
539 would be expected at the middle of chromosome 1 and at the proximal end of each of the other
540 chromosomes. The *AvaI* repeat element, considered as a marker of the centromeres [52] was not
541 found on all chromosomes and even when found, the number of repeats in the array could be as
542 small as four, such as the repeat on chromosome 1 (Fig. 2). With the exception of chromosome 11
543 for which an *AvaI* repeat was found at the position 5 Mb, the *AvaI* elements, when detected on a
544 chromosome, were found within 2.5 Mb of the chromosome ends, reflecting the results found on
545 HAv3.1 [34]. However, although the positions of the *AvaI* repeats is identical between the two
546 assemblies, the number of repeat elements vary for each given position. For the moment, the exact
547 position of the centromeres remains uncertain, but the criteria of the eventual presence of an *AvaI*
548 element remains a plausible indication, especially as these seem to be coincident with other specific
549 characteristics, such as low (GC) content [50] or low levels of polymorphism and recombination
550 rates [46]. If these characteristics are indicators of the centromere positions, then chromosome 11
551 and perhaps also chromosome 7 should be considered sub-metacentric, although in this case,
552 TTAGG repeats would be expected at both of the extremities of these chromosomes, which is not
553 the case in any of the studies to date. Further improvements in genome sequencing and assembly
554 and in obtaining higher-resolution cytogenetic metaphase chromosome preparations will be
555 necessary to elucidate this question.

556 *Comparing the genomes of two honey bee subspecies.*

557 The HAv3.1 assembly is based on a sample from the DH4 line, thought to be mainly of *A. m.*
558 *ligustica* descent [9]. The comparison with our *Apis mellifera mellifera* AMelMel1.1 assembly
559 allows for the detection of rearrangements occurring between these two distinct genetic types, that
560 can't be detected through short read sequencing.

561 Short sequence fragments repeated in tandem, such as the 91 bp and 371 bp repeats described here,
562 tend to vary in copy number through non-allelic homologous recombination (NAHR) or unequal
563 crossing-over [53]. A rapid observation of the LAST alignment data between the two assemblies
564 suggests that the 371 bp repeat element can vary greatly in copy number and the 91 bp element to a
565 much lesser extent, although these preliminary observations will require more thorough analyses.
566 No obvious function was found for these elements to date, except for the fact that a BLAST search
567 found that the 91 bp element shows similarity of sequence to one out of two exons of *Apis dorsata*
568 and *Apis florea* lncRNAs, suggesting these are incomplete and consequently not active in the repeat
569 arrays. However, the annotation of the lncRNAs in *Apis dorsata* and *Apis florea* is only based on
570 the alignment of short reads RNA-seq. More work is needed to confirm this finding concerning the
571 91 bp repeat and further comparisons with other bee genomes whose sequences are underway [54]
572 will help understand these interesting genome elements. The 5S ribosomal RNA genes are another
573 interesting case of variation in gene number and studies in mouse and human have shown that this
574 variation may be important for a balanced dosage of rRNA, that can have possible implications in
575 diseases [37,38]. It would be interesting to see if the variations of 5S gene numbers observed here is
576 a difference between the two honey bee subspecies investigated or if intra-population variation can
577 be found.

578 After screening out rearrangements that could be due to errors associated with assembly problems,
579 such as inversions of complete small contigs, thirteen inversions larger than 1 kb were detected
580 between the two genomes. Out of these, a large 1.6 Mb inversion on chromosome 7 is likely an
581 error in HAv3.1, as it was also seen when sequencing a closely related sample from the *Apis*
582 *mellifera ligustica* subspecies [40]. Out of the twelve remaining inversions, some involve genes,
583 present either at one of the breakpoints, having inversions within their structure (usually introns) or
584 whose structure remains intact, but are in reverse orientation. As usual, interesting functions that
585 may explain some of the phenotypic differences found between the two subspecies represented by
586 our dataset will be found (see Additional file 2: Table S10). Even when restricting to genes for
587 which functions were observed in insects, three genes stand out. One is *Bric-a-brac 1-like*, whose
588 implication in body pigmentation in *Drosophila* [41], could be linked to our two reference genomes
589 representing light (yellow) and dark coloured honey bee subspecies. Another is *Rhomboid*,
590 previously shown to be involved in the formation in wing veins in *Drosophila* [42]. A third is the
591 hormone receptor *HR38*, shown to be involved in the synchronisation of reproductive activity in the
592 moth *Agrotis ipsilon* and the larval-pupal transition in the Colorado potato beetle *Leptinotarsa*
593 *decemlineata* [43,44].

594 *Nuclear mitochondrial DNA segments and perspectives for pangenomics.*

595 To test the utility of having two reference genomes for genotyping structural variants, we tried
596 genotyping two NUMTs, present in one or the other HAv3.1 and AMelMel1.1 assembly, with
597 GraphTyper2. Results show that GraphTyper2 could not call genotypes for all samples. In the first
598 instance, this is surprising, given the fact that this test of presence or absence of a 745 bp fragment
599 in the case of NUMT_Chr2 and a 576 bp one for NUMT_Chr10 is done on haploid samples,
600 simplifying the problem, as each of the NUMTs should be either present or absent in each

601 individual tested. This may be caused by the fact that Graphtyper2 extracts reads that were
602 previously mapped to the structural variant regions on a linear reference genome, thus possibly
603 introducing a bias. It is however surprising, that when HAv3.1 was used as reference for the
604 primary mapping of reads, NUMT_Chr10 could not be genotyped at all. This reference-bias could
605 be overcome by using more recent methods in which the reads for the genomes to genotype are
606 mapped directly on the pan-genome graph, although such methods are more complex to use in
607 practice, due to problems such as the definition of genome coordinates [55].

608

609 **Conclusions**

610 In conclusion, we present here a genome assembly for the honey bee *Apis mellifera* that is from a
611 different subspecies than the current reference genome. One originality of the assembly process was
612 to use recombination data rather than optical maps or HiC for scaffolding contigs into
613 chromosomes. We characterise for the first time long tandem repeats that are present in the genome
614 and are responsible for most sequence discontinuities and show that these belong to two main
615 repeats families yet to be further characterised and whose potential function in the genome remains
616 to be investigated. Finally, we show the interest of having two reference-quality genomes for the
617 detection of structural variants, such as inversions and insertions-deletions and demonstrate the
618 possibility of using a pan-genome approach for genotyping such variants in honey bee populations.

619

620 **Declarations**

621 **Ethics approval and consent to participate**

622 Not applicable

623 **Consent for publication**

624 Not applicable

625 **Availability of data and materials**

626 The AMelMel1.1 assembly has been deposited on the NCBI under the accession number
627 GCA_003314205. The reads of the 36 corresponding PACBIO_SMRT runs are in SRA under the
628 accessions SRR9587836 to SRR9593684. Scripts and supplementary description of bioinformatic
629 analyses are available in GitHub: <https://github.com/avignal5/PacificBee/tree/main>.

