Chromosome-scale assembly of the yellow mealworm genome

Evangelia Eleftheriou[1] , Jean-Marc Aury[1] , Benoit Vacherie[2] , Benjamin Istace[1] , Caroline Belser[1] , Benjamin Noel[1] , Yannick Moret[3] , Thierry Rigaud[3] , Fabrice Berro[4], Sona Gasparian[4], Karine Labadie-Breteau[2] , Thomas Lefebvre[4] , and Mohammed-Amin Madoui[1*],

[1]Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, 91057 Evry, France

[2]Genoscope, Institut de biologie François Jacob, Commissariat à l'Energie Atomique (CEA), Université Paris-Saclay, 91057 Evry France

[3]Équipe Écologie Évolutive, UMR CNRS 6282 BioGéoSciences, Université Bourgogne-Franche Comté, 21000 Dijon, France

[4]Ynsect, 91000 Evry, France

[*]corresponding author: amadoui@genoscope.cns.fr

Table of Contents

Supplementary Table 1: Genomic data

Nanopore reads from different runs on PromethION and MinION devices were merged.

	BG (male pupa)	BG (male pupa)	BG (male pupa)	BO (male pupa)
Sequencing Technology	ONT Promethion	ONT Minion	HiSeq 4000	HiSeq 4000
Library Protocol	1D genomic DNA by ligation	1D gDNA by selecting for long reads	PCR free	Dovetail Hi-C
Nb of Reads	2,264,874	217,053	162,443,183 x 2	124,589,387 x 2
Read length (mean for ONT*)	19,129	9,763	151 x 2	151 x 2
Max Read length (only for ONT)	247,043	181,352		
Cummulative size	45,443,965,268		47,496,139,168	37,141,198,341
N50 (only for ONT)	34,818			
Coverage *	146 x		153x	119x
GC%	39.75%		39.29%	39.30%

** for a genome of 310Mb based on GenomeScope Profile*
**ONT: Oxford Nanopore Technologies*

Supplementary Table 2: Transcriptomic data

AA and AB samples are pools of 150 insects each and available for downloading at NCBI. The rest six samples were sequenced at Genoscope, Institut François Jacob, CEA in Evry, France.

	AA (pool of 150 insects) NCBI	AB (pool of 150 insects) NCBI	AE (female pupa)	AF (female adult)	AH (sterile larva)	AI (sterile male adult)	AJ (sterile juvenile)	AK (sterile male pupa)
Sequencing Technology	Illumina HiSeq 1000	Illumina HiSeq 1000	NovaSeq 6000	NovaSeq 6000	NovaSeq 6000	NovaSeq 6000	NovaSeq 6000	NovaSeq 6000
Library Protocol	Maxima H minus first strand	Maxima H minus first strand	TruSeq reversely-stranded	TruSeq reversely-stranded	TruSeq reversely-stranded	TruSeq reversely-stranded	TruSeq reversely-stranded	TruSeq reversely-stranded
Nb of Reads	48,301,407 x 2	65,990,251 x 2	77,007,578 x 2	60,203,841 x 2	71,885,055 x 2	62,293,152 x 2	64,087,701 x 2	78,125,398 x 2
Read length (x2 paired-end)	100 x 2	100 x 2	151 x 2	151 x 2	151 x 2	151 x 2	151 x 2	151 x 2
Cummulative size	9,389,557,947	12,814,810,793	22,909,528,124	17,922,385,942	22,909,528,124	18,447,453,373	18,884,945,164	22,815,151,154
Coverage *	30x	41x	73x	57x	73x	59x	60x	73x
GC%	34.69%	35.27%	45.28%	44.07%	44.74%	43.82%	44.26%	45.21%

** for a genome of 310Mb based on GenomeScope Profile*

Supplementary Table 3: Metrics for long reads, contigs and scaffolds through different steps

	Long Reads	Long Reads (after YACRD)	Long Reads (after NECAT Pre-filtering)	NECAT Assembly	NECAT Assembly (after Polishing)	Haplotype (after Haplo-Merger2)	Polished Haplotype	Scaffolds (with Salsa2)	Final Assembly
# Contigs	2,481,927	2,372,861	250,277	527	527	187	187	138	112 (110 nuclear + 2 mitochondrial)
Max Contig Length	247,043	200,175	155,166	8,326,708	8,396,068	20,298,552	20,297,691	47,681,210	33,042,542
Mean Contig Length	18,310	15,807	49,546	814,527	819,341	1,541,657	1,541,607	2,089,196	2,570,819
Cumulative size	45,443,965,268	37,508,352,859	12,400,190,549	429,255,720	431,792,740	288,289,869	288,280,515	288,309,015	287,931,689
N50 (L50)	34,818 (448,059)	31,788 (398,831)	49,311 (100,050)	2,426,793 (57)	2,452,408 (57)	6,812,004 (12)	6,811,711 (12)	22,440,466 (5)	21,885,684 (6)
N90 (L90)	11,951 (1,274,614)	10,097 (1,170,536)	37,464 (216,023)	559,678 (179)	562,731 (179)	1,377,763 (48)	1,377,747 (48)	5,674,206 (15)	5,674,206 (16)
auN	37,542	34,524	52,859	2,778,417	2,798,350	8,771,730	8,771,103	22,535,111	18,643,178
GC%	39.75%	39.43%	38.57%	36.73%	36.66%	36.71%	36.72%	36.72%	36.72%

Supplementary Table 4: Repeats

Repeats for *T. molitor 2020* and *T. castaneum* were detected using same tools and parameters as for *T. molitor 2021 (*see main article*).* The portion of genome covered by repeats is 5-6% for the three assemblies.

