The shared genomic architecture of human nucleolar organizer regions

Ioanna Floutsakou[1,4], Saumya Agrawal[2,4], Thong T. Nguyen[3,4], Cathal Seoighe[3], Austen R.D. Ganley[2], and Brian McStay[1,5]

[1] *Centre for Chromosome Biology, School of Natural Sciences, National University of Ireland, Galway, Galway, Ireland.* [2] *Institute of Natural and Mathematical Sciences, Massey University, Auckland, New Zealand.* [3] *School of Mathematics, Statistics and Applied Mathematics, National University of Ireland, Galway, Galway, Ireland.* [4] These authors contributed equally to this work. [5]Corresponding author: E-mail brian.mcstay@nuigalway.ie

SUPPLEMENTAL INFORMATION

Supplemental Methods (page 3)

Supplemental Tables (page 12)

Table 1. Putative gene models from the DJ
Table 2. Putative gene models from the PJ
Table 3. Segmentally duplicated regions from the DJ contig
Table 4. Segmentally duplicated regions from the PJ contig
Table 5. Fold increase of DJ sequences present in BAC array clones
Table 6. Sequencing statistics for the DJ and PJ cosmids
Table 7. Assembly statistics for the DJ and PJ cosmid assemblies
Table 8. DJ and PJ primers for mapping, ChIP, and RT-PCR

Supplemental Figures

Fig. 1. Positions of cosmids and BACs in the DJ
Fig. 2. Positions of cosmids and BACs in the PJ
Fig. 3. PJ sequences are conserved among the acrocentric chromosomes
Fig. 4. Structure of DJ and PJ cosmids
Fig. 5. Workflow for junction verification mapping pipeline
Fig. 6. Alu and 147bp ACRO1 repeat inter-chromosomal PJ variation
Fig. 7. Construction scheme for the DJ and PJ contigs
Fig. 8. Repeat composition of the DJ and PJ regions.
Fig. 9. 48bp CER repeats are found distal to the NOR on all five acrocentric chromosomes.
Fig. 10. HMM Logo for the DJ ACRO138 repeats
Fig. 11. Similarity plot of the DJ inverted repeat arms.
Fig. 12. Positions of sequence features in the DJ and PJ
Fig. 13. Comparison of PJ and DJ positioning in human cells.
Fig. 14. UBF and rDNA colocalise in nucleolar caps upon AMD treatment.
Fig. 15. Construction of DJ BAC arrays on metacentric chromosomes.

Fig. 16. Heat map of the association between chromatin marks and chromatin states
Fig. 17. H3K4me3 peaks overlap with open chromatin domains identified by DNaseI-HS-seq and FAIRE-seq
Fig. 18. Occurrence of CTCF DNA motifs within the DJ
Fig. 19. Transcription profiling of the DJ
Fig. 20. Quantitation of DJ transcript levels.
Fig. 21. RTPCR confirms the presence of a transcript from the DJ 138 kb promoter

Supplemental Spreadsheet and Video Descriptions

Data File 1. DJ and PJ intra- and inter-chromosomal sequence comparison matrices
Data File 2. DJ sequence FASTA file
Data File 3. DJ sequence feature file (GFF3)
Videos 1 and 2. 3D arrangement of the DJ in the nucleus

Supplemental Methods

BACs

BACs spanning the DJ; RP11-337M7 (AL592188, chr22), CH507-535F5 (CT476834, chr21), CH507-145C22 (CU633906, chr21), RP11-272E10 (AC011841) and those spanning the PJ; bP-2171c21 (CR392039, chr21), bP-2154M18 (CR381535, chr21), were obtained from the BACPAC Resource Centre (Children's Hospital Oakland). Q-arm BACs; RP11-420H1 (AC018739, chr13), RP11-115A12 (AC010766, chr14), RP11-32B5 (AC068446, chr15), RPCI-11-89H21 (AZ517782, chr21) and RP11-278E23 (AC013360, chr22) were also obtained from the BACPAC Resource Centre.

Cosmid sequencing

Four PJ cosmids and two DJ cosmids (LA13 165F6, LA14 101B3, LA15 25H3, N 29M24, LA13 133H12, and LA15 64C10) were sequenced using Illumina sequencing, and the sequencing statistics are given in Supplemental Table 6. Sequences for the cosmids were de-multiplexed and assembled individually, using the assembly parameters below. The assembly statistics are presented in Supplemental Table 7.

Parameters for *de novo* assembly using ABySS

Different values for parameter k-mer size (-k) and minimum mean k-mer coverage of a unitig (-c) were used for ABySS to obtain the optimum assembly (Supplemental Fig. 4) for each cosmid as follows:

LA13 165F6: ABYSS -k 35 -c 88 -e 2 -o LA13_165F6_assembly.fasta <LA13_165F6 NGS_data.fastq>

LA14 101B3: ABYSS -k 35 -c 99 -e 2 -o LA14_101B3_assembly.fasta <LA14_101B3 NGS_data.fastq>

LA15 25H3: ABYSS -k 35 -c 75 -e 2 -o LA15_25H3_assembly.fasta <LA15_25H3 NGS_data.fastq>

N 29M24: ABYSS -k 35 -c 114 -e 2 -o N29M24_assembly.fasta <N29M24 NGS_data.fastq>

LA13 133H12: ABYSS -k 35 -c 100 -e 2 -o LA13_133H12_assemlby.fasta <LA13_133H12 NGS_data.fastq>

LA15 64C10: ABYSS -k 35 -c 200 -e 2 -o LA15_64C10_assembly.fasta <LA15_64C10 NGS_data.fastq>

Bioinformatic screen for proximal rDNA junctions

We developed a five step mapping method to identify putative rDNA-PJ junctions (Supplemental Fig. 5). The basis of this approach is that whole genome shotgun sequencing data should contain reads that cross the rDNA-PJ junction and have one part from the rDNA and the other part from the PJ.

Data Acquisition: Human WGS data from The Center for Applied Genomics (CRA) was downloaded from the Ensemble trace archive (currently available through the NCBI trace archive using the query: species_code="HOMO SAPIENS"AND CENTER_NAME = "CRA" AND STRATEGY = "WGA"). All 27,449,655 reads were used for the analysis.

Reference sequence preparation. Our method searches for junctions between the rDNA (bases 122,273-167,354 of BAC AL353644 were taken as the rDNA reference) and a given reference sequence. To search for PJ junctions, all four PJ cosmids and the PJ BAC clone CR392039 (with the rDNA sequence trimmed off) were used as reference sequences. The DJ contig was used as the reference sequence for the DJ region (positive control).

Pipeline: The steps described below were repeated for each flanking reference sequence individually (see also Supplemental Fig. 5).

Step 1: Reads were mapped to the flanking reference sequence using gsMapper v2.3 (454 Roche) using 95% identity and 100 bp minimum alignment length as the cut offs. All reads were considered as single end reads during the mapping. Reads with “Partial” mapping status from the “454ReadStatus.txt” output file were selected from the mapped reads for the next step (the term “Partial” is used in gsMapper to denote reads that only a part maps to the reference sequence).

Step 2: All reads obtained from step 1 were then mapped to the human rDNA. Reads that partially mapped to the rDNA with 95% identity and 100 bp minimum alignment length were selected.

Step 3: Reads that partially mapped to both the flanking reference and the rDNA (in steps 2 and 3) were clustered together according to their position in the flanking region. The flanking regions were divided into partially overlapping bins, and the reads were placed into these bins.

Step 4: Reads were removed if they fell in one of the following categories:

a) Same region of the read is partially mapping to both the flanking and the rDNA sequences.

b) Entire read does not match the reference sequences on both sides of the potential junction point.

c) Entire read also matches other regions of the human genome.

d) Only one read represents the cluster.

Step 5: To check that the potential junction sequences found are not restricted to one sequencing center, the cluster sequences that remained from step 4 were used to search complete human whole genome sequencing data present in the NCBI trace database. This includes all reads submitted by the different sequencing centers, including the CRA dataset. All hits found across the junction of the cluster sequences were manually checked for matches ≥100 bp to both the rDNA and the flanking region. Reads that did not match ≥100 bp were end-trimmed using quality scores (as BLAST does not take quality score into account) and BLAST performed again to the respective reference sequences.

The mapping results are shown below. The mapping method predicted two potential PJ junctions, one in the 18S and other in ITS-1. These PJ junctions are the same as those identified from the BAC and cosmid sequences, and no evidence for additional junction regions was found. The same procedure was repeated for the DJ to check the accuracy of the method. The DJ results predict a single junction, the same as that found in the BAC and cosmid sequences.

Results for different steps of the junction mapping pipeline.

Flanking reference sequence	**Reference coordinates used for mapping**	**# reads mapped to flanking sequence (Step 1)**	**# reads mapped to the rDNA (Step 2)**	**# of clusters (Step 3)**	**# of clusters after filter (Steps 4 & 5)**
LA13 165F6 (13)	1 - 20,193	325,570	417	6	1
LA14 101B3 (14)	1 - 27,432	318,834	1,307	12	2
LA15 25H3 (15)	1 - 35,460	332,822	746	15	1
N 29M24 (22)	1 - 32,522	448,750	45,556	10	1
CR392039 (21)	1 - 155,929	828,792	102,938	17	1
Distal contig	1 - 379,046	554,888	19,023	8	1

PCR

PCR primers (Supplemental Table 8) were designed with the aid of RepeatMasker (Smit) to filter out repetitive sequences. Primers were then selected from the region of interest using

Primer Blast (http://www.ncbi.nlm.nih.gov/tools/primer-blast/). In DJ and PJ mapping experiments 50ng of genomic DNA prepared from somatic cell hybrids was used as template. For the transcript analysis by PCR, cDNA was prepared from total cellular or nucleolar RNA samples with a Protoscript first strand cDNA synthesis kit (New England Biolabs) using either random hexamer or oligo-dT primers.

Inter-chromosomal identity of DJ and PJ regions

To determine DJ interchromosomal homogeneity, pairwise comparisons of all interchromosomal cosmid and BAC sequences were generated (Supplemental Data File 1) using the Stretcher global alignment tool from EMBOSS (Rice et al. 2000) and YASS (Noe and Kucherov 2005). However, because most of these clones derive from chr21, we chose a representative clone for each of the five acrocentric chromosomes to assess sequence conservation: LA13 133H12 (chr13), LA14 138-F10 (chr14), LA15 64C10 (chr15), (CT476837 (chr21), and AL353644 (chr22). These sequences show an average of 99.1% identity to each other (Fig. 2B). This analysis was biased towards the rDNA end of the DJ contig. To determine whether sequence conservation is reduced further away from the rDNA, the sequence identities between the misannotated BAC clone AC011841 and DJ BAC clones from chr21 were calculated. The average level of identity was 98.5%, indicating that DJ conservation is not restricted to the rDNA junction but extends towards the telomere.

To investigate PJ interchromosomal conservation four cosmids [LA13 165F6 (chr13), LA14 101B3 (chr14), LA15 25H3 (chr15), and N 29M24 (chr22)] and one BAC [CR392039 (chr21)] were chosen as representatives for the five acrocentric chromosomes. Pairwise comparisons were performed to determine the level of sequence identity and this varies from 86-99%, with the average level of identity being 93.3% (Fig. 2B). Similar to the DJ analysis, BACs further away from the rDNA were compared to see if the level of identity altered. Pairwise identities between CR392039 (chr21) and CR381535 (chr21), and AC145212 (unplaced) are 96.5% and 97.9% respectively. This higher level of identity indicates that the PJ is also conserved away from the rDNA. The lower level of interchromosomal identity in the PJ compared to the DJ largely results from an indel of ~2.7 kb that includes three Alu elements, an Alu that seems to have inserted into the PJ of some acrocentric chromosomes, and copy number and sequence variation of the ACRO1 147 bp repeats near the rDNA junction (Supplemental Fig. 6).

Repeat content analysis of the DJ and PJ

The repeat content of the contigs was determined using RepeatMasker. Novel tandem repeats in DJ and PJ contigs were searched using Tandem repeat finder (Benson 1999) and BLAST. mVISTA (Frazer et al. 2004) used to generate the similarity plot of the two arms of the inverted repeat. MAFFT (Katoh et al. 2009) was used for ACRO138 repeat multiple sequence alignments.

Gene prediction pipeline for the DJ and PJ

We designed a four-stage pipeline to determine the presence of potential gene coding regions in the DJ and PJ contigs. We used both RepeatMasker masked and unmasked DJ and PJ contigs to predict the potential genes.

Step 1: *Ab inito* gene prediction tools Genscan (Burge and Karlin 1997), Fgenesh (Salamov and Solovyev 2000), glimmerHMM (Majoros et al. 2004) and GeneMark (Besemer and Borodovsky 2005) were used to predict gene signals.

Step 2: Homology searches of the DJ and PJ contigs were performed using BLAST against the following GenBank datasets: non-redundant protein and reference mRNA sequences specific restricted to primates, and EST sequences specific to human. Hits were then remapped to the contigs using two spliced alignment tools: Exonerate (Slater and Birney 2005) and PASA (Program to Assemble Spliced Alignments (Haas et al. 2003)). Exonerate was used for protein, reference mRNA and EST sequences while PASA was only used for EST sequences. For both tools 90% identity was used as the filter cutoff.

Step 3: All the evidence from Steps 1 and 2 was combined using EVM (EVidence Modeller (Haas et al. 2008)). To optimize the weightings of the *de novo* gene prediction, protein, and EST evidence streams for merging, EVM simulations were performed using variable weights. Any EVM gene models that did not contain any database evidence or contained one-or-more exons that lacked any database evidence were removed.

Step 4: In the final curation step, EVM gene models, as well as protein sequence matches from Exonerate that were not included by EVM in step 3, were mapped to the contigs to compile the complete set of putative gene models. Apollo v1.11.6 (Lewis et al. 2002) was used for visualization and refinement of the gene models.

For the DJ, the gene prediction pipeline predicted eight potential genes from the unmasked and four from the masked sequence (Supplemental Fig. 12, Supplemental Table 1). All the potential genes except one are single exon genes. For the PJ, six genes were predicted from the unmasked and four from the masked sequence (Supplemental Fig. 12, Supplemental Table 2). One of the gene models from the PJ has alternative transcripts. Two of these gene models are multi-exonic while four of them have a single exon only. In all cases the masked gene models were a subset of the unmasked models.

