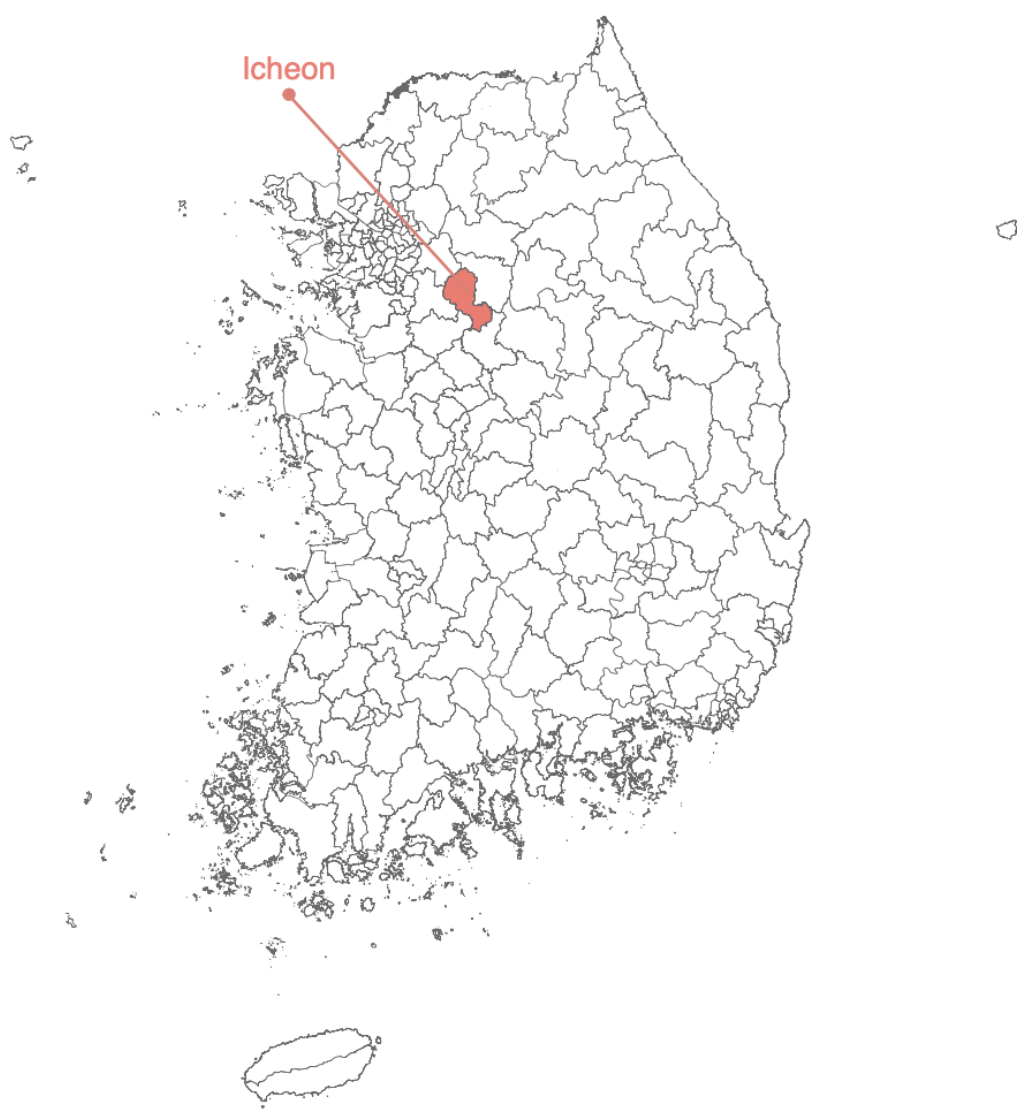
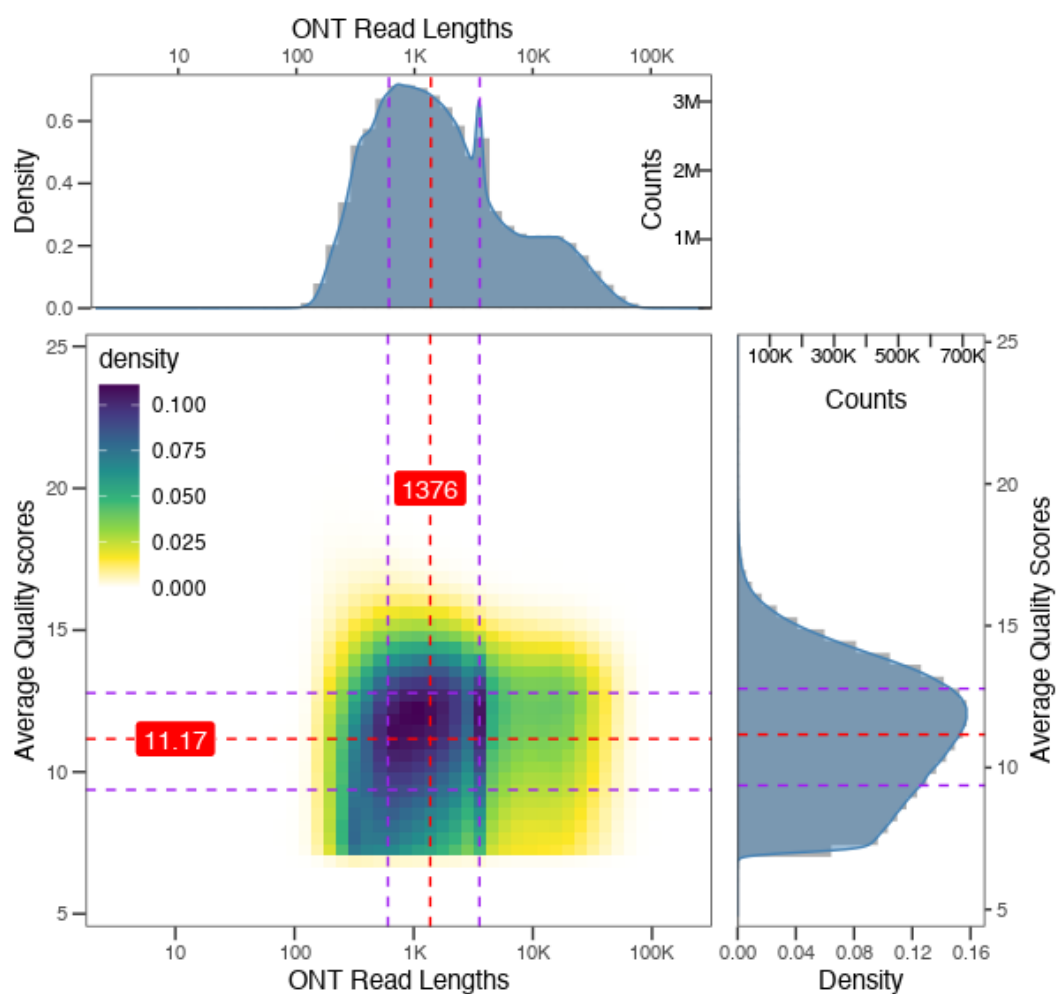