630 **Competing interests**

631 The authors declare that they have no competing interests

632 **Funding**

633 This study was financially supported by the INRA Département de Génétique Animale (INRA
634 Animal Genetics division) “PacificBee” grant. It was performed in collaboration with the GeT core
635 facility, Toulouse, France (<http://get.genotoul.fr>), and was supported by France Génomique
636 National infrastructure, funded as part of “Investissement d’avenir” program managed by Agence
637 Nationale pour la Recherche (contract ANR-10-INBS-09) and by the GET-PACBIO program («
638 Programme opérationnel FEDER-FSE MIDI-PYRENEES ET GARONNE 2014-2020 »).

639 **Authors’ contributions**

640 KC-T, WM and AV performed sampling and high molecular weight DNA extraction; CV, CR, CD
641 performed the sequencing; CK performed assembly into contigs and contig quality checks; SEE, BS
642 and AV did the chromosome-level assembly using recombination data; AV did the tandem repeats
643 analyses; QB and AV did the NUMT detection and analysis. SEE and AV drafted the manuscript.
644 All authors read and approved the final manuscript.

645

646 **References**

- 647 1. Sheppard W, Meixner MD. *Apis mellifera pomonella*, a new honey bee subspecies from Central Asia.
648 *Apidologie*. 2003;34:367–75.
- 649 2. Whitfield CW, Behura SK, Berlocher SH, Clark AG, Johnston JS, Sheppard WS, et al. Thrice out of
650 Africa: ancient and recent expansions of the honey bee, *Apis mellifera*. *Science*. 2006;314:642–5.
- 651 3. Han F, Wallberg A, Webster MT. From where did the Western honeybee (*Apis mellifera*) originate? *Ecol*
652 *Evol*. 2012;2:1949–57.
- 653 4. Cridland JM, Tsutsui ND, Ramírez SR. The Complex Demographic History and Evolutionary Origin of
654 the Western Honey Bee, *Apis Mellifera*. *Genome Biology and Evolution*. 2017;9:457–72.
- 655 5. Dogantzis KA, Tiwari T, Conflitti IM, Dey A, Patch HM, Muli EM, et al. Thrice out of Asia and the
656 adaptive radiation of the western honey bee. *Sci Adv*. 2021;7:eabj2151.
- 657 6. Wragg D, Eynard SE, Basso B, Canale-Tabet K, Labarthe E, Bouchez O, et al. Complex population
658 structure and haplotype patterns in the Western European honey bee from sequencing a large panel of
659 haploid drones. *Molecular Ecology Resources*. 2022;22:3068–86.
- 660 7. Pieplow JT, Brauße J, Praagh JP, Moritz RFA, Erler S. A scientific note on using large mixed sperm
661 samples in instrumental insemination of honeybee queens. 2017;1–3.
- 662 8. Pinto MA, Henriques D, Chávez-Galarza J, Kryger P, Garnery L, van der Zee R, et al. Genetic integrity of
663 the Dark European honey bee (*Apis mellifera mellifera*) from protected populations: a genome-wide
664 assessment using SNPs and mtDNA sequence data. *Journal of Apicultural Research*. 2014;53:269–78.
- 665 9. Weinstock GM, Robinson GE, Worley KC, Hartfelder K, Zdobnov EM, Hartfelder K, et al. Insights into
666 social insects from the genome of the honeybee *Apis mellifera*. *Nature*. 2006;443:931–49.
- 667 10. Elsik CG, Worley KC, Bennett AK, Beye M, Camara F, Childers CP, et al. Finding the missing honey
668 bee genes: lessons learned from a genome upgrade. *BMC Genomics*. 2014;15:86.
- 669 11. Wallberg A, Bunikis I, Vinnere Pettersson O, Mosbech M-B, Childers AK, Evans JD, et al. A hybrid de
670 novogenome assembly of the honeybee, *Apis mellifera*, with chromosome-length scaffolds. *bioRxiv*.
671 2018;1–37.
- 672 12. Talenti A, Powell J, Hemmink JD, Cook EAJ, Wragg D, Jayaraman S, et al. A cattle graph genome
673 incorporating global breed diversity. *Nat Commun*. 2022;13:910.
- 674 13. Wang T, Antonacci-Fulton L, Howe K, Lawson HA, Lucas JK, Phillippy AM, et al. The Human
675 Pangenome Project: a global resource to map genomic diversity. *Nature*. 2022;604:437–46.
- 676 14. Garnery L, Franck P, Baudry E, Vautrin D, Cornuet J-M, Solignac M. Genetic diversity of the west
677 European honey bee (*Apis mellifera mellifera* and *A. m. iberica*) I. Mitochondrial DNA. *Genetics Selection*
678 *Evolution*. 1998;30:S31.

- 679 15. Garnery L, Franck P, Baudry E, Vautrin D, Cornuet J-M, Solignac M. Genetic diversity of the west
680 European honey bee (*Apis mellifera mellifera* and *A. m. iberica*) II. Microsatellite loci. *Genetics Selection*
681 *Evolution*. 1998;30:S49.
- 682 16. Liu H, Zhang X, Huang J, Chen J-Q, Tian D, Hurst LD, et al. Causes and consequences of crossing-over
683 evidenced via a high-resolution recombinational landscape of the honey bee. *Genome Biol*. 2015;16:15.
- 684 17. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate
685 long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res*. 2017;27:722–36.
- 686 18. Frith MC, Kawaguchi R. Split-alignment of genomes finds orthologies more accurately. *Genome Biol*.
687 2015;16:1–17.
- 688 19. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis
689 Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*.
690 2010;20:1297–303.
- 691 20. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*.
692 2010;26:589–95.
- 693 21. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv:12073907 [q-
694 bio] [Internet]. 2012 [cited 2019 Apr 3]; Available from: <http://arxiv.org/abs/1207.3907>
- 695 22. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and
696 VCFtools. *Bioinformatics*. 2011;27:2156–8.
- 697 23. Petit M, Astruc J-M, Sarry J, Drouilhet L, Fabre S, Moreno C, et al. Variation in Recombination Rate and
698 Its Genetic Determinism in Sheep Populations. *Genetics*. 2017;genetics.300123.2017.
- 699 24. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, et al. BUSCO
700 Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Molecular Biology and*
701 *Evolution*. 2018;35:543–8.
- 702 25. Zdobnov EM, Tegenfeldt F, Kuznetsov D, Waterhouse RM, Simão FA, Ioannidis P, et al. OrthoDB v9.1:
703 cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral
704 orthologs. *Nucleic Acids Research*. 2017;45:D744–9.
- 705 26. Zhao H, Sun Z, Wang J, Huang H, Kocher J-P, Wang L. CrossMap: a versatile tool for coordinate
706 conversion between genome assemblies. *Bioinformatics*. 2014;30:1006–7.
- 707 27. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*.
708 1999;27:573–80.
- 709 28. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in
710 Performance and Usability. *Molecular Biology and Evolution*. 2013;30:772–80.
- 711 29. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2--a multiple
712 sequence alignment editor and analysis workbench. *Bioinformatics*. 2009;25:1189–91.
- 713 30. Li H. Minimap2: pairwise alignment for nucleotide sequences. Birol I, editor. *Bioinformatics*.
714 2018;34:3094–100.