	Tenebrio molitor 2021	**Tenebrio molitor 2020**	**Tribolium Castaneum**
assembly size	287,931,689 bp	280,780,514 bp	165,944,485 bp
contigs	112	31,390	2,082
GC level	36.72%	36.03%	33.86%
bases masked	17,298,313 bp (6.01 %)	14,784,139 bp (5.27 %)	10,284,212 bp (6.20 %)
	number of / length occupied / percentage of sequence elements*	number of / length occupied / percentage of sequence elements*	number of / length occupied / percentage of sequence elements*
DNA transposons:	**29,182 / 5,584,839 bp / 1.94%**	**38,447 / 6,101,177 bp / 2.17%**	**21,026/ 5,757,743 bp / 3.47%**
hAT-Charlie	354 / 75,994 bp / 0.03%	495 / 80,945 bp / 0.03%	129 / 20,631 bp / 0.01%
TcMar-Tigger	196 / 28,919 bp / 0.01%	293 / 37,247 bp / 0.01%	1,057 / 80,257 bp / 0.05%
SINEs:	539 / 32,334 bp / 0.01%	1,070 / 58,098 bp / 0.02%	281 / 31,050 bp / 0.02%
ALUs	1 / 50 bp / 0.00%	3 / 382 bp / 0.00%	1 / 45 bp / 0.00%
MIRs	2 / 65 bp / 0.00%	6 / 405 bp / 0.00%	0 / 0 bp / 0.00%
LINEs:	11,417 / 4,005,113 bp / 1.39%	12,754 / 3,472,876 bp / 1.24%	4,398 / 1,301,528 bp / 0.78%
LINE1	329 / 20,275 bp / 0.01%	634 / 37,990 bp / 0.01%	163 / 10,386 bp / 0.01%
LINE2	1,099 / 149,969 bp / 0.05%	1,381 / 168,228 bp / 0.06%	1,068 / 306,642 bp / 0.18%
L3/CR1	1,688 / 424,660 bp / 0.15%	2,181 / 455,200 bp / 0.16%	1,637 / 675,386 bp / 0.41%
LTR elements:	7,950 / 3,597,958 bp / 1.25%	10,290 / 2,306,092 bp / 0.82%	3,273 / 824,802 bp / 0.50%
ERVL	18 / 1,057 bp / 0.00%	43 / 2,179 bp / 0.00%	11 / 634 bp / 0.00%
ERVL-MaLRs	2 / 106 bp / 0.00%	4 / 224 bp / 0.00%	4 / 192 bp / 0.00%
ERV_classI	119 / 6,401 bp / 0.00%	279 / 14,905 bp / 0.01%	83 / 4,607 bp / 0.00%
ERV_classII	79 / 4,504 bp / 0.00%	150 / 8,170 bp / 0.00%	65 / 3,994 bp / 0.00%
Unclassified	1,790 / 228,410 bp / 0.08%	2,553 / 281,171 bp / 0.10%	547 / 76,839 bp / 0.05%
Small RNA	2,851 / 939,516 bp / 0.33%	612 / 75,855 bp / 0.03%	646 / 73,705 bp / 0.04%
Satellites	605 / 273,554 bp / 0.10%	404 / 78,560 bp / 0.03%	211 / 24,706 bp / 0.01%
Simple repeats	**49,992 / 2,094,562 bp / 0.73%**	**47,642 / 1,874,361 bp / 0.67%**	**36,699 / 1,708,043 bp / 1.03%**
Low complexity	11,960 / 575,359 bp / 0.20%	11,941 / 566,664 bp / 0.20%	10,271 / 504,269 bp / 0.30%

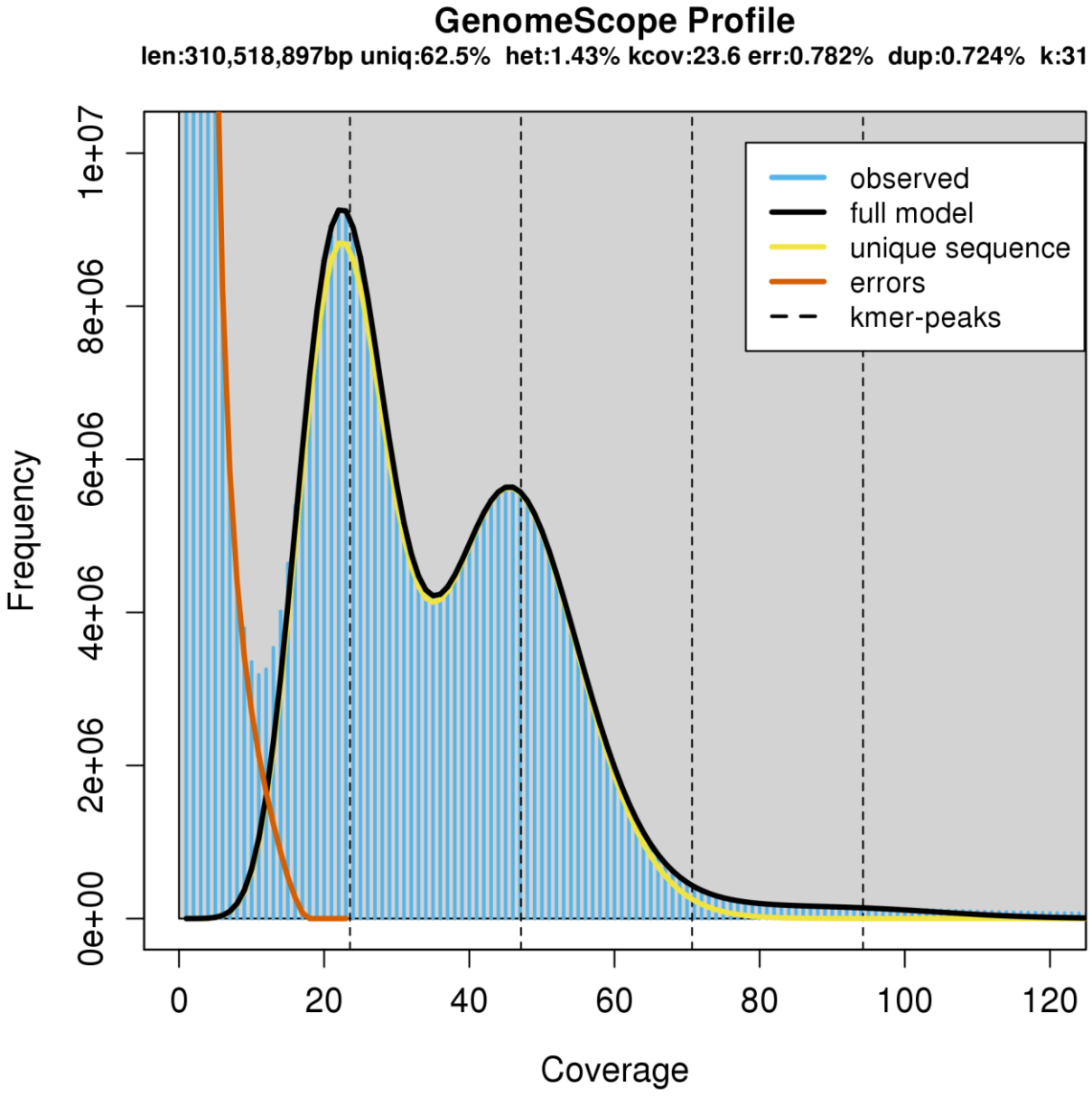
**most repeats fragmented by insertions or deletions have been counted as one element*

Supplementary Note 1: Gmove

Gmove is an easy-to-use predictor with no need of a pre-calibration step. Briefly, putative exons and introns, extracted from alignments, are used to build a graph, where nodes and edges represent, respectively, exons and introns. Gmove extracts all paths from the graph and searches open reading frames that are consistent with the protein evidence.

