nature | methods

Characterization of missing human genome sequences and copy-number polymorphic insertions

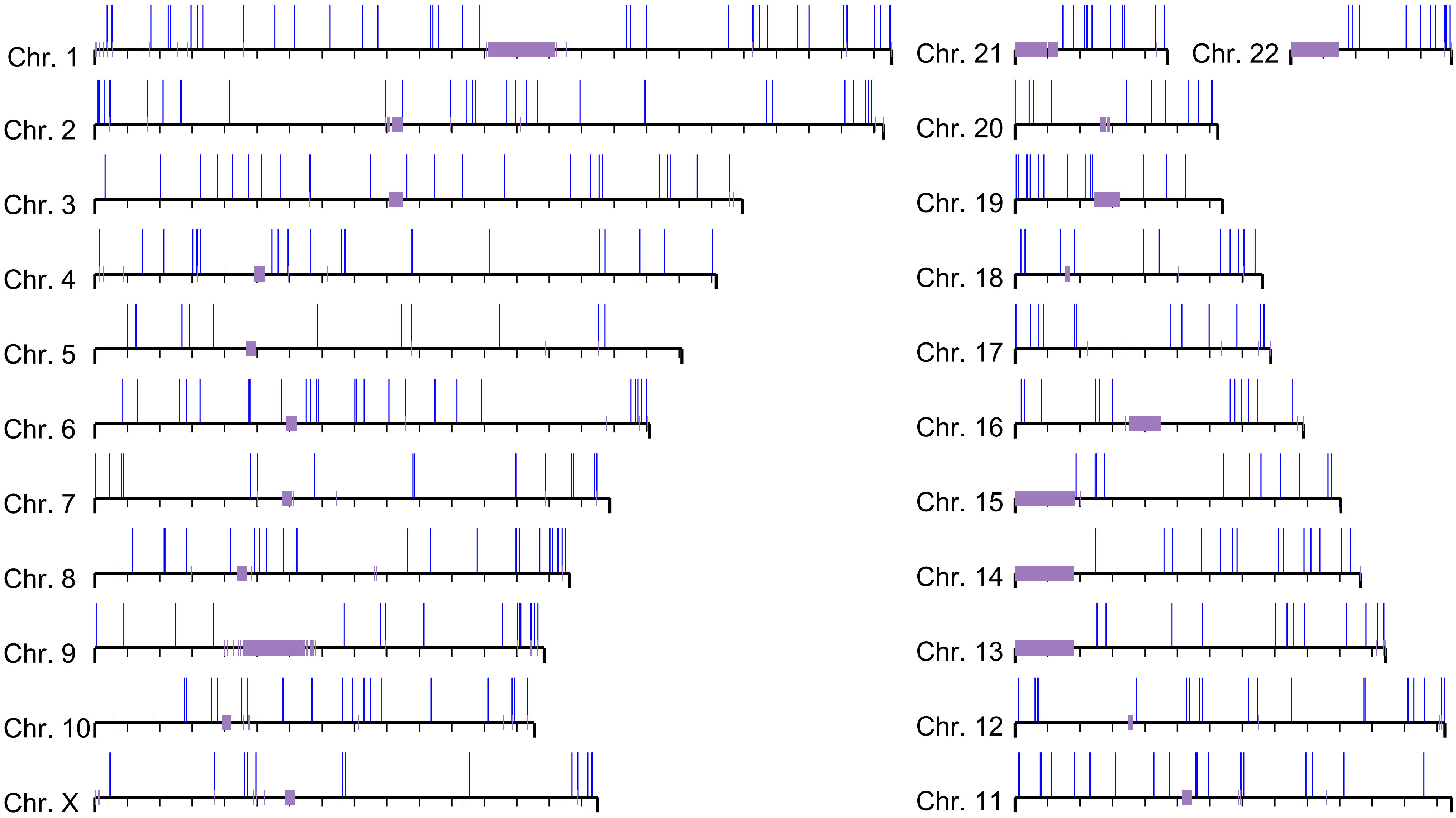
Jeffrey M Kidd, Nick Sampas, Francesca Antonacci, Tina Graves, Robert Fulton, Hillary S Hayden, Can Alkan, Maika Malig, Mario Ventura, Giuliana Giannuzzi, Joelle Kallicki, Paige Anderson, Anya Tsalenko, N Alice Yamada, Peter Tsang, Rajinder Kaul, Richard K Wilson, Laurakay Bruhn & Evan E Eichler

Supplementary figures and text:

Supplementary Note	Detailed descriptions of methods
Supplementary Figure 1	Genomic distribution of novel insertions
Supplementary Figure 2	FISH mapping of NA15510 assembled contigs
Supplementary Figure 3	Distribution of FST values for novel sequence contigs
Supplementary Figure 4	Comparison of VST and FST
Supplementary Figure 5	Annotated images of sequenced insertions
Supplementary Figure 6	Size distribution for sequenced insertions
Supplementary Figure 7	Distribution of constraint for conserved elements
Supplementary Table 2	Map locations of anchored loci
Supplementary Table 3	FISH analysis of orphan clones
Supplementary Table 5	Noise-multiplier results
Supplementary Table 9	Novel insertions with high VST
Supplementary Table 11	Composition of sequenced insertions
Supplementary Table 12	Comparison of sequenced insertions with GRCh37

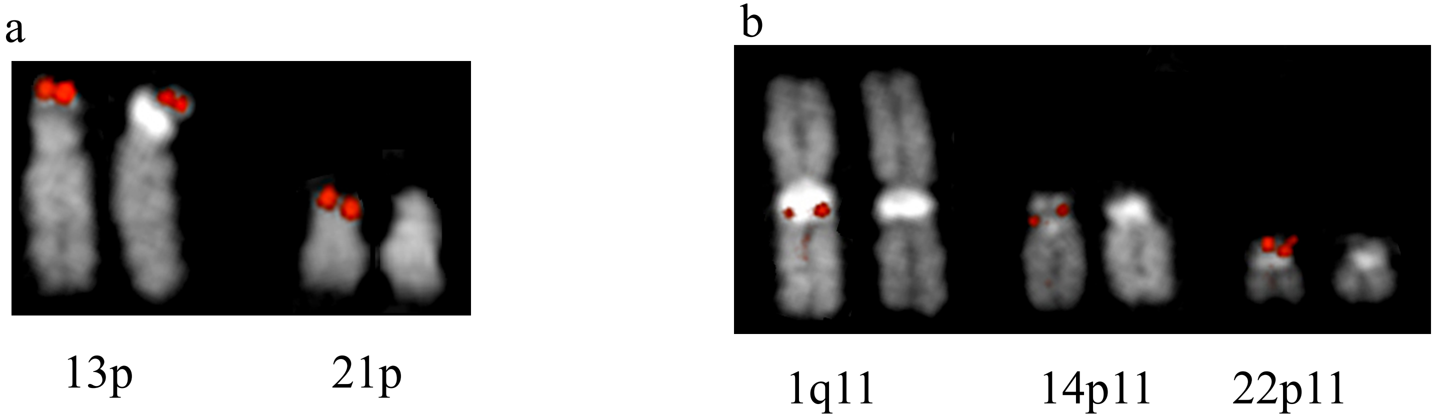
Note: Supplementary Tables 1, 4, 6–8, 10, 12 and 14 are available on the Nature Methods website.

Nature Methods: doi:10.1038/nmeth.1451

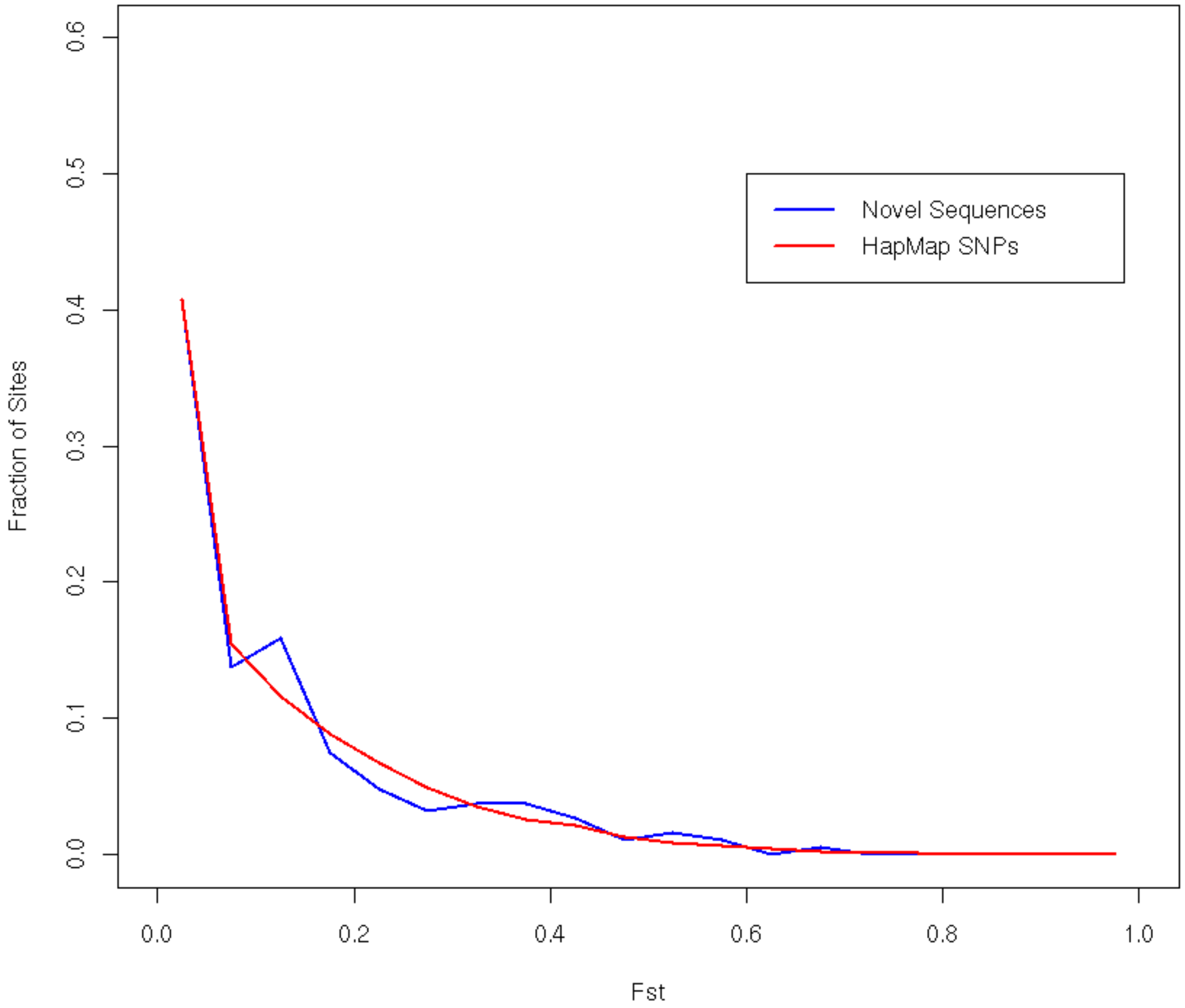


Supplementary Figure 1 Genomic distribution of novel insertions
The diagram depicts the locations (blue lines) of 400 new insertion loci mapped to the human genome (build35) by one-end anchored end-sequence placements. Purple boxes represent locations of known gaps.

Nature Methods: doi:10.1038/nmeth.1451

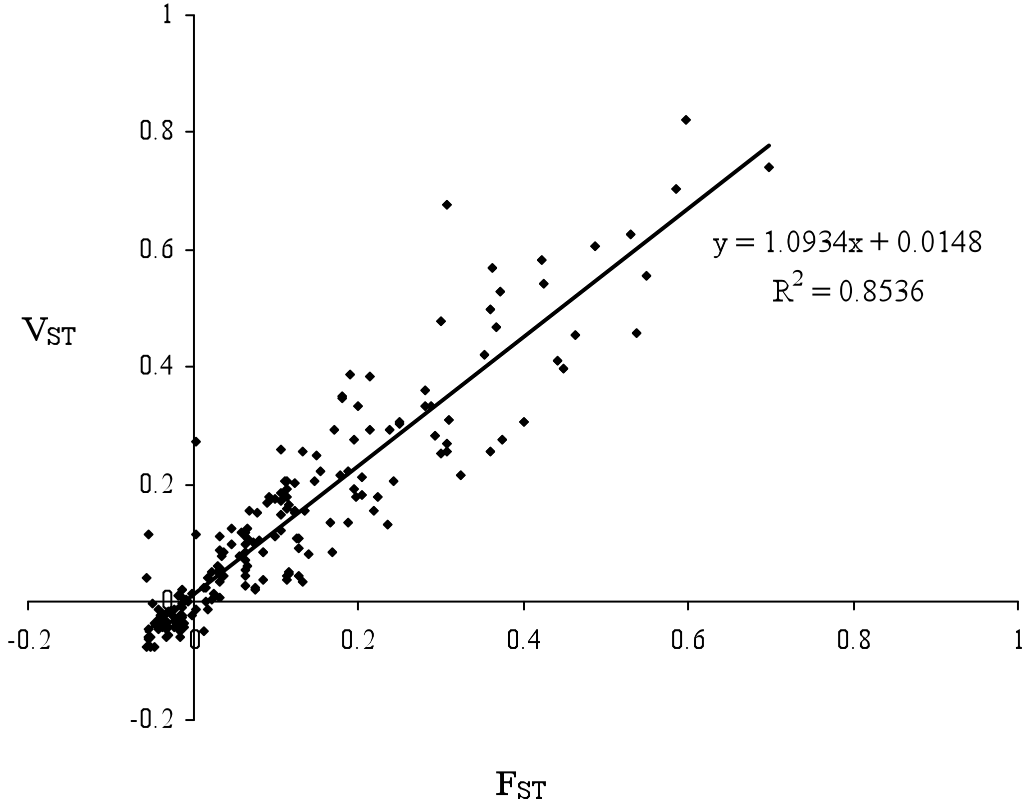


Supplementary Figure 2 FISH mapping of NA15510 assembled contigs
(a) Contig #74 (probe WIBR2-3212B04) maps to 13p and 21p. (b) Contig #140 (probe WIBR2-1011K06) maps to 1q11, 14p11, and 22p11. Note that in both cases hybridization does not occur on each homologous chromosome, indicating that the contigs are copy-number polymorphic.



Supplementary Figure 3 Distribution of F_{ST} values for novel sequence contigs
Global F_{ST} was calculated for 189 loci with contigs that form bi-allelic genotypes (blue line), as well as for 2,122,433 HapMap SNPs that are polymorphic in the same individuals (red line).

Nature Methods: doi:10.1038/nmeth.1451



Supplementary Figure 4 Comparison of V_{ST} and F_{ST}
Values are shown for one contig from each of the 189 loci used in Supplementary Figure 3

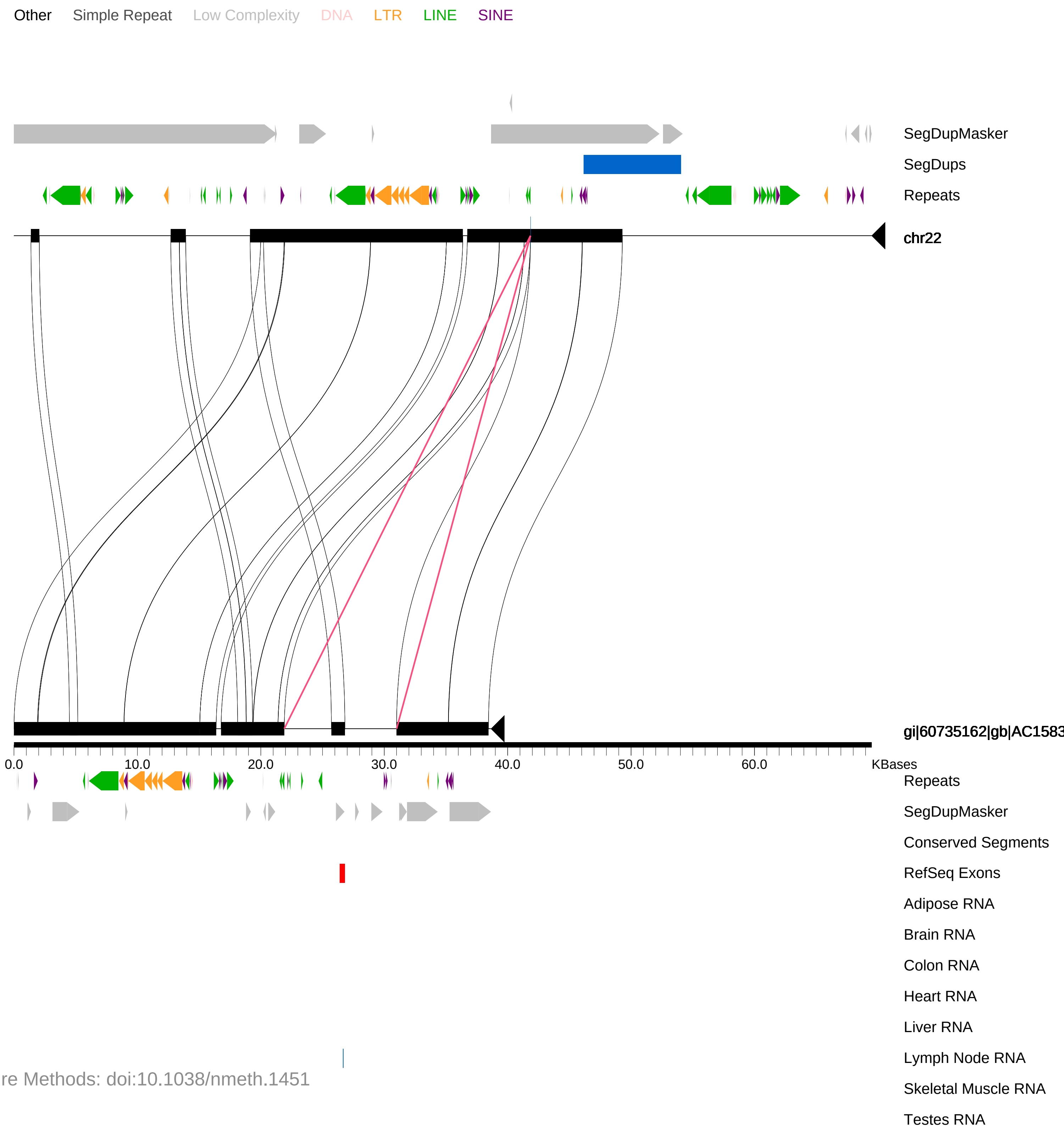
Nature Methods: doi:10.1038/nmeth.1451

Supplementary Figure 5 Annotated images of sequenced insertions

The sequence of each fosmid insert (lower black line) is compared against the build36 genome assembly (upper black line). Black boxes and lines connect matching sequence segments. The magenta lines indicate the breakpoints determined by sequence alignment. When applicable, yellow boxes indicate the extent of matching sequence on each side of the insertion, and the blue box indicates uncertainty in position of the breakpoint on the chromosome sequence. Common repeats (RepeatMasker) and predicted (DupMasker) and annotated duplications are depicted as indicated. The positions of RefSeq exons are shown in red above each chromosome. Additional annotation is located below each clone sequence. These annotations were created specifically for the inserted sequence, not for the clone as a whole. The annotations correspond to conserved segments (green), matching hits from the RefSeq database (red), and regions containing three or more mRNA-seq reads obtained from Wang et al. (blue). Only mRNA-seq reads that do not map against the build36 genome were considered. Conserved segments and RefSeq exon matches were only determined for the portion of the clone that represents the insertion relative to build36.