- 715 31. Heller D, Vingron M. SVIM-asm: structural variant detection from haploid and diploid genome
716 assemblies. Robinson P, editor. *Bioinformatics*. 2021;36:5519–21.
- 717 32. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map
718 format and SAMtools. *Bioinformatics*. 2009;25:2078–9.
- 719 33. Eggertsson HP, Kristmundsdottir S, Beyter D, Jonsson H, Skuladottir A, Hardarson MT, et al.
720 GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nat*
721 *Commun*. 2019;10:5402.
- 722 34. Wallberg A, Bunikis I, Pettersson OV, Mosbech M-B, Childers AK, Evans JD, et al. A hybrid de novo
723 genome assembly of the honeybee, *Apis mellifera*, with chromosome-length scaffolds. *BMC Genomics*.
724 2019;20:275.
- 725 35. Robertson HM, Gordon KHJ. Canonical TTAGG-repeat telomeres and telomerase in the honey bee, *Apis*
726 *mellifera*. *Genome Res*. 2006;16:1345–51.
- 727 36. Vignal A, London J, Rahuel C, Cartron JP. Promoter sequence and chromosomal organization of the
728 genes encoding glycoporphins A, B and E. *Gene*. 1990;95:289–93.
- 729 37. Gibbons JG, Branco AT, Godinho SA, Yu S, Lemos B. Concerted copy number variation balances
730 ribosomal DNA dosage in human and mouse genomes. *Proc Natl Acad Sci USA*. 2015;112:2485–90.
- 731 38. Hall AN, Turner TN, Queitsch C. Thousands of high-quality sequencing samples fail to show meaningful
732 correlation between 5S and 45S ribosomal DNA arrays in humans. *Sci Rep*. 2021;11:449.
- 733 39. Ding Q, Li R, Ren X, Chan L, Ho VWS, Xie D, et al. Genomic architecture of 5S rDNA cluster and its
734 variations within and between species. *BMC Genomics*. 2022;23:238.
- 735 40. Cao L, Zhao X, Chen Y, Sun C. Chromosome-scale genome assembly of the high royal jelly-producing
736 honeybees. *Sci Data*. 2021;8:302.
- 737 41. De Castro S, Peronnet F, Gilles J-F, Mouchel-Vielh E, Gibert J-M. bric à brac (bab), a central player in
738 the gene regulatory network that mediates thermal plasticity of pigmentation in *Drosophila melanogaster*.
739 Kopp A, editor. *PLoS Genet*. 2018;14:e1007573.
- 740 42. Sturtevant MA. The *Drosophila* rhomboid gene mediates the localized formation of wing veins and
741 interacts genetically with components of the EGF-R signaling pathway. *Genes & Development*.
742 1993;7:961–73.
- 743 43. Gassias E, Durand N, Demondion E, Bourgeois T, Bozzolan F, Debernard S. The insect HR38 nuclear
744 receptor, a member of the NR4A subfamily, is a synchronizer of reproductive activity in a moth. *FEBS J*.
745 2018;285:4019–40.
- 746 44. Shen C-H, Xu Q-Y, Mu L-L, Fu K-Y, Guo W-C, Li G-Q. Involvement of *Leptinotarsa* hormone receptor
747 38 in the larval-pupal transition. *Gene*. 2020;751:144779.
- 748 45. NCBI Genome [Internet]. [cited 2023 Aug 28]. Available from:
749 <https://www.ncbi.nlm.nih.gov/datasets/genome/?taxon=7460>
- 750 46. Wragg D, Eynard SE, Basso B, Canale-Tabet K, Labarthe E, Bouchez O, et al. Complex population

- 751 structure and haplotype patterns in Western Europe honey bee from sequencing a large panel of haploid
752 drones [Internet]. Genetics; 2021 Sep. Available from:
753 <http://biorxiv.org/lookup/doi/10.1101/2021.09.20.460798>
- 754 47. Nurk S, Koren S, Rhie A, Rautiainen M, Bizikadze AV, Mikheenko A, et al. The complete sequence of a
755 human genome. Science. 2022;376:44–53.
- 756 48. Beye M, Gattermeier I, Hasselmann M, Gempe T, Schioett M, Baines JF, et al. Exceptionally high levels
757 of recombination across the honey bee genome. Genome Res. 2006;16:1339–44.
- 758 49. Solognac M, Mougel F, Vautrin D, Monnerot M, Cornuet JM. A third-generation microsatellite-based
759 linkage map of the honey bee, *Apis mellifera*, and its comparison with the sequence-based physical map.
760 Genome Biol. 2007;8:R66.
- 761 50. Wallberg A, Glémin S, Webster MT. Extreme Recombination Frequencies Shape Genome Variation and
762 Evolution in the Honeybee, *Apis mellifera*. PLoS Genet. 2015;11:e1005189.
- 763 51. Kaskinova M, Yunusbayev B, Altinbaev R, Raffiudin R, Carpenter MH, Kwon HW, et al. Improved *Apis*
764 *mellifera* reference genome based on the alternative long-read-based assemblies. G3
765 Genes|Genomes|Genetics. 2021;11:jkab223.
- 766 52. Beye M, Moritz RF. Characterization of honeybee (*Apis mellifera* L.) chromosomes using repetitive
767 DNA probes and fluorescence in situ hybridization. J Hered. 1995;86:145–50.
- 768 53. Hastings PJ, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. Nat Rev
769 Genet. 2009;10:551–64.
- 770 54. Pennisi E. Sequencing all life captivates biologists. Science. 2017;355:894–5.
- 771 55. Sirén J, Monlong J, Chang X, Novak AM, Eizenga JM, Markello C, et al. Pangenomics enables
772 genotyping of known structural variants in 5202 diverse genomes. Science. 2021;374:abg8871.
- 773
- 774
- 775

776 **Figures**

777 **Figure 1. Orientation of the AMelMel1.1 contig presenting an inversion on chromosome 7** 778 **when compared to HAv3.1.**

779 The repeats present at the boundary between the contigs were used to orient the AMelMel1.1 contig
780 on chromosome 7. Assemblies with one or the other orientation of the contig were self-aligned with
781 LAST. Left: orientation from AMelMel1.0 and right, orientation from AMelMel1.1. For each pair
782 of alignments, only the junction between contigs are shown: the two ends of the contig to orient, the
783 end of the previous and the start of next contigs. Results clearly show the orientation in
784 AMelMel1.1 is the correct one.

785 **Figure 2. Comparison of Amel4.5 and AMelMel1.1 assemblies for chromosome 3.**

786 Abscissa: AMelMel, ordinate: Amel4.5. AMelMel contig borders are represented with vertical
787 dotted lines. Additionally, for both Amel4.5 and AMelMel, the position and number of
788 recombination events detected along the chromosome are represented for each interval flanked by
789 informative markers in the meiosis analyzed. Average SNP density and recombination rate are
790 given for 1Mb windows. Regions indicated in red on the Amel4.5 assembly represent
791 recombination ‘hotspots’ regions where number of recombination events between two informative
792 SNPs is higher than five. See supplementary data for the other chromosomes.