Supplementary Figure 1: GenomeScope Profile for *T. molitor*

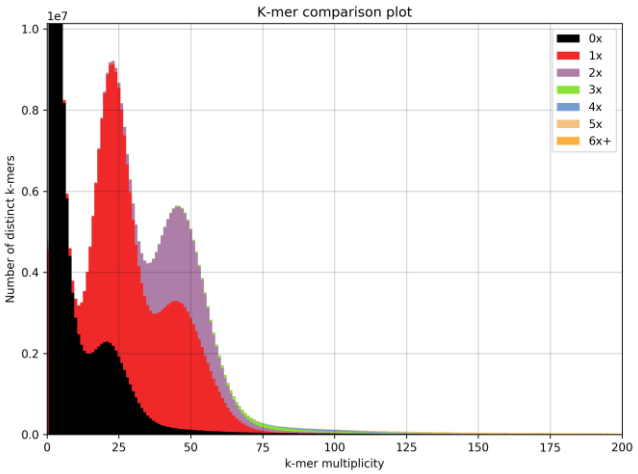
Gemome size is estimated at ~310Mb with an heterozygosity rate of 1.43%.



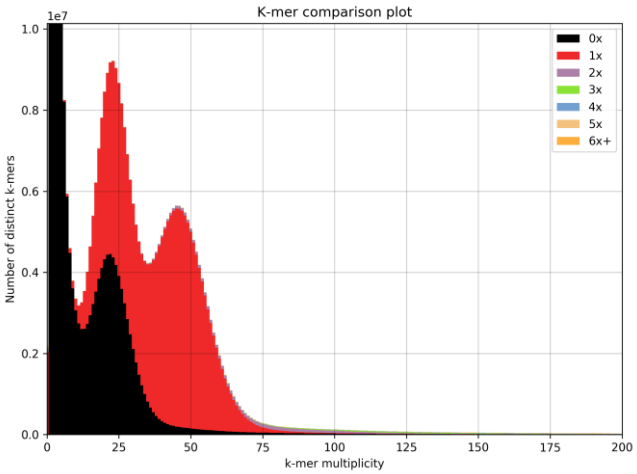
Supplementary Figure 2: K-mer plot before and after Haplomerger

Haplomerger2 efficiently eliminated the duplicated part of the assembly (purple curve on image a)

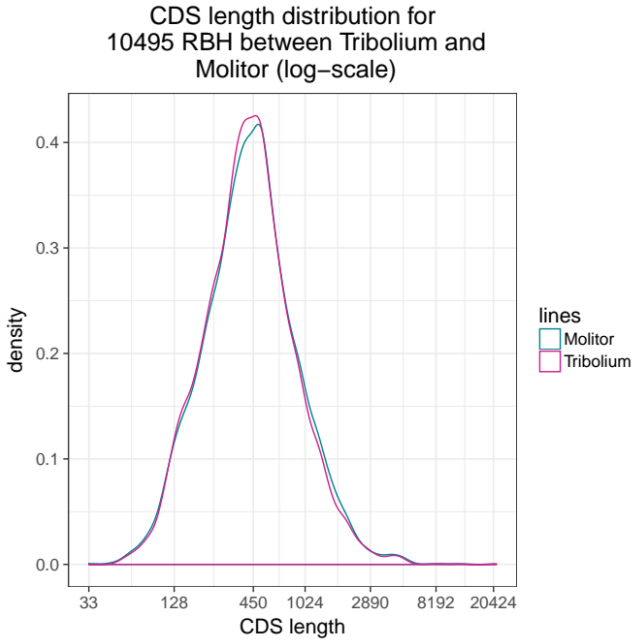
a) Before Haplomerger

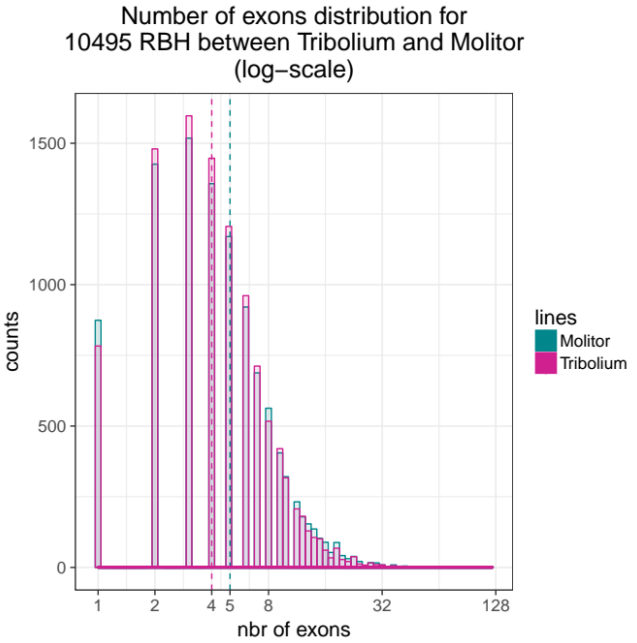


b) After Haplomerger



Supplementary Figure 3: Comparison of CDS lengths and number of exons of orthologous genes between *T. molitor* and *T. castaneum*

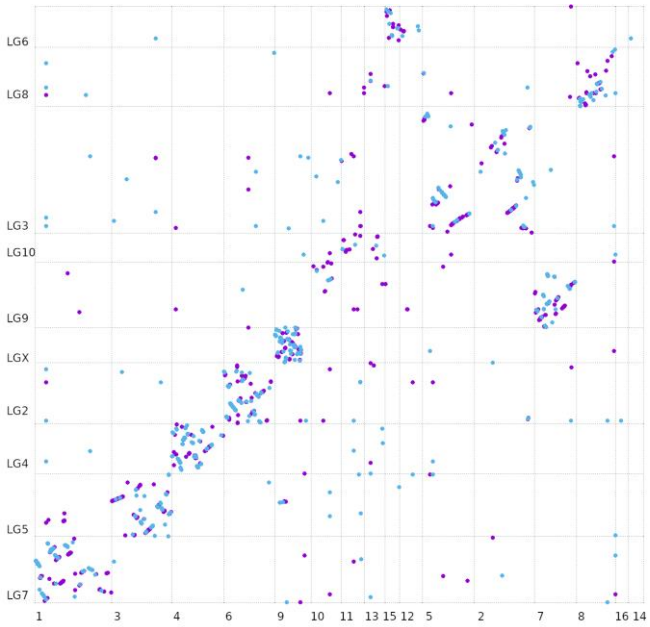




CDS lengths are log-transformed but we maintained real length values on the x-axis. The pink coloured curve corresponds to *T. castaneum* values and the turquoise to *T. molitor*. Mean CDS length is at 428 amino acids for *T. molitor* and 418 for *T. castaneum* . The trends of the two curves are quite similar.