Chromatin profiling method

Mapping

The short read length (36 bp for ChIP-seq and 75 bp for RNA-seq) means it is likely that many reads will map to both the DJ and the human genome. To map the reads uniquely to the DJ, we created a custom human genome that includes the latest human assembly hg19 and DNA sequences of the DJ and human rDNA repeat (extracted from BAC clone RP11-337M7, GenBank accession number AL592188). We mapped the ChIP-seq data onto this custom genome using bowtie (v0.12.7) (Langmead et al. 2009) with parameters -l 34 -a --best --strata -m 1 to take into account of sequencing quality and to yield the best mapping rate. Only uniquely aligning reads were kept for further analyses, with potential duplicates being removed from the alignments using Picard (http://picard.sourceforge.net/). Mapped reads from all replicates for each chromatin mark were combined to obtain the highest read coverage.

Signal profiling

For each mapped read, we created a tag by extending 200 bp (the known expected fragment size) from 5’ end towards 3’ end of the read. We then calculated the number of tags overlapping each base (also known as coverage depth) across the custom genome. This coverage depth was then normalized per million total mapped reads. Finally, to smooth the signal profile, we ran a sliding 200 bp window (with step size of 10 bp) across the DJ and calculated the average normalized coverage depth for each window. Signal profiles for open chromatin markers (FAIRE and DNaseI) were processed using F-Seq (Boyle et al. 2008) following a published procedure (Song et al. 2011).

Peak calling

We used MACS (Zhang et al. 2008) with mostly default parameters (-g hs --nomodel --shiftsize 100 -p 1e-5) to call peaks for each chromatin mark. Since the expected fragment size is known we did not apply shifting model in MACS (--nomodel), but instead applied a shift size of 100 bp (--shiftsize 100) that is a half of the fragment size. We ran MACS for each chromatin mark in each cell type separately and used the corresponding Input DNA as a control (Fig. 6A).

Chromatin state analysis

To depict the chromatin landscape from the combination of chromatin marks, the multivariate Hidden Markov Model (HMM) software, ChromHMM, was used with default parameters (Ernst and Kellis 2012) to segment the custom genome into different chromatin states. The seven cell types we used were chosen because these have comprehensive data for all 10 chromatin marks. We ran the HMM model with 15 states (Supplemental Fig. 16) as this number is enough to characterize the whole human genome (Ernst et al. 2011).

References

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**(2): 573-580.

Besemer J, Borodovsky M. 2005. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res* **33**(Web Server issue): W451-454.

Boyle AP, Guinney J, Crawford GE, Furey TS. 2008. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics* **24**(21): 2537-2538.

Burge C, Karlin S. 1997. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268**(1): 78-94.

Ernst J, Kellis M. 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**(3): 215-216.

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M et al. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**(7345): 43-49.

Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. 2004. VISTA: computational tools for comparative genomics. *Nucleic Acids Res* **32**(Web Server issue): W273-279.

Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Jr., Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD et al. 2003. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* **31**(19): 5654-5666.

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9**(1): R7.

Katoh K, Asimenos G, Toh H. 2009. Multiple alignment of DNA sequences with MAFFT. *Methods Mol Biol* **537**: 39-64.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**(3): R25.

Lewis SE, Searle SMJ, Harris N, Gibson M, Iyer VR, Richter J, Wiel C, Bayraktaroglu L, Birney E, Crosby MA et al. 2002. Apollo: a sequence annotation editor. *Genome Biol* **3**(12): 1-14.

Majoros WH, Pertea M, Salzberg SL. 2004. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**(16): 2878-2879.

Noe L, Kucherov G. 2005. YASS: enhancing the sensitivity of DNA similarity search. *Nucleic Acids Res* **33**(Web Server issue): W540-543.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**(6): 276-277.

Salamov AA, Solovyev VV. 2000. Ab initio gene finding in Drosophila genomic DNA. *Genome Res* **10**(4): 516-522.

Slater GS, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**: 31.

Smit AFA, Hubley, R. and Green, P. RepeatMasker.

Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, Lee BK, Sheffield NC, Graf S, Huss M, Keefe D et al. 2011. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res* **21**(10): 1757-1767.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**(9): R137.

Supplemental Tables

Supplemental Table 1. Putative gene models from the DJ

Gene model number	Start coordinate	End coordinate	Length	Strand
dg1.1.1	33924	34145	221	+
dg2.1.1*	321828	322277	449	+
dg3.1.1*	327412	327705	293	+
dg4.1.1	330230	330427	197	+
dg5.1.1	333297	332554	743	-
dg6.1.1*	150885	150689	196	-
dg6.1.2*	155079	154917	162	-
dg7.1.1*	137251	136850	401	-
dg8.1.1	33974	33600	374	-

* predicted in both masked and unmasked sequences

Supplemental Table 2. Putative gene models from the PJ

Gene model number	Start coordinate	End coordinate	Length	Strand
pg1.1.1*	68543	68599	57	+
pg1.1.2*	125644	125874	231	+
pg1.1.3*	126489	126590	102	+
pg1.2.1	112302	113033	732	+
pg1.3.1	119489	119629	141	+
pg2.1.1*	127300	127674	375	+
pg3.1.1*	159642	159861	220	+
pg3.1.2*	159946	160145	200	+
pg4.1.1*	107900	107694	205	-

* predicted in both masked and unmasked sequences

Supplemental Table 3. Segmentally duplicated regions from the DJ contig

Contig	Start	End	Length	Duplicate chromosome location	Start	End	% identity
DJ	104572	107241	2670	Chromosome 1	68626275	68628991	85.9
DJ	321433	324062	2630	Chromosome 1	68626302	68628991	86.8
DJ	339231	340355	1125	Chromosome 2	133118377	133119492	86.9
DJ	133743	134999	1257	Chromosome 3	75692439	75693702	90.5
DJ	187100	192124	5025	Chromosome 3	75683443	75688394	86.5
DJ	192335	195685	3351	Chromosome 3	75679409	75682745	88.5
DJ	229508	233248	3741	Chromosome 3	75679017	75682746	88.7
DJ	233455	235987	2533	Chromosome 3	75683438	75685973	89.4
DJ	291986	295386	3401	Chromosome 3	75690329	75693702	88.8
DJ	133743	134999	1257	Chromosome 4	190967606	190968868	90.2
DJ	189397	192119	2723	Chromosome 4	190975243	190978018	87.3
DJ	192335	194410	2076	Chromosome 4	190978791	190980900	89.1
DJ	194452	196632	2181	Chromosome 4	190980897	190983106	89.3
DJ	228930	233248	4319	Chromosome 4	190978790	190983106	88.5
DJ	233455	235987	2533	Chromosome 4	190975469	190978029	88.4
DJ	291986	294160	2175	Chromosome 4	190968819	190970979	87.7
DJ	104624	106629	2006	Chromosome 5	97912035	97914081	87.2
DJ	322240	324062	1823	Chromosome 5	97912219	97914060	85.7
DJ	104572	105914	1343	Chromosome 10	126676918	126678260	90.4
DJ	322759	324062	1304	Chromosome 10	126676945	126678269	91.5
DJ	133743	134999	1257	Chromosome 10	135459443	135460705	89.9
DJ	189397	192124	2728	Chromosome 10	135466940	135469711	87.7
DJ	192335	196632	4298	Chromosome 10	135470341	135474655	88.8
DJ	228930	233248	4319	Chromosome 10	135470340	135474655	89.1
DJ	233455	235987	2533	Chromosome 10	135467165	135469716	88.9
DJ	291986	295386	3401	Chromosome 10	135459443	135462818	87.9
DJ	322516	323669	1154	Chromosome 11	43544396	43545549	85.6
DJ	191180	192208	1029	Chromosome 12	38482027	38483048	85.3
DJ	233375	234434	1060	Chromosome 12	38482027	38483075	85.2
DJ	104453	105987	1535	Chromosome Y	59001390	59002913	91.1
DJ	322768	324223	1456	Chromosome Y	59001463	59002943	88.6

Supplemental Table 4. Segmentally duplicated regions from the PJ contig

Contig	Start	End	Length	Duplicate chromosome position	Start	End	% identity
PJ	1	105260	105260	Chromosome 1	143533694	143428925	96.5
PJ	1	73800	73800	Chromosome 1	142553006	142568640	94.6
PJ	1	73523	73523	Chromosome 1	142957003	142882933	92.6
PJ	74595	105050	30456	Chromosome 1	142632870	142662577	92
PJ	83766	89181	5416	Chromosome 1	142867130	142861830	93.6
PJ	45647	50040	4394	Chromosome 1	826486	830889	92.6
PJ	100926	105105	4180	Chromosome 1	143290746	143286566	94.1
PJ	142793	146966	4174	Chromosome 1	143236907	143241077	96.5
PJ	142793	146963	4171	Chromosome 1	143351481	143347325	96.4
PJ	100926	105050	4125	Chromosome 1	143154903	143159025	95.5
PJ	142793	145163	2371	Chromosome 1	142797349	142799737	96
PJ	100942	103157	2216	Chromosome 1	142853069	142850868	94
PJ	77250	79225	1976	Chromosome 1	143134409	143136392	95.1
PJ	105310	143700	38391	Chromosome 2	95520400	95559252	90.4
PJ	166722	191289	24568	Chromosome 2	132522403	132547076	90.9
PJ	157565	166067	8503	Chromosome 2	132513740	132522358	93.2
PJ	176920	180292	3373	Chromosome 2	132577642	132580992	92.7
PJ	162850	166067	3218	Chromosome 2	132555152	132558361	94.5
PJ	103245	105238	1994	Chromosome 2	132391991	132393988	84.9
PJ	103700	105238	1539	Chromosome 2	131424391	131425925	87.6
PJ	205180	206424	1245	Chromosome 2	162137144	162138410	89.1
PJ	103245	104369	1125	Chromosome 2	132034672	132035800	85.2
PJ	103245	104369	1125	Chromosome 2	131209987	131211114	84.8
PJ	103245	104369	1125	Chromosome 2	131207537	131208665	85.6
PJ	103245	104369	1125	Chromosome 2	131428006	131429134	85.8
PJ	103254	104365	1112	Chromosome 2	131427378	131428485	85.1
PJ	103254	104365	1112	Chromosome 2	131208186	131209296	85.7
PJ	103254	104364	1111	Chromosome 2	132034042	132035151	85.4
PJ	103245	104298	1054	Chromosome 2	130819347	130820398	85.1
PJ	103350	104369	1020	Chromosome 2	130822611	130823632	85.2
PJ	103245	104260	1016	Chromosome 2	130823142	130824161	86.7
PJ	20453	71388	50936	Chromosome 3	75829445	75880859	92
PJ	360	20452	20093	Chromosome 3	75887065	75906840	91.1
PJ	71560	88466	16907	Chromosome 3	75809054	75826395	90.1
PJ	77250	94391	17142	Chromosome 4	49180639	49198010	93.5
PJ	77250	88799	11550	Chromosome 4	49608353	49620270	91.1
PJ	99701	105105	5405	Chromosome 4	49202173	49207588	94.5
PJ	99701	105105	5405	Chromosome 4	49588922	49594333	94.4
PJ	64214	67839	3626	Chromosome 4	49161181	49164805	93
PJ	64214	66367	2154	Chromosome 4	49628635	49630784	92.9
PJ	204592	206423	1832	Chromosome 4	49282737	49284629	86.1
PJ	204592	206423	1832	Chromosome 4	49294922	49296797	87.7
PJ	204592	206423	1832	Chromosome 4	49301077	49302933	87.9
PJ	204592	206423	1832	Chromosome 4	49307357	49309231	87.9
PJ	204592	206423	1832	Chromosome 4	49514283	49516172	87.2
PJ	204592	205855	1264	Chromosome 4	49311042	49312353	87.7
PJ	143988	145443	1456	Chromosome 7	97499612	97501087	88.5
PJ	132568	133633	1066	Chromosome 7	97488744	97489797	88.2

PJ	161526	199611	38086	Chromosome 9	42259637	42297833	92
PJ	161526	192870	31345	Chromosome 9	70655139	70686615	93.3
PJ	161526	192869	31344	Chromosome 9	45400618	45431950	93.6
PJ	175341	191110	15770	Chromosome 9	68322537	68338220	94.1
PJ	192869	199611	6743	Chromosome 9	45394186	45400585	90
PJ	193195	199610	6416	Chromosome 9	70686619	70693040	94.9
PJ	193198	199611	6414	Chromosome 9	43203986	43210444	94.4
PJ	161526	166068	4543	Chromosome 9	68308254	68312765	94.7
PJ	106413	109844	3432	Chromosome 9	44117027	44120374	87.7
PJ	108137	111348	3212	Chromosome 9	42364995	42368241	90.6
PJ	189660	192870	3211	Chromosome 9	43210454	43213696	94.1
PJ	158413	161573	3161	Chromosome 9	45431993	45435142	95.7
PJ	108232	111348	3117	Chromosome 9	69378612	69381771	90.9
PJ	108232	111348	3117	Chromosome 9	43133754	43136914	91
PJ	108235	111348	3114	Chromosome 9	67923394	67926548	90.9
PJ	166072	168898	2827	Chromosome 9	68312893	68315739	94.9
PJ	158413	160331	1919	Chromosome 9	70653271	70655188	96.1
PJ	158413	160331	1919	Chromosome 9	42257769	42259686	95.8
PJ	158413	160260	1848	Chromosome 9	68306368	68308214	96
PJ	106413	107996	1584	Chromosome 9	43132024	43133529	88.2
PJ	198109	199611	1503	Chromosome 9	68338785	68340315	94.5
PJ	102087	103220	1134	Chromosome 9	38567007	38568129	86.5
PJ	168902	191112	22211	Chromosome 14	19758417	19780416	91.7
PJ	158413	168513	10101	Chromosome 14	19817857	19828180	93
PJ	161526	168901	7376	Chromosome 14	19750845	19758348	93.2
PJ	162863	168513	5651	Chromosome 14	19361516	19367278	92.5
PJ	103982	105265	1284	Chromosome 14	19600050	19601322	85.1
PJ	103982	105265	1284	Chromosome 14	19974090	19975362	85.1
PJ	176920	180678	3759	Chromosome 18	14989355	14993073	93
PJ	201937	203471	1535	Chromosome 19	44916370	44917954	84.9
PJ	204592	206073	1482	Chromosome 19	44962091	44963591	89.8
PJ	202036	203471	1436	Chromosome 19	44959179	44960652	87
PJ	204672	206073	1402	Chromosome 19	44913397	44914821	89.2
PJ	16075	71345	55271	Chromosome 21	10113638	10168755	94.6
PJ	158413	199611	41199	Chromosome 21	10604062	10645045	92.8
PJ	74600	95600	21001	Chromosome 21	10178710	10199160	92.6
PJ	74595	94392	19798	Chromosome 21	9662432	9682082	94.4
PJ	1	15750	15750	Chromosome 21	10097985	10113638	94.1
PJ	94695	105103	10409	Chromosome 21	9682085	9692595	90.4
PJ	64214	69025	4812	Chromosome 21	9645775	9650595	93.1
PJ	99060	103157	4098	Chromosome 21	10199865	10203948	95.1
PJ	71425	73760	2336	Chromosome 21	9653425	9655770	92
PJ	69195	71350	2156	Chromosome 21	9650595	9652751	93.6
PJ	159895	191112	31218	Chromosome 22	16062555	16093911	92
PJ	162863	168513	5651	Chromosome 22	16460105	16465867	92.5
PJ	103982	105265	1284	Chromosome 22	16241735	16243007	85.1
PJ	103982	105265	1284	Chromosome 22	16246649	16247921	85.9
PJ	1	141920	141920	Chromosome Y	13295249	13436504	95.6
PJ	164404	171133	6730	Chromosome Y	13239281	13246083	96.1
PJ	144448	146947	2500	Chromosome Y	13292733	13295248	96.45