793 **Figure 3. Tandem repeats of period size 90-371 bp detected in the AMelMel1.1 assembly.** The
794 colour scale represents the period size of the repeat elements and the Y axis the total length of the
795 repeat array. Vertical dotted lines represent the contig boundaries in the AMelMel1.1 assembly. The
796 position of *AluI* and *AvaI* repeats are indicated with the number of repeats in parentheses. The
797 figure shows clearly, that most contigs are separated by tandem repeats of period size close to 371
798 bp, of length in the order of 10 kb or more. See also Supplementary file 1: Fig S11 for repeats of

799 longer period size. (1000-2000 bp). Although not represented on the graph (period size = 5),
800 TTAGG telomere are indicated with the number of repeats in parentheses, when present at a
801 chromosome end.

802 **Figure 4. Phylogenetic trees for the tandem repeats of period size 91-93 and 367-371 bp.**

803 Only tandem repeats with ten or more elements, such as detected by Tandem Repeat Finder, were
804 considered. Left: phylogenetic tree for the 74 sequences with a period size of 367-371 bp; right:
805 phylogenetic tree for the 43 sequences with a period size of 91-93 bp. The vertical red lines indicate
806 the cut-off that was used to define the groups of sequence based on similarity.

807 **Figure 5. Differences in copy numbers for 5S RNA ribosomal genes.** Two of the loci containing
808 5S RNA genes, present at 15 kb distance on chromosome 3 are shown. Top: screenshot of the NCBI
809 genome viewer for the region showing the annotation for the 5S RNA genes. Bottom: dotplot
810 alignment of HAv3.1 (x-axis) and AMelMel1.1 (y-axis) in the region. The first group of genes in
811 the bottom left contains seven genes in HAv3.1 and twenty in AMelMel1.1 on the forward strand.
812 The second in the top right contains eleven genes in HAv3.1 and eight in AMelMel1.1 on the
813 reverse strand. The red lines off diagonal show the sequence similarity between the two groups of
814 genes and indicate the two gene clusters are in reverse orientation.

815 **Figure 6. A 10 kb inverted duplication on chromosome 3 between HAv3.1 and AMelMel1.1.**

816 Bottom right: a dot plot representation of the alignment with LAST of AMelMel1.1 to HAv3.1
817 show a 10 kb inversion on chromosome 3. Self-alignments of AMelMel1.1 (left) and HAv3.1 (top)
818 show that the latter has an inverted repeated sequence in the region. The vertical yellow lines show
819 the position of repeats that were previously detected and shown in the NCBI annotation (grey
820 boxes, bottom) and are also found in our LAST alignments. NCBI annotation of genes are in green.

821 **Figure 7. Insertions and deletions in *Apis mellifera* subspecies.**

822 Analysis of NUMT insertions detected in only one assembly. Top: dotplot representation of LAST
823 alignments between the two assemblies show a 745 bp variant present in AMelMel1.1 on
824 chromosome 2 and absent in HAv3.1 (left) and a 576 bp variant present in HAv3.1 chromosome 10
825 and absent in AMelMel1.1 (right). For each variant, sequencing depths were evaluated on the
826 reference in which it is present. Middle: mean sequencing depth over 80 samples (red) shows a drop
827 coinciding with the position of the variants, suggesting that a significant proportion of samples may
828 lack the corresponding segment and standard deviation (blue) increases in the same region,
829 confirming the heterogeneity of the samples for the presence or absence of the variant. Bottom:
830 mean sequencing depth per subspecies, with *A. m. caucasia* (15 samples) in green, *A. m. ligustica*
831 (30 samples) in yellow and *A. m. mellifera* (35 samples) in black. Results suggest most of the *A. m.*
832 *mellifera* samples contain the insertion present in the AMelMel1.1 assembly on chromosome 2, as
833 the sequencing depth remains constant throughout the region, and not the one present in the HAv3.1
834 assembly on chromosome 10, as indicated by a sequencing depth close to zero. Inversely, most of
835 the *A. m. ligustica* samples contain the insertion present in the HAv3.1 assembly on chromosome 10
836 and not the one in the AMelMel1.1 assembly on chromosome 2. Most *A. m. caucasia* samples lack
837 the insertion present in the AMelMel1.1 assembly and a few seem to have the insertion present in
838 the HAv3.1 assembly.

839 **Figure 8: Comparing the indel variant calling between sequencing depth analysis and**

840 **GraphTyper2.** The Presence or absence of the NUMTs in the samples was evaluated by the
841 pangenome graph approach with GraphTyper2 (x-axis) and by estimating the sequencing depth at
842 the position of the NUMTs on the genome in which it is present (y-axis). Sequencing depths were
843 normalised by calculating the ratio between sequencing depth at the position of the NUMT
844 sequence and that of the flanking sequence. Nine out of 80 samples (11 %) could not be called for

845 NUMT_Chr2 and 19 (24 %) for NUMT_Chr10. When alleles could be called by Graphtyper2,
846 results agreed with the data based on sequencing depth.

847

848

849 **Tables**

850 **Table 1: Literature comparison of *Apis mellifera* genetic maps**

851

	Data	Physical size (Mb)	Genetic size (M)	CO/chromosome	cM/Mb
Hunt and Page (1995)	microsatellites	178	34.5	4.3	19.4
Solignac et al. (2004)	microsatellites	178	40.6	-	22.8
Solignac et al. (2007)	microsatellites	186	40	-	22.04
Beye et al. (2006)	microsatellites	238	45.5	5.7	19
Liu et al. (2015)	SNP	220	81.4	5.1	37
Wallberg et al. (2015)	SNP	229	59.5	-	26
Wallberg et al. (2019)	SNP	219	47.3	-	21.6
AMelMel1.1	SNP	220	50	3.1	23