**Chromosome level de novo hybrid assembly of
Asian Honeybee, *Apis cerana* Koreana
Supplementary Figure & Table**

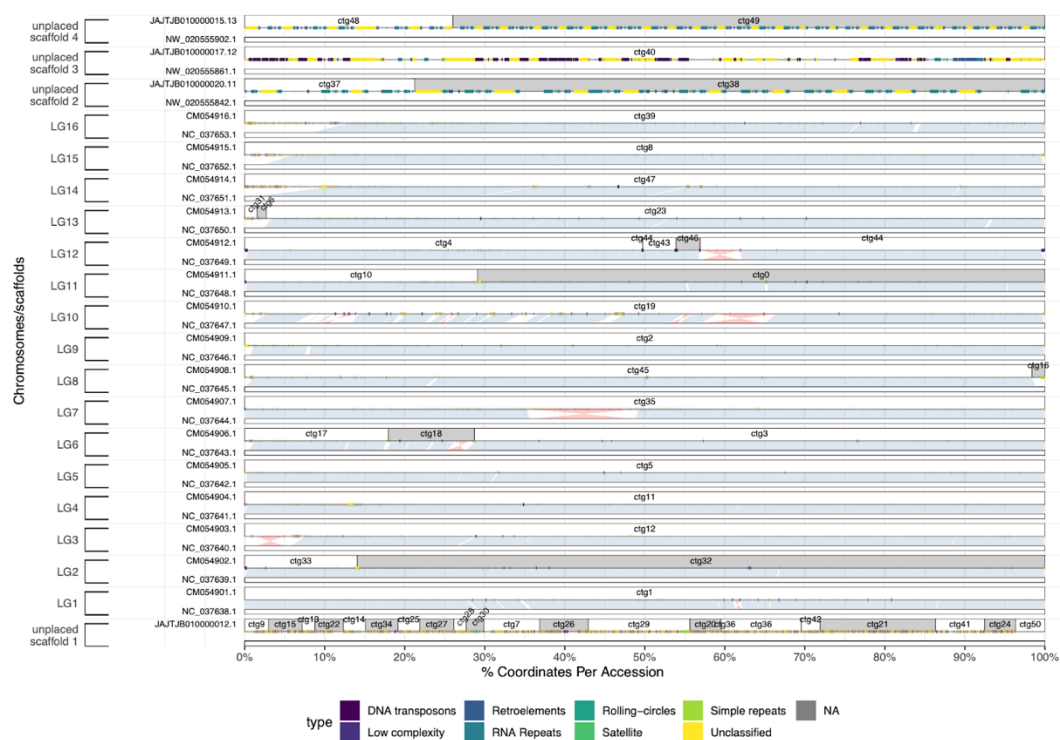
**Byung-Ju Kim¹, Jeong Hyeon Lee^{1,2}, Chi Hwan Kim³, Gyongju Nah³,
Olga Frunze¹, and Hyung Wook Kwon^{1,2,*}**



Supplementary Figure S1. Sample collection site. We collected *Apis cerana* from an apiary at Icheon-si, Gyeonggi-do 2019 for long-read sequencing using Oxford nanopore technology.



Supplementary Figure S2. Read length and average Read quality distribution of raw ONT reads. The processed data has 4,527,921 reads and a median length of 1376 bps (average: 4,324.20 bps). The median read quality is 11.17. X-axis was log-scaled for visualization purposes.



Supplementary Figure S3. Scaffolding pattern, repeat, and synteny pattern against *Apis mellifera* The scaffolding pattern, repeat element distribution, and synteny relationships of the *A. cerana* genome compared to the *A. mellifera* genome were illustrated. Chromosomes and unplaced scaffolds of *A. cerana* (AcerK1.0) are shown along the vertical axis, aligned with their syntenic regions in *A. mellifera*. Repeats are categorized by type, including DNA transposons (purple), RNA repeats (red), rolling-circle elements (yellow), satellites (green), and simple repeats (light green). Regions of low complexity and unclassified sequences are also depicted.

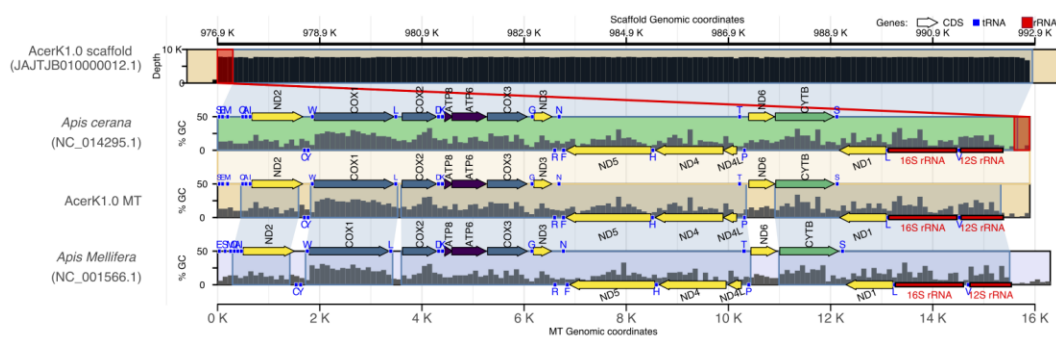
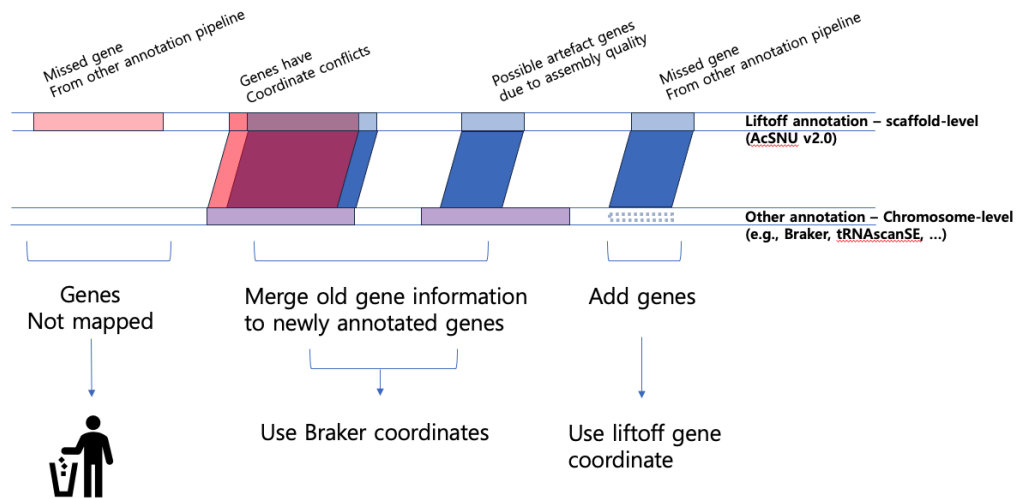
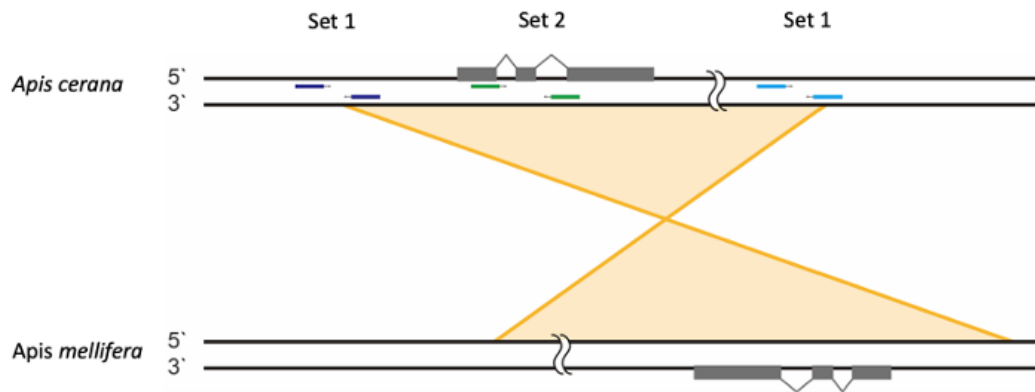


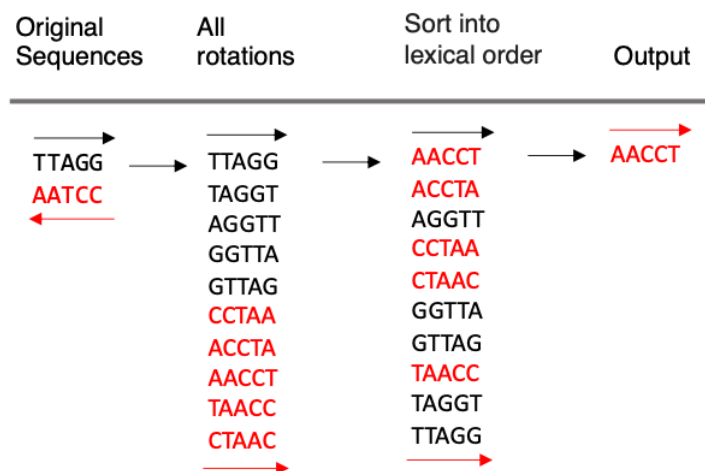
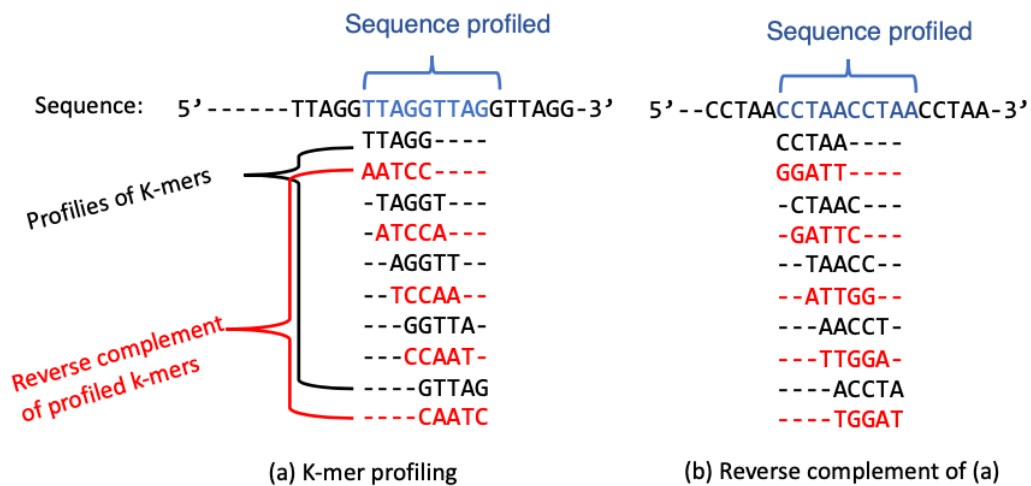
Figure S4. Mitochondrial extraction and comparison to *A. cerana* and *A. mellifera* The *A. cerana* reference genome (OR936096; 15,890 bp) was mapped to a scaffold containing multiple mitochondrial (MT) regions and repeat contamination, and one continuous block starting before the first base of the *A. cerana* reference MT (NC_014295.1) was identified. Read-depth and sequence identity to the reference were compared to resolve nucleotide conflicts between the frontal and backward blocks, and a full-length MT genome was constructed. The resulting genome includes 13 CDS, 22 tRNAs, and 2 rRNAs, with a high sequence identity (99.11%) to the *A. cerana* reference mtDNA (NC_014295.1; 15,895 bp).