Supplemental Table 5. Fold increase of DJ sequences present in BAC array clones

Location of primer pairs	Clone 1*	Clone 2*
40.6 kb	5.0	3.5
93.9 kb	10.0	21.5
153.5 and 275.9 kb	5.0	13.5
196.4 and 229.2 kb	7.6	25.2
340.2 kb	8.7	12.6

* determined by qPCR, see Figure 5 for description of clones

Supplemental Table 6. Sequencing statistics for the DJ and PJ cosmids

Cosmid name	Yield (Mbp)	% PF[a]	# Reads	% of >= Q30 Bases (PF)	Mean Quality Score (PF)	Mean length of reads (bp)
LA13 165F6	233	91.62	4,621,240	94.67	37.59	55
LA14 101B3	203	91.23	4,036,040	94.18	37.25	55
LA15 25H3	217	92.92	4,237,550	95.5	37.88	55
N 29M24	281	92.39	5,530,692	94.99	37.67	55
LA13 133H12	438	88.40	9,013,761	93.28	37.12	55
LA15 64C10	407	91.45	8,105,879	95.13	37.77	55

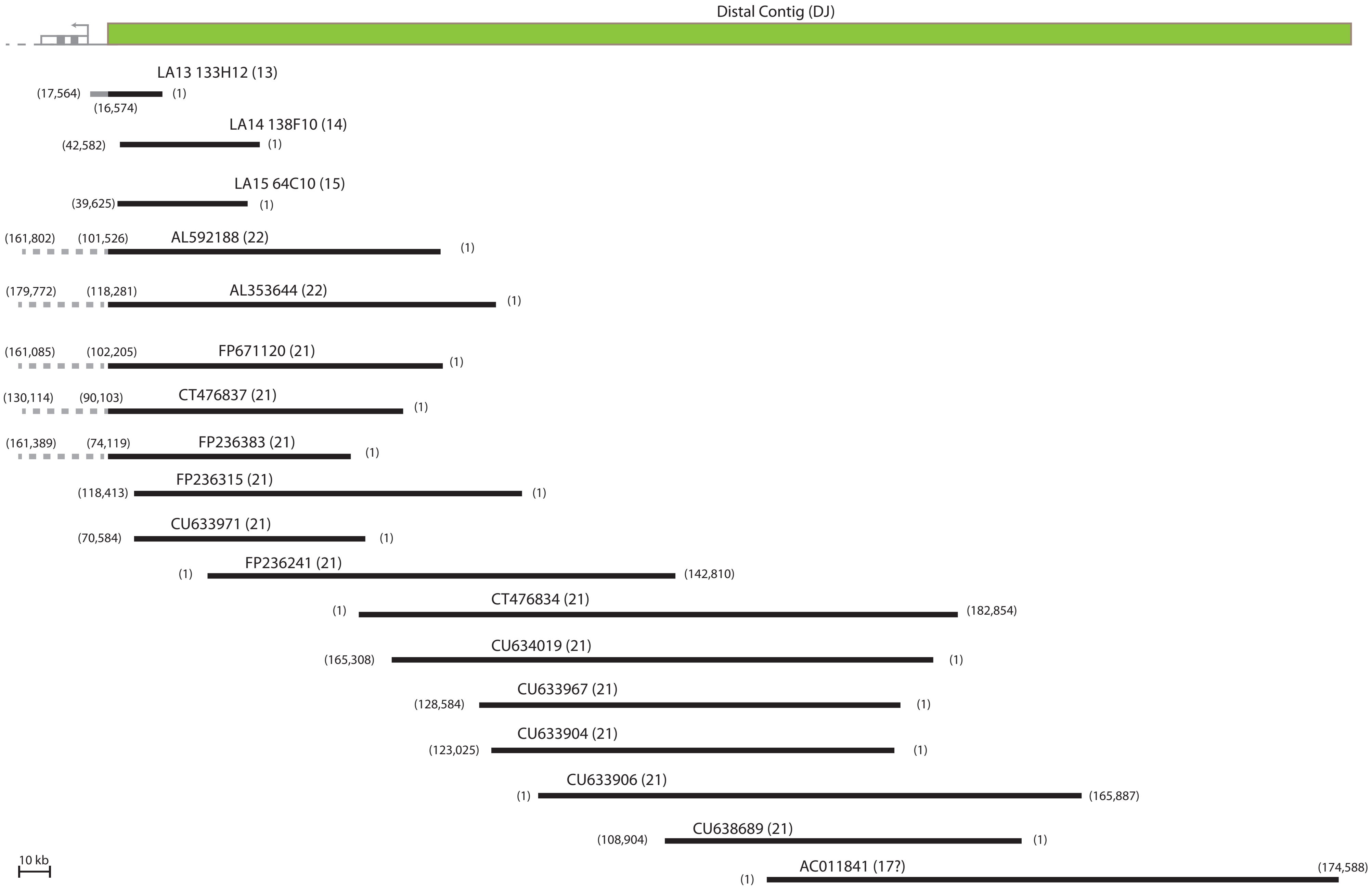
a. PF = The number of detected clusters that meet the filtering criterion.

Supplemental Table 7. Assembly statistics for the DJ and PJ cosmid assemblies

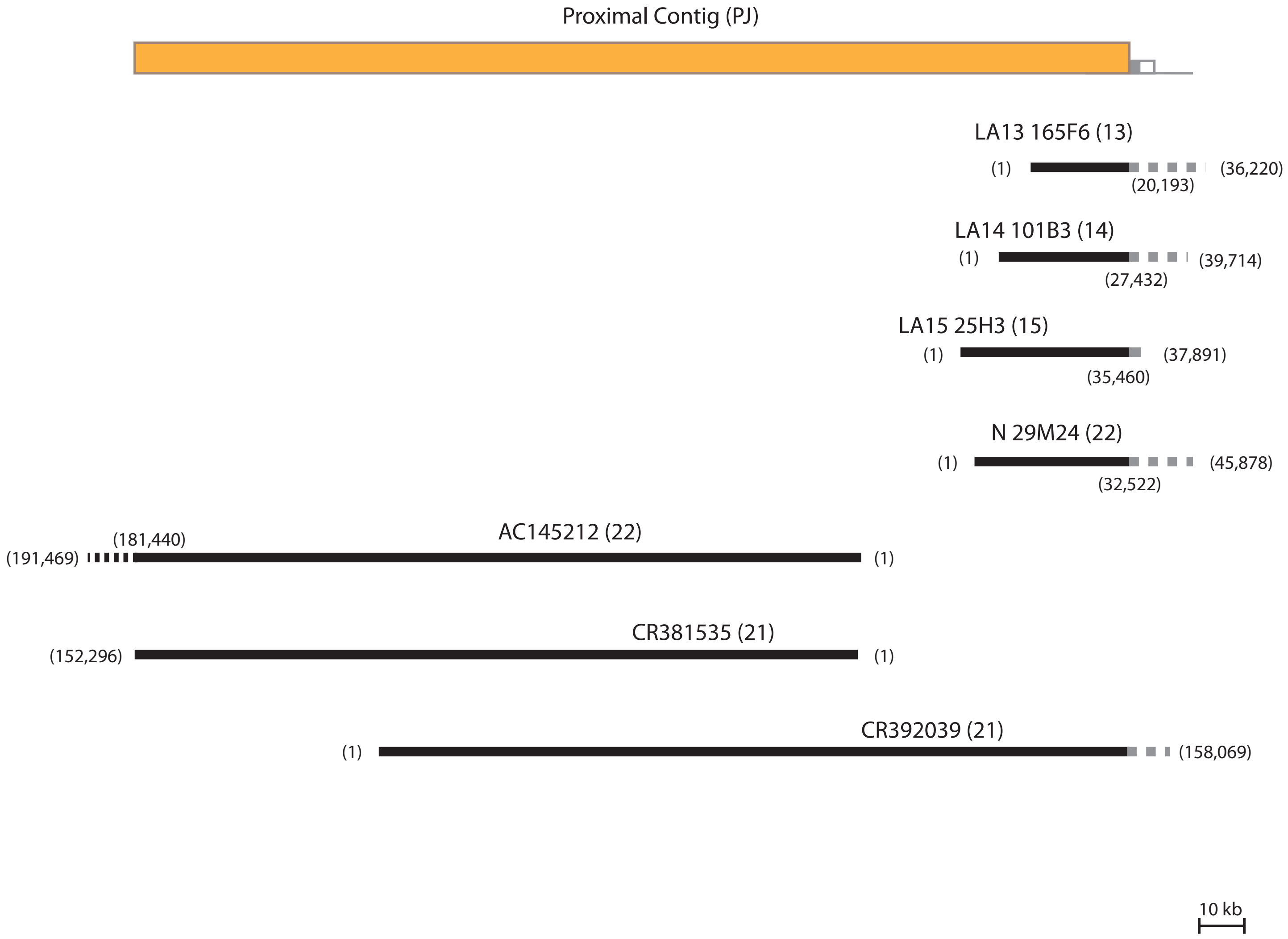
Cosmid name	Average Coverage	# of contigs	# of contigs > 100	# of contigs > N50	Max contig length (bp)	Mean (bp)	N50	Total size of contigs (bp)
LA13 165F6	4,172	83	24	4	12224	2067	4351	49619
LA14 101B3	2,932	68	25	4	12163	2138	5585	53460
LA15 25H3	4,363	30	9	2	21892	4979	6521	44817
N 29M24	4,681	609	91	13	3286	438	812	39879
LA13 133H12	10,736	265	52	10	4427	814	1182	42337
LA15_64C10	7,554	304	66	14	2573	719	1114	47490

Supplemental Table 8. DJ and PJ primers for mapping, ChIP, and RT-PCR

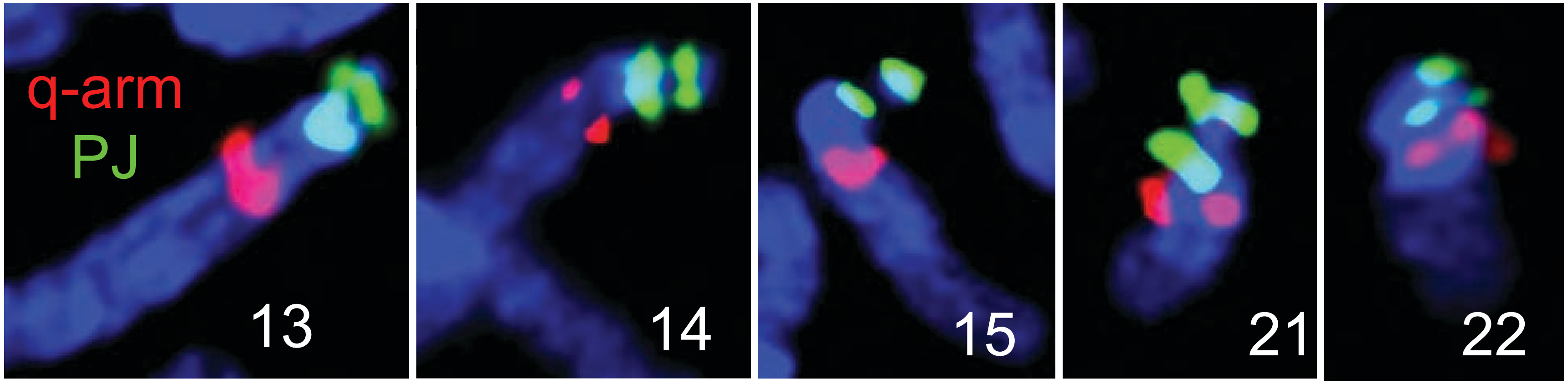
Kb distal to rDNA	**Forward primer**	**Reverse primer**
4.4	ACAACGCAAGGAAAAAGCGACACC	GGCAAGCGTTGGACTTGACCGT
40.6	TGCTGATTCCCCTTTTTGTC	GAAGGTGGTGTCGGTGAGAT
93.9	CCCTTCAGTGCTACACACCG	CCAGCACGTTGATCGCAAGG
106.1	TCTCCAGTGACACCTGCTGGCT	GACGTGAACGAGTCGCAGCC
138.0	TCCCAGCAAACGCAGCGAGG	AATCCCAGCATGCACGGGGC
153.5 and 275.9	GTGTCAAATCACATATGTCAC	CCCATTCTTCTCCAGCTGTGC
196.4 and 229.2	GCCCCAATGCCTCCCGCAAC	CAGCTCAAGAGCCTGGACCC
204.1	AACAGGCACCTGCATTGGGGA	AGCACACTCAAAGAGGCTGAGGA
292.0	TGGGGTCAGGGACAGTCCGC	TGCGACCAAAGGGCTGGGAG
340.2	GGCATTCGCCTCTGTGTCCC	CACAGGCCGGGCATTGTCAG
Kb proximal to rDNA	**Forward primer**	**Reverse primer**
6.6	CATACTACCTTGTCCTCCAG	GTCTGTTCAACATTTCCATG
Transcript	**DJ location/Exon and number if applicable**	**Sequence**
disnor138	137.6 forward	CCCCAGTCGTTTGGTCGCAGG
disnor138	137.4 reverse	GCCCCGTGCATGCTGGAATTG
disnor187 (AK026938)	155.0 /exon3 /forward	ACAGACCTGGGAGACATGCTACAC
disnor187 (AK026938)	150.7/exon4 /reverse	CCCGGGCTCTGAGTGCATCC
disnor 238 (BX647680)	238.5/exon1/forward	ACAGCCCCTGTTGCCCTGCG
disnor 238 (BX647680)	242.6/exon3/reverse	CTGCGTCTCACCAAGGCTCACC
Transcript	**Primer name**	**Sequence**
disnor 238 (BX647680)	DJ242.5f	GAGACCTTCCTTCCCCACGGGTT
disnor 238 (BX647680)	DJ244.3r	TCGCTGGGGGAATCGGGGATG