852

853

854 **Additional files**

855 **Additional file 1 Supplementary methods and Supplementary Figures S1-S18**

856 Format: pdf

857 Title: Supplementary methods and supplementary figures S1-S11

858 Description:

859 **Additional file 2 Tables S1-S11**

860 Format: Excel file

861 Title: Supplementary tables S1-S11

862 Description:

863 **Additional file 3 AMelMel-Hav3**

864 Format: pdf

865 Title: Comparison of AMelMel1.1 and HAv3.1 genome assemblies.

866 Description: Dot plot alignments of the AMelMel1.1 and HAv3.1 genome assemblies.

867 **Additional file 4 AMelMel-Amel4_5**

868 Format: pdf

869 Title: Comparison of AMelMel1.1 and Amel4.5 assemblies

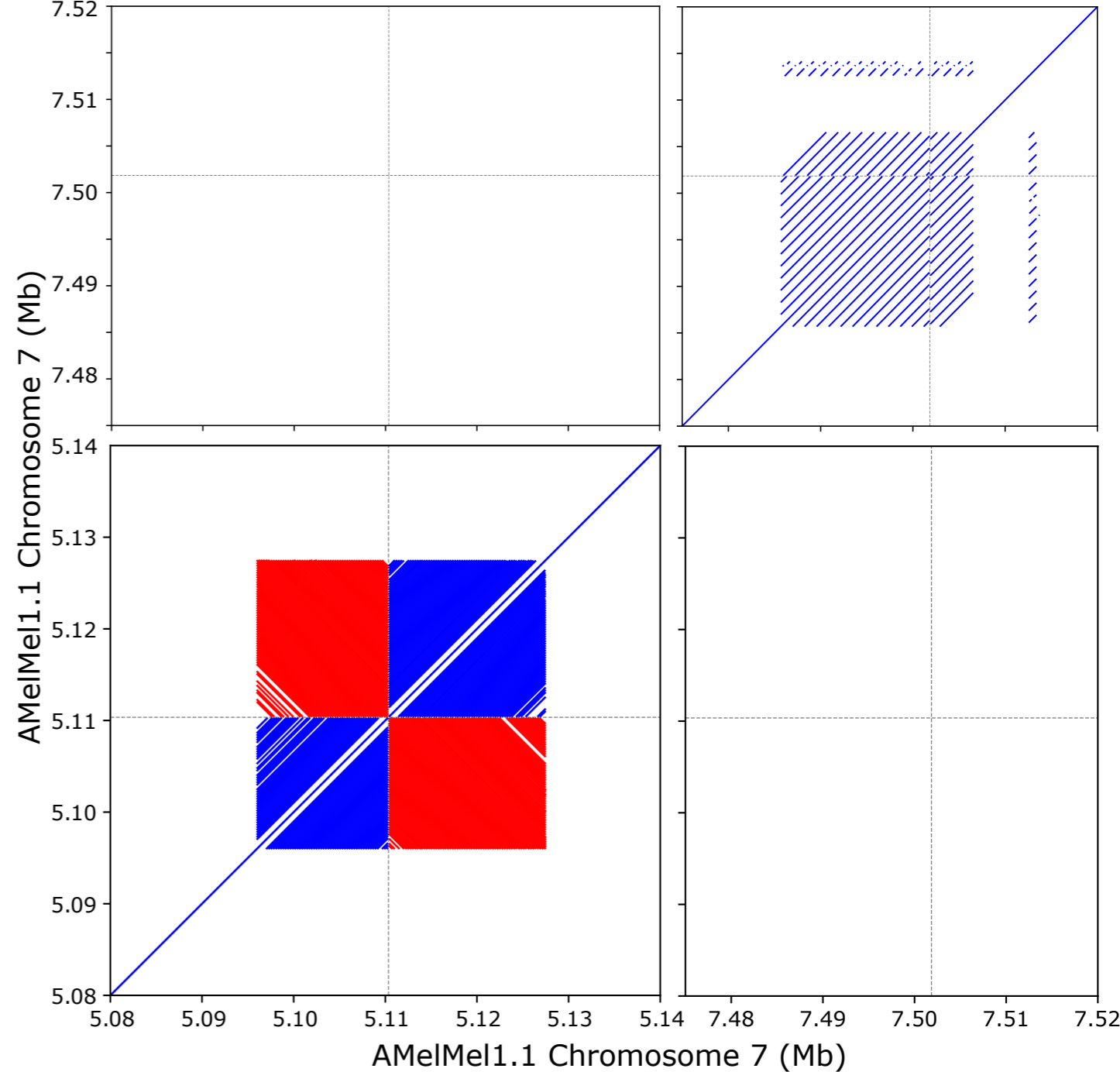
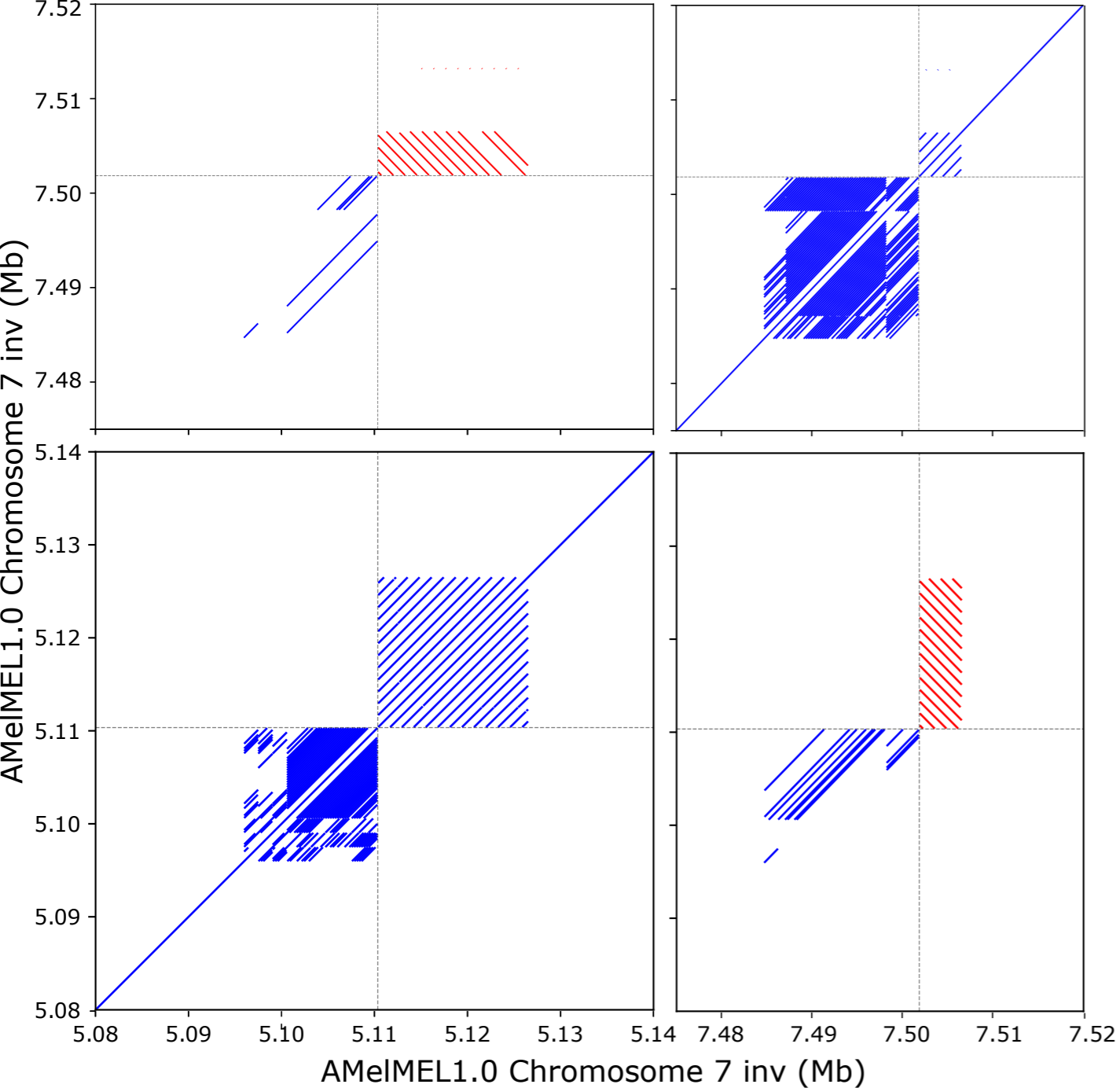
870 Description: Dot plot alignments of the AMelMel1.1 and HAv3.1 genome assemblies.