Values for plotting are log-transformed but we maintained the real numbers on the axis. Pink and turquoise coloured curves correspond to *T. castaneum* and *T. molitor*, respectively. Dashed vertical lines represent the medan values (5 exons for *T. molitor* and 4 for *T. castaneum*).

Supplementary Figure 4: Aligning *T. molitor* to *T. castaneum*



Alignment (using NUCmer) between the 16 longest scaffolds of *T. molitor* (x-axis) and the 10 chromosomes of *T. castaneum* (y-axis).

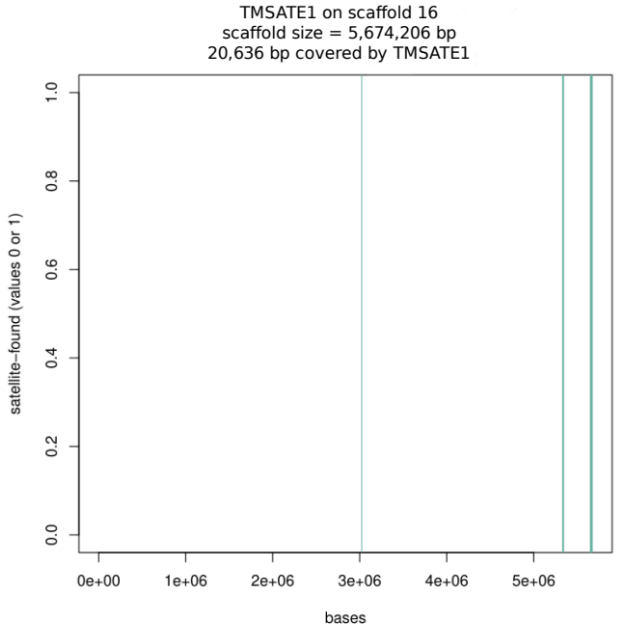
The alignment shows several clear associations between :
- scaffold 1 and LG7,
- scaffold 3 and LG5,
- scaffolds 2, 5 and LG3 (as if chromosome LG3 underwent a fission in *T. molitor*)
- scaffold 4 and LG4
- scaffold 6 and LG2
- scaffold 8 and LG8
- scaffolds 7 , 10 and LG9
- scaffold 9 and LGX

as well as some more ambiguous associations between:
- scaffolds 11, 13 and LG10
- scaffolds 12, 14, 15, 16 and LG6

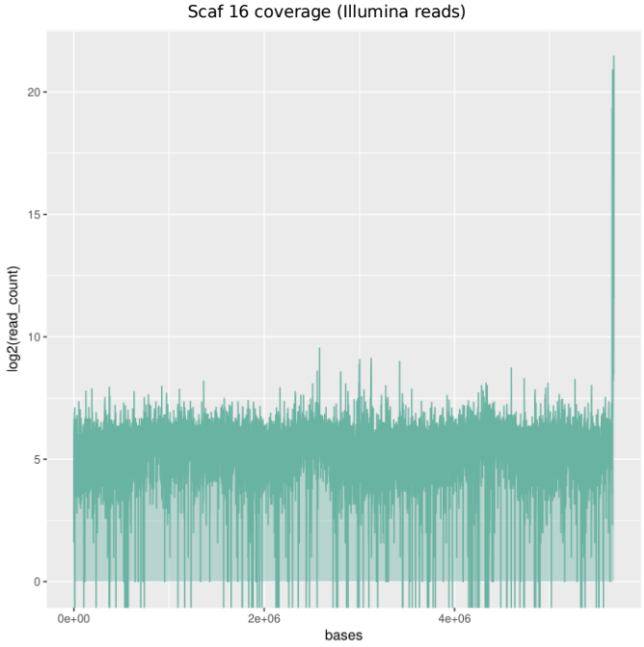
Supplementary Figure 5: Position of the 142 bp satellite (TMSATE1) on scaffolds 16, 58, 99, 23 and their coverage by Illumina Reads

We performed a BLASTn analysis for the 142-bp satellite sequence (BLASTn overlap >80%, identity ≥90%). The satellite was detected in 18 scaffolds (the 9 longest scaffolds 1-9 and 11, 13, 14, 16, 35, 46, 58, 99, 108) and extends to 174,807 bp accounting for 0.06% of the assembly.
(According to RepeatMasker results , the satellite extends to 248,412 bp accounting for 0.08 % of the assembly).
While it ranges from 1 to 17 instances (average 4.35) in 15 scaffolds , it is highly present in scaffolds 16, 58, 99 (135, 534, 511 instances respectively).
More particularly, in scaffold 58, under the BLASTn constraints of overlap >80% and identity score ≥90%, the satellite covers 77,501 bp (or ~47% of the scaffold's length) (Fig. 5e). According to RepeatMasker the satellite covers 84,915 bp or ~51.8% of the scaffold's length .
In scaffold 99 the region covered by the satellite measures 72,136 bp or almost 90% of the scaffold's length (Fig. 5g).
However, at the same time, scaffolds 58, 99 have an extremely high coverage by Illumina reads (Figures 5f, 5h),
probably indicating that different regions of the genome (containing the satellite) have been collapsed into single scaffolds (misassembled scaffolds due to erroneous sequence collapses).
RepeatMasker detected the 142bp satellite (TMSATE1) also in scaffolds 10, 12, 15, 17, 18, 19, 21, 23. BLASTn didn't find it there simply because of the strict constraints of overlap and identity score we set.
It is worth noting that in scaffold 23 a variant sequence of the satellite accounts for 4% of scaffold's proper length (Fig. 5c). This portion is much higher than the median 0.02%, calculated on scaffolds that have the satellite (excluding the extreme cases of scaffolds 58, 99).