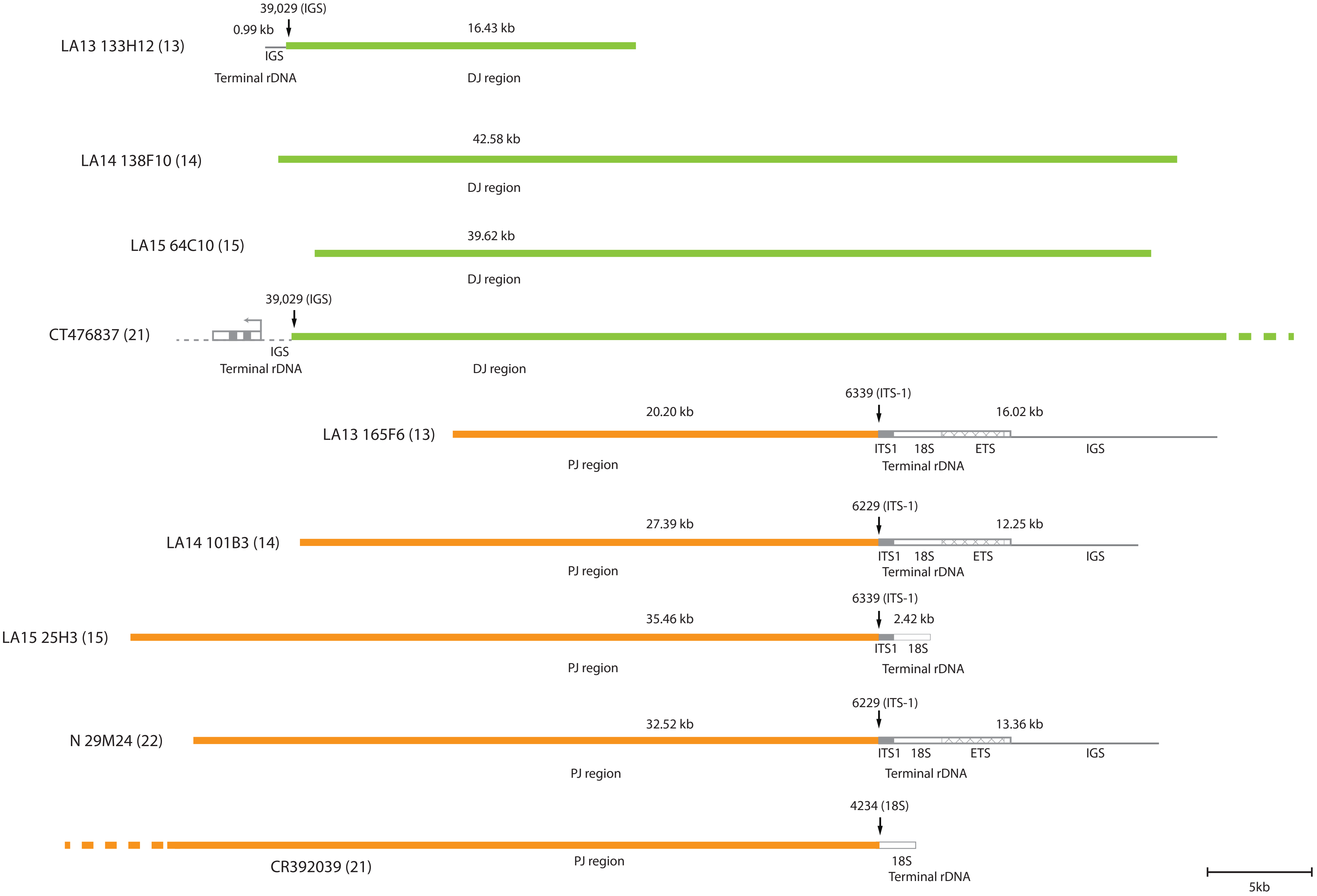
Supplemental Figure 1. Positions of cosmids and BACs in the DJ. Sequences of all DJ cosmids and BACs (heavy black lines) identified in this study are mapped onto a representative DJ sequence (green). The rDNA is indicated by dotted grey lines. The start, finish, and rDNA junction positions of each clone are indicated in parentheses (base pairs). Cosmid/BAC names are shown, with the chromosome of origin in parentheses. Scale is shown at the bottom



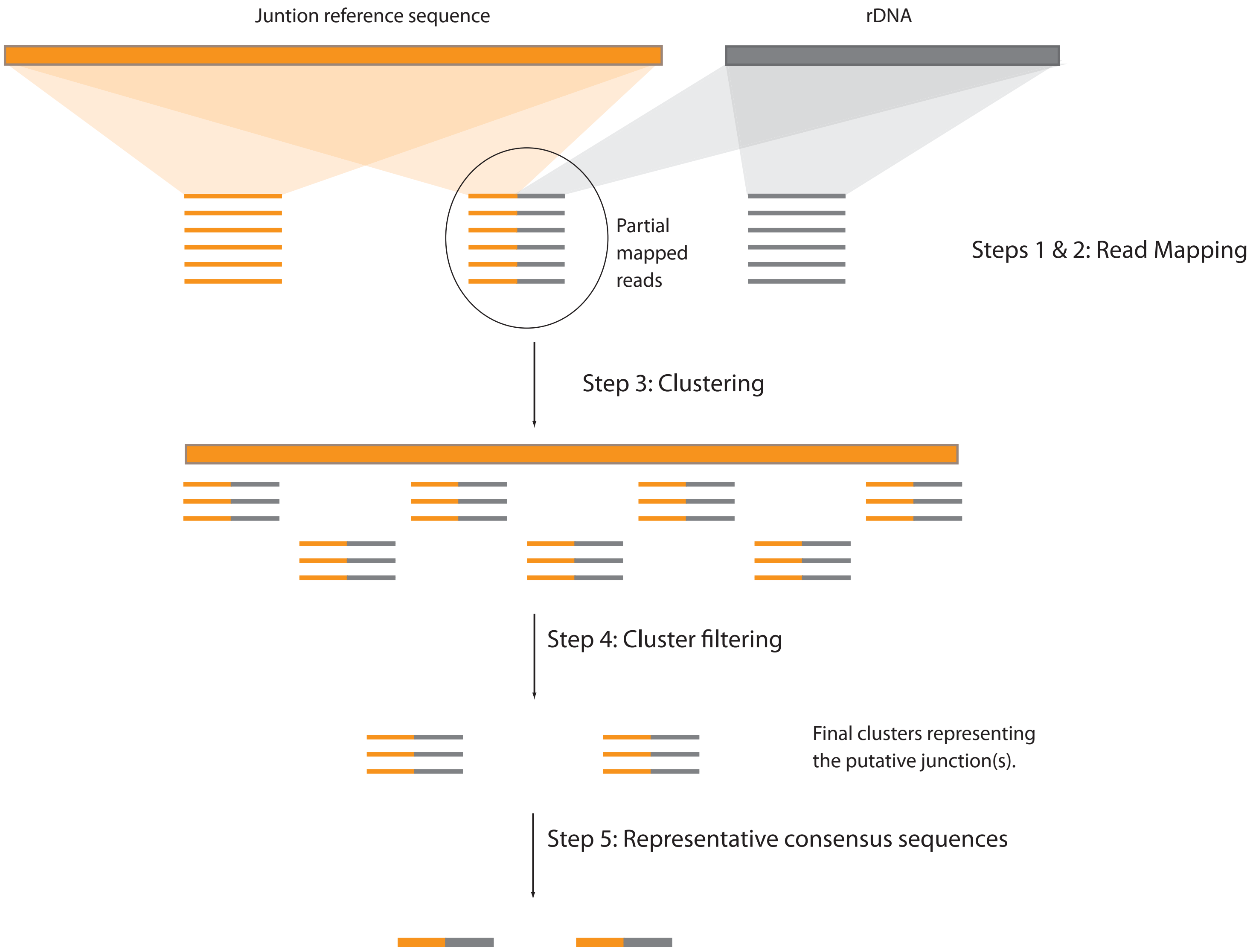
Supplemental Figure 2. Positions of cosmids and BACs in the PJ. Sequences of all PJ cosmids and BACs (heavy black lines) identified in this study are mapped onto a representative PJ sequence (orange). The rDNA is indicated by dotted grey lines. The start, finish, and rDNA junction positions of each clone are indicated in parentheses (base pairs). Cosmid/BAC names are shown, with the chromosome of origin in parentheses. Scale is shown at the bottom.



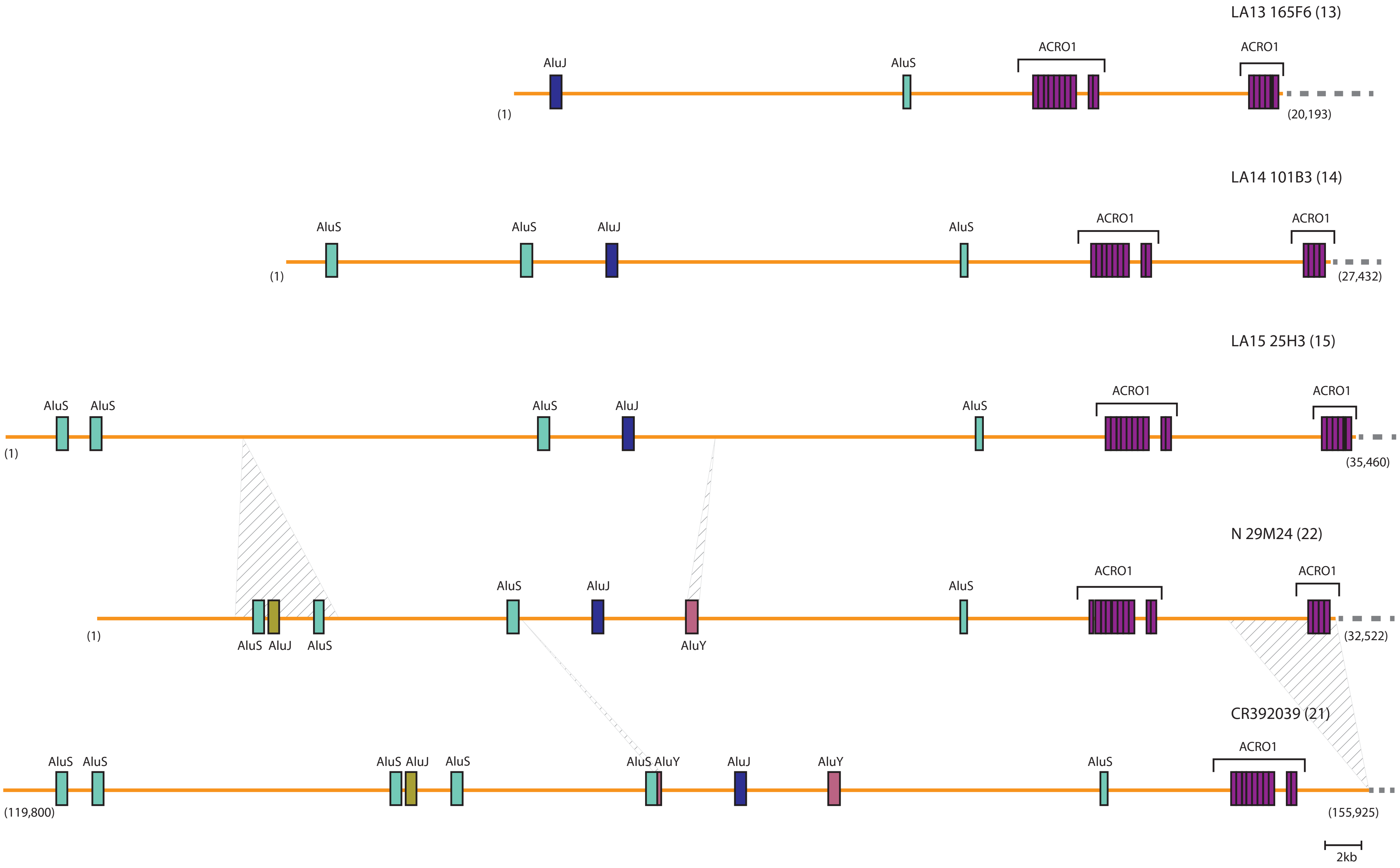
Supplemental Figure 3. PJ sequences are conserved among the acrocentric chromosomes. Human metaphase chromosomes were probed with PJ Bacmid CR381535 (labelled in green) and acrocentric specific q-arm bacmid probes (labelled in red).



