Science Advances
AAAS

advances.sciencemag.org/cgi/content/full/5/6/eaau3648/DC1

Supplementary Materials for

Unprecedented reorganization of holocentric chromosomes provides insights into the enigma of lepidopteran chromosome evolution

Jason Hill*, Pasi Rastas, Emily A. Hornett, Ramprasad Neethiraj, Nathan Clark, Nathan Morehouse, Maria de la Paz Celorio-Mancera, Jofre Carnicer Cols, Heinrich Dircksen, Camille Meslin, Naomi Keehnen, Peter Pruisscher, Kristin Sikkink, Maria Vives, Heiko Vogel, Christer Wiklund, Alyssa Woronik, Carol L. Boggs, Sören Nylin, Christopher W. Wheat*

*Corresponding author. Email: jason.hill@imbim.uu.se (J.H.); chris.wheat@zoologi.su.se (C.W.W.)

Published 12 June 2019, *Sci. Adv.* **5**, eaau3648 (2019)
DOI: 10.1126/sciadv.aau3648

The PDF file includes:

Fig. S9. Comparative assessment of genome assemblies and chromosomal evolution across, using a BLAST-like approach.
References (*69*–*85*)

Other Supplementary Material for this manuscript includes the following:

(available at advances.sciencemag.org/cgi/content/full/5/6/eaau3648/DC1)

Note S4 (.zip format). Functional annotation summary.
Table S2 (Microsoft Excel format). Genomic data from Lepbase used in the comparative assessment of genome assemblies.
Table S4 (.tsv format). GO term enrichment in syntenic blocks.

Supplementary materials

Note S1. Genome assembly, mtDNA assembly, HaploMerger results, and BUSCO analysis.

Nuclear genome assembly

The *P. napi* initial assembly used Illumina paired-end (PE) reads and mate-pair (MP) reads to generate scaffolds directly and was composed of 7,829 scaffolds with an N50-length of 300 kb and total length of 347 Mb. A second round of scaffolding was done using an added Illumina long mate-pair (LMP) library generated from a sibling of each individual used in the initial sequencing and assembly and SSPACE. This resulted in a *P. napi* assembly of 5860 scaffolds with an N50-length of 477 kb and total length of 350 Mb. The final scaffolding step used a third sibling of each species and a Chicago Illumina variable insert size library which brought the *P. napi* assembly to 3005 scaffolds with an N50-length of 4.2 Mb and a total length of 350 Mb.

Assembly and annotation of the mitochondrial genome.

Contigs containing mitochondrial genome sequence were identified by a BLASTN search using published cytochrome oxidase I (COI) sequences against the *P. napi* assembled genome. The identified contigs were then imported into Geneious (version 5.6.6.) and assembled to form the whole mitochondrial genome. Sequencing and assembly errors were manually investigated and corrected by mapping sequencing reads to the assembled mitochondrial genome using CLC Genomics Workbench v.4.

In order to annotate the mtDNA, the protein coding genes (PCGs) were first predicted using Genewise(*69*) and then the Open Reading Frames (ORFs) were manually checked through alignment to the mitochondrial genome (NCBI Genbank acc. HM156697) of the closely related butterfly, *Pieris rapae*. Available *P. napi* partial mitochondrial sequences were also aligned to aid annotation of both protein coding and rRNA genes (Genbank accessions: AF170861; AM236011;

GQ148917; DQ150035; DQ150071; LC090589). tRNA features were initially identified using tRNAscan-SE(*70*) and manually checked through alignment with the *P. rapae* mitochondrial genome.

The assembled mitochondrial genome of P. napi is 14945bp in length with a total AT content of 79.9%. These values fall well within the range of previously sequenced Lepidopteran mitogenomes. The mtDNA consists of the 37 genes typical of animals, 13 of which are protein coding, 22 are transfer RNAs (tRNAs), and the remaining two are ribosomal RNAs (rRNAs). In addition there is a non-coding AT-rich control region. The sequence of the control region typically does not assemble well and may therefore be incomplete. Few rearrangements have been observed among arthropod mitochondrial DNA, and usually these consist of translocations of tRNAs. Here, when compared to three butterflies (Pieris rapae HM156697; Pieris melete NC_010568; *Melitaea cinxia* HM243592) and one moth (Bombyx mori AY048187), perfect synteny in gene order and orientation within the mitogenome is observed. For P. napi the start codon of 12/13 PCGs is the typical ATN (ATT: NAD2, ATP8, NAD3, NAD5, NAD6; ATG: COII, ATP6, COIII, NAD4, NAD4L, CYTB or ATA: NAD1). However the putative start codon for COI is CGA, which appears to be common across insects. The stop codons of P. napi PCGs are either the common TAA (NAD2, ATP8, ATP6, COIII, NAD5, NAD4, NAD4L, NAD6, CYTB, NAD1), TAG (NAD3), or a single T as an incomplete stop codon, which has been found in several other Lepidopteran mitochondrial genes (e.g. *Melitaea cinxia* HM243592).

Haplomerger2 results.

Our results indicate that ~9.5 Mbp of sequence was identified as duplicated assembly content, and that it was contained in smaller scaffolds. To further verify this we aligned the repeat masked chromosome assembly to itself, and to the haplomerger2 generated

scaffolds. In alignment the chromosomal assembly had no alignments greater than 1000bp and visual inspection of the haplomerger2 alignment to the chromosome assembly revealed no merged scaffolds that aligned to more than one place in the chromosomal assembly. Thus, our chromosomal level assembly is haploid. However, there is likey to be some duplication in our non-placed scaffolds, which will be addressed in future assembly versions.

Haploid assessment of scaffold and chromosome assemblies.

The initial HiRise scaffold assembly statistics were as follows:

AssemblyQC Result

Contigs Generated : 2,969

Maximum Contig Length : 15,427,984

Minimum Contig Length : 109

Average Contig Length : 117,804.0 ± 1,128,208.3

Median Contig Length : 2,479.0

Total Contigs Length : 349,759,982

Total Number of Non-ATGC Characters : 78,612,647

Percentage of Non-ATGC Characters : 22.476

Contigs >= 100 bp : 2,969

Contigs >= 200 bp : 2,968

Contigs >= 500 bp : 2,968

Contigs >= 1 Kbp : 2,841

Contigs >= 10 Kbp : 581

Contigs >= 1 Mbp : 37

N50 value : 12,597,868

Haplomerger2 processed this assembly and the results were as follows:

AssemblyQC Result

Contigs Generated : 1,478

Maximum Contig Length : 15,377,620

Minimum Contig Length : 548

Average Contig Length : 230,216.3 ± 1,580,553.8

Median Contig Length : 5,902.5

Total Contigs Length : 340,259,761

Total Number of Non-ATGC Characters : 74,846,293

Percentage of Non-ATGC Characters : 21.997

Contigs >= 100 bp : 1,478

Contigs >= 200 bp : 1,478

Contigs >= 500 bp : 1,478

Contigs >= 1 Kbp : 1,418

Contigs >= 10 Kbp : 355

Contigs >= 1 Mbp : 38

N50 value : 12,541,812

The Haplomerger2 results are as follows:

summary of haplomerger2 run

#	original	haplo_only
Total Contigs Length :	349,759,982	340,259,761
Contigs Generated :	2,969	1,478
Maximum Contig Length :	15,427,984	15,426,992

BUSCO analysis

We assessed our genome assembly and annotation completeness using Benchmarking Universal Single-Copy Orthologs (BUSCO) version is: 3.0.2, run in mode genome with the lineage dataset: insecta_odb9 (Creation date: 2016-02-13, number of species: 42, number of BUSCOs: 1658). Below are shown the output for the full genome assembly, followed by the results for the AllPaths scaffolds only (i.e. before scaffolding). This demonstrates that the full assembly has a couple more complete SCOs, which were brought together by our scaffold joins during chromosome level assembly.

C:94.5%[S:92.3%,D:2.2%],F:2.7%,M:2.8%,n:1658

1567 Complete BUSCOs (C)

1531 Complete and single-copy BUSCOs (S)

36 Complete and duplicated BUSCOs (D)

44 Fragmented BUSCOs (F)

47 Missing BUSCOs (M)

1658 Total BUSCO groups searched

C:94.3%[S:92.3%,D:2.0%],F:2.8%,M:2.9%,n:1658

1563 Complete BUSCOs (C)

1530 Complete and single-copy BUSCOs (S)

33 Complete and duplicated BUSCOs (D)

46 Fragmented BUSCOs (F)

49 Missing BUSCOs (M)

1658 Total BUSCO groups searched

Note S2. HiRise scaffolding and misassembly correction.

The HiRise(*37*) scaffolding step provided a method to detect and break low confidence input scaffolds from SSPACE(*41*) and Allpaths-LG(*40*) using an independent set of reads. 90 scaffold mis-joins in 44 scaffolds were detected and broken before assembly continued.

Note S3. Linkage group correction of misassemblies.

113 scaffolds of the genome assembly contained multiple segregating markers which allowed for the manual identification of 60 misassemblies within scaffolds and merging of corrected scaffolds into chromosomes. K-mer estimation of genome size and heterozygosity rate with GenomeScope(*71*) was as follows:

k = 15

property	min	max
Heterozygosity	1.99681%	2.01325%
Genome Haploid Length	261,121,754 bp	261,578,934 bp
Genome Repeat Length	192,745,333 bp	193,082,797 bp
Genome Unique Length	68,376,421 bp	68,496,136 bp
Model Fit	81.9164%	95.0115%
Read Error Rate	0.0419424%	0.0419424%

Note S4. Functional annotation summary.

Annotation predicted 13,622 gene models (119,909 exons) derived from 20,325 mRNA transcripts respectively, with 9,346 genes annotated using a combination of KEGG(*72*), PFAM(*73*), InterPro(*74*), GO(*75*), MetaCyc(*76*), UniPathway(*77*), and Reactome(*78*) (Supplementary_annotation_tables.zip).

Information about Coding Genes:

Number of genes 13622

Number of mrnas 20325

Number of mrnas with utr both sides 9467

Number of mrnas with at least one utr 16893

Number of cdss 20325

Number of five_prime_utrs 13690

Number of three_prime_utrs 12670

Number of exons 119909

Number of introns 99584

Number of exon of cds 111587

Number of intron of cds 91262

Number of exon of five_prime_utr 17985

Number of intron of five_prime_utr 4295

Number of exon of three_prime_utr 16525

Number of intron of three_prime_utr 3855

mean mrnas per gene 1.5

mean cdss per mrna 1.0

mean five_prime_utrs per mrna 0.7

mean three_prime_utrs per mrna 0.6

mean exons per mrna 5.9

Total gene length 81632528

Total mrna length 122624718

Total cds length 20172833

Total five_prime_utr length 2661035

Total three_prime_utr length 7454905

Total exon length 30288280

mean gene length 5992

mean mrna length 6033

mean cds length 992

mean five_prime_utr length 194

mean three_prime_utr length 588

mean exon length 252

Longest genes 230888

Longest mrnas 230888

Longest cdss 38541

Longest five_prime_utrs 3377

Longest three_prime_utrs 8137

Longest exons 13013

Shortest genes 24

Shortest mrnas 24

Shortest cdss 6

Shortest five_prime_utrs 1

Shortest three_prime_utrs 1

Shortest exons 1

Functional inference for genes and transcripts was performed using the translated CDS features of each coding transcript. Each predicted protein sequences was blasted against the Uniprot/Swissprot(*67*) reference data set (downloaded 2014-05-15) in order to retrieve the gene name and the protein function as well as run against InterProscan version 5.7-48(*79*) in order to retrieve Interpro(*74*), PFAM(*73*), and GO(*75*) data. Outputs from both analyses have been parsed using the Annie annotation tool(*80*) to extract and reconcile relevant meta data into predictions for canonical protein names and functional predictions.

Database origin	Total Term Number	mRNA number referred by a term	Gene number referred by a term
PFAM	22931	17471	10105
Interpro	30268	17922	10337
GO	30590	12583	7243

Moreover, functional information was retrieved for 18413 transcripts (10624 genes), and 8060 transcripts (5283 genes) don’t have any functional information.

Note S5. RBH of orthologs.

The *B. mori* Geneset A from KAIKOBASE v3.2.2(*53*) was filtered for single copy orthologs (SCOs) as defined in orthoDB v9.1(*52*). 3101 SCOs were then compared to the annotated protein set of *P. napi* and 2743 genes were called as orthologs by reciprocal best hit (RBH) in blastp between *B. mori* and *P. napi*. To achieve a higher resolution map of the synteny relationship between species the full gene set of *B. mori* was compared to *P. napi* and 8176 genes were identified by RBH as orthologs.

Note S6. Whole-genome sequencing and RNA-based linkage map.

A second linkage map for *P. napi* was constructed from 16 full-sib offspring and the parents of a cross between a mother from Abisko, Sweden and father from Barcelona, Spain. None of these individuals were used in either the original genome assembly or the first linkage map based on RAD-tag sequencing. Illumina RNAseq reads for 12 individual offspring and Illumina whole genome sequencing reads for 4 individual offspring and 2 parents were used to construct a WGS linkage map of 106,362 SNP markers. Maternally inherited markers confirmed linkage of scaffolds within chromosomes due to achiasmatic recombination.

See Supplemental Figure S1 for these results per chromosome.

Code used to prepare data:

samtools mpileup -r Chromosome_"$i" -A -gu -Q 15 -t DP -f Pieris_napi_repmask_fullAsm.fasta -b list_of_bamfiles.txt | bcftools call -cv --skip-variants indels - | vcftools --vcf - --maf .1 --max-missing .85 --min-meanDP 6 --minDP 3 --recode --out consensus_Chr"$i"

Note S7. MP spanning.

35,492,386 read pairs from the 3 kb insert MP library, 39,072,074 read pairs from the 7 kb library, and 2,648,844 read pairs from the 40 kb library mapped to the chromosomes with a mapq > 20, in the proper orientation, and with the expected insert size. Every base pair position of the assembly was assigned a count of 0 for each mate pair library. Using a custom awk script, each read pair was taken in turn and the counter of every base pair of the assembly between the reads was incremented by one for that library. After all reads were processed each base pair position had a count of the number of reads that spanned that position for each library. The number of total mate pair spans used to diagnose potential misassemblies was the sum of the 3kb, 7kb, and 40kb library counts.

Mate pair spanning of a position is expected to fall towards zero if there is a misassembly or the region around that position contains low complexity or repeat rich sequence that decreases the mapq below the threshold of mapq <= 20.

See Supplemental Figure S1 for these results per chromosome.

Note S8. Alignment between *P. napi* and *P. rapae.*

The 75 largest scaffolds of *P. napi* and 101 largest scaffolds of *P. rapae* containing 90% of the content of each genome were aligned with each other using last aligner v. 714.

```
lastal -m10 -r5 -q97 -a0 -b97 -C2 -l100 PrapaeN90 " \
< Pieris_napi_chromosomes.fa > PnapiChr_PrapaeN90.maf
```

And visualized in dotplot format

```
/data/programs/last-714/scripts/last-dotplot -s 6 -x 5000 -y 5000 Pnapi_chm_N90.maf
Pnapi_chm_N90.png
```

Visual inspection of alignments between scaffolds showed 47 places where two *P. napi* scaffolds which were collinear on the linkage map were also collinearly aligned to a single *P. rapae* scaffold.

Note S9. Syntenic block support of scaffold joins.

Blocks of synteny were defined from 7109 of the genes identified as reciprocal best blast hits between *B. mori* geneset A and *P. napi* final annotation that occurred on the chromosomes of the *P. napi* assembly. Adjacent blocks were merged by recursively removing genes that shared no neighbors from the same *B. mori* chromosome, then blocks of size 2, 3, 4, and 5. This reduced the number of genes in syntenic blocks to 6,839, and the number of blocks from 499 to 99. While it is possible that a single gene or group of 5 genes translocated from a different chromosome or existed as a small fragment prior to the ancestral fusion events, it seemed more likely that misidentification in the reciprocal best blast hit of these non-single copy orthologs was responsible for the attribution of these small blocks of synteny.

See Supplemental Figure S1 for these results per chromosome.

Note S10. SNP generation from nextRAD library sequence.

Used trimmomatic(*81*) to trim the adapters from the reads

```
system("java -jar Trimmomatic-0.32/trimmomatic-0.32.jar SE $file $outfile
ILLUMINACLIP:Trimmomatic-0.32/adapters/NexteraPE-PE.fa:1:30:5 MINLEN:75
HEADCROP:$start_trim CROP:$trim");
```

Read in the sequence files with a Perl script and track the number of occurrences of each read.

Keep all the reads with counts above a threshold. This is the first set of loci.

Call the reads with a count above a threshold “repetitive” and create a fasta file of the repeats.

Align the first set of loci to the repeat set allowing many mismatches (we use bwa-mem or BBMap). Remove all loci that align to the repeats. The remaining loci are the second set of loci.

```
system("bwa mem -t 6 -B 2 -O 3 -k 13 -a $repeat_fasta_file $fastq_file >
$repeat_align_file_mt");
```

Align the second set of loci to itself. Reads that map to each other are likely alleles at the same locus. Pick one and remove the others. The winnowed set of loci are the reference set.

```
system("java -ea -Xmx31g -cp align2.BBMapPacBioSkimmer in=" . $fastq_file . " out=" .
$align_file . " minid=" . $align_threshold . " maxsites=150 ambig=all secondary=true ow sssr=" .
$sssr . " slow idtag pambig=f maxindel=15 nodisk k=9 minhits=1 expectedsites=40");
```

The above is the “rad” specific set of steps. We now try to use a general pipeline for SNP calling.

Align all the reads to the reference set (bwa-mem or BBMap)

```
$t = "java -ea -Xmx31g -cp align2.BBMap in=" . $gf . " ref=" . $ref_fasta_file . " nodisk
out=stdout minid=" . $align_threshold . " ow slow k=9 idtag maxindel=40 | samtools view -bSu - |
samtools sort -@ 8 - " . $sort_file;
```

Sort the resulting bam files

```
samtools view -bSu - | samtools sort -@ 8 - " . $sort_file;
```

Use samtools and bcftools to create a vcf genotype table

```
samtools mpileup -gu -Q 15 -t DP -f ref_2014_txt.fasta -b ref_txt.align_samples | bcftools call -cv -
 > 2014.vcf
```

Then filter the vcf on minimum read depth and allele frequency

```
# vcftools --vcf 2014.vcf --maf .1 --max-missing .85 --min-meanDP 6 --minDP 3 --recode
```

Note S11. Comparison of *P. napi* versus *B. mori* collinear blocks within Lepidoptera genomes.

Each scaffold of the 24 Lepidoptea genomes was searched for SCOs from the Lepidoptera specific OrthoDB v9. Each SCO was identified with a *P. napi* and *B. mori* chromosome. If a scaffold contained SCOs from a single *P. napi* chromosome and 2 or more *B. mori* chromosome, that scaffold was called napi-like. Conversely if a scaffold contained SCOs from a single *B. mori* chromosome and 2 or more *P. napi* chromosomes that scaffold was called mori-like. The limitations of this approach are two-fold, misassemblies can and in the case of the HI-rise *P. napi* assembly have joined within a scaffold collinear blocks that should not be. In the case of *P. napi* these misassemblies were corrected with a linkage map that removed all mori-like structures within the chromosomes.

Note S12. Estimation of genomic rearrangements.

To estimate the number of chromosomal rearrangements between B. mori and P. napi, we used two different approaches available in the grimm software. First, we directly estimated the optimal scenario of rearrangement events without sign orientation of genes (“-u -s”), which gave our upper values of translocations and inversions (57 and 6, respectively). We then predicted the optimal orientation of the individual genes in our dataset using an approximation algorithm and different numbers of iterations (“-U 10000”, “-U 100000”, and “-U 200000”), and then ran this assuming orientation was known. All of these analyses gave the same results, which were the lower end of the reported estimates, with 47 translocations and 3 inversions estimated.

Note S13. Lepidopteran synteny cutoff selection.

These cutoffs were used for the reported results because we believed them to be the best trade-off between type I and type II error in identifying truly syntenic genes between Lepidopteran species. We found that due to BLAST's algoithm, HSPs were generally exons and if less than three were used to assign identity genes that highly conserved domains caused more than 20% of the single copy orthologs to have more than one potential ortholog in other species which is extremely unlikely to be true. If more than 3 or more HSPs gave similar results but more than 3 excluded orthologs with more than 3 exons. Similarly using 5 genes to determine a scaffold was found to be a conservative enough cutoff to compensate for errors of gene identification that the BLAST driven algorithm generated without being so conservative that smaller scaffolds and shorter collinear blocks were missed. In the more fragmented assemblies 5 genes was indeed too many to assign identity to some scaffolds. Changing these cutoffs to other similar values did not change the general pattern of results but did alter the ratios slightly.

Table S1. Read data used in assembly and analysis.

Library	Mean insert size	Source	# raw read pairs	# clean read pairs	Coverage
Illumina 180 bp PE	-20bp overlap	Female pupae #1	350 M	348 M	180x
Illumina 3 kb MP	3600bp	Female pupae #1	242 M	61 M	34x
Illumina 7 kb MP	6500bp	Female pupae #1	264 M	70 M	40x
Illumina 40kb MP	40kb	Unsexed pupae #2	8.9 M	8.9 M	12x
Chicago	variable	Unsexed pupae #3	13 M	13 M	19x
Total				500 M	285 X

Table S2. Genomic data from Lepbase used in the comparative assessment of genome assemblies.

Lepbase_genomes.xlsx

Table S3. Lepidopteran phylogeny calibrations.