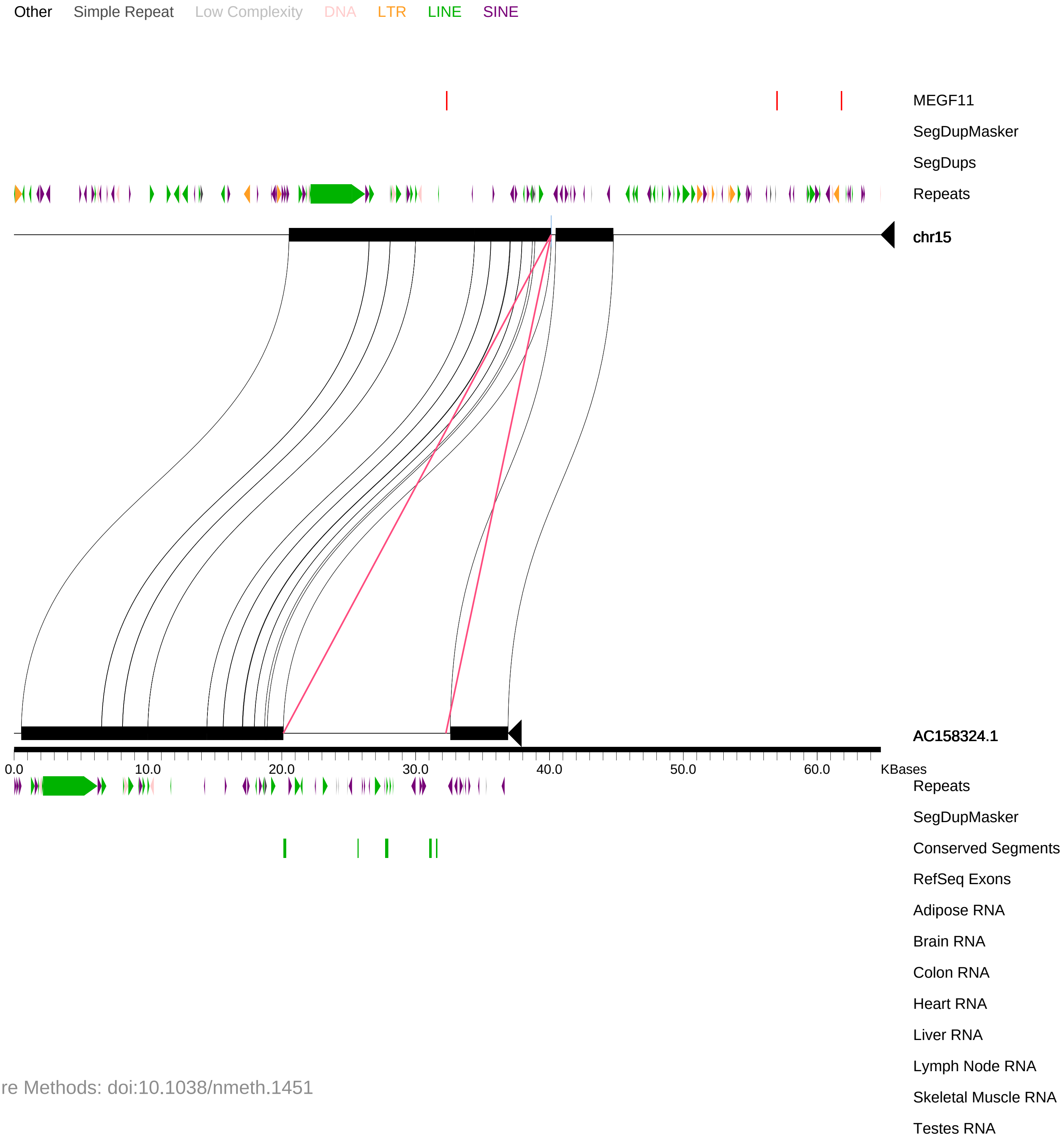
Clone file = AC158320.fa

Insertion Size: 9105



Clone file = AC158324.rc.fa

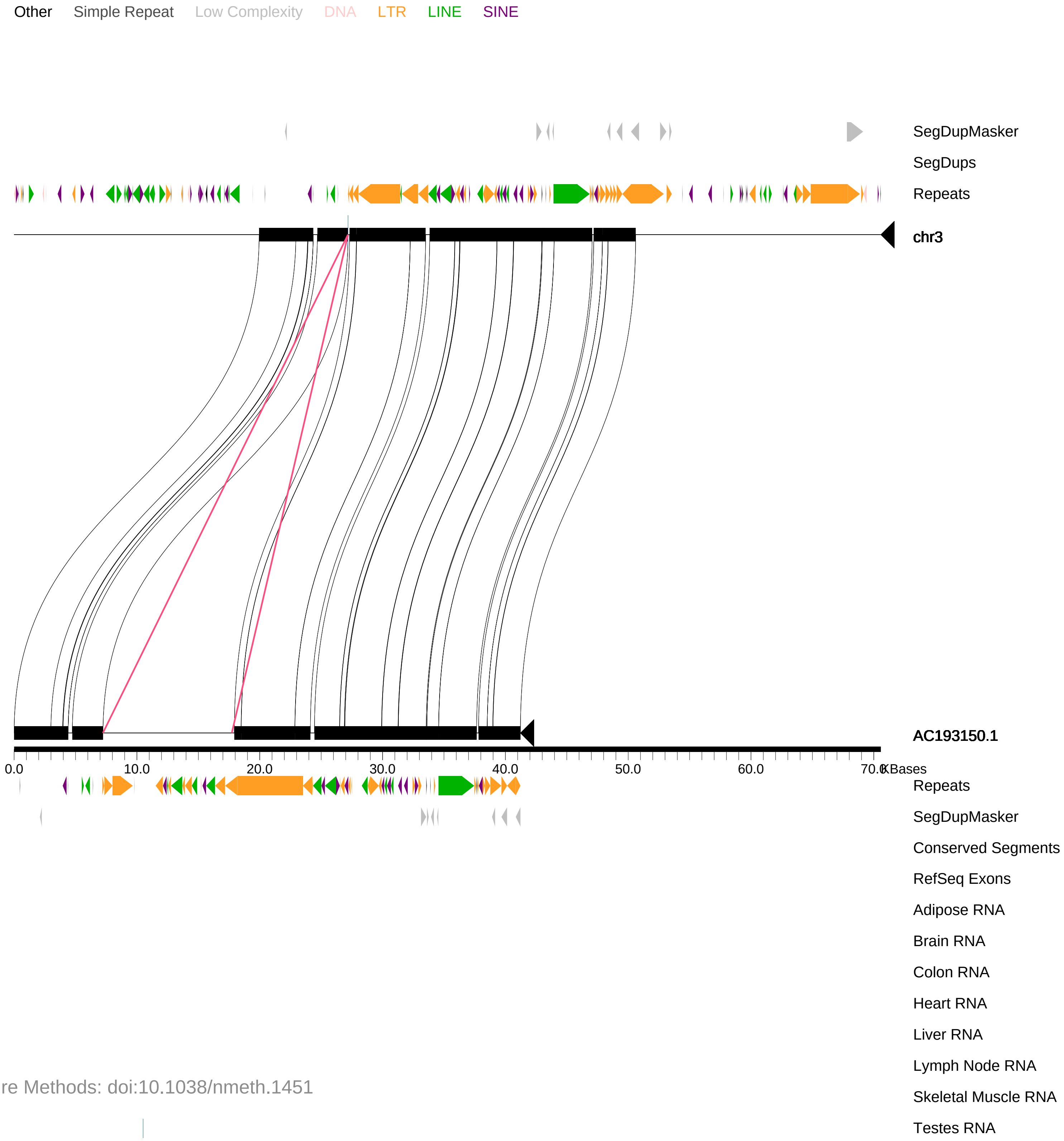
Insertion Size: 12150



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC193150.rc.fa

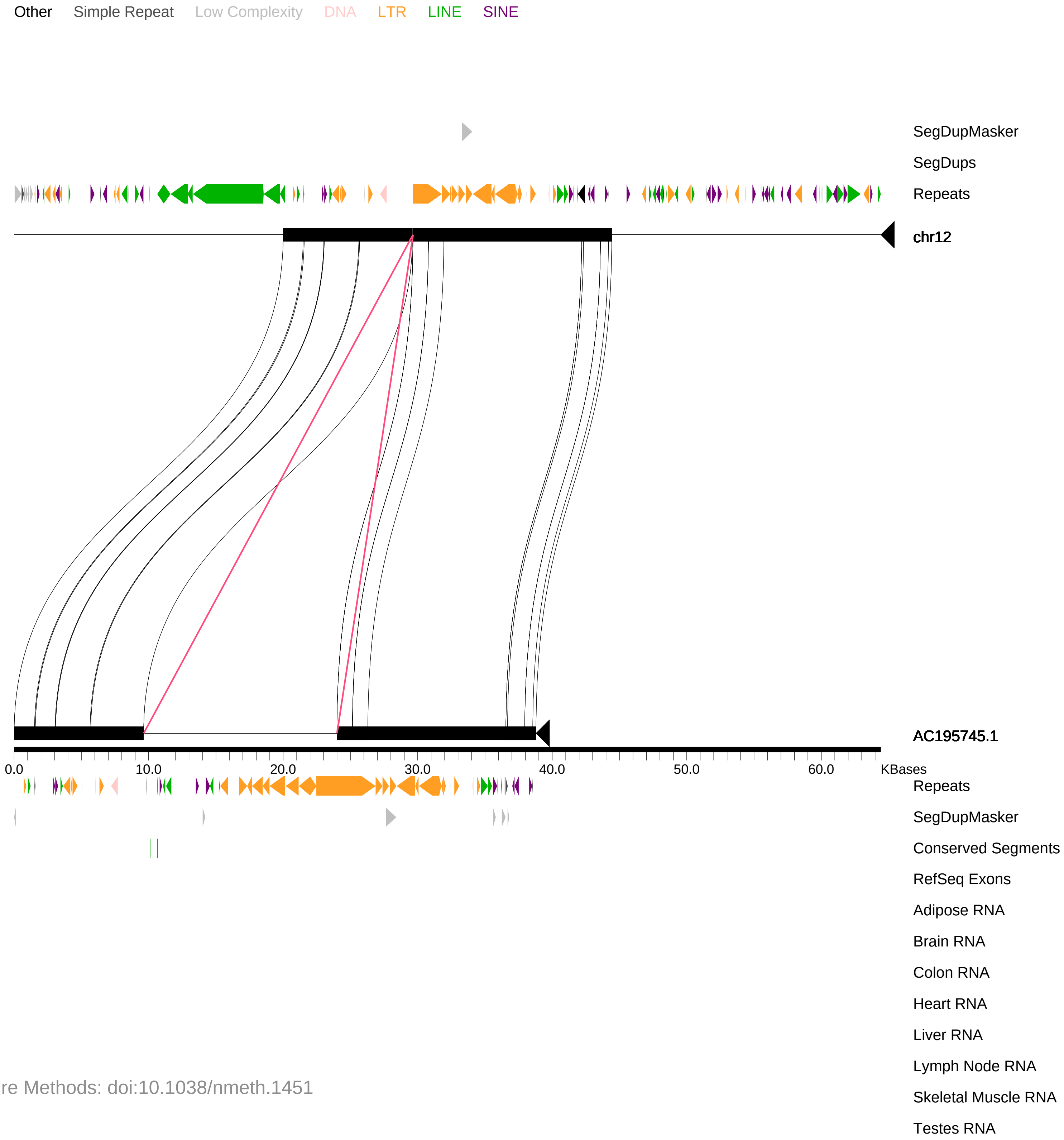
Insertion Size: 10496



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC195745.rc.fa

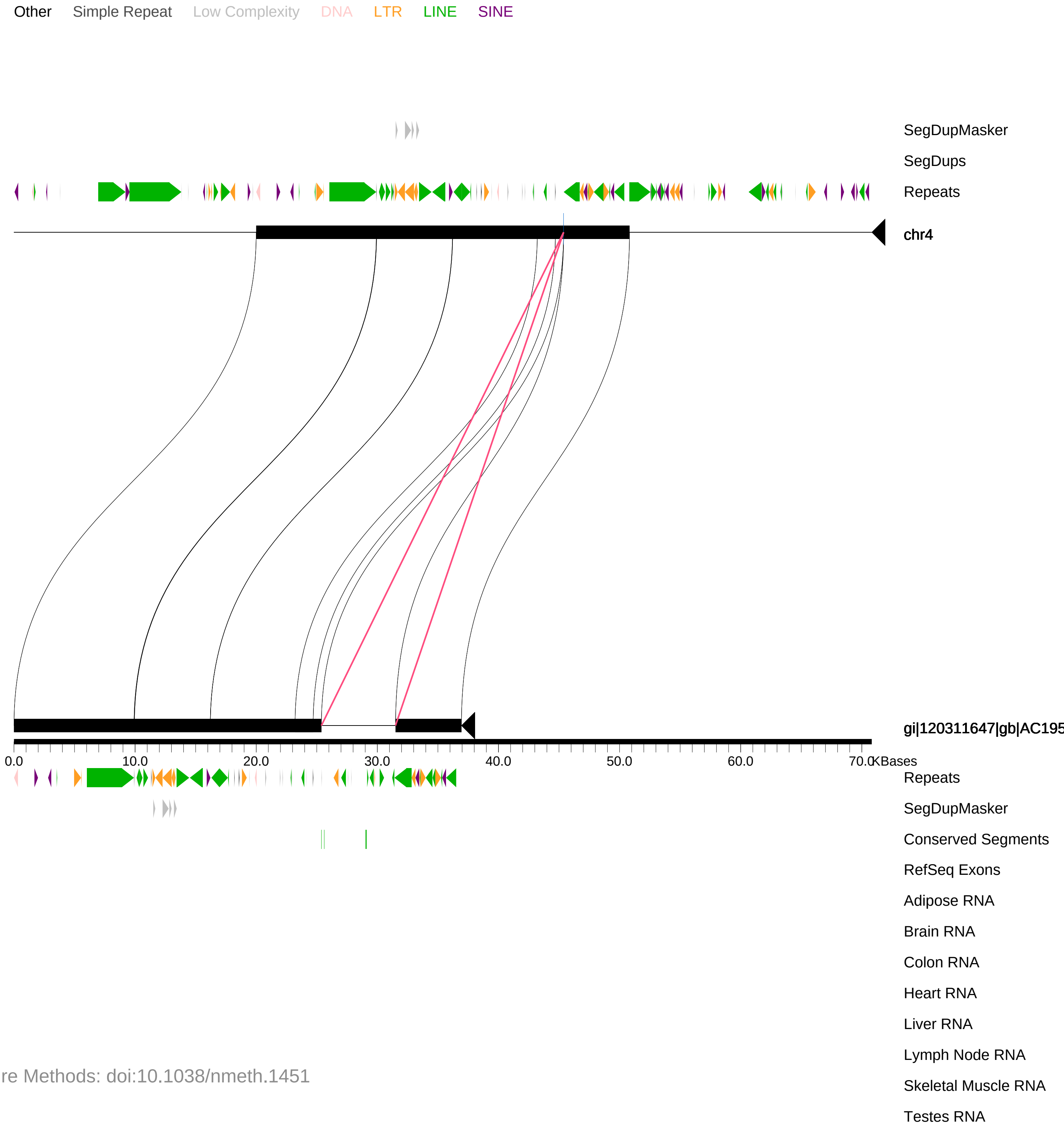
Insertion Size: 14370



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC195766.fa

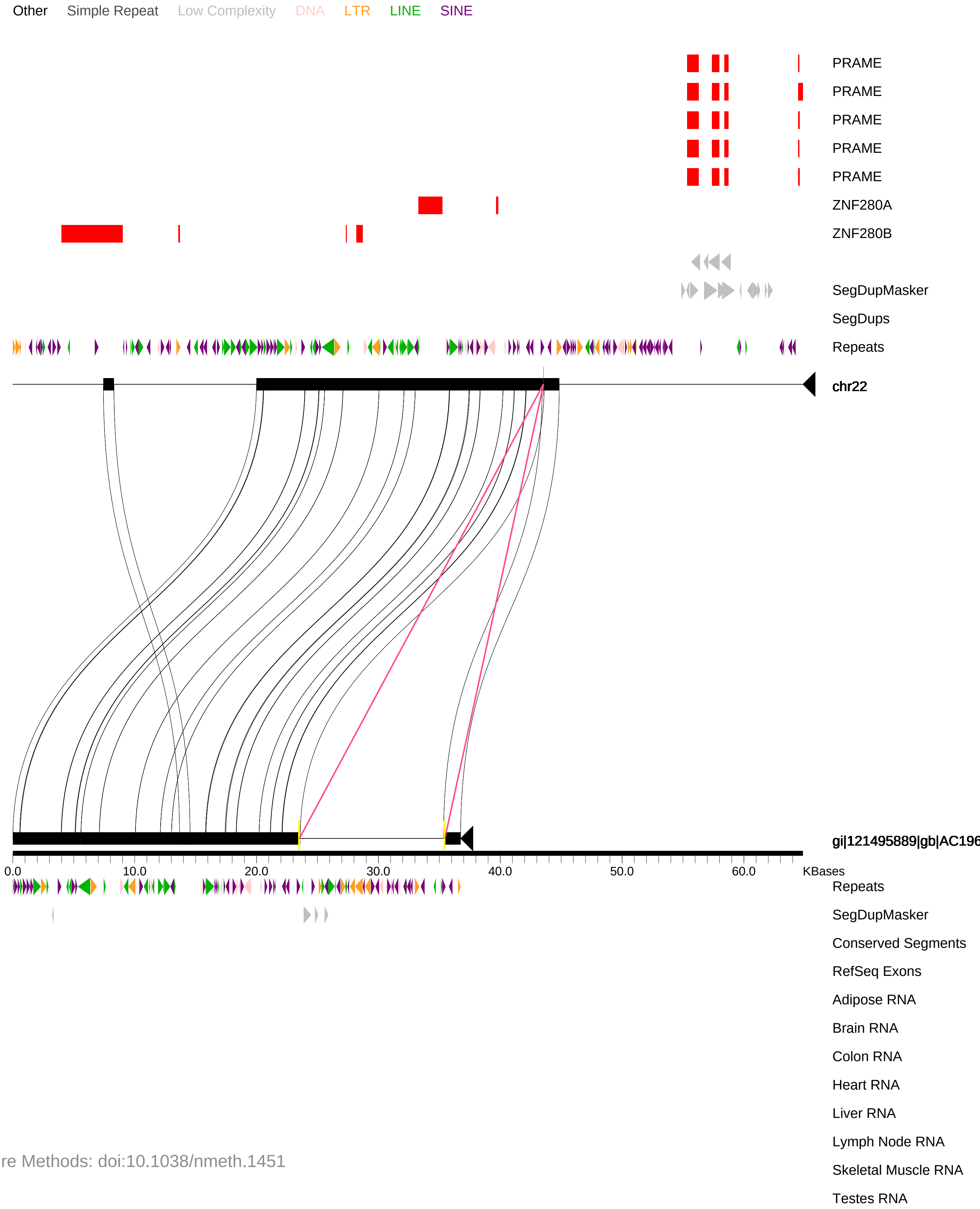
Insertion Size: 6117



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC196513.fa

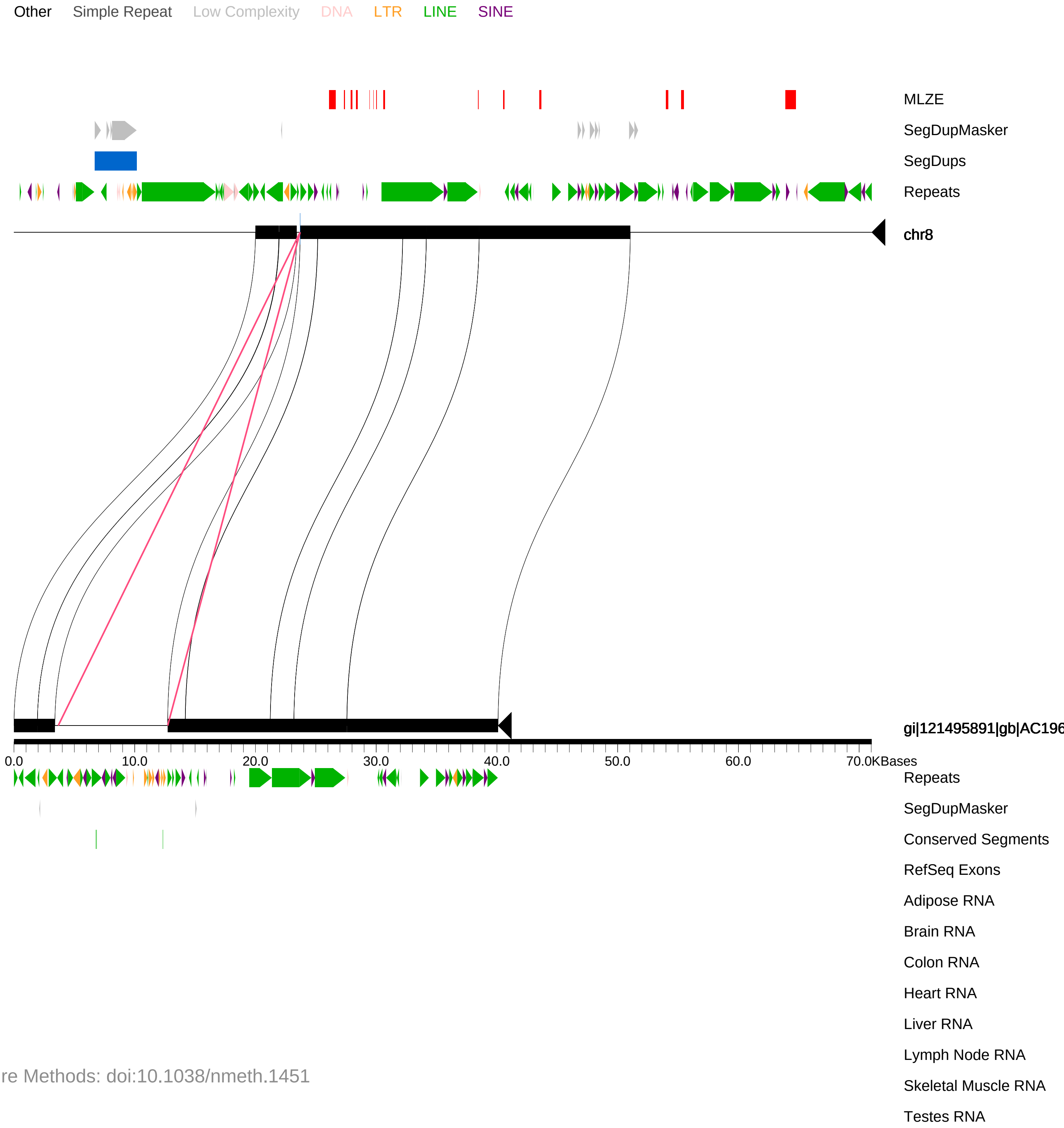
Insertion Size: 11940



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC196515.fa

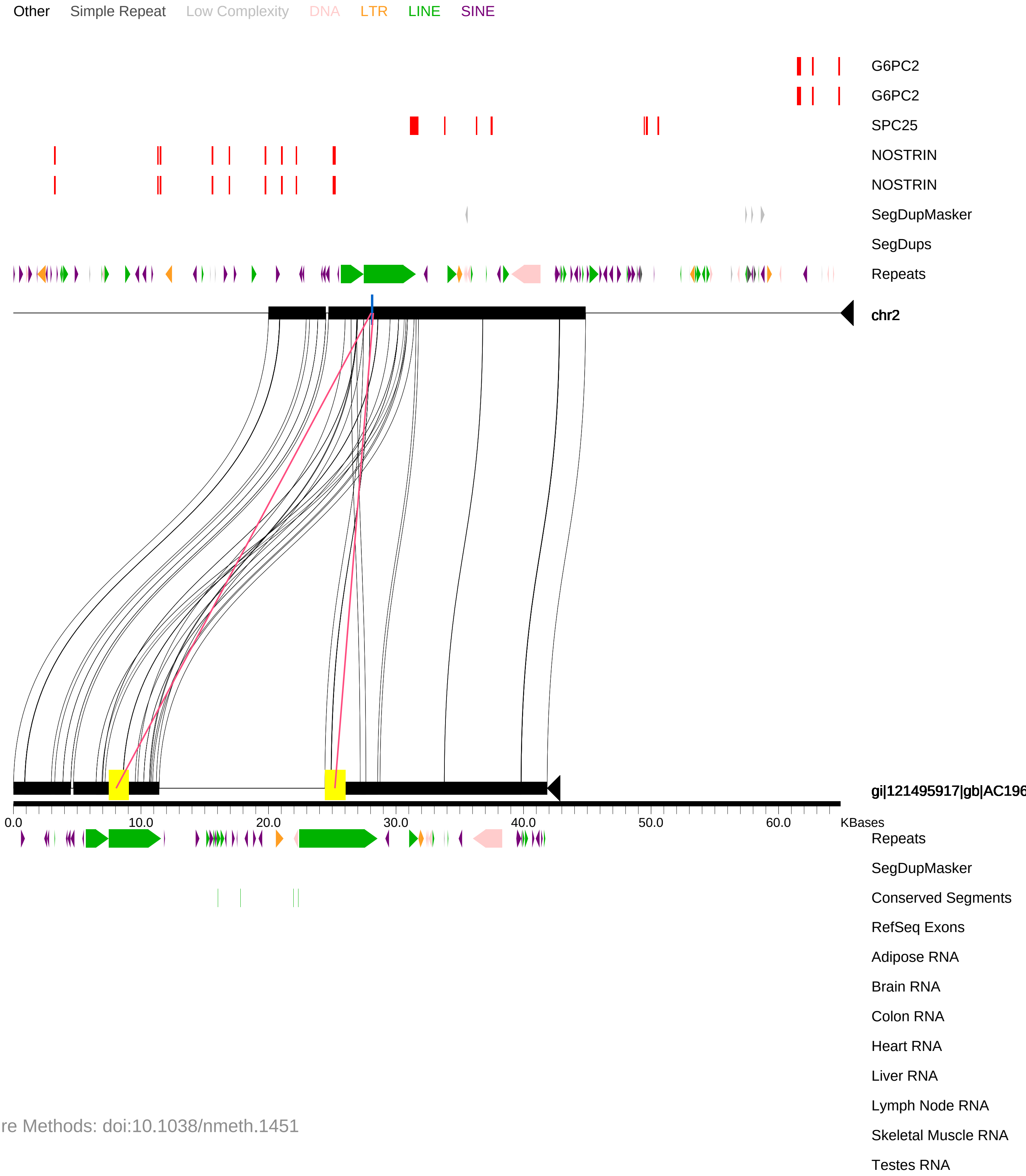
Insertion Size: 9064



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC196541.fa

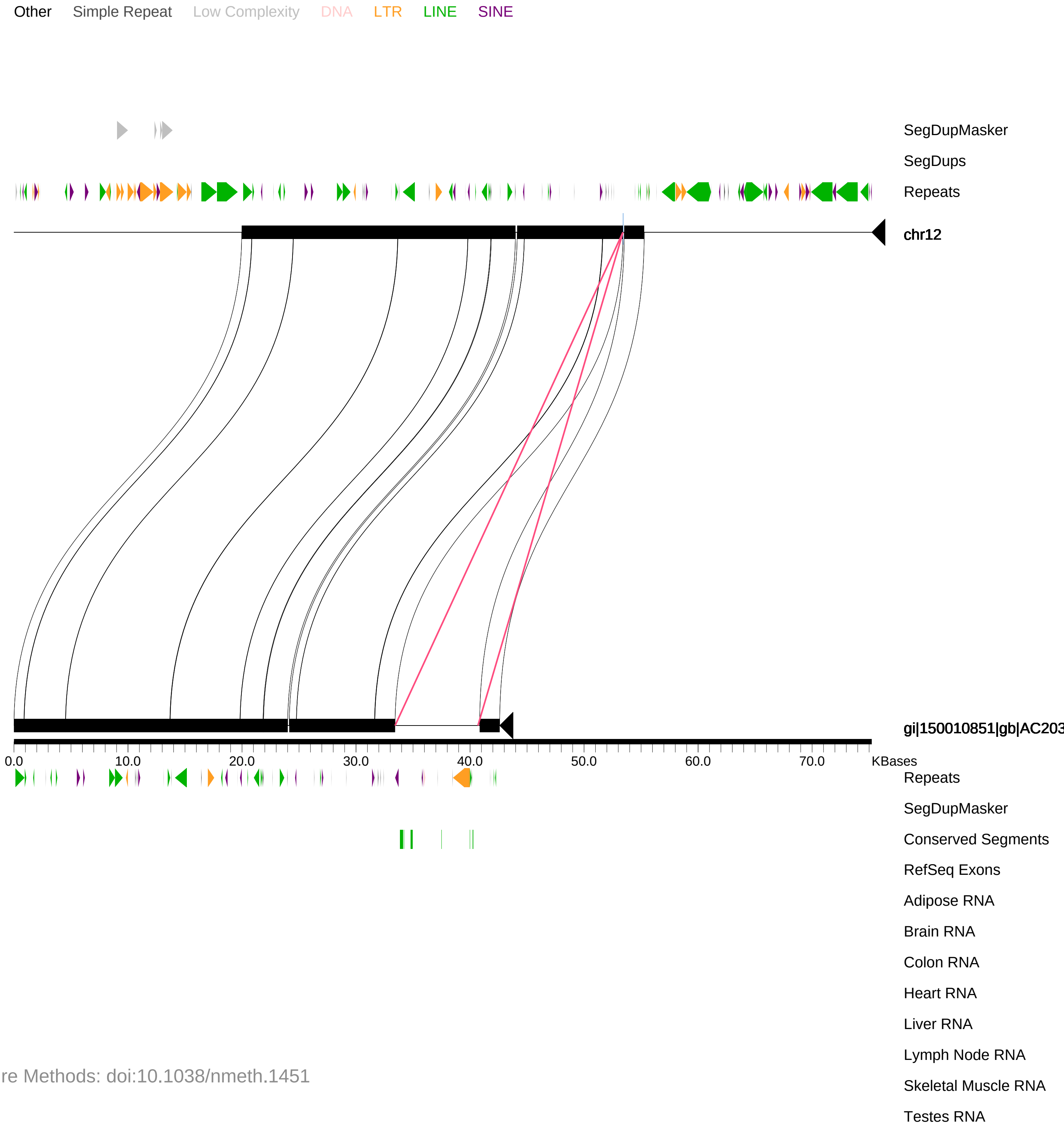
Insertion Size: 17157



Nature Methods: doi:10.1038/nmeth.1451

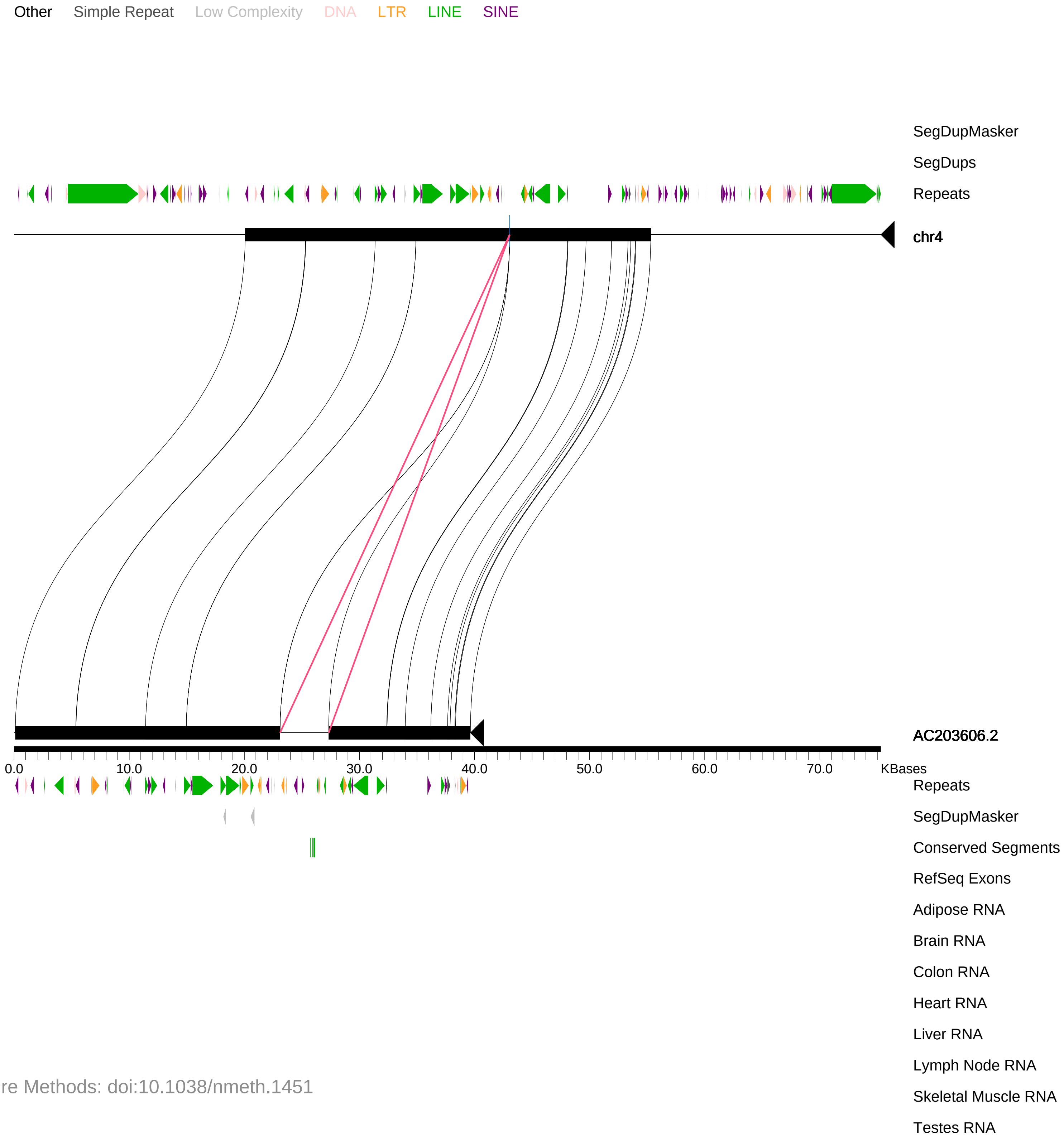
Clone file = AC203605.fa

Insertion Size: 7269



Clone file = AC203606.rc.fa

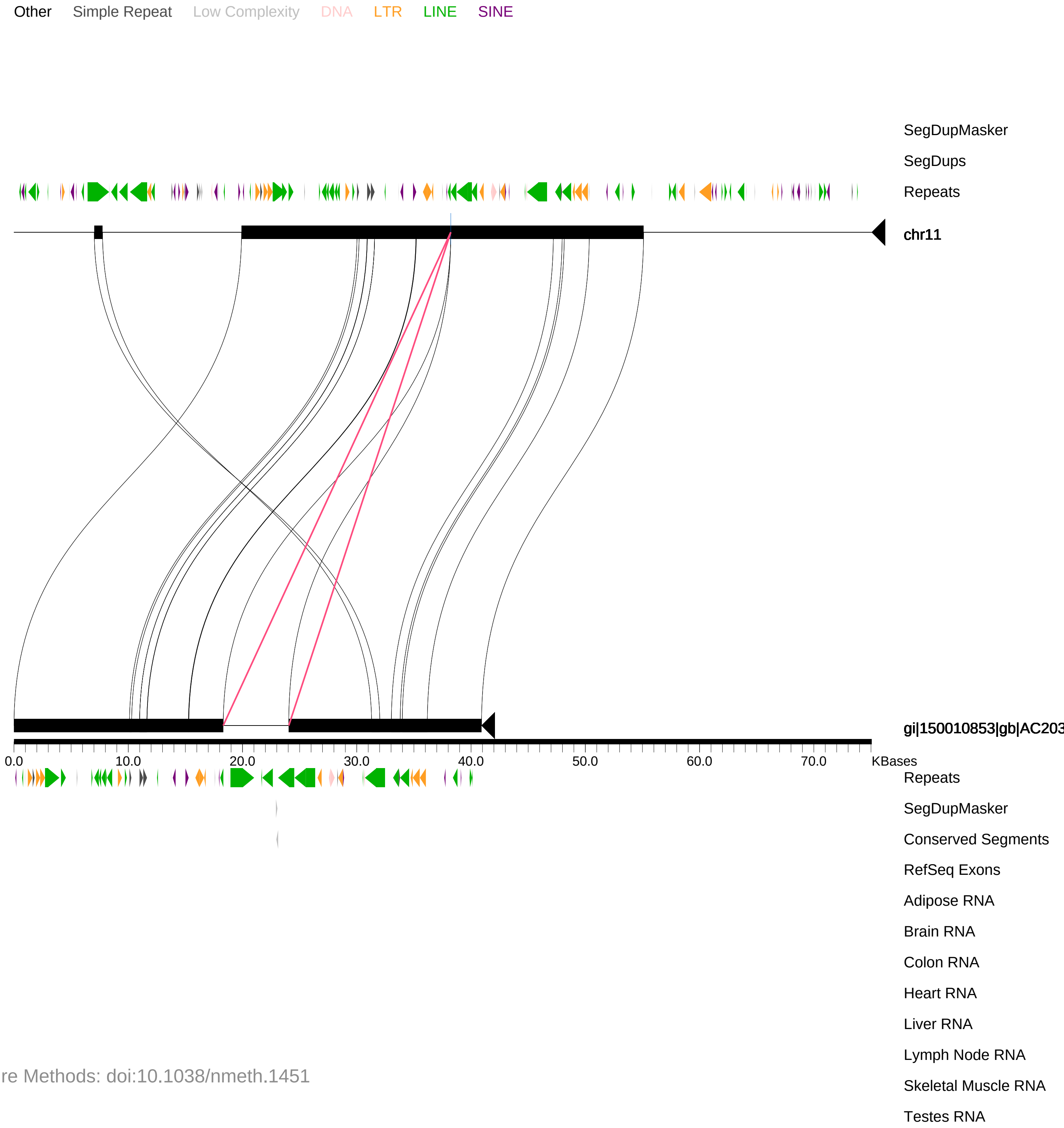
Insertion Size: 4231



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC203610.fa

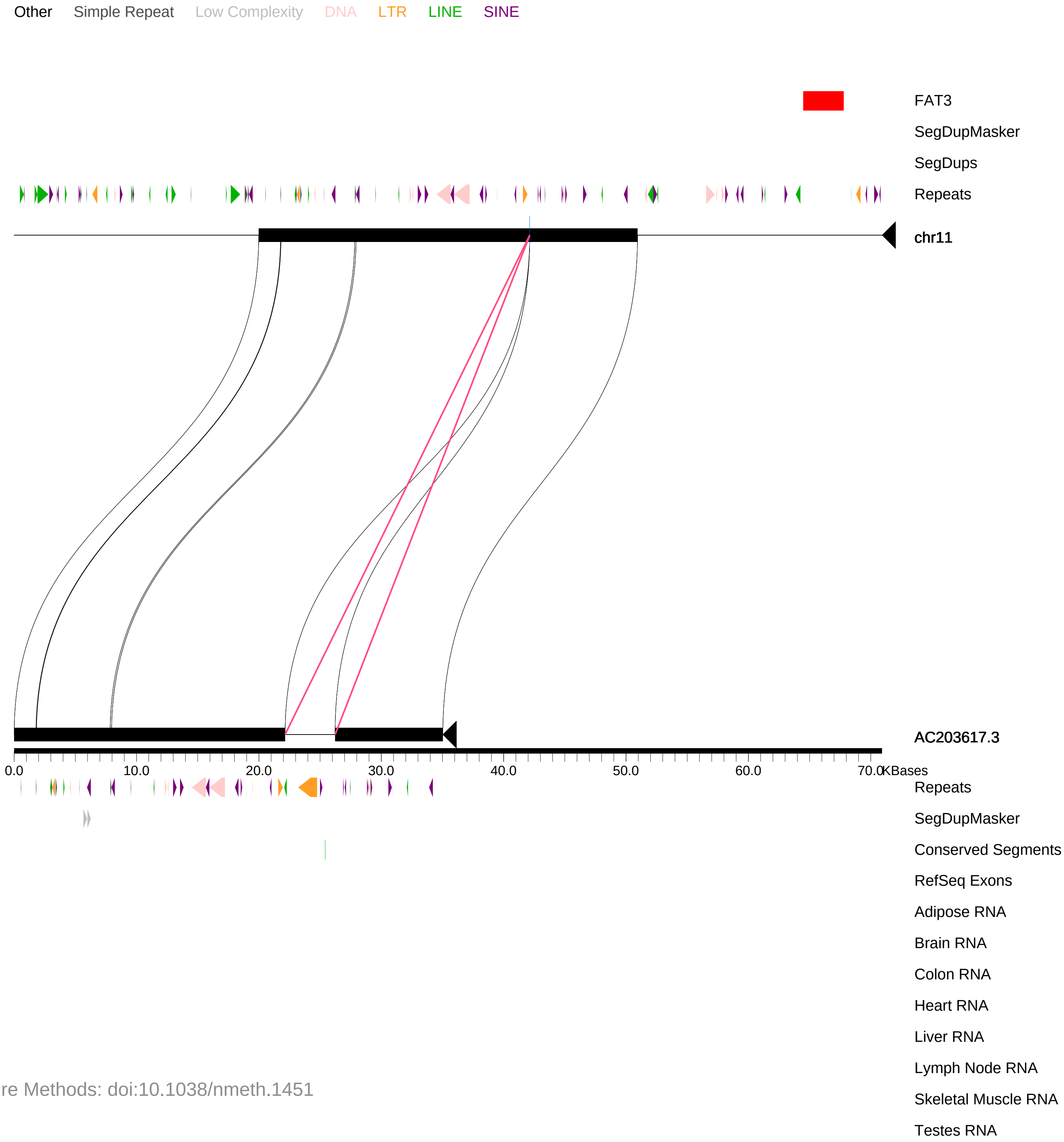
Insertion Size: 5738



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC203617.rc.fa

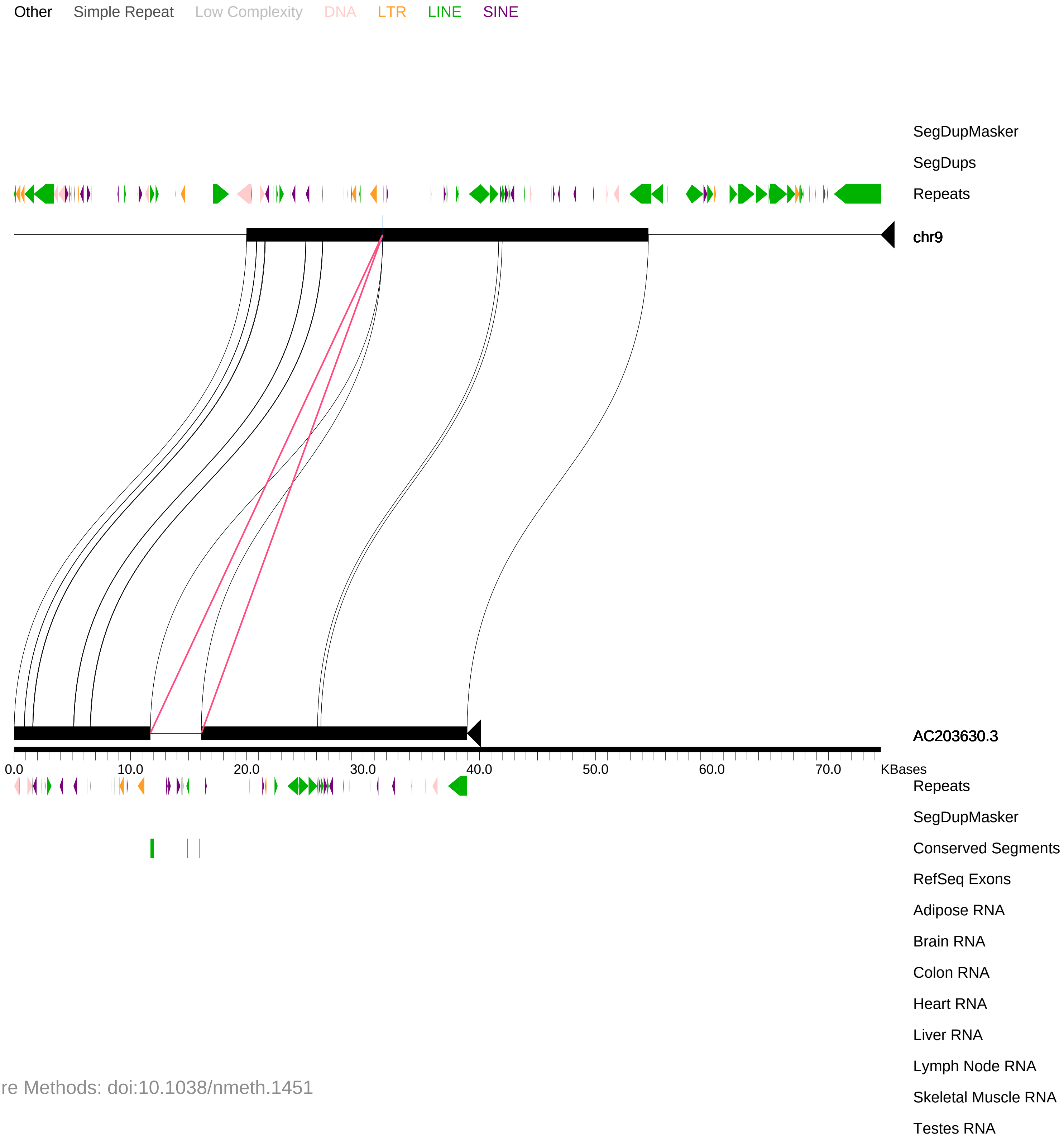
Insertion Size: 4087



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC203630.rc.fa

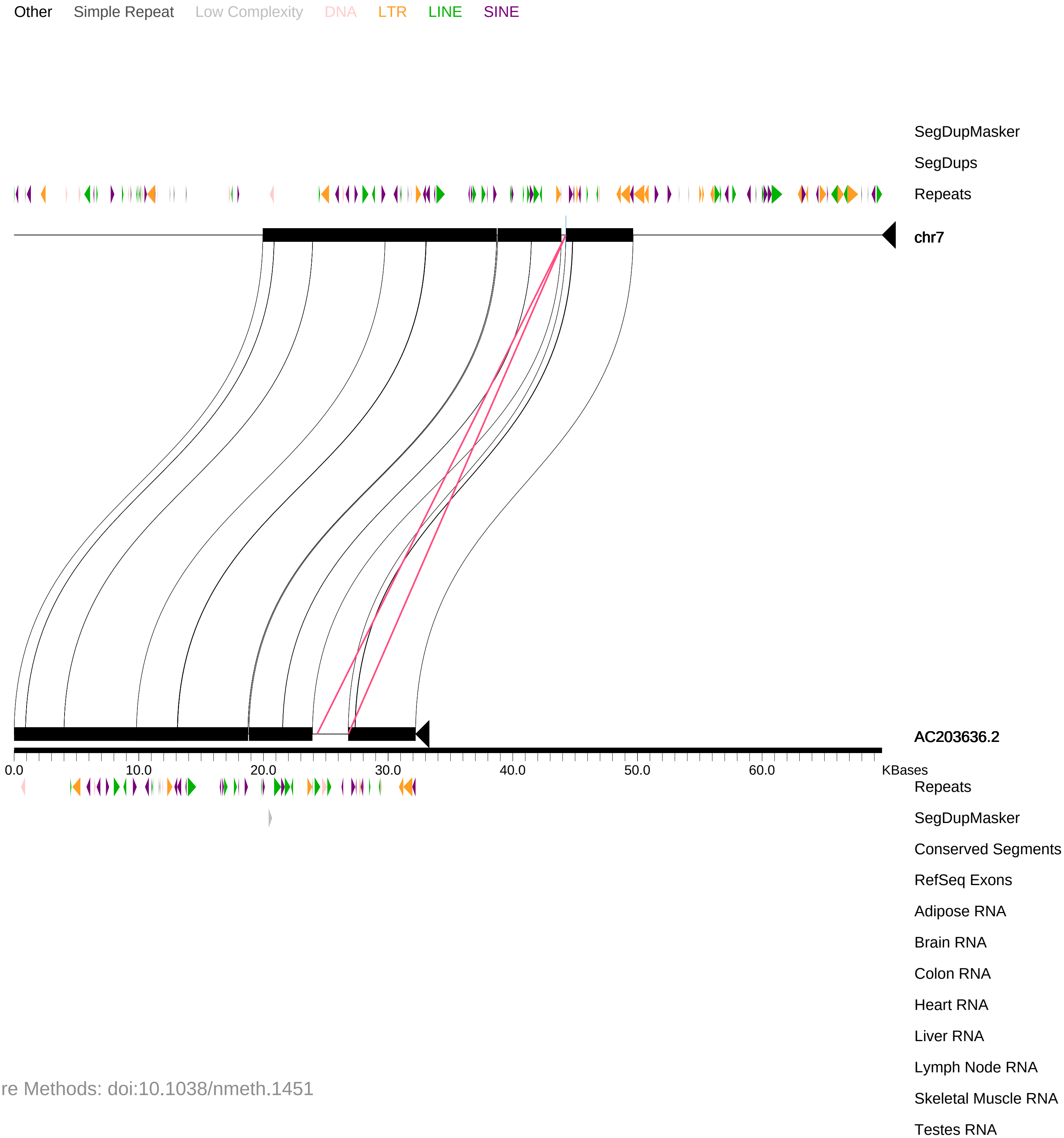
Insertion Size: 4385



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC203636.rc.fa