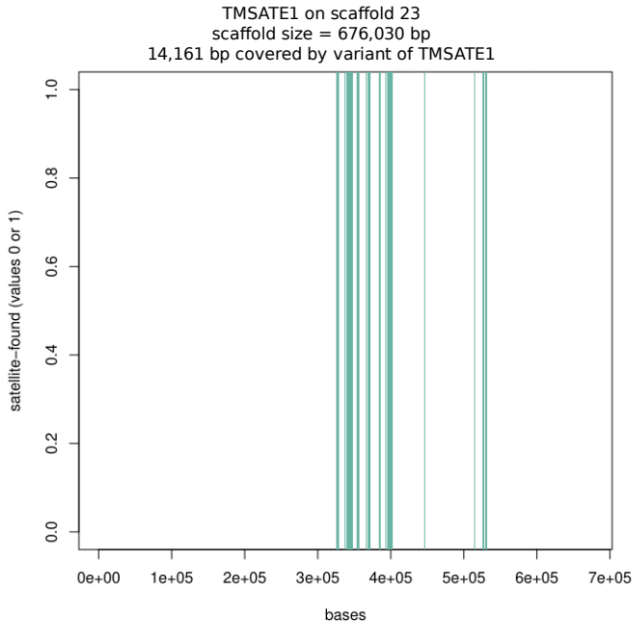
a


TMSATE1 on scaffold 16
scaffold size = 5,674,206 bp
20,636 bp covered by TMSATE1
satellite−found (values 0 or 1)
0.0
0.2
0.4
0.6
0.8
1.0
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
bases

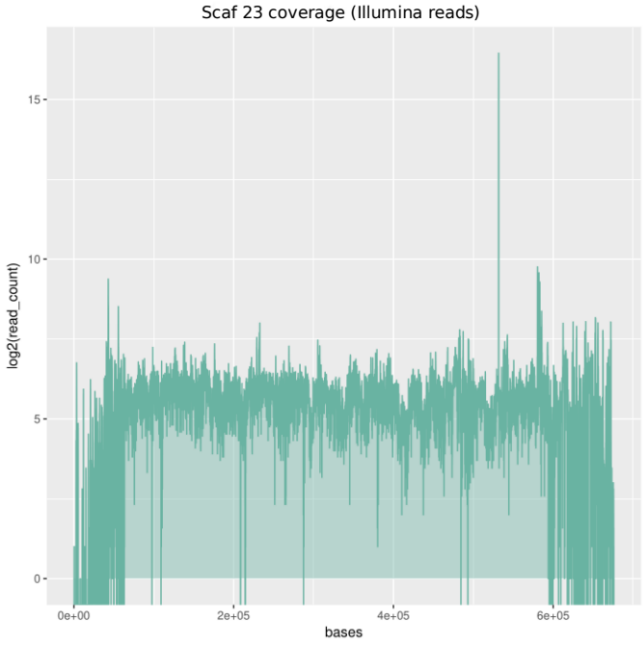
b


Scaf 16 coverage (Illumina reads)
log2(read_count)
0
5
10
15
20
0e+00
2e+06
4e+06
bases

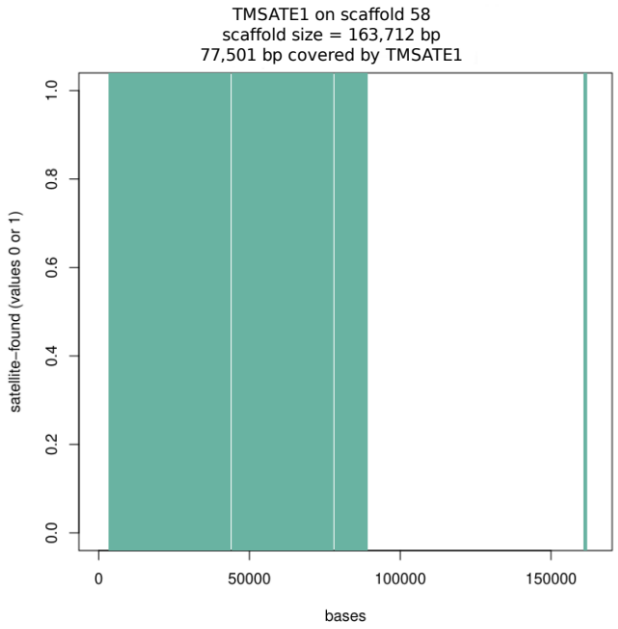
c


TMSATE1 on scaffold 23
scaffold size = 676,030 bp
14,161 bp covered by variant of TMSATE1
satellite−found (values 0 or 1)
0.0
0.2
0.4
0.6
0.8
1.0
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
7e+05
bases

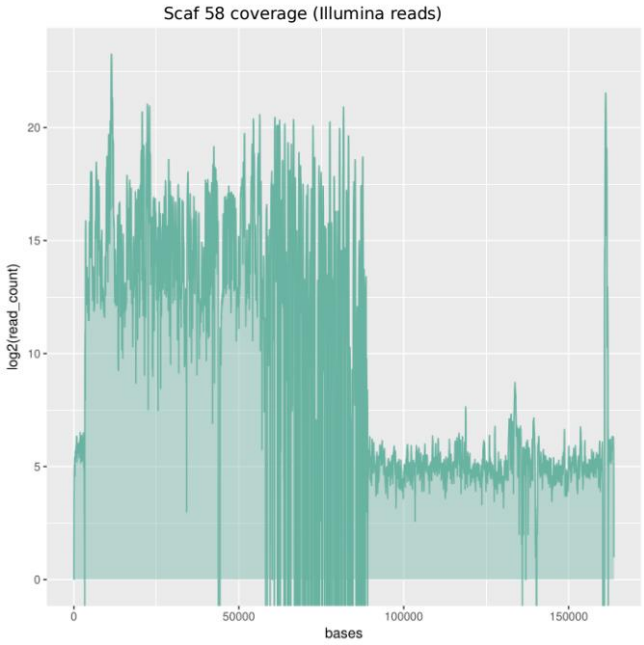
d


Scaf 23 coverage (Illumina reads)
log2(read_count)
0
5
10
15
0e+00
2e+05
4e+05
6e+05
bases

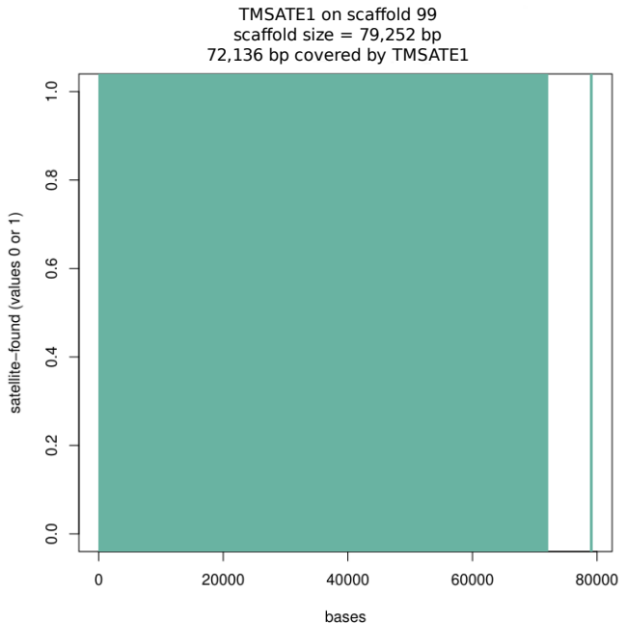
e


TMSATE1 on scaffold 58
scaffold size = 163,712 bp
77,501 bp covered by TMSATE1
satellite−found (values 0 or 1)
0.0
0.2
0.4
0.6
0.8
1.0
0
50000
100000
150000
bases