Supplementary Figure S5. Strategy to merge gene annotation. Unmapped genes that could not be aligned to the genome were discarded to ensure annotation accuracy. Existing annotations were aligned and merged with newly annotated genes using Braker coordinates. Previously unannotated genes were incorporated into the final annotation using Liftoff gene coordinates.



Supplementary Figure S6 Primer design Primers were designed across key regions of interest, including syntenic blocks (Set 1) and structural variation junctions (Set 2). Set 1: Primers target conserved regions to confirm synteny between the two species. Set 2: Primers span inversion breakpoints or translocation junctions to verify the presence of structural variations. The orange lines indicate the alignment and corresponding structural relationships between the *A. cerana* and *A. mellifera* genomes.



(c) Burrow-wheeler + lexical sorting

Supplementary Figure S7 Strategy to count telomere motifs based on K-mer. (a-b) At a given anonymous sequence, the direction of the sequence is unknown. However, the profiled k-mers in both directions within the same range are expected to be identical. (c) The Burrows-Wheeler transform was applied to the reverse complement of the sequence, allowing circular repeats to be merged into a single repeat motif.

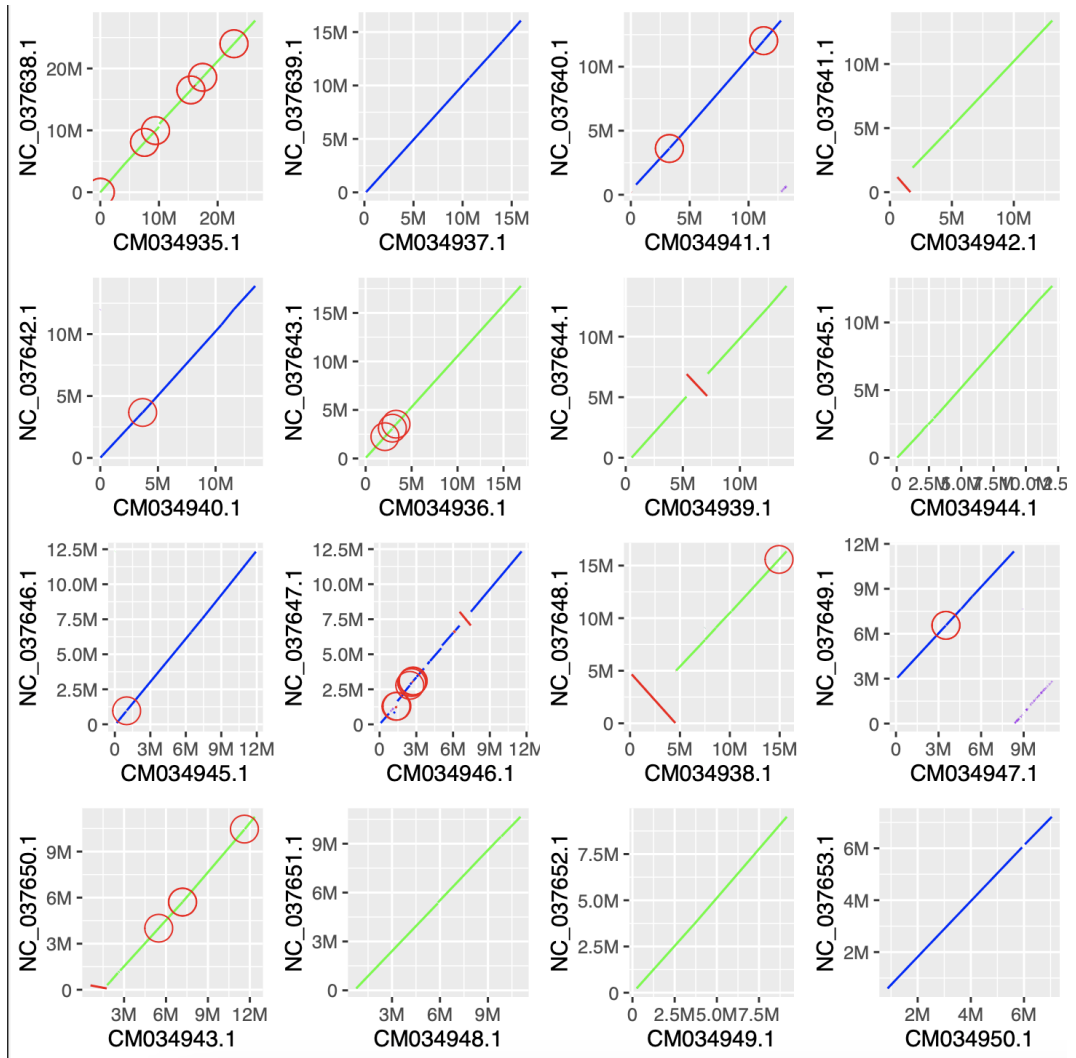
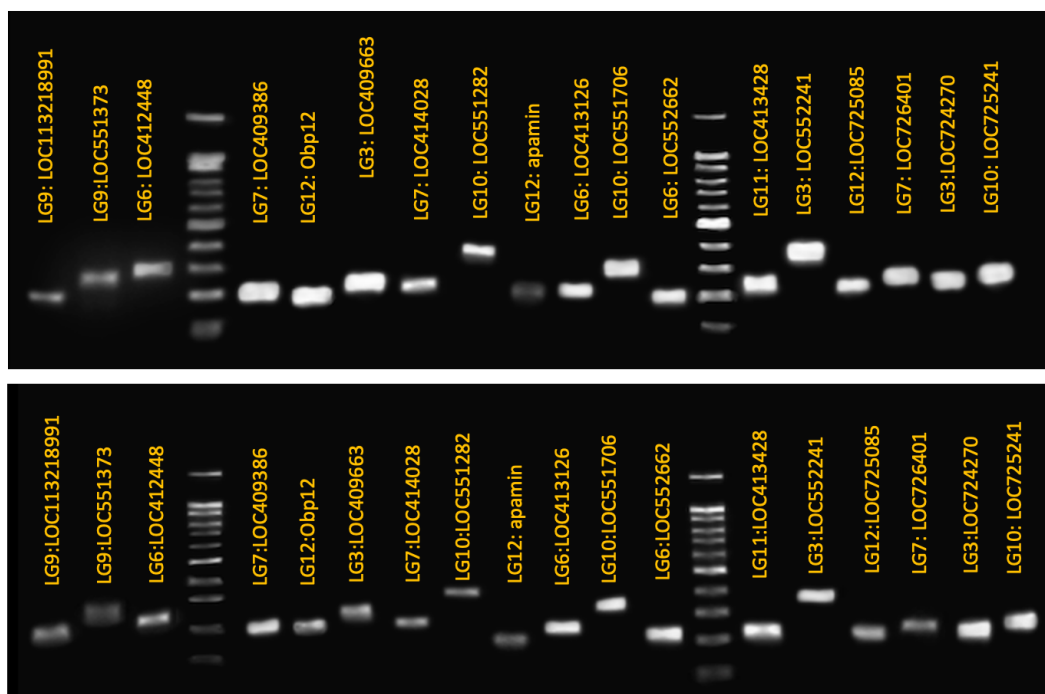
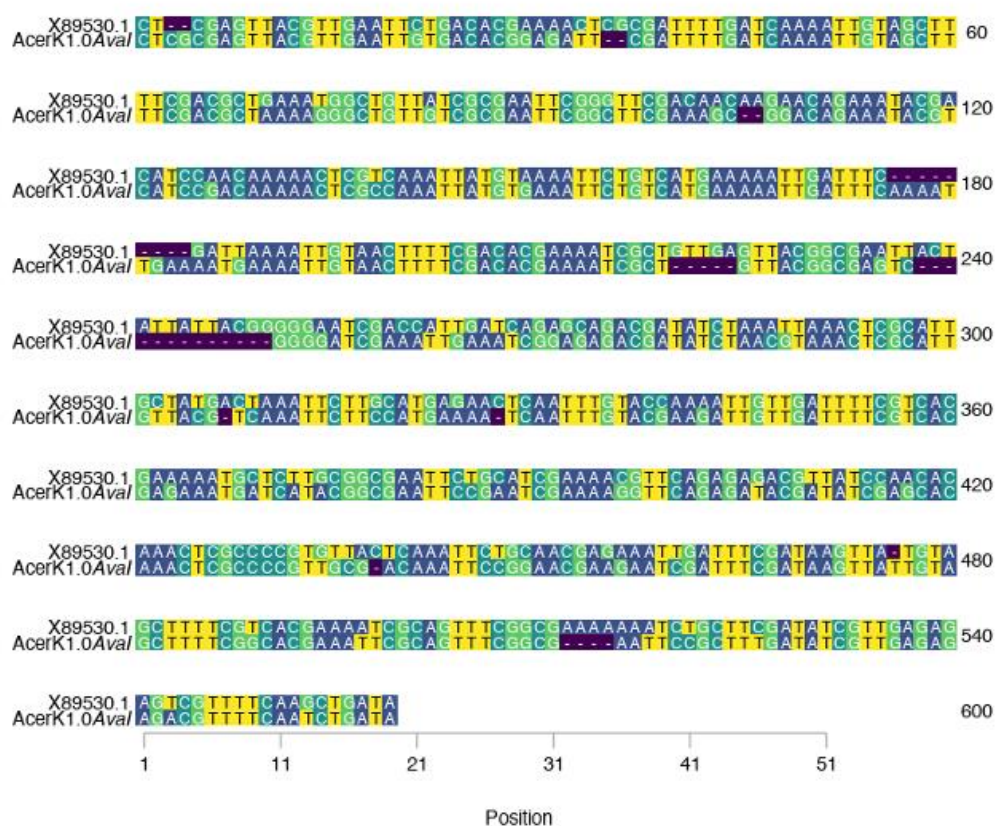