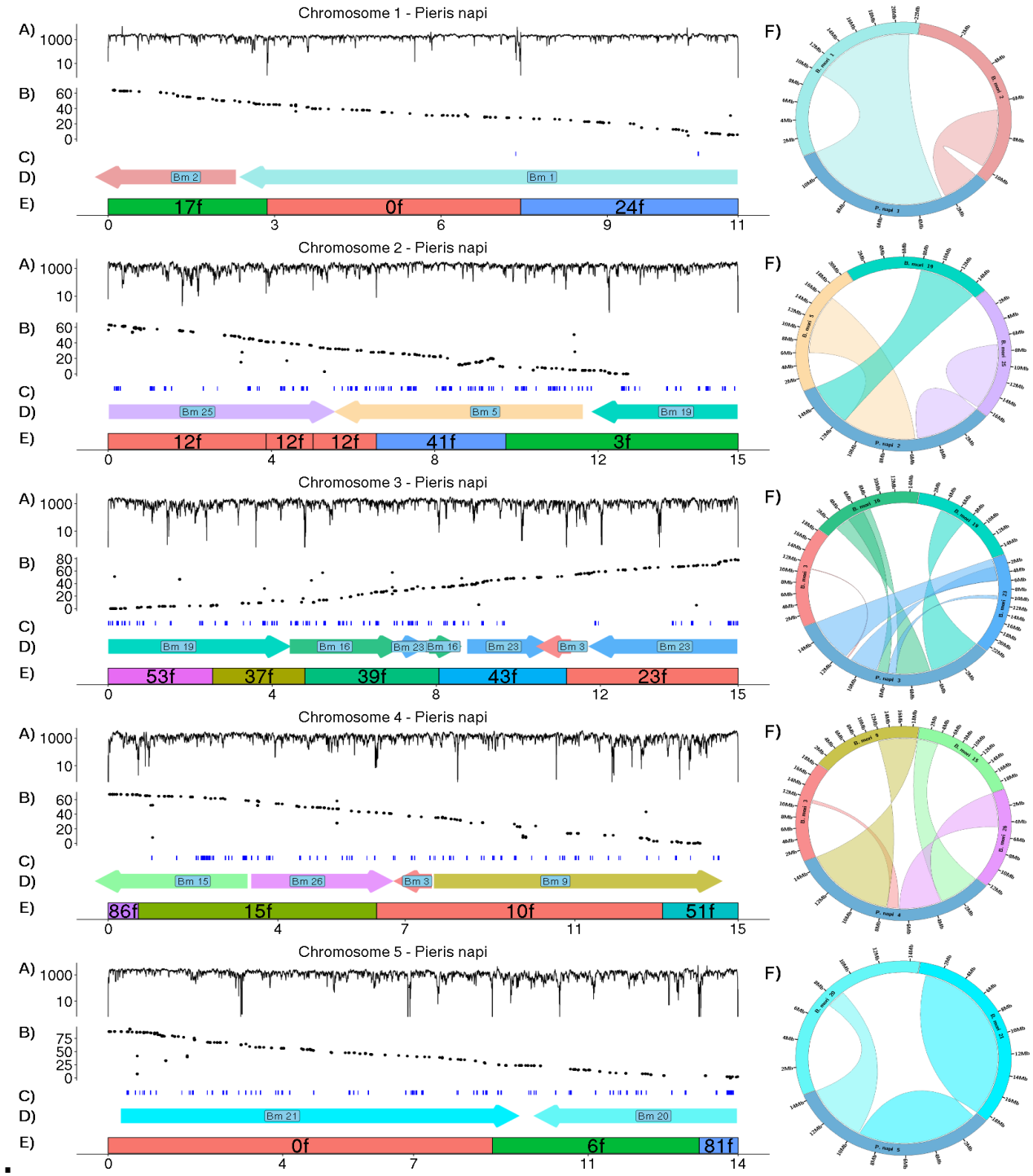
Node ages in millions of years ago (MA). Node numbers correspond to Supplemental Fig. S8. Reference for each node age in that figure is provided below.

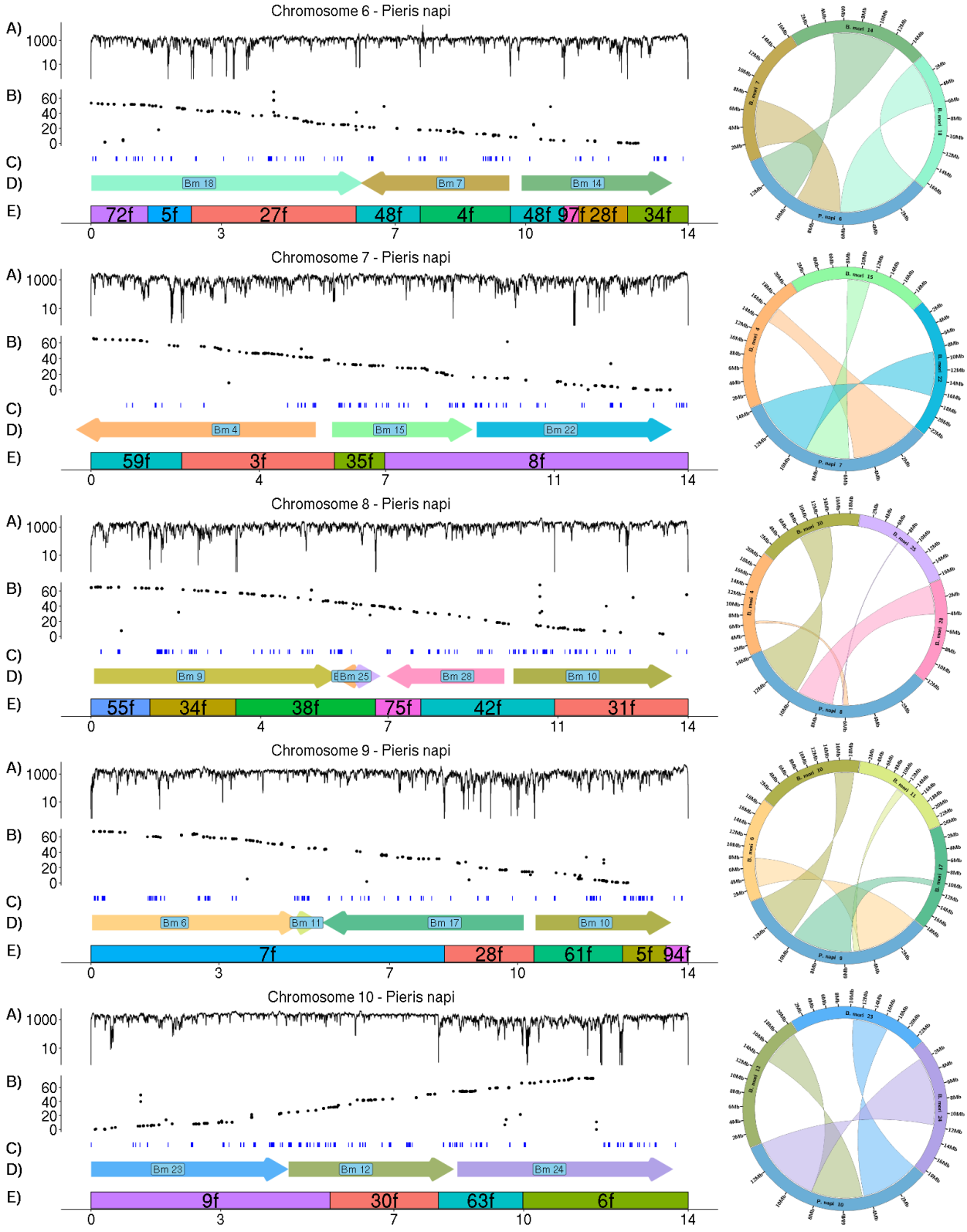
Node	Age (MA)	Reference
25	140.47	(*83*)
26	116.74	(*83*)
27	109.78	(*83*)
28	94.05	(*83*)
29	83.85	(*83*)
30	90.70	(*83*)
31	28.24	(*83*, *85*)
32	92.74	(*83*)
33	110.87	(*65*)
34	30.92	(*82*)
35	24.79	(*82*)
36		
37	109.46	(*65*)
38	104.73	(*65*)
39	86.18	(*26*)
40	75.39	(*26*)
41	15.79	(*26*)
42	101.41	(*65*)
43	89.79	(*83*)
44	85.35	(*83*)
45	78.55	(*83*)
46	10.75	(*84*)

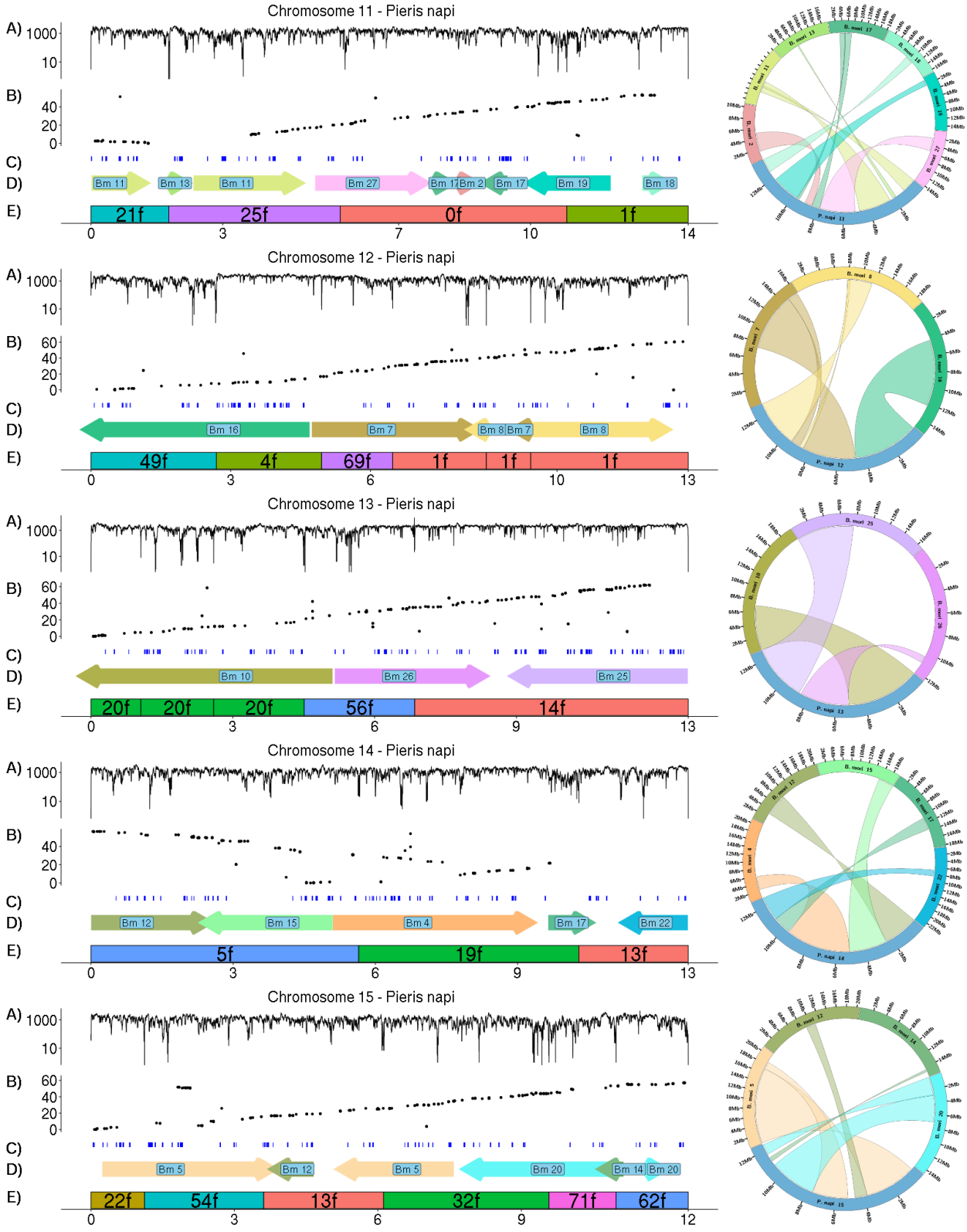
Table S4. GO term enrichment in syntenic blocks.

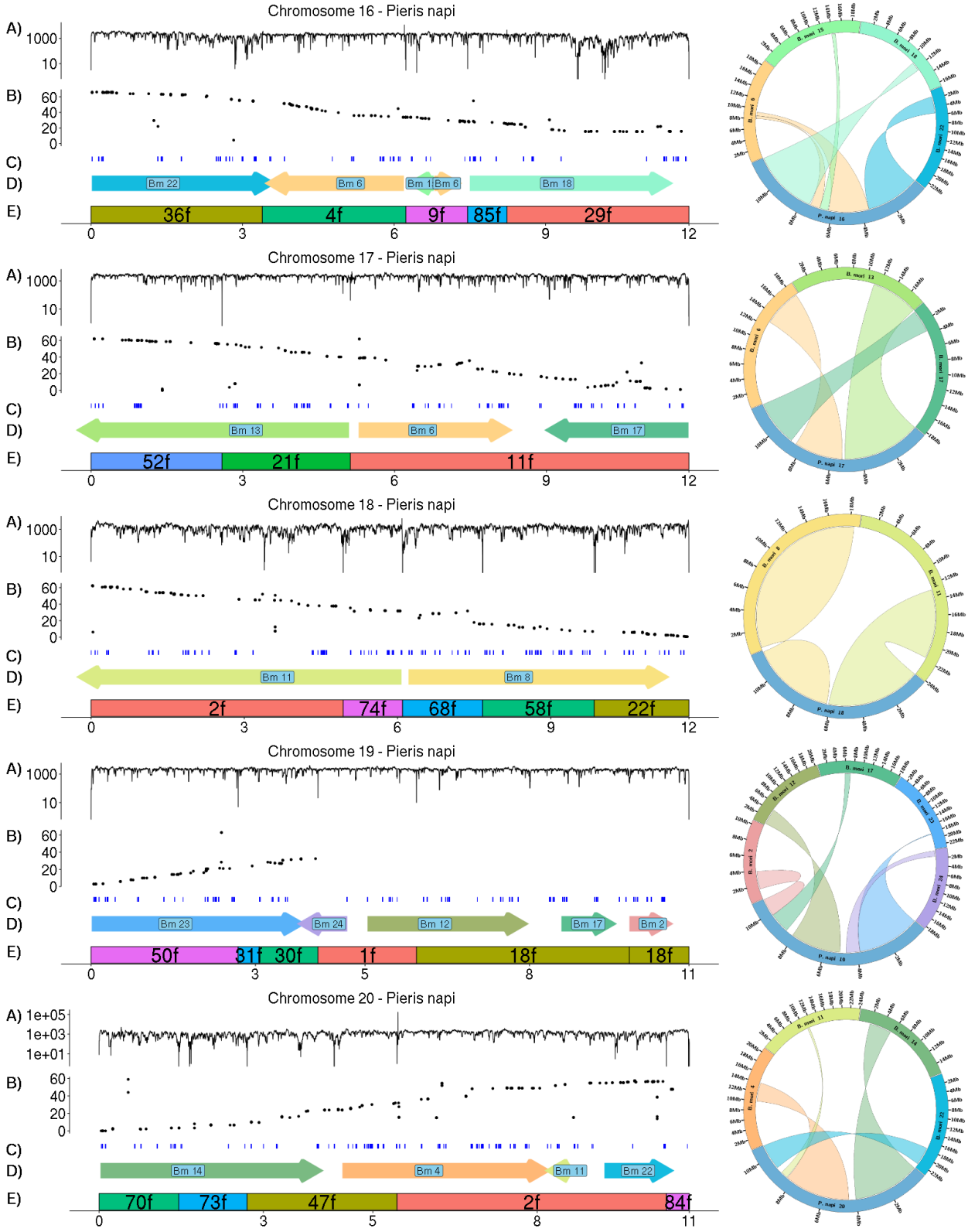
A GO term was considered enriched within a syntenic block if at least 3 genes were assigned that GO term and it was overrepresented compared to the rest of the genome by a fisher exact test with a p-value < 0.01. The GO.ID and Term identify the GO category analyzed, the number of total genes Annotated within that category, and the number of Significant genes within a particular syntenic block are tabulated along with the Fisher exact test significance (classicFisher) of that category's enrichment. The syntenic block id (sbid) and that blocks genomic coordinates indicate the region containing that significant GO term.

Table of results: GOenrichment.tsv


Chromosome 1 - Pieris napi
A)
1000
10
B)
60
40
20
0
C)
D)
Bm 2
Bm 1
E)
17f
0f
24f
0
3
6
9
11
F)
Chromosome 2 - Pieris napi
A)
1000
10
B)
60
40
20
0
C)
D)
Bm 25
Bm 5
Bm 19
E)
12f
12f
12f
41f
3f
0
4
8
12
15
F)
Chromosome 3 - Pieris napi
A)
1000
10
B)
80
60
40
20
0
C)
D)
Bm 19
Bm 16
Bm 23
Bm 16
Bm 23
Bm 3
Bm 23
E)
53f
37f
39f
43f
23f
0
4
8
12
15
F)
Chromosome 4 - Pieris napi
A)
1000
10
B)
60
40
20
0
C)
D)
Bm 15
Bm 26
Bm 3
Bm 9
E)
86f
15f
10f
51f
0
4
7
11
15
F)
Chromosome 5 - Pieris napi
A)
1000
10
B)
75
50
25
0
C)
D)
Bm 21
Bm 20
E)
0f
6f
81f
0
4
7
11
14
F)


Chromosome 6 - Pieris napi
A)
1000
10
B)
60
40
20
0
C)
D)
Bm 18
Bm 7
Bm 14
E)
72f
5f
27f
48f
4f
48f
97f
28f
34f
0
3
7
10
14
Chromosome 7 - Pieris napi
A)
1000
10
B)
60
40
20
0
C)
D)
Bm 4
Bm 15
Bm 22
E)
59f
3f
35f
8f
0
4
7
11
14
Chromosome 8 - Pieris napi
A)
1000
10
B)
60
40
20
0
C)
D)
Bm 9
Bm 25
Bm 28
Bm 10
E)
55f
34f
38f
75f
42f
31f
0
4
7
11
14
Chromosome 9 - Pieris napi
A)
1000
10
B)
60
40
20
0
C)
D)
Bm 6
Bm 11
Bm 17
Bm 10
E)
7f
28f
61f
5f
94f
0
3
7
10
14
Chromosome 10 - Pieris napi
A)
1000
10
B)
60
40
20
0
C)
D)
Bm 23
Bm 12
Bm 24
E)
9f
30f
63f
6f
0
3
7
10
14


Chromosome 11 - Pieris napi
A)
1000
10
B)
40
20
0
C)
D)
Bm 11
Bm 13
Bm 11
Bm 27
Bm 17
Bm 2
Bm 17
Bm 19
Bm 18
E)
21f
25f
0f
1f
0
3
7
10
14
Chromosome 12 - Pieris napi
A)
1000
10
B)
60
40
20
0
C)
D)
Bm 16
Bm 7
Bm 8
Bm 7
Bm 8
E)
49f
4f
69f
1f
1f
1f
0
3
6
10
13
Chromosome 13 - Pieris napi
A)
1000
10
B)
60
40
20
0
C)
D)
Bm 10
Bm 26
Bm 25
E)
20f
20f
20f
56f
14f
0
3
6
9
13
Chromosome 14 - Pieris napi
A)
1000
10
B)
40
20
0
C)
D)
Bm 12
Bm 15
Bm 4
Bm 17
Bm 22
E)
5f
19f
13f
0
3
6
9
13
Chromosome 15 - Pieris napi
A)
1000
10
B)
60
40
20
0
C)
D)
Bm 5
Bm 12
Bm 5
Bm 20
Bm 14
Bm 20
E)
22f
54f
13f
32f
71f
62f
0
3
6
9
12


Chromosome 16 - Pieris napi
A)
1000
10
B)
60
40
20
0
C)
D)
Bm 22
Bm 6
Bm 1
Bm 6
Bm 18
E)
36f
4f
9f
85f
29f
0
3
6
9
12
Chromosome 17 - Pieris napi
A)
1000
10
B)
60
40
20
0
C)
D)
Bm 13
Bm 6
Bm 17
E)
52f
21f
11f
0
3
6
9
12
Chromosome 18 - Pieris napi
A)
1000
10
B)
60
40
20
0
C)
D)
Bm 11
Bm 8
E)
2f
74f
68f
58f
22f
0
3
6
9
12
Chromosome 19 - Pieris napi
A)
1000
10
B)
60
40
20
0
C)
D)
Bm 23
Bm 24
Bm 12
Bm 17
Bm 2
E)
50f
31f
30f
1f
18f
18f
0
3
5
8
11
Chromosome 20 - Pieris napi
A)
1e+05
1e+03
1e+01
B)
60
40
20
0
C)
D)
Bm 14
Bm 4
Bm 11
Bm 22
E)
70f
73f
47f
2f
84f
0
3
5
8
11