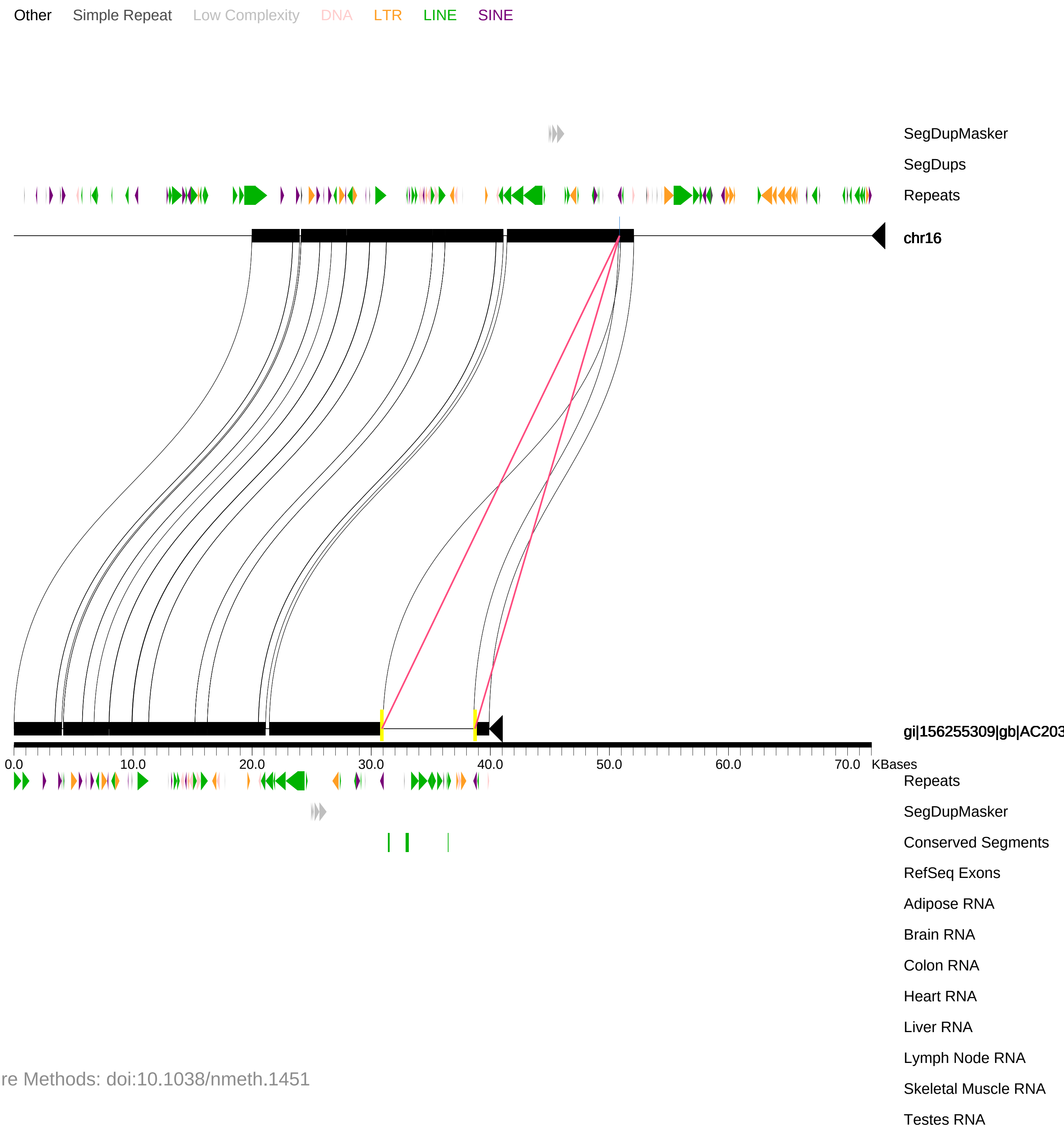
Insertion Size: 2500



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC203638.fa

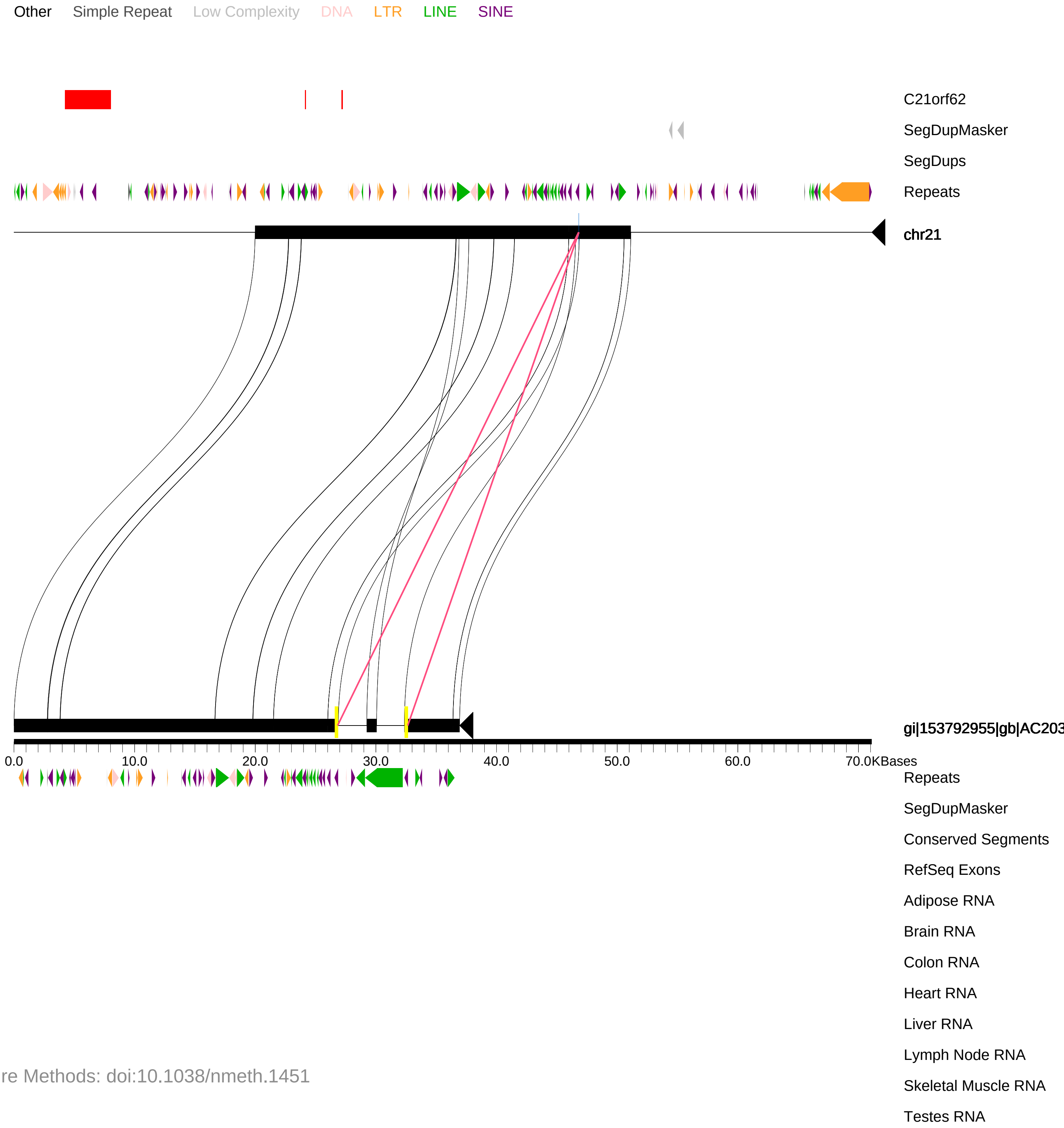
Insertion Size: 7822



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC203640.fa

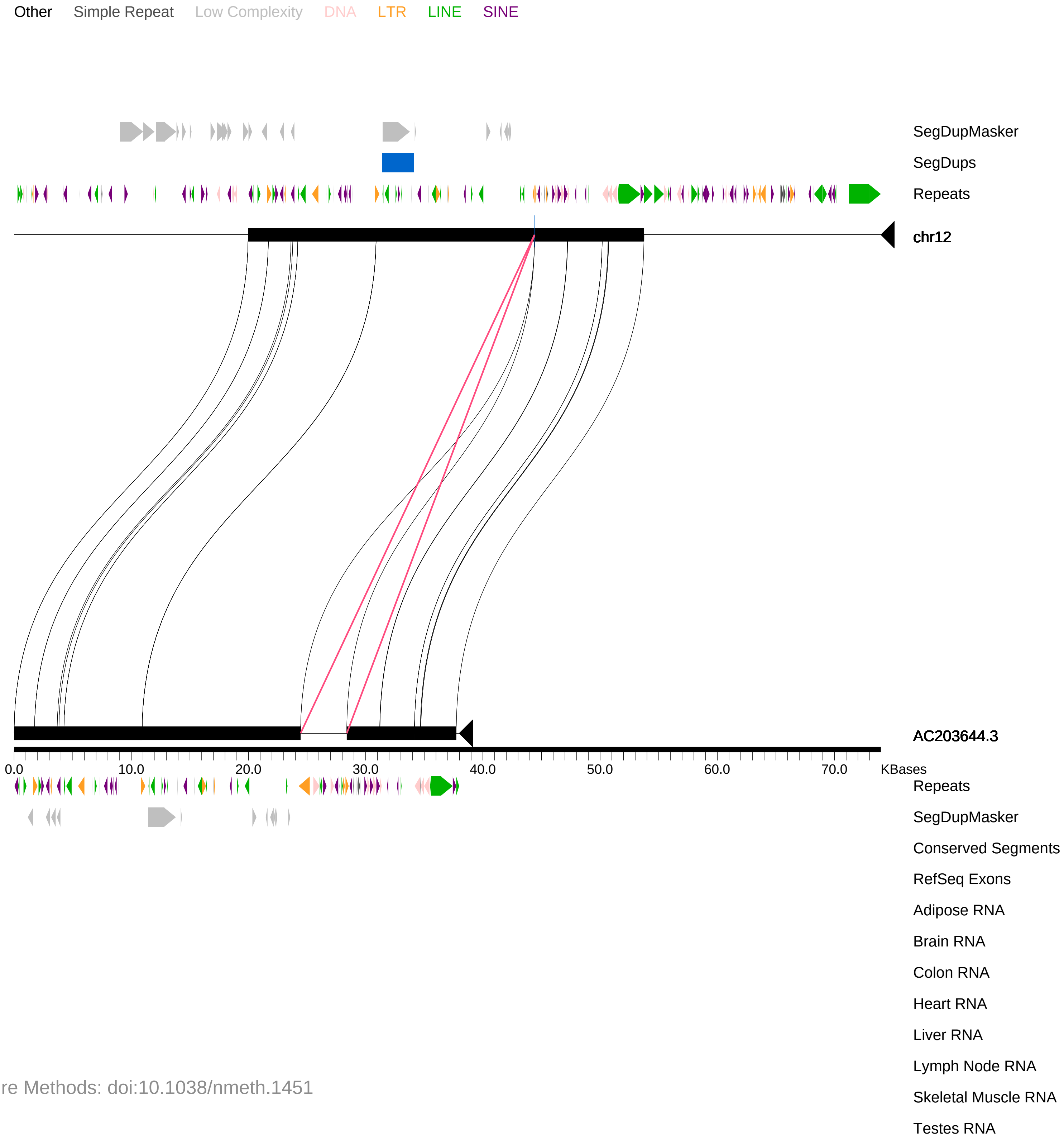
Insertion Size: 5817



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC203644.rc.fa

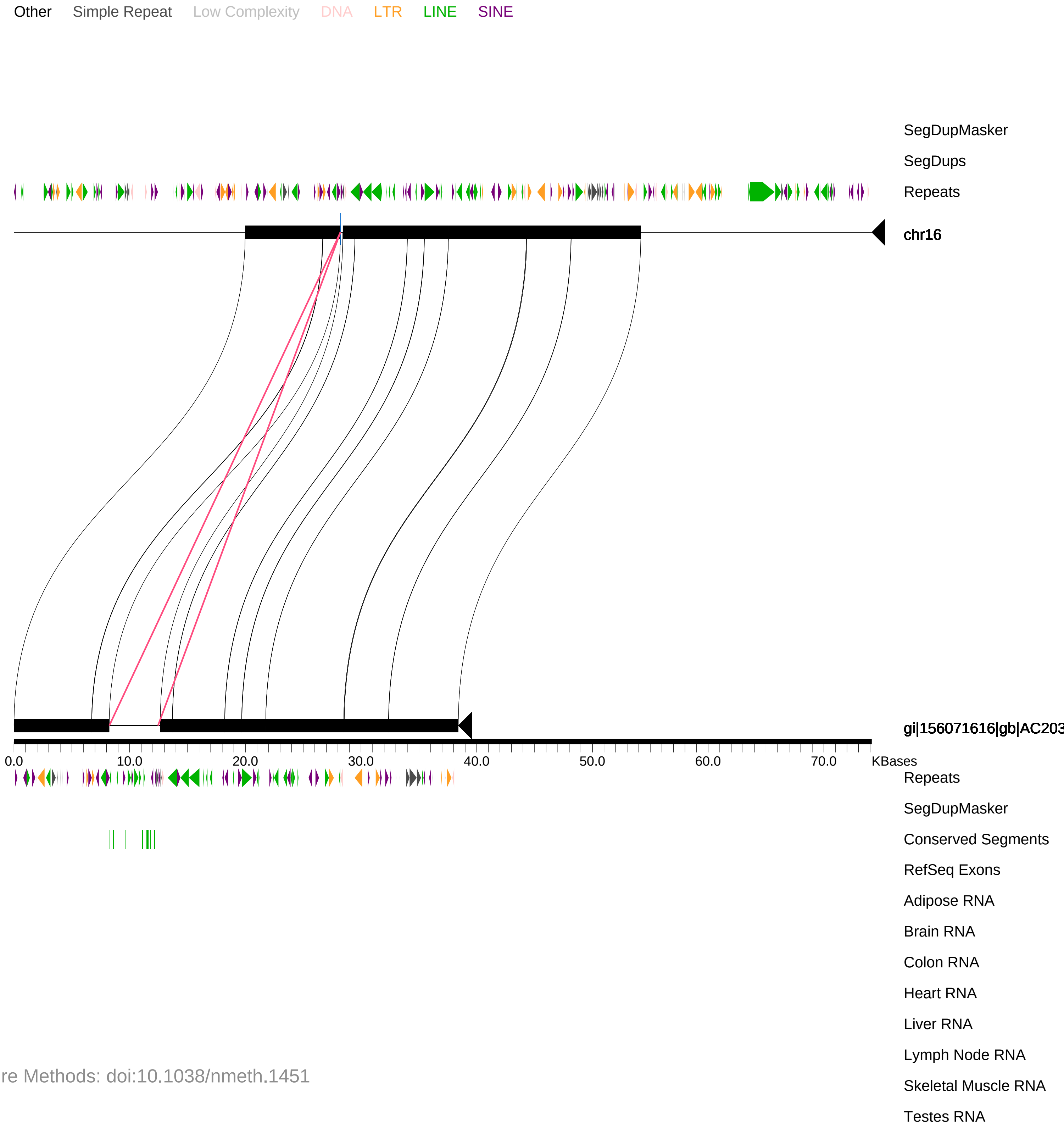
Insertion Size: 3946



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC203665.fa

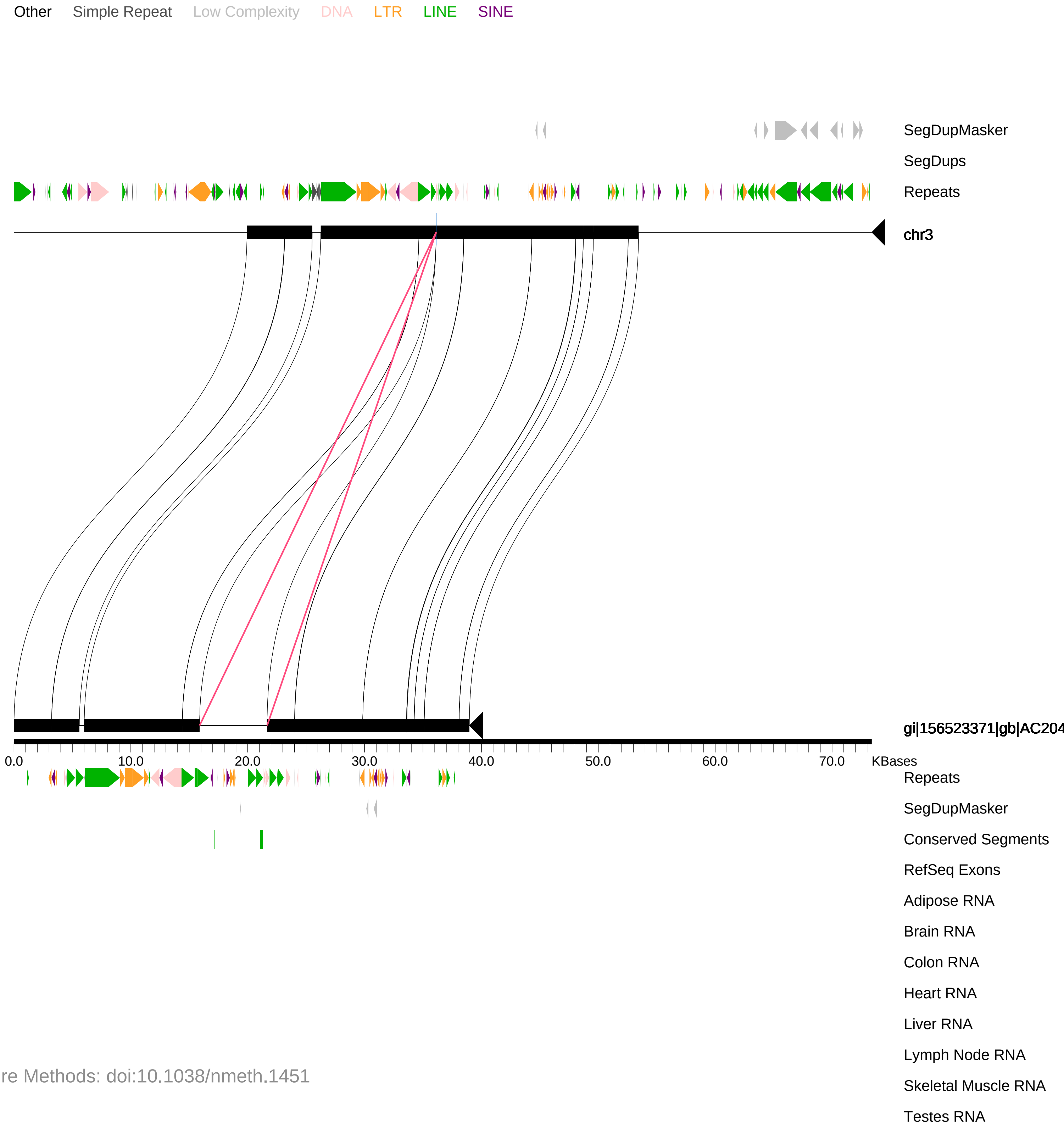
Insertion Size: 4213



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC204963.fa

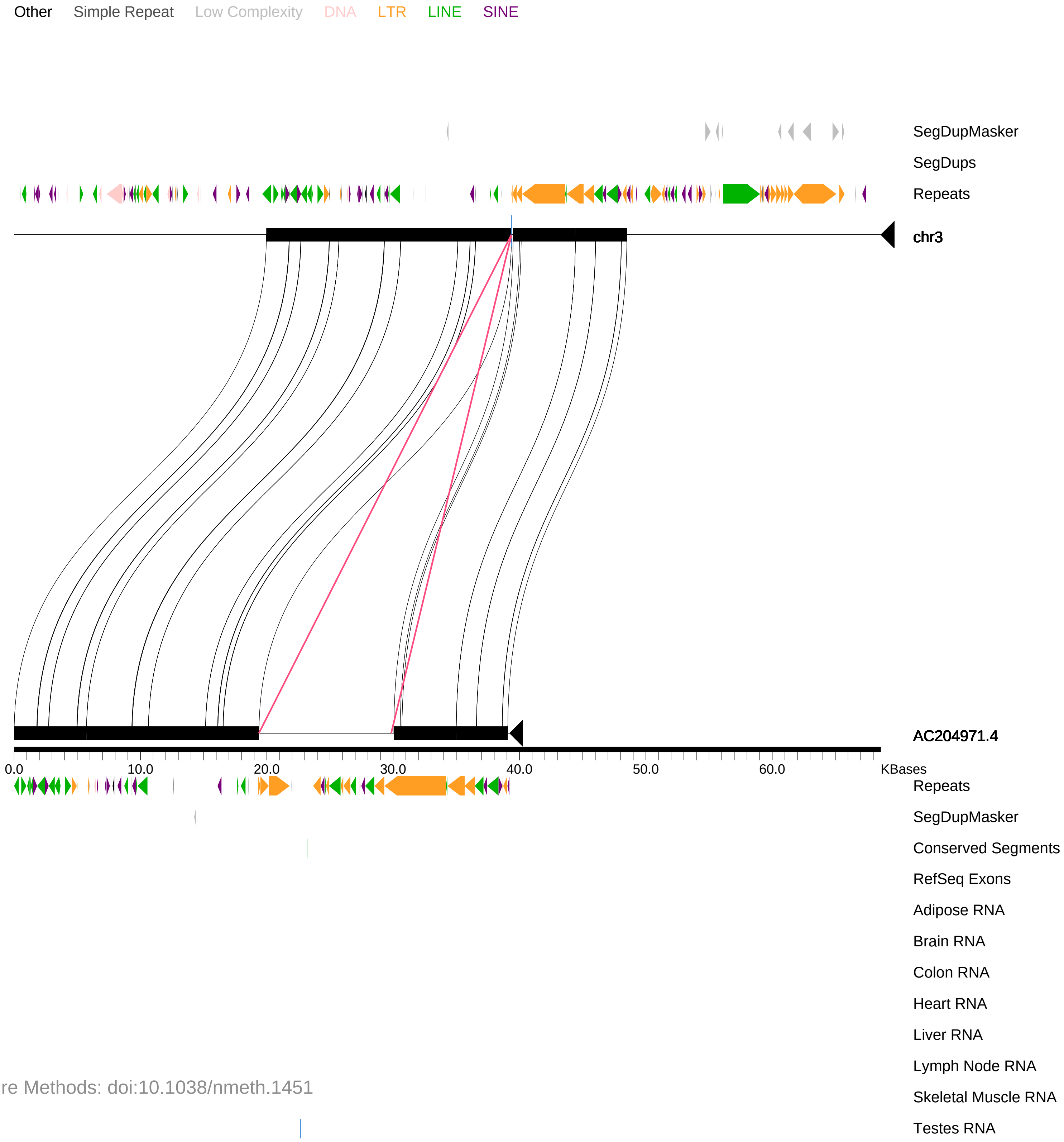
Insertion Size: 5779



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC204971.rc.fa

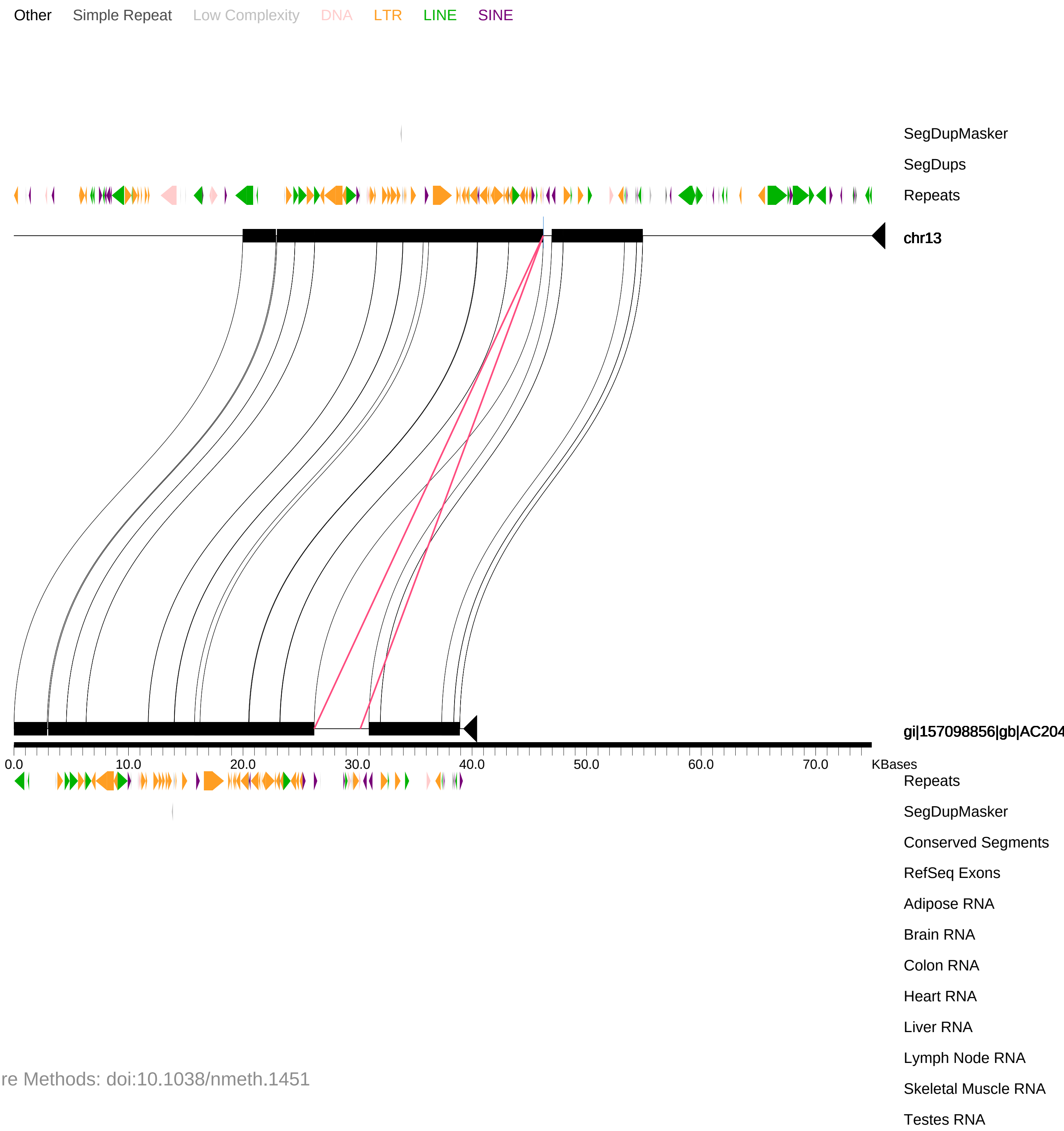
Insertion Size: 10473



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC204972.fa

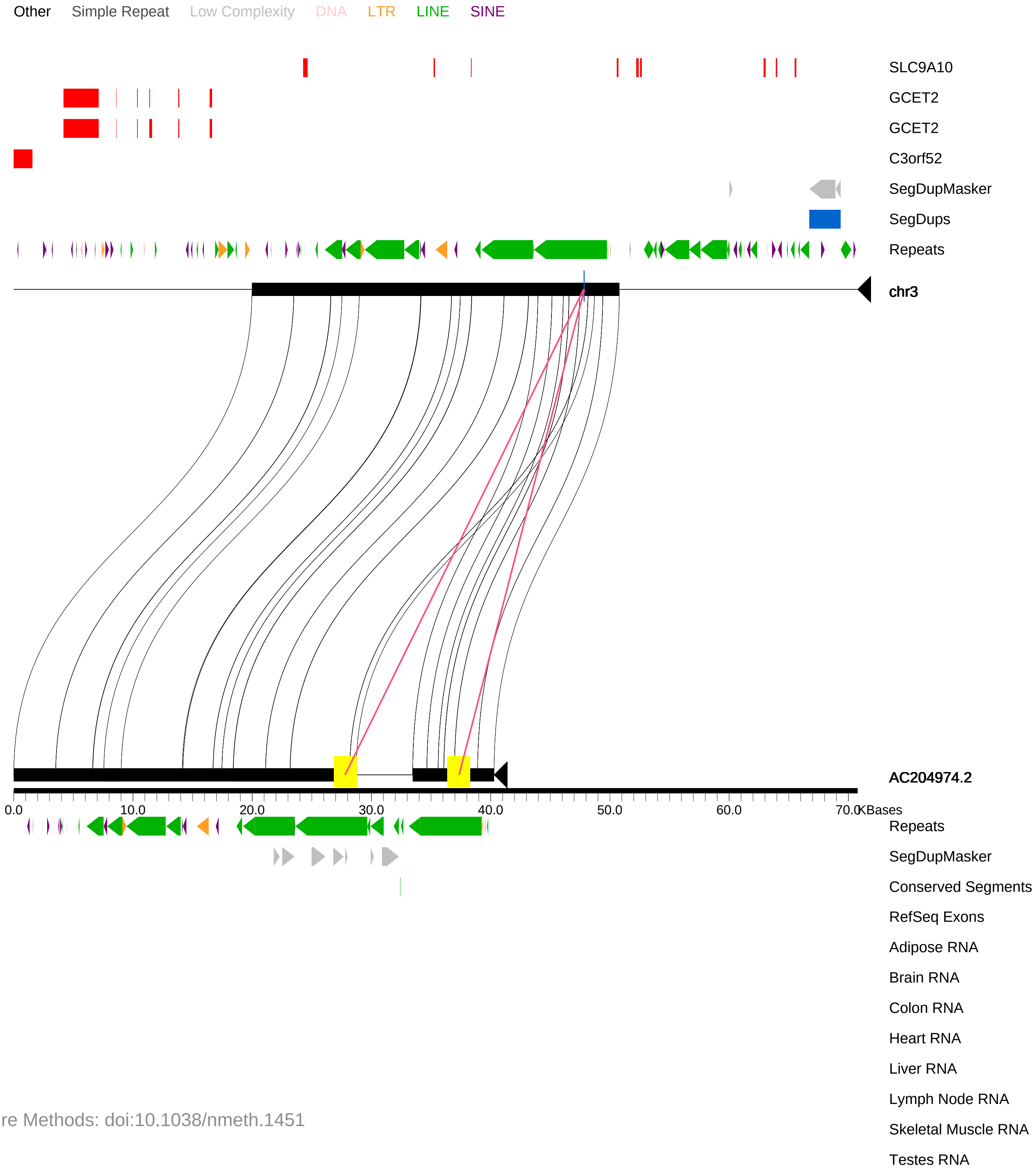
Insertion Size: 4026



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC204974.rc.fa

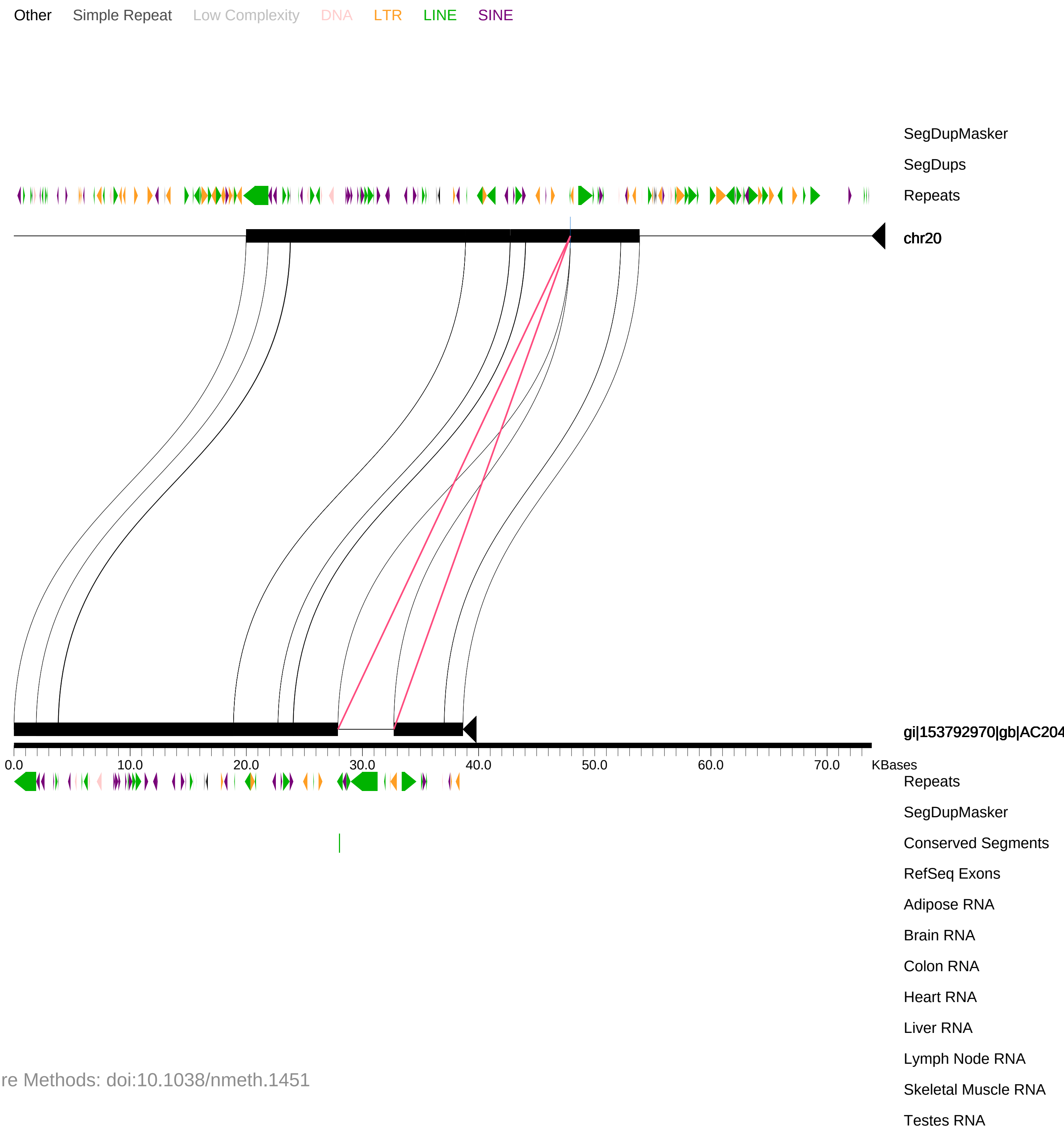
Insertion Size: 9569



Nature Methods: doi:10.1038/nmeth.1451

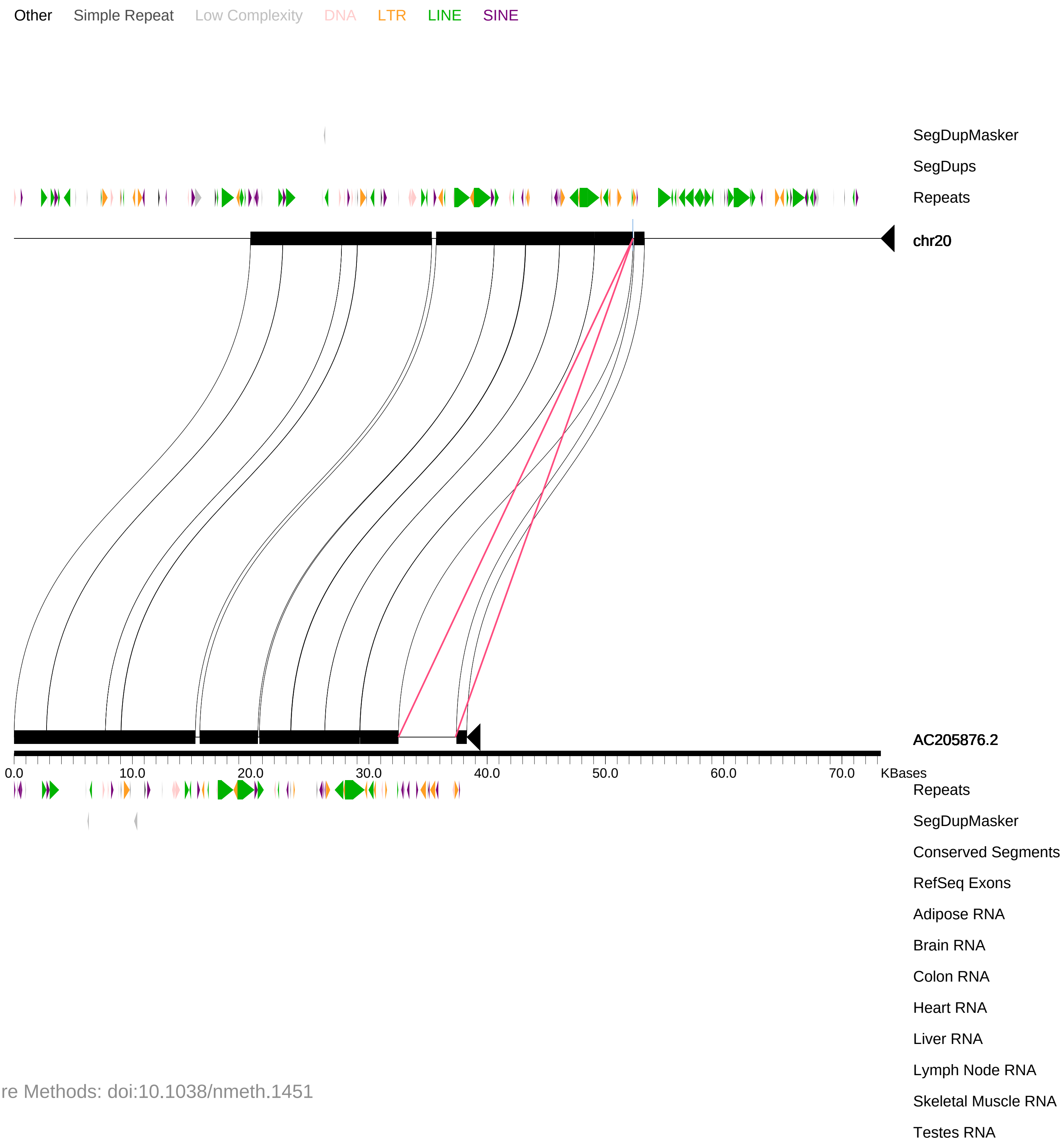
Clone file = AC204980.fa

Insertion Size: 4812



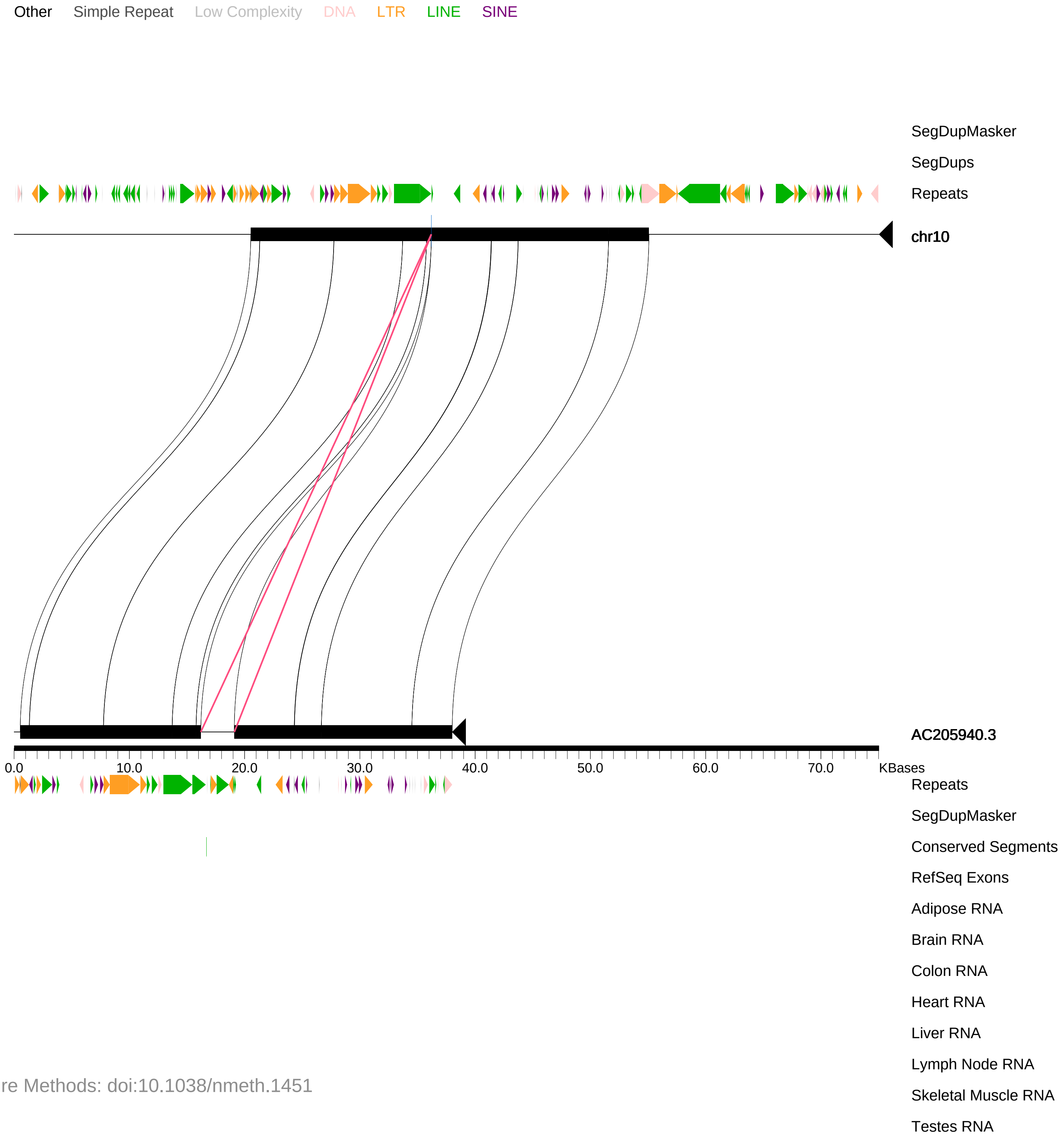
Clone file = AC205876.rc.fa

Insertion Size: 4821



Clone file = AC205940.rc.fa

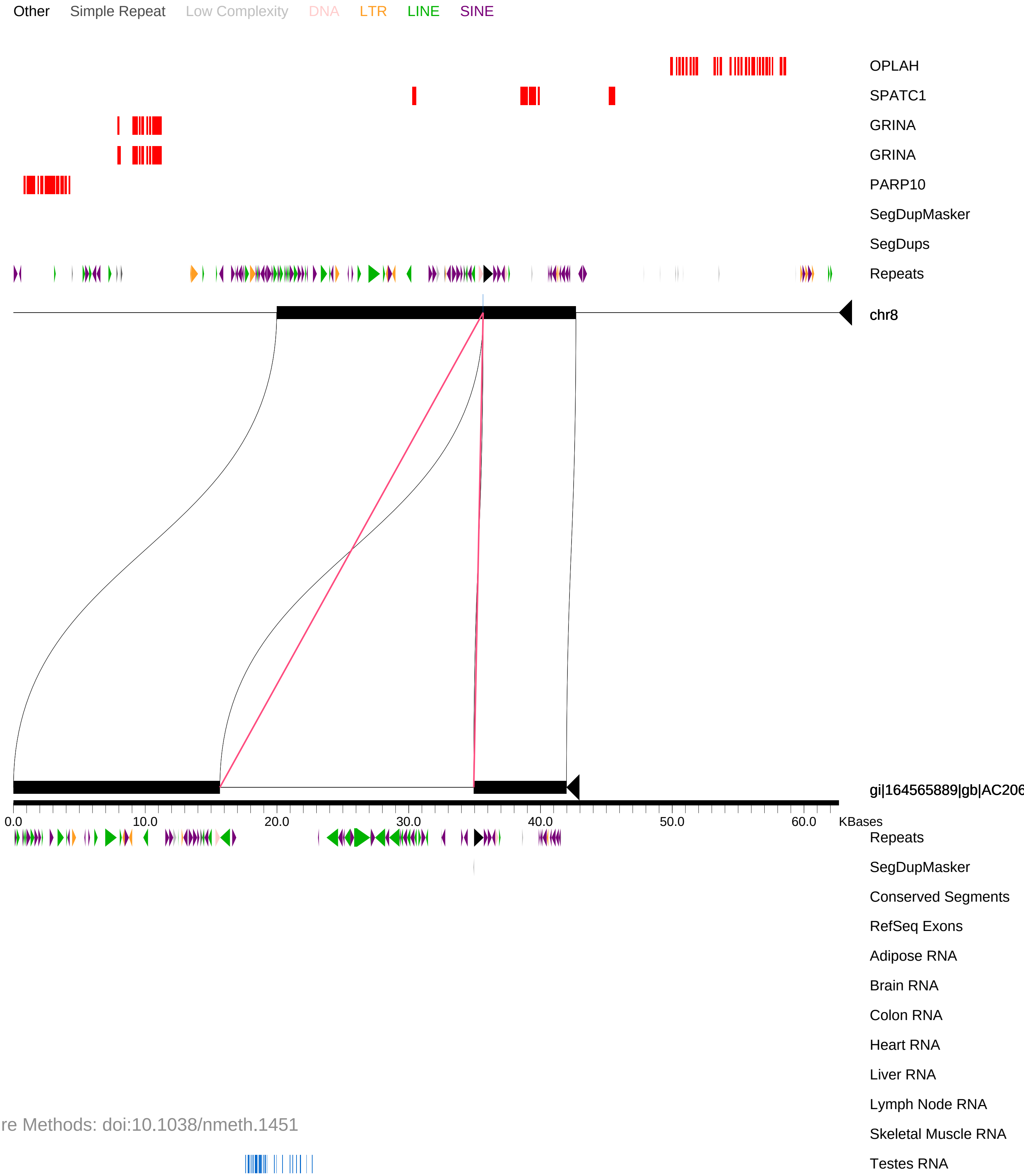
Insertion Size: 2912



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC206437.fa

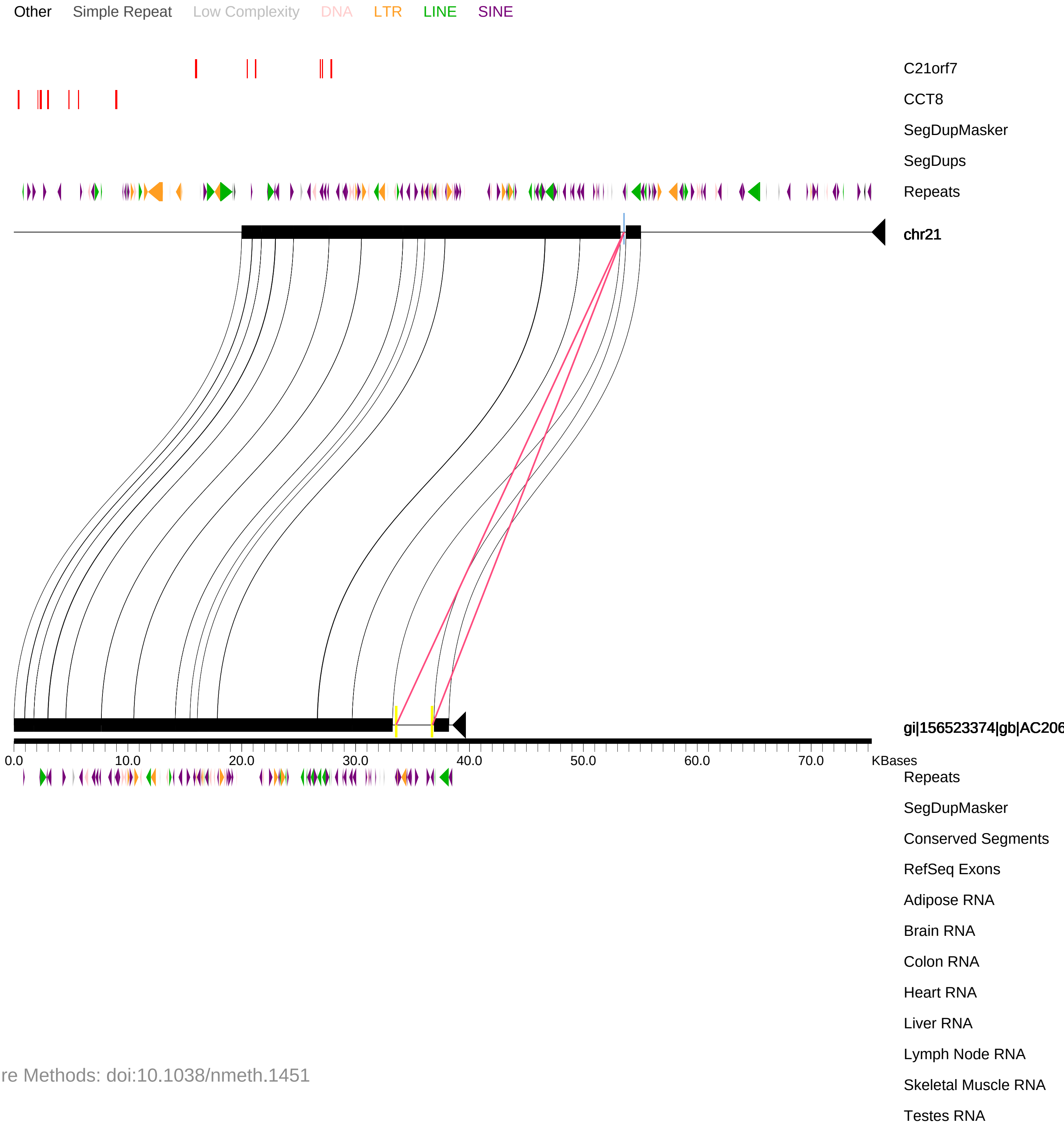
Insertion Size: 19270



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC206474.fa