Figure S8. dotplot between *A. cerana* FLBSL-2018 and *A. mellifera* Amel_HAv3.1
 Diagonal blue and green lines represent regions of high collinearity and synteny, while red lines and disruptions indicate inversions, translocations, or other structural rearrangements. Circular markers highlight key structural differences, such as duplications or unaligned regions.



Supplementary Figure S9. Electrophoresis bands found within inverted regions between *A. cerana* and *A. mellifera* Bands were labeled with their respective chromosome (LG) and gene identifiers for clarity. The upper and lower panels represent different sets of inverted regions.



Supplementary Figure S10. *AcerK1.0* *AvaI* (530 bp) sequence was identified by Blast search and MAFFT alignment. The alignment of the *AcerK1.0* *AvaI* sequence (530 bp) was identified through a BLAST search and was aligned using MAFFT. Sequence conservation and variation across positions are highlighted with color-coded nucleotide bases for clarity, and key regions of similarity and divergence are indicated by numbered positions.

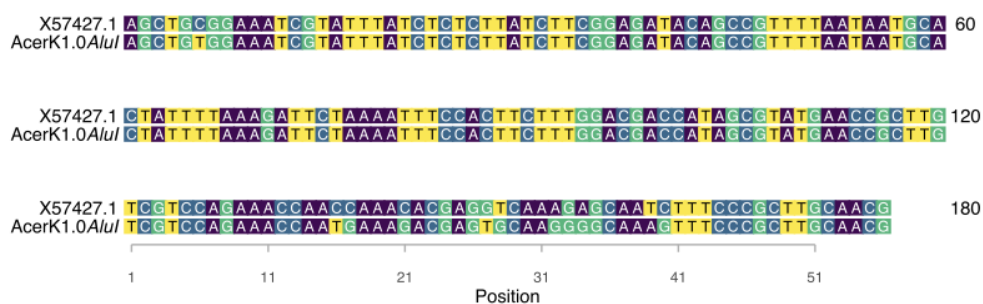
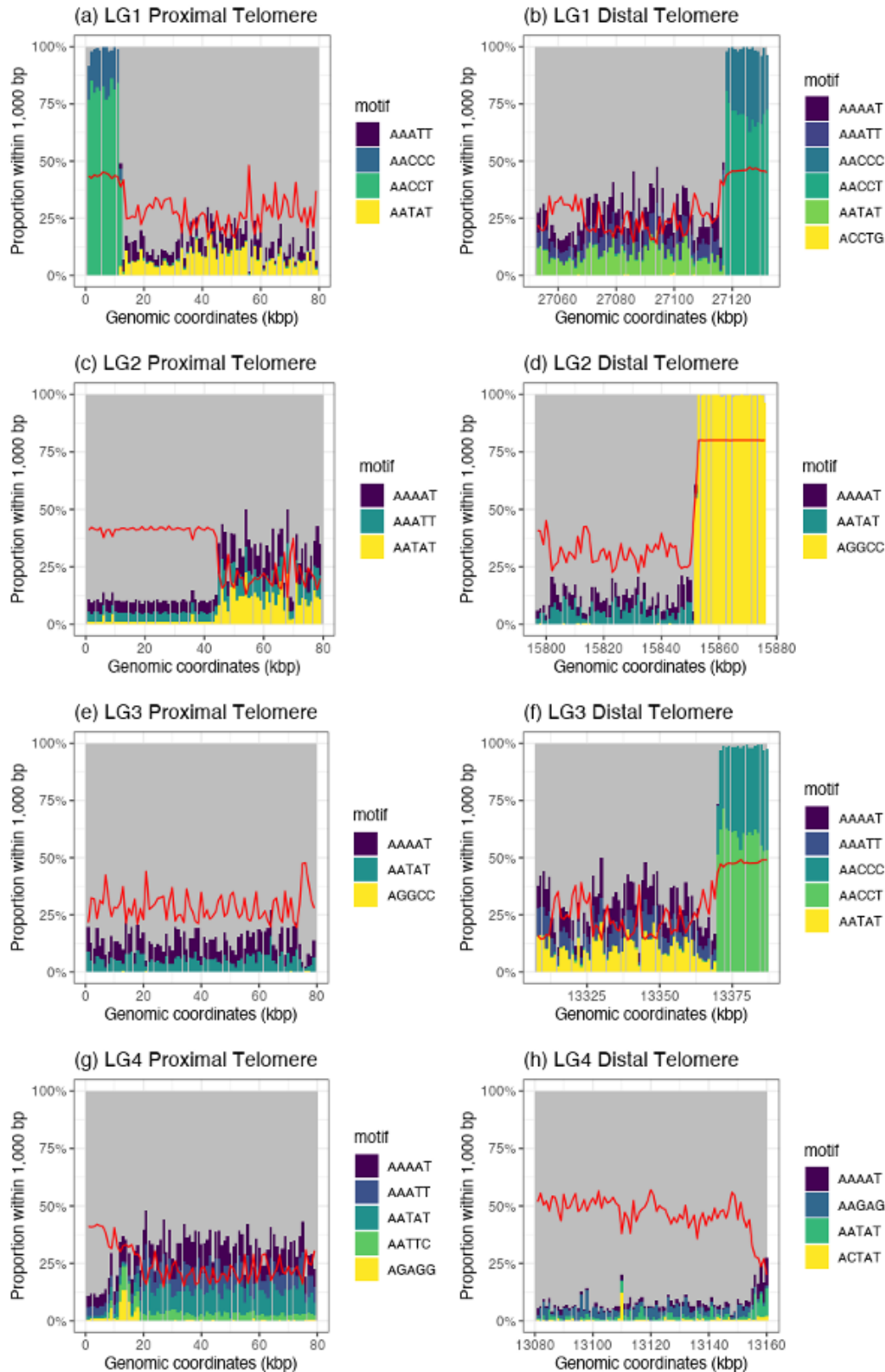
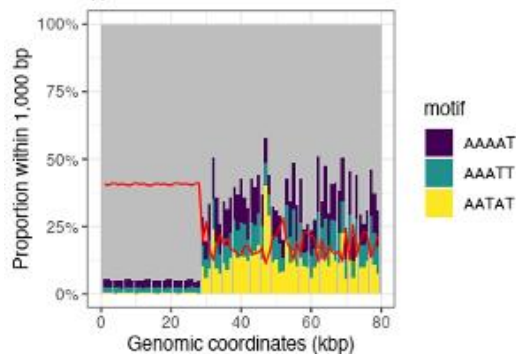