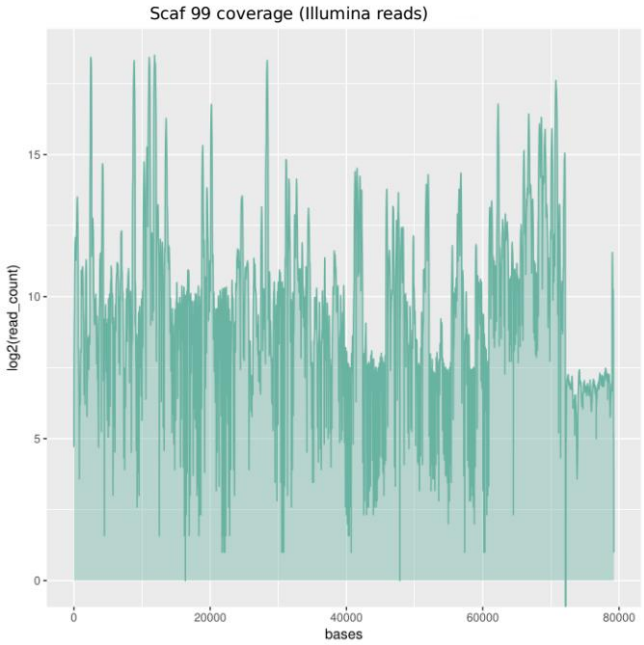
f


Scaf 58 coverage (Illumina reads)
log2(read_count)
0
5
10
15
20
0
50000
100000
150000
bases

g

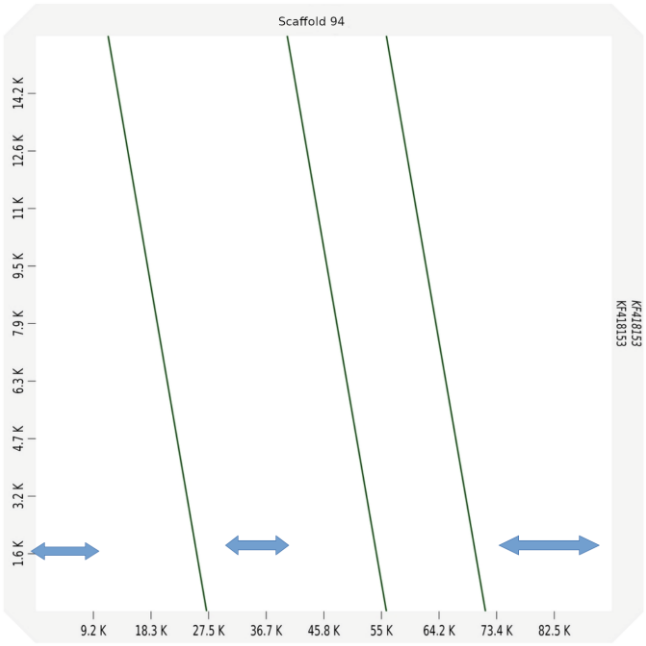

TMSATE1 on scaffold 99
scaffold size = 79,252 bp
72,136 bp covered by TMSATE1
satellite−found (values 0 or 1)
0.0
0.2
0.4
0.6
0.8
1.0
0
20000
40000
60000
80000
bases

h


Scaf 99 coverage (Illumina reads)
log2(read_count)
0
5
10
15
0
20000
40000
60000
80000
bases

Supplementary Figure 6: Presence of mitochondrial genome on scaffold 94

We aligned, with minimap2, the mitochondrial genome of *T. molitor* (with name KF418154 at NCBI) to the assembly. Then, we plotted the alignment using the d-genies platform of Toulouse (http://dgenies.toulouse.inra.fr/) (Fig. 6a).



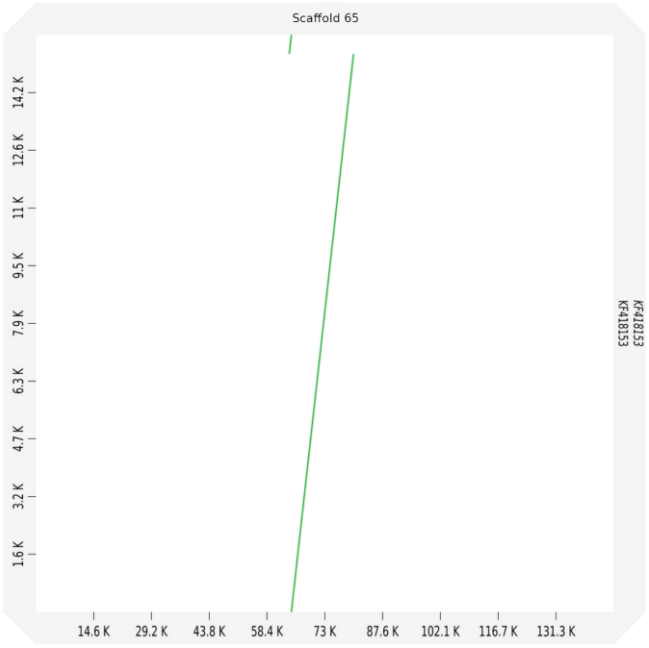
The figure on the left shows the alignment between the scaffold 94 (x-axis) and the mitochondrial genome (y-axis). The two sequences align with identity score 85-89% at three regions of the scaffold 94 (represented by green lines).

Then, we extracted the regions of scaffold 94 where mitochondrial genome did not align (represented by blue double arrows) and aligned the latter separately to each of these regions. The alignment identity scores were lower this time (lines in red text at the Table 5, right below).