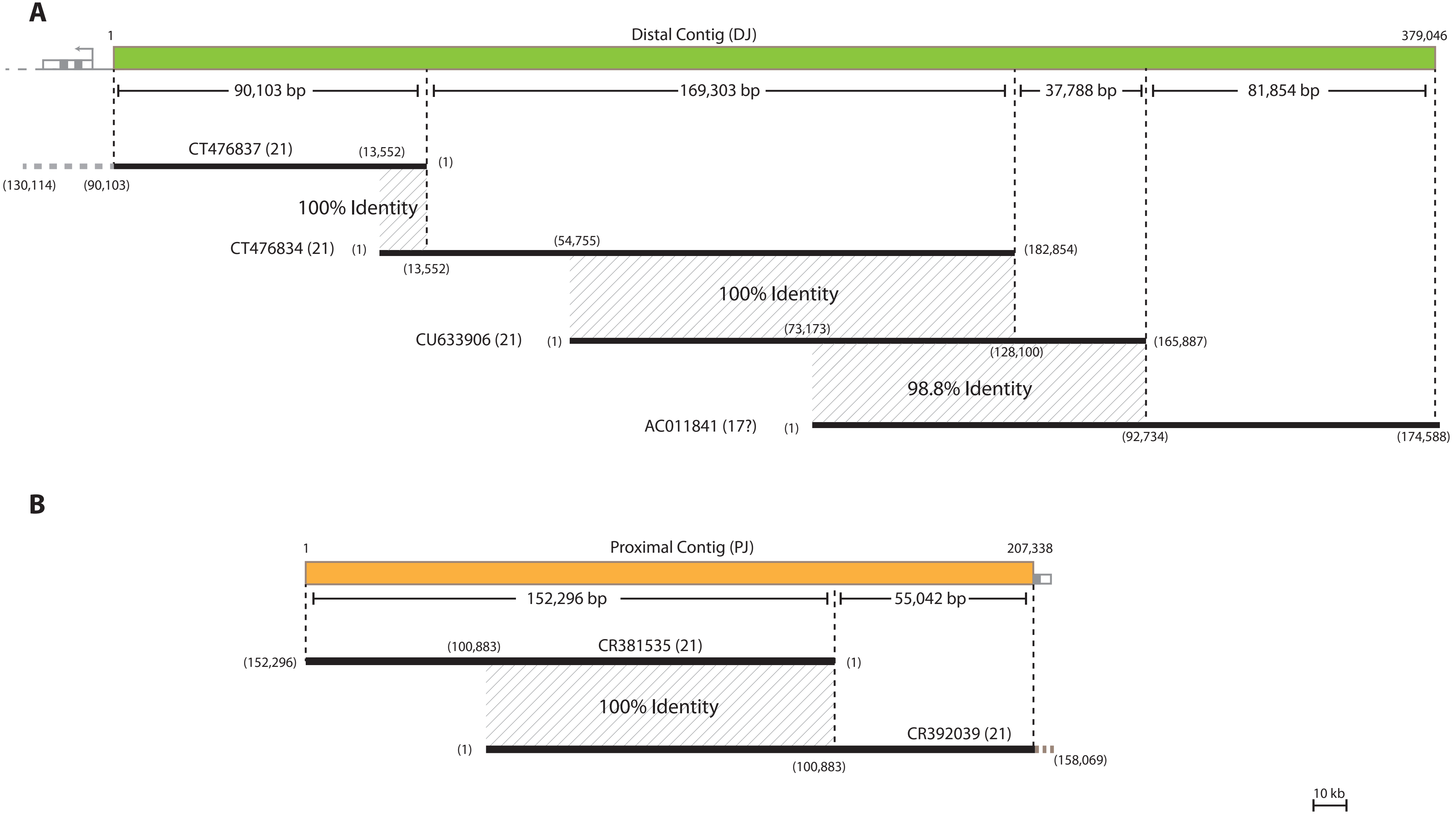
Supplemental Figure 4. Structure of DJ and PJ cosmids. The DJ (green) and PJ (orange) regions are represented by colored lines, and the rDNA by grey lines. Arrows indicate the junction point. Cosmid names are indicated with chromosome of origin shown in parentheses. The length of the flanking region and rDNA in each cosmid is given in kb. The position of the rDNA junction point is indicated in base pairs relative to the human rDNA unit in Genbank (U13369), with the region in which the junction falls given in parentheses. BACs CT476837 and CR392039 are included to compare cosmid and BAC junction positions. Scale is shown at the bottom.


Juntion reference sequence
rDNA
Partial mapped reads
Steps 1 & 2: Read Mapping
Step 3: Clustering
Step 4: Cluster filtering
Final clusters representing the putative junction(s).
Step 5: Representative consensus sequences

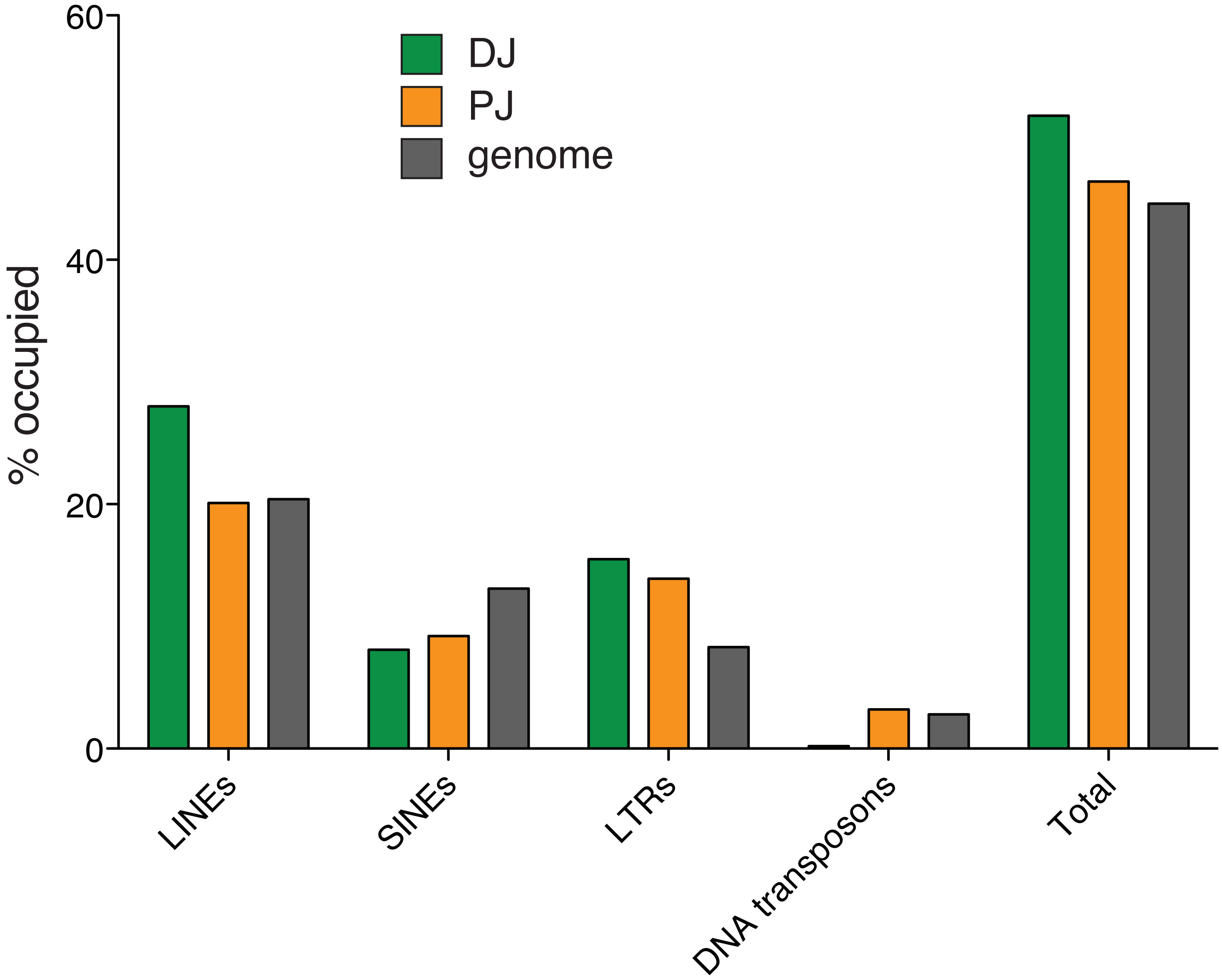
Supplemental Figure 5. Workflow for junction verification mapping pipeline. Steps 1 & 2: whole genome shotgun reads (lines) were mapped to a reference sequence (orange box) and to the rDNA (grey box) with cut-offs of 95% identity and 100 bp minimum alignment length. Step 3: Partially mapped reads that matched both the reference sequence and the rDNA (orange-grey hybrid lines) were clustered according to their position on the reference sequence. Step 4: Clusters were filtered using various criteria (Supplemental Methods). Step 5: Consensus sequences were made for each of the resulting clusters.



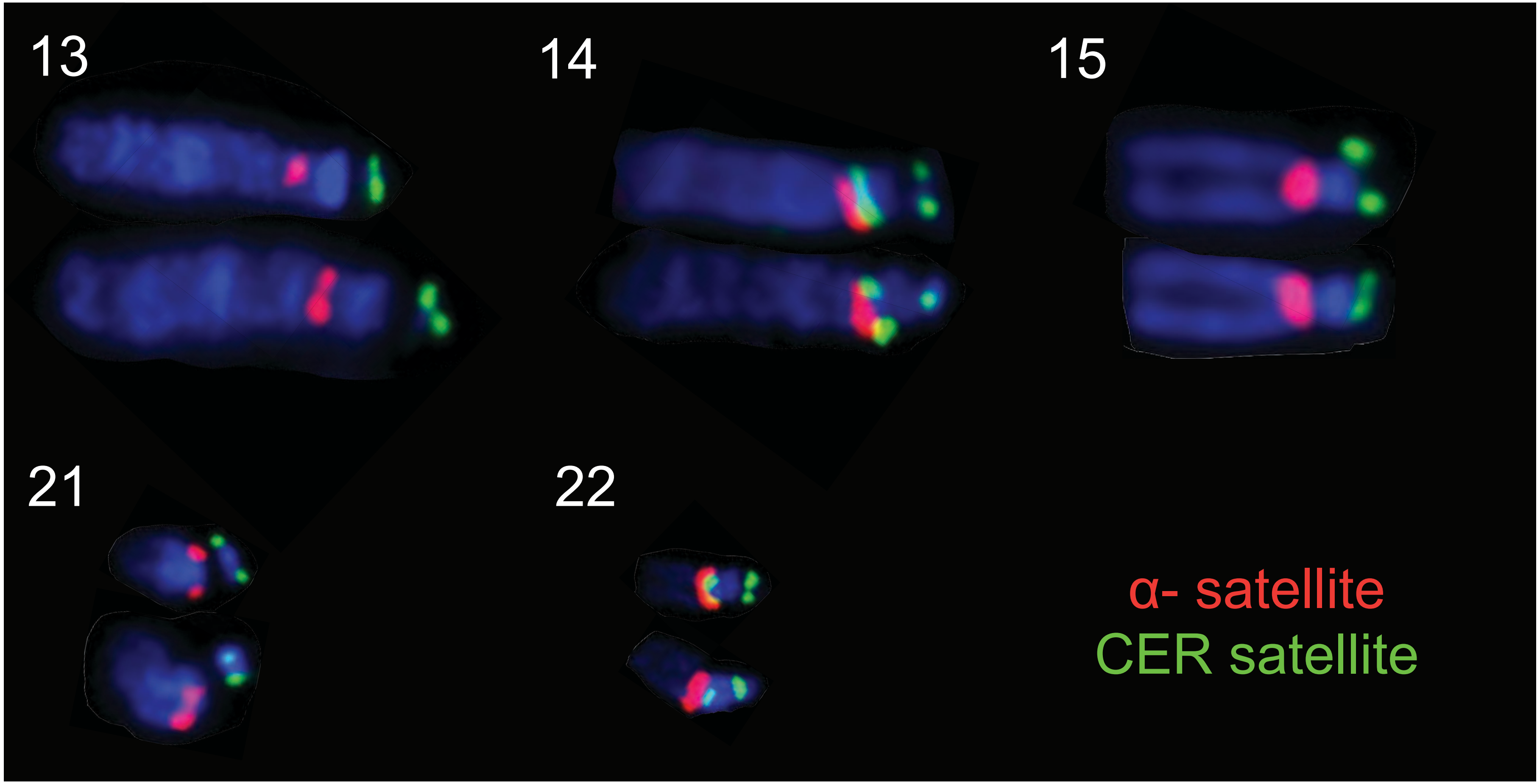
Supplemental Figure 6. Alu and 147 bp ACRO1 repeat inter-chromosomal PJ variation. The positions of the Alu and ACRO1 repeats in the PJ clones (orange lines) are indicated by colored boxes. Dotted grey lines represent the rDNA. Shaded regions represent indel events that include Alu or 147 bp ACRO1 repeats and result in interchromosomal length variation. Clone name, chromosome of origin (parentheses), and repeat name are all indicated. The positions in the clones where the rDNA begins are given in base pairs (parentheses). Scale is shown at the bottom.