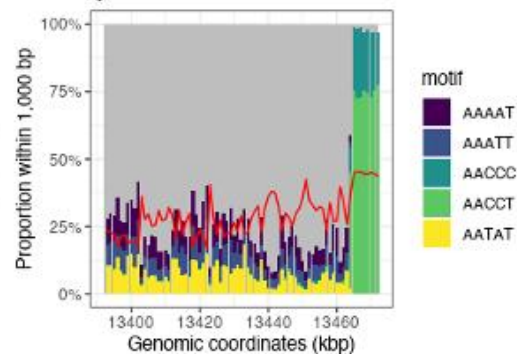
Figure S11. AcerK *AluI* (176 bp) sequence was identified by Blast search and MAFFT alignmnet The alignment of the *AcerK1.0* *AluI* sequence (176 bp) was identified using a BLAST search and was aligned with MAFFT. Conserved and variable nucleotide positions are highlighted with color-coded bases for clarity, and positions are numbered to assist in identifying specific regions of interest.



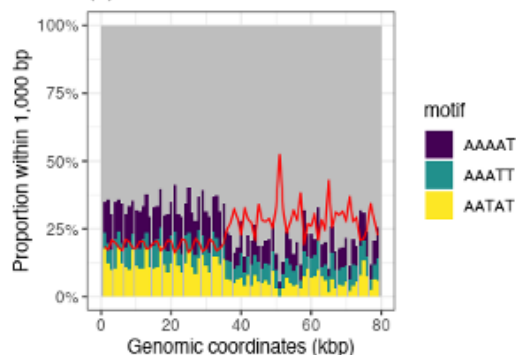
(i) LG5 Proximal Telomere



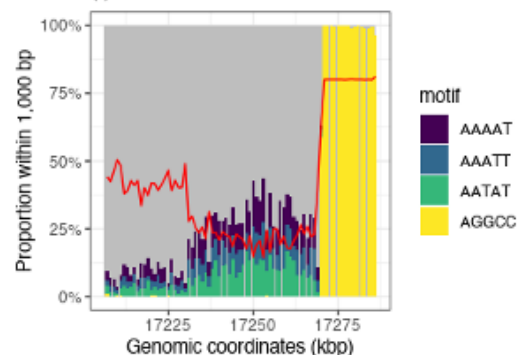
(j) LG5 Distal Telomere



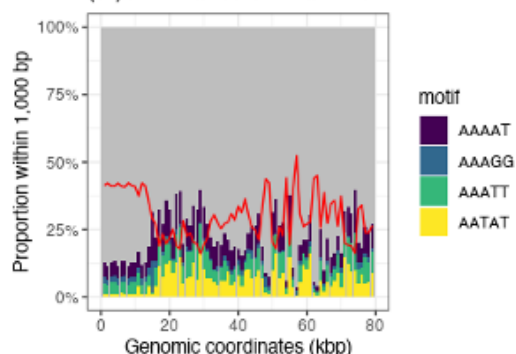
(k) LG6 Proximal Telomere



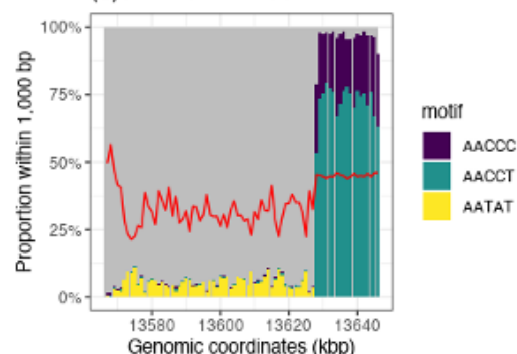
(l) LG6 Proximal Telomere



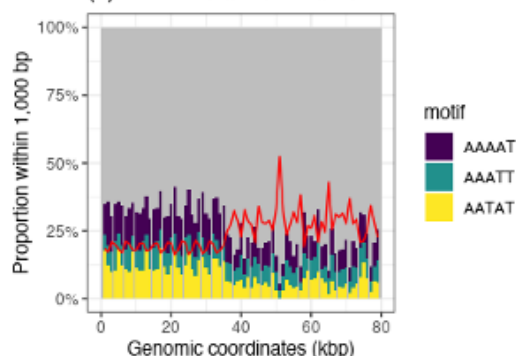
(m) LG7 Proximal Telomere



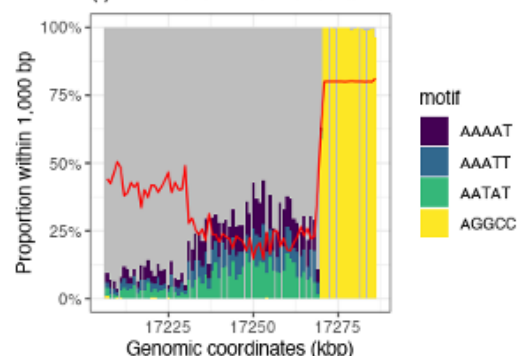
(n) LG7 distal Telomere



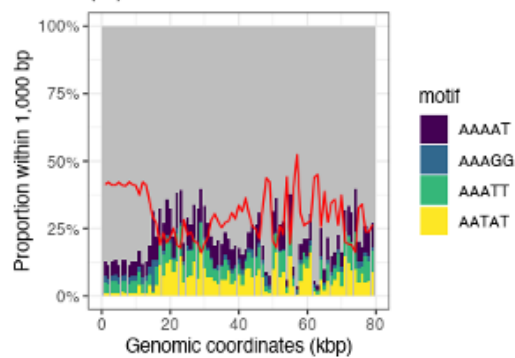
(k) LG6 Proximal Telomere



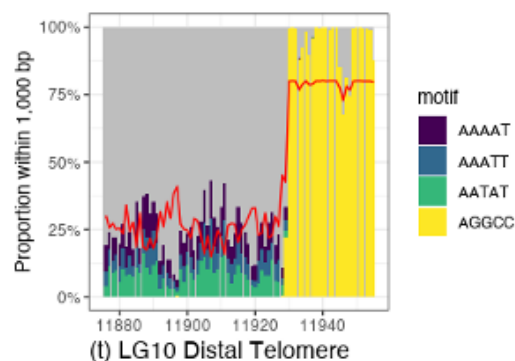
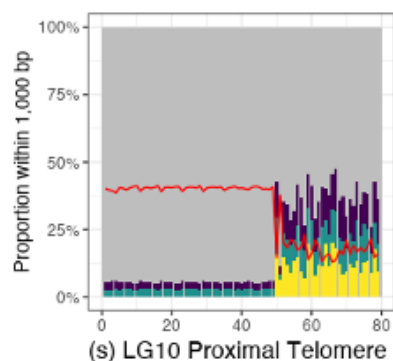
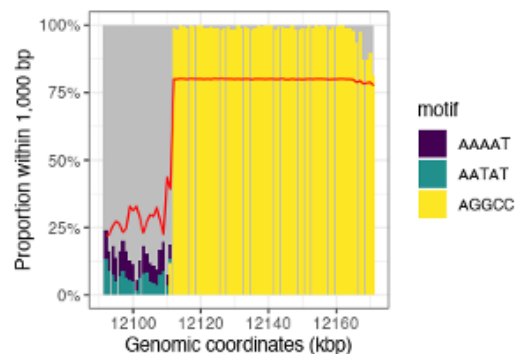
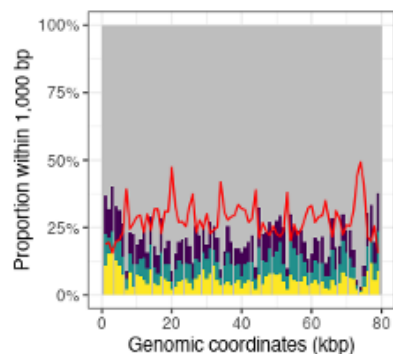
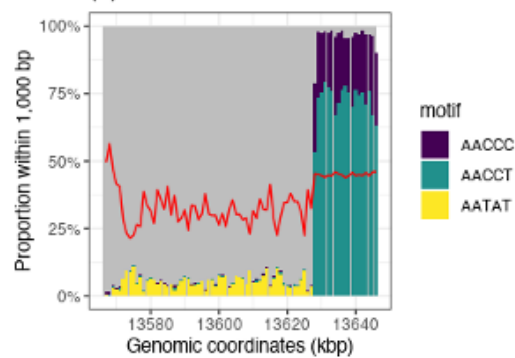
(l) LG6 Proximal Telomere