871 **Additional file 5 Inversions**

872 Format: pdf

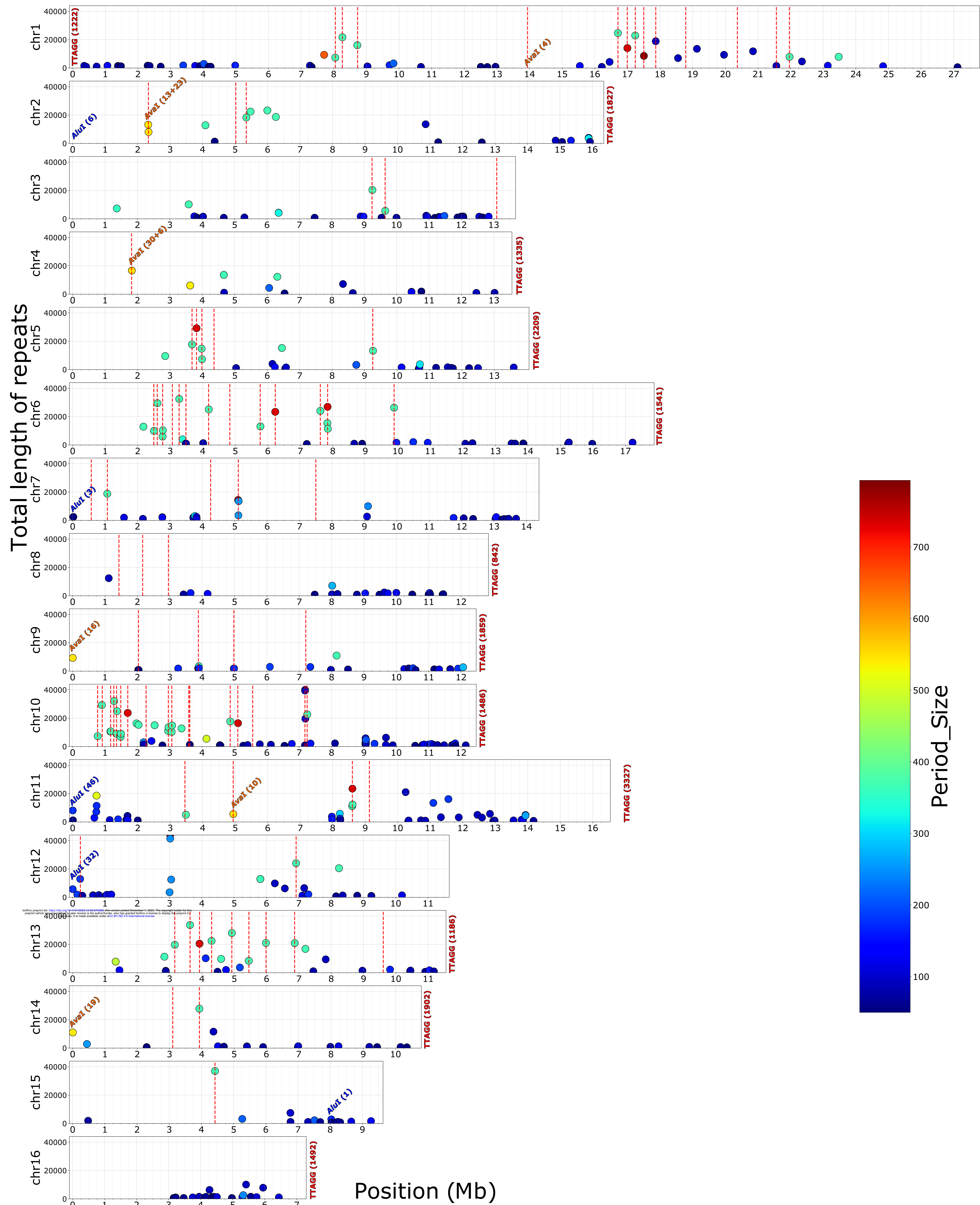
873 Title: Inversions larger than 1 kb detected between the AMelMel1.1 and HAv3.1 genome
874 assemblies.

875 Description: Inversion structural variants larger than 1 kb, detected after aligning the AMelMel1.1
876 and HAv3.1 genome assemblies with LAST.

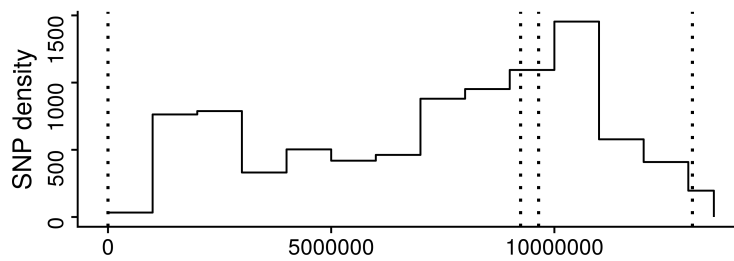
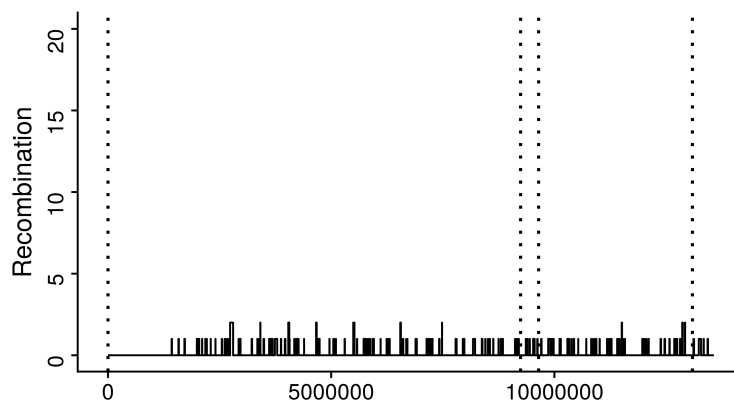
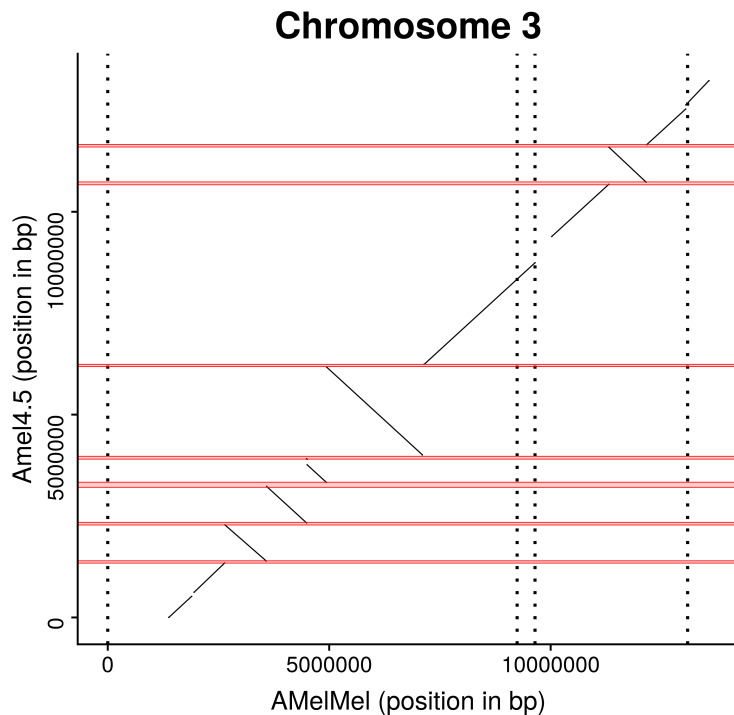
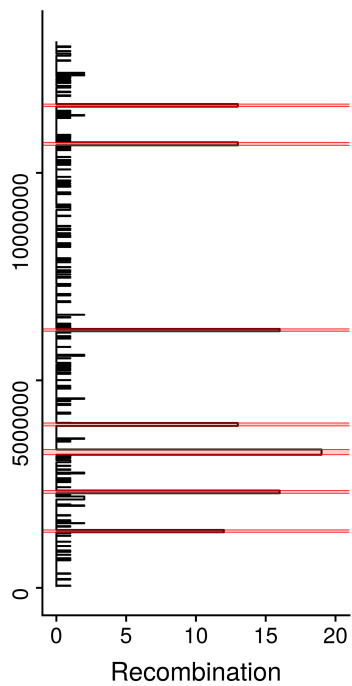
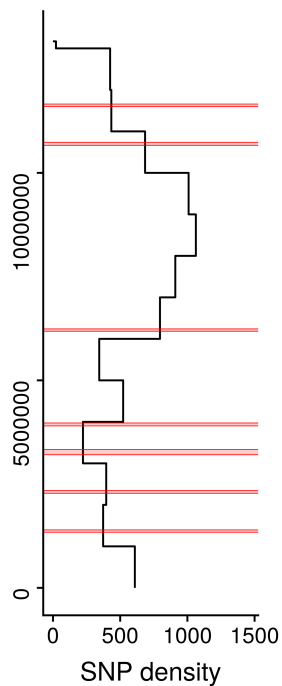