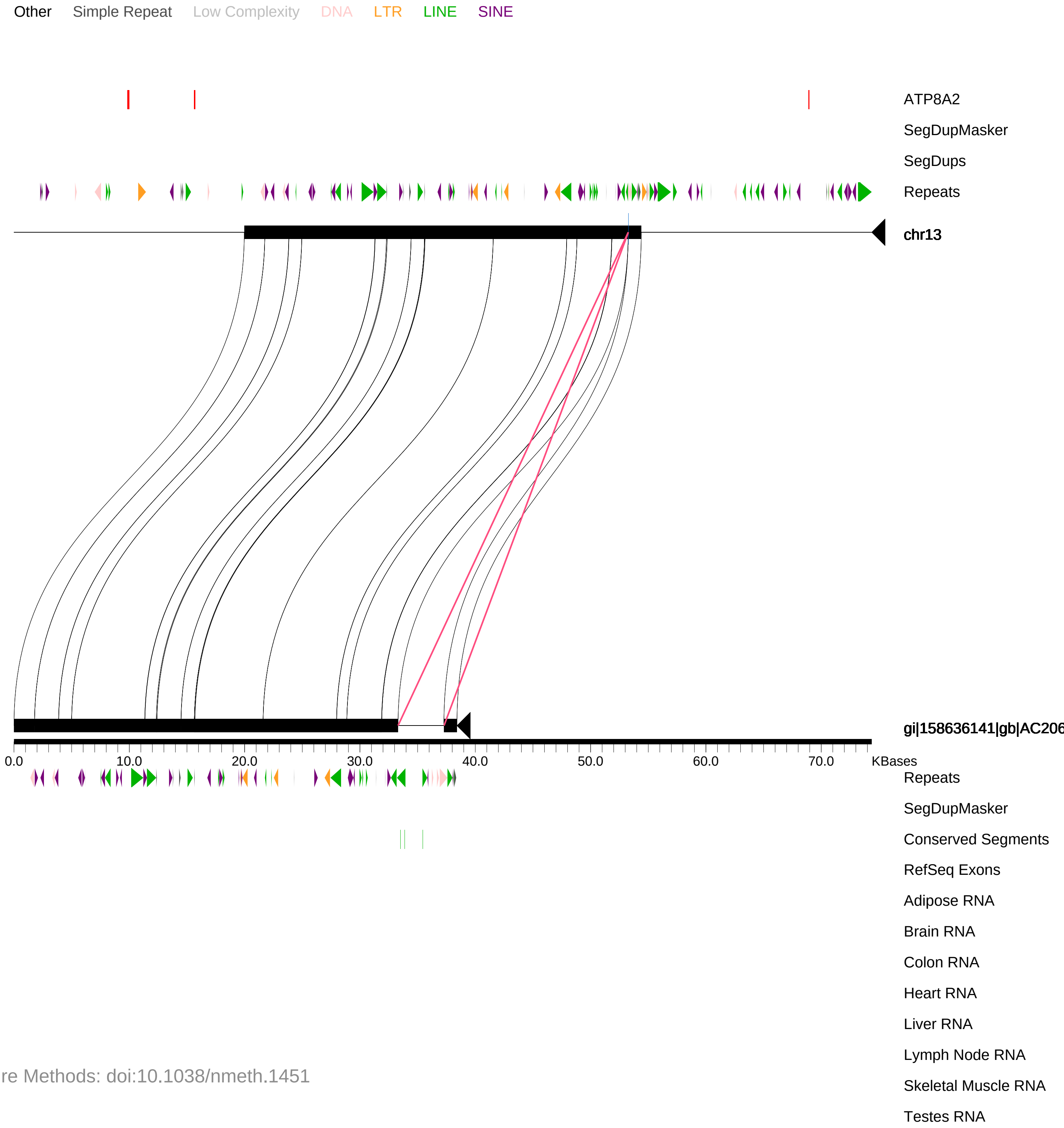
Insertion Size: 3187



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC206479.fa

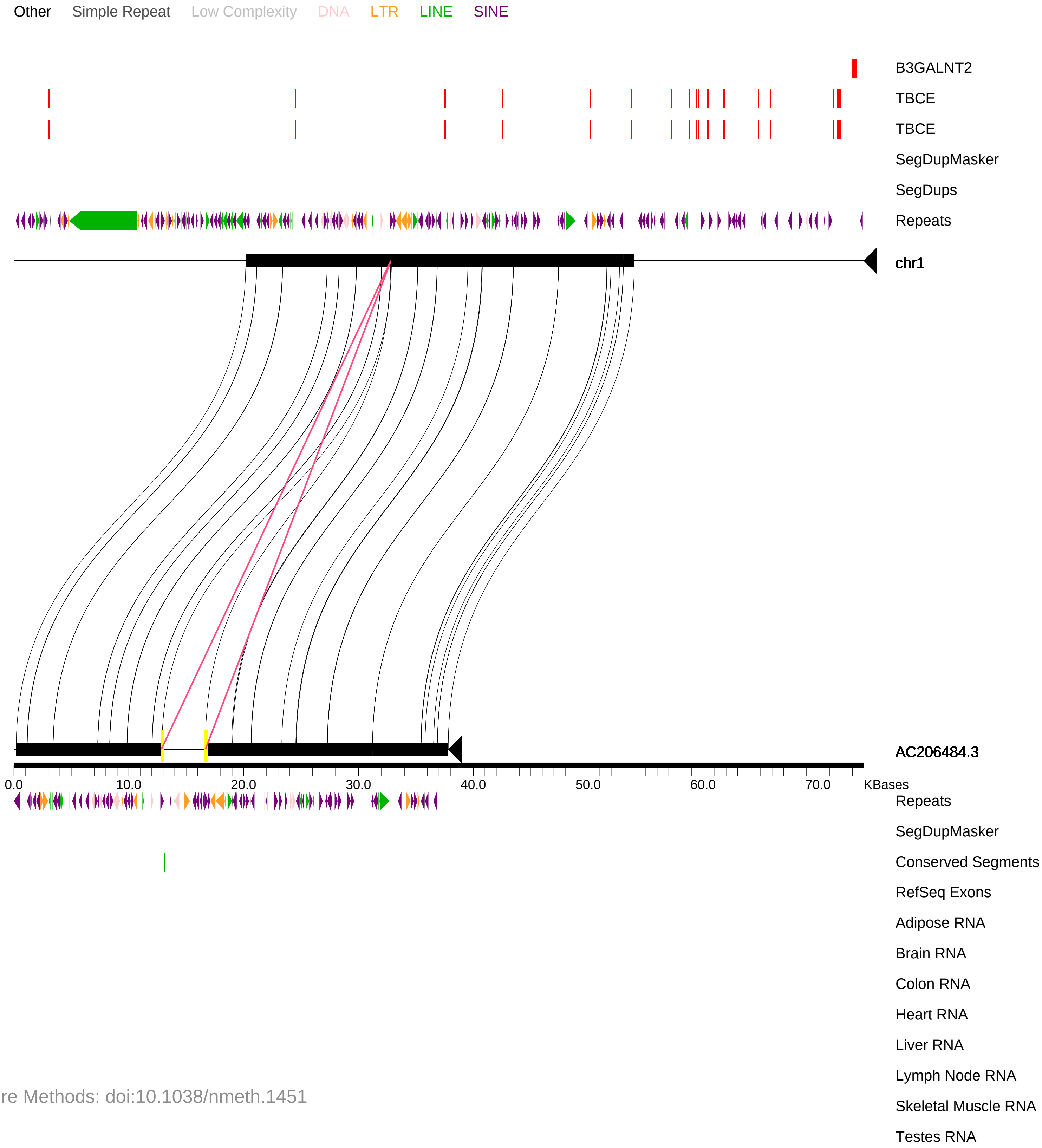
Insertion Size: 3986



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC206484.rc.fa

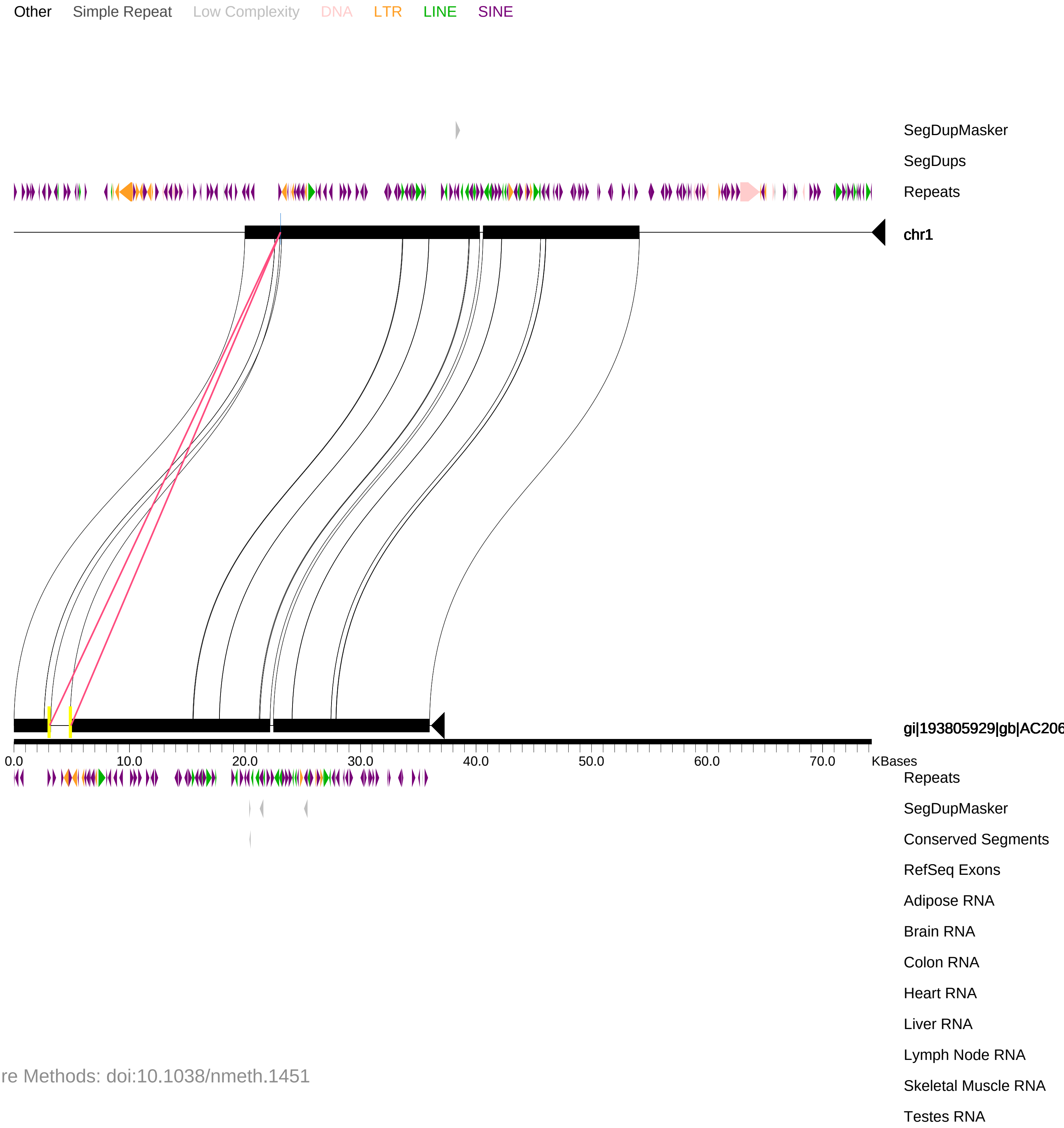
Insertion Size: 3848



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC206609.fa

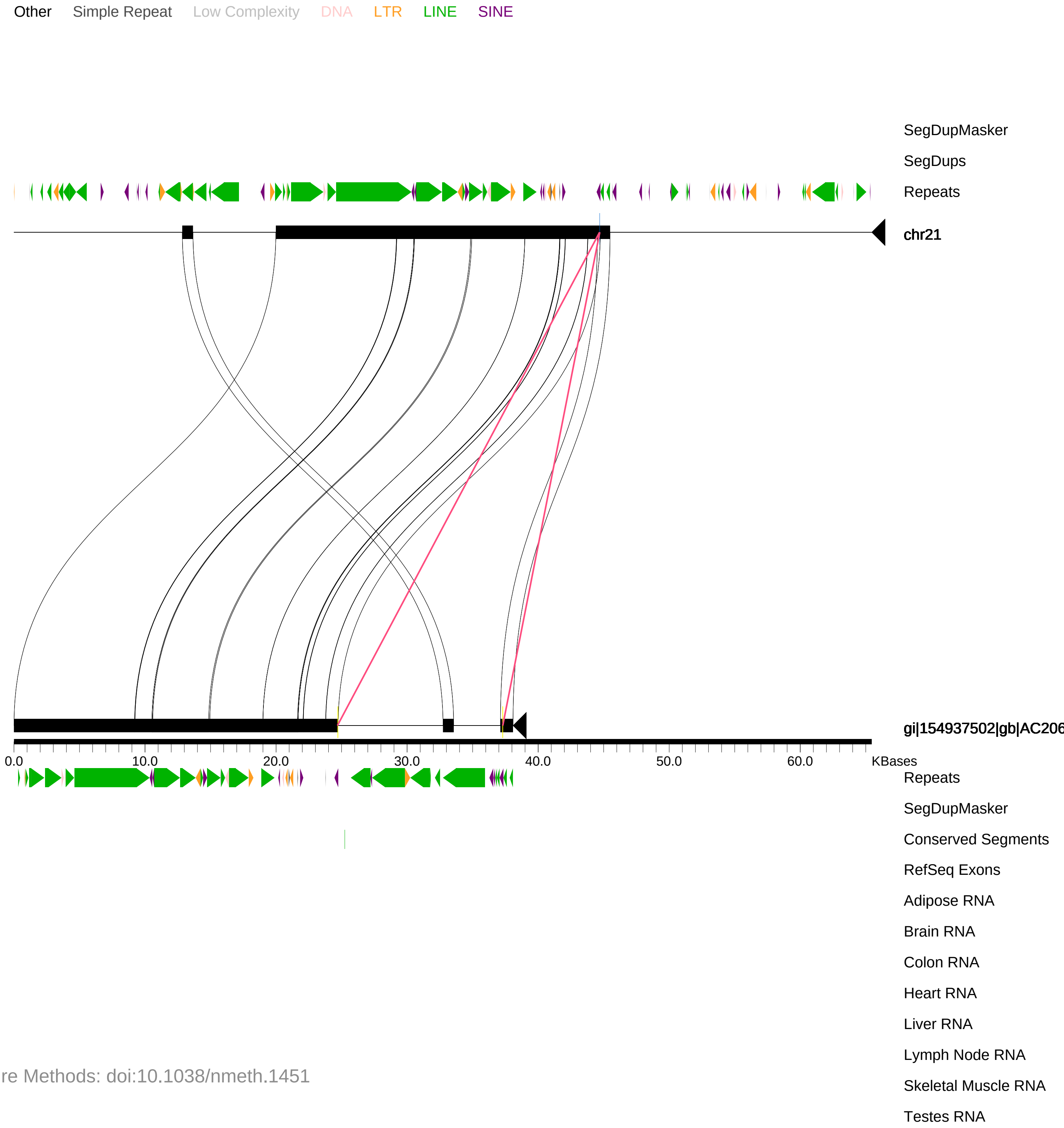
Insertion Size: 1853



Nature Methods: doi:10.1038/nmeth.1451

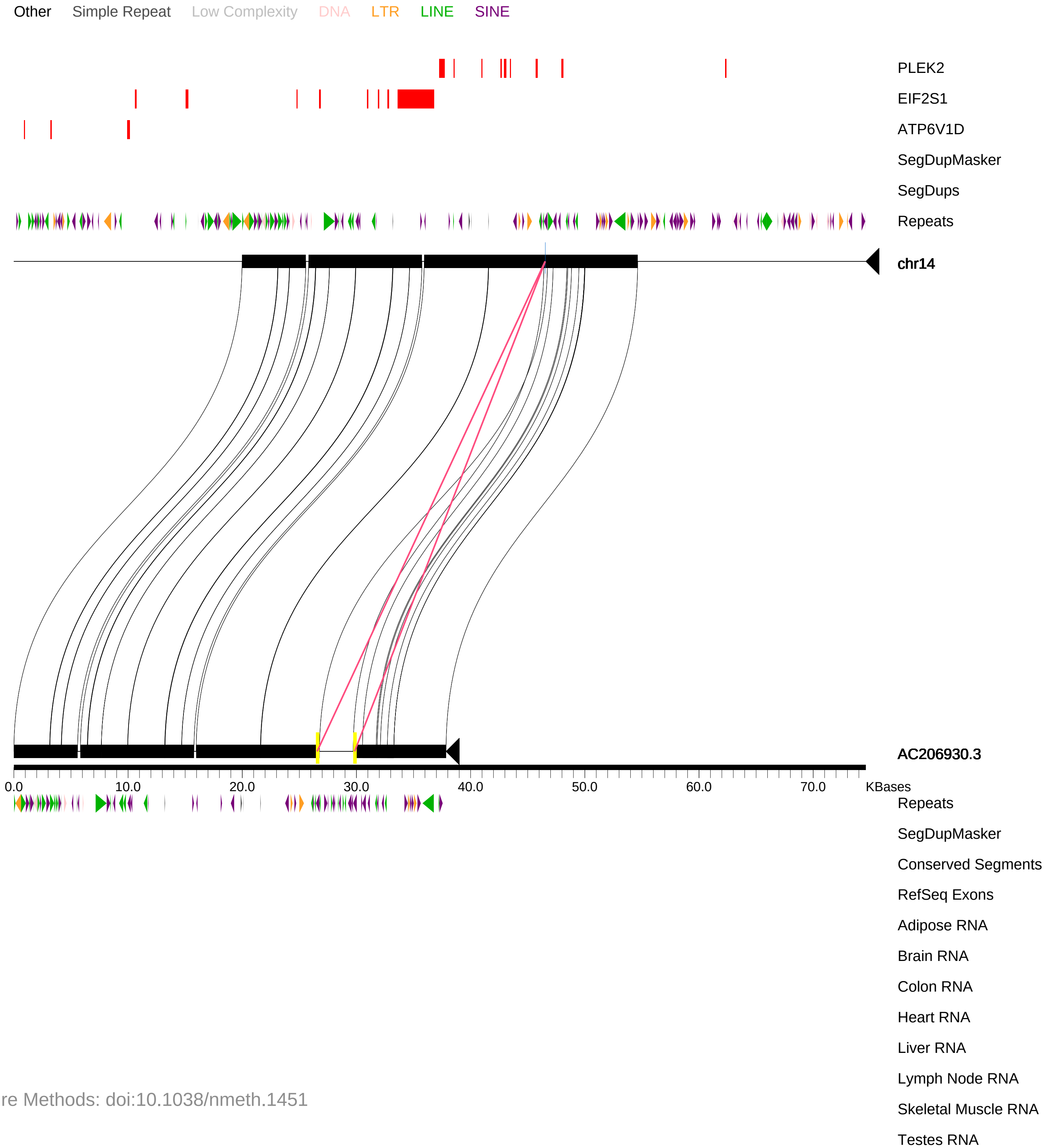
Clone file = AC206743.fa

Insertion Size: 12617



Clone file = AC206930.rc.fa

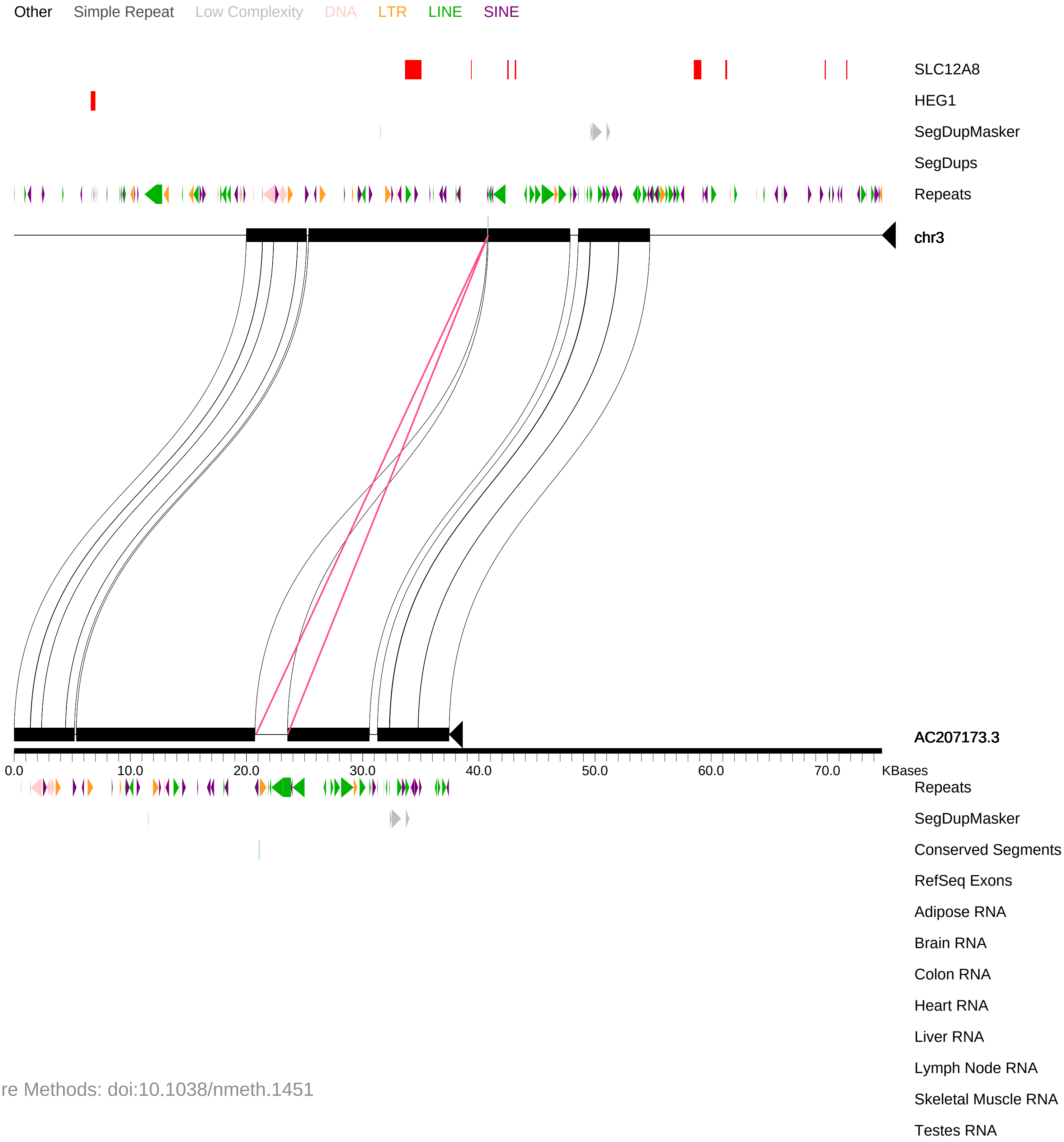
Insertion Size: 3284



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC207173.rc.fa

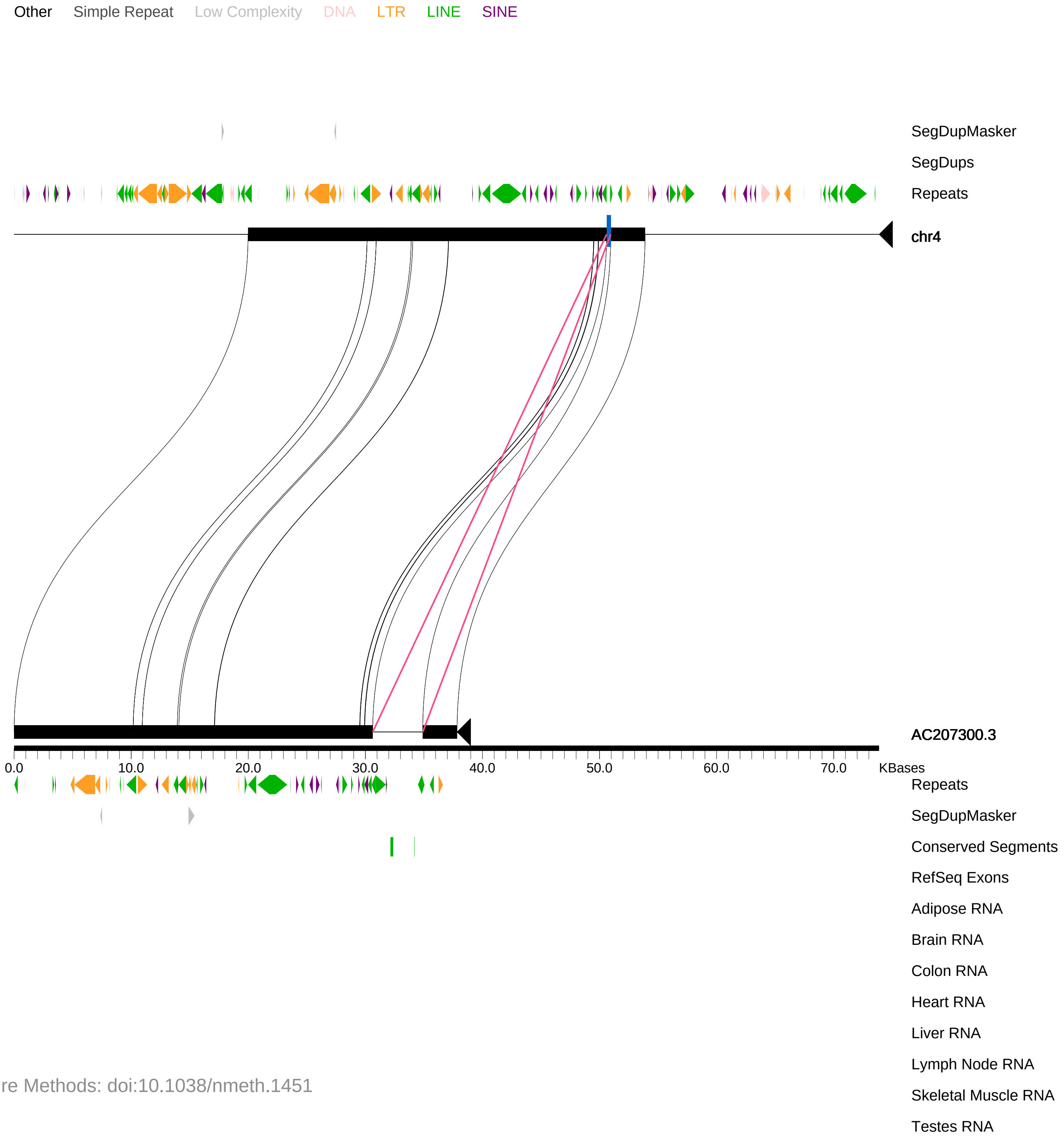
Insertion Size: 2721



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC207300.rc.fa

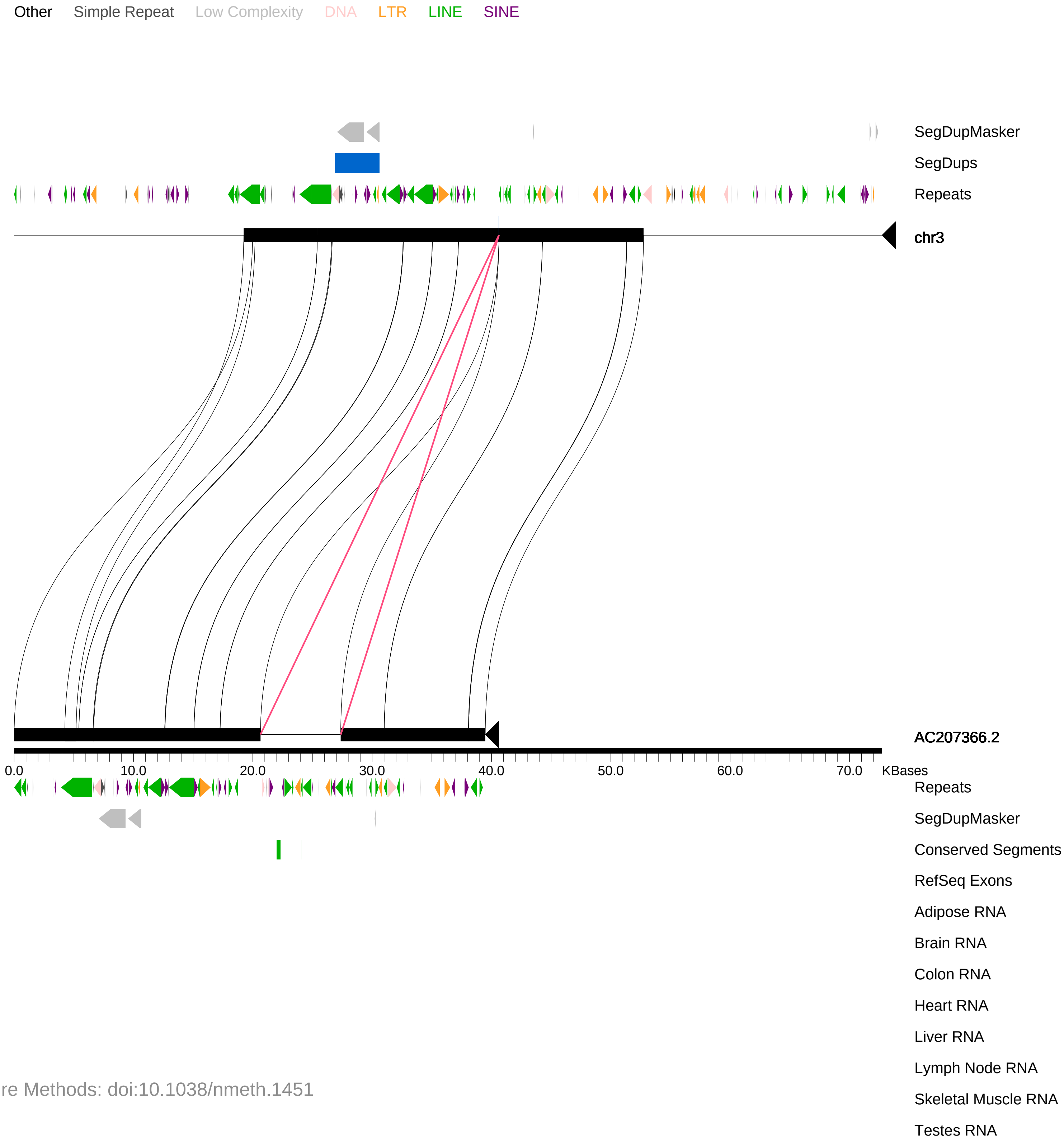
Insertion Size: 4274



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC207366.rc.fa

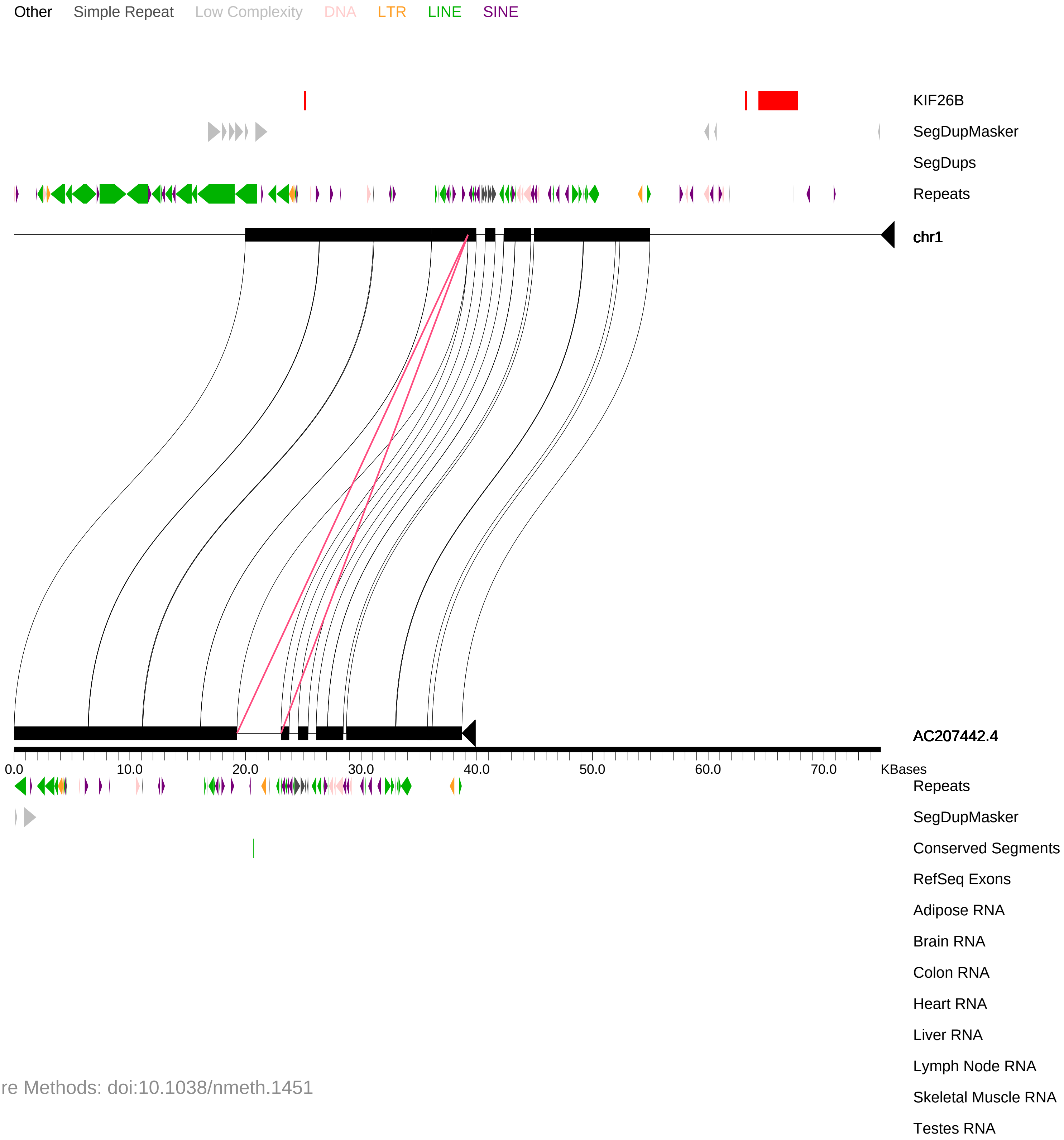
Insertion Size: 6730



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC207442.rc.fa

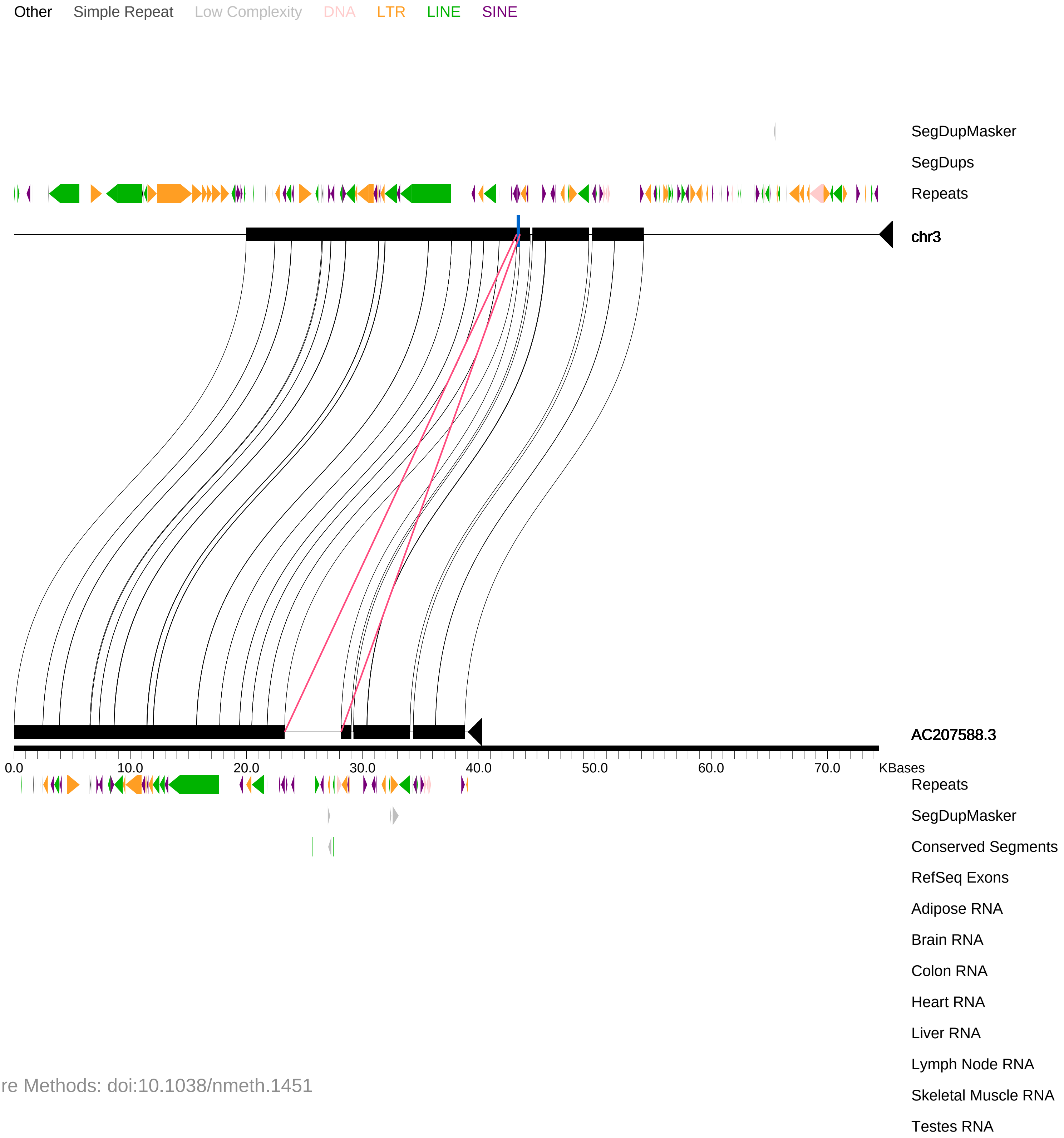
Insertion Size: 3811



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC207588.rc.fa

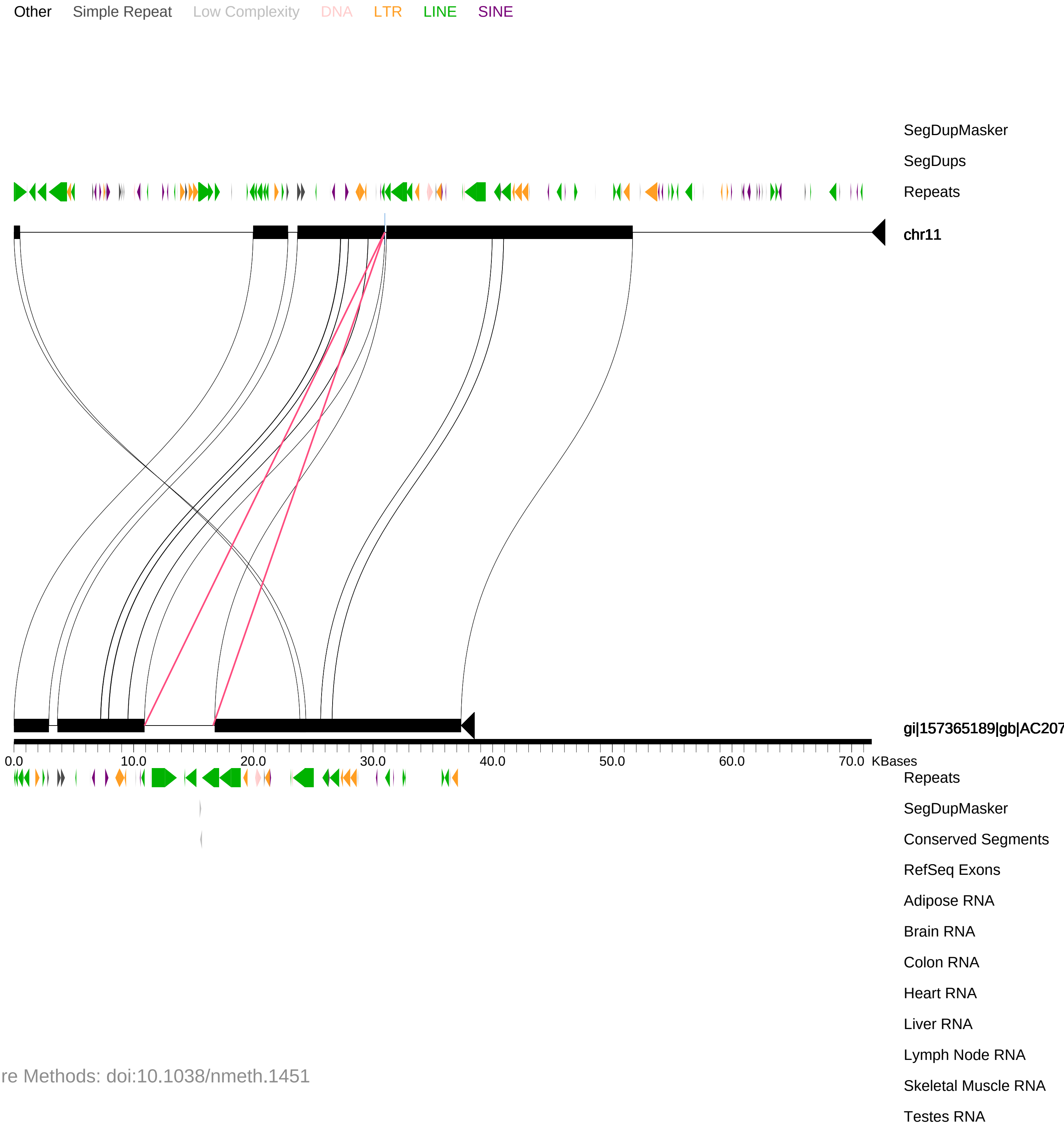
Insertion Size: 4860



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC207607.fa

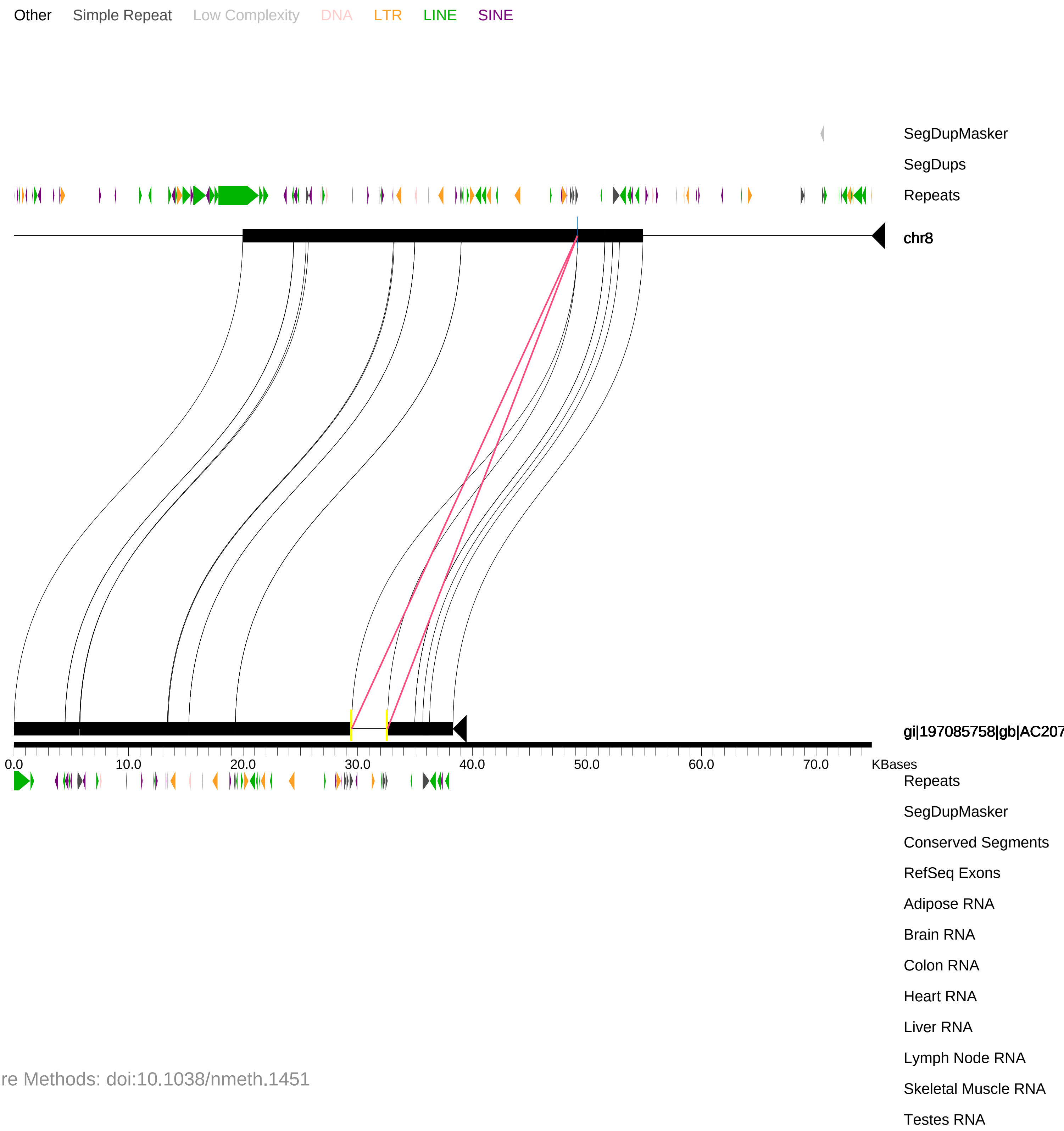
Insertion Size: 5739



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC207611.fa

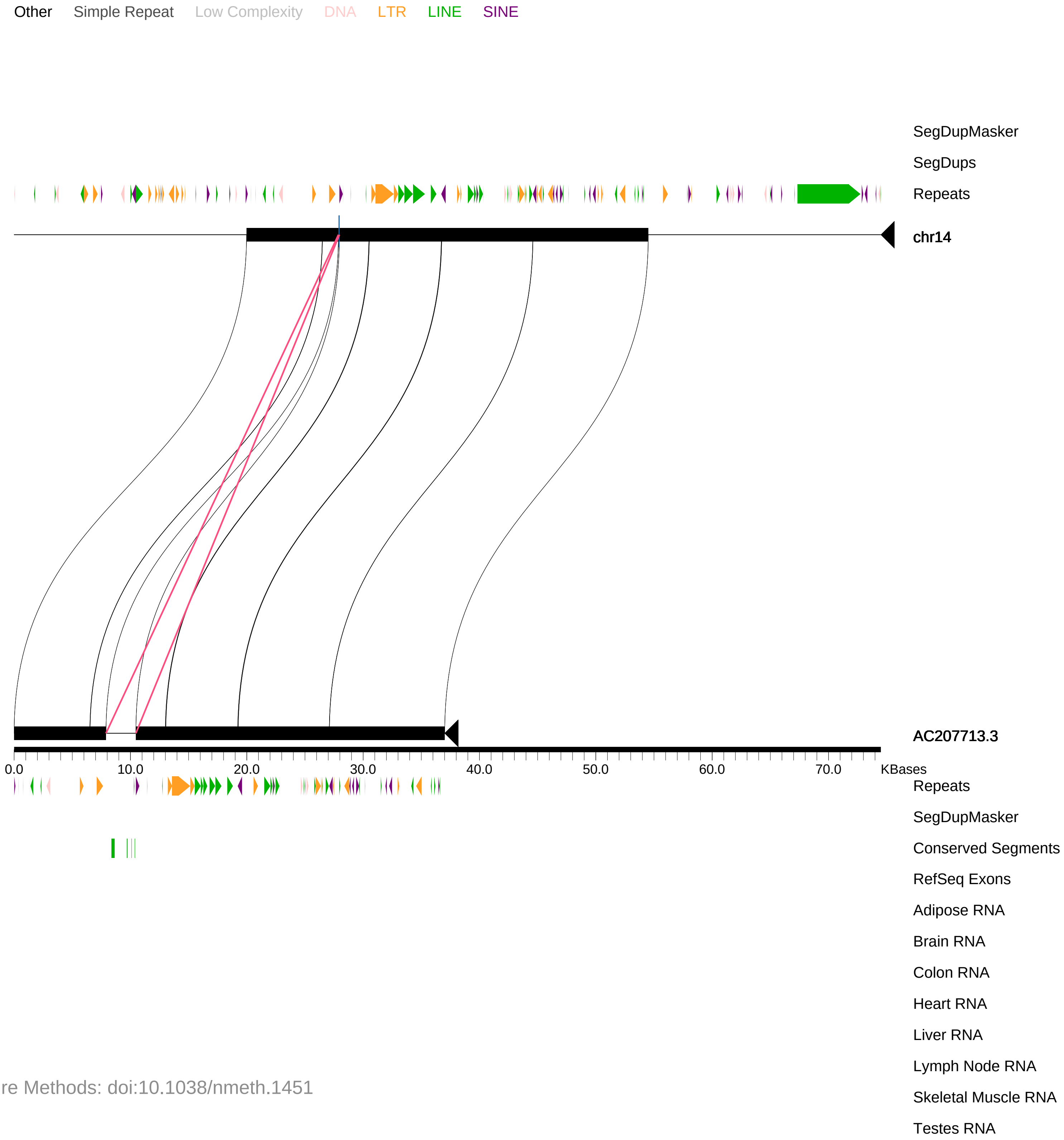
Insertion Size: 3133