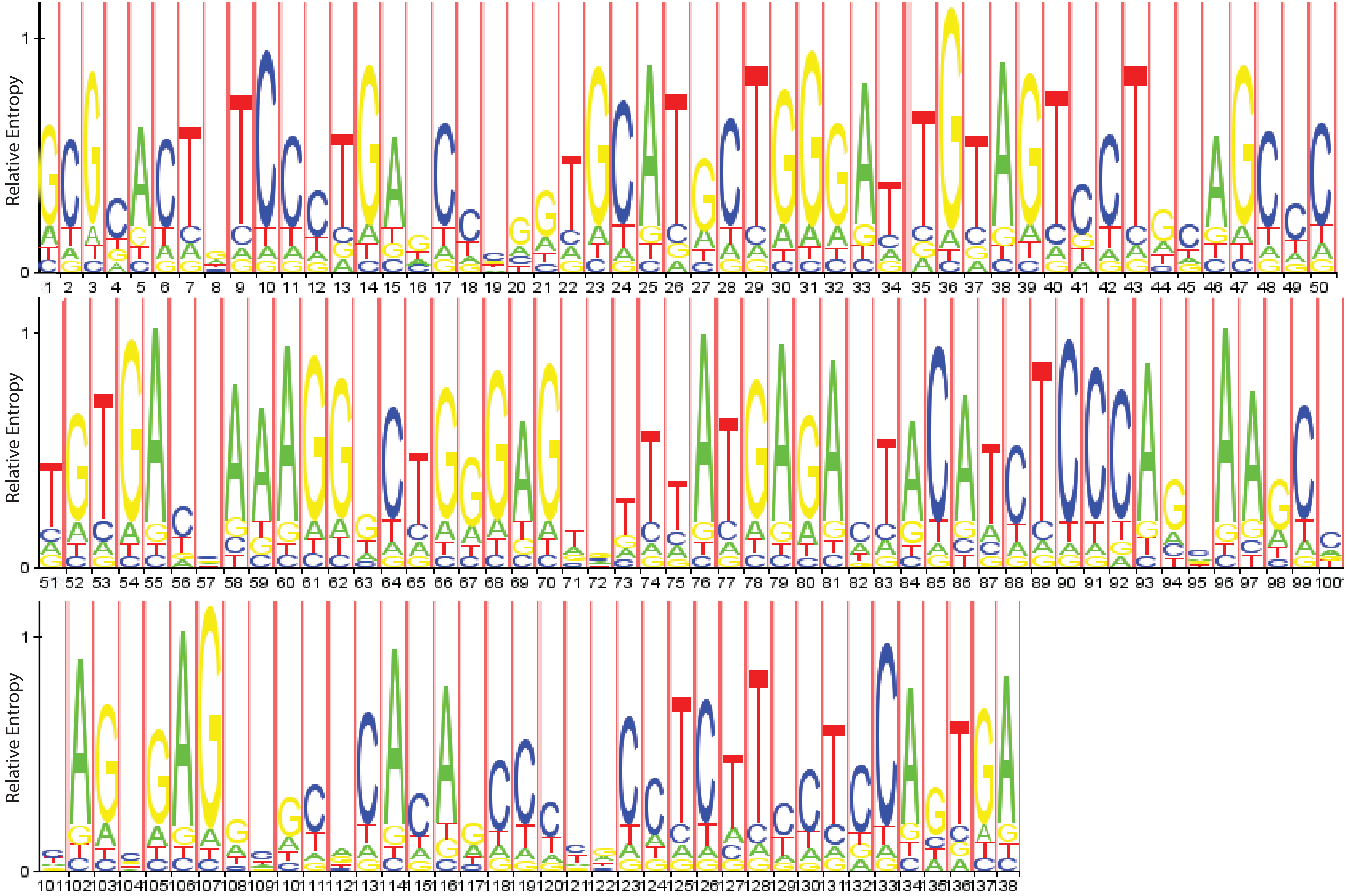
Supplemental Figure 7. Construction scheme for the DJ and PJ contigs. A. The DJ contig (green box) was obtained by merging four BACs (black lines; clone name and chromosome of origin are indicated) from two different chromosomes, and is 379,046 bp in length. **B.** The PJ contig (orange box) was obtained by merging two BACs from chr21, but the small piece of rDNA at the end was removed for our analyses. The resulting contig is 207,338 bp in length and is 100% identical to an unplaced human genome contig (GenBank Accession number NT_113958). Diagonal shading represents regions of BAC overlap. % identity and overlap start and stop positions (in parentheses) are shown. The dotted vertical lines demarcate parts of the BACs used in the contig formation, with the length of each fragment indicated immediately below the schematic contig. The rDNA is shown in grey. Scale is at the bottom.



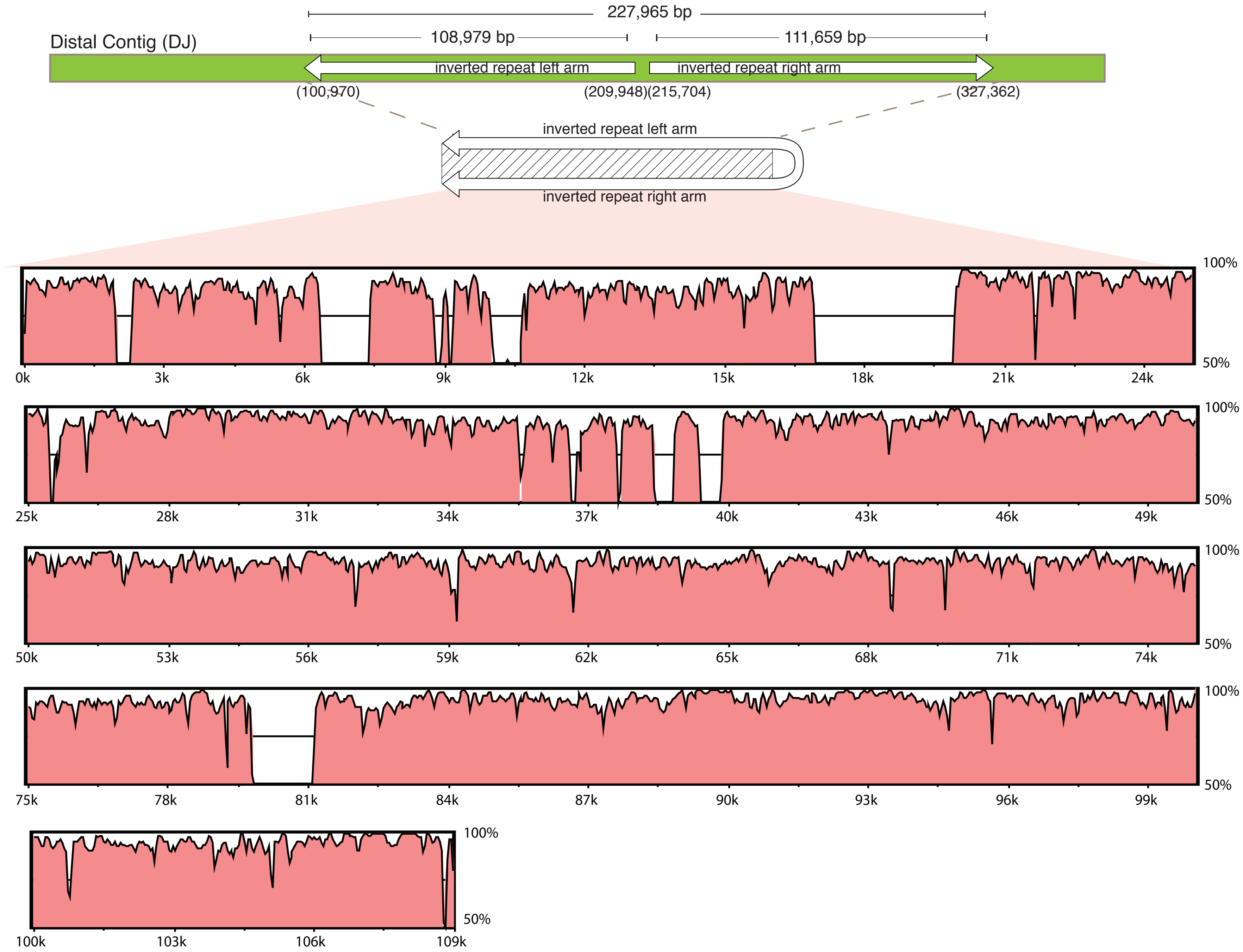
Supplemental Figure 8. Repeat composition of the DJ and PJ regions. Fraction of the DJ, PJ and total human genome occupied by different classes of transposon, as well as total repeat content, are plotted. Overall repeat content is broadly comparable between the PJ/DJ and the whole genome.



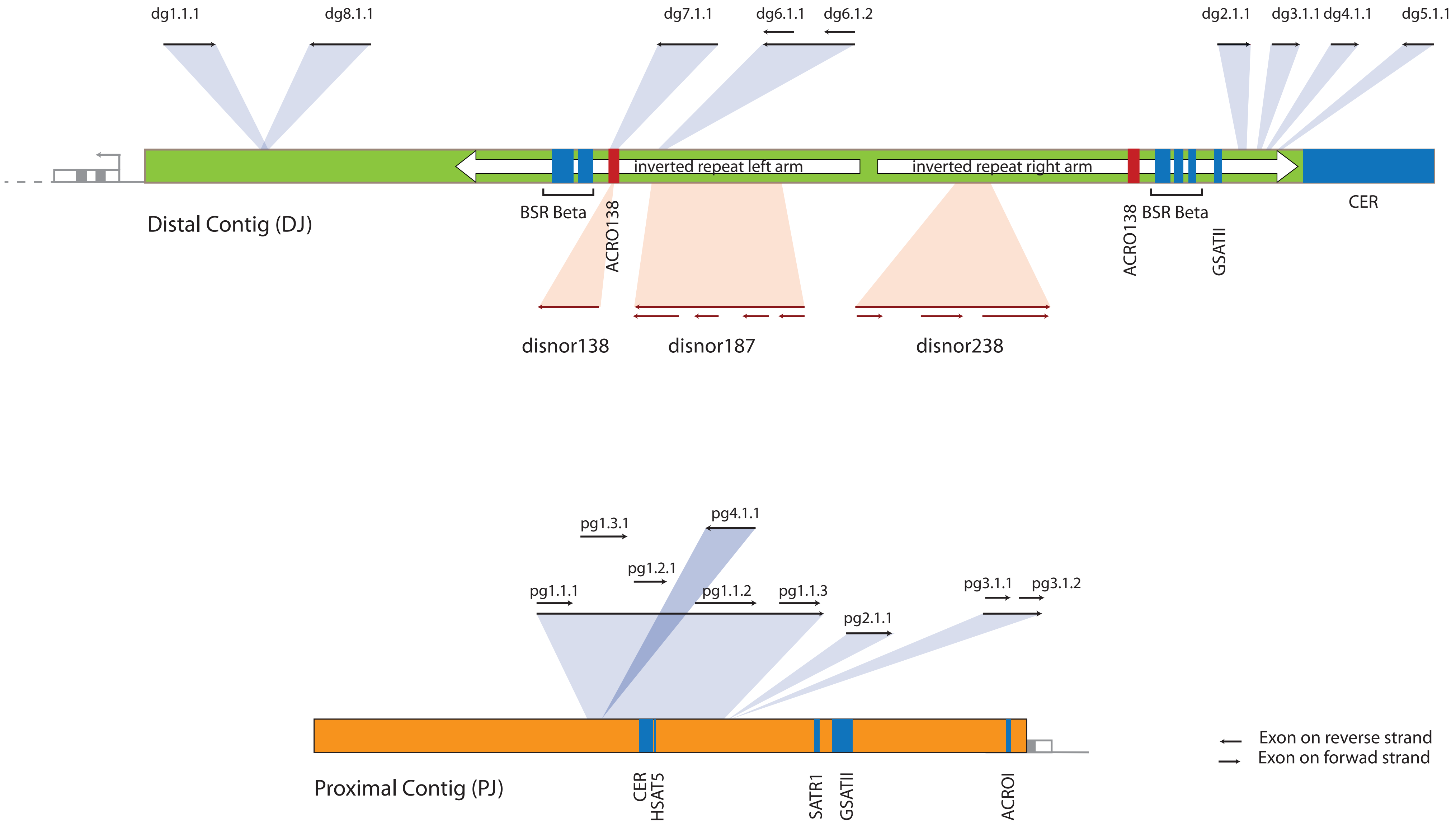
Supplemental Figure 9. 48bp CER repeats are found distal to the NOR on all five acrocentric chromosomes. Human metaphase chromosomes were probed with acrocentric specific centromeric α- satellite probes (labelled in red) and a 48 bp CER satellite probe (labelled in green). Note that peri-centromeric CER blocks are present on chromosomes 14 and 22.



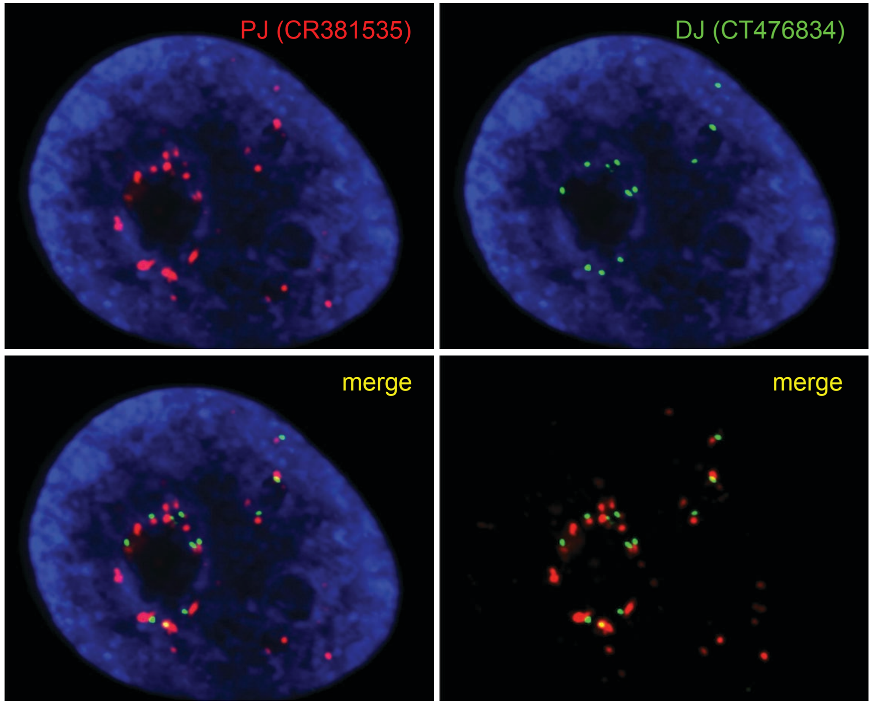
Supplemental Figure 10. HMM Logo for the DJ ACRO138 repeat. The logo was generated using repeat sequences from the 136 kb and 289 kb repeat blocks. It represents the consensus for the repeat. The height of each base represent the emission probability of that base at that specific position. The thickness of red vertical line shows the number of bases expected to be inserted at that position.