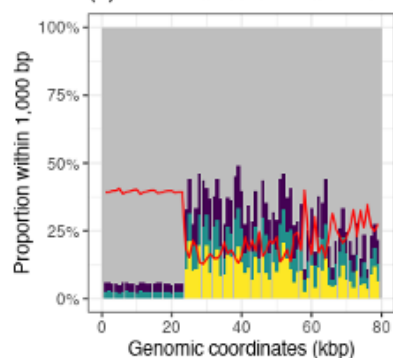
(m) LG7 Proximal Telomere



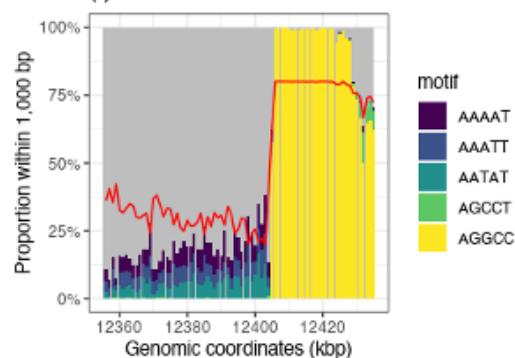
(n) LG7 distal Telomere

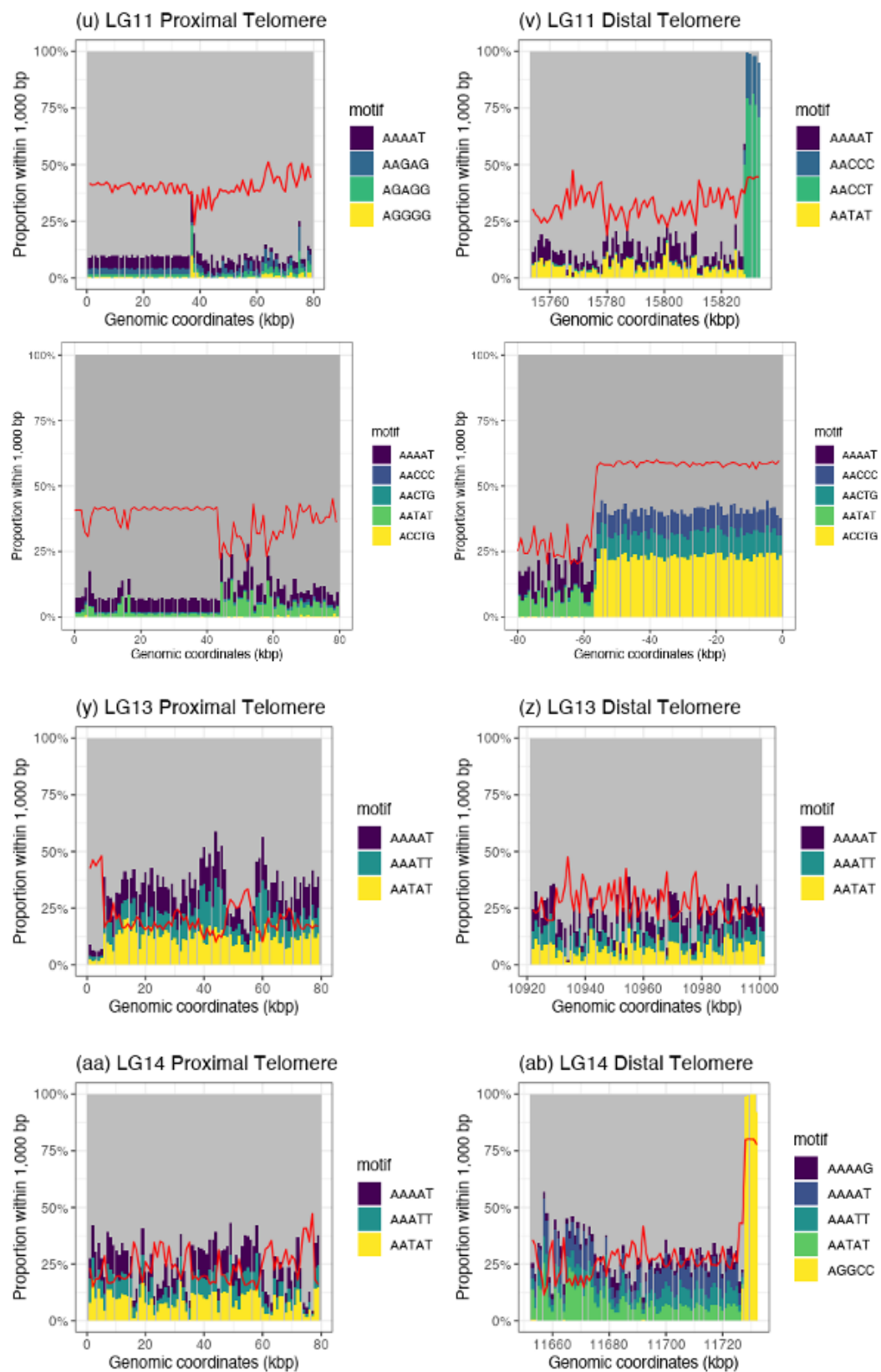


(s) LG10 Proximal Telomere

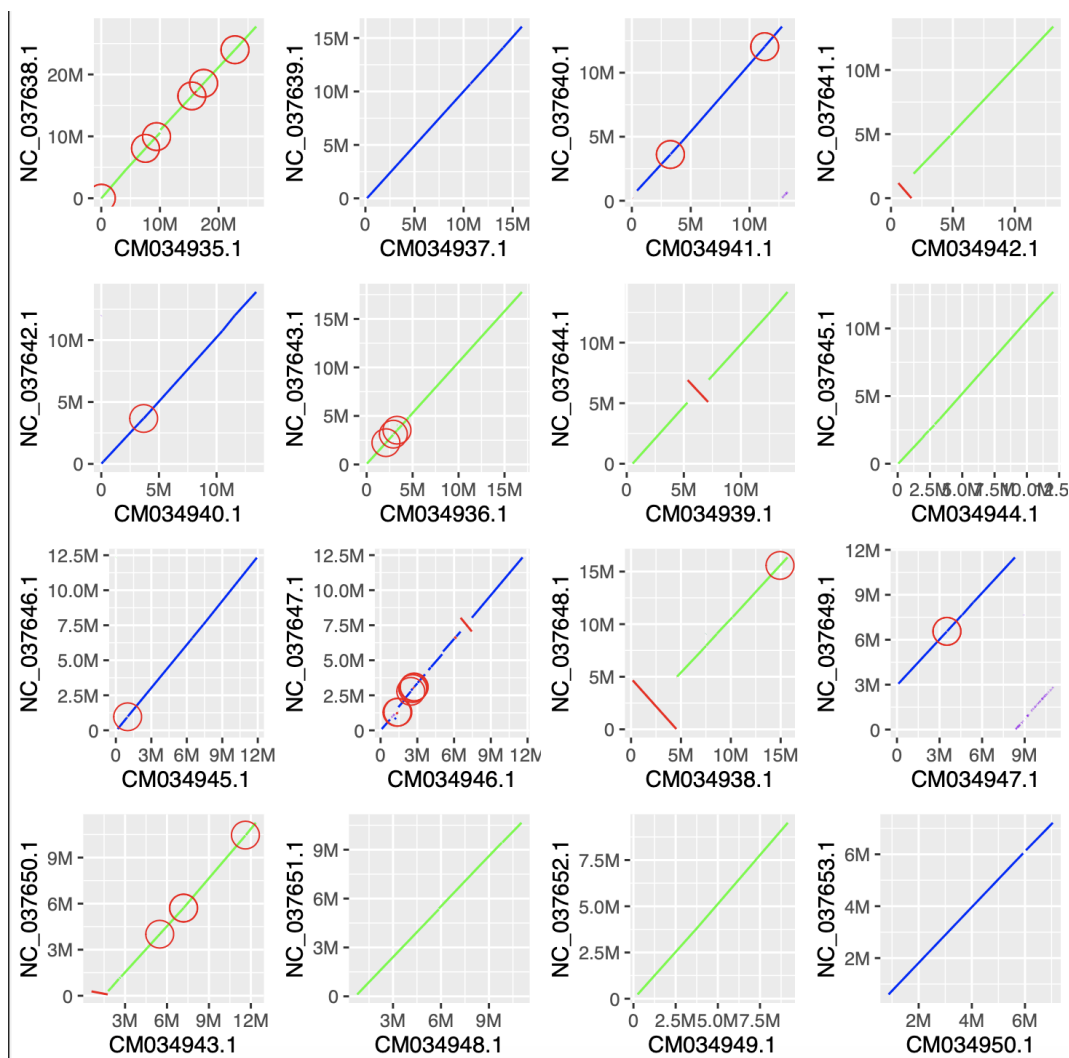


(t) LG10 Distal Telomere





Supplementary Figure S12. 5-mer profile distribution of Proximal-, Distal-telomere region of *A. cerana koreana*