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC207713.rc.fa

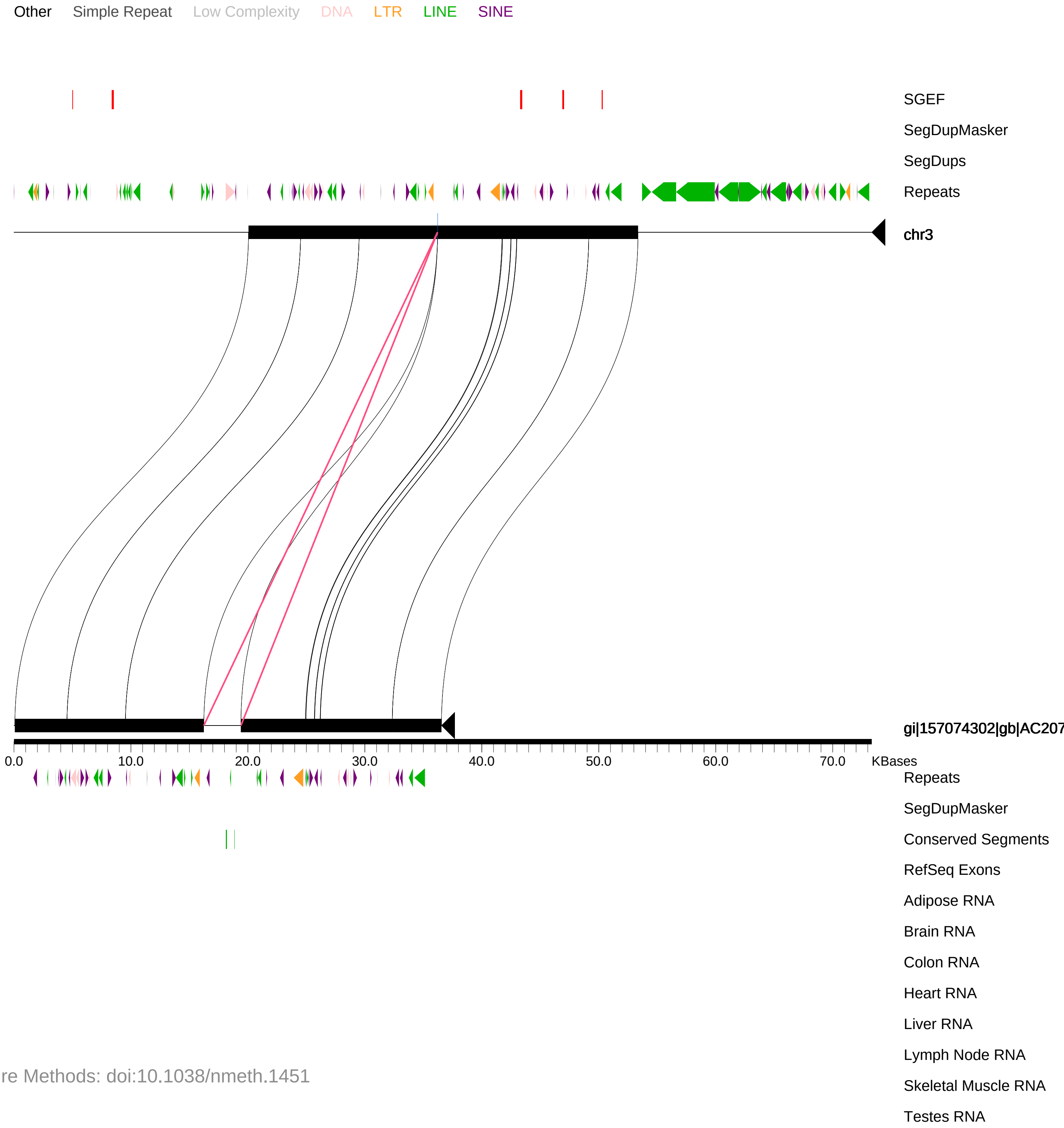
Insertion Size: 2571



Nature Methods: doi:10.1038/nmeth.1451

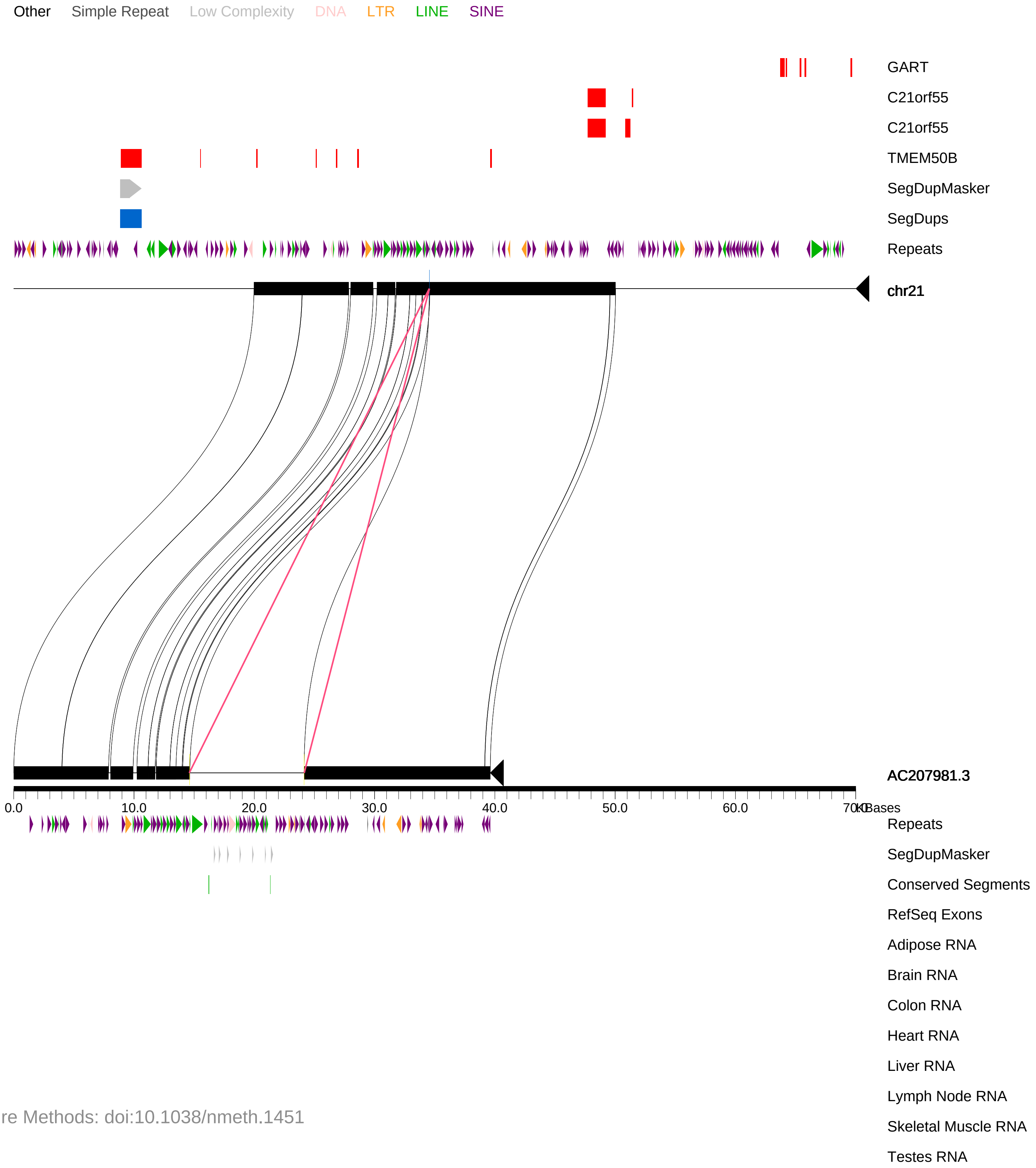
Clone file = AC207777.fa

Insertion Size: 3173



Clone file = AC207981.rc.fa

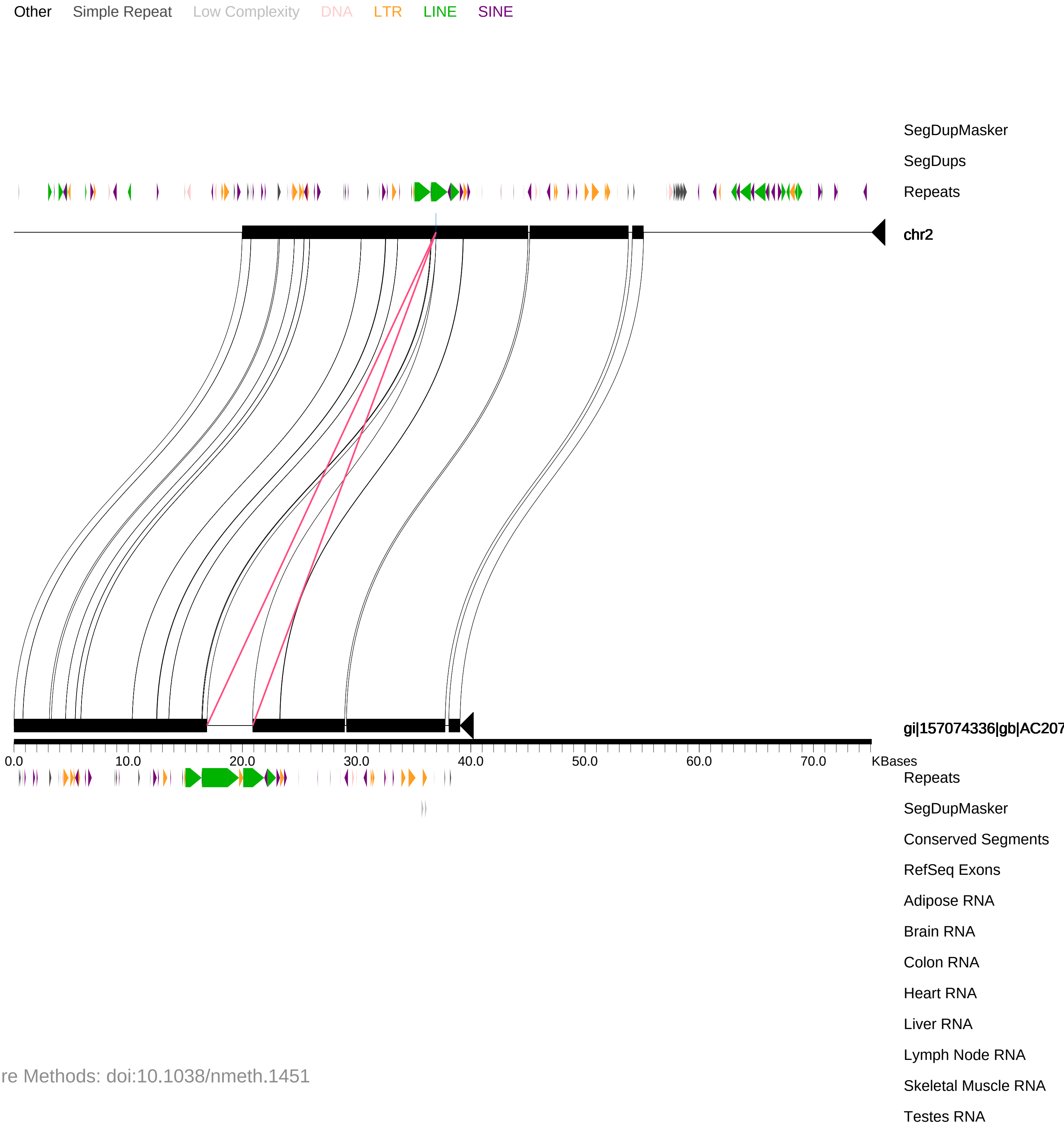
Insertion Size: 9557



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC207999.fa

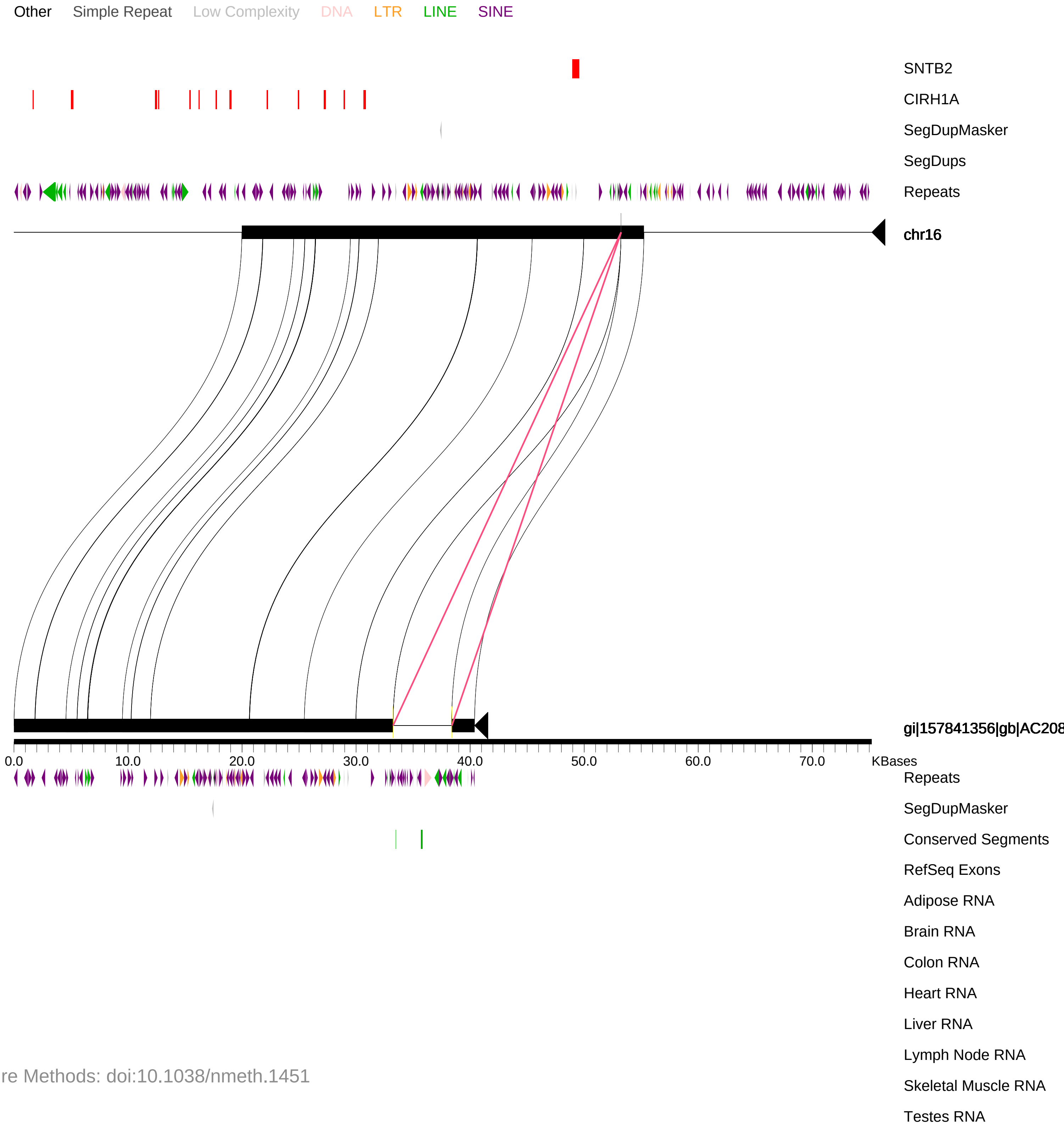
Insertion Size: 4006



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC208009.fa

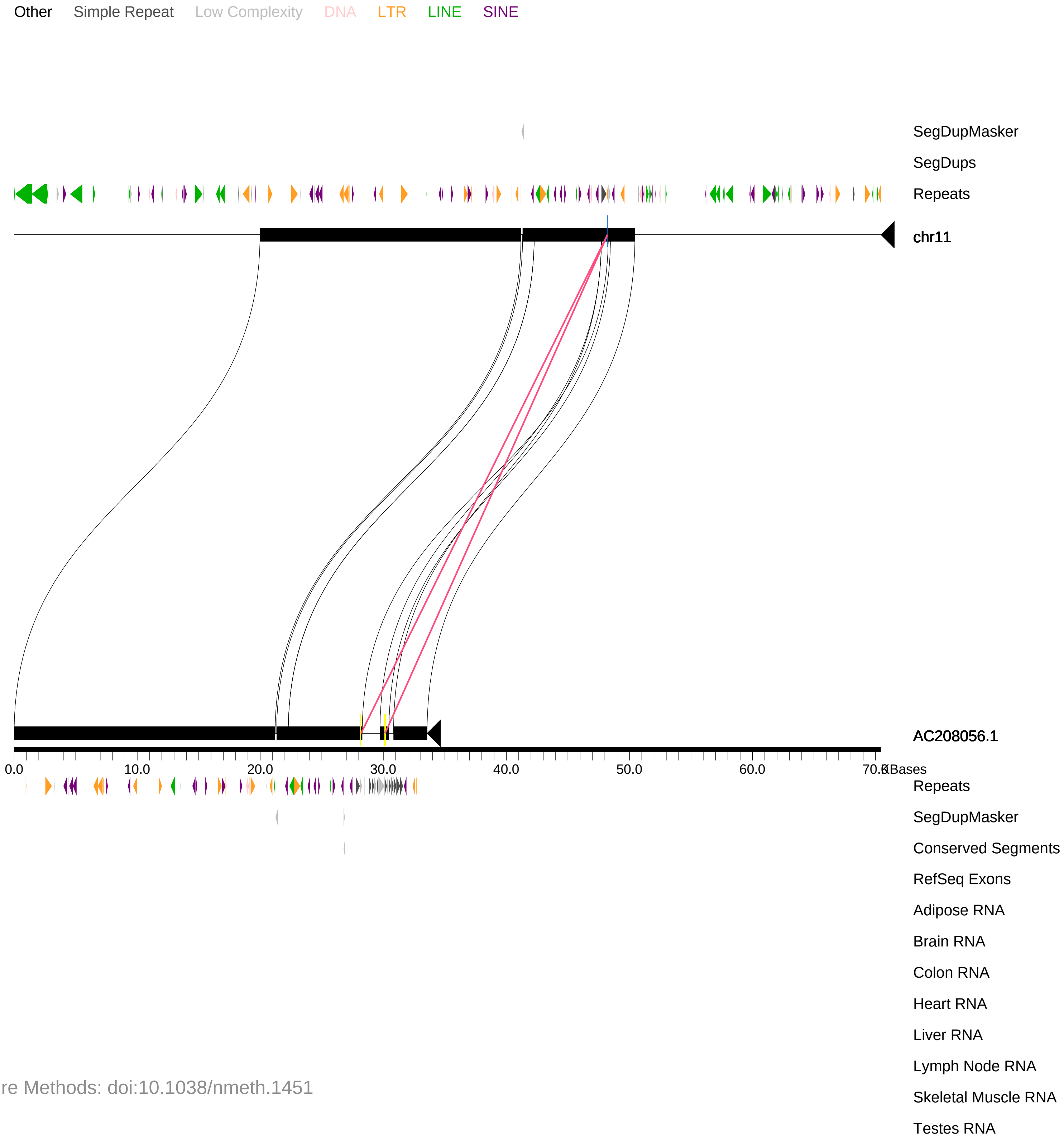
Insertion Size: 5157



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC208056.rc.fa

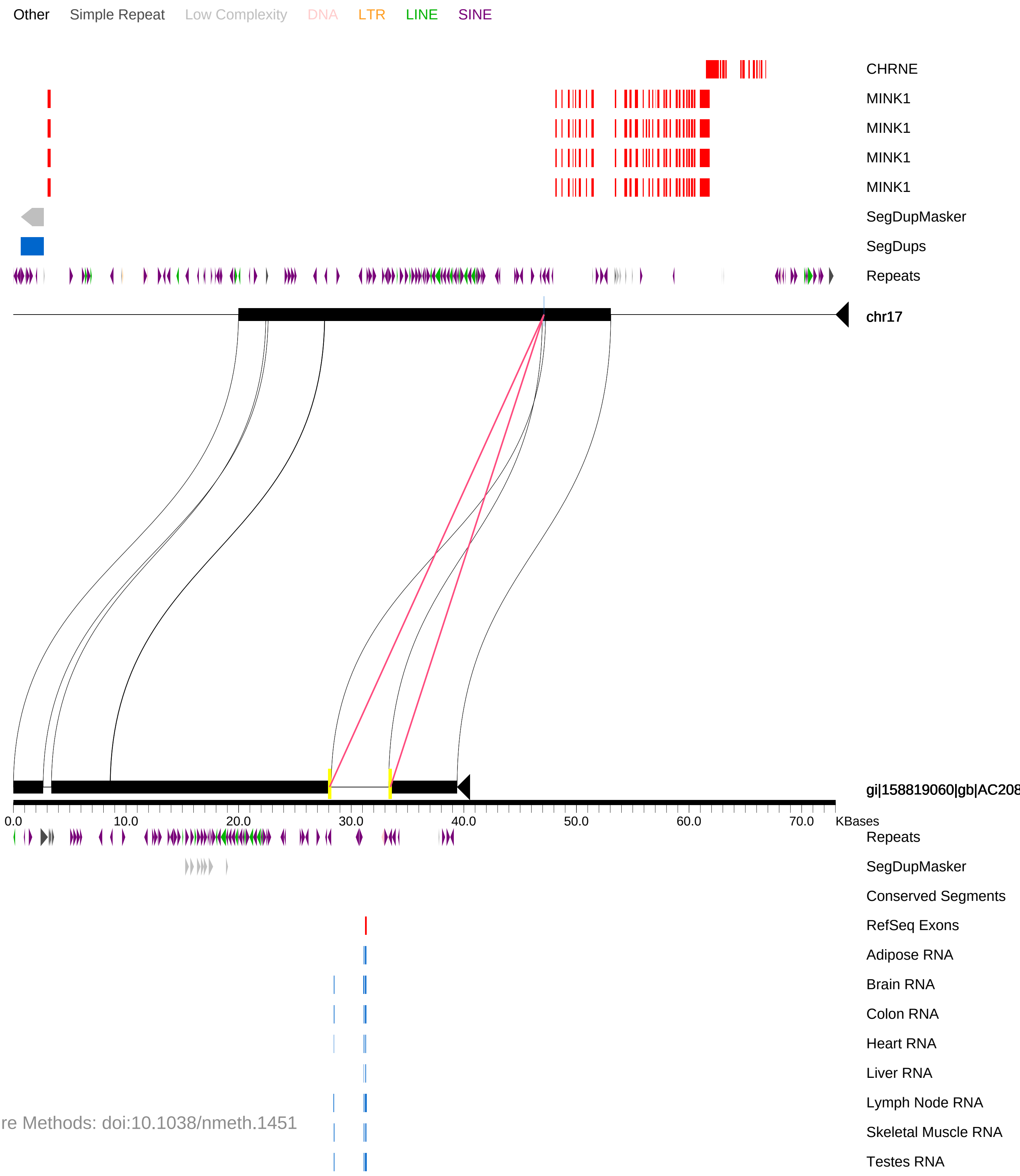
Insertion Size: 1953



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC208058.fa

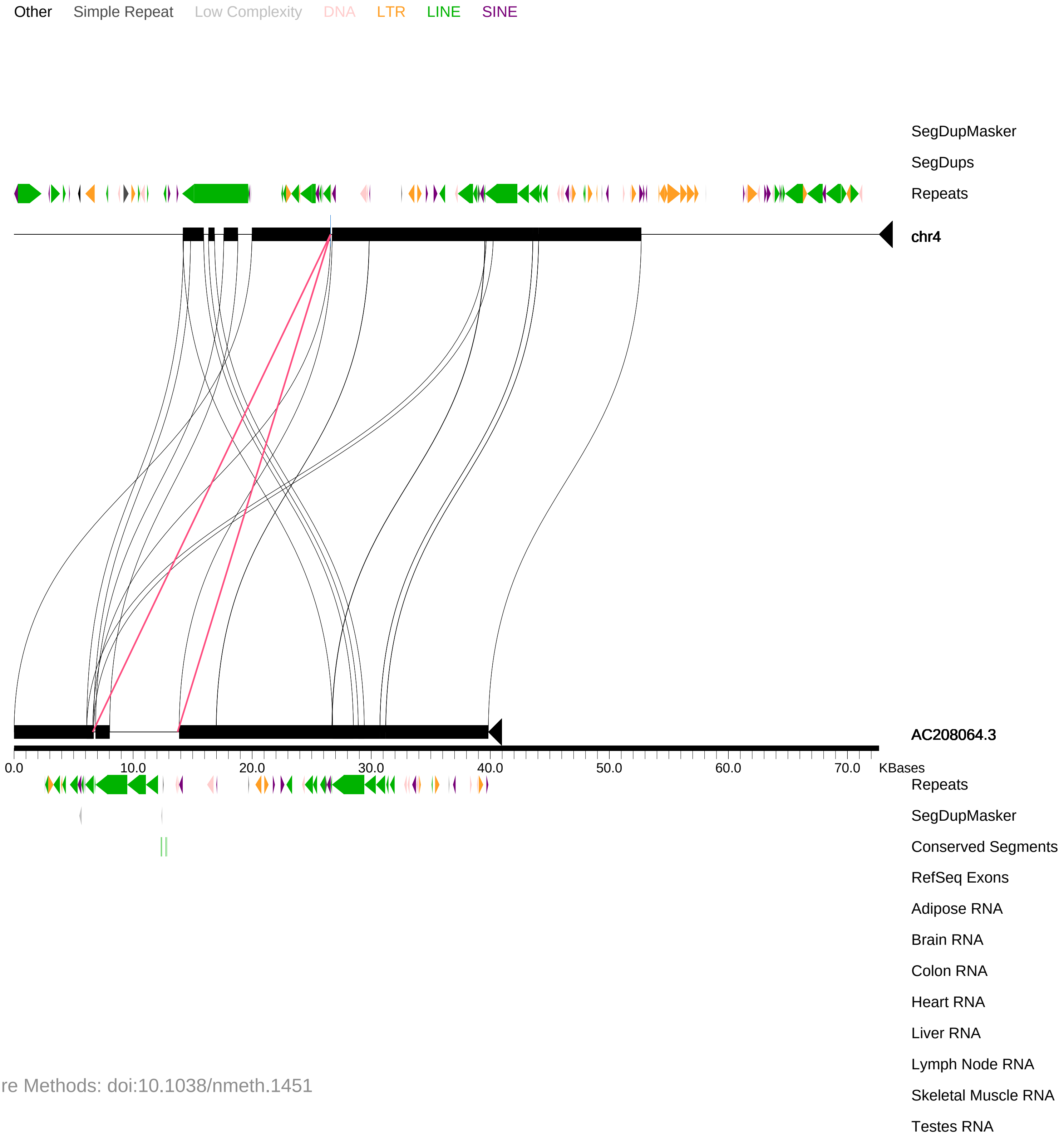
Insertion Size: 5405



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC208064.rc.fa

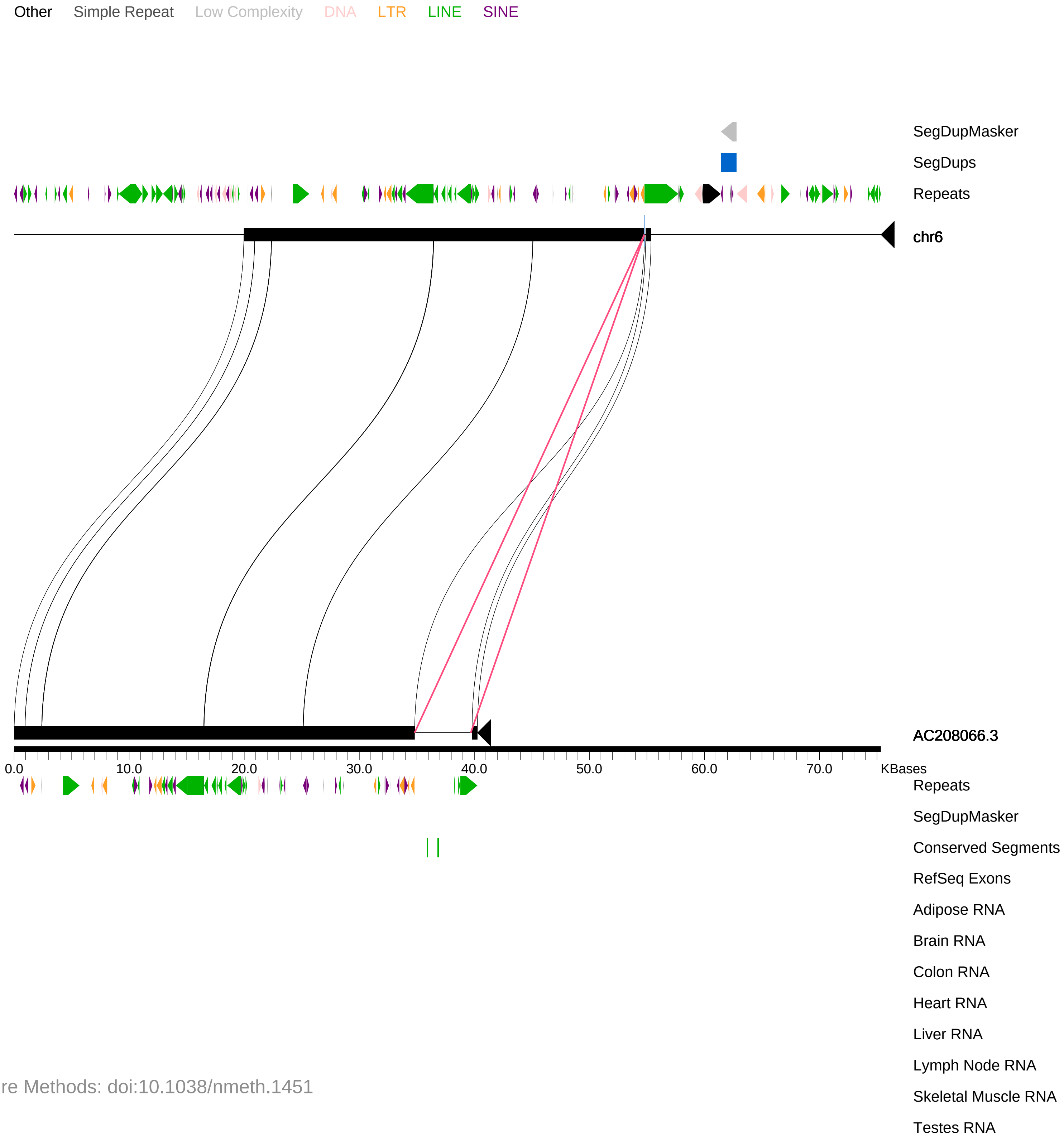
Insertion Size: 7131



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC208066.rc.fa

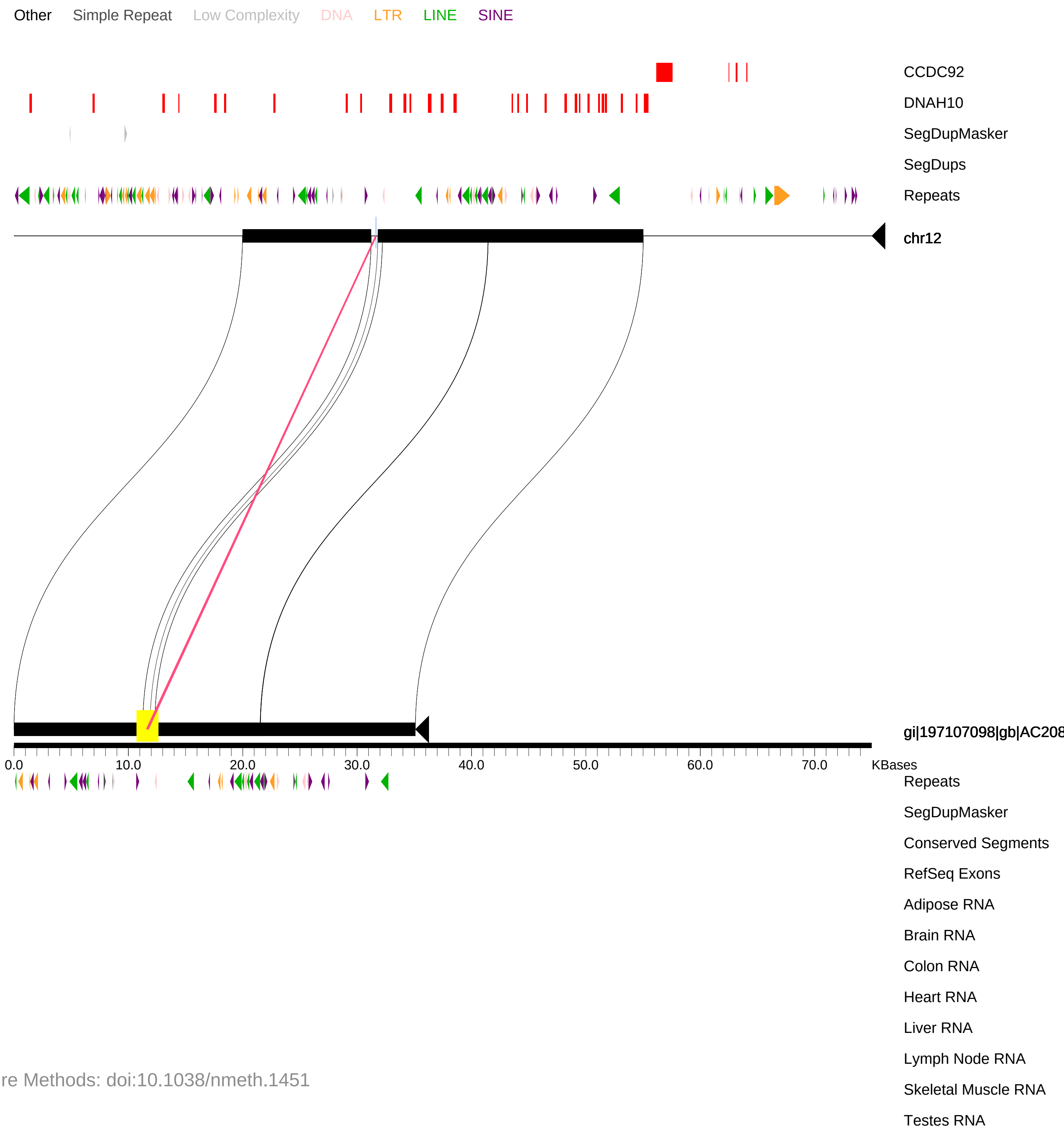
Insertion Size: 4881



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC208069.fa

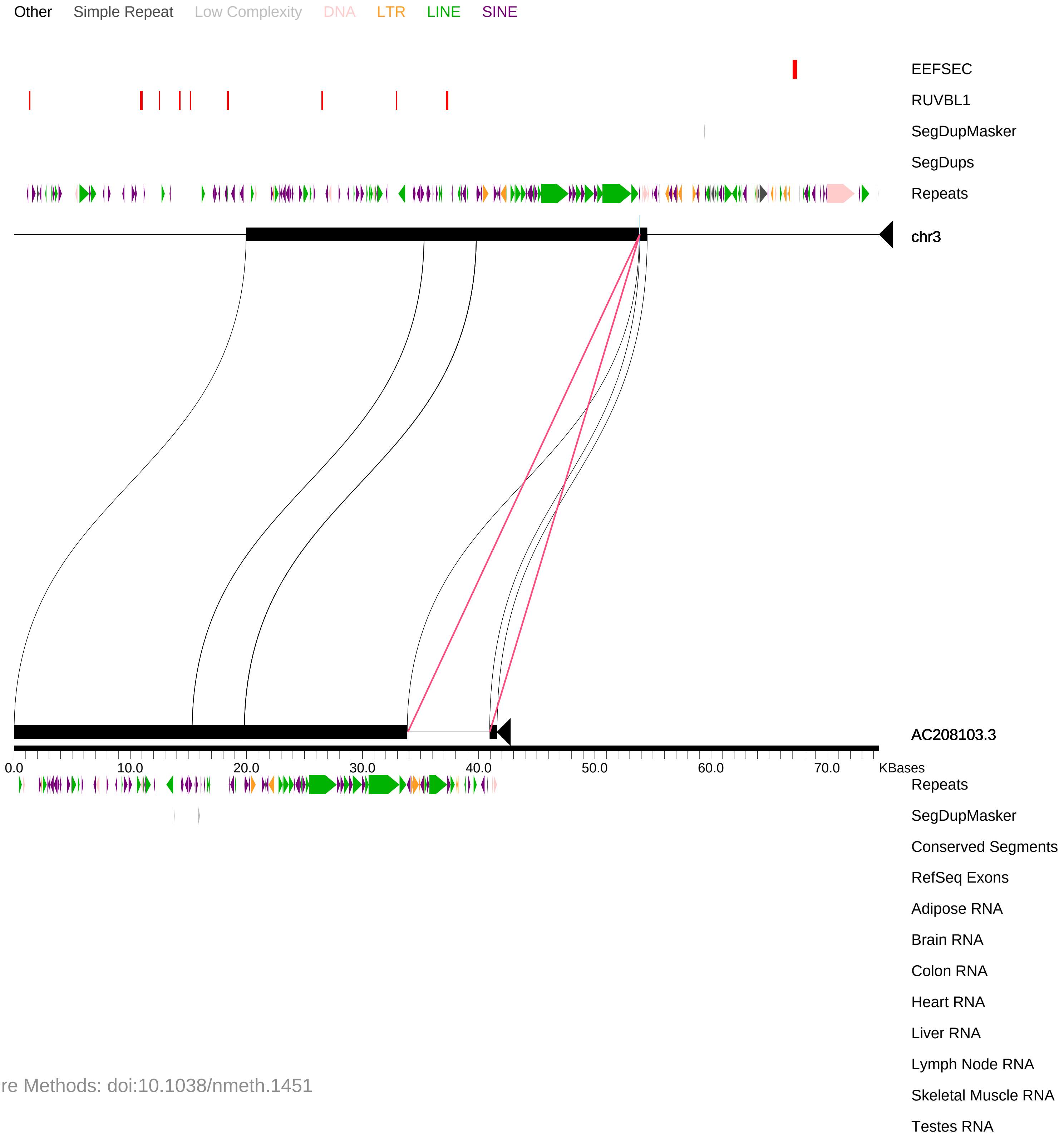
Insertion Size: 106



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC208103.rc.fa

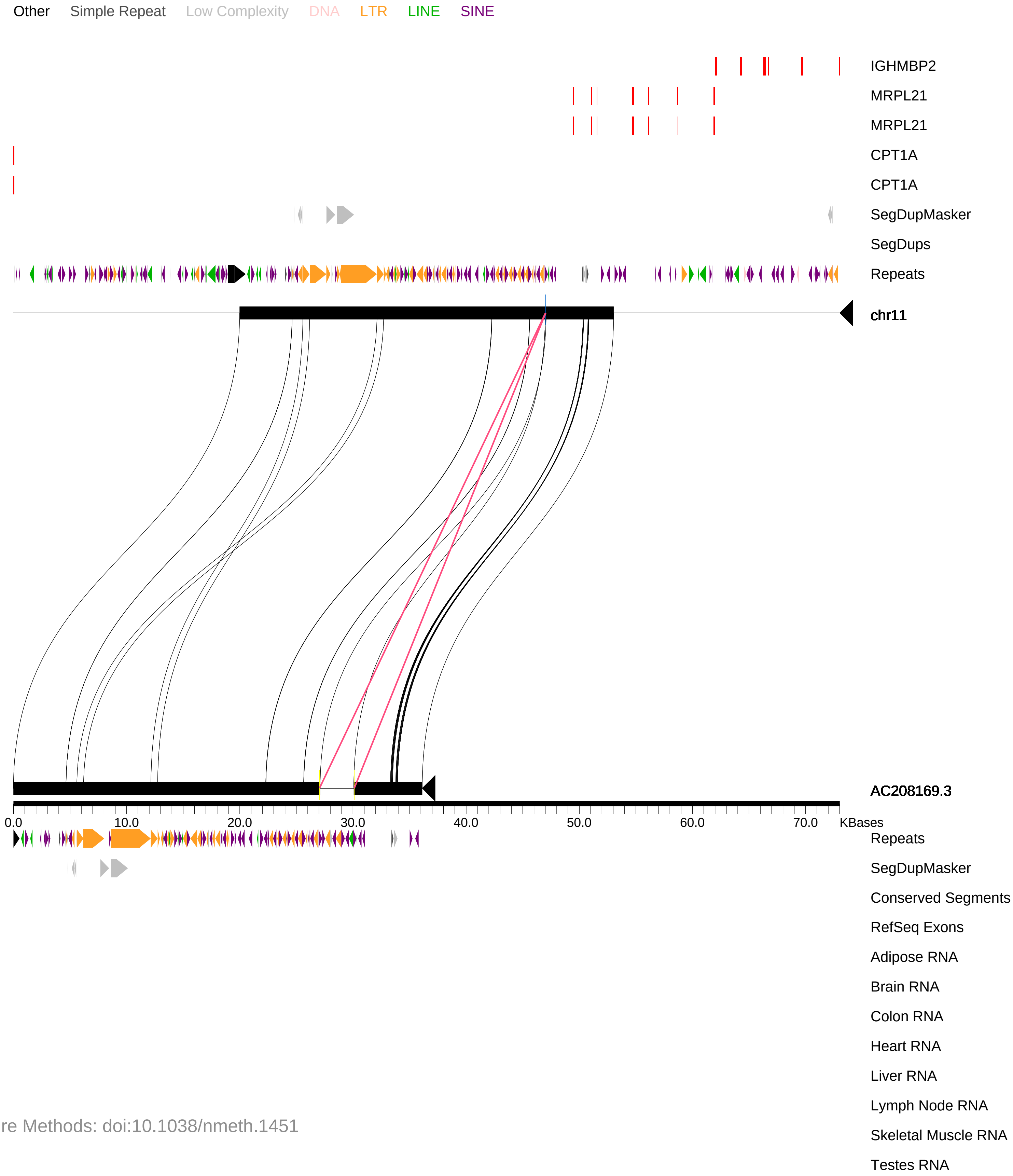
Insertion Size: 7049



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC208169.rc.fa

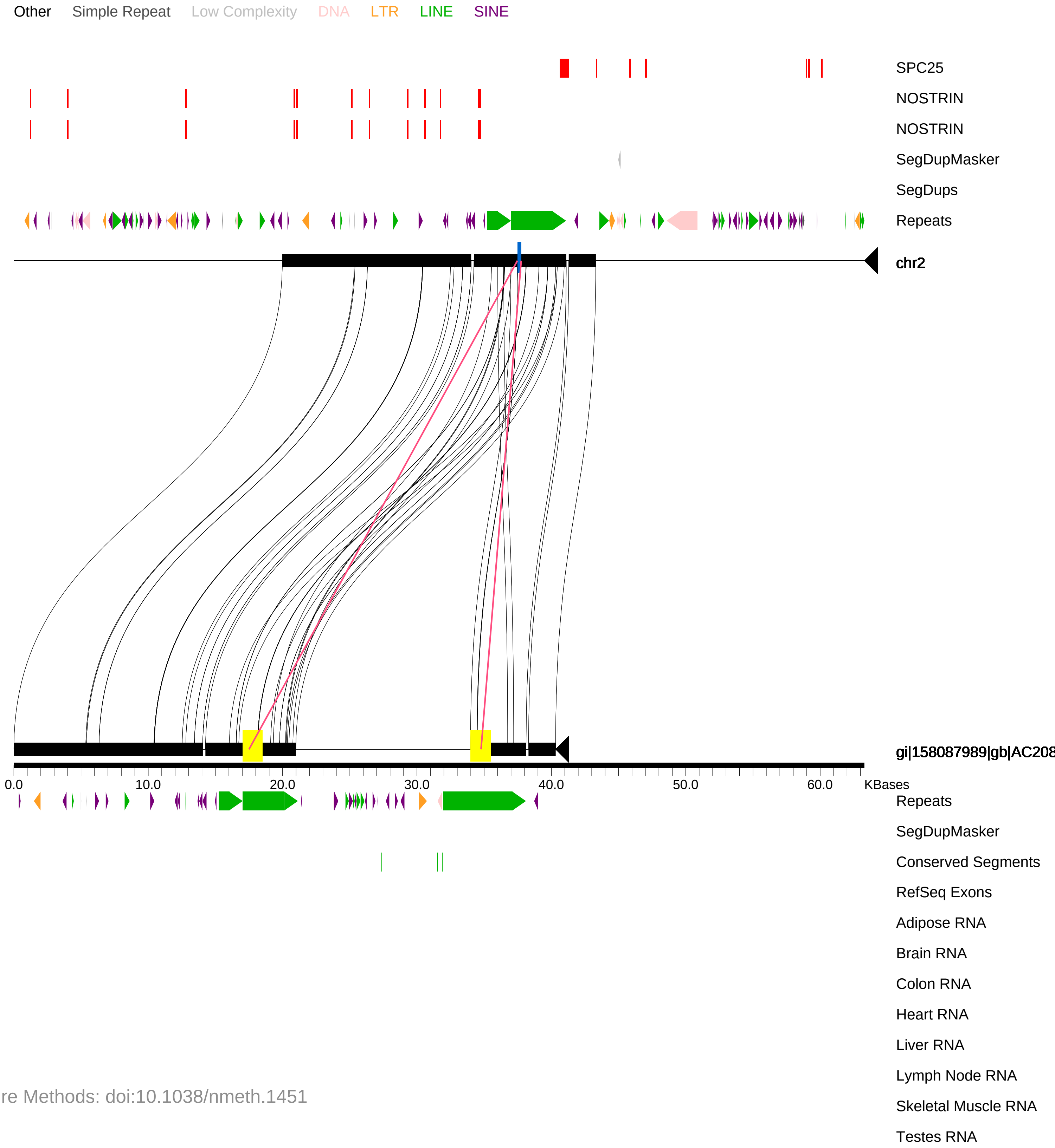
Insertion Size: 3070



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC208170.fa

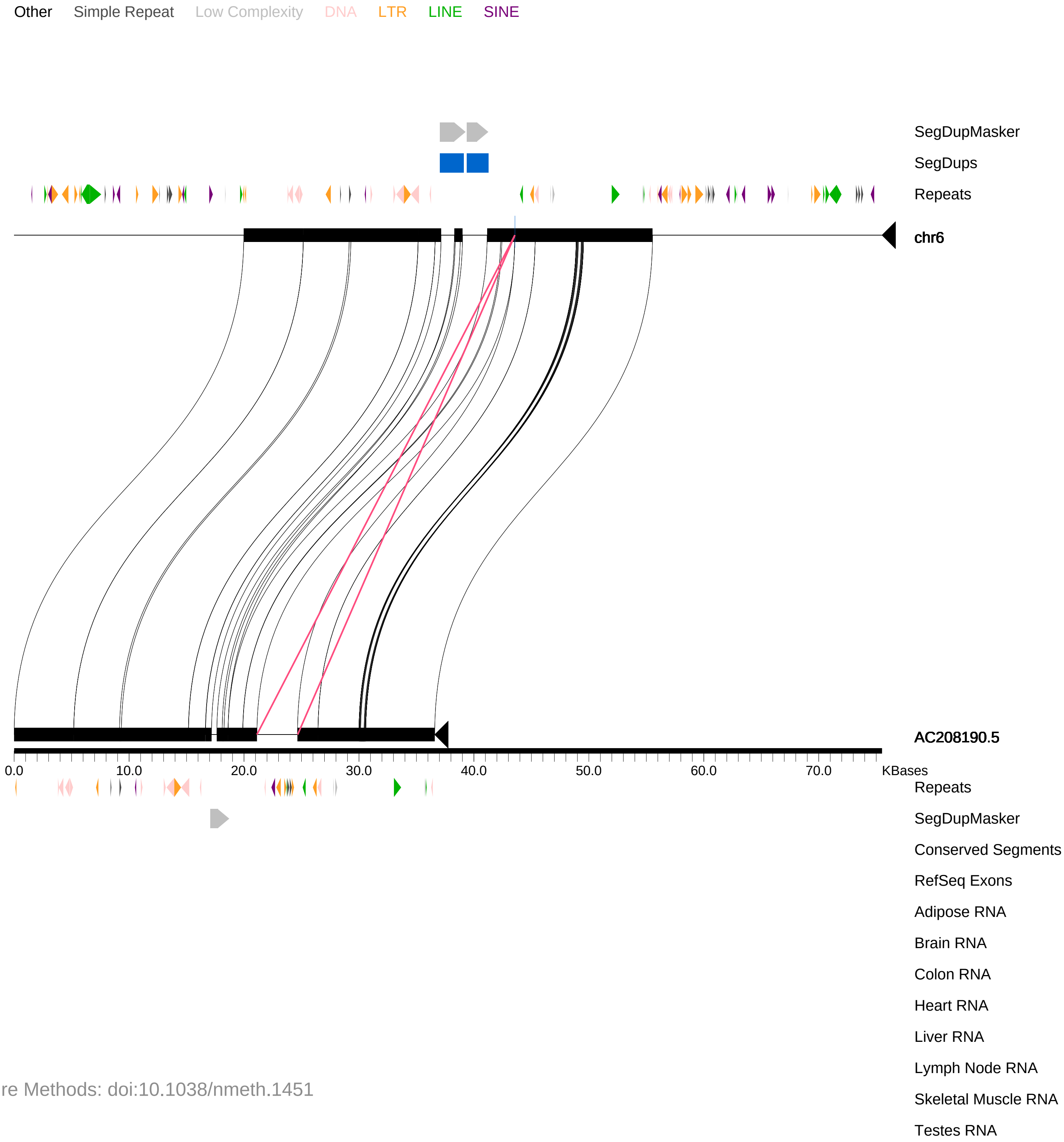