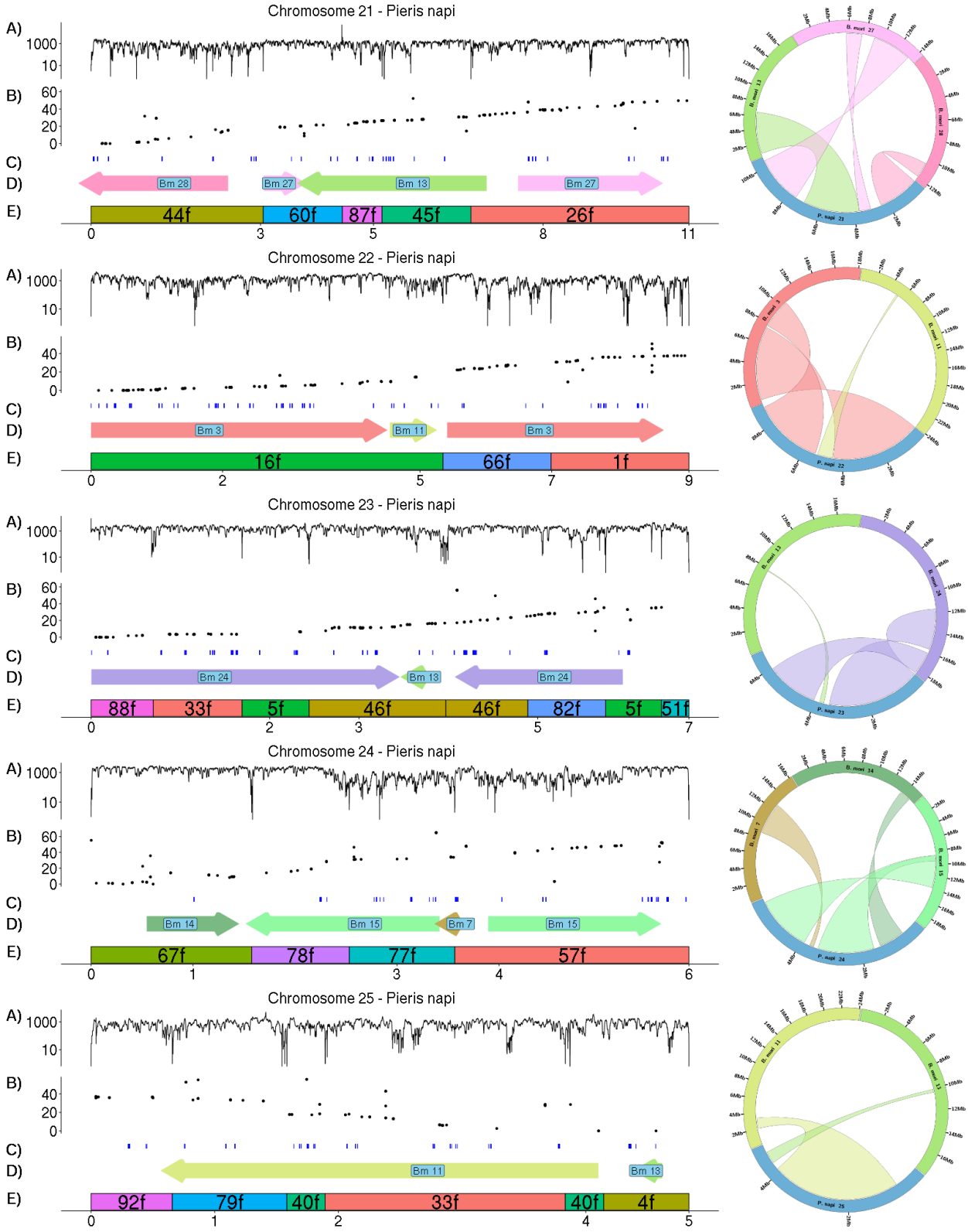


Fig. S1. Chromosomal assembly validation figures. As in Fig. 2 of main text. Validation of the largest four *P. napi* chromosomes. Within each, **a)** mate pair spanning depth is shown across each chromosome, summed from the 3kb, 7kb, and 40kb libraries (genome averaged = 1356). Of the

scaffold join positions 74 of 97 were spanned by > 50 properly paired reads (mean = 117.8, S.D. = 298.7), while the remaining 23 scaffold joins had 0 mate pair spans. **b)** black dots represent RAD-seq linkage markers and their recombination distance along chromosomes from the first linkage map **c)** Results from the second linkage map of maternally inherited markers (RNA-seq and whole genome data), where all markers within a chromosome are completely linked due to suppressed recombination in females (i.e. recombination distance is not shown on Y axis). **d)** *B. mori* collinear blocks, colored and labeled by their chromosomal origin, along with orientation by arrow, as in Fig. 1a. **e)** *P. napi* scaffolds comprising each chromosome, labeled to indicate scaffold number and orientation. **f)** To the right of each *P. napi* chromosome is a circos plot showing the location and orientation of the collinear blocks from each *B. mori* donor chromosome that comprise a given *P. napi* chromosome, colored as in Fig. 1a. A twist in the ribbon indicates a reversal of the 5' to 3' orientation of the *B. mori* relative to the *P. napi* chromsomes. Ribbon width on the *P. napi* chromosome is relative to the size of the collinear block. Remaining chromosomes shown in Supplementary Fig. 2.

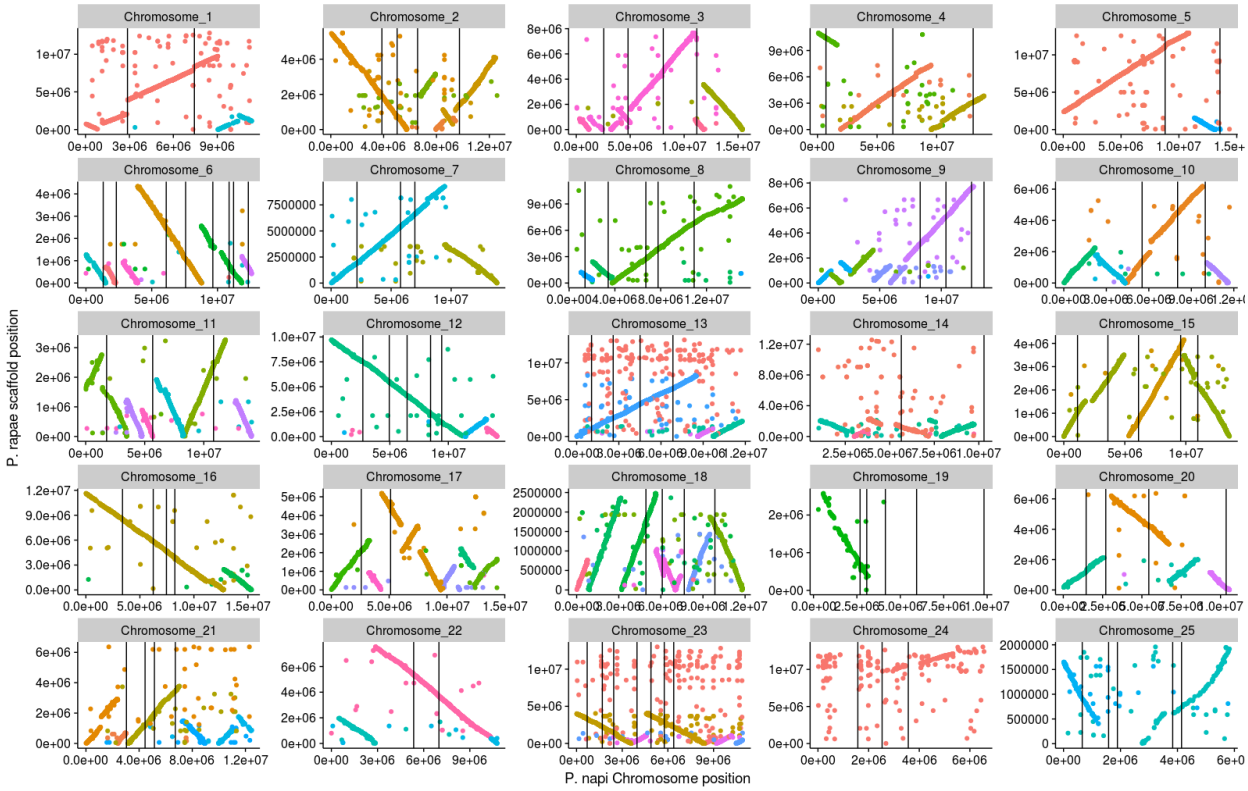


Fig. S2. Alignment of *P. rapae* scaffolds to *P. napi* chromosomes. Scaffold joins in the chromosomal assembly of *P. napi* were validated by alignment with *P. rapae* scaffolds. LAST aligner v. 714 with default settings aligned the assemblies and found agreement between exons of the two species (mean size 151 bp). Verticle lines represent boundaries between *P. napi* scaffolds, showing where they were joined by the linkage map for chromosomal level assembly (e.g. compare with Supplemental Figure S1). While complete synteny between these two species is very unlikely we assumed that if a HiRise scaffold of *P. rapae* spans two scaffolds of *P. napi* that were joined by the linkage map that the relationship between the *P. napi* scaffolds has greater support. Noise was reduced by filtering out alignments if they had a score < 300 or if they belonged to a *P. rapae* scaffold that had less than 150 other alignments to a giver *P. napi* chromosome.

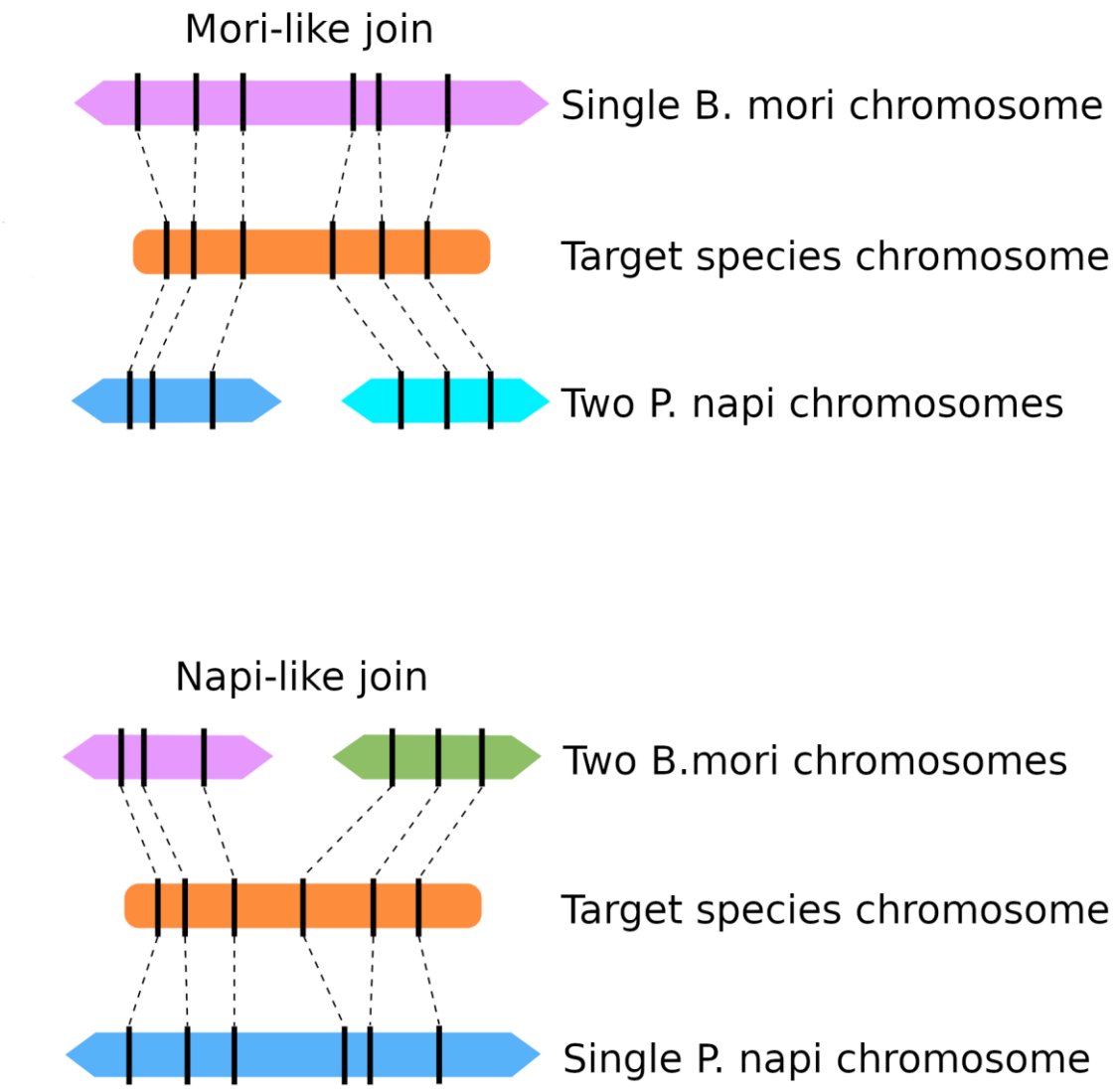


Fig. S3. Example of criteria used to determine napi-like and mori-like joins. Two alternative examples of how 6 orthologous genes (represented by black rectangles) shared between *P. napi*, *B. mori*, and a target species could allow for the inference of the a napi-like or mori-like chromosome level synteny. In the top example 6 genes called by blastx reside on a single target species scaffold. If their orthologs on *B. mori* reside on a single chromosome and their orthologs on *P. napi* reside on two chromosomes the scaffold "joins" two *P. napi* chromosomes in the same manner as *B. mori* and is counted as

supporting a mori-like chromosome. Conversely in the bottom example the 6 *P. napi* orthologs reside on a single chromosome and the *B. mori* orthologs are split, indicating the scaffold supports a napi-like scaffold.

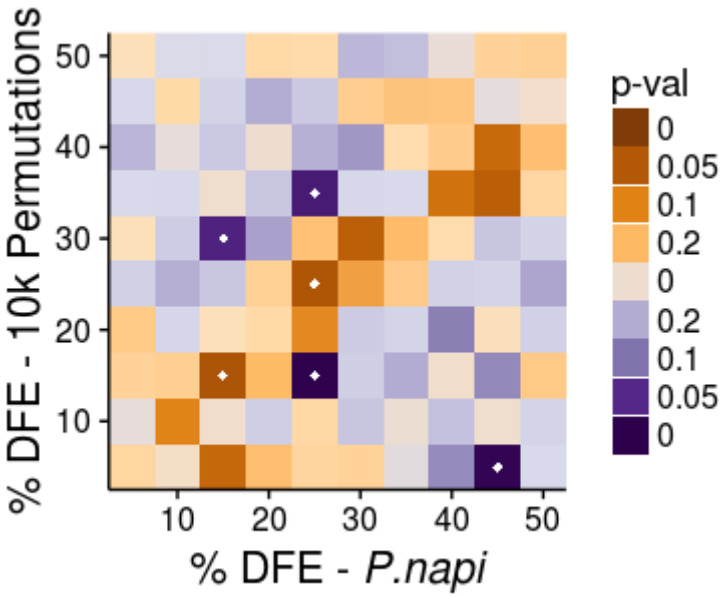
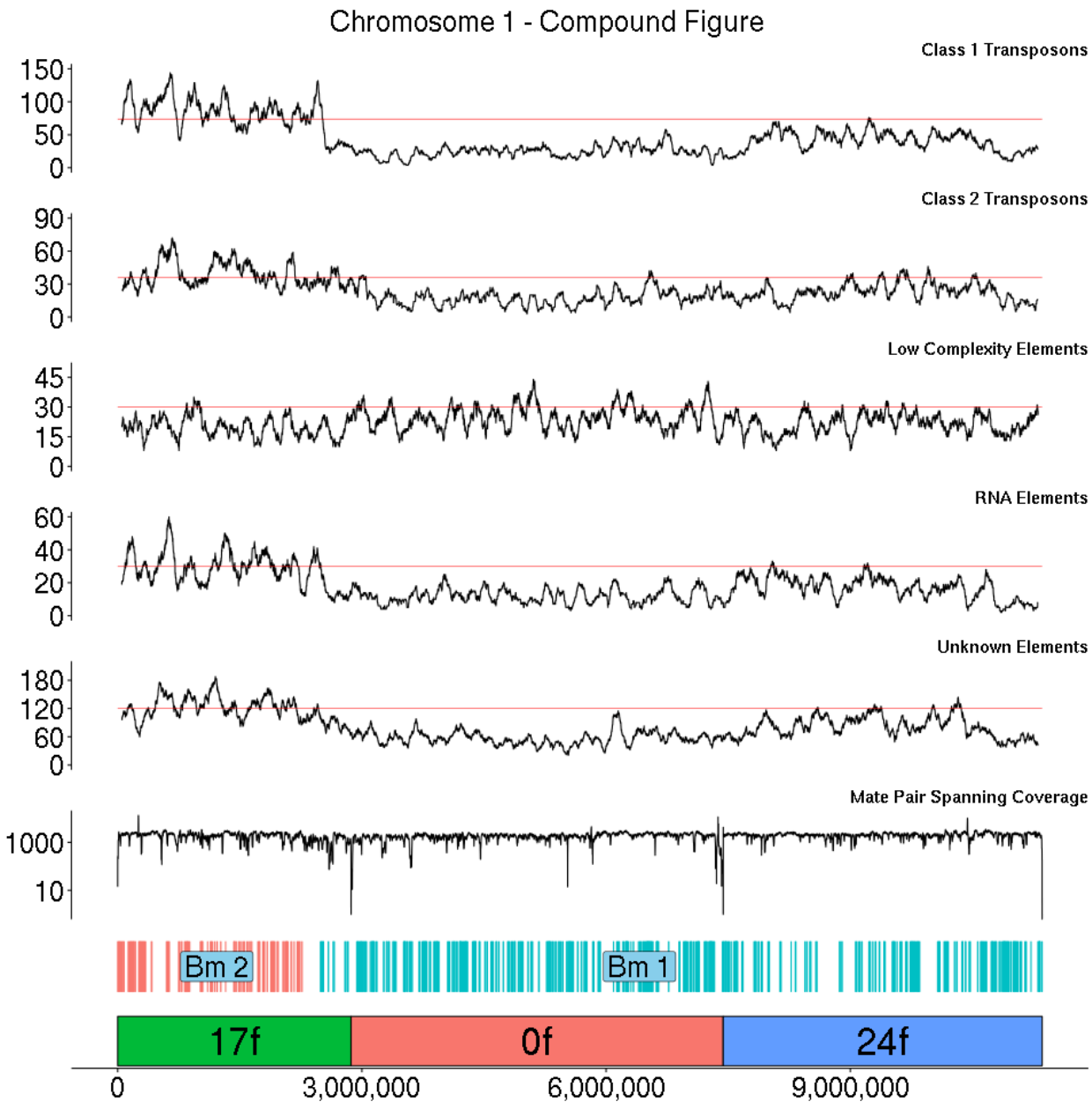
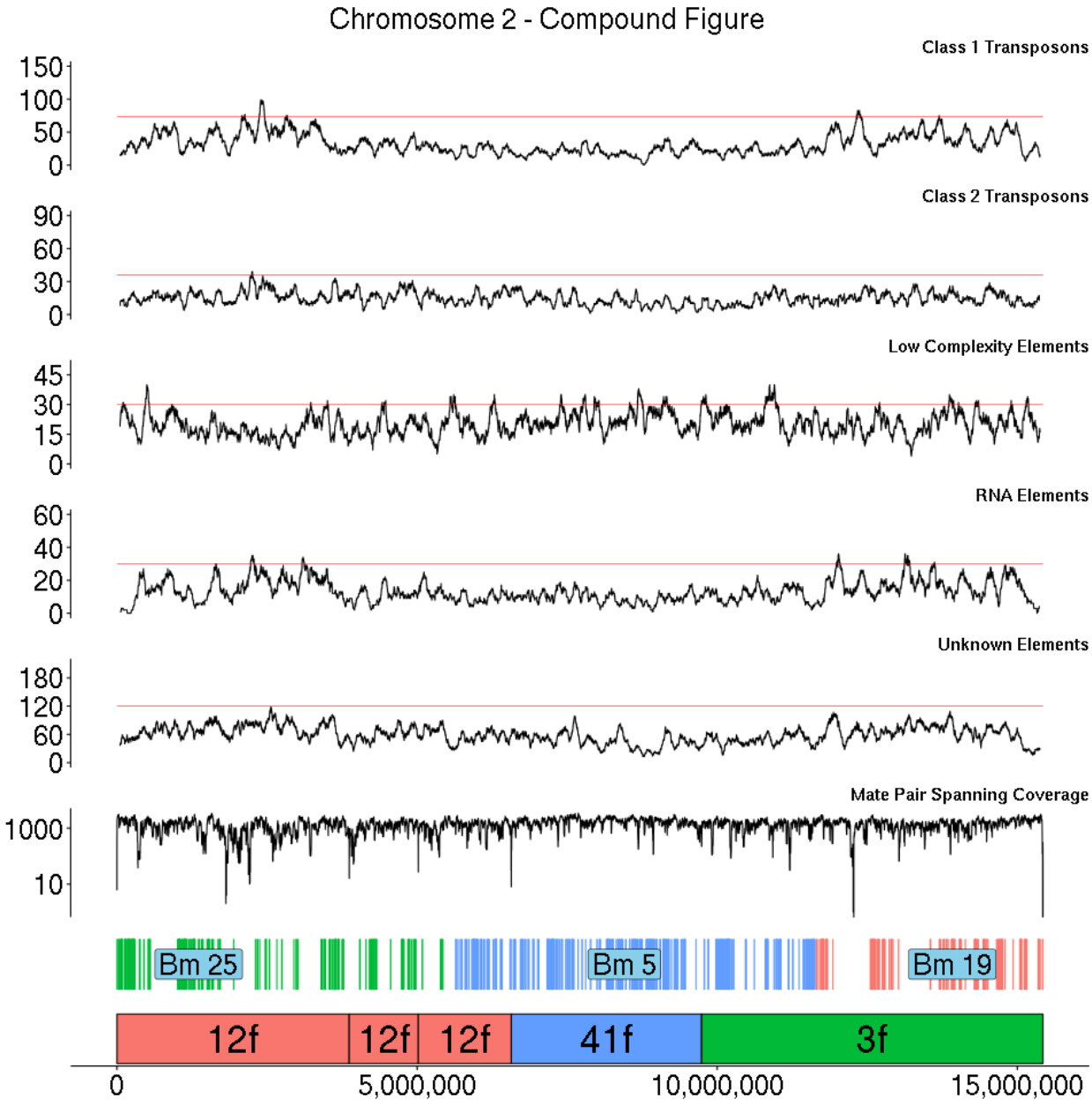
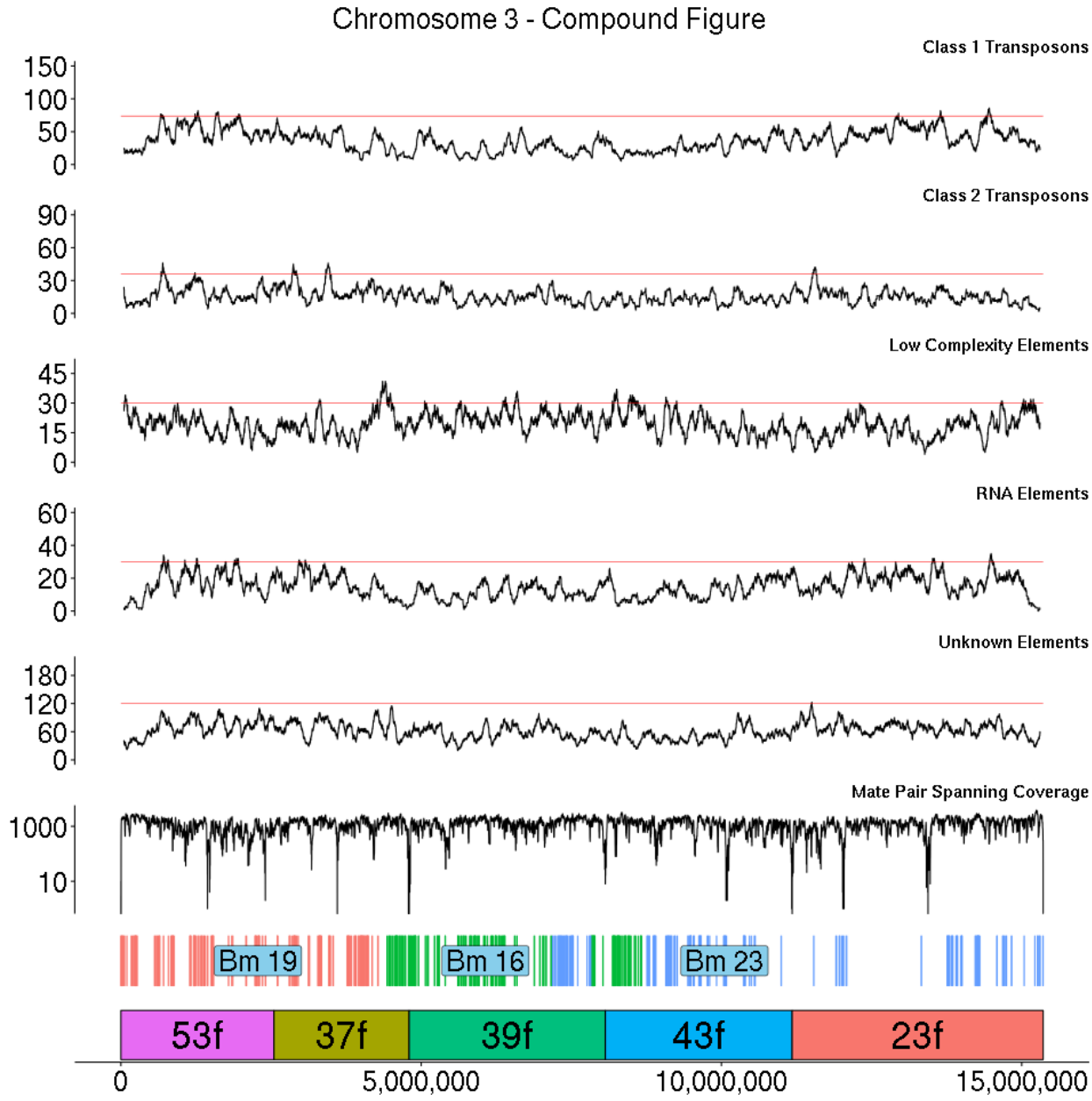
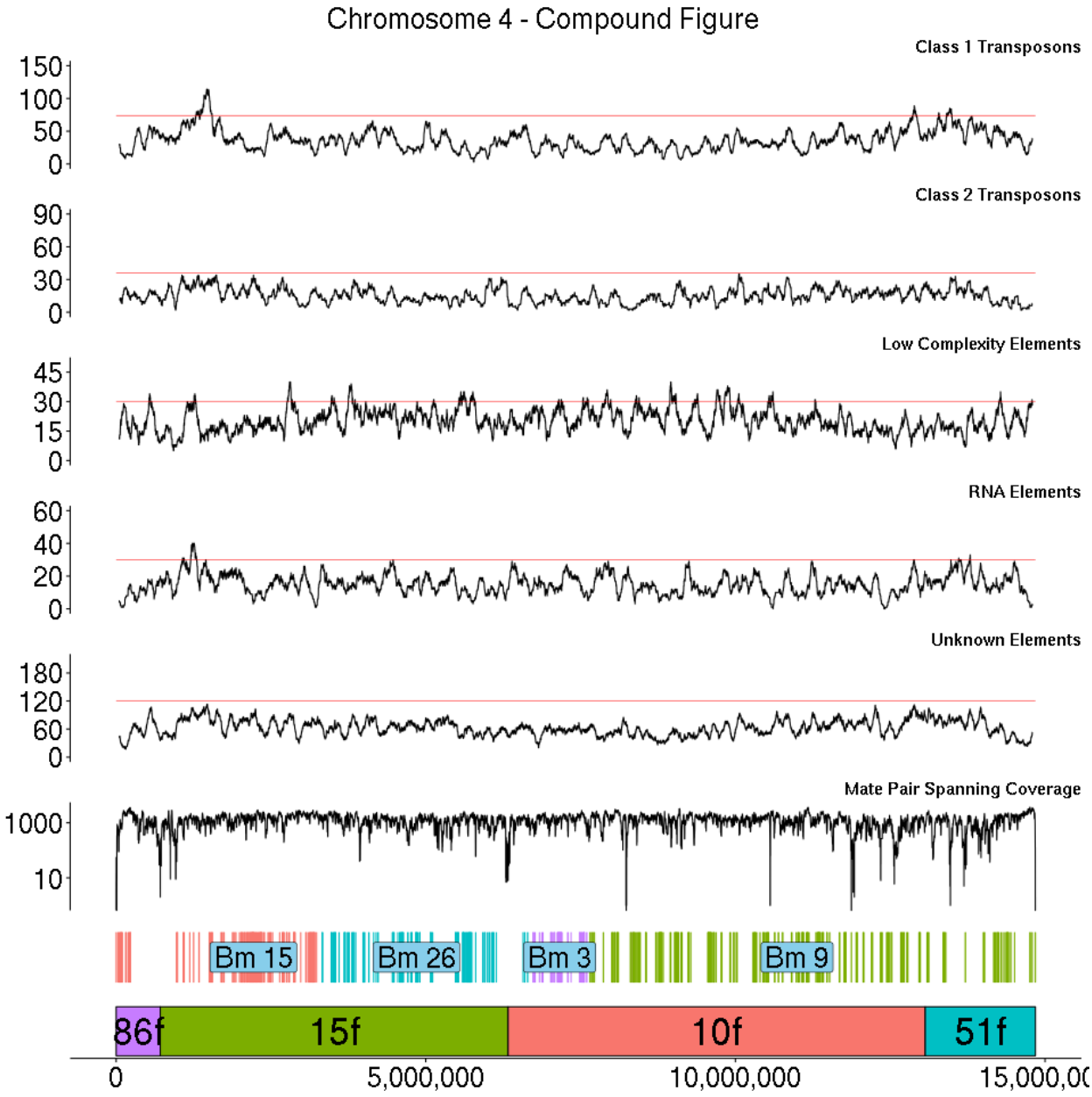