Supplementary Figure S13. dotplot between *A. cerana* YURSC-2021 and *A. mellifera* Amel_HAv3.1 The proportion of 5-mer motifs within a 1000 bp window along genomic coordinates is represented in each panel. Key telomeric motifs, such as AAAAT, AAAT, and AATAT, are highlighted in color-coded legends. Variability in the enrichment of specific motifs between proximal and distal telomeres across linkage groups is demonstrated, with noticeable spikes in telomeric motifs observed at distal regions, suggesting conserved telomeric repeat patterns.

Supplementary Table S1. SRA data used in RNA mapping and annotation. *SRR01653580 is excluded in the analysis since it shows < 40% mapping ratio for all the tested assemblies.

Bioproject	N	Run
PRJNA161655	1	SRR536773
PRJNA383614	2	SRR5481633; SRR5481633
PRJNA235974	9	SRR1380970; SRR1380976; SRR1380979; SRR1380984
		SRR1388774; SRR1406762; SRR01653580*; SRR1653592
		SRR1653605
PRJNA277991	2	SRR1914350; SRR1914351
PRJNA339910	12	SRR4084079; SRR4084080; SRR4084081; SRR4084082
		SRR4084083; SRR4084084; SRR4084085; SRR4084086
		SRR4084087; SRR4084088; SRR4084089; SRR4084090
PRJNA431178	9	SRR6505206; SRR6505209; SRR6505211; SRR6505782
		SRR6505783; SRR6505784; SRR6505785; SRR6506058
		SRR6506157
PRJNA690015	22	SRR13367560; SRR13367561; SRR13367562; SRR13367563
		SRR13367564; SRR13367565; SRR13367566; SRR13367567
		SRR13367568; SRR13367569; SRR13367570; SRR13367571
		SRR13367572; SRR13367573; SRR13367574; SRR13367575
		SRR13367576; SRR13367577; SRR13367578; SRR13367579
		SRR13367580; SRR13367581
7 Project	57 Run	

Supplementary Table S2. Raw read statistics

(a) Basic raw sequence statistics

Category	Base Pair
Mean read length	4,324.20
Mean read quality	11.10
Median read length	1,376.00
Median read quality	11.20
Number of reads	4,527,921.00
Read length N50	15,127.00
STDEV read length	7,900.90
Total base	19,579,799,717.00

(b) Number, percentage and megabases of reads above quality cutoffs

Quality Cutoff	Number of reads	Percentage	Reads
>Q5	4527920	100.00%	19579.8Mb
>Q7	4526480	100.00%	19578.9Mb
>Q10	3025210	66.80%	13795.9Mb
>Q12	1675711	37.00%	7762.8Mb
>Q15	159428	3.50%	434.7Mb

(c) Top 5 highest mean basecall quality scores and their read lengths

Rank	mean basecall quality	read length
1	25.2	305
2	24.7	387
3	23.4	211
4	23.3	149
5	23.2	114

(d) Top 5 longest reads and their mean basecall quality score

Rank	read length	mean basecall quality
1	281,275	7.3
2	173,954	9.4
3	163,311	12.5
4	157,451	8.6
5	155,863	9.1

Supplementary Table S3. Scaffolding status and accession matching between Amel_HAv3.1 and AcerK1.0

*MT and contaminant reads were located in this scaffold

Linkage Group	Amel_HA v3.1	FLSBL-2018	Acer_K1.0	# Contigs to map chromosome
(Chromosome) based on Amel HAv3.1	Accession	Accession	Accession	
LG1	NC_037638.1	CM022002.1	CM054901.1	1
LG2	NC_037639.1	CM022011.1	CM054902.1	2
LG3	NC_037640.1	CM022016.1	CM054903.1	1
LG4	NC_037641.1	CM022015.1	CM054904.1	1
LG5	NC_037642.1	CM022014.1	CM054905.1	1
LG6	NC_037643.1	CM022010.1	CM054906.1	3
LG7	NC_037644.1	CM022013.1	CM054907.1	1
LG8	NC_037645.1	CM022017.1	CM054908.1	2
LG9	NC_037646.1	CM022003.1	CM054909.1	1
LG10	NC_037647.1	CM022004.1	CM054910.1	1
LG11	NC_037648.1	CM022012.1	CM054911.1	1
LG12	NC_037649.1	CM022006.1	CM054912.1	5
LG13	NC_037650.1	CM022007.1	CM054913.1	3
LG14	NC_037651.1	CM022005.1	CM054914.1	1
LG15	NC_037652.1	CM022008.1	CM054915.1	1
LG16	NC_037653.1	CM022009.1	CM054916.1	1
unmapped scaffold	NW_020555842.1	Not available	JAJTJB010000020.1	2
unmapped scaffold	NW_020555861.1	Not available	JAJTJB010000020.1	1
unmapped scaffold	NW_020555902.1	Not available	JAJTJB010000015.1	2
Not available (MT containing contigs + plant contamination)	Not available	Not available	JAJTJB010000012.1*	20

Table S4. Mitochondrial mapping report on JAJTJB010000012) Key alignment metrics include the start and end positions on both target and query sequences, target and query lengths, percentage identity, and strand orientation for each mapped region. High identity values (e.g., 98.36% and 99.11%) are observed in specific regions, indicating strong alignment between the target and query sequences. Notable regions with long query lengths (e.g., 15659 bp) highlight conserved mitochondrial regions. Strand orientation (+1 or -1) indicates alignment directionality, with mismatches primarily seen in inversions or partial alignments.

Start in TargetSeq	End in TargetSeq	Start in QuerySeq	End in QuerySeq	Target Length	Query Length	% identity	Target Strand	Query Strand	Target Seq	Query Seq
389610	389825	8074	8280	216	207	93.09	1	1	JAJTJB010000012.1	NC_014295.1
487209	487430	8280	8074	222	207	90.58	1	-1	JAJTJB010000012.1	NC_014295.1
976690	976993	15596	15895	304	300	98.36	1	1	JAJTJB010000012.1	NC_014295.1
976994	992644	1	15659	15651	15659	99.11	1	1	JAJTJB010000012.1	NC_014295.1
2205926	2206147	8074	8280	222	207	90.58	1	1	JAJTJB010000012.1	NC_014295.1
2247946	2248170	11157	10932	225	226	90.27	1	-1	JAJTJB010000012.1	NC_014295.1
2339665	2339889	10932	11157	225	226	90.27	1	1	JAJTJB010000012.1	NC_014295.1
2645335	2645552	8280	8074	218	207	92.24	1	-1	JAJTJB010000012.1	NC_014295.1
2686760	2687035	10931	11206	276	276	90.94	1	1	JAJTJB010000012.1	NC_014295.1
3076367	3076591	11157	10932	225	226	90.27	1	-1	JAJTJB010000012.1	NC_014295.1
3159549	3159773	10932	11157	225	226	90.27	1	1	JAJTJB010000012.1	NC_014295.1

Supplementary Table S5. MT genomes used in the phylogenetic study. The mitochondrial genomes included in the phylogenetic analysis of *A. cerana* and *A. mellifera*, detailing GenBank accession numbers, geographic origins, genome lengths, and specific annotations were summarized.