Insertion Size: 17268



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC208190.rc.fa

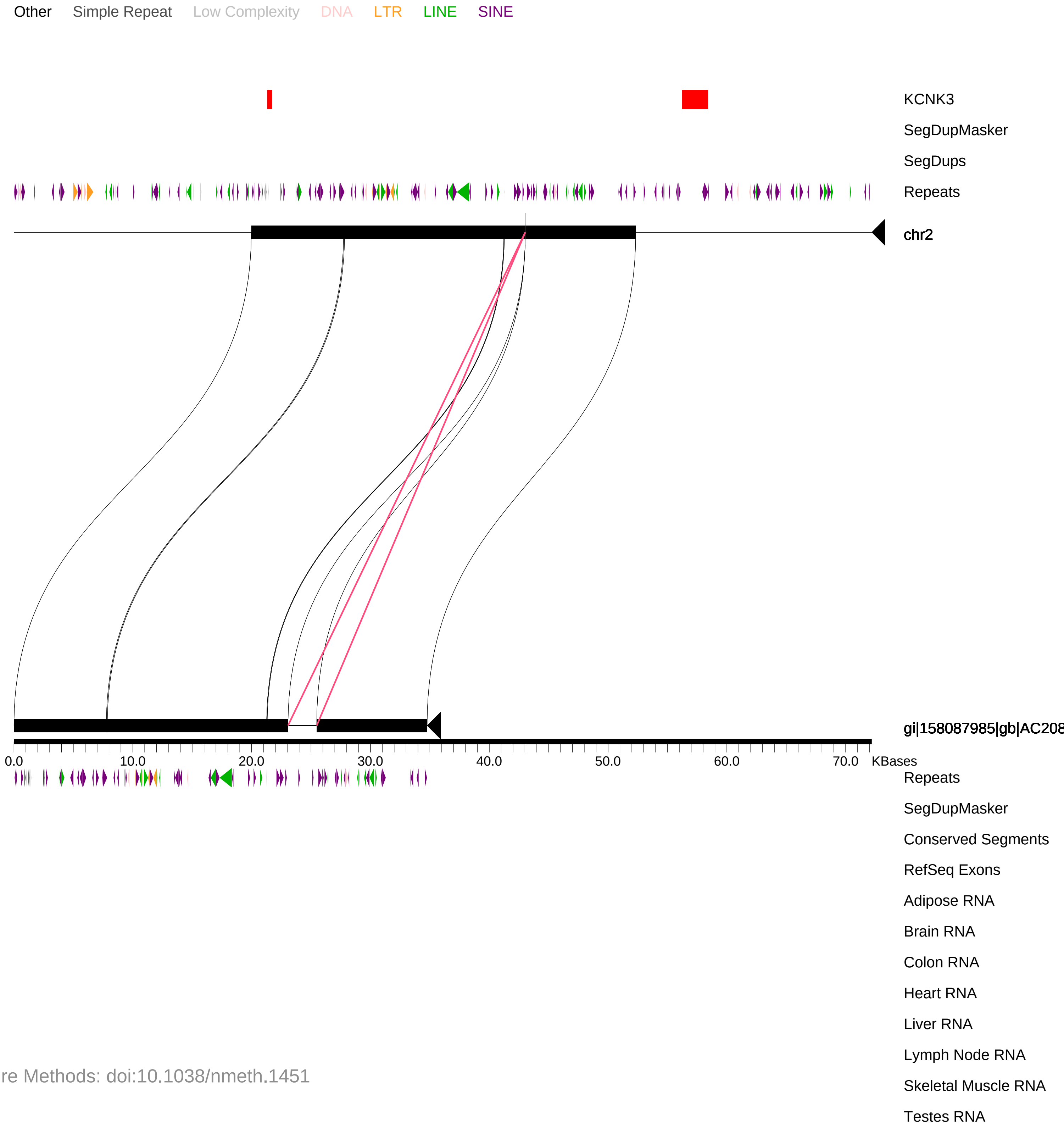
Insertion Size: 3547



Nature Methods: doi:10.1038/nmeth.1451

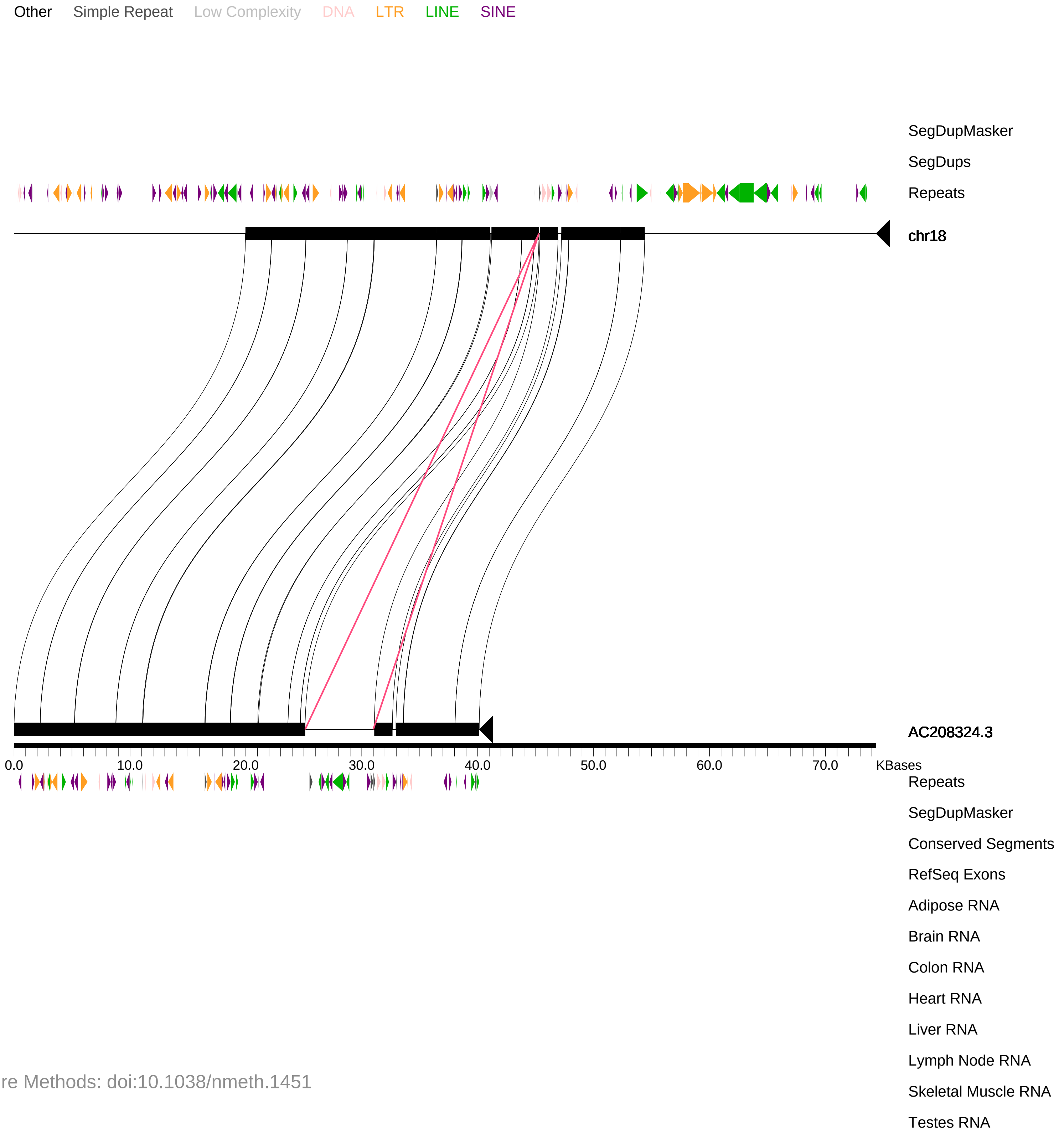
Clone file = AC208323.fa

Insertion Size: 2419



Clone file = AC208324.rc.fa

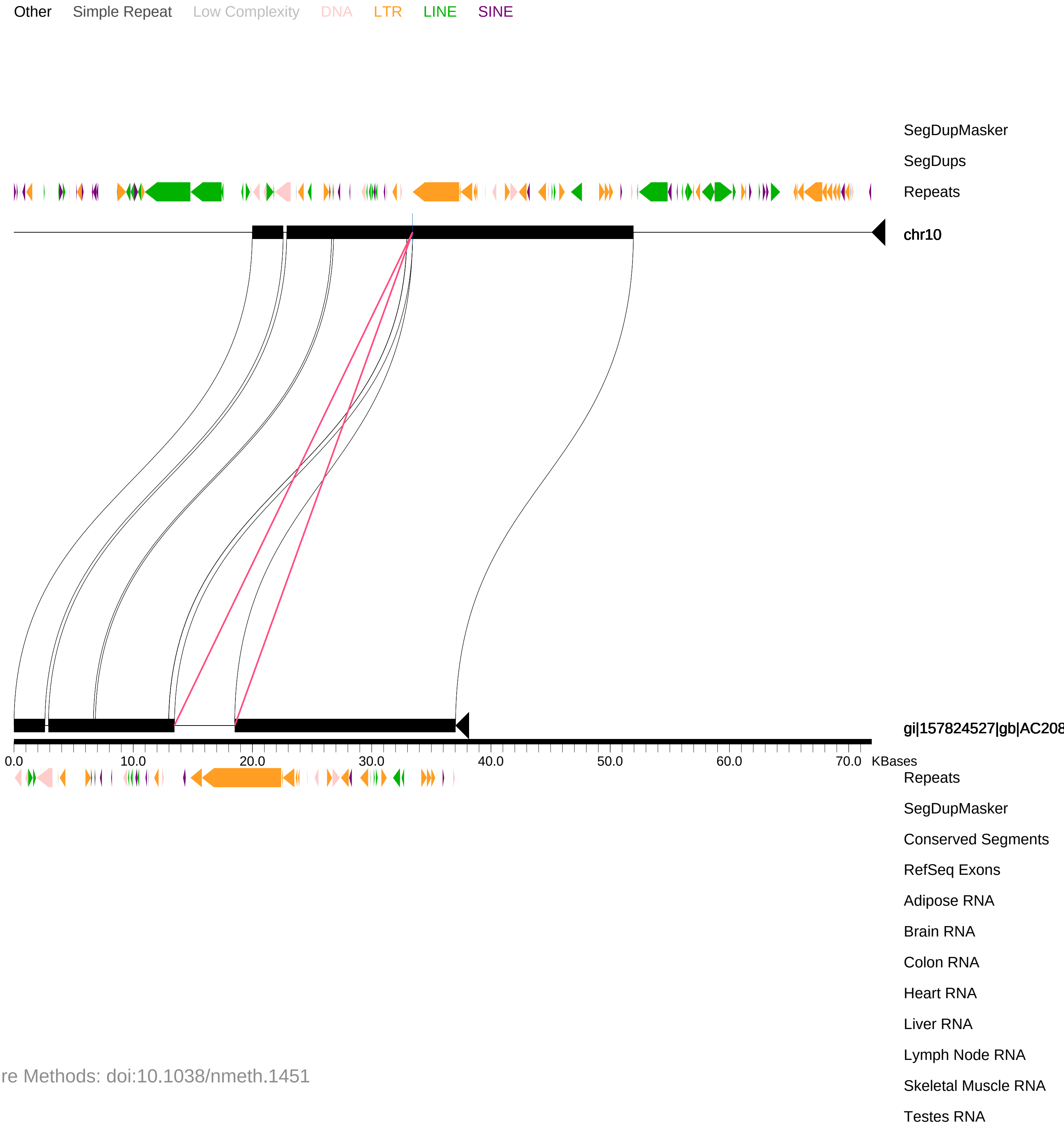
Insertion Size: 5878



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC208502.fa

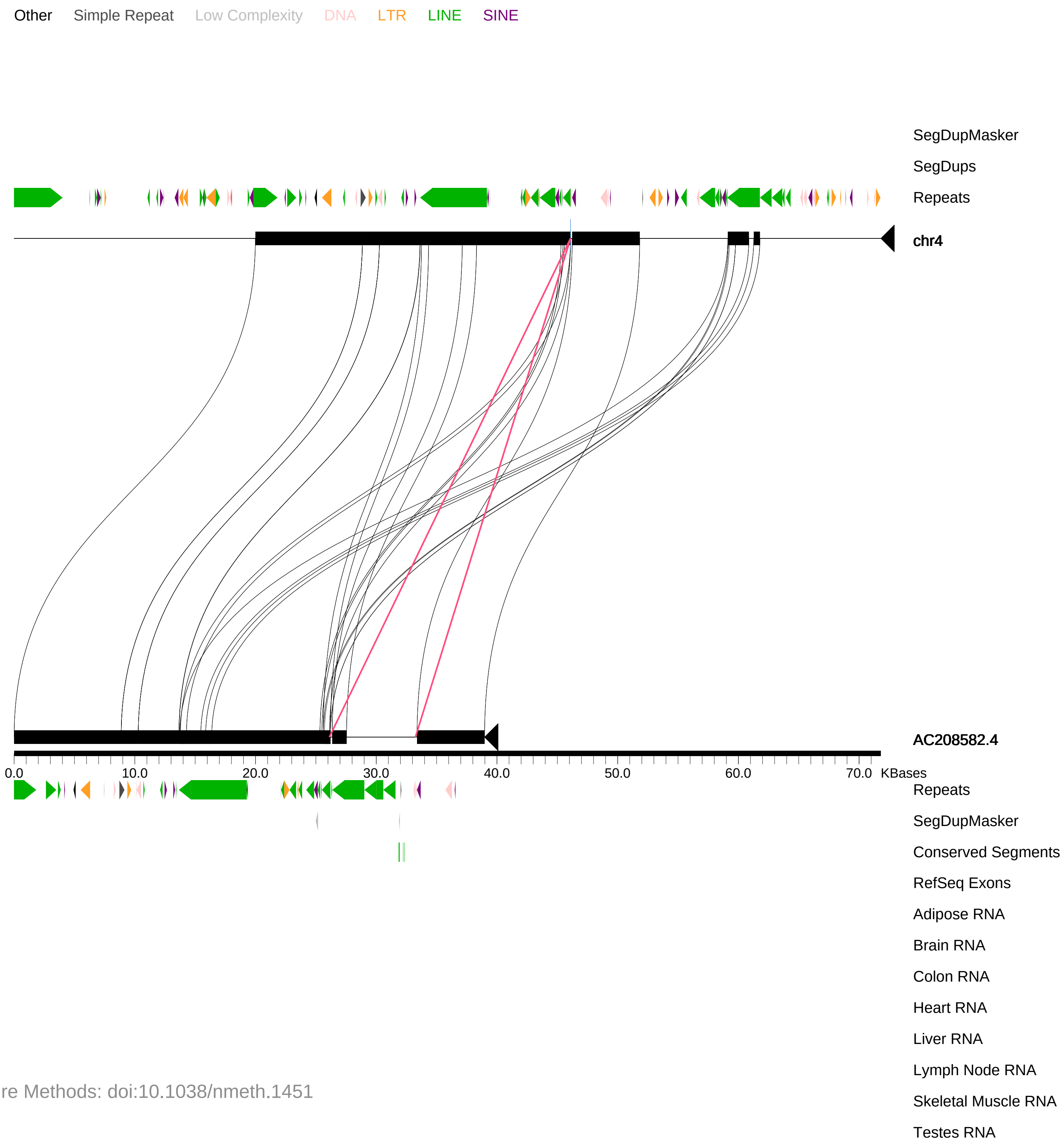
Insertion Size: 5067



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC208582.rc.fa

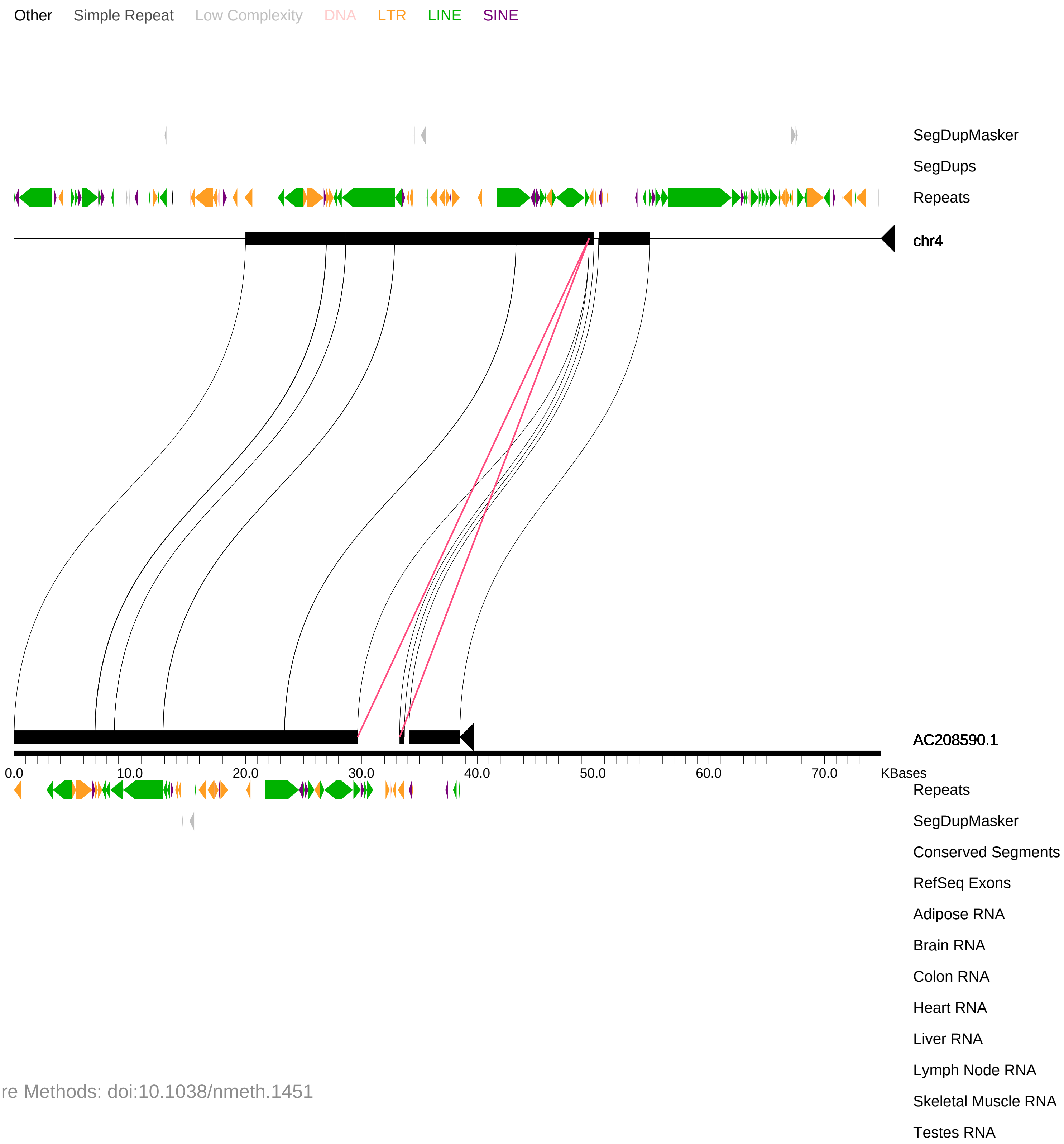
Insertion Size: 7129



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC208590.rc.fa

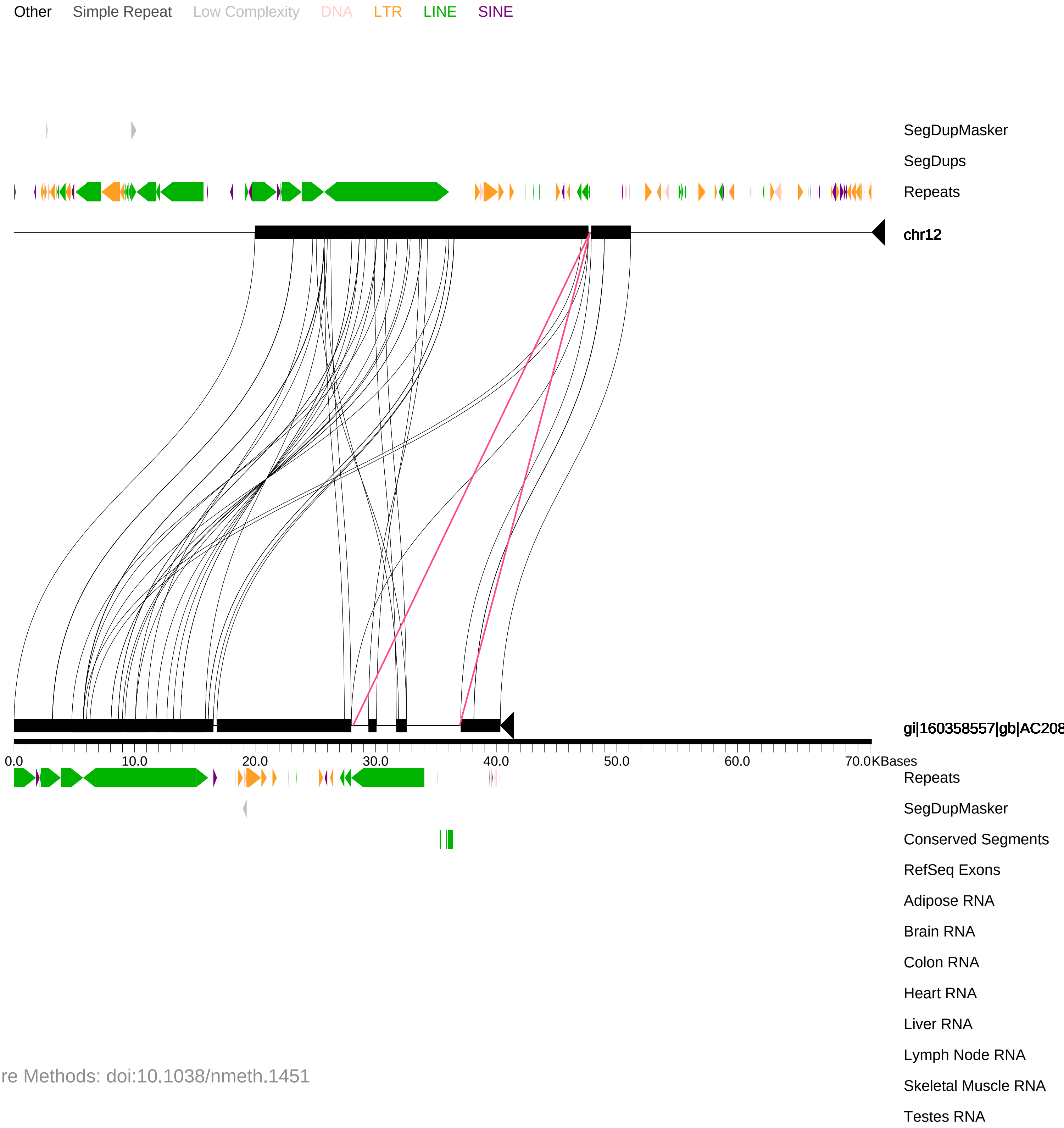
Insertion Size: 3634



Nature Methods: doi:10.1038/nmeth.1451

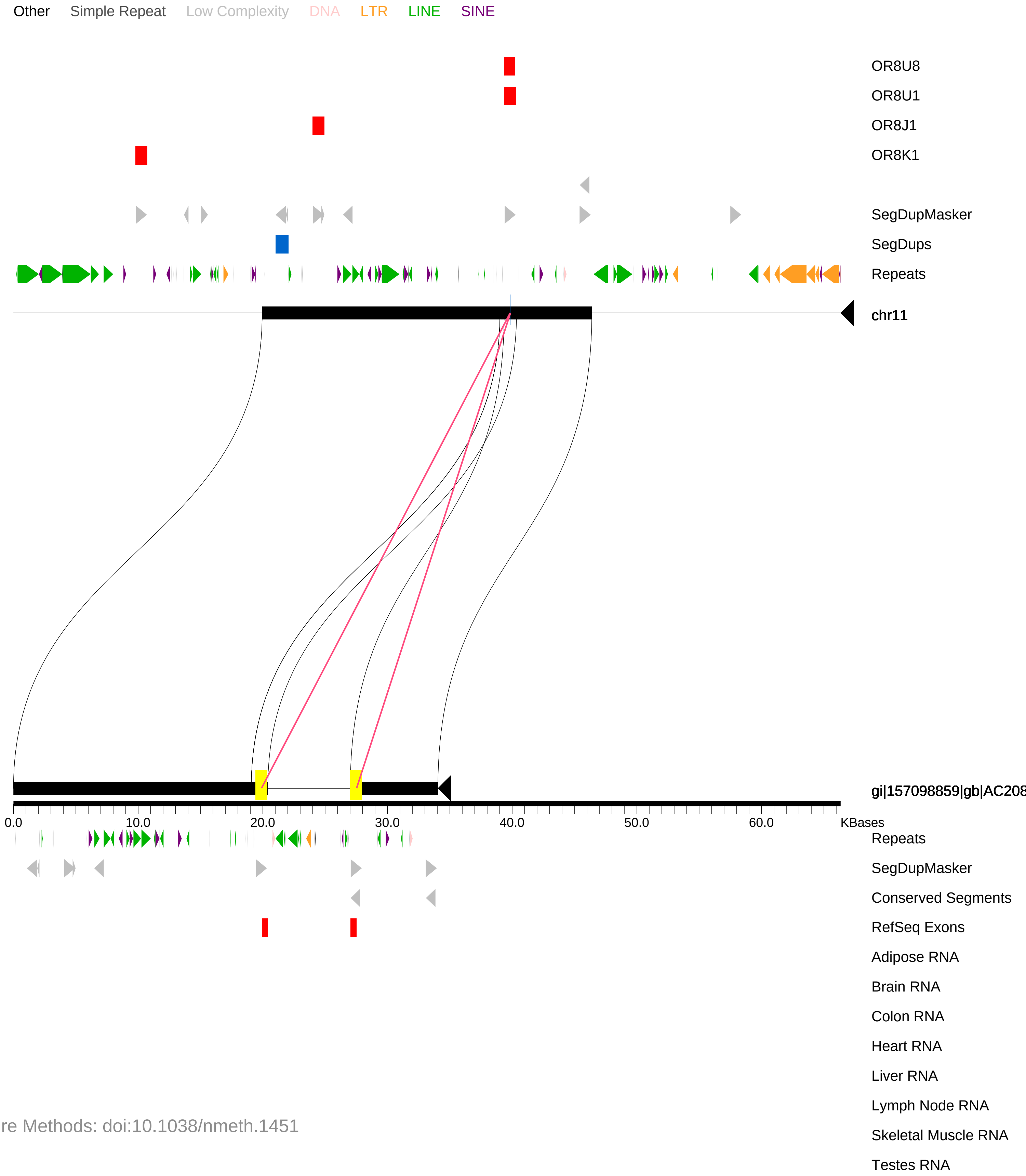
Clone file = AC208716.fa

Insertion Size: 8849



Clone file = AC208786.fa

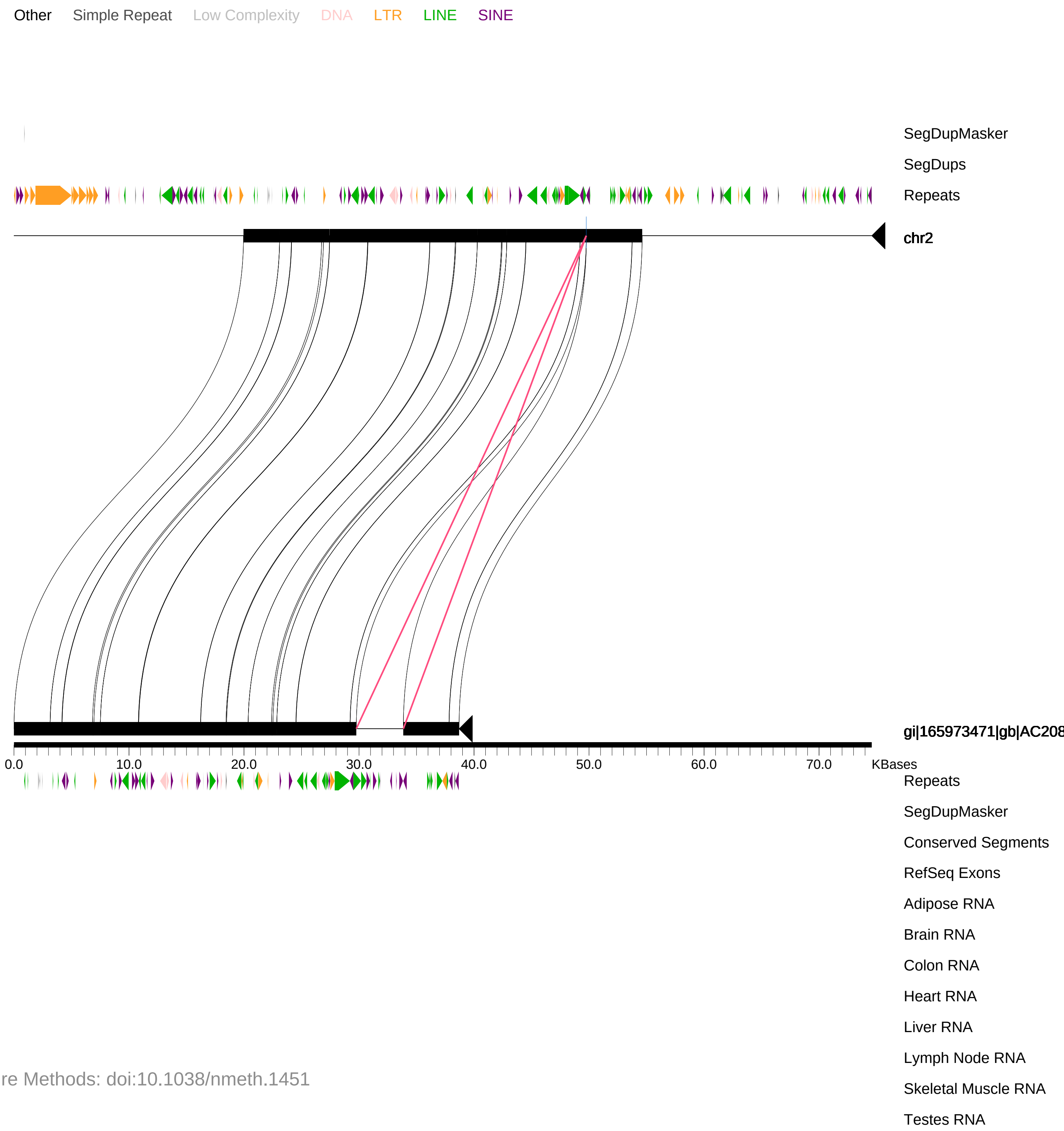
Insertion Size: 7628



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC208871.fa

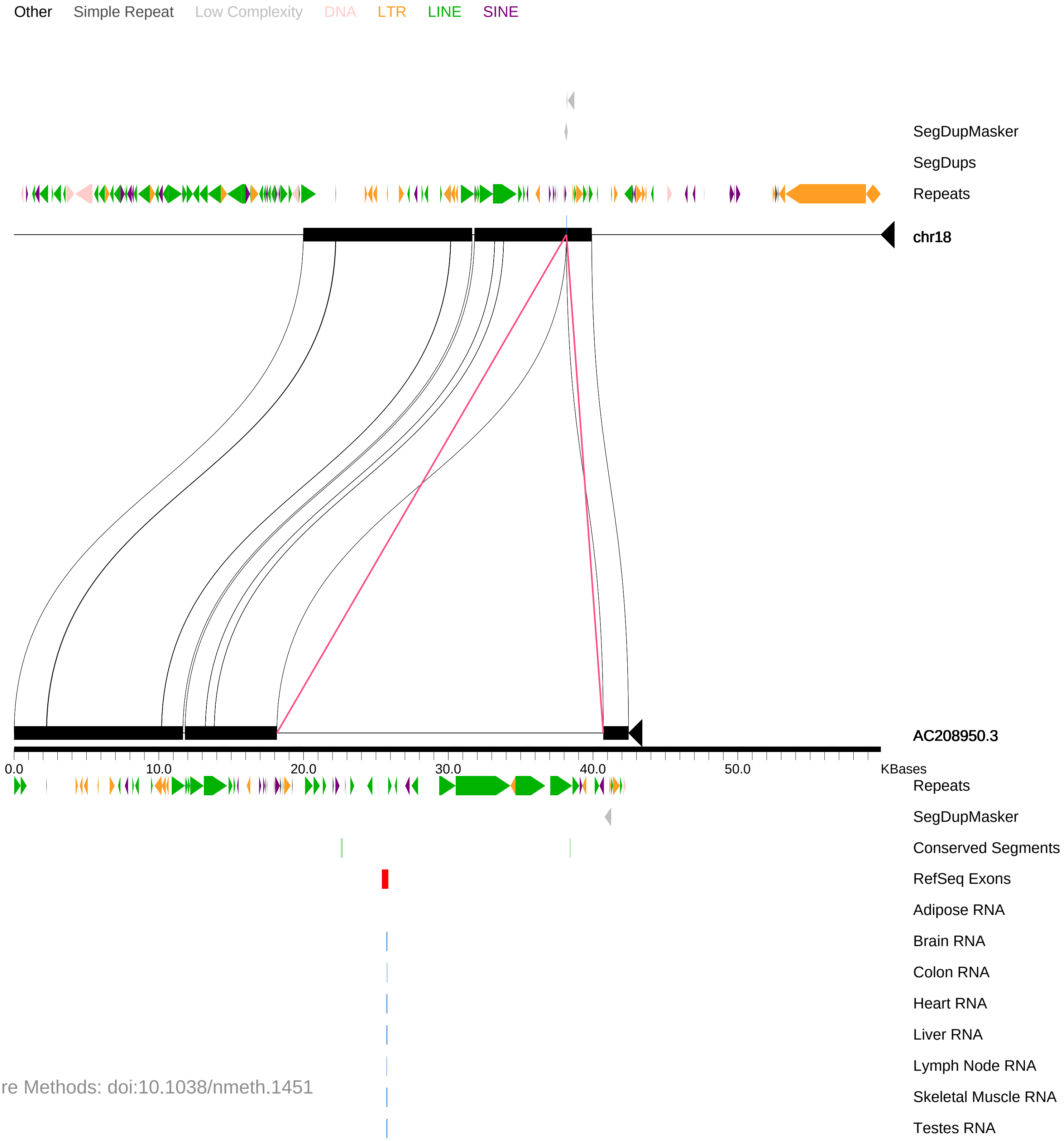
Insertion Size: 4094



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC208950.rc.fa

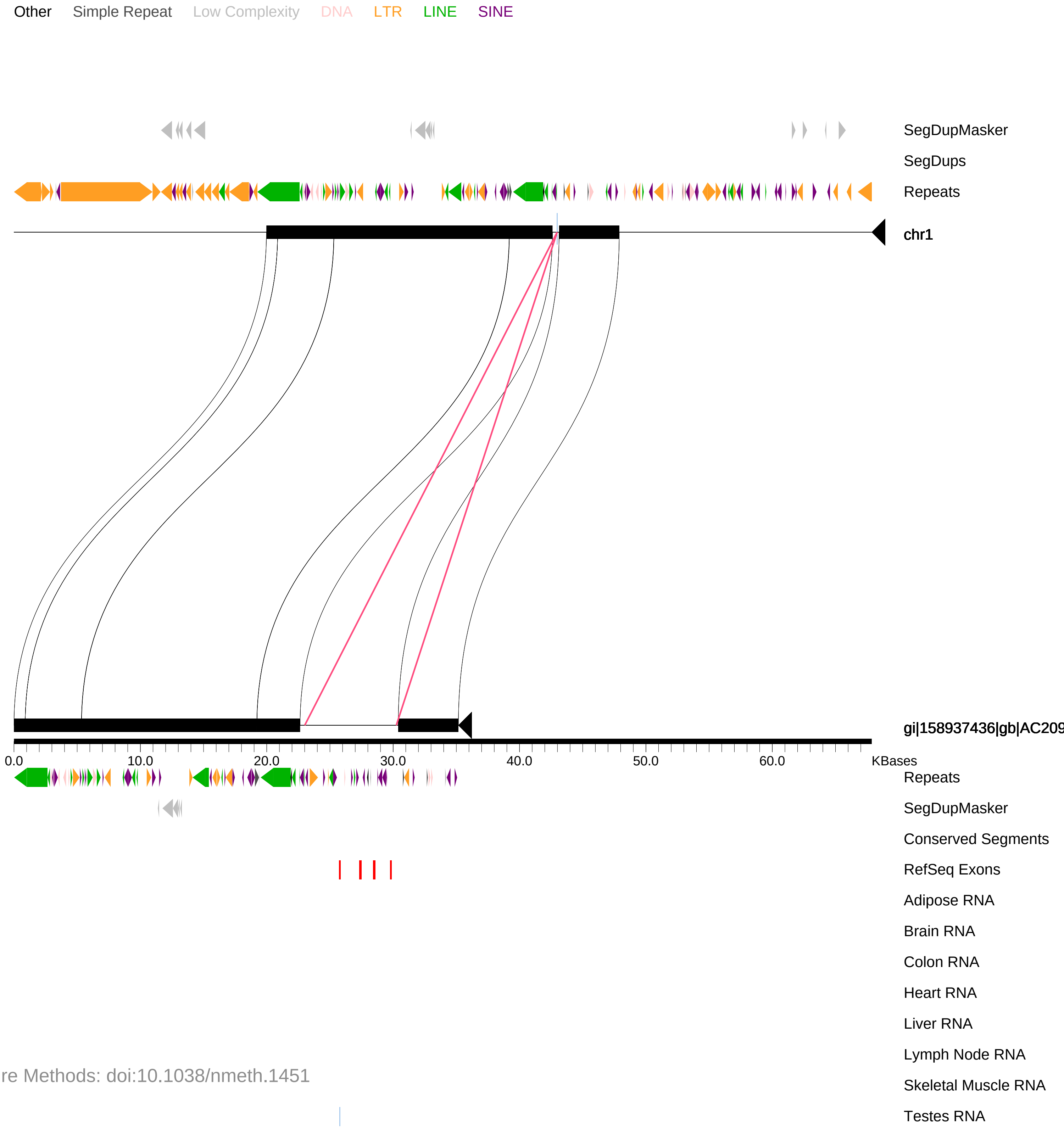
Insertion Size: 22557



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC209007.fa

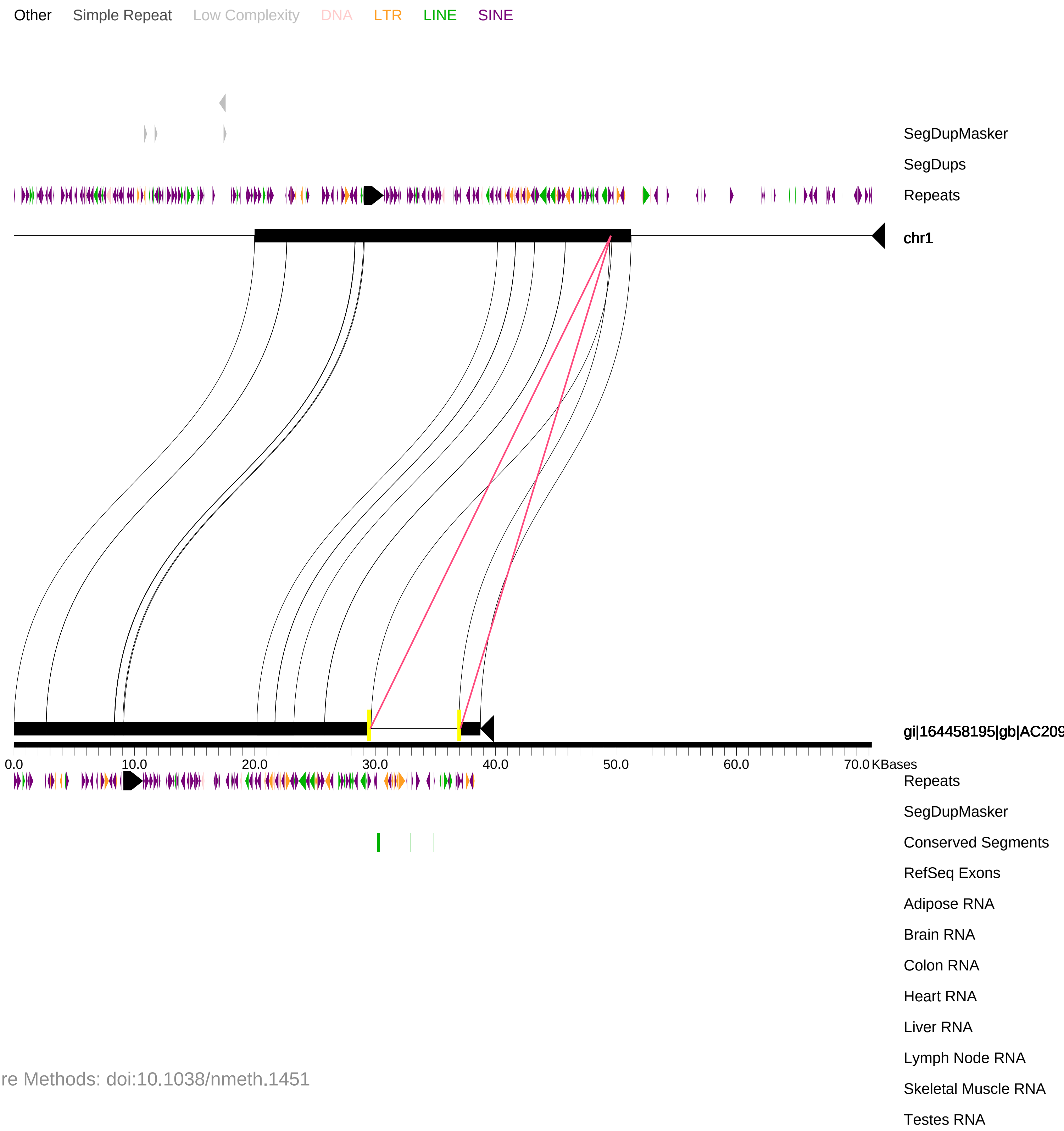
Insertion Size: 7246



Nature Methods: doi:10.1038/nmeth.1451

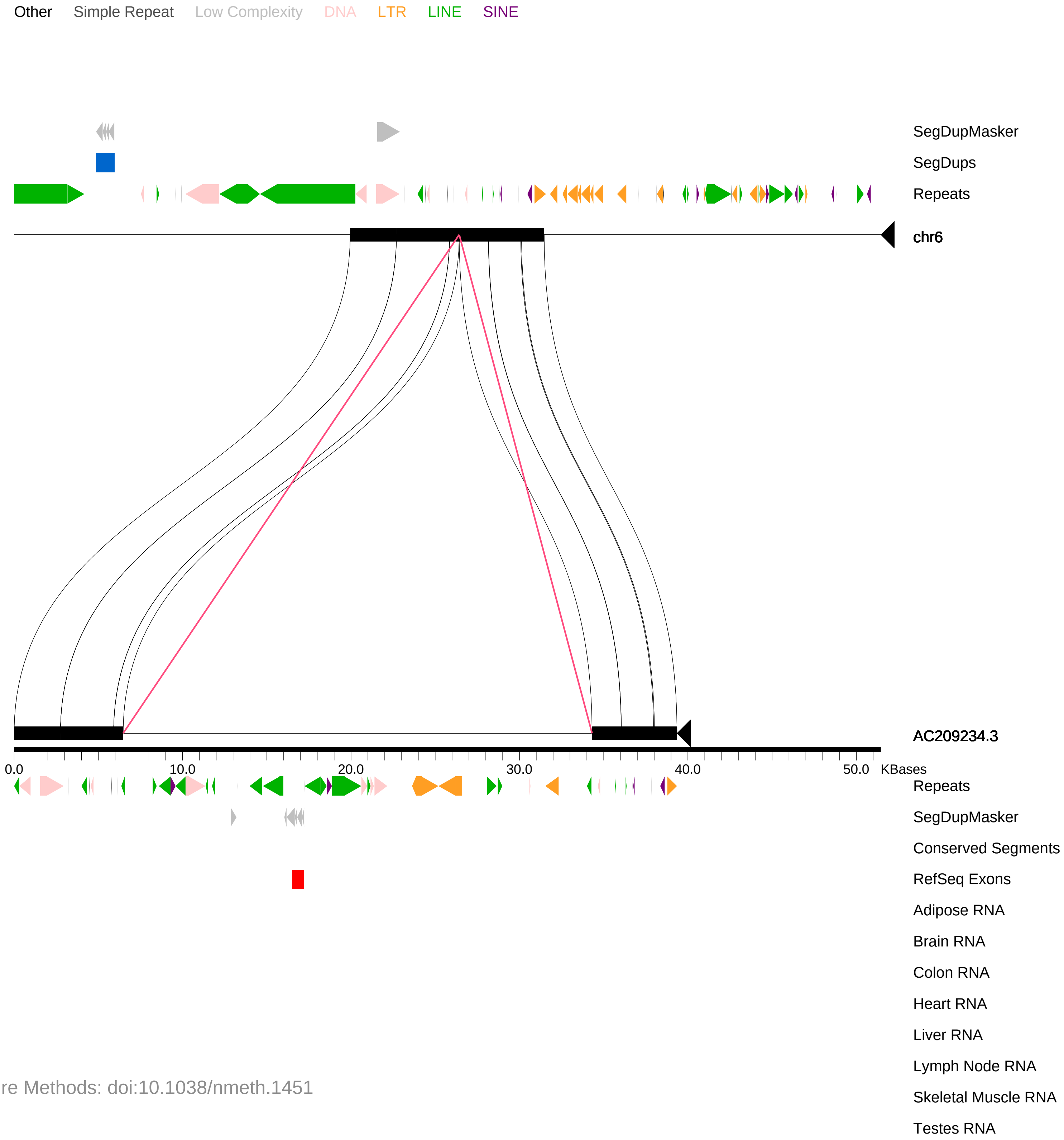
Clone file = AC209232.fa

Insertion Size: 7488



Clone file = AC209234.rc.fa

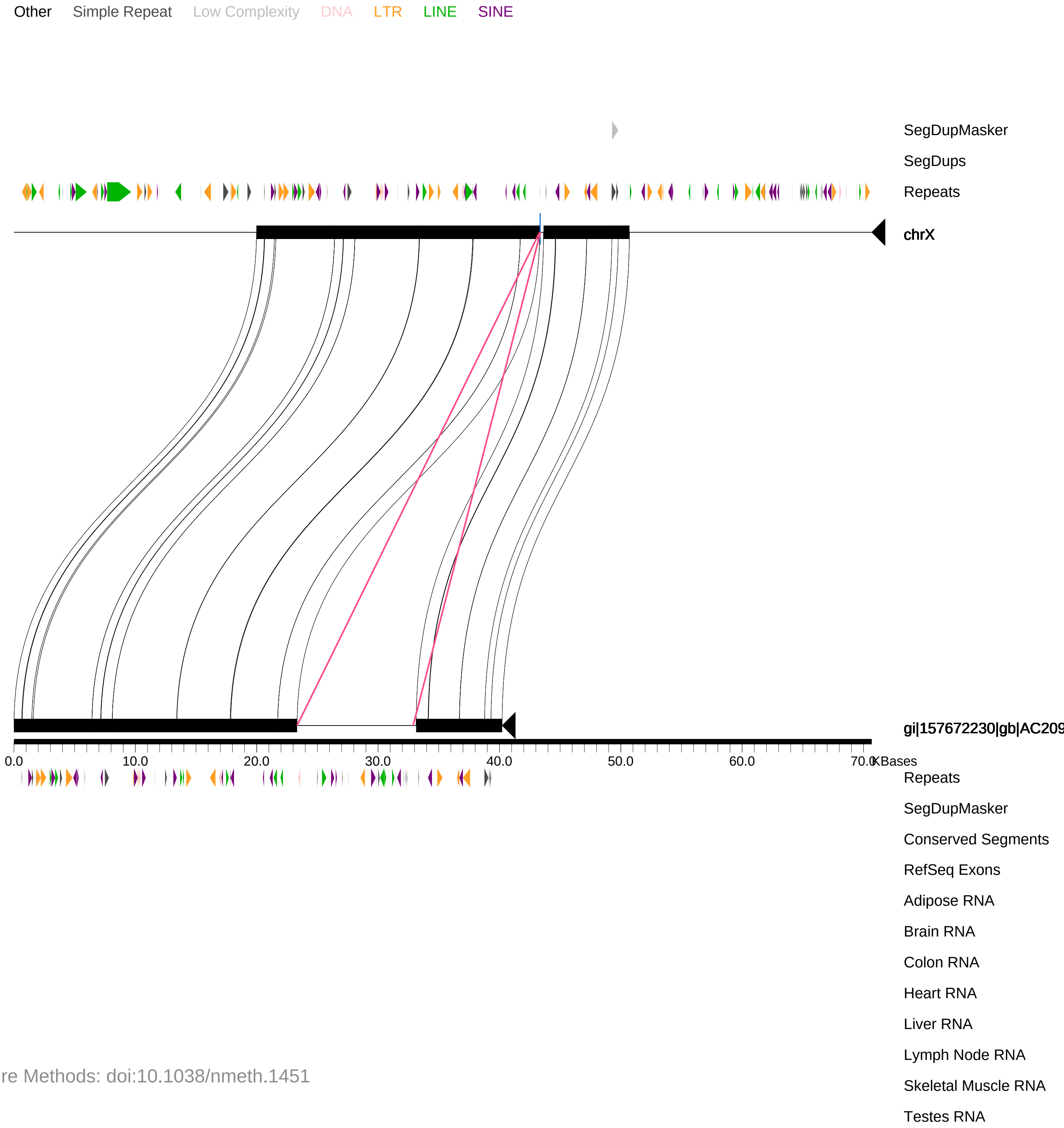
Insertion Size: 27841