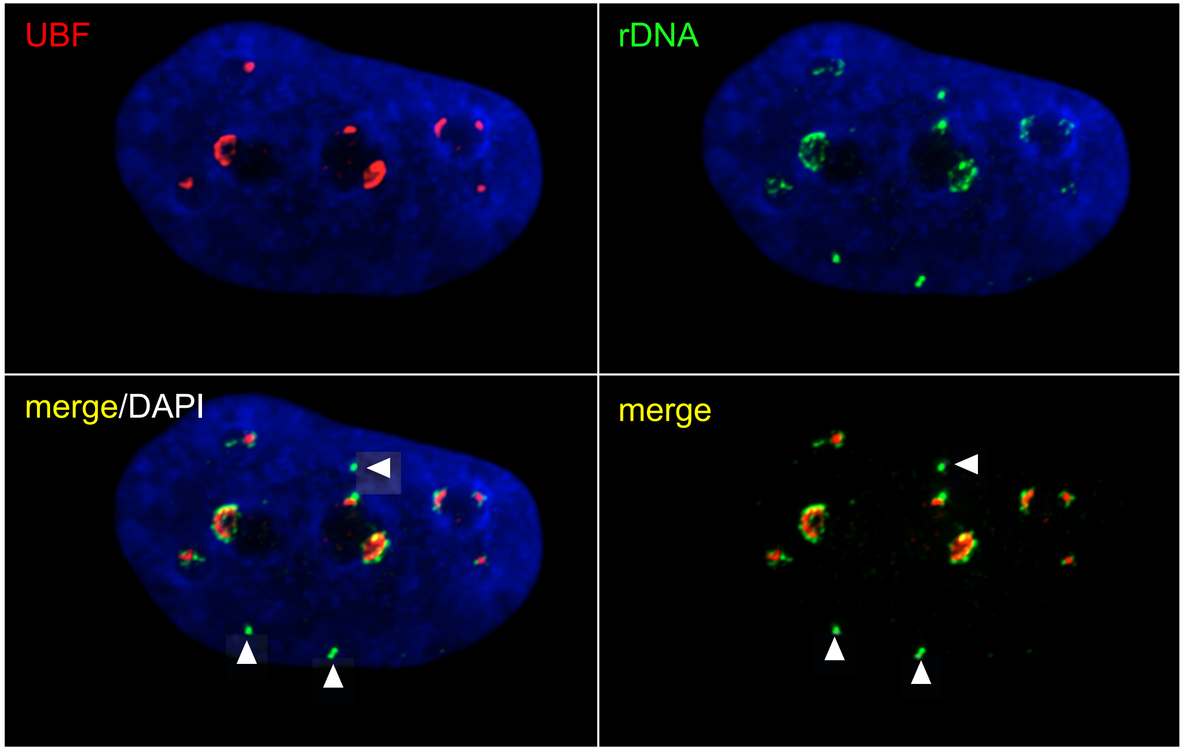
Supplemental Figure 11. Similarity plot of the DJ inverted repeat arms. Location and arrangement of the large inverted repeat (white arrows) in the DJ contig (green) is shown at the top. VISTA plot with a sliding window of 100 bp, showing the level of sequence identity between the two inverted repeat arms (represented by the folded repeat arms and hatched region below the contig), is shown below. Average sequence identity is 79.5%. Sizes and positions of the inverted repeats are indicated in base pairs.



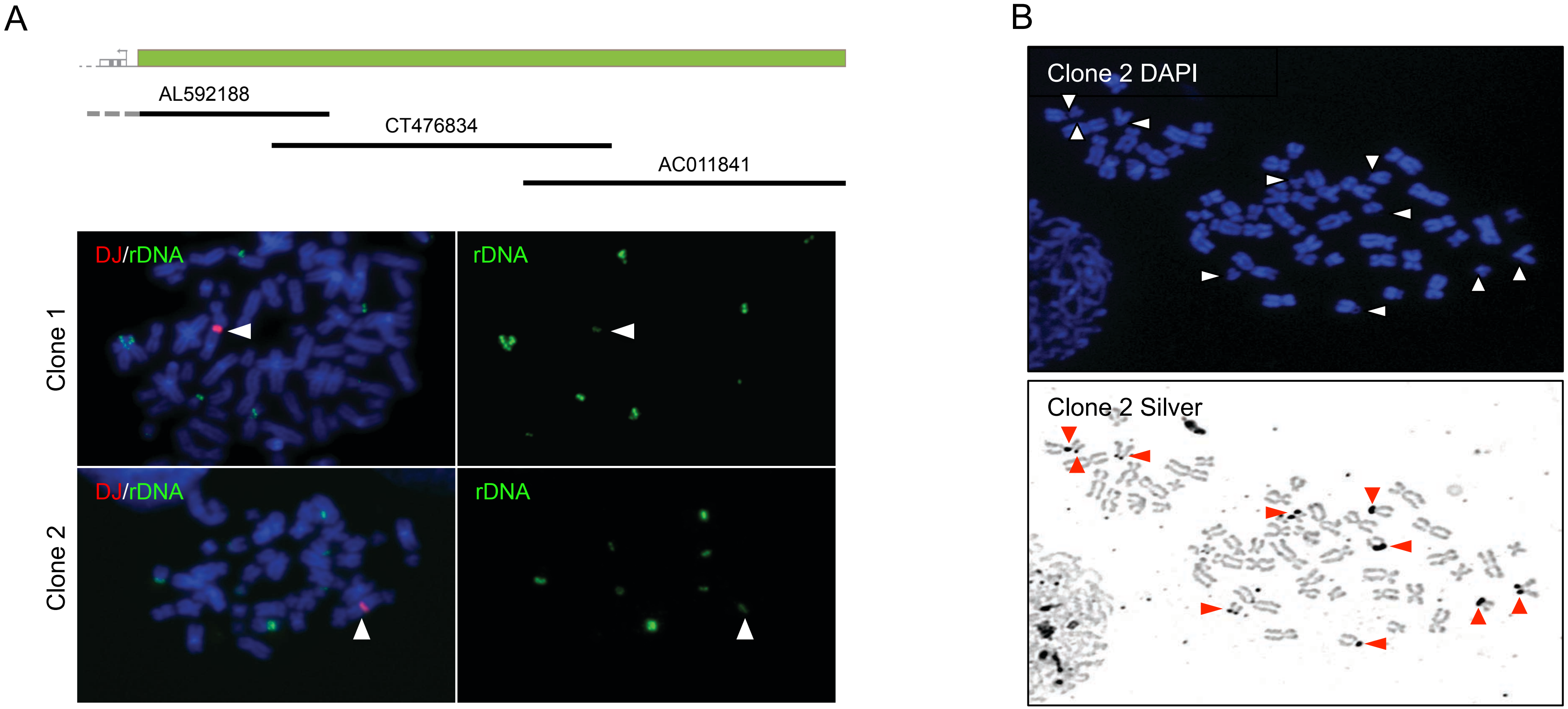
Supplemental Figure 12. Positions of sequence features in the DJ and PJ. Gene models from the gene prediction pipeline are shown as black arrows with their locations indicated by blue shading in the DJ (dg gene numbers) and PJ (pg gene numbers). Exons are indicated by small arrows above the main gene model arrow. Gene model pg1 has three alternative transcripts (pg1.1, pg1.2 and pg1.3), as indicated by different tiers of small arrows. Transcripts in the DJ from the transcriptome analysis are shown as red arrows with their locations indicated by red shading. The novel ACRO138 repeat is shown as a red block, with other satellite repeats shown in blue.



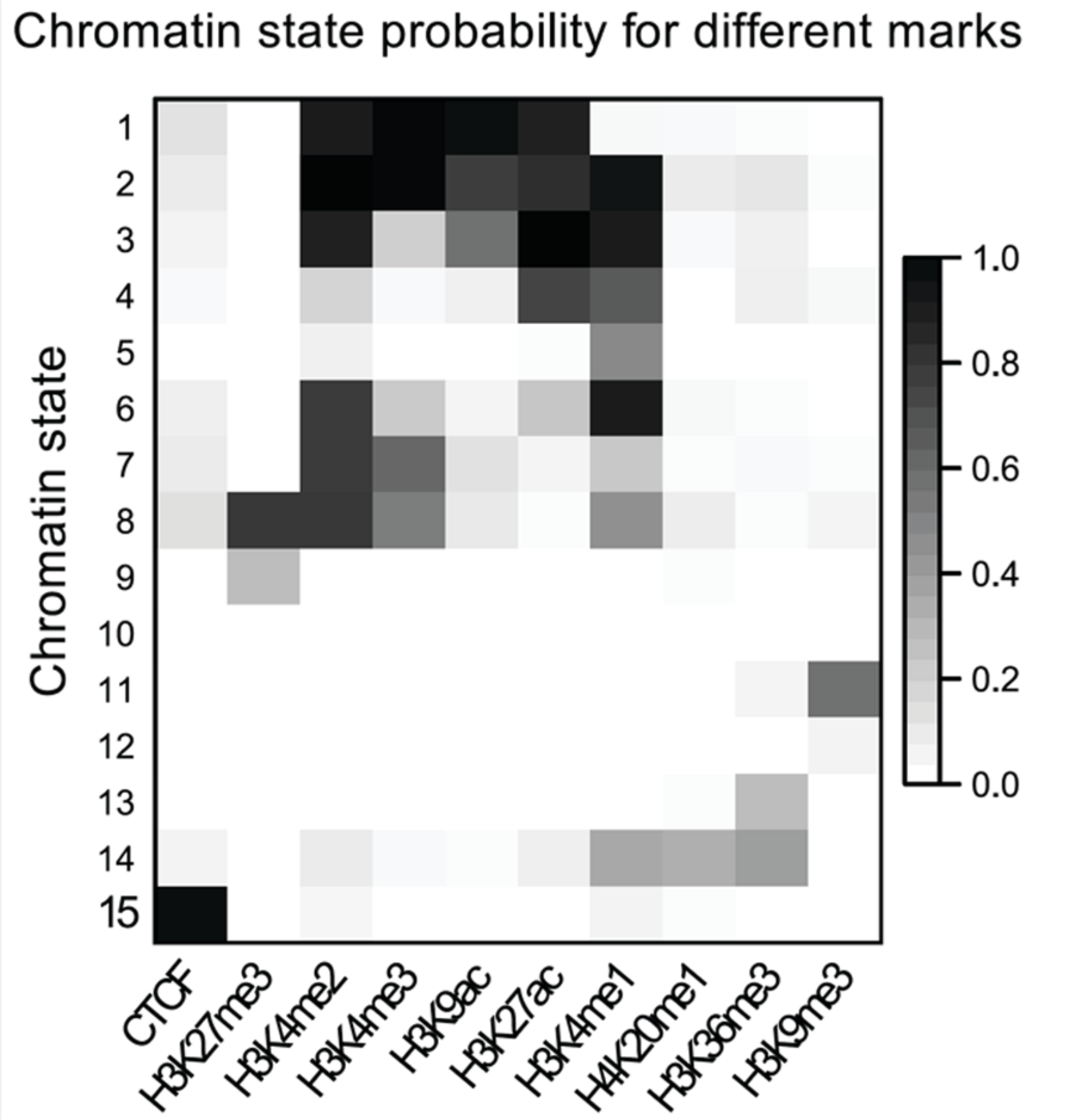
Supplemental Figure 13. Comparison of PJ and DJ positioning in human cells. FISH was performed on human HT1080 cells using PJ BAC (CR381535) and DJ BAC (CT476834) labelled in red and green respectively. DJ and PJ signals are excluded from nucleoli (DAPI free regions). The PJ hybridisation pattern is more complex than that obtained with the DJ probe. Note the presence of isolated PJ signals.



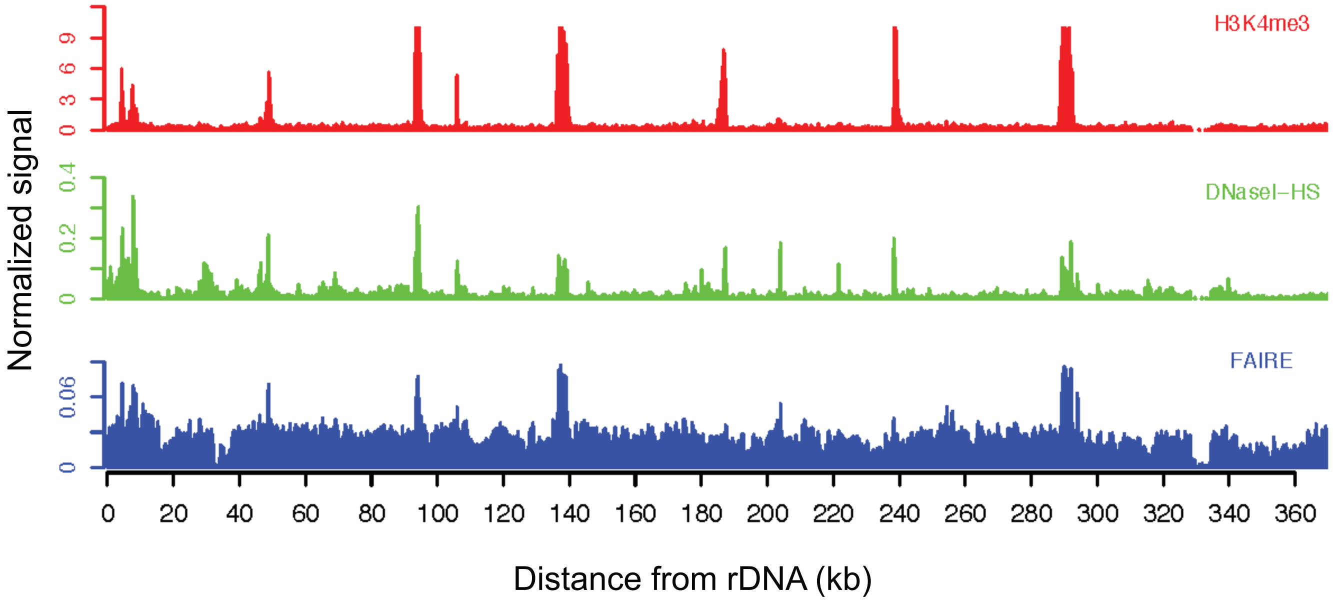
Supplemental Figure 14. UBF and rDNA colocalise in nucleolar caps upon AMD treatment. 3D-immuno FISH was performed on AMD treated HeLa cells using antibodies agains UBF (red) and a rDNA IGS probe. Note the presence of silent (non-UBF associated) NORs (arrowheads).