Genbank Accession	Country/location	length (bp)	note
OR936096	Korea	15,890	<i>Acer</i> K1.0 mtDNA
AP017314.1	Japan	15,917	
AP017941.1	Japan/Amami-Oshimalsl and	15,788	tRNA ^{Ser} starts at 327 bp.
AP017983.2	China	15,460	
AP017984.2	Taiwan	15,376	
AP018149.1	Sarah-Borneo	15,884	tRNA ^{Asn} is missing
AP018431.1	Korea	15,925	
AP018450.1	Russia	15,919	
KM244704.1	China	15,712	tRNA ^{Ser} starts at 262 bp.
KX908206.1	Korea	15,904	
LC640345.1	Russia	15,718	tRNA ^{Ser} starts at 261 bp.
LC640346.1	Russia	15,756	tRNA ^{Ser} starts at 260 bp.
			tRNA ^{Glu} was duplicated.
LC640347.1	Russia	15,612	tRNA ^{Ser} starts at 238 bp.
LC640348.1	Russia	15,668	tRNA ^{Ser} starts at 238 bp.
LC640349.1	Russia	15,651	tRNA ^{Ser} starts at 261 bp.
LC640350.1	Russia	15,663	tRNA ^{Ser} starts at 240 bp.
LC640351.1	Russia	15,778	tRNA ^{Ser} starts at 359 bp.
MW309837.1	Korea	15,907	
GQ162109.1	China	15,895	Reference genome(NC_014925.1)
NC_001566.1	Italia	16343	<i>Apis mellifera</i> MT Reference.

Supplementary Table S6. Primer Design

(a) Primer designs for Inverted region.

LG	NCBI_Symbol (Amel)	Gene description	target stran d	F/R	Sequence	mer	Tm	GC%	Start/ End	Size	Tm ave rage
LG9	LOC113218991	mitochondrial 2-oxodicarboxylate carrier-like	+	F	GGTGCAATGCAATCTAATCG	21	57.4	42.86	6524	191	55.45
				R	TTTAACCAAGGAATGATGGAA	21	53.5	33.33	6715		
LG9	LOC551373	ectonucleoside triphosphate diphosphohydrolase 5	-	F	GTTCACAATCAGAATGCA	20	56.4	45	5174	273	56.05
				R	TATTAAATGCAACAAAGACATGC	23	55.7	30.43	5447		
LG6	LOC412448	zinc transporter ZIP1	-	F	TTTCTGAATTGCAGGGATCG	20	56.4	45	992	297	56.4
				R	GCGCCAATAAAAGTCCCATATA	20	56.4	45	1289		
LG7	LOC409386	ubiquitin-conjugating enzyme E2 N	+	F	TAGATTGGGCAGGATTTGCT	20	56.4	45	1790	191	56.4
				R	CGAGTCCATTCTTTGGCATT	20	56.4	45	1981		
LG12	Obp12	obp12	-	F	CGTCGGAGGCCAAGATAAAA	20	56.4	45	592	171	56.4
				R	CTCAATTTCGGAGCCATTTCAT	20	56.4	45	763		
LG3	LOC409663	lysozyme	+	F	ATCTGTTTGGGACCATTTC	20	56.4	45	1405	209	57.4
				R	GAGCTGCACAATAAGGATCG	20	58.4	50	1614		
LG7	LOC414028	sorting nexin-4	+	F	TTGCCGCTTCGTAGATTTT	20	54.3	40	65	186	57.4
				R	GTACGATGAGCGTCGAGTGA	20	60.5	55	251		
LG10	LOC551282	ribokinase	+	F	TCGACAGTGGTAAGTAAGCTTT	23	59.3	39.13	896	358	57.85
				R	TTTTCATAGCAGGTGCTCCA	20	56.4	45	1254		
LG12	apamin	apamin	+	F	TTTGTGACACAGTAATGGTA	23	57.5	34.78	89	200	57.95
				R	CGCAAAGTGCAGTTTCTGGT	20	58.4	50	289		
LG6	LOC413126	serine/threonine-protein kinase Tao	+	F	CATGGTAGTTTTGGGGCTGT	20	58.4	50	4771	207	58.4
				R	AAATTACCCAGGCAGTGTGG	20	58.4	50	4978		
LG10	LOC551706	unconventional myosin-IXb	+	F	GGACATCCCGAGAAAGATGA	20	58.4	50	3697	285	58.4
				R	AATGAGAAGATCGGCGTGAC	20	58.4	50	3982		
LG6	LOC552662	non-canonical poly(A) RNA polymerase PAPD5-like	-	F	TGCTGTTGTTCAATGTCGT	20	56.4	45	13	191	58.45
				R	GAAGTCTCGGTTGATACC	20	60.5	55	204		
LG11	LOC413428	uncharacterized LOC413428	+	F	GGGGTTAAGGAATTGGAA	20	56.4	45	6331	227	58.45
				R	GCGGGTAATCTGAACAGAGC	20	60.5	55	6558		
LG3	LOC552241	60S ribosomal protein	+	F	TACTCAGCTGCTTTGATGC	20	58.4	50	1346	203	59.75
				R	CACCAGTTACTTCAGTTGTTCCA	23	61.1	43.48	1549		
LG12	LOC725085	basic helix-loop-helix neural transcription factor TAP	-	F	CAGTGCGCAACGTGTTATTT	20	56.4	45	3625	341	58.45
				R	GATAAGAGCACTGGCGGAGA	20	60.5	55	3966		
LG7	LOC726401	ras-specific guanine nucleotide-releasing factor 2	+	F	GAACATGTCCCAAGGCACTG	21	61.3	52.38	126529	235	58.85
				R	TCGTTTACGCATGCTTTCTG	20	56.4	45	126764		
LG3	LOC724270	Crustacean hyperglycemic hormones	+	F	GCGGGAGATCGTTAGTTTTG	20	58.4	50	14193	208	59.45
				R	AAGTGGTGGGATAGGCACAG	20	60.5	55	14401		
LG10	LOC725241	uncharacterized LOC725241	+	F	ATCTTGGCCAAAAGCTGAAG	20	56.4	45	54930	215	59.45
				R	GGCAGACAGACTGGCCTTAG	20	62.5	60	55145		

(b) Primer designs for Inversion boundary

LG	F/R	Primer Sequence	Tm	Starts in	PCR block size	Tm Average
LG3	F	AAAATCGATCTTCAAACATATACCA	58.4	658805	519	58.4
	R	CTAAATATATTATGTTTATCAGGGTG	58.4	659324		
LG5-1	F	CCTTGGAGAGTTAAATCGTACG	60.3	3676358	536	59.35
	R	CTTGACCGAAGGCAGCTAAA	58.4	3676894		
LG5-2	F	TGGAATGGATTTCGGCTTTAC	56.4	3679164	524	58.35
	R	CATGTATGGCAGAAAAGCACTG	60.3	3679688		
LG0	F	CTGAAAGATAATTTTGAAGATTCCC	60	971129	382	59.2
	R	TTTGTGCTGGATTCTTAACCTGG	58.4	971511		
LG16	F	ATGGCGTTCCATTATCGTTC	56.4	5474287	527	57.4
	R	AAAGAAACCTCTAGGTCCCA	58.4	5474814		

Supplementary Table S7. RNA-seq mapping statistics for 5 assembly

Author/Year	min	Q25	med	Q75	max	avg	sd
Park/2015	37.4	86.95	94.35	97.85	99.2	87.28	18.11
Daio/2018	82.5	94	97.85	99.26	99.7	96.07	4.26
Yakoi/2018	83.3	96.36	99	99.33	99.7	97.34	3.30
Wang/2020	83.3	96.48	99.15	99.5	99.9	97.42	3.38
KIM/2024	83.4	96.48	99.2	99.5	99.9	97.58	3.28

Supplementary Table S8. *AvaI* and *AluI* repeats were found by blast search.

(a) AcerK1.0

Seqs	AcerK1.0 Accession	AluI	AvaI
Unplace	JAJTJB010000012.1	19	24
LG2	CM054902.1	247	133
LG3	CM054903.1	1	14
LG4	CM054904.1	52	0
LG5	CM054905.1	0	51
LG6	CM054906.1	0	2
LG7	CM054907.1	77	0
LG9	CM054909.1	0	8
LG10	CM054910.1	0	43
LG11	CM054911.1	195	179
LG12	CM054912.1	238	0
LG14	CM054914.1	11	172
LG16	CM054916.1	0	1
Unplace	NW_020555861.1	2	0
total		842	627

(b) FLSBL-2018 GCA_011100585.1

Seqs	AcerK1.0 Accession	AluI	AvaI
LG14	CM022005.1	0	22
LG12	CM022006.1	70	0
LG7	CM022013.1	8	0
LG4	CM022015.1	83	0
LG8	CM022016.1	0	10
Unplaced	WJBF01000105.1	0	12
Unplaced	WJBF01000114.1	0	9
Unplaced	WJBF01000084.1	0	1
Unplaced	WJBF01000036.1	22	0
Unplaced	WJBF01000021.1	5	0
Total		188	54