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC209283.fa

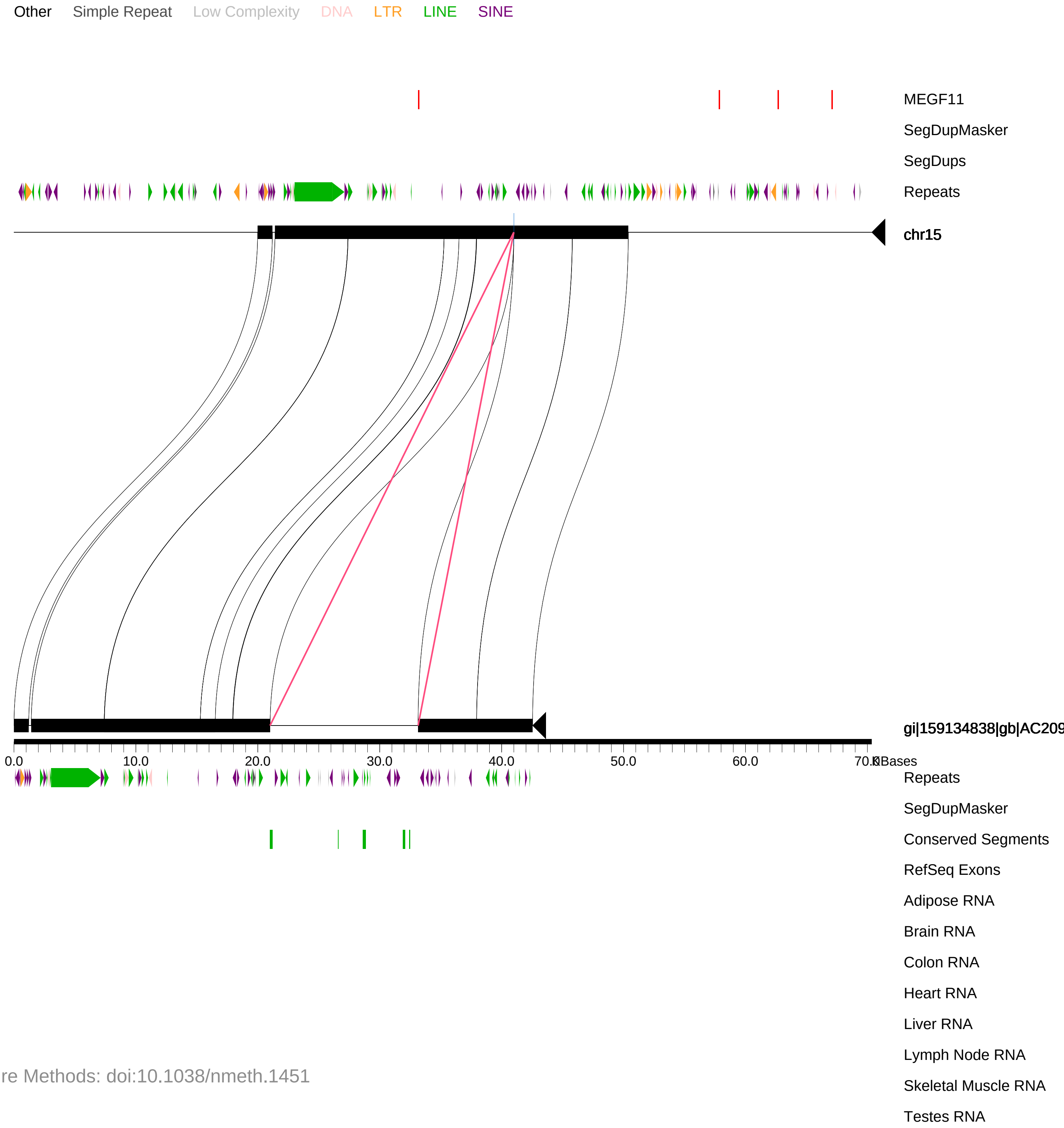
Insertion Size: 9563



Nature Methods: doi:10.1038/nmeth.1451

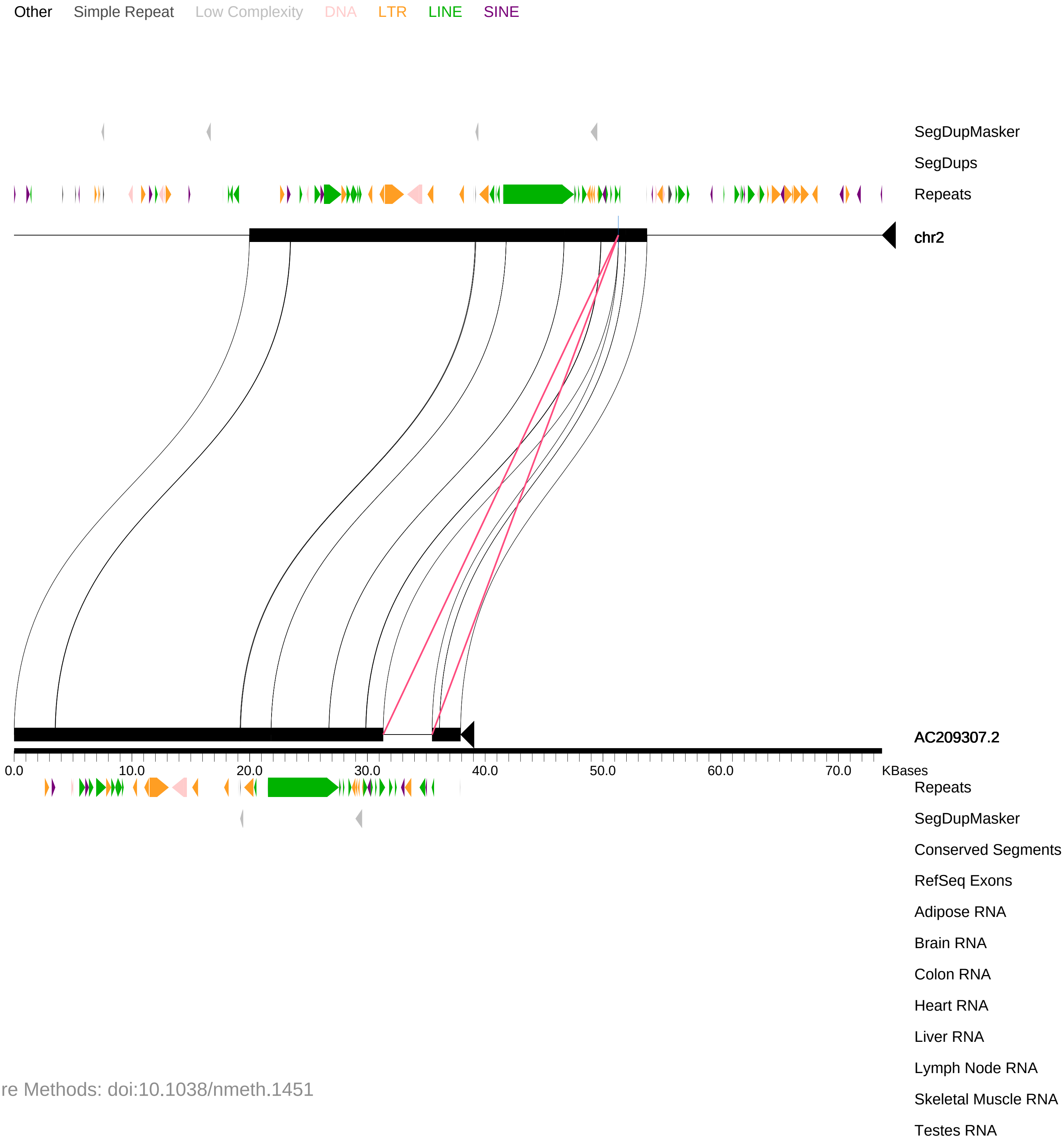
Clone file = AC209298.fa

Insertion Size: 12146



Clone file = AC209307.rc.fa

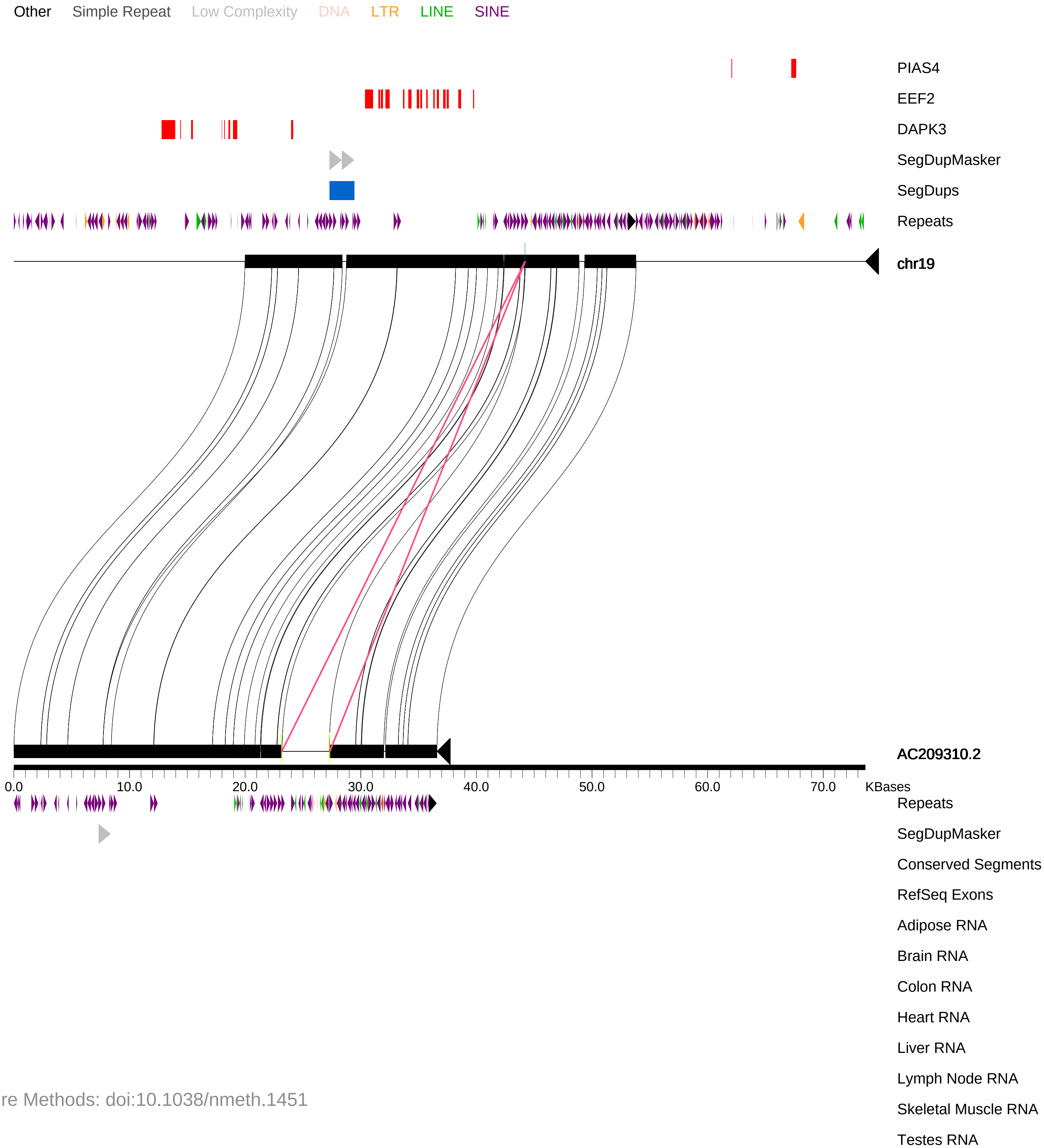
Insertion Size: 4150



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC209310.rc.fa

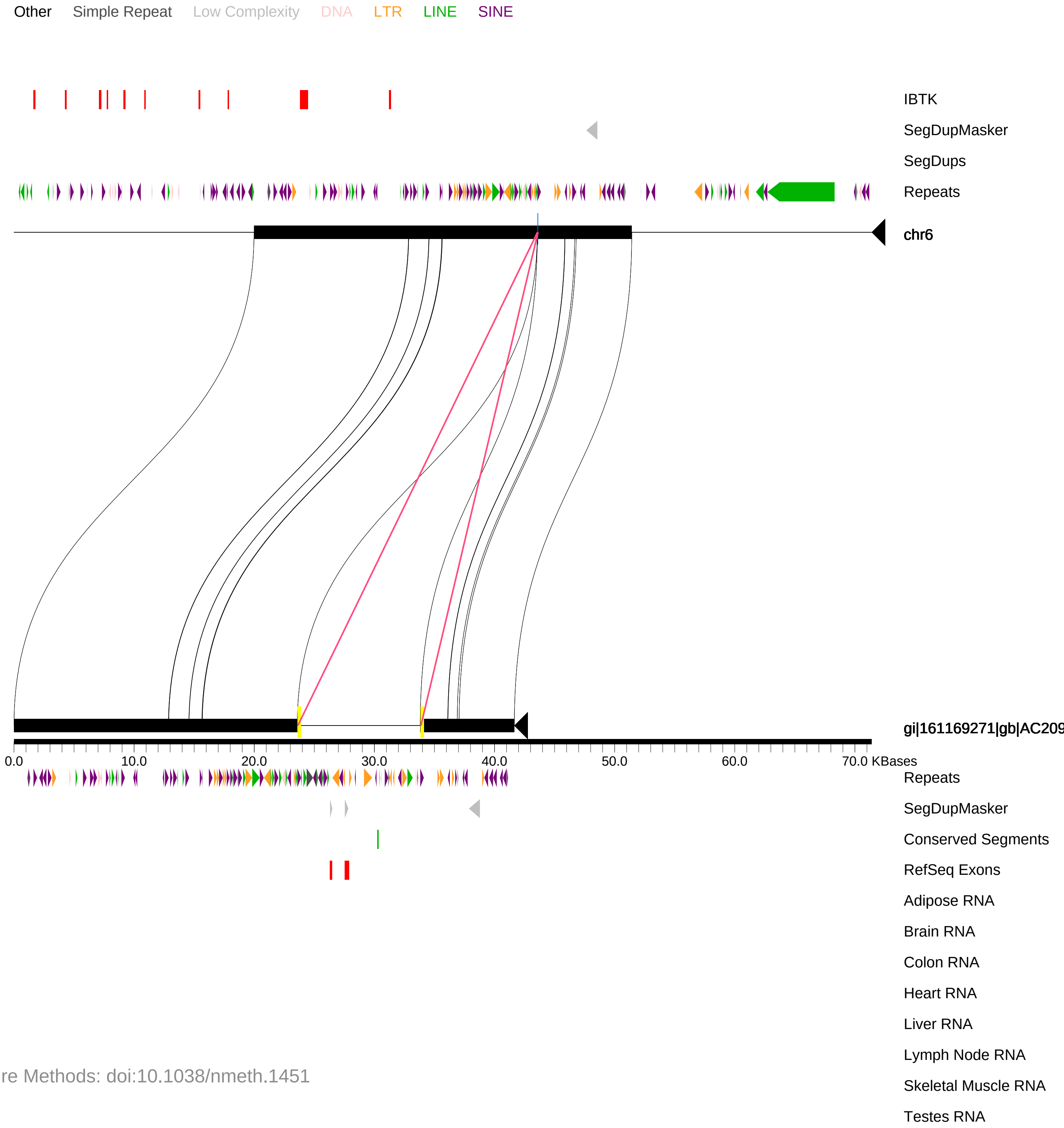
Insertion Size: 4158



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC209420.fa

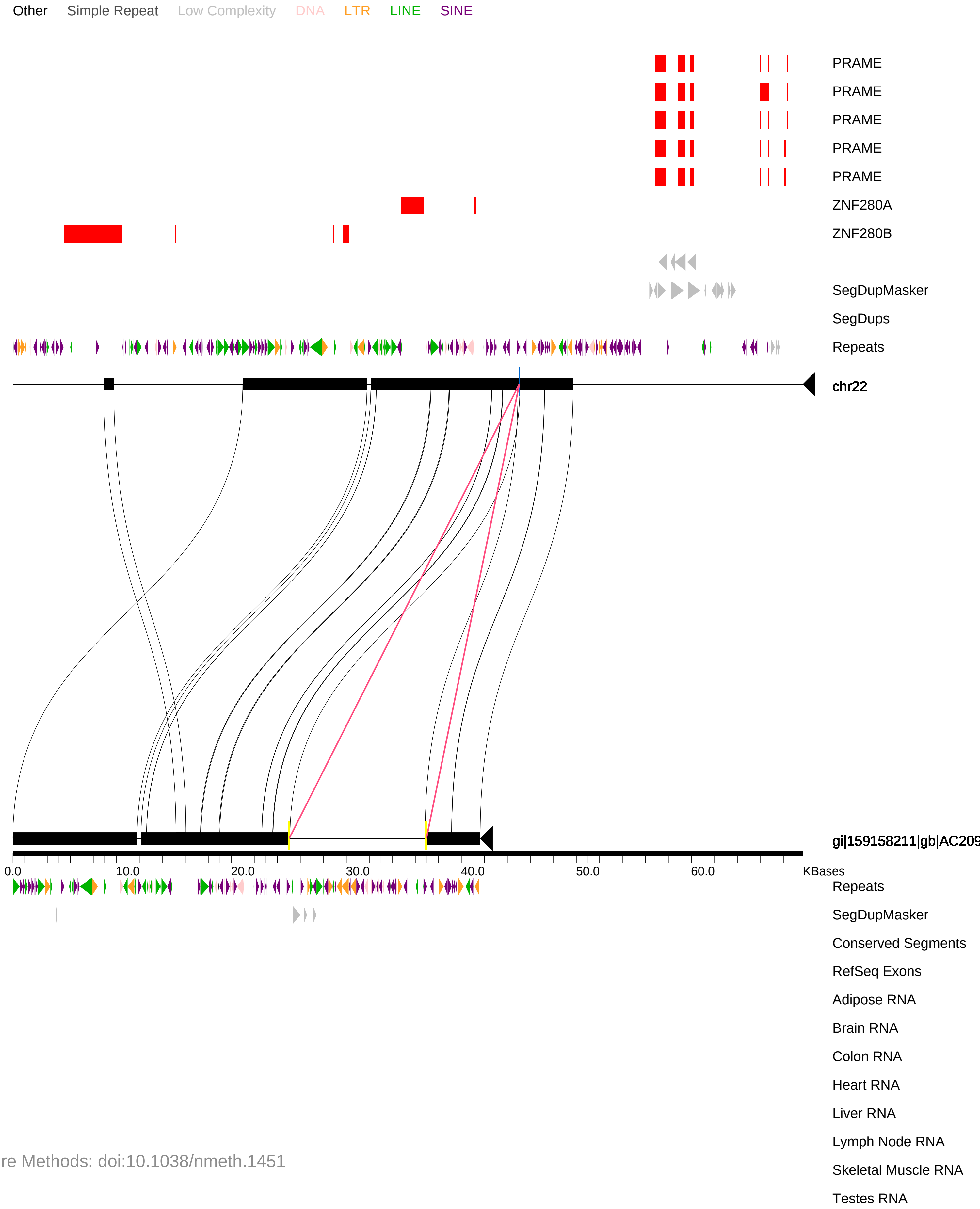
Insertion Size: 10243



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC209546.fa

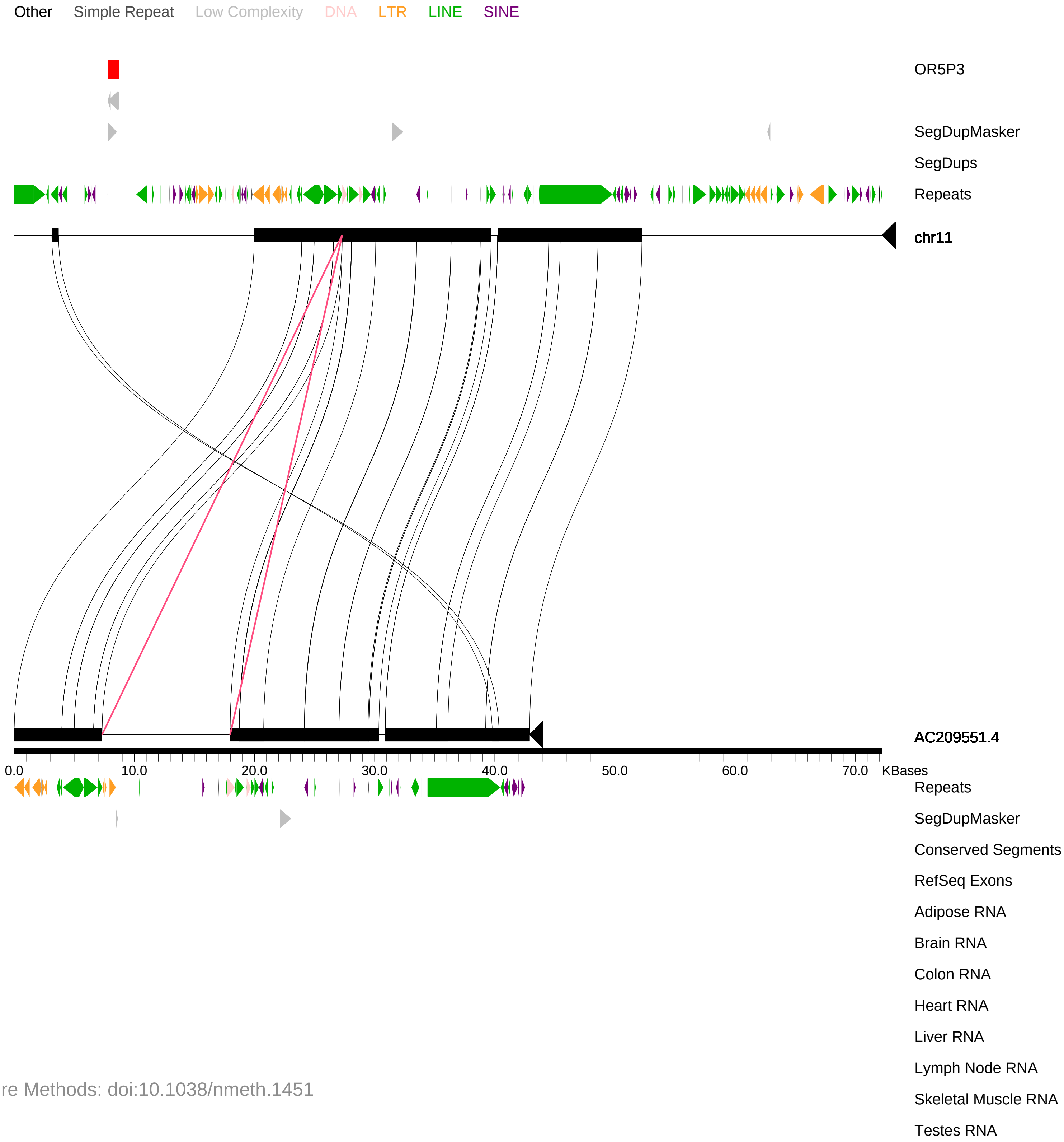
Insertion Size: 11929



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC209551.rc.fa

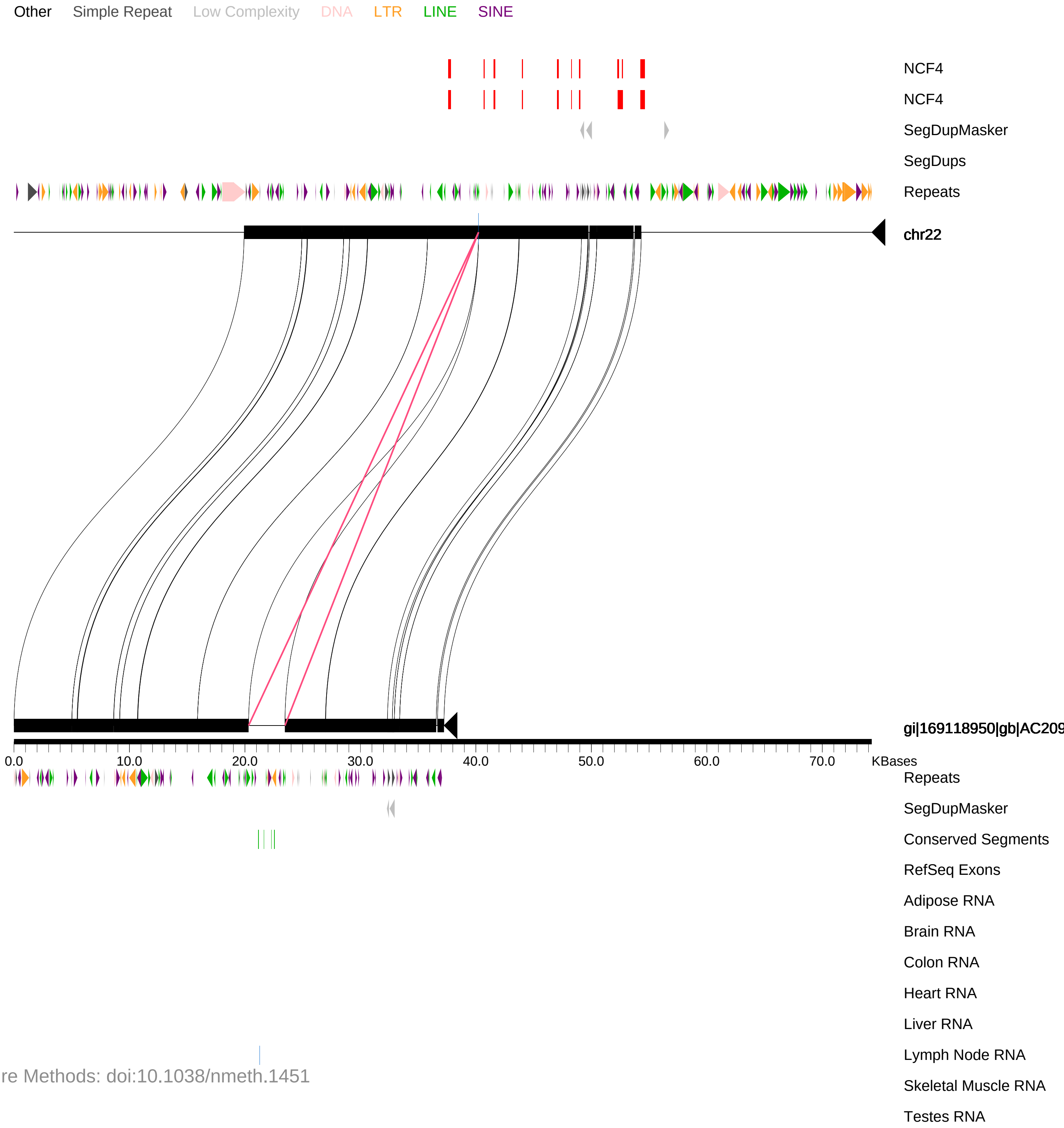
Insertion Size: 10658



Nature Methods: doi:10.1038/nmeth.1451

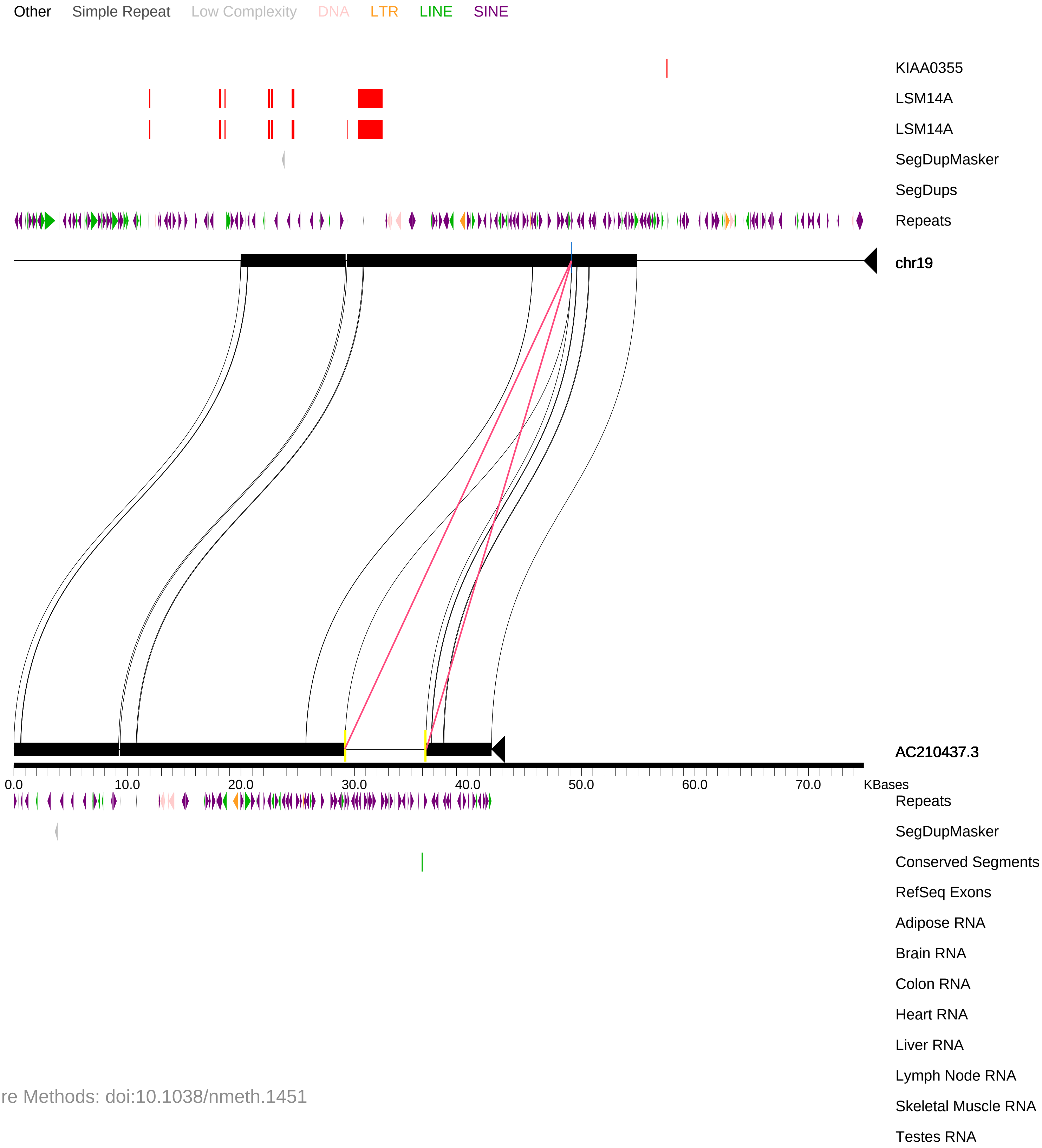
Clone file = AC209618.fa

Insertion Size: 3151



Clone file = AC210437.rc.fa

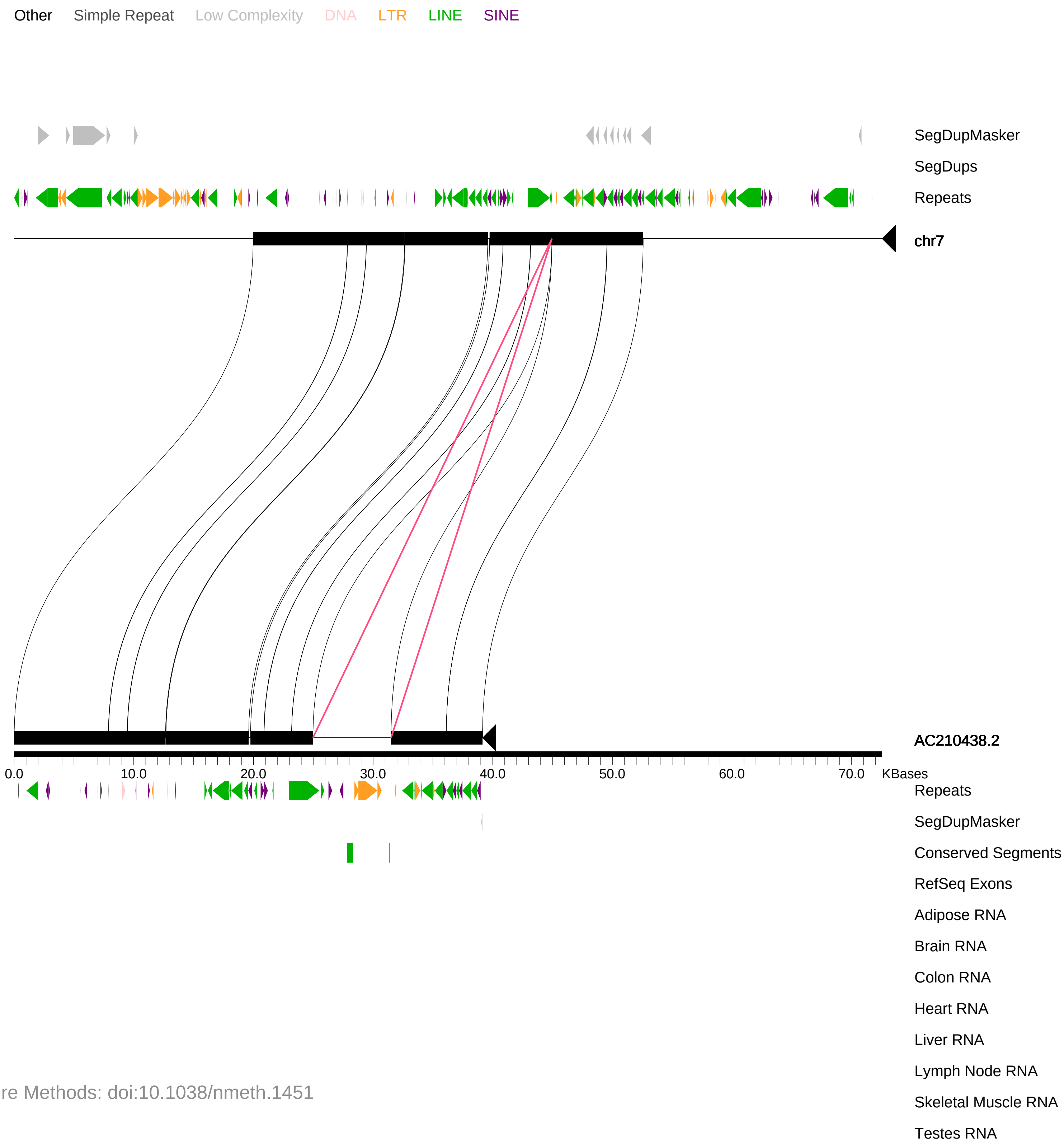
Insertion Size: 7192



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC210438.rc.fa

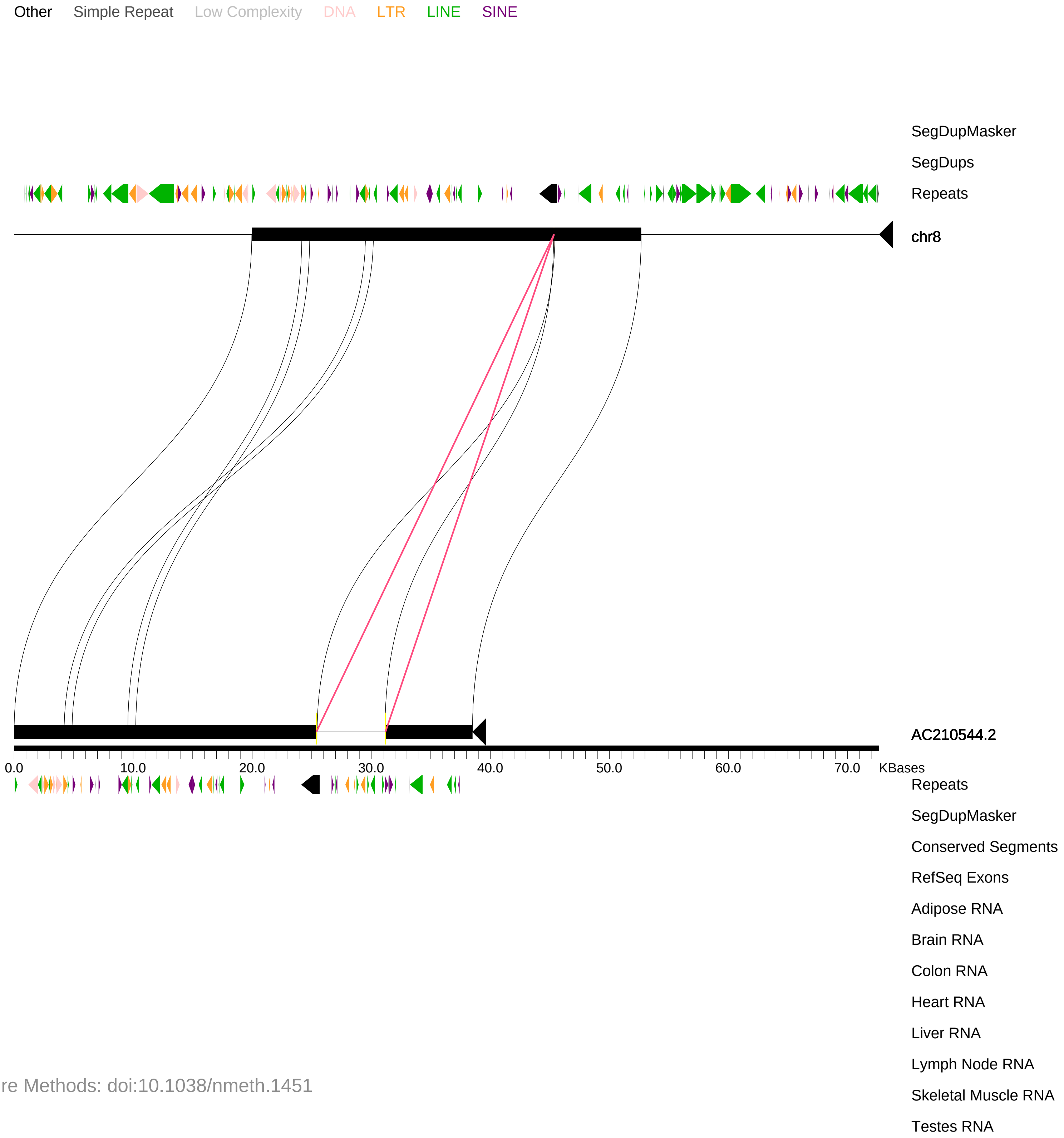
Insertion Size: 6530



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC210544.rc.fa

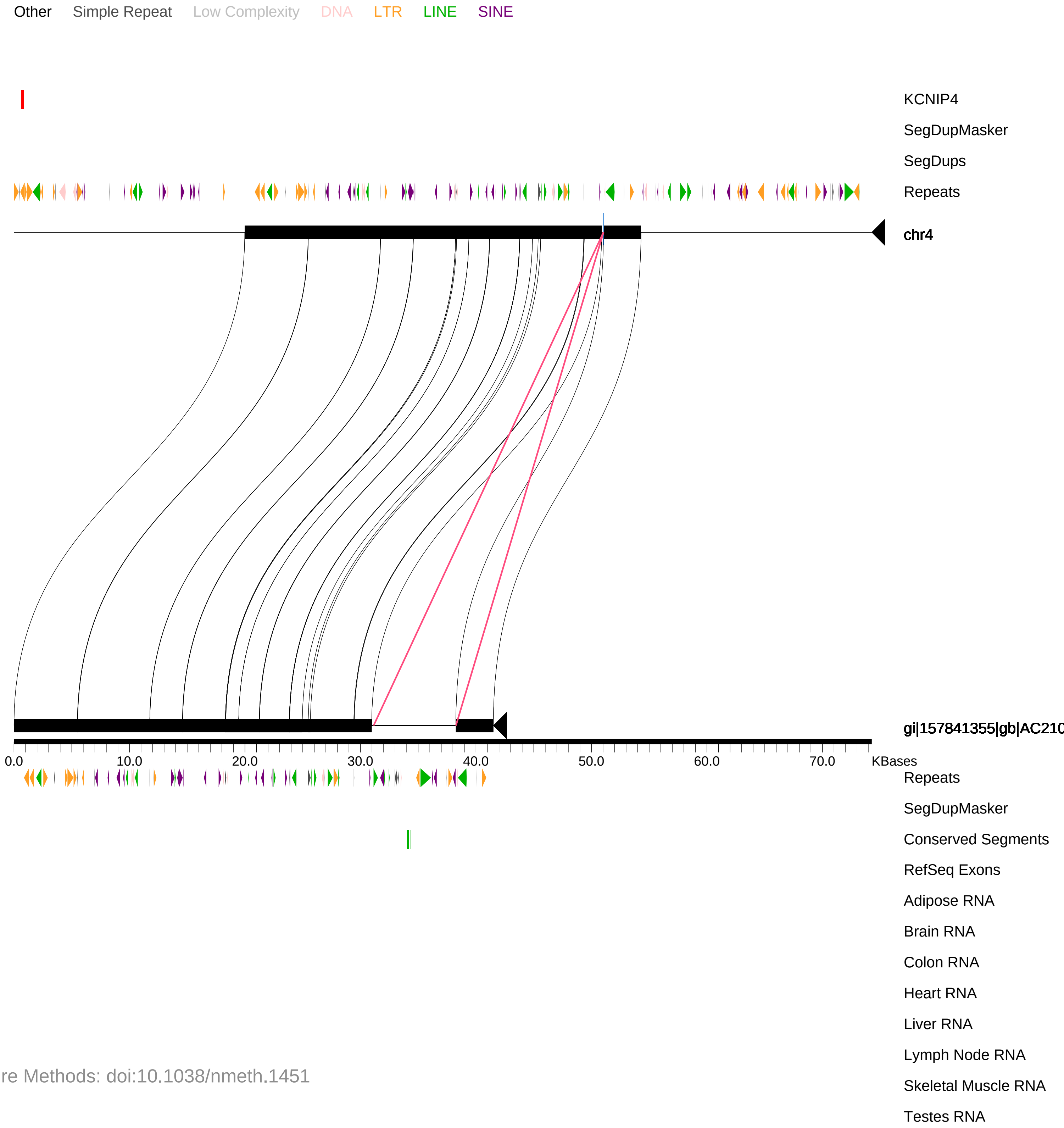
Insertion Size: 5805



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC210756.fa

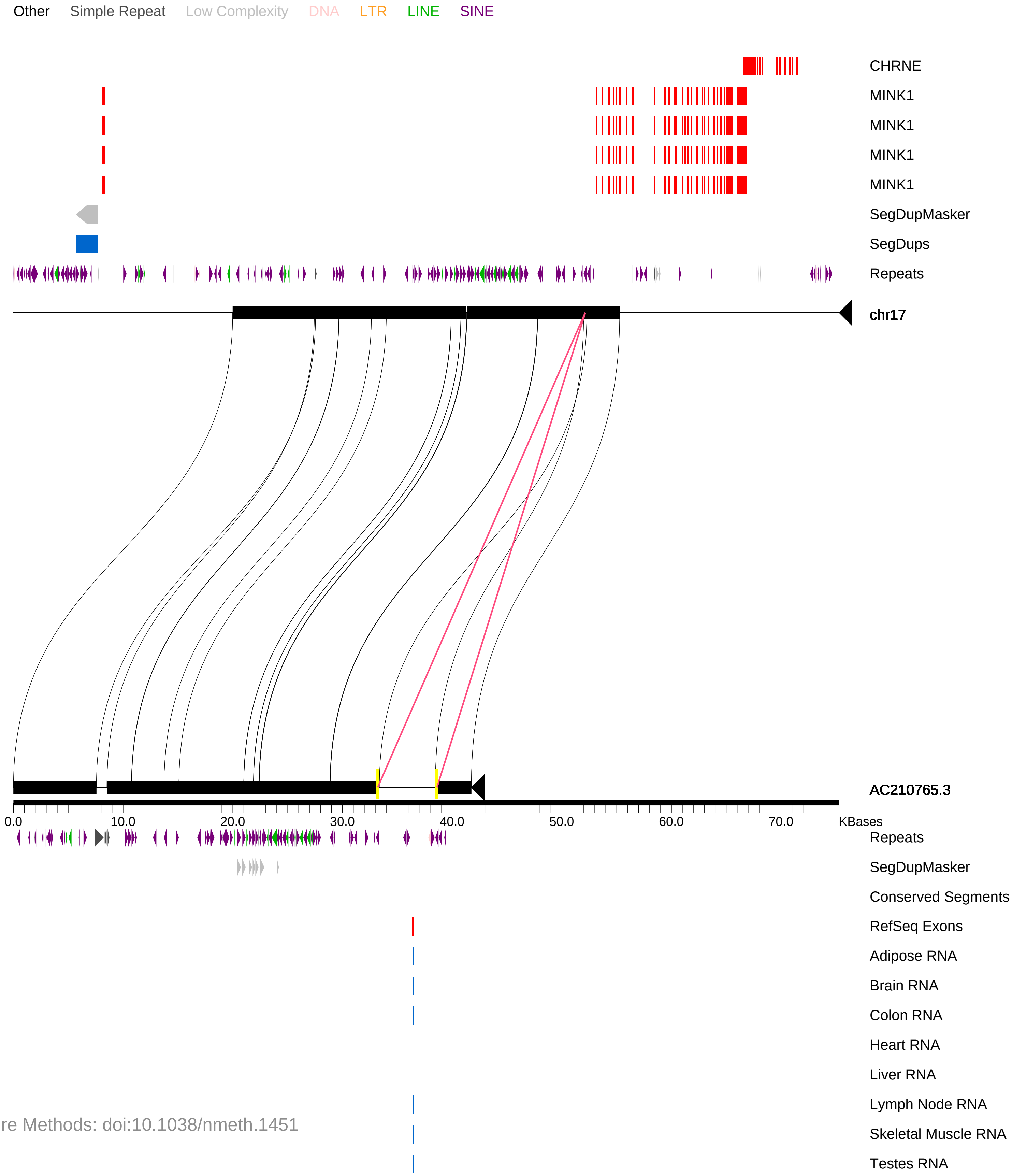
Insertion Size: 7128



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC210765.rc.fa

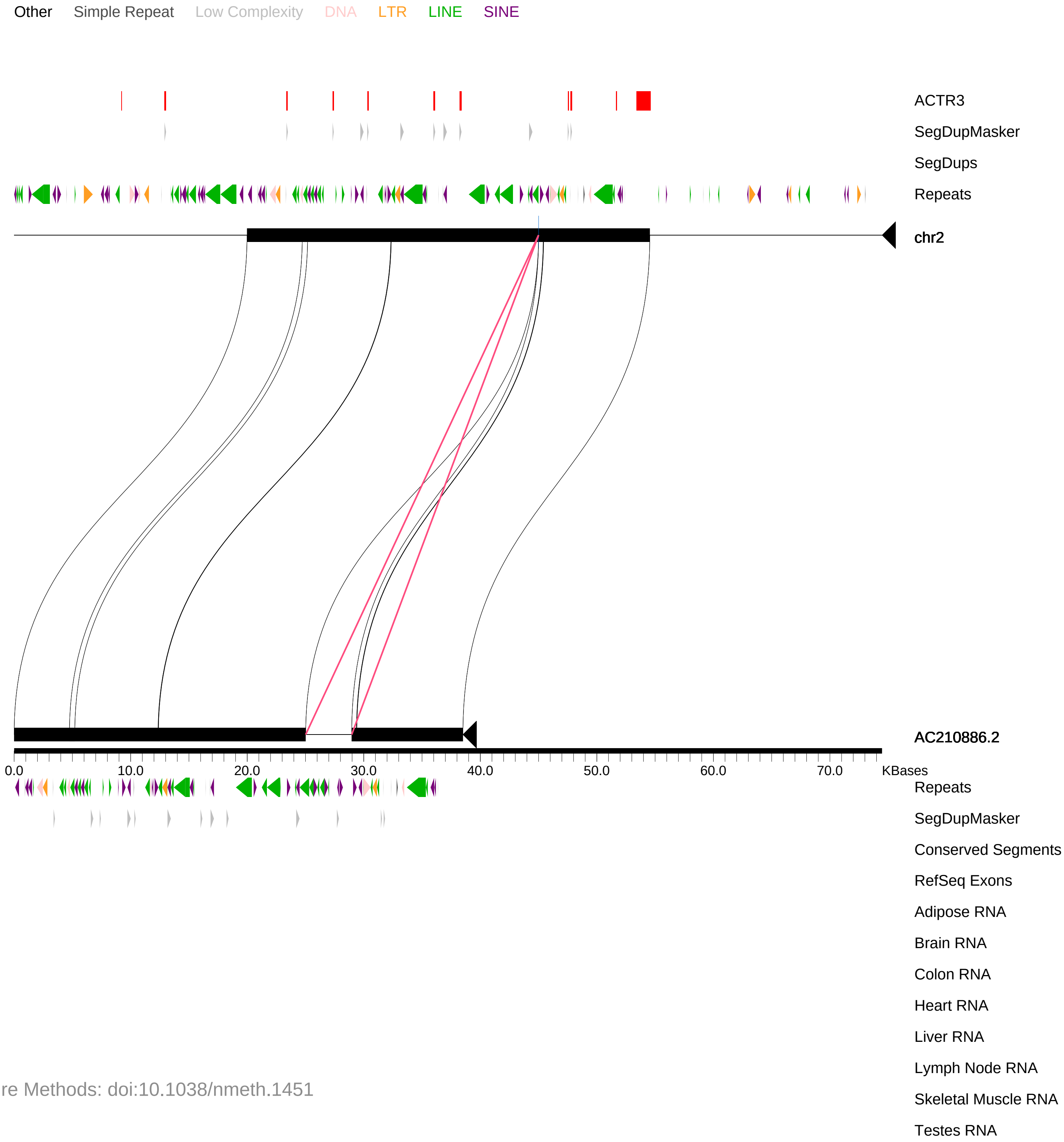
Insertion Size: 5405