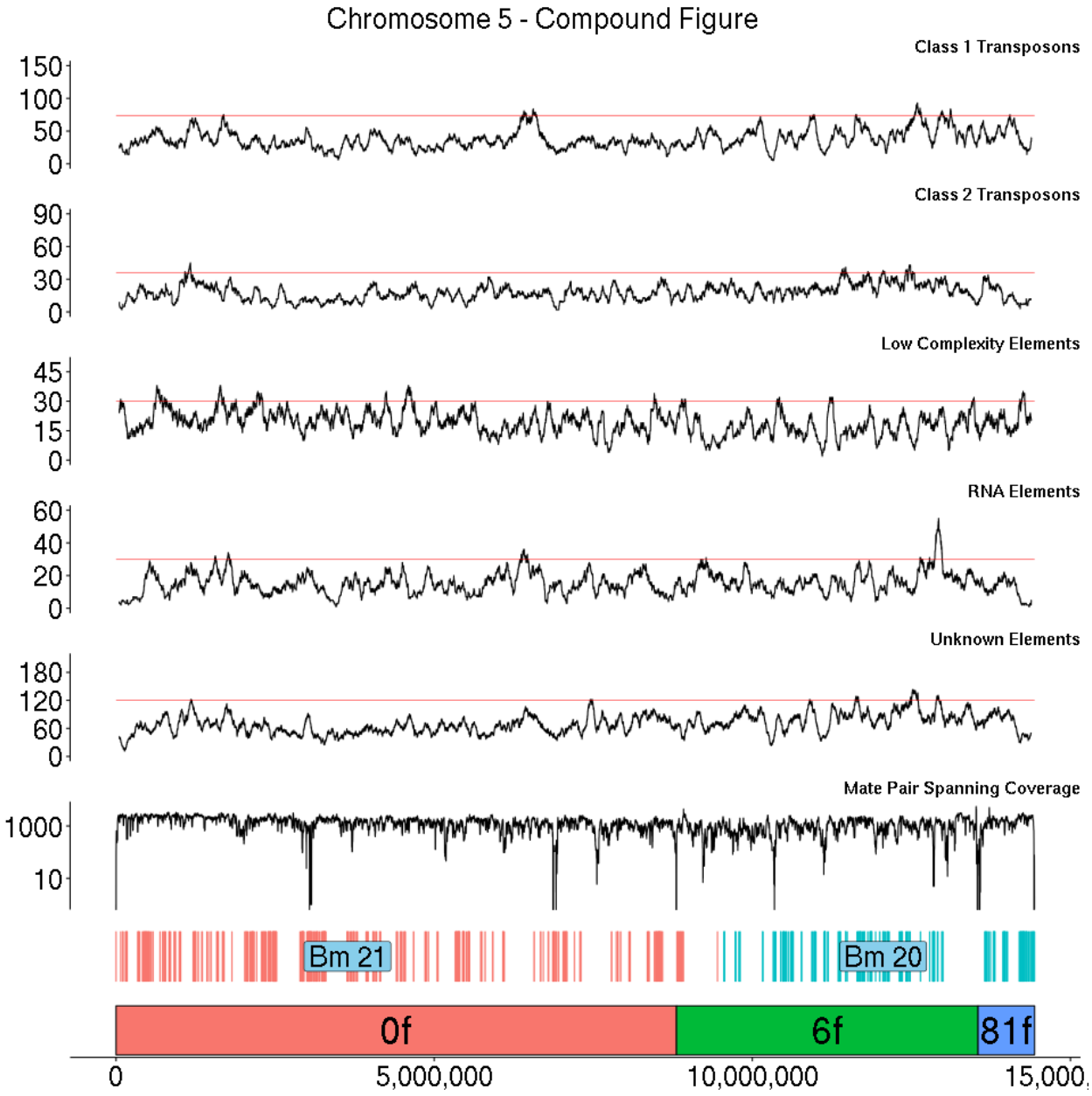
Fig. S4. Permutation analysis of chromosomal terminal ends. Comparison between the observed DFE distribution and the expected distribution generated from 10,000 genomes of 25 chromosomes constructed from the random fusion of the observed collinear blocks. Bins in which more genes occur in the observed genomes than the expected distribution are in orange, less genes in blue, $P < 0.05$ in either direction are denoted by a white dot. SCO spatial distribution was significantly higher than expected along the diagonal (two bins with $p < 0.05$), while significantly lower than expected off the diagonal (four bins with $p < 0.05$).

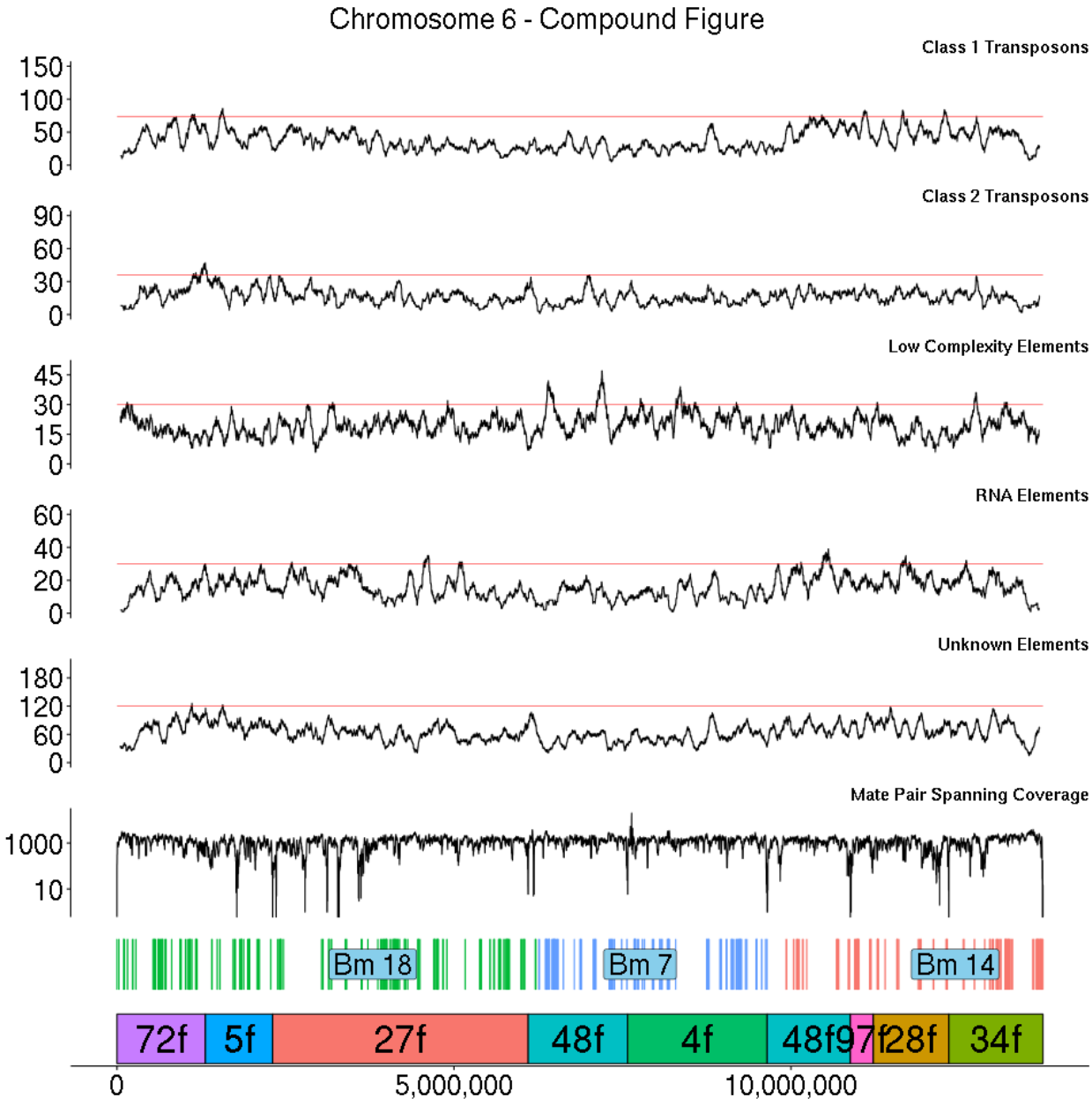

Chromosome 1 - Compound Figure
Class 1 Transposons
150
100
50
0
Class 2 Transposons
90
60
30
0
Low Complexity Elements
45
30
15
0
RNA Elements
60
40
20
0
Unknown Elements
180
120
60
0
Mate Pair Spanning Coverage
1000
10
Bm 2
Bm 1
17f
0f
24f
0
3,000,000
6,000,000
9,000,000

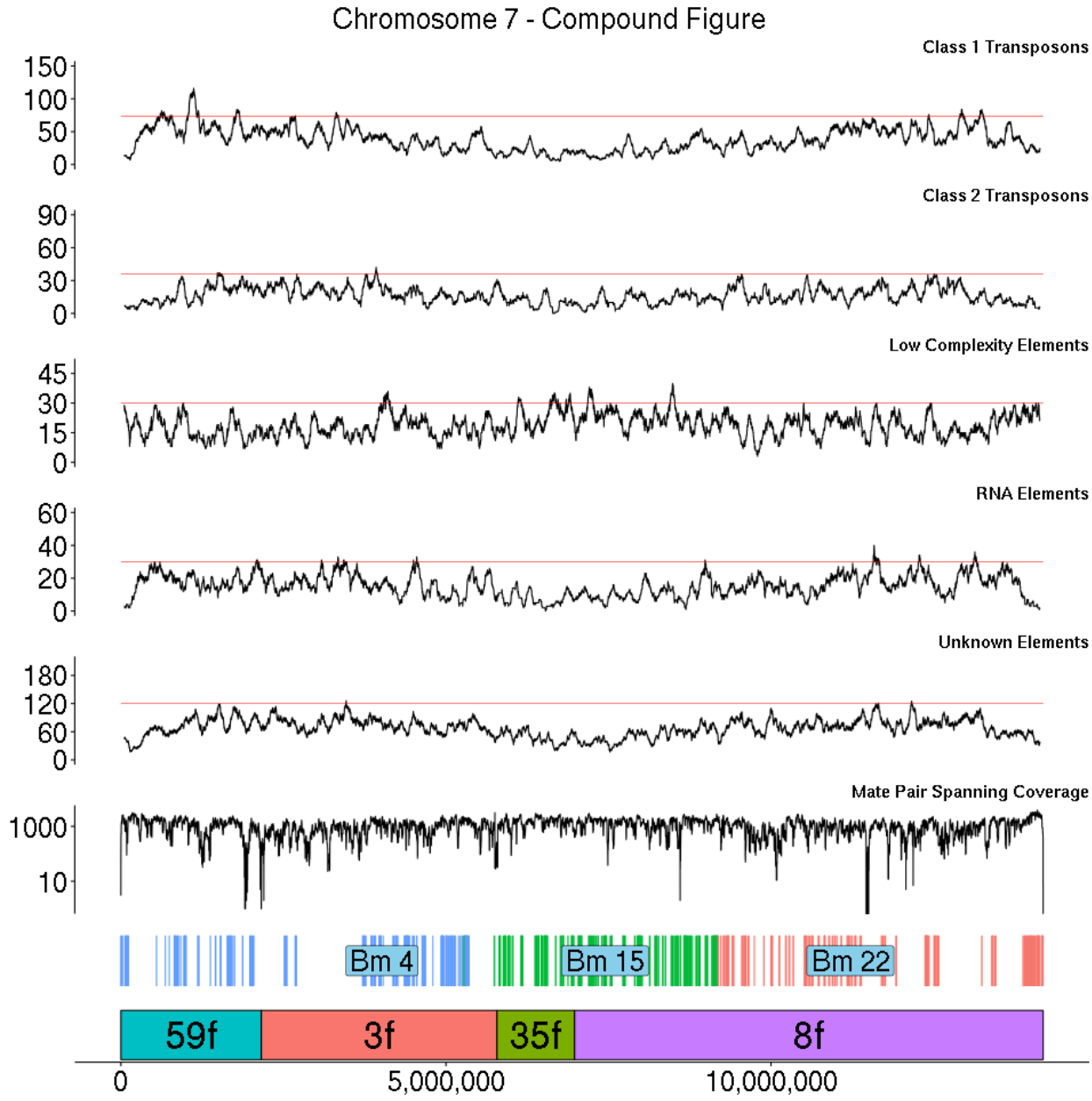

Chromosome 2 - Compound Figure
Class 1 Transposons
150
100
50
0
Class 2 Transposons
90
60
30
0
Low Complexity Elements
45
30
15
0
RNA Elements
60
40
20
0
Unknown Elements
180
120
60
0
Mate Pair Spanning Coverage
1000
10
Bm 25
Bm 5
Bm 19
12f
12f
12f
41f
3f
0
5,000,000
10,000,000
15,000,000

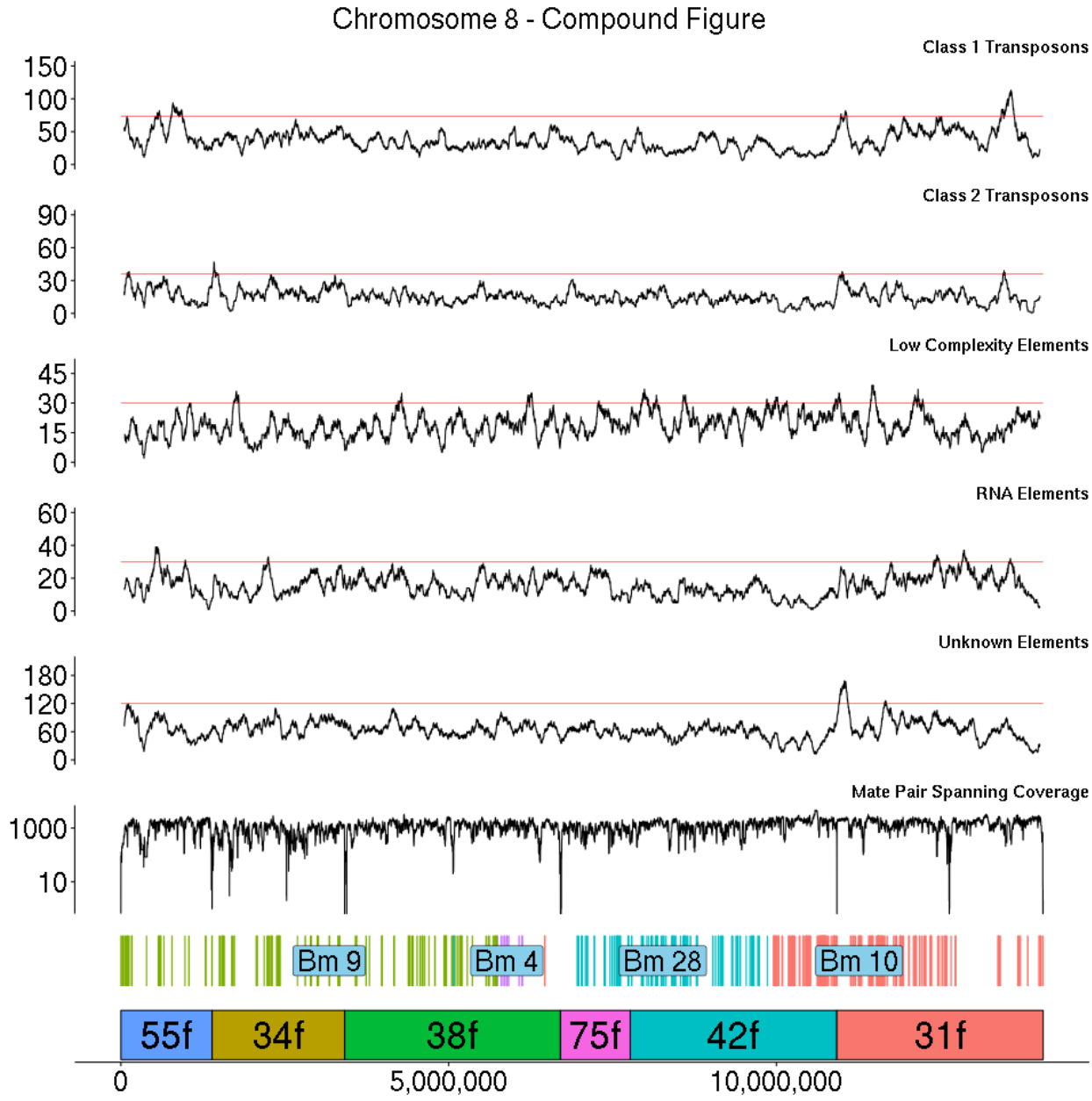

Chromosome 3 - Compound Figure
Class 1 Transposons
150
100
50
0
Class 2 Transposons
90
60
30
0
Low Complexity Elements
45
30
15
0
RNA Elements
60
40
20
0
Unknown Elements
180
120
60
0
Mate Pair Spanning Coverage
1000
10
Bm 19
Bm 16
Bm 23
53f
37f
39f
43f
23f
0
5,000,000
10,000,000
15,000,000