Tandem repeats

Period size: 50-1000 bp. Nb repeats: 10-1000

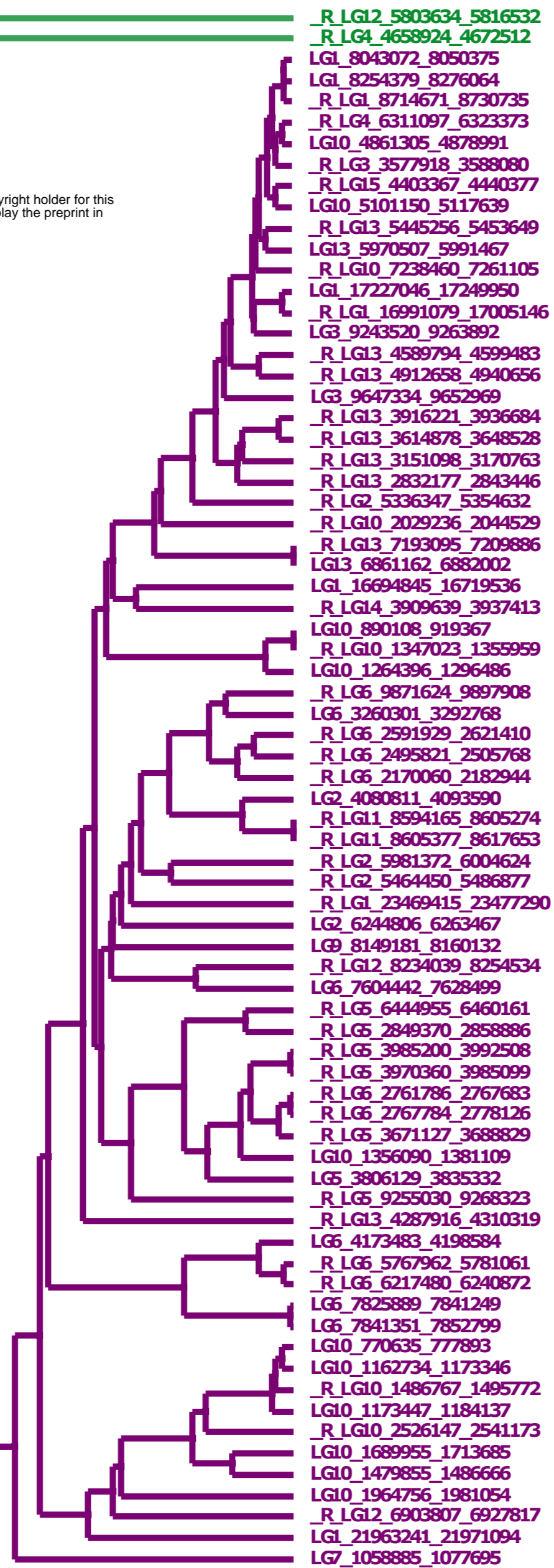


bioRxiv preprint doi: <https://doi.org/10.1101/2020.12.08.358888>; this version posted December 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