Supplementary Table 5: Alignment between the mitochondrial genome and the scaffold 94

Mito seq	Length of mito seq (bp)	Start position on mito seq	End position on mito seq	Strand	Target	Length of target seq (bp)	Start position on scaffold	End position on scaffold	Length of matched seq (bp)	Length of aligned seq (bp)	Identity Score
KF418153	15785	1	12188	-	Scaf_94: **1-11513**	11513	21	11511	8938	12209	73%
KF418153	15785	16	15784	-	Scaf_94	91698	11513	27172	13564	15773	85%
KF418153	15785	1	15784	-	Scaf_94: **27172-40025**	12854	17	12852	11063	15785	70%
KF418153	15785	1	15784	-	Scaf_94	91698	40025	55803	14095	15793	89%
KF418153	15785	1	15784	-	Scaf_94	91698	55805	71572	13954	15788	88%
KF418153	15785	1	15784	-	Scaf_94: **71572-91698**	20127	3	13020	10779	15785	53%

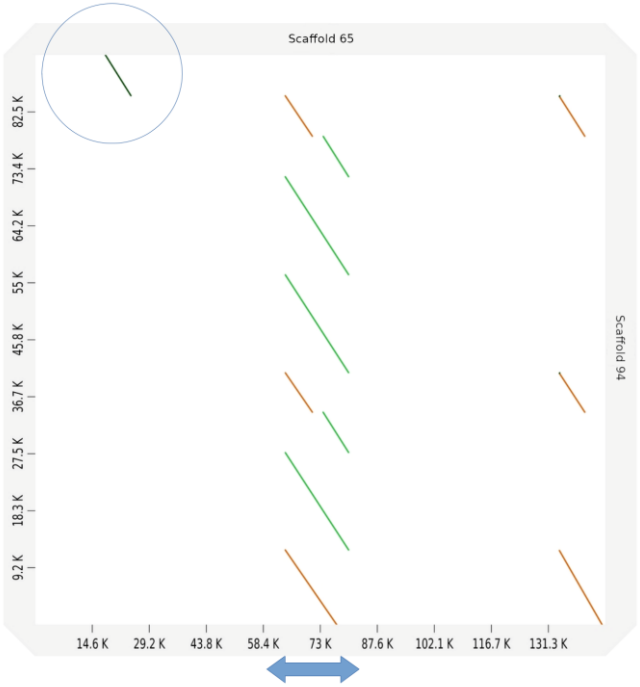
Supplementary Figure 7: Presence of mitochondrial genome on scaff 65



The figure on the left shows the alignment between the scaffold 65 (x-axis) and the mitochondrial genome (y-axis). The two sequences align with identity score 50-75%, lower than this of scaffold 94.

Supplementary Figure 8: Alignment of scaffolds 94, 65

The figure below shows how scaffolds 94 and 65 align to each other.



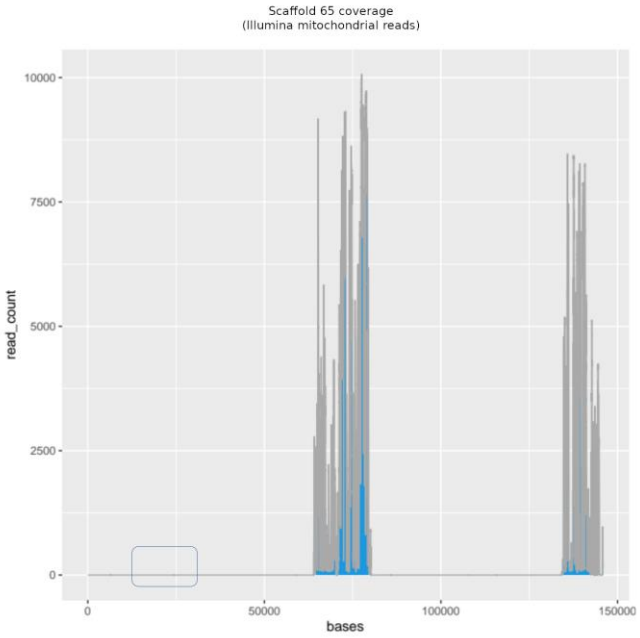
Scaffold 65 and 94 lie on x-axis and y-axis, respectively.
The part in the middle (blue double arrow) corresponds to the 15kb mitochondrial sequence while we also observe a better alignment between the two scaffolds located somewhere at the beginning of the scaffold 65 (dark green line in circle) and at the end of scaffold 94. The two sequences align also at the end of scaffold 65.

From the last line of the Table 5 (in red text) we observe that the mitochondrial sequence aligned from position 71,575 to 84,595 on scaffold 94. This means that there are almost 7kb bases at the end of scaffold 94 where mitochondrial genome does not align to and this region seems to be the common part between scaffolds 65 and 94 that we see on the figure in dark green line.

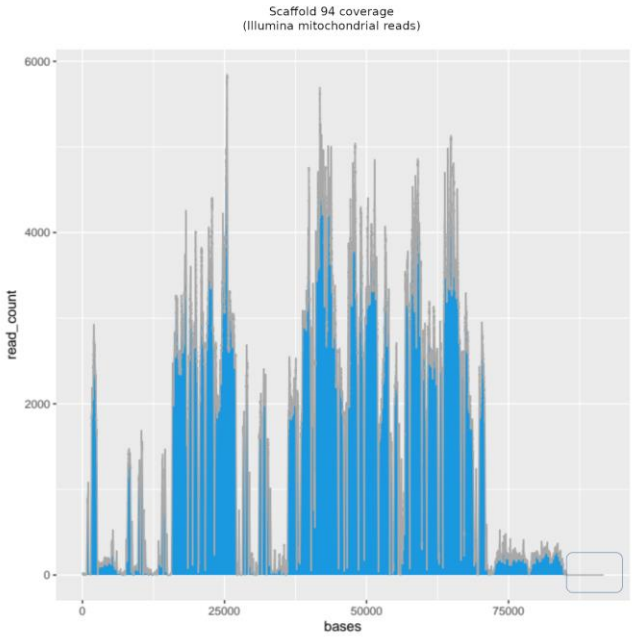
Supplementary Figure 9: Coverage of scaffolds 65, 94 by Illumina mitochondrial reads

We also aligned, with BWA-mem (parameter -B 8, identity score 90% and 80% overlap), Illumina reads to the mitochondrial sequence KF418153 in order to create a "pool" of high confidence mitochondrial reads. The figures below show the coverage of scaffolds 65 and 94 by mitochondrial reads. We confirm, once again, the existence of a common sequence between scaffolds 65 and 94 that is 7kb long (represented by a rectangular on the figures below) and where mitochondrial reads do not align to. This sequence could be an unknown, so far, part of *T. molitor* mitochondrial genome.