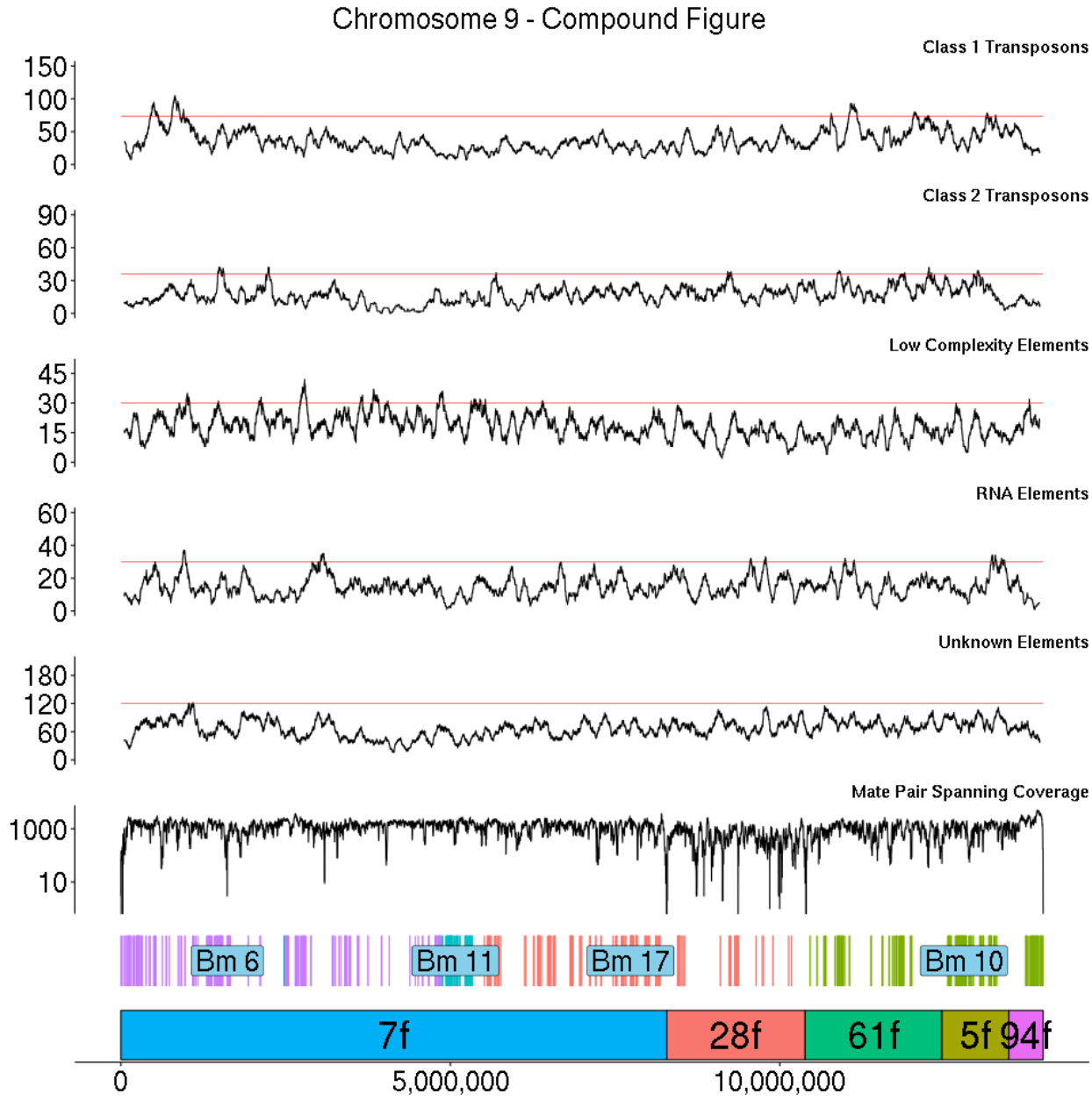

Chromosome 4 - Compound Figure
Class 1 Transposons
150
100
50
0
Class 2 Transposons
90
60
30
0
Low Complexity Elements
45
30
15
0
RNA Elements
60
40
20
0
Unknown Elements
180
120
60
0
Mate Pair Spanning Coverage
1000
10
Bm 15
Bm 26
Bm 3
Bm 9
86f
15f
10f
51f
0
5,000,000
10,000,000
15,000,00

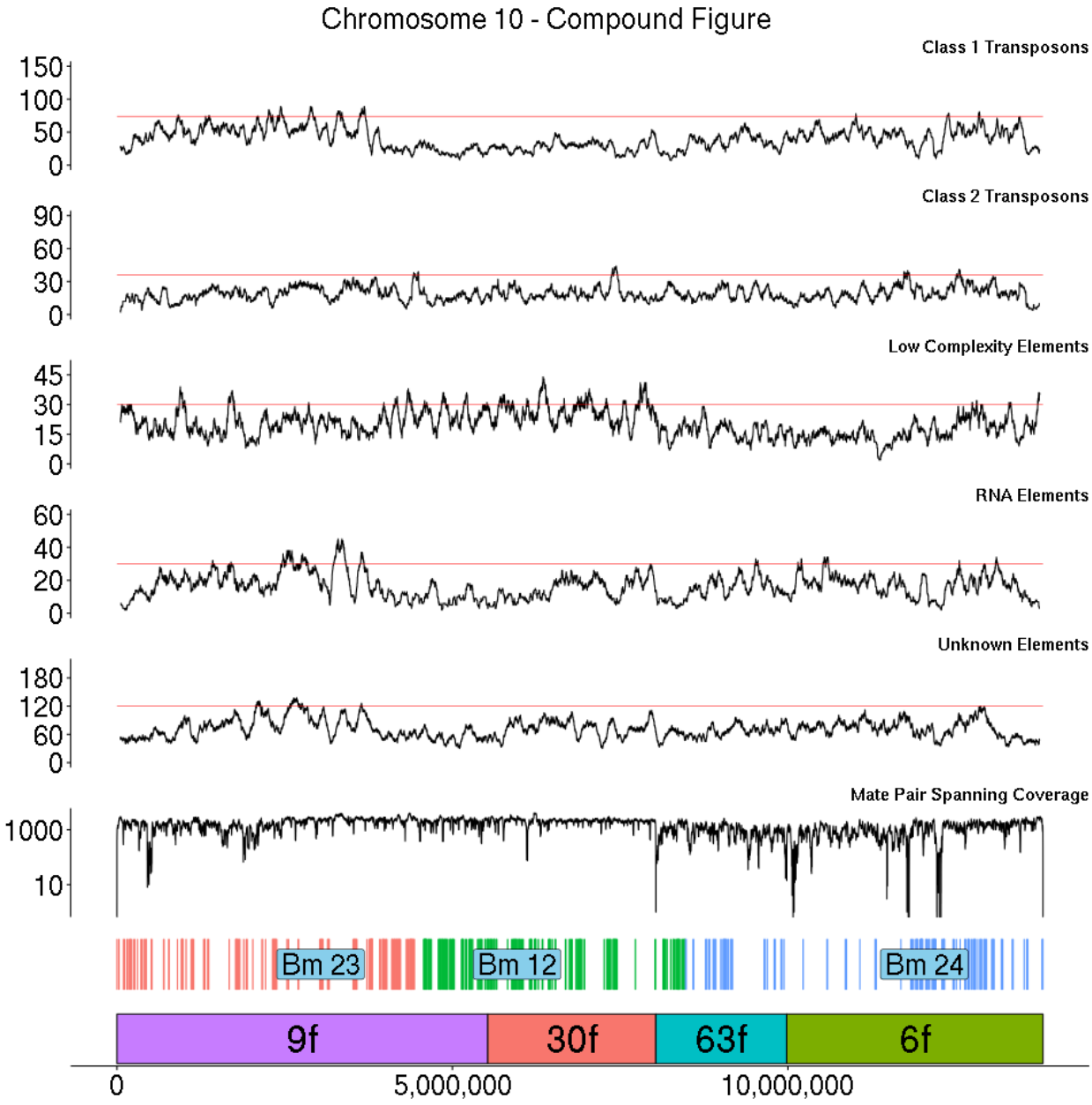

Chromosome 5 - Compound Figure
Class 1 Transposons
150
100
50
0
Class 2 Transposons
90
60
30
0
Low Complexity Elements
45
30
15
0
RNA Elements
60
40
20
0
Unknown Elements
180
120
60
0
Mate Pair Spanning Coverage
1000
10
Bm 21
Bm 20
0f
6f
81f
0
5,000,000
10,000,000
15,000,

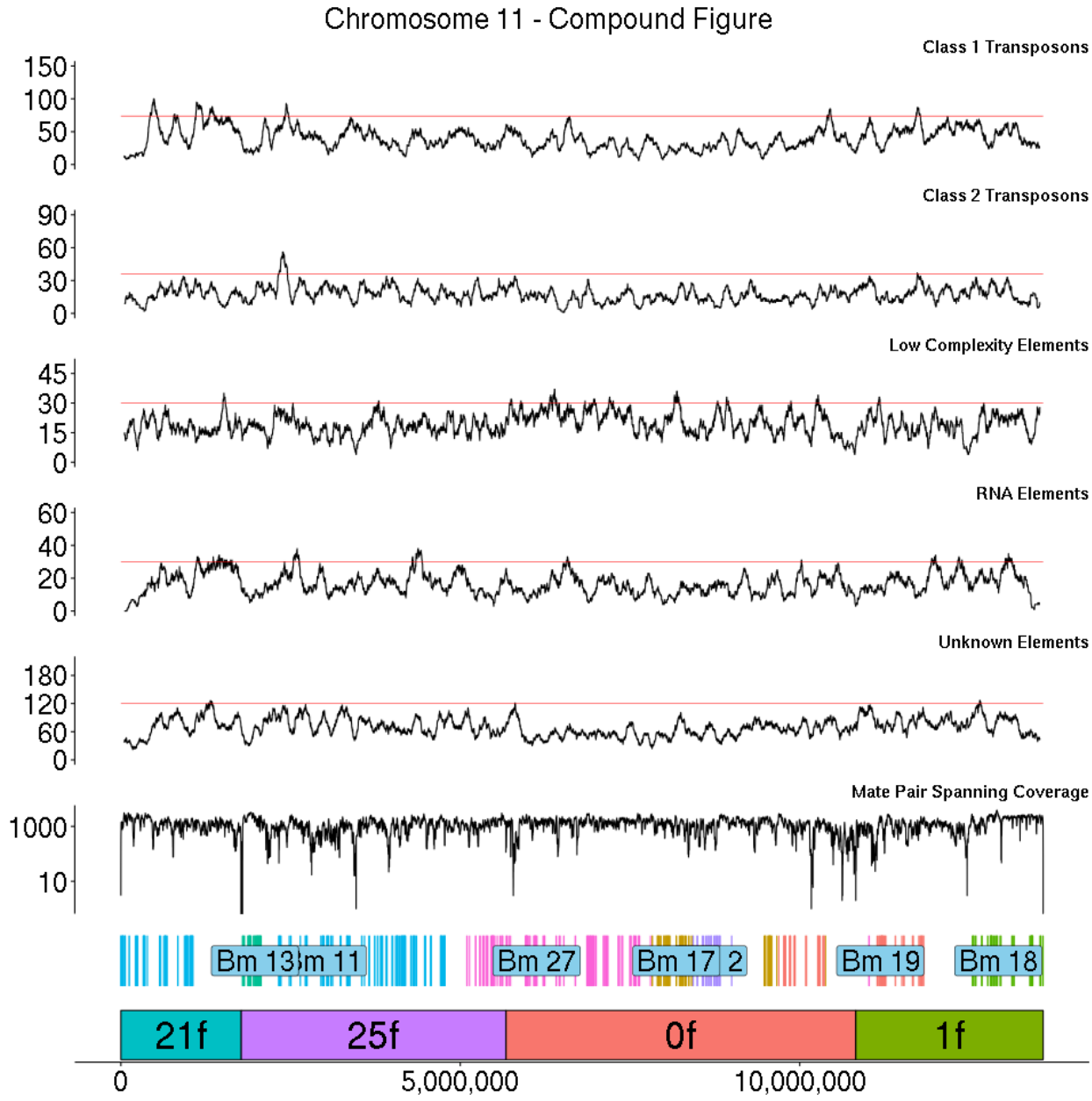

Chromosome 6 - Compound Figure
Class 1 Transposons
150
100
50
0
Class 2 Transposons
90
60
30
0
Low Complexity Elements
45
30
15
0
RNA Elements
60
40
20
0
Unknown Elements
180
120
60
0
Mate Pair Spanning Coverage
1000
10
Bm 18
Bm 7
Bm 14
72f
5f
27f
48f
4f
48f
97f
28f
34f
0
5,000,000
10,000,000

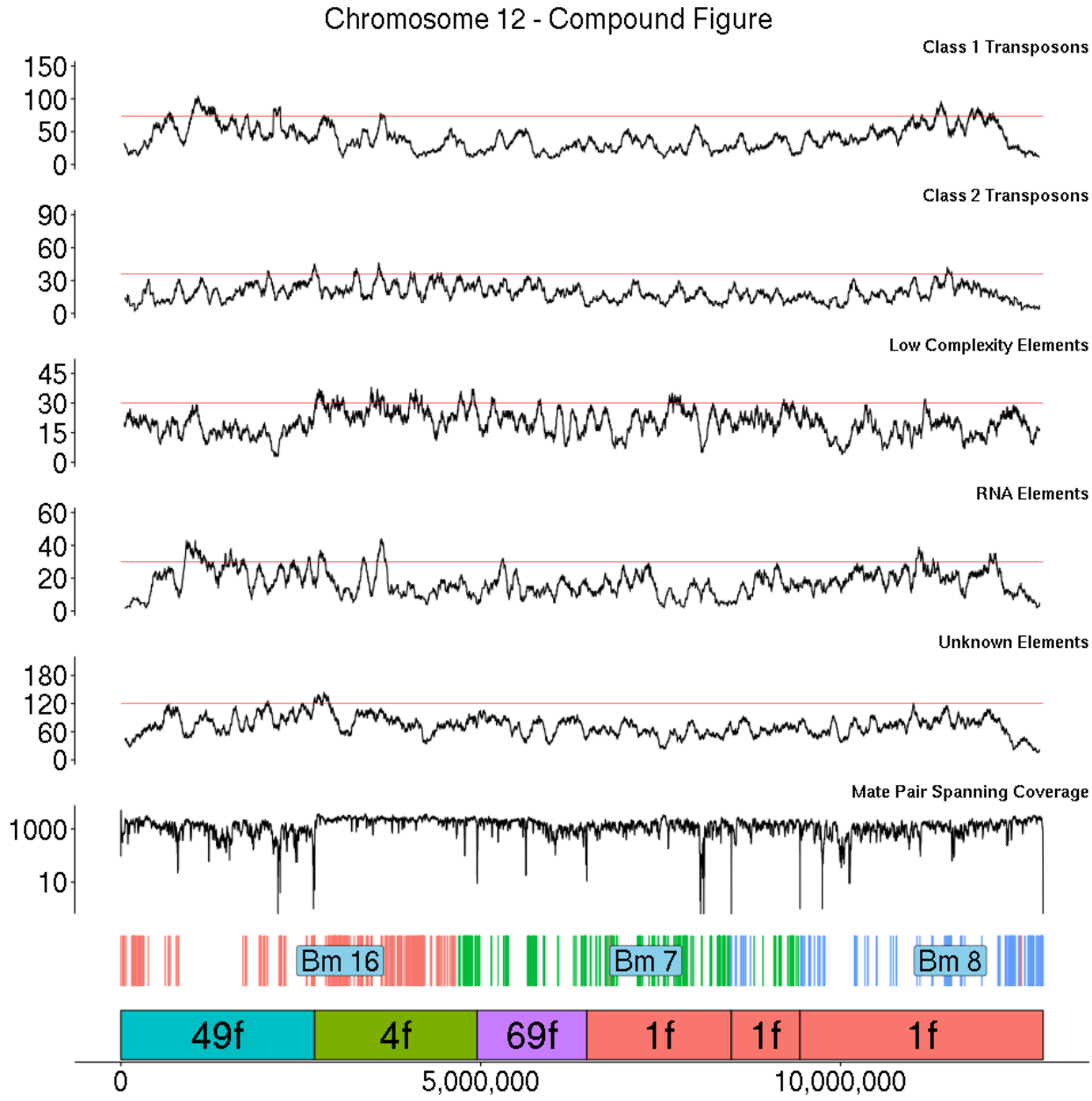

Chromosome 7 - Compound Figure
Class 1 Transposons
150
100
50
0
Class 2 Transposons
90
60
30
0
Low Complexity Elements
45
30
15
0
RNA Elements
60
40
20
0
Unknown Elements
180
120
60
0
Mate Pair Spanning Coverage
1000
10
Bm 4
Bm 15
Bm 22
59f
3f
35f
8f
0
5,000,000
10,000,000

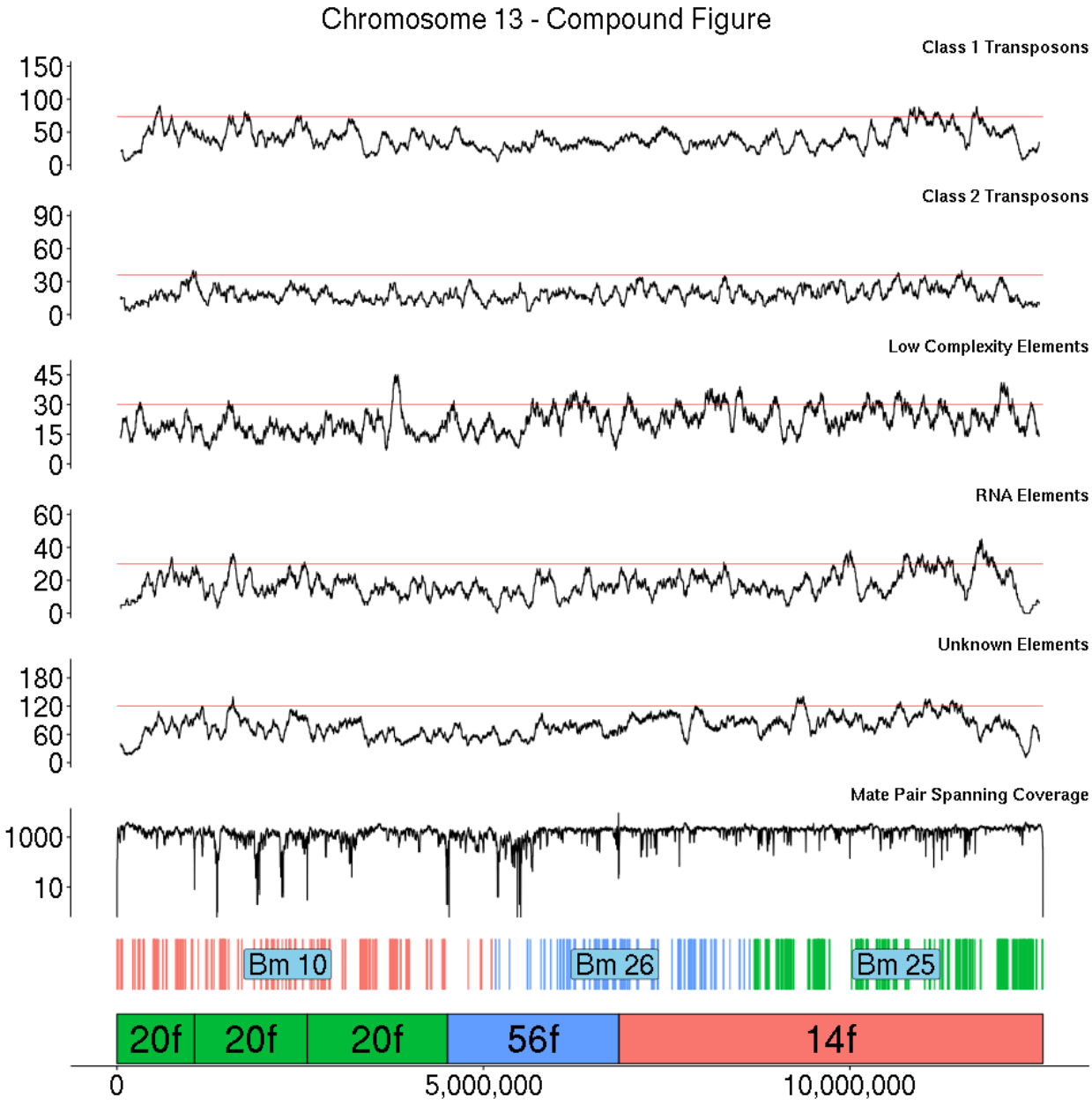

Chromosome 8 - Compound Figure
Class 1 Transposons
150
100
50
0
Class 2 Transposons
90
60
30
0
Low Complexity Elements
45
30
15
0
RNA Elements
60
40
20
0
Unknown Elements
180
120
60
0
Mate Pair Spanning Coverage
1000
10
Bm 9
Bm 4
Bm 28
Bm 10
55f
34f
38f
75f
42f
31f
0
5,000,000
10,000,000