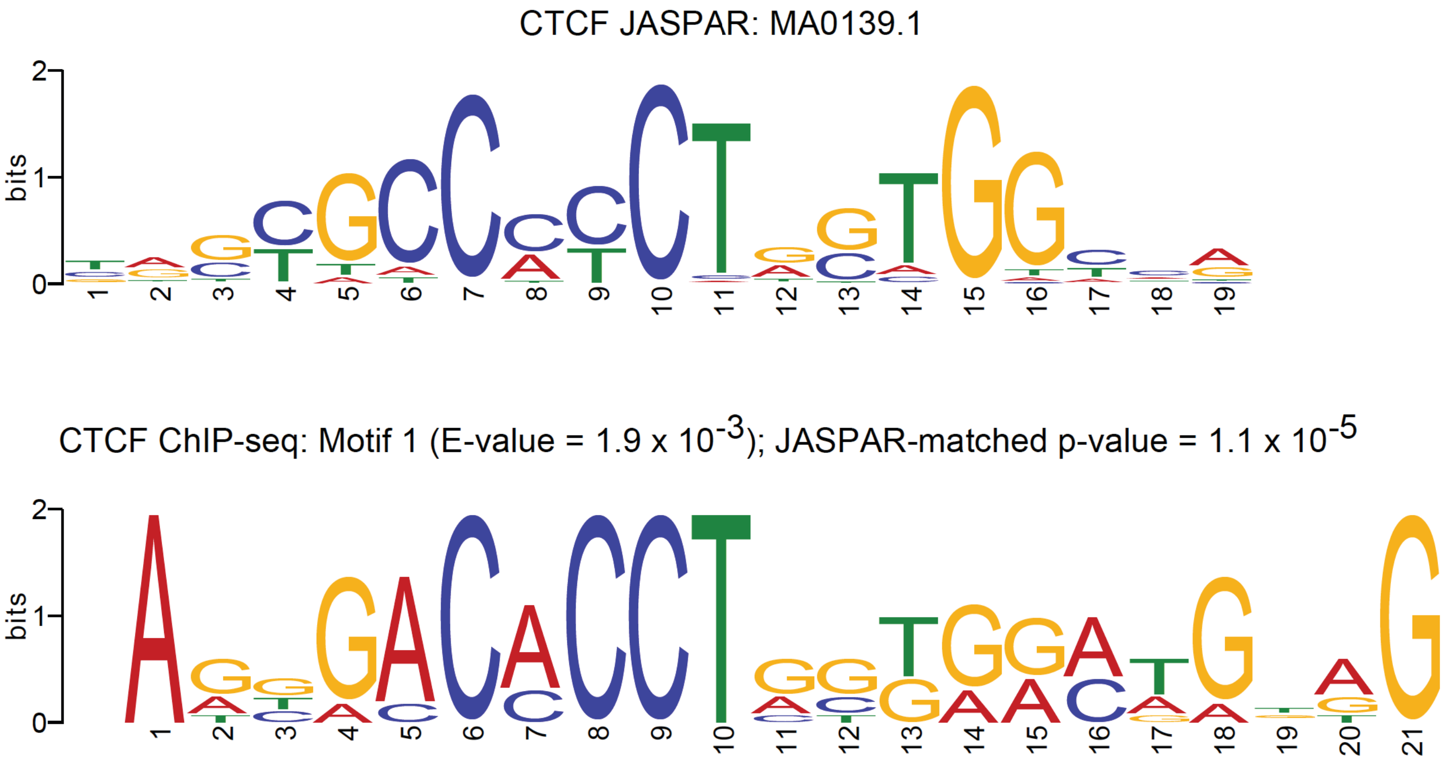
Supplemental Figure 15. Construction of DJ BAC arrays on metacentric chromosomes. A. Three BACs encompassing the DJ contig were transfected into HT1080 cells. FISH was performed on metaphase chromosome spreads prepared from two clones (1 and 2) with rDNA (green) and DJ BAC CT476834 (red) probes, demonstrating that in both cases DJ arrays have integrated into non-NOR bearing metacentric chromosomes. Note the presence of small amounts of rDNA at these ectopic arrays. **B.** Silver-staining of chromosome spreads reveals that the rDNA present within ectopic DJ arrays is inactive. Multiple spreads (> 20) from each clone were examined. A representative spread from clone 2 is shown. The silver associated with endogenous active NORs is indicated by arrowheads.



Supplemental Figure 16. Heat map of the association between chromatin marks and chromatin states. Each row shows the association of a chromatin sate with different chromatin marks and the frequencies between 0 and 1 with which they occur. These frequencies are equivalent to the emission probability parameters of the multivariate Hidden Markov Model learned across the custom human genome during model training.



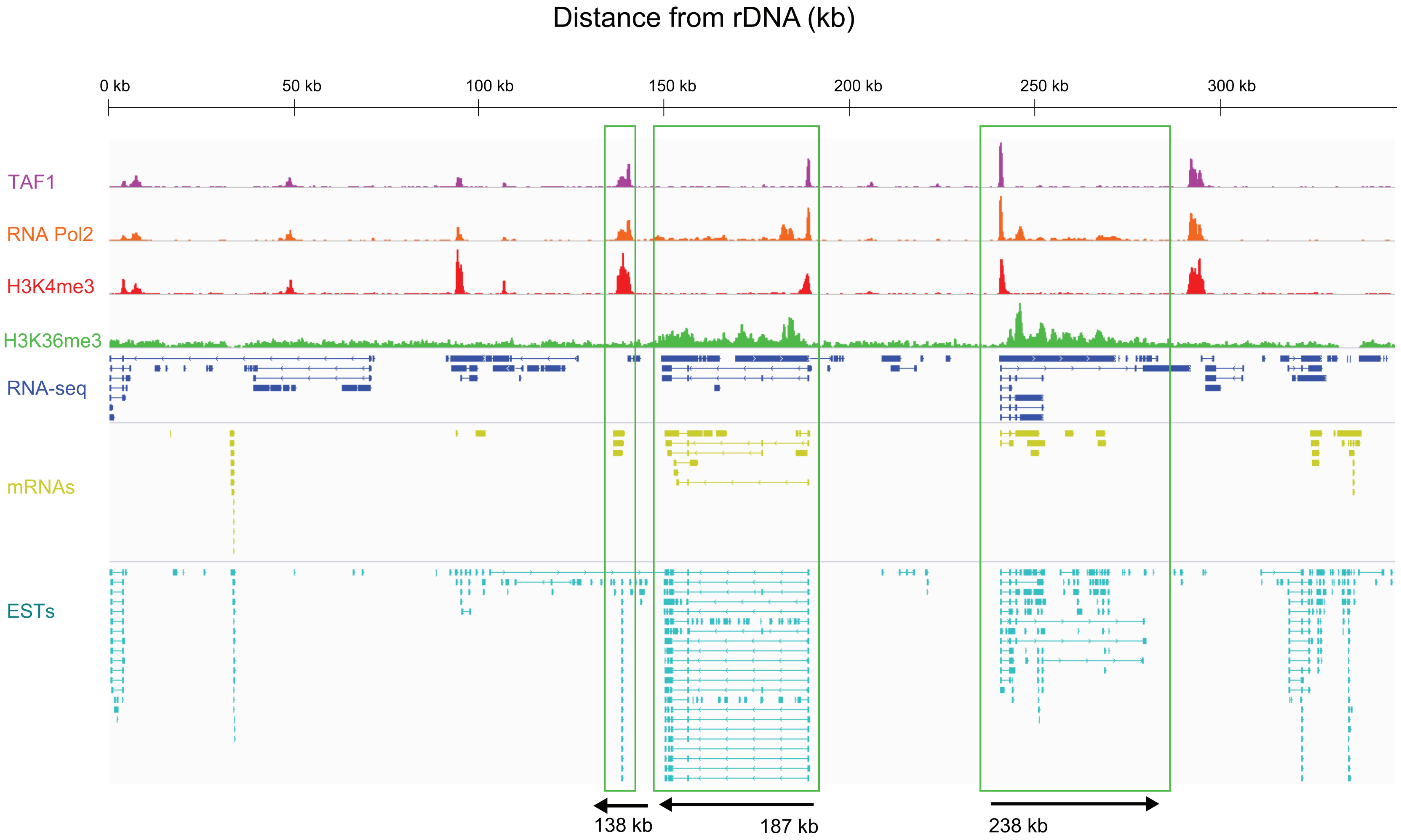
Supplemental Figure 17. H3K4me3 peaks overlap with open chromatin domains identified by DNaseI-His-seq and FAIRE-seq. Top track shows the enrichment signal of H3K4me3 that was normalized to tags per million mapped reads (also shown in Fig. 4a). The two bottom tracks show enrichment signals of two open chromatin marks, DNaseI-HS (DNase I hypersensitive sites) and FAIRE (Formaldehyde Assisted Isolation of Regulatory Elements), calculated using F-Seq software (Supplemental Methods).



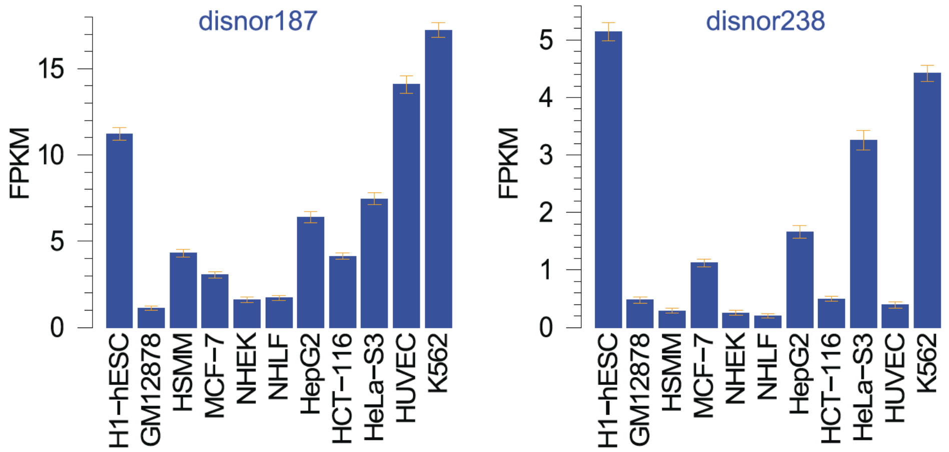
CTCF JASPAR: MA0139.1

CTCF ChIP-seq: Motif 1 (E-value = 1.9 x 10^{-3}); JASPAR-matched p-value = 1.1 x 10^{-5}

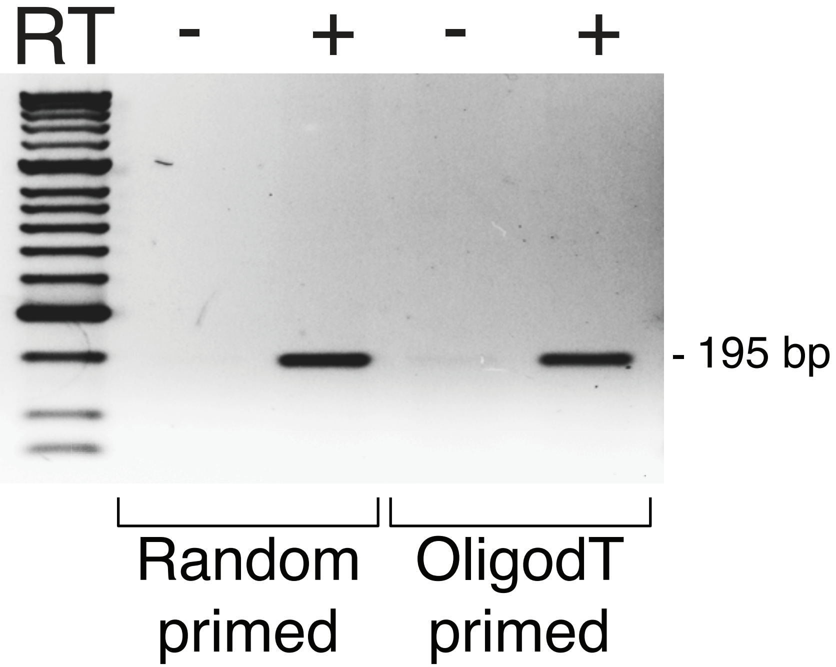
Supplemental Figure 18. Occurrence of CTCF DNA motifs within the DJ. Upper panel shows a CTCF DNA motif discovered from DNA sequences associated with the ChIP-seq peaks of CTCF across the DJ. The lower panel shows a known CTCF motif model (obtained from JASPAR) that is matched against the CTCF DNA motif obtained from ChIP-seq data that is shown in the upper panel.



Supplemental Figure 19. Transcription profiling of the DJ. The top four tracks show ChIP-seq signals of four chromatin marks (TAF1, RNA PolII, H3K4me3, and H3K36me3); and the bottom tracks show the structures of all DJ transcripts assembled from RNA-seq, mRNA, and EST data. Transcripts originating from promoters at 138 kb, 187 kb and 238 kb in the DJ (boxed in green) were validated.



Supplemental Figure 20. Quantitation of DJ transcript levels. Transcript abundances for disnor187 and disnor238 were estimated from RNA-seq data, measured as FPKM (Fragments Per Kilobase of transcript per Million mapped reads) from a variety of different cell types (bottom). This shows that disnor187 and disnor238 are transcribed at low to moderate levels.



Supplemental Figure 21. RTPCR confirms the presence of a transcript from the DJ 138 kb promoter. PCR was performed with cDNA prepared from HT1080 total RNA primed with either random hexamers or Oligo dT. Sequencing of the expected 195 bp product confirmed that it is a precise match with the DJ contig.

Supplemental Data and Video Descriptions

Supplemental Data 1. Four matrices showing the % sequence identities and % gaps from pairwise comparisons between DJ-containing sequences, and between PJ-containing sequences. Interchromosomal and intrachromosomal comparisons are provided, using cosmids and BACs.

Supplemental Data 2. FASTA format file of the DJ contig sequence used in this study.

Supplemental Data 3. GFF format file of the features identified in the DJ contig sequence from this study.

Supplemental Video 1. Video showing the 3D arrangement of the DJ sequences in the nucleus. 3D-immuno FISH was performed on HT1080 cells. DJ sequences were detected using BAC CT476834 (green) and nucleoli were revealed using antibodies to UBF (red). Nuclei were stained with DAPI (blue). DJ and DAPI signals are merged

Supplemental Video 2. Video showing the 3D arrangement of the DJ sequences in the nucleus. 3D-immuno FISH was performed on HT1080 cells. DJ sequences were detected using BAC CT476834 (green) and nucleoli were revealed using antibodies to UBF (red). Nuclei were stained with DAPI (blue). DJ and UBF signals are merged.