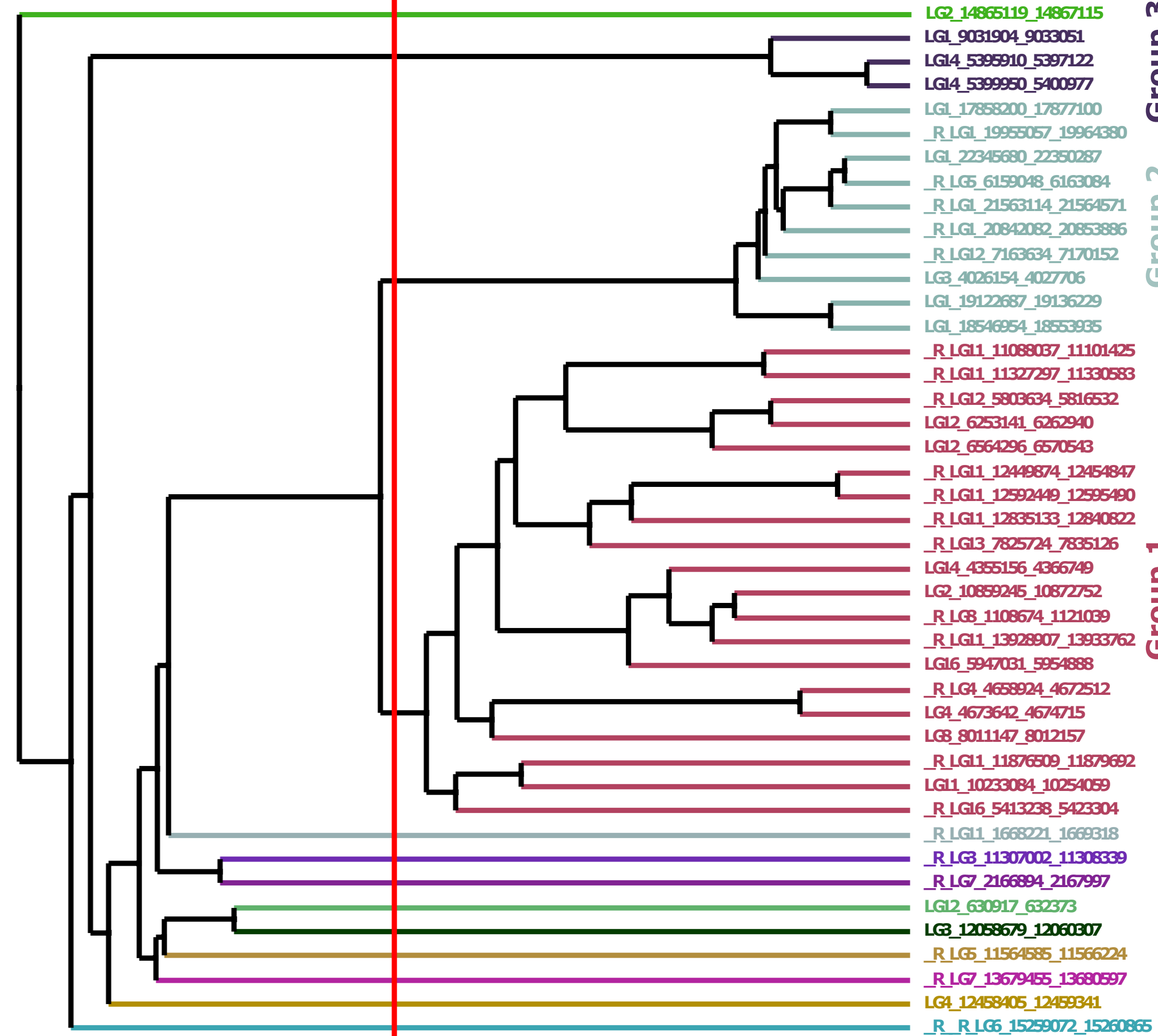


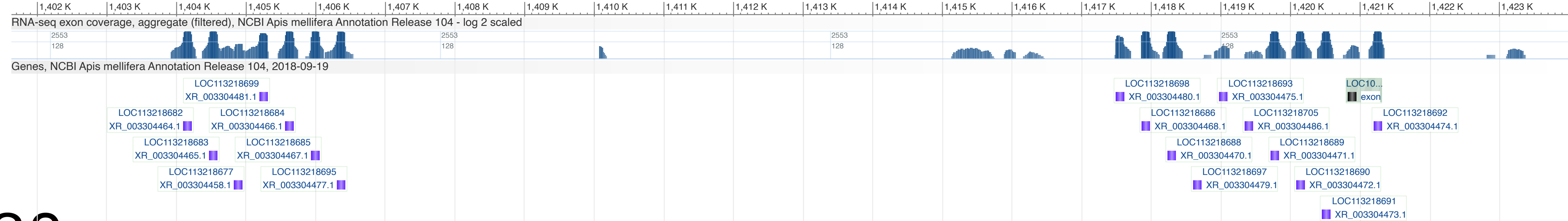
371 bp repeat

bioRxiv preprint doi: <https://doi.org/10.1101/2023.12.06.570386>; this version posted December 7, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.



91 bp repeat





AMElMel1.1 chromosome 3 (Mb)

1.380

1.375

1.370

1.365

1.360

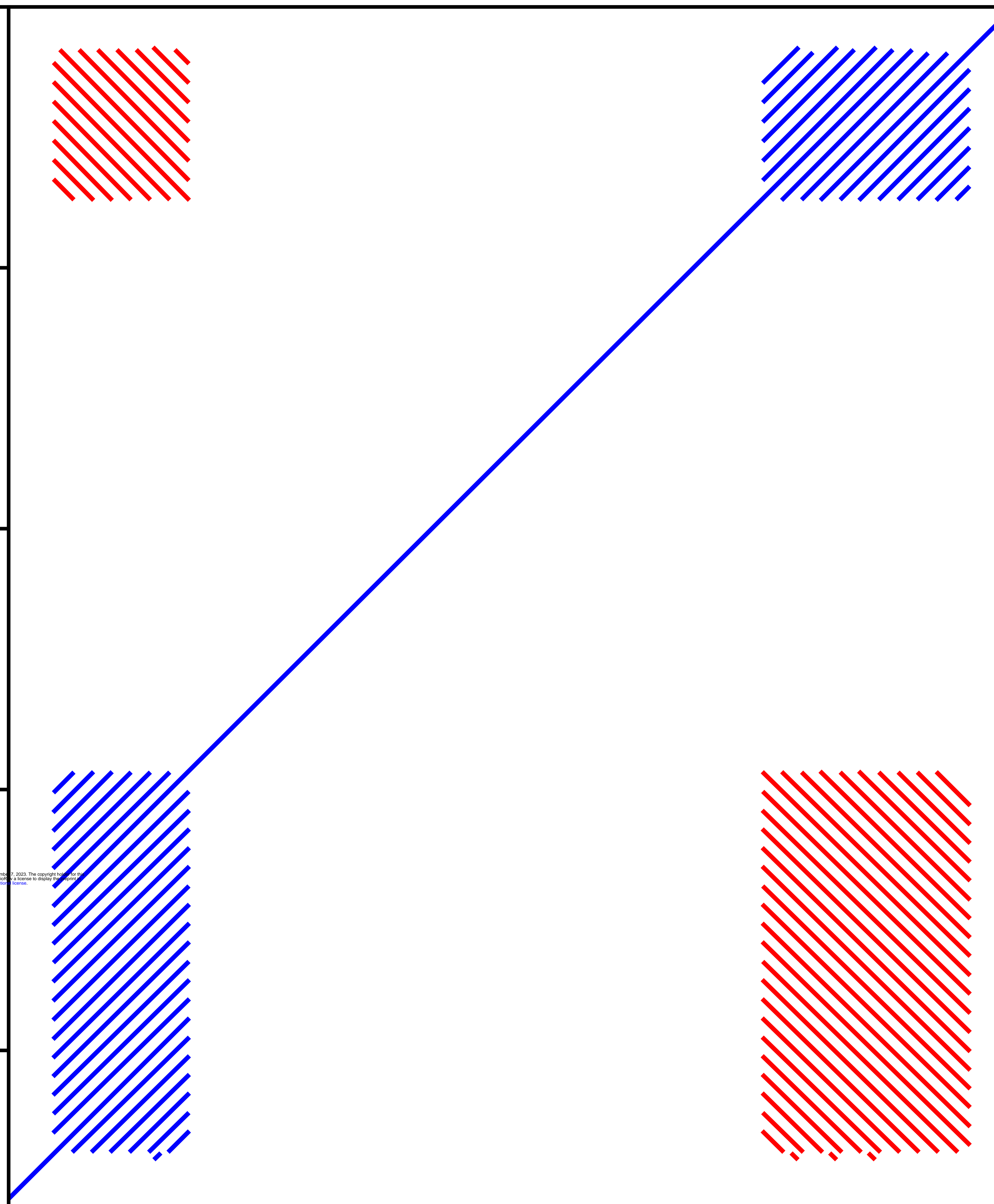
1.405

1.410

1.415

1.420

HAv3.1 chromosome 3 (Mb)



bioRxiv preprint doi: <https://doi.org/10.1101/2023.12.06.570386>; this version posted December 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