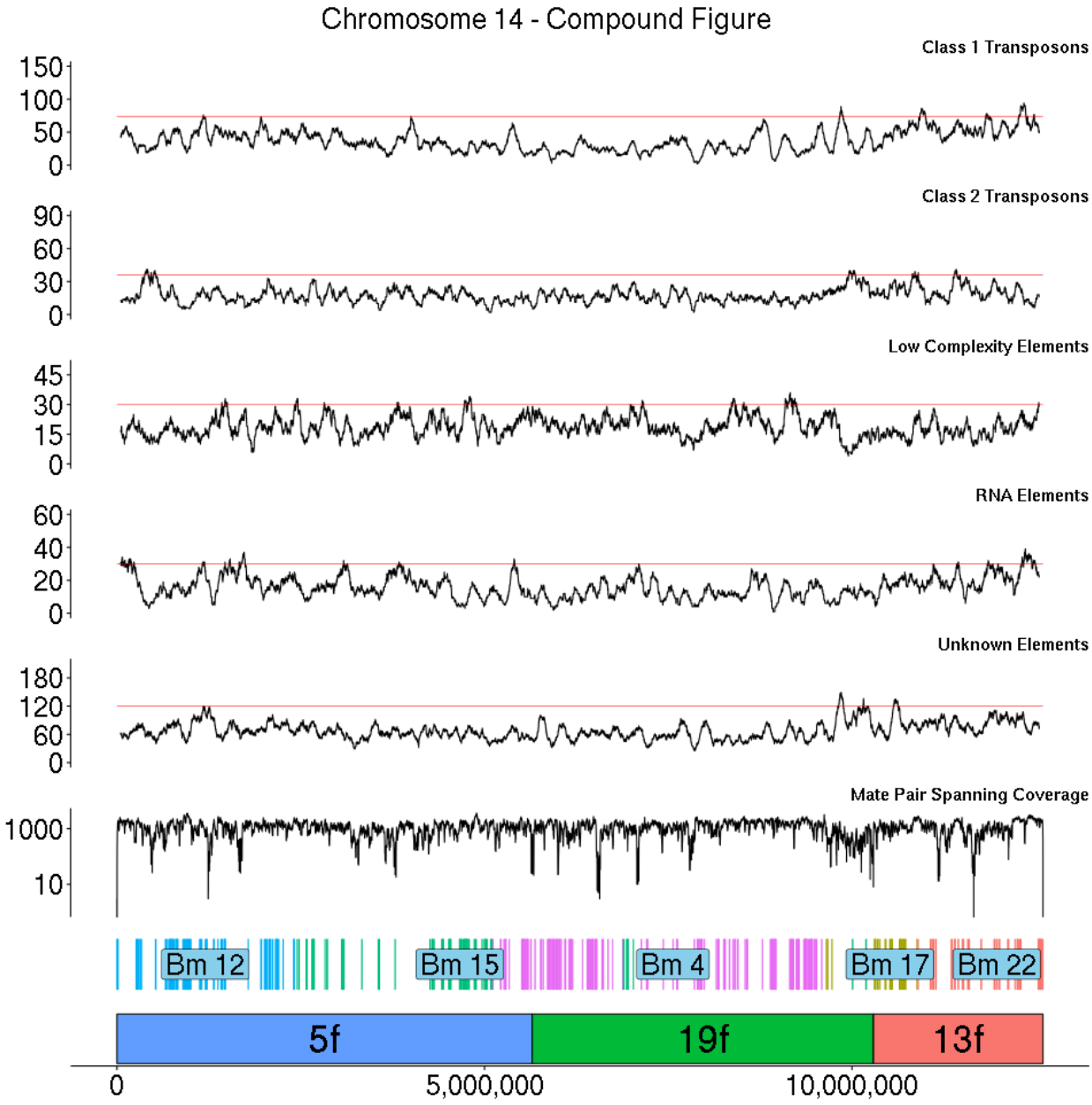

Chromosome 9 - Compound Figure
Class 1 Transposons
150
100
50
0
Class 2 Transposons
90
60
30
0
Low Complexity Elements
45
30
15
0
RNA Elements
60
40
20
0
Unknown Elements
180
120
60
0
Mate Pair Spanning Coverage
1000
10
Bm 6
Bm 11
Bm 17
Bm 10
7f
28f
61f
5f
94f
0
5,000,000
10,000,000

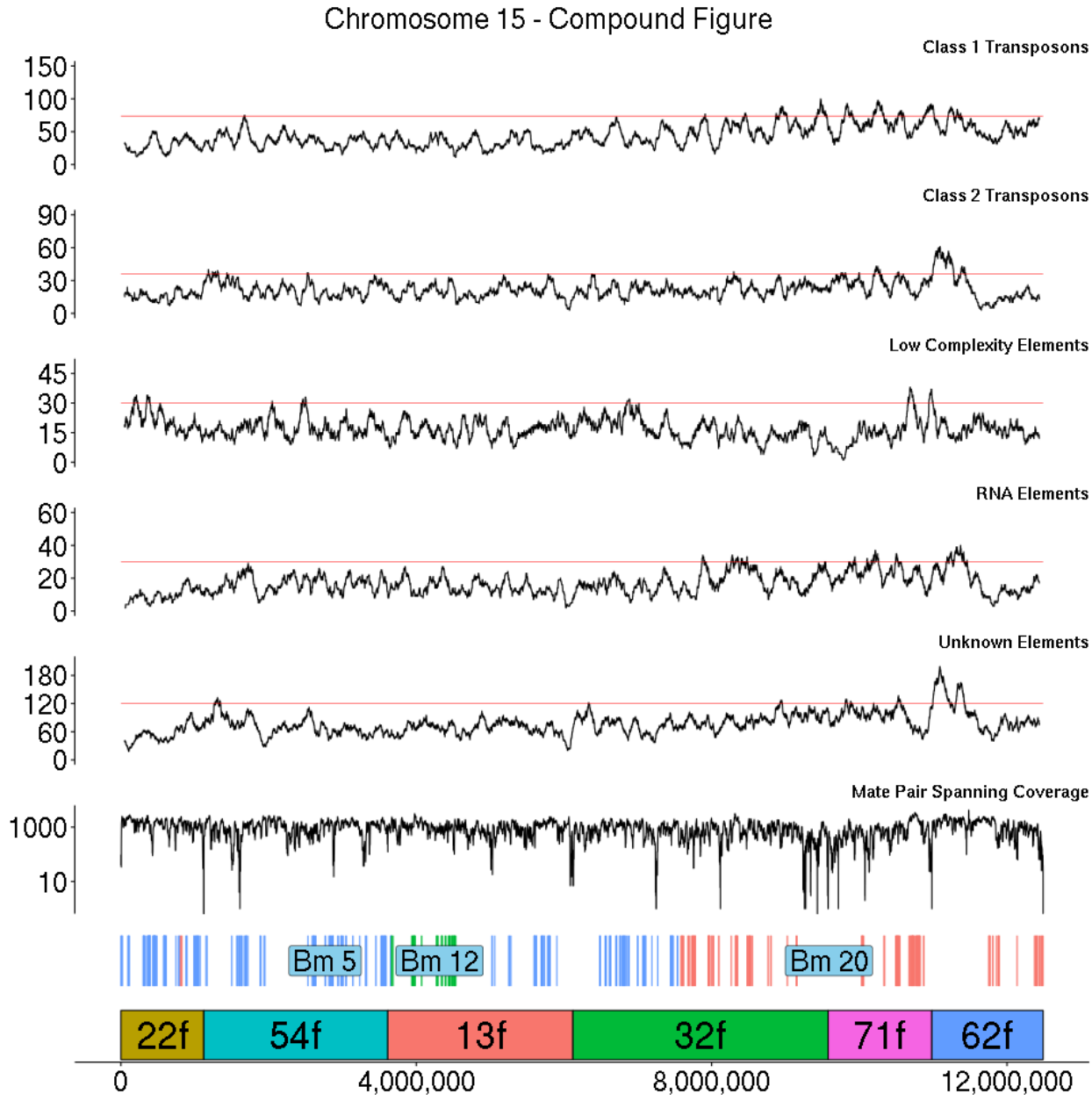

Chromosome 10 - Compound Figure
Class 1 Transposons
150
100
50
0
Class 2 Transposons
90
60
30
0
Low Complexity Elements
45
30
15
0
RNA Elements
60
40
20
0
Unknown Elements
180
120
60
0
Mate Pair Spanning Coverage
1000
10
Bm 23
Bm 12
Bm 24
9f
30f
63f
6f
0
5,000,000
10,000,000

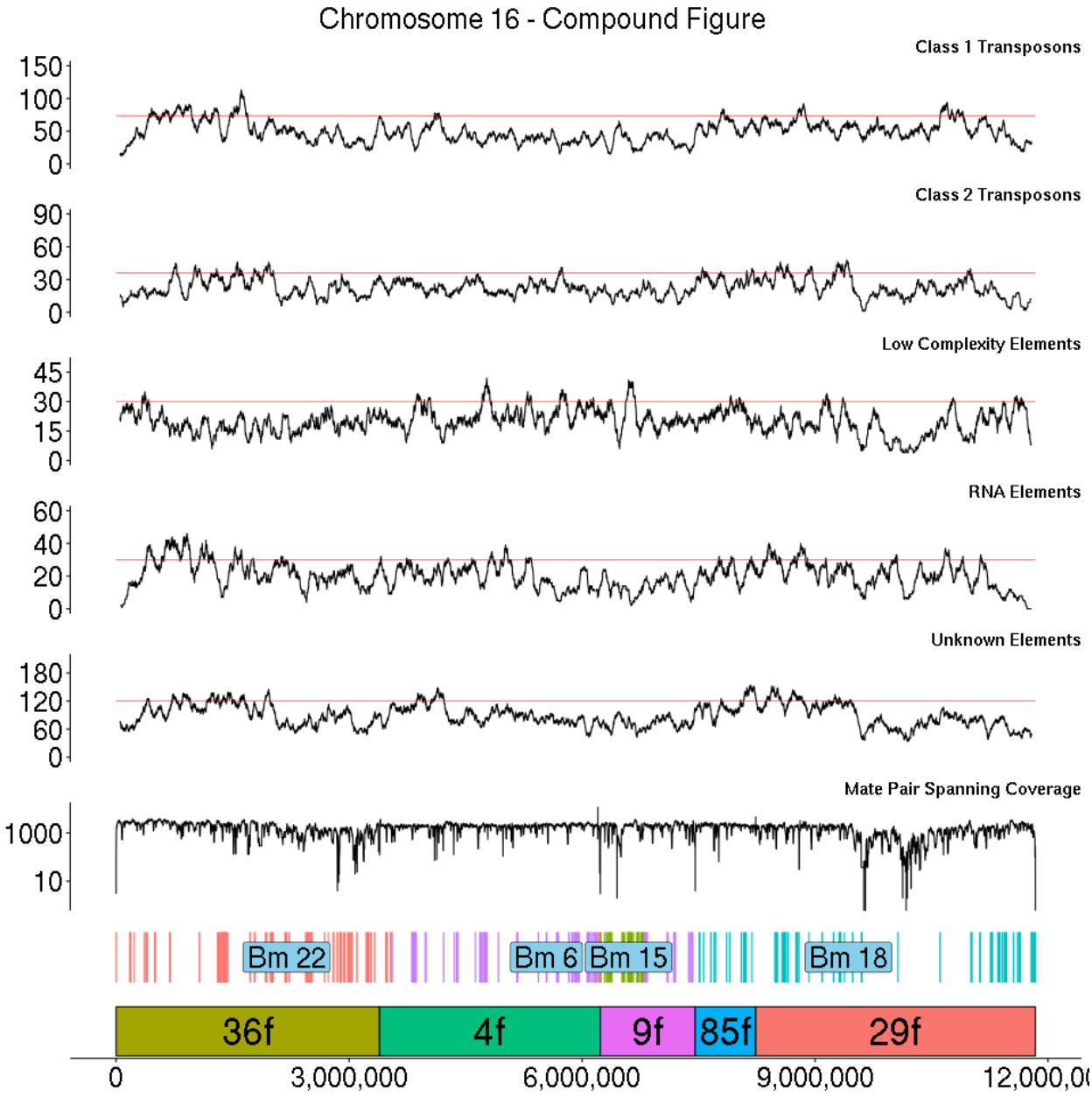

Chromosome 11 - Compound Figure
Class 1 Transposons
150
100
50
0
Class 2 Transposons
90
60
30
0
Low Complexity Elements
45
30
15
0
RNA Elements
60
40
20
0
Unknown Elements
180
120
60
0
Mate Pair Spanning Coverage
1000
10
Bm 13
m 11
Bm 27
Bm 17
2
Bm 19
Bm 18
21f
25f
0f
1f
0
5,000,000
10,000,000

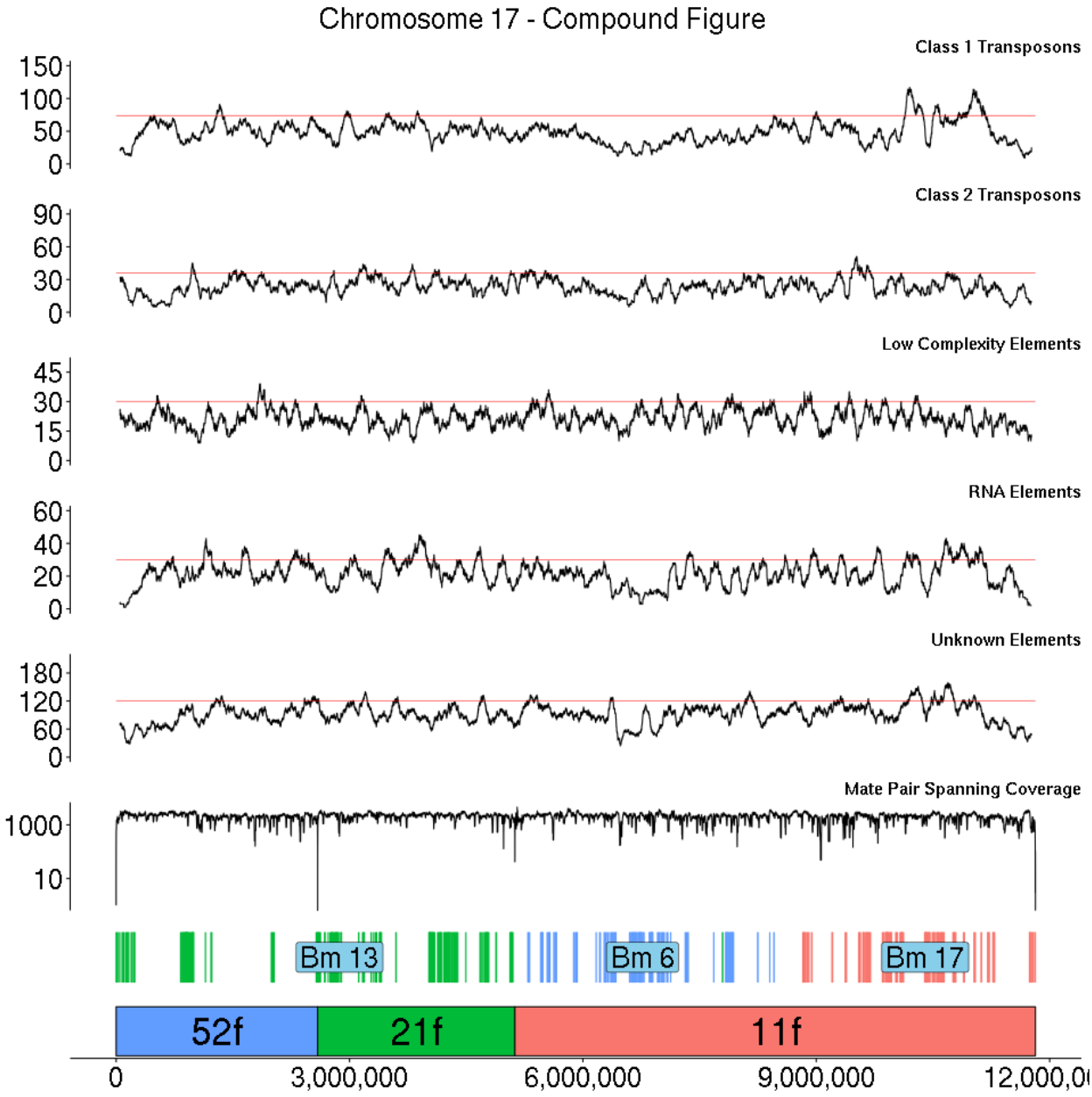

Chromosome 12 - Compound Figure
Class 1 Transposons
150
100
50
0
Class 2 Transposons
90
60
30
0
Low Complexity Elements
45
30
15
0
RNA Elements
60
40
20
0
Unknown Elements
180
120
60
0
Mate Pair Spanning Coverage
1000
10
Bm 16
Bm 7
Bm 8
49f
4f
69f
1f
1f
1f
0
5,000,000
10,000,000

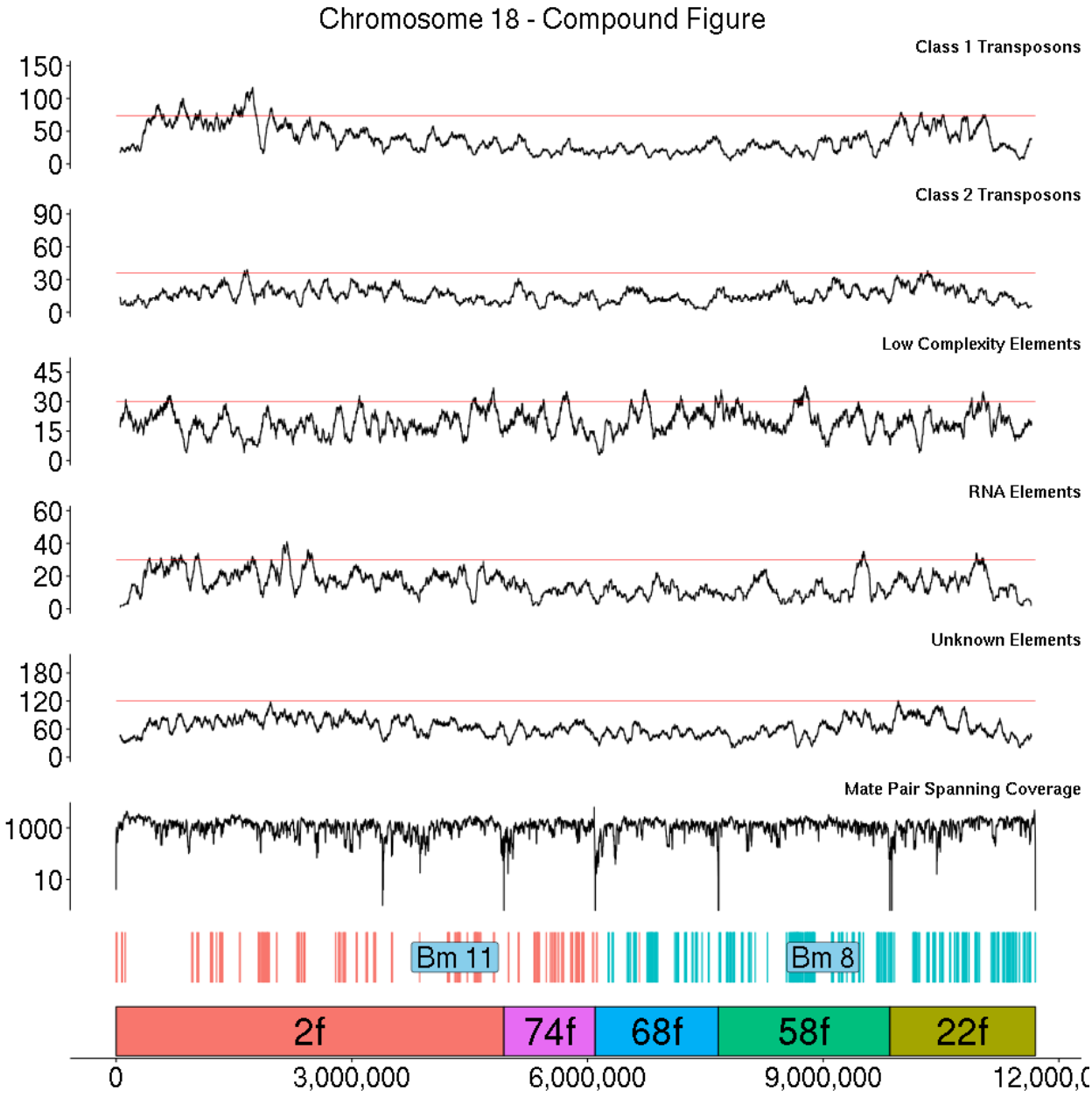

Chromosome 13 - Compound Figure
Class 1 Transposons
150
100
50
0
Class 2 Transposons
90
60
30
0
Low Complexity Elements
45
30
15
0
RNA Elements
60
40
20
0
Unknown Elements
180
120
60
0
Mate Pair Spanning Coverage
1000
10
Bm 10
Bm 26
Bm 25
20f
20f
20f
56f
14f
0
5,000,000
10,000,000