The 7kb region is located at the beginning of scaffold 65



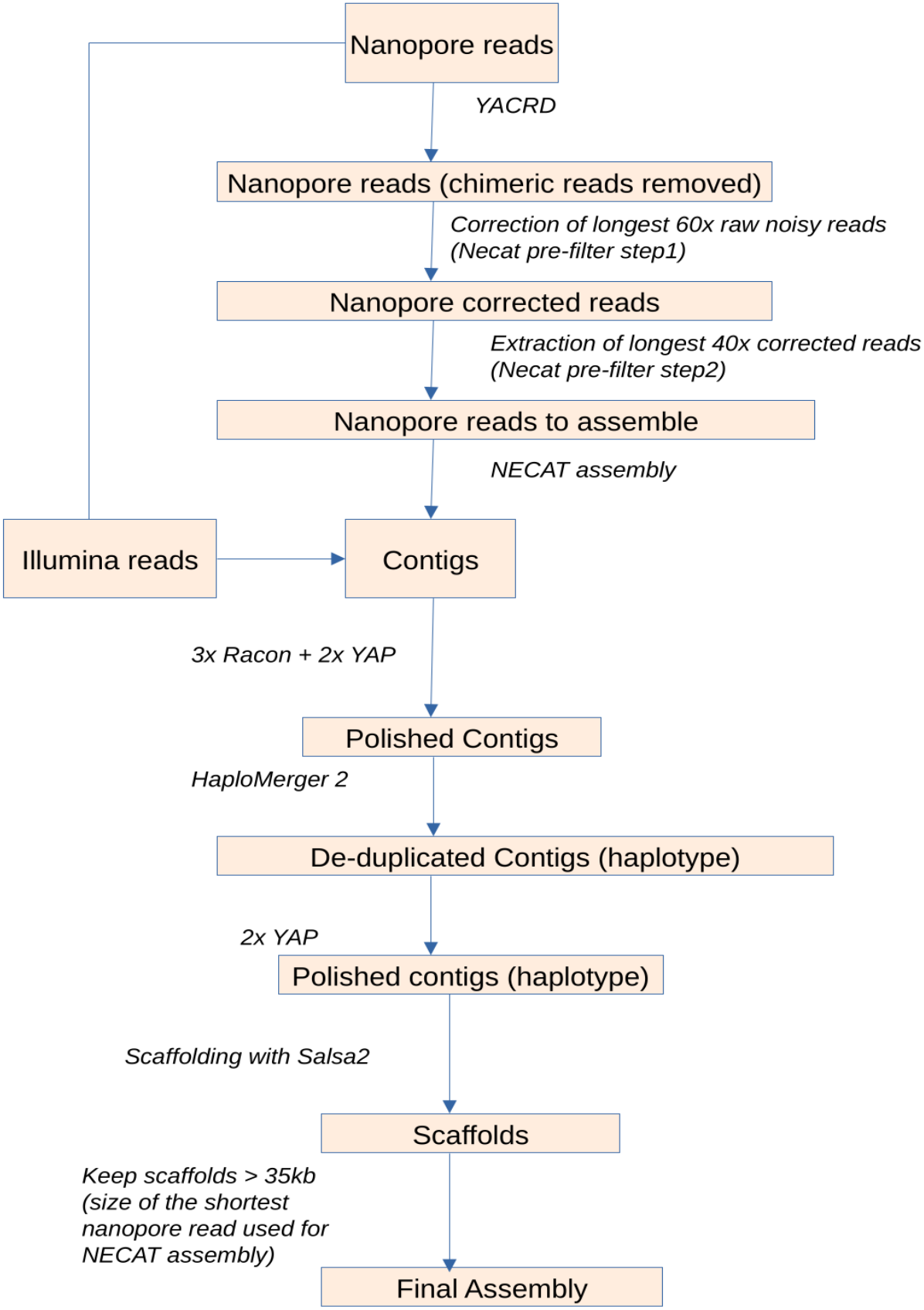
The 7kb region is located at the end of scaffold 94



Supplementary Table 6: Samples' accession numbers

Sample Type	Sex and Developmental stage	Sample Name	Sequencing Technology	Study Accession	Submission Accession	Sample Accession	Experiment Accession	Run or Analysis Accession
DNA	Male pupa	CPD_BG_ONT_1_FAL24795_A	Oxford Nanopore	ERP128758	ERA4143595	ERS6344234	ERX5504889	ERR5859001
DNA	Male pupa	CPD_BG_ONT_1_PAD99440_A	Oxford Nanopore	ERP128758	ERA4143595	ERS6344234	ERX5504890	ERR5859002
DNA	Male pupa	CPD_BGOSDE_6_HFHC5BBXY.12B	Illumina PCR-free	ERP128758	ERA4143595	ERS6344234	ERX5504891	ERR5859003
DNA	Male pupa	CPD_BGOSDE_4_HFWM7BBXY.12B	Illumina PCR-free	ERP128758	ERA4143595	ERS6344234	ERX5504892	ERR5859004
DNA	Male pupa	CPD_BOOSDF_8_HGJYYB	Hi-C	ERP128758	ERA4185802		ERX5513830	ERR5870148
RNA	Female pupa	CPF_AEOSRB_1_CVTP3.12BA090	NovaSeq	ERP128775	ERA4146275	ERS6348205	ERX5508089	ERR5862256
RNA	Female pupa	CPF_AEOSRB_4_H2TK3DSXY.12B	NovaSeq	ERP128775	ERA4146275	ERS6348205	ERX5508095	ERR5862262
RNA	Female adult	CPF_AFOSRB_1_CVTP3.12BA091	NovaSeq	ERP128775	ERA4146275	ERS6348205	ERX5508090	ERR5862257
RNA	Female adult	CPF_AFOSRB_4_H2TK3DSXY.12B	NovaSeq	ERP128775	ERA4146275	ERS6348205	ERX5508096	ERR5862263
RNA	Sterile larva	CPF_AHOSRB_1_CVTP3.12BA093	NovaSeq	ERP128775	ERA4146275	ERS6348205	ERX5508091	ERR5862258
RNA	Sterile larva	CPF_AHOSRB_4_H2TK3DSXY.12B	NovaSeq	ERP128775	ERA4146275	ERS6348205	ERX5508097	ERR5862264
RNA	Sterile male adult	CPF_AIOSRB_1_CVTP3.12BA094	NovaSeq	ERP128775	ERA4146275	ERS6348205	ERX5508092	ERR5862259
RNA	Sterile male adult	CPF_AIOSRB_4_H2TK3DSXY.12B	NovaSeq	ERP128775	ERA4146275	ERS6348205	ERX5508098	ERR5862265
RNA	Sterile juvenile	CPF_AJOSRB_1_CVTP3.12BA095	NovaSeq	ERP128775	ERA4146275	ERS6348205	ERX5508093	ERR5862260
RNA	Sterile juvenile	CPF_AJOSRB_4_H2TK3DSXY.12B	NovaSeq	ERP128775	ERA4146275	ERS6348205	ERX5508099	ERR5862266
RNA	Sterile male pupa	CPF_AKOSRB_1_CVTP3.12BA096	NovaSeq	ERP128775	ERA4146275	ERS6348205	ERX5508094	ERR5862261
RNA	Sterile male pupa	CPF_AKOSRB_4_H2TK3DSXY.12B	NovaSeq	ERP128775	ERA4146275	ERS6348205	ERX5508100	ERR5862267
-	Genome Assembly + Annotation	-	-	ERP128837	-	ERS6376675	-	ERZ2140416

Supplementary Figure 10: Assembly workflow



Supplementary Figure 11: Annotation workflow