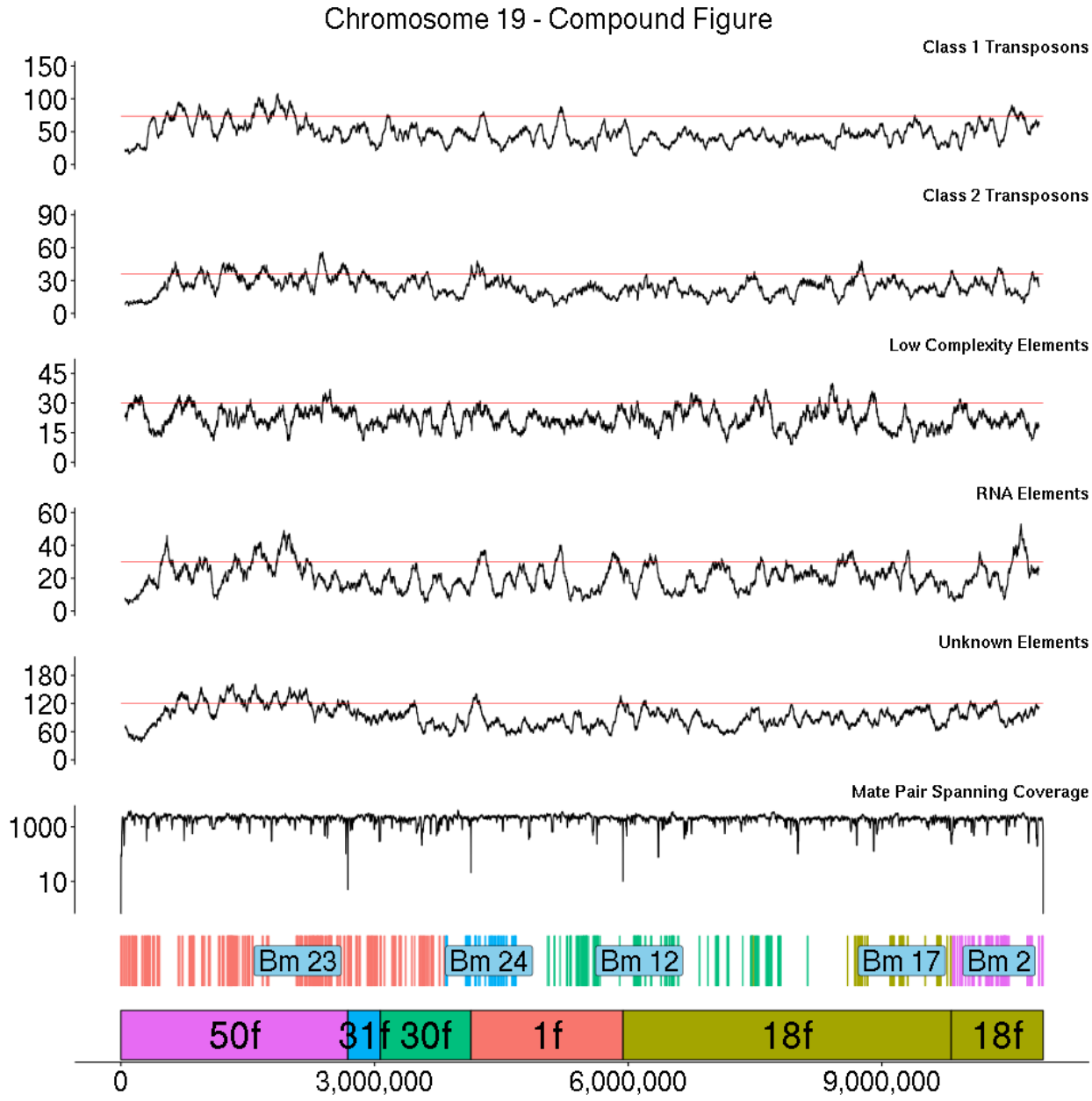

Chromosome 14 - Compound Figure
Class 1 Transposons
150
100
50
0
Class 2 Transposons
90
60
30
0
Low Complexity Elements
45
30
15
0
RNA Elements
60
40
20
0
Unknown Elements
180
120
60
0
Mate Pair Spanning Coverage
1000
10
Bm 12
Bm 15
Bm 4
Bm 17
Bm 22
5f
19f
13f
0
5,000,000
10,000,000

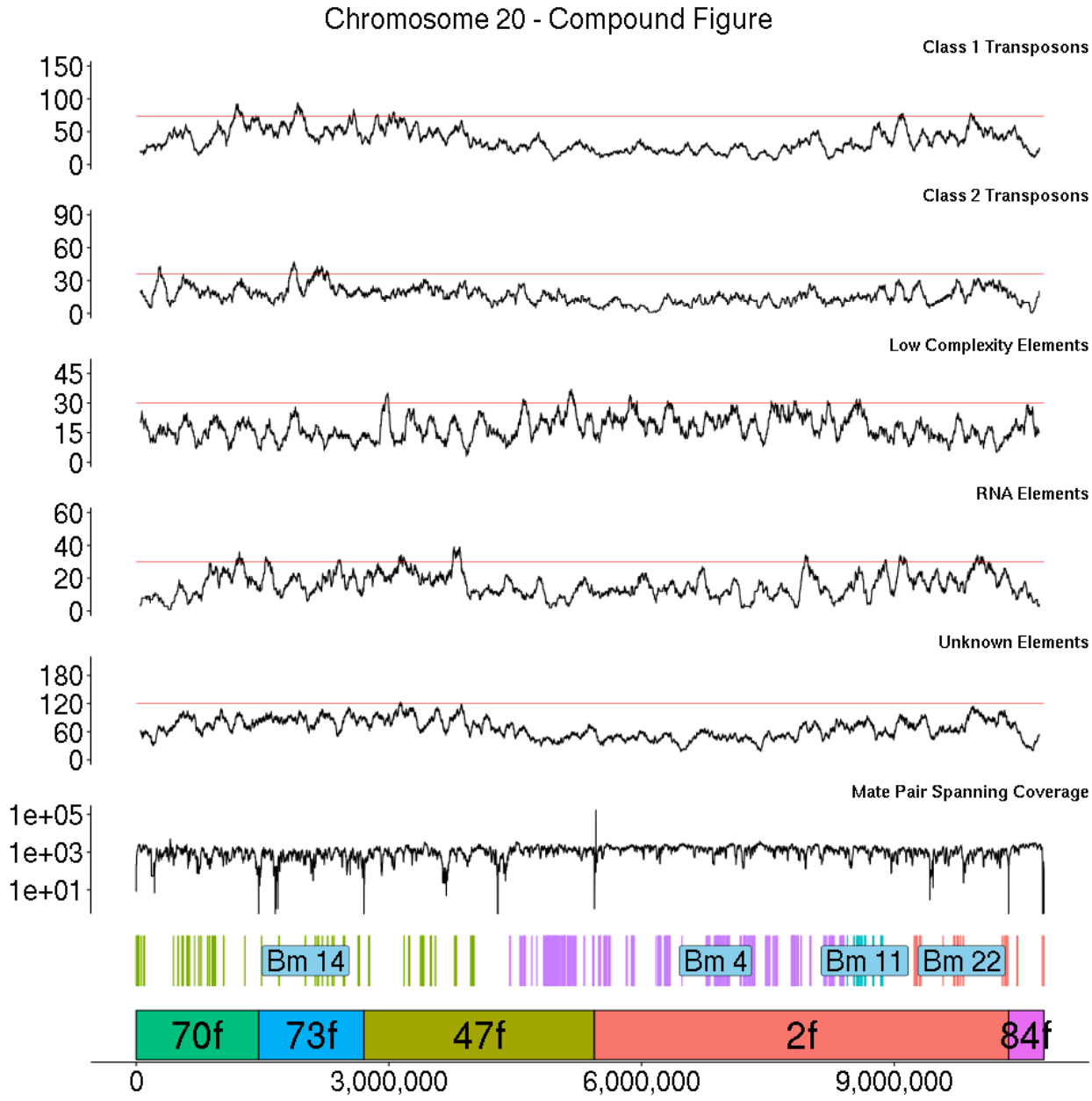

Chromosome 15 - Compound Figure
Class 1 Transposons
150
100
50
0
Class 2 Transposons
90
60
30
0
Low Complexity Elements
45
30
15
0
RNA Elements
60
40
20
0
Unknown Elements
180
120
60
0
Mate Pair Spanning Coverage
1000
10
Bm 5
Bm 12
Bm 20
22f
54f
13f
32f
71f
62f
0
4,000,000
8,000,000
12,000,000

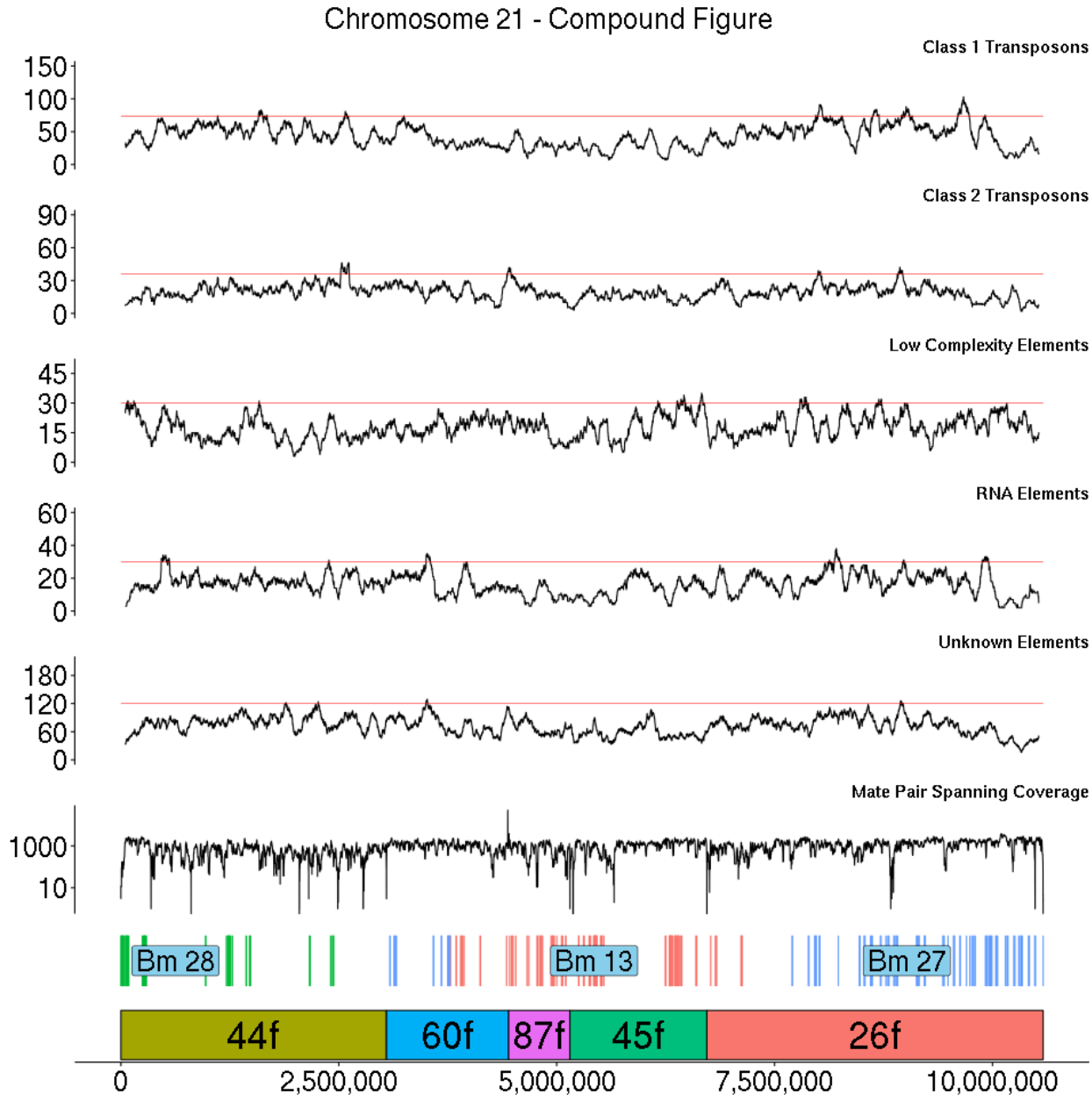

Chromosome 16 - Compound Figure
Class 1 Transposons
150
100
50
0
Class 2 Transposons
90
60
30
0
Low Complexity Elements
45
30
15
0
RNA Elements
60
40
20
0
Unknown Elements
180
120
60
0
Mate Pair Spanning Coverage
1000
10
Bm 22
Bm 6
Bm 15
Bm 18
36f
4f
9f
85f
29f
0
3,000,000
6,000,000
9,000,000
12,000,0

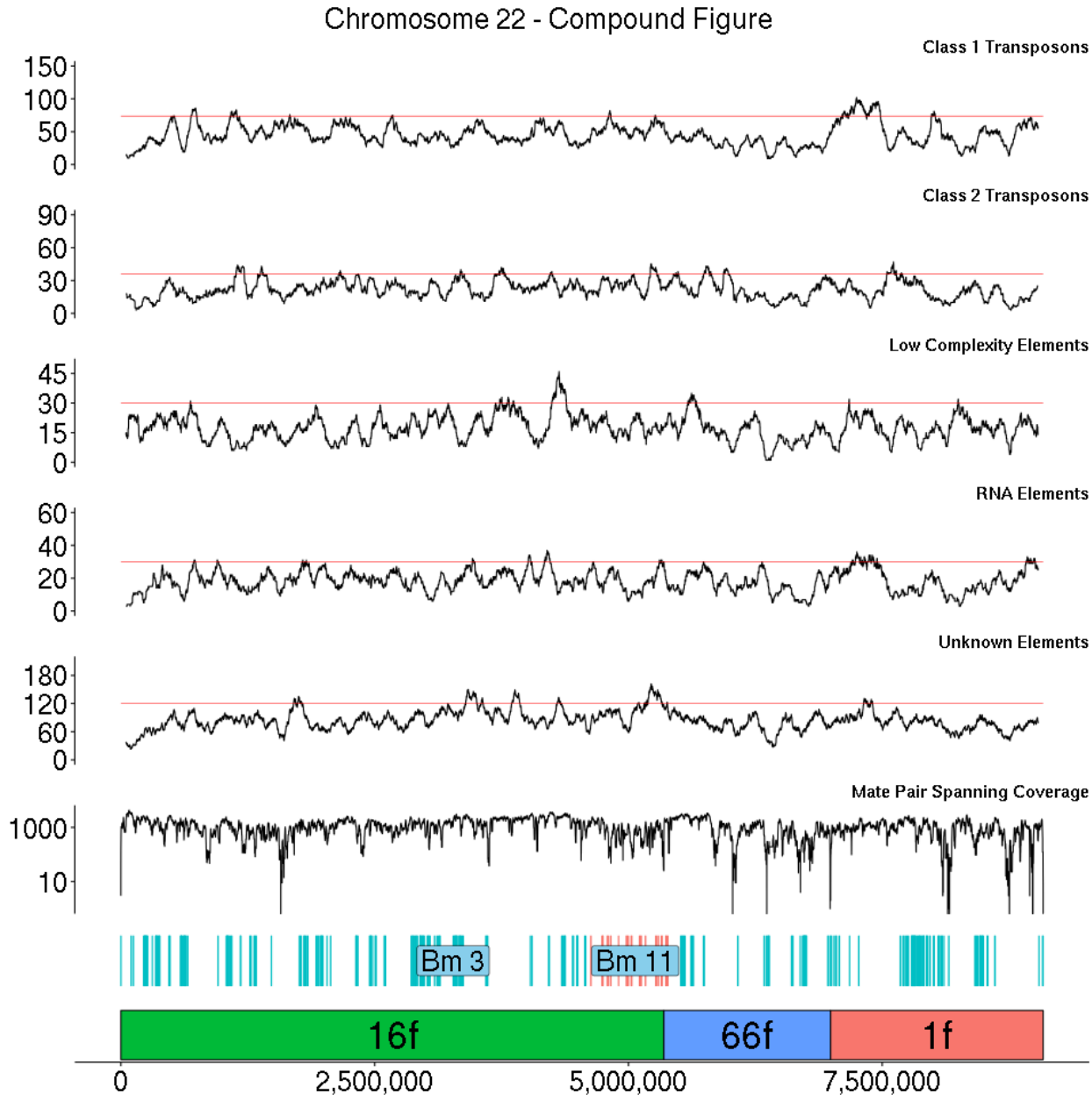

Chromosome 17 - Compound Figure
Class 1 Transposons
150
100
50
0
Class 2 Transposons
90
60
30
0
Low Complexity Elements
45
30
15
0
RNA Elements
60
40
20
0
Unknown Elements
180
120
60
0
Mate Pair Spanning Coverage
1000
10
Bm 13
Bm 6
Bm 17
52f
21f
11f
0
3,000,000
6,000,000
9,000,000
12,000,0

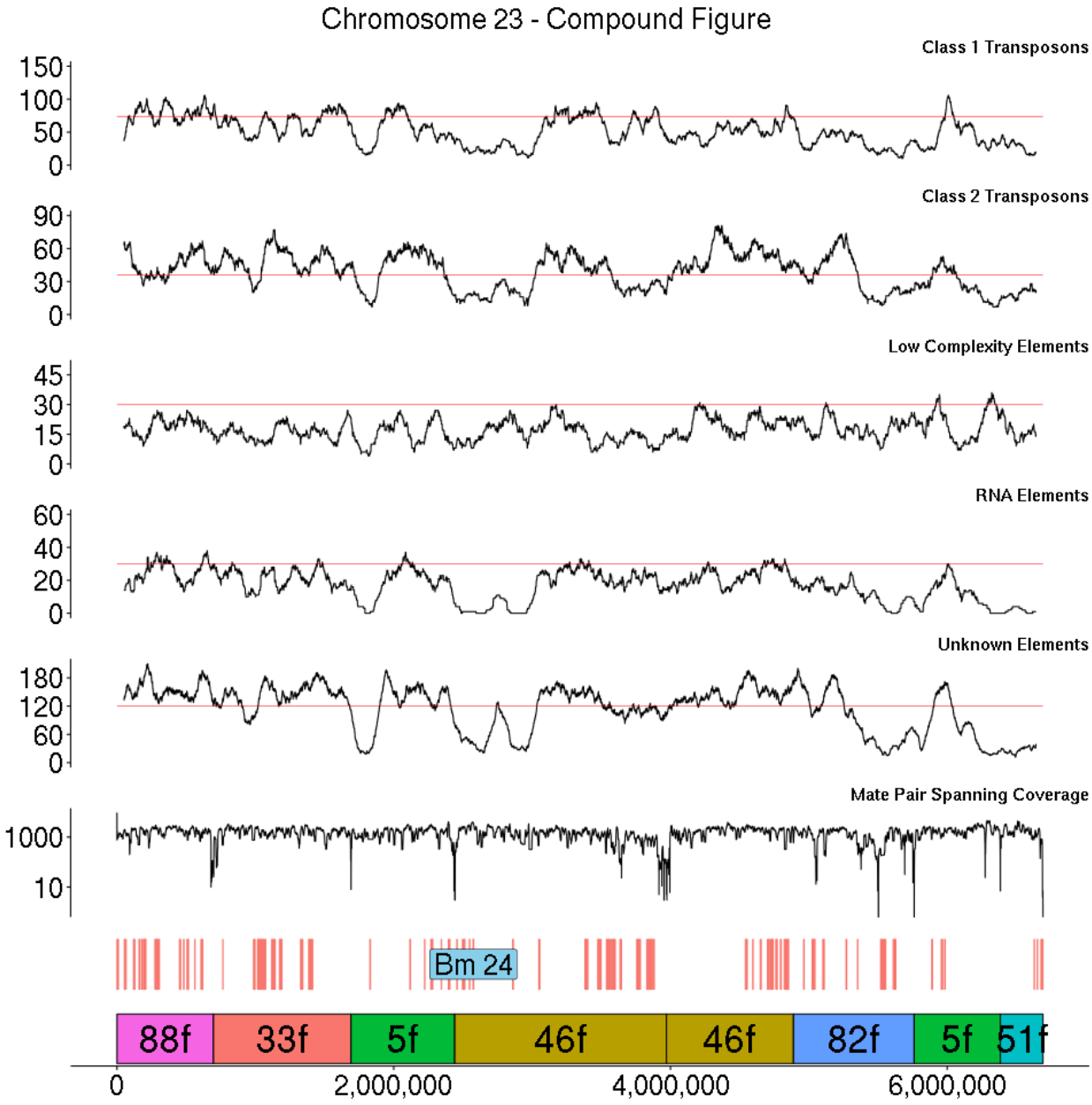

Chromosome 18 - Compound Figure
Class 1 Transposons
150
100
50
0
Class 2 Transposons
90
60
30
0
Low Complexity Elements
45
30
15
0
RNA Elements
60
40
20
0
Unknown Elements
180
120
60
0
Mate Pair Spanning Coverage
1000
10
Bm 11
Bm 8
2f
74f
68f
58f
22f
0
3,000,000
6,000,000
9,000,000
12,000,0

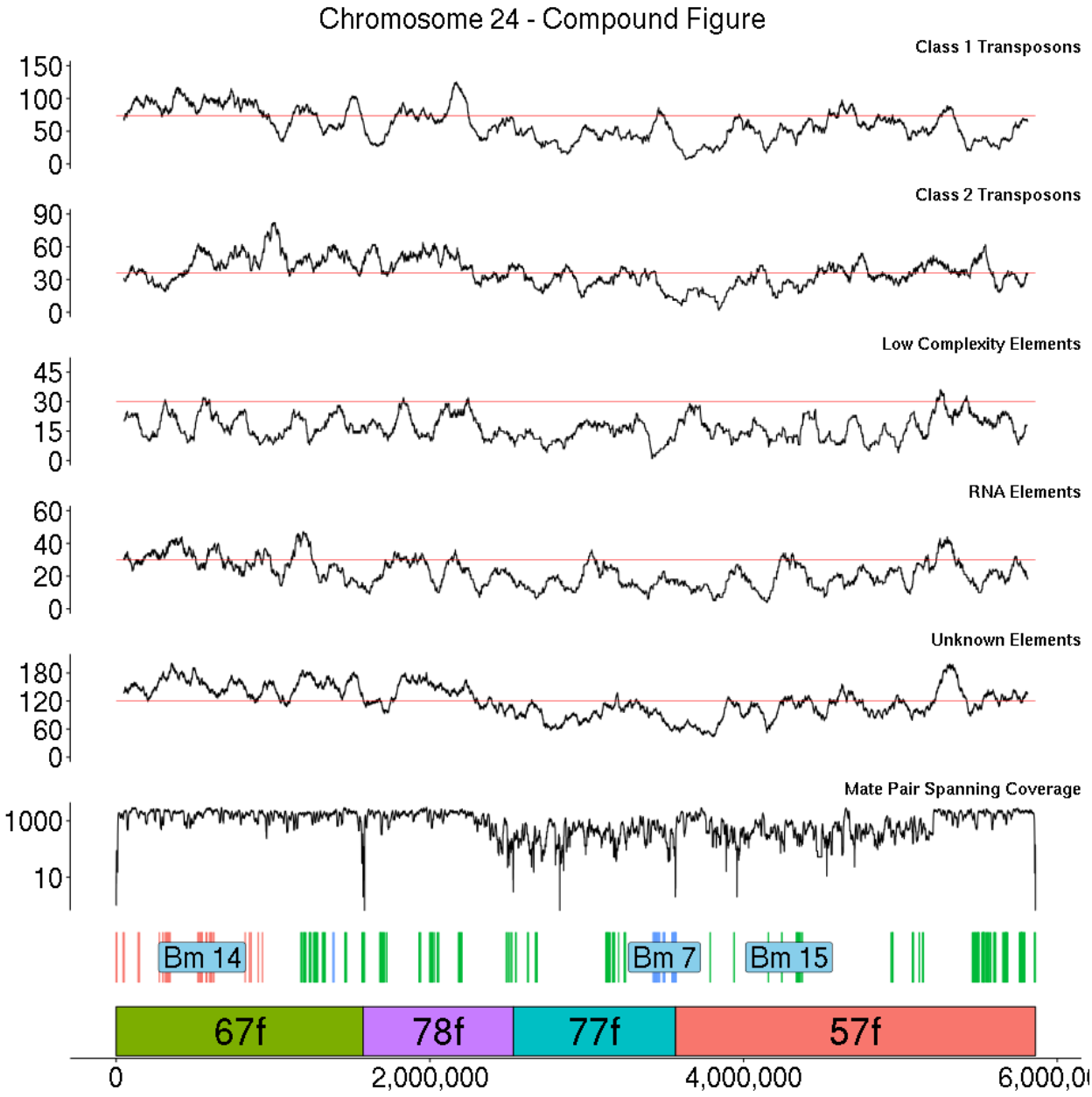

Chromosome 19 - Compound Figure
Class 1 Transposons
150
100
50
0
Class 2 Transposons
90
60
30
0
Low Complexity Elements
45
30
15
0
RNA Elements
60
40
20
0
Unknown Elements
180
120
60
0
Mate Pair Spanning Coverage
1000
10
Bm 23
Bm 24
Bm 12
Bm 17
Bm 2
50f
31f
30f
1f
18f
18f
0
3,000,000
6,000,000
9,000,000


Chromosome 20 - Compound Figure
Class 1 Transposons
150
100
50
0
Class 2 Transposons
90
60
30
0
Low Complexity Elements
45
30
15
0
RNA Elements
60
40
20
0
Unknown Elements
180
120
60
0
Mate Pair Spanning Coverage
1e+05
1e+03
1e+01
Bm 14
Bm 4
Bm 11
Bm 22
70f
73f
47f
2f
84f
0
3,000,000
6,000,000
9,000,000


Chromosome 21 - Compound Figure
Class 1 Transposons
150
100
50
0
Class 2 Transposons
90
60
30
0
Low Complexity Elements
45
30
15
0
RNA Elements
60
40
20
0
Unknown Elements
180
120
60
0
Mate Pair Spanning Coverage
1000
10
Bm 28
Bm 13
Bm 27
44f
60f
87f
45f
26f
0
2,500,000
5,000,000
7,500,000
10,000,000


Chromosome 22 - Compound Figure
Class 1 Transposons
150
100
50
0
Class 2 Transposons
90
60
30
0
Low Complexity Elements
45
30
15
0
RNA Elements
60
40
20
0
Unknown Elements
180
120
60
0
Mate Pair Spanning Coverage
1000
10
Bm 3
Bm 11
16f
66f
1f
0
2,500,000
5,000,000
7,500,000


Chromosome 23 - Compound Figure
Class 1 Transposons
150
100
50
0
Class 2 Transposons
90
60
30
0
Low Complexity Elements
45
30
15
0
RNA Elements
60
40
20
0
Unknown Elements
180
120
60
0
Mate Pair Spanning Coverage
1000
10
Bm 24
88f
33f
5f
46f
46f
82f
5f
51f
0
2,000,000
4,000,000
6,000,000


Chromosome 24 - Compound Figure
Class 1 Transposons
150
100
50
0
Class 2 Transposons
90
60
30
0
Low Complexity Elements
45
30
15
0
RNA Elements
60
40
20
0
Unknown Elements
180
120
60
0
Mate Pair Spanning Coverage
1000
10
Bm 14
Bm 7
Bm 15
67f
78f
77f
57f
0
2,000,000
4,000,000
6,000,0

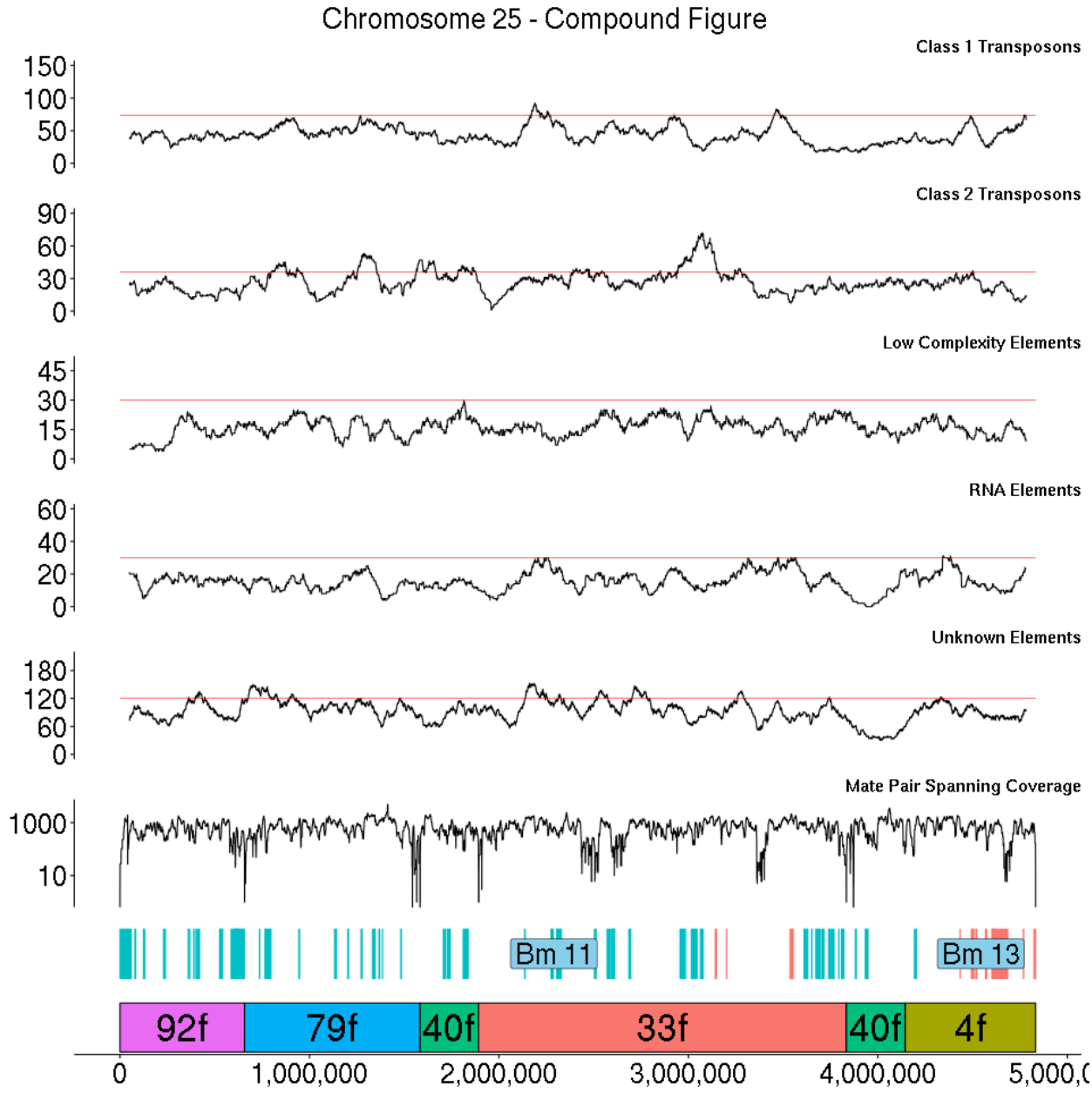


Fig. S5. Repetitive element distribution across chromosomes. Repetitive elements were classified using Repeat masker v.4.0.5. The number of repeats in 10,000 base pair windows are plotted along each chromosome for Class 1 and Class 2 transposons, low complexity elements, RNA elements, and unknown elements. Mate-pair spanning depth, collinear block identity, and component scaffold identity are also shown.

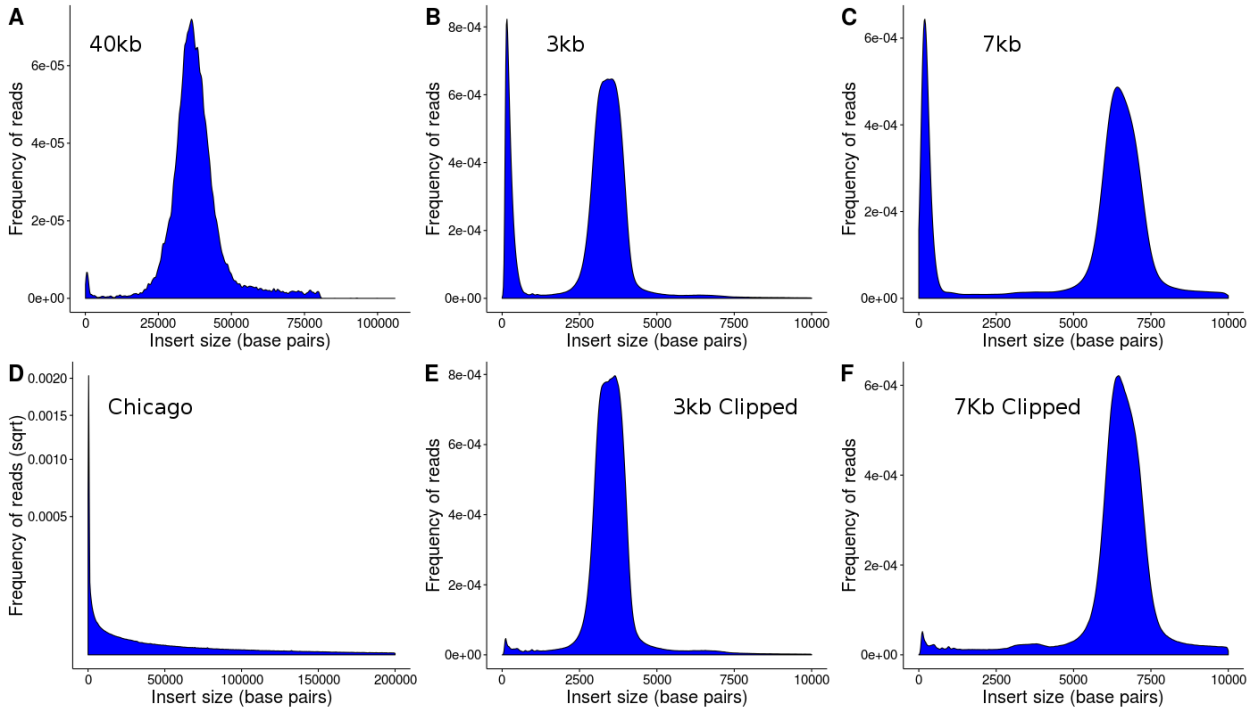


Fig. S6. Insert size distribution. Insert size distribution of mate pair libraries used in genome construction and assessment. Insert sizes determined after mapping libraries back to final assembly with bbmap(*39*) after libraries were quality and adapter trimmed with bbduk. A) 40kb library, B) 3kb library, C) 7kb library, D) Chicago library, E) 3kb library after filtering out paired end reads that were included as an unwanted byproduct of library construction, and F) 7kb library after nextclip(*38*) filtering.

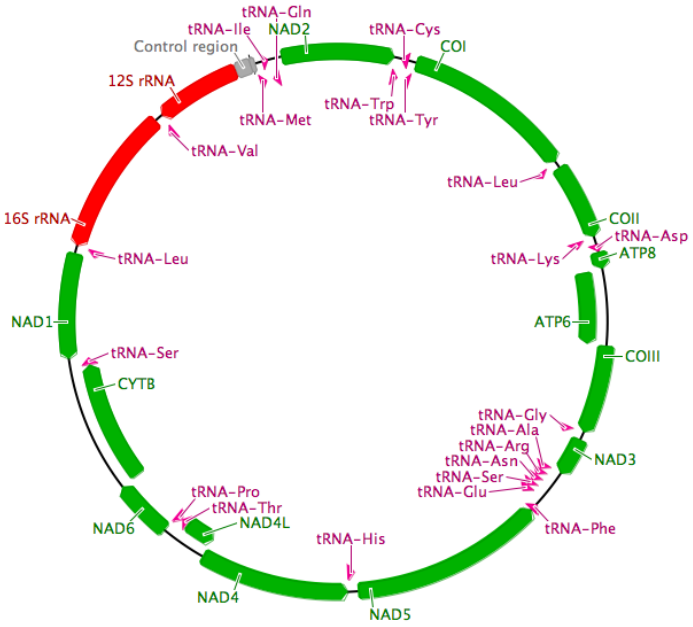


Fig. S7. The mtDNA of *P. napi*.

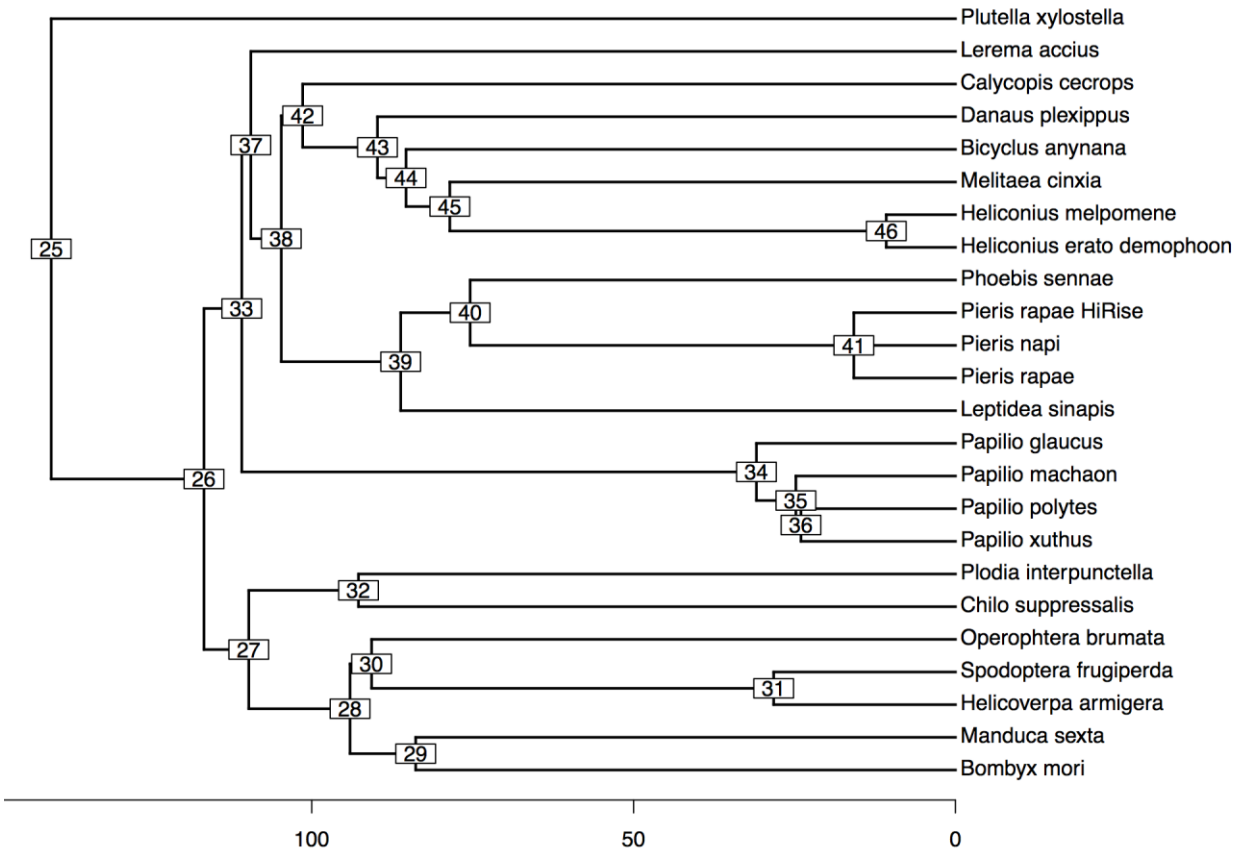
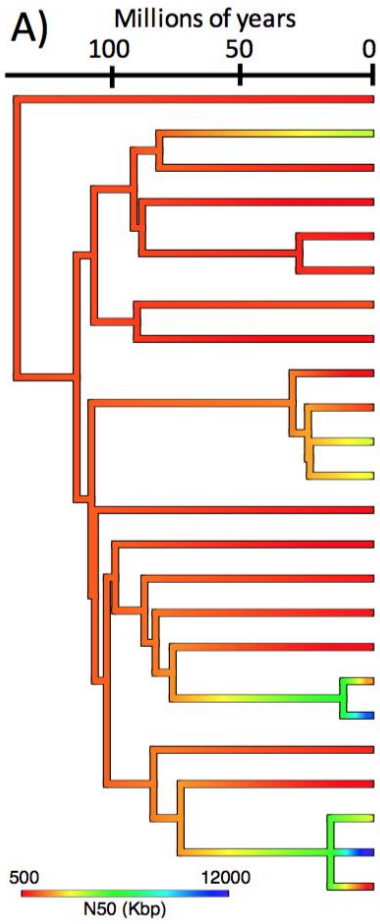


Fig. S8. Chronogram of lepidopteran genomes with node labels. A chronogram of currently available Lepidopteran genomes (n=24) with nodes identified that correspond to Supplemental Table S3, with time in million years ago (MYA).



B)

Species	Family	N50 (kb)	Scaffold Joins *napi,mori*	Scaffold Joins *napi / mori*
Plutella xylostella	Plutellidae	737	8,31	0.26
Bombyx mori	Bombycidae	4008	4,37	0.11
Manduca sexta	Sphingidae	664	0,0	NA
Operophtera brumata	Geometridae	65.6	1,6	0.17
Spodoptera frugiperda	Noctuidae	601	10,17	0.59
Helicoverpa armigera	Noctuidae	1000	5,34	0.15
Plodia interpunctella	Pyralidae	1300	4,25	0.16
Chilo suppressalis	Crambidae	5.2	0,0	NA
Papilio glaucus	Papilionidae	230.3	7,38	0.18
Papilio machaon	Papilionidae	1200	7,22	0.32
Papilio polytes	Papilionidae	3700	2,21	0.10
Papilio xuthus	Papilionidae	3400	1,13	0.08
Lerema accius	Hesperiidae	525.3	6,30	0.20
Calycopis cecrops	Lycaenidae	233.5	2,30	0.07
Danaus plexippus	Nymphalidae	716	5,31	0.16
Bicyclus anynana	Nymphalidae	638.3	4,30	0.13
Melitaea cinxia	Nymphalidae	331	2,8	0.25
Heliconius melpomene	Nymphalidae	2103	3,19	0.02
Heliconius erato	Nymphalidae	10700	2,9	0.22
Leptidea sinapis	Pieridae	857	0,4	0.00
Phoebis sennae	Pieridae	256.7	7,31	0.23
Pieris rapae Shen	Pieridae	617	25,12	2.08
Pieris rapae Nallu	Pieridae	3706	26,5	5.20
Pieris napi noLM	Pieridae	300	21,14	1.50

Fig. S9. Comparative assessment of genome assemblies and chromosomal evolution across, using a BLAST-like approach. A) A chronogram of currently available Lepidopteran genomes (n=24) with nodes identified that correspond to Supplemental Table S3, with time in million years ago (MYA), showing N50 genome size. B) Table of the genome assembly N50 for each species, followed by estimates of their chromosomal similarity relative to *B. mori* vs. *P. napi*. In each scaffold of each genome, SCOs were identified that were shared with *B. mori* and *P. napi*. Then we quantified the number of times a scaffold contained SCOs that from two separate chromosomes of *B. mori*, but from a single *P. napi* chromosome (napi-like scaffold), or vice versa (mori-like scaffold). Ratios > 1 indicate support for a *P. napi* like chromosomal structure (see note S11 for more details).