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC210886.rc.fa

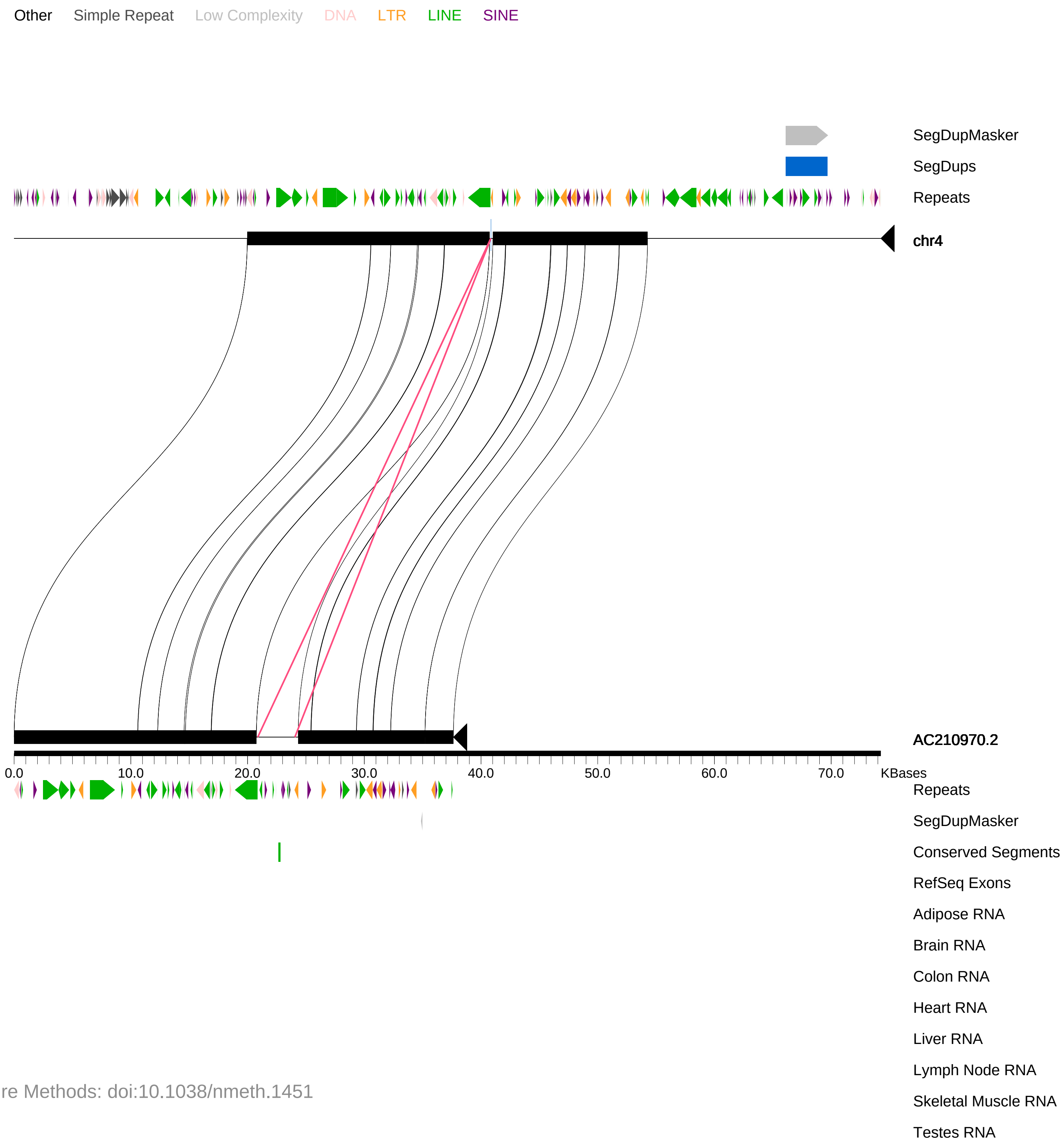
Insertion Size: 3949



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC210970.rc.fa

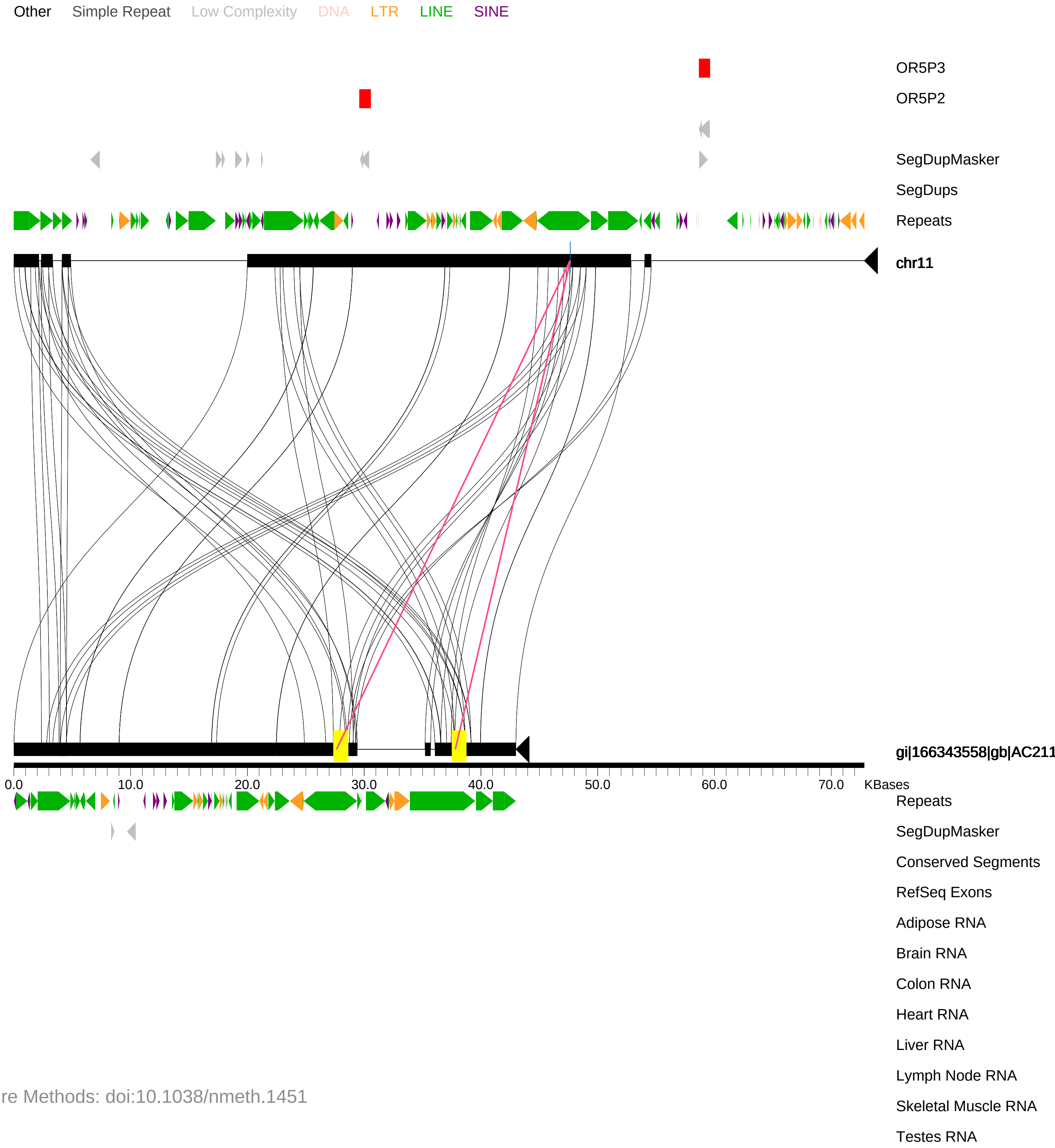
Insertion Size: 3179



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC211399.fa

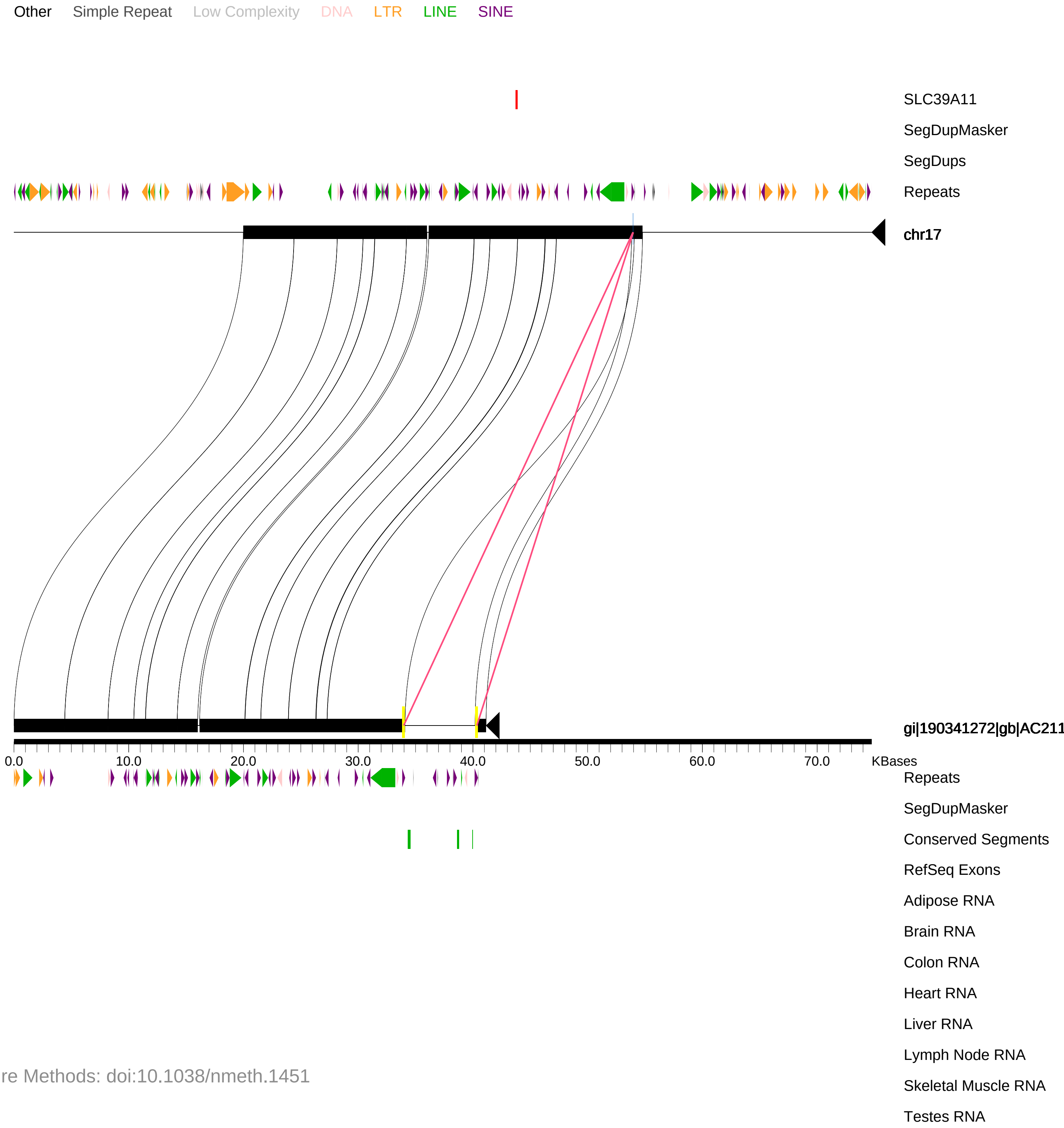
Insertion Size: 10168



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC211712.fa

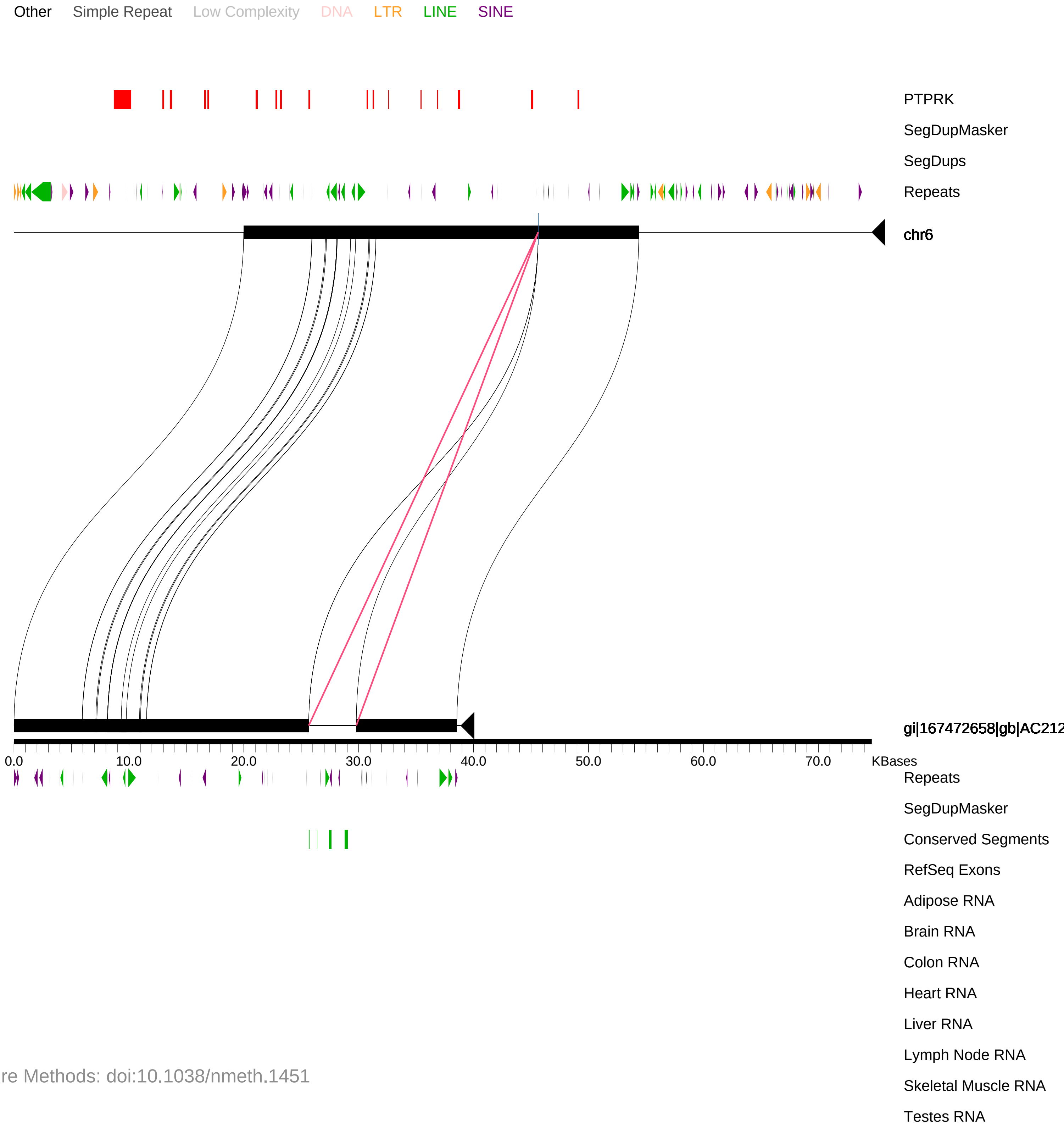
Insertion Size: 6378



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC212491.fa

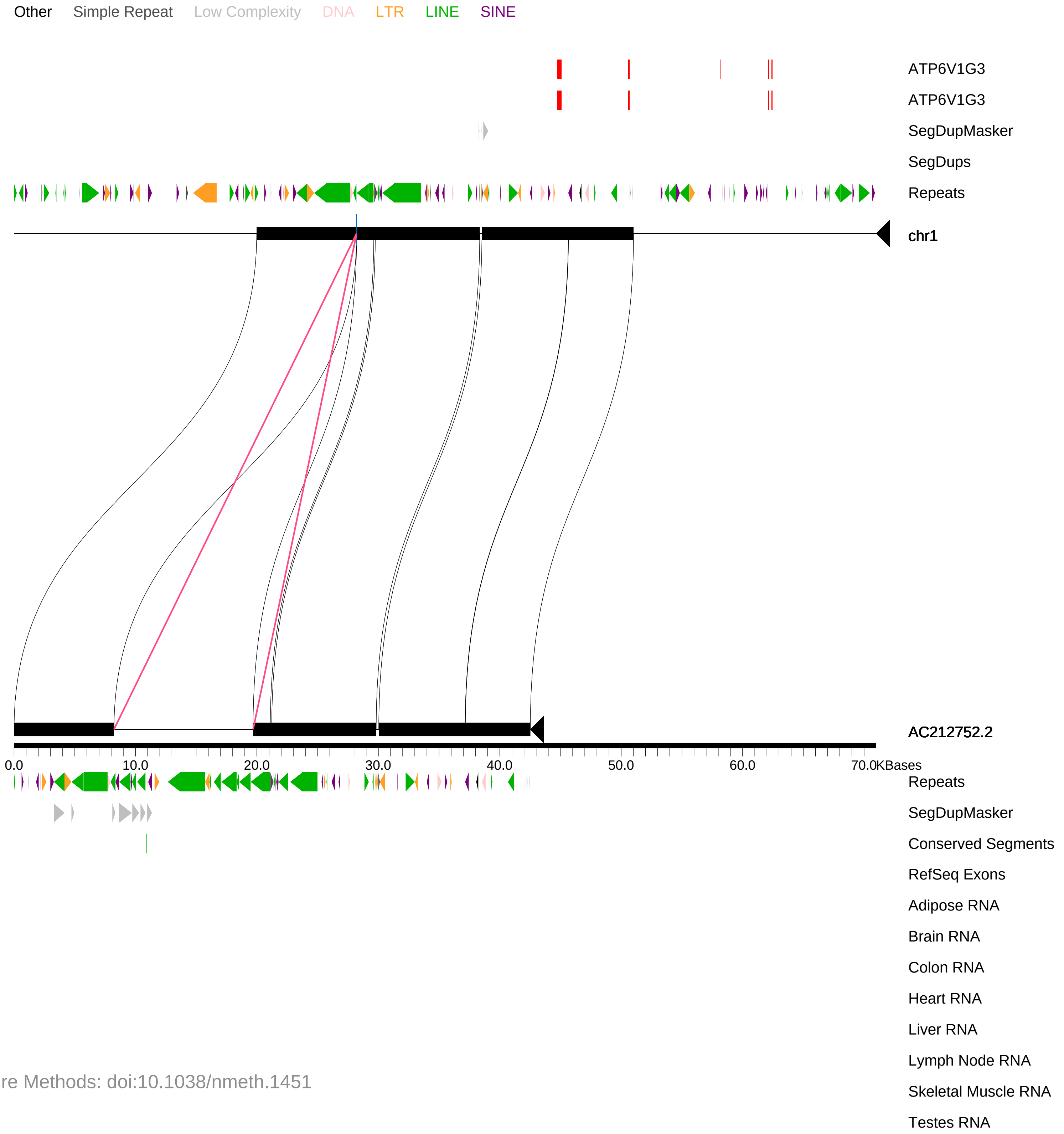
Insertion Size: 4137



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC212752.rc.fa

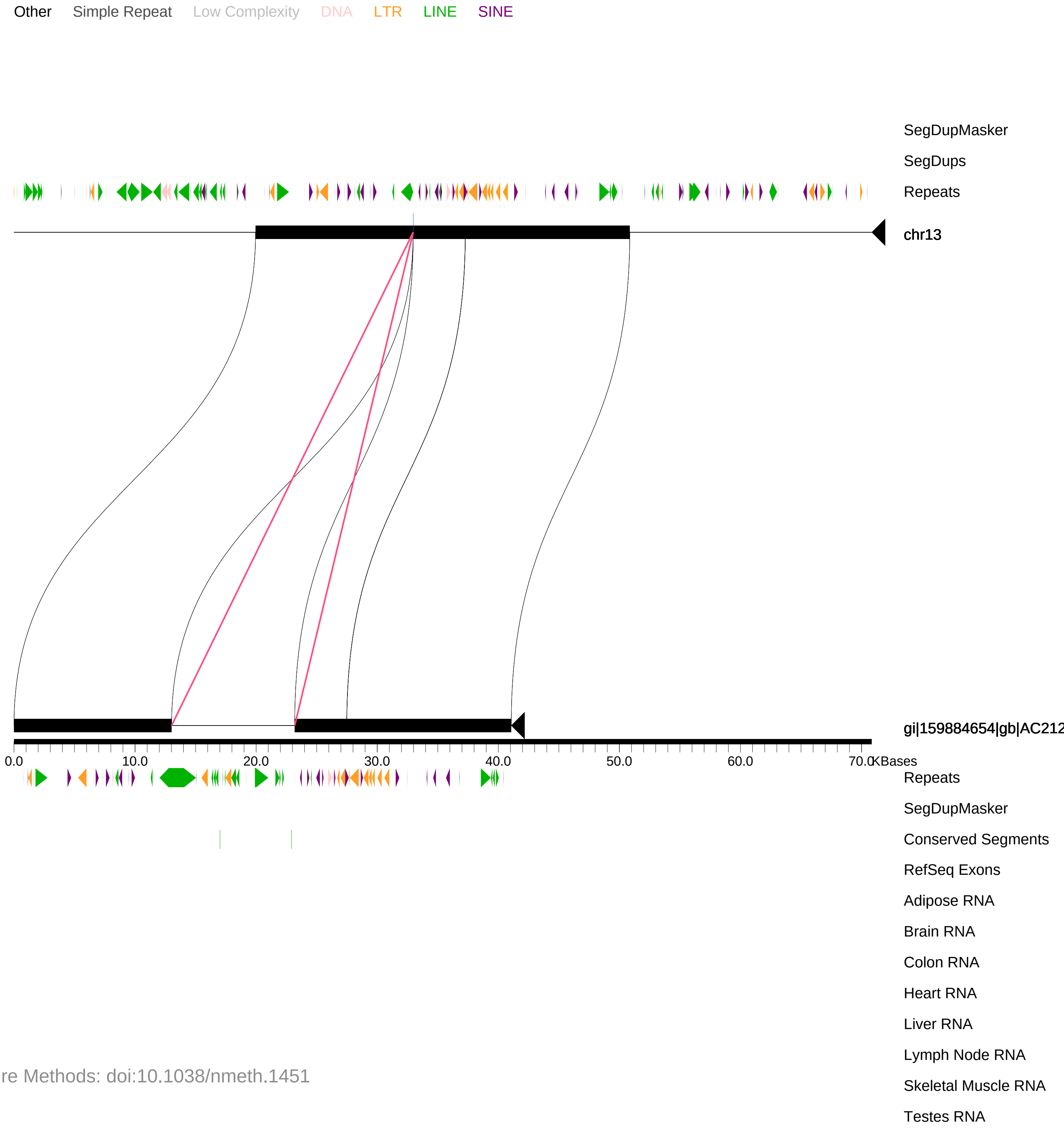
Insertion Size: 11461



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC212759.fa

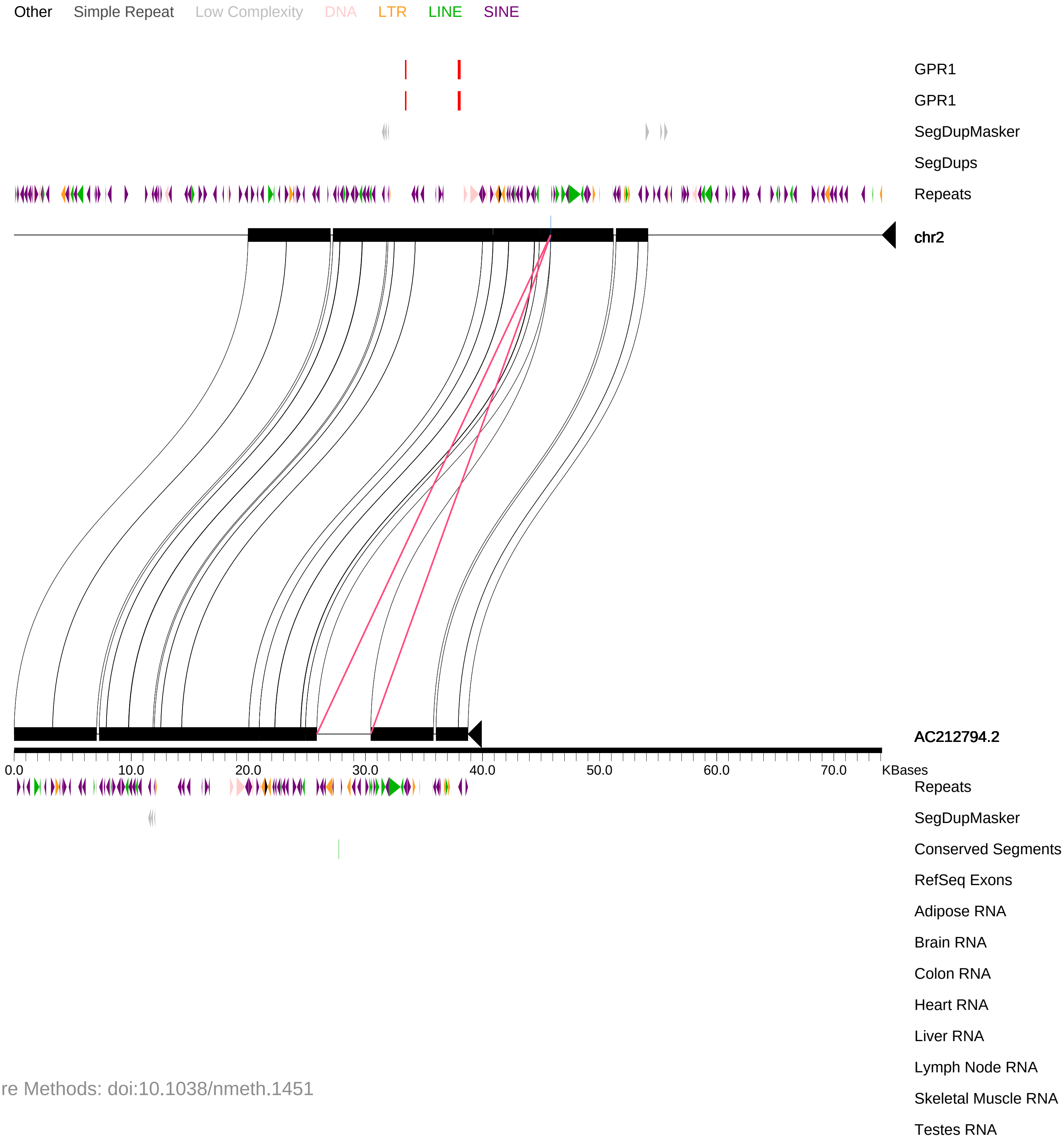
Insertion Size: 10171



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC212794.rc.fa

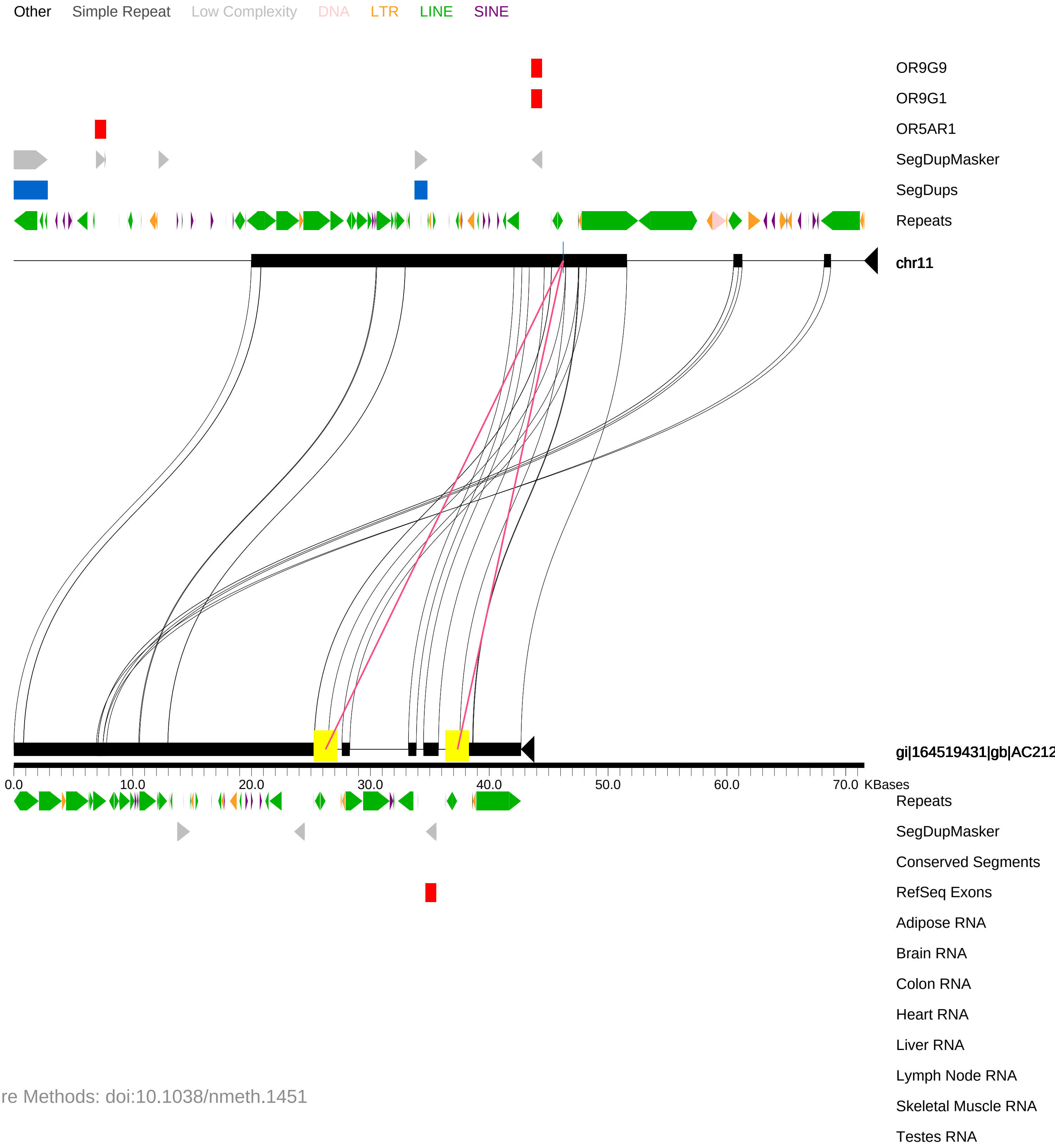
Insertion Size: 4609



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC212901.fa

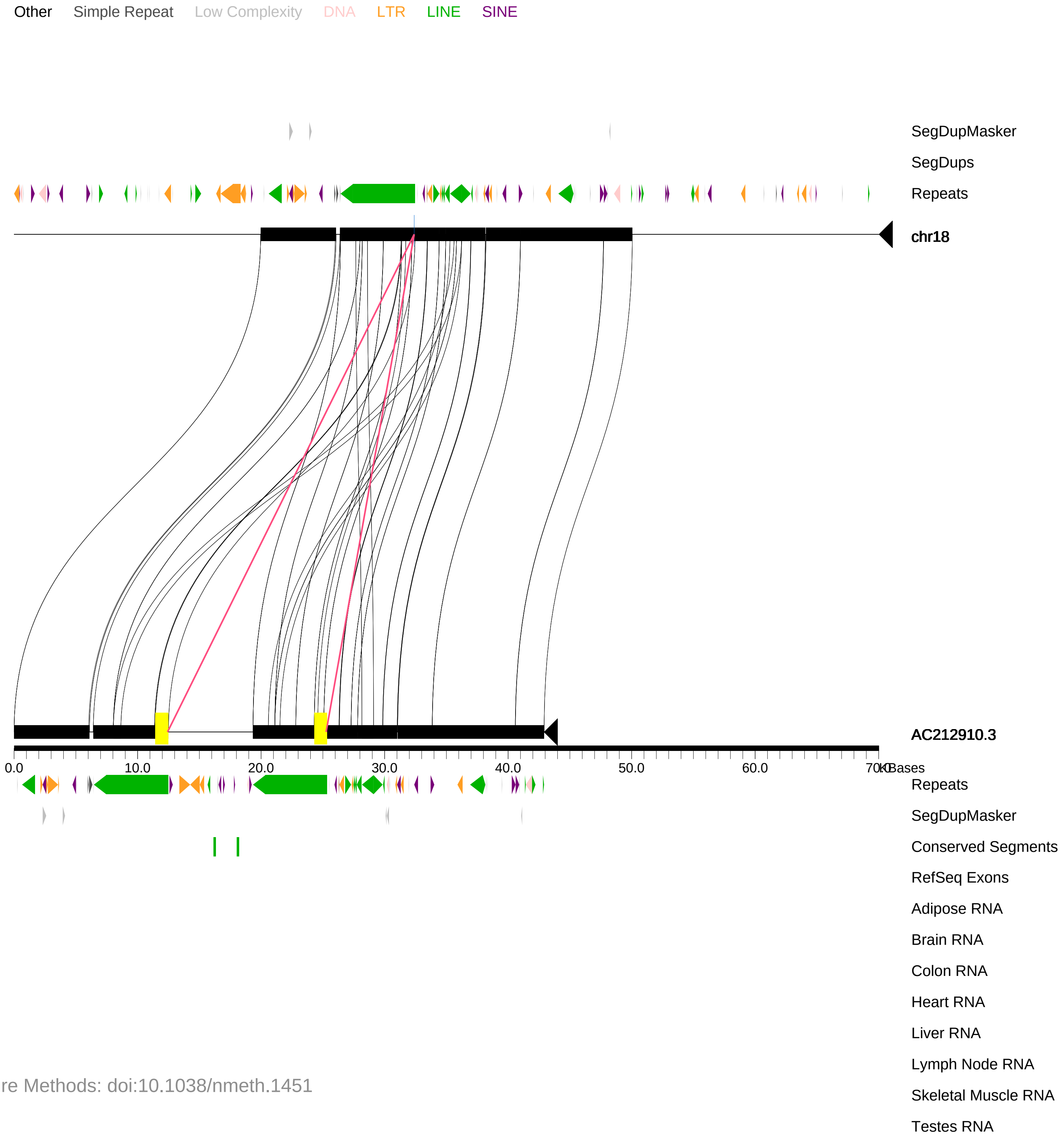
Insertion Size: 11103



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC212910.rc.fa

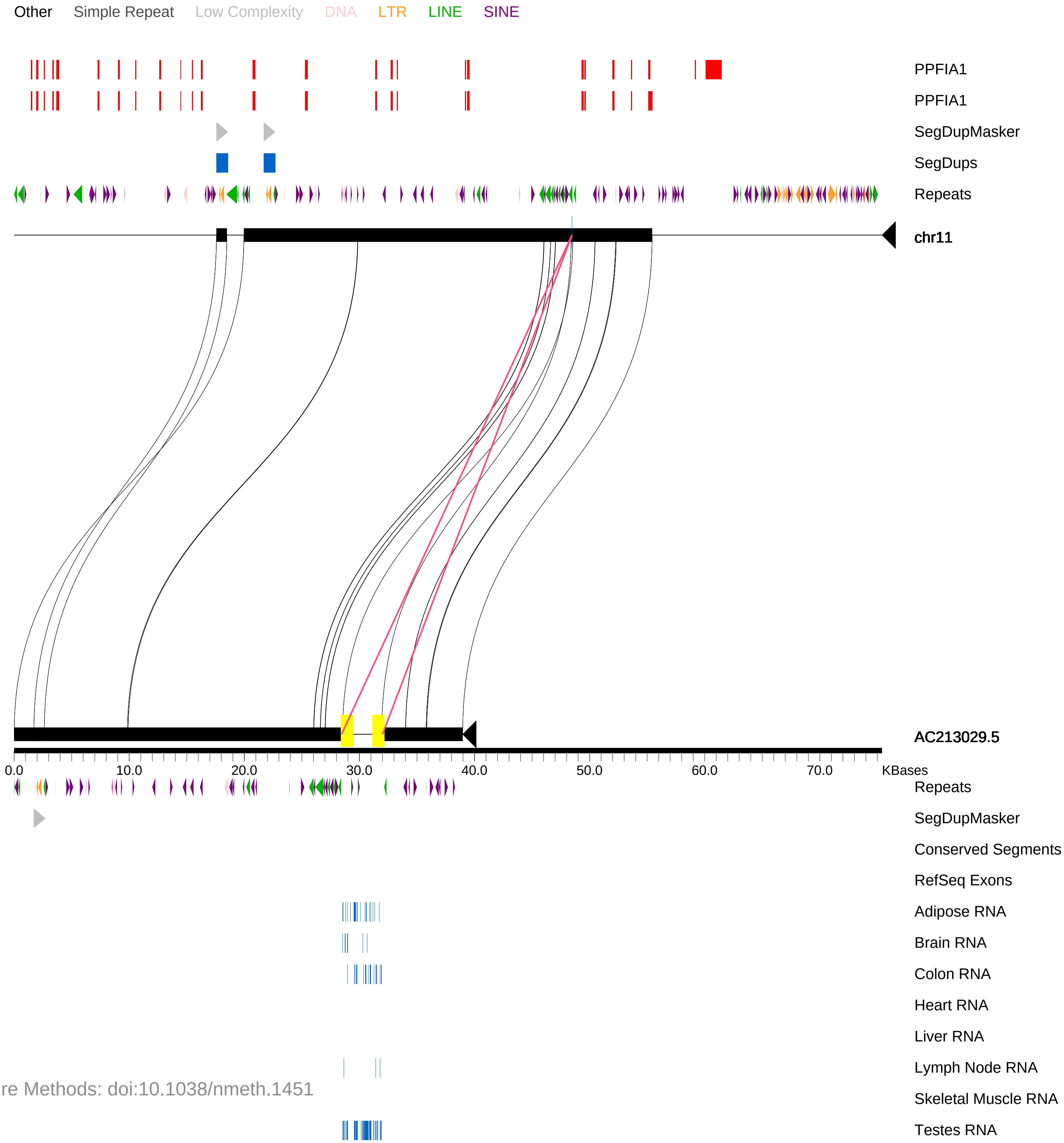
Insertion Size: 12842



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC213029.rc.fa

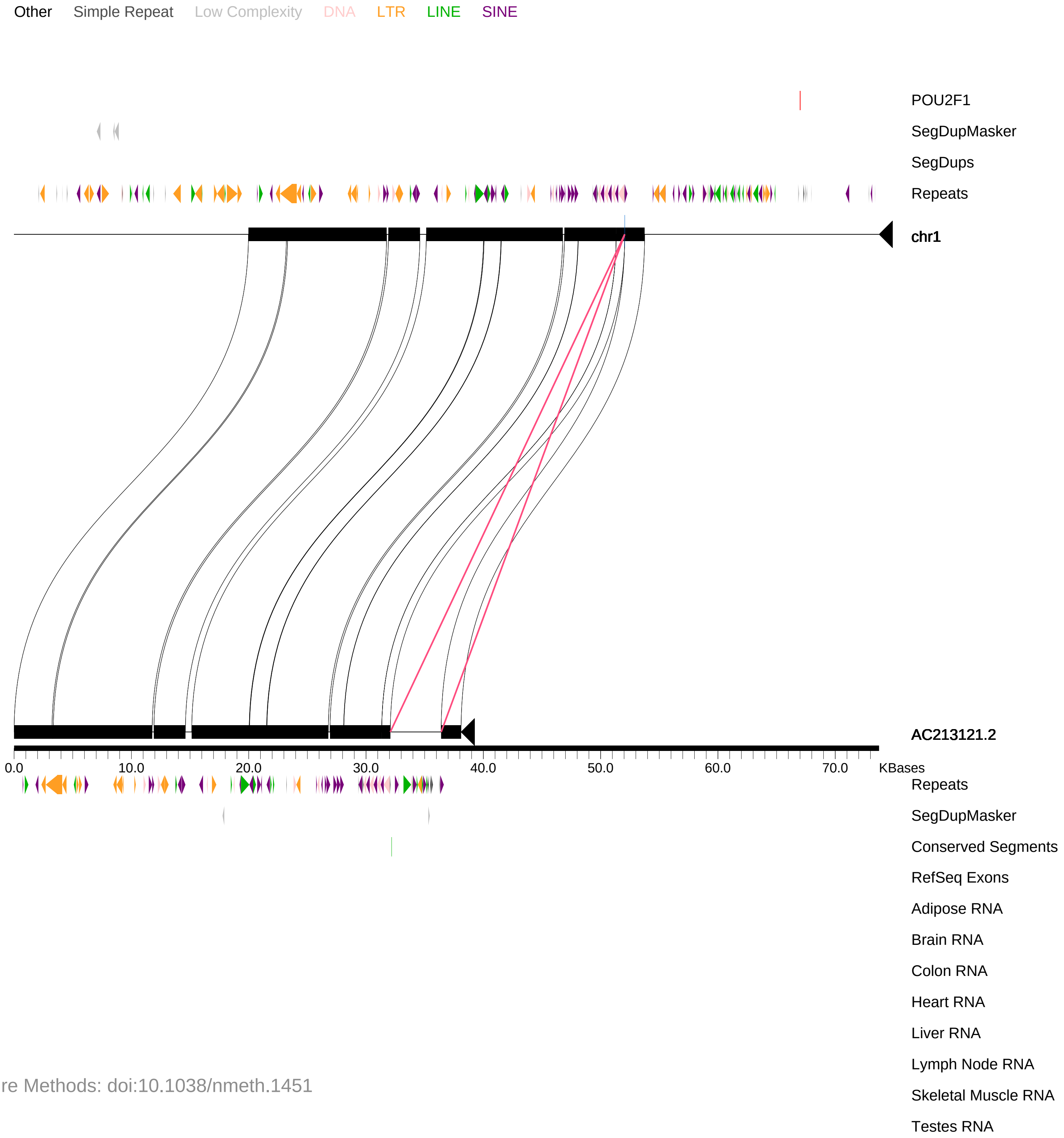
Insertion Size: 3558



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC213121.rc.fa

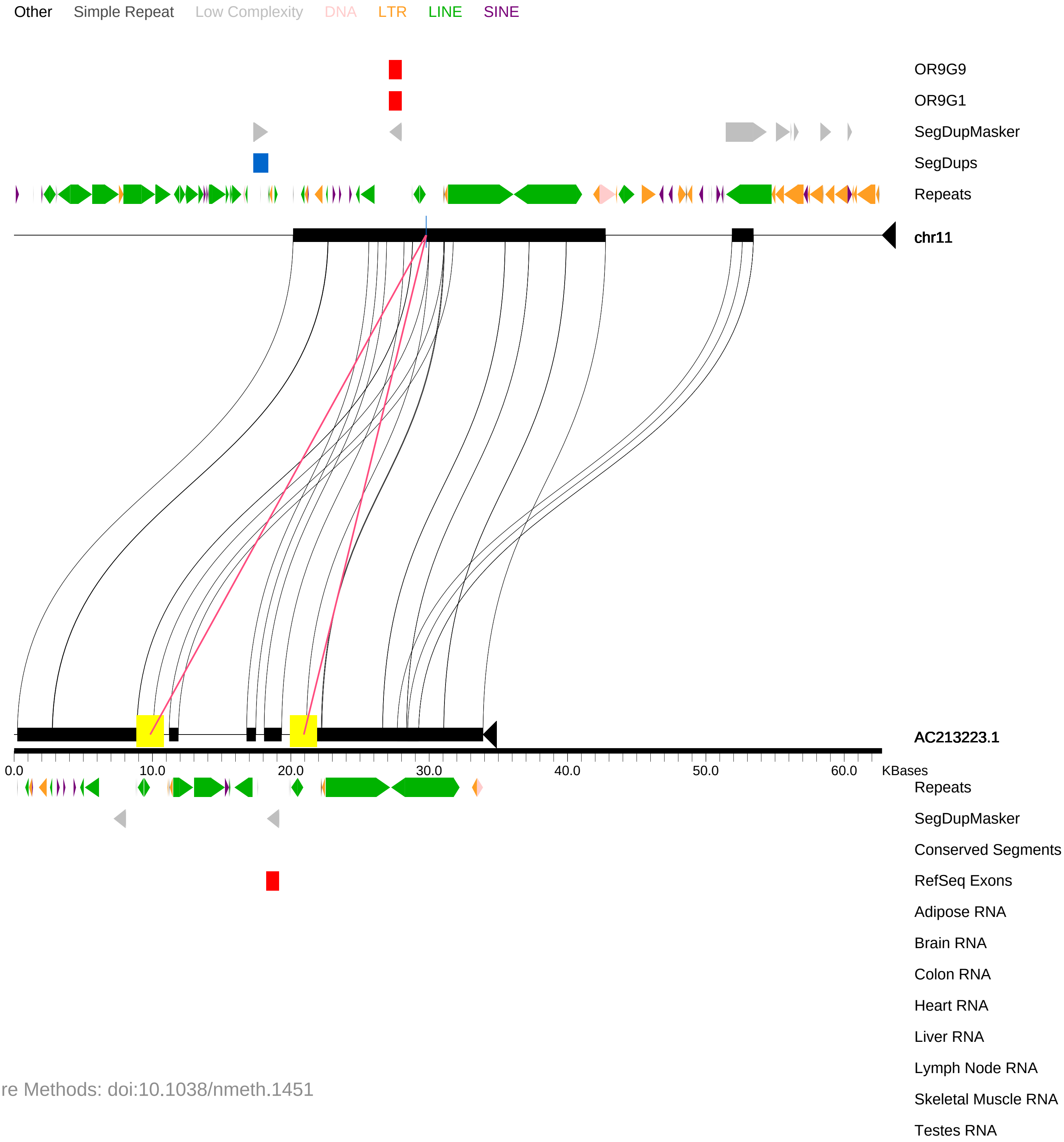
Insertion Size: 4338



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC213223.rc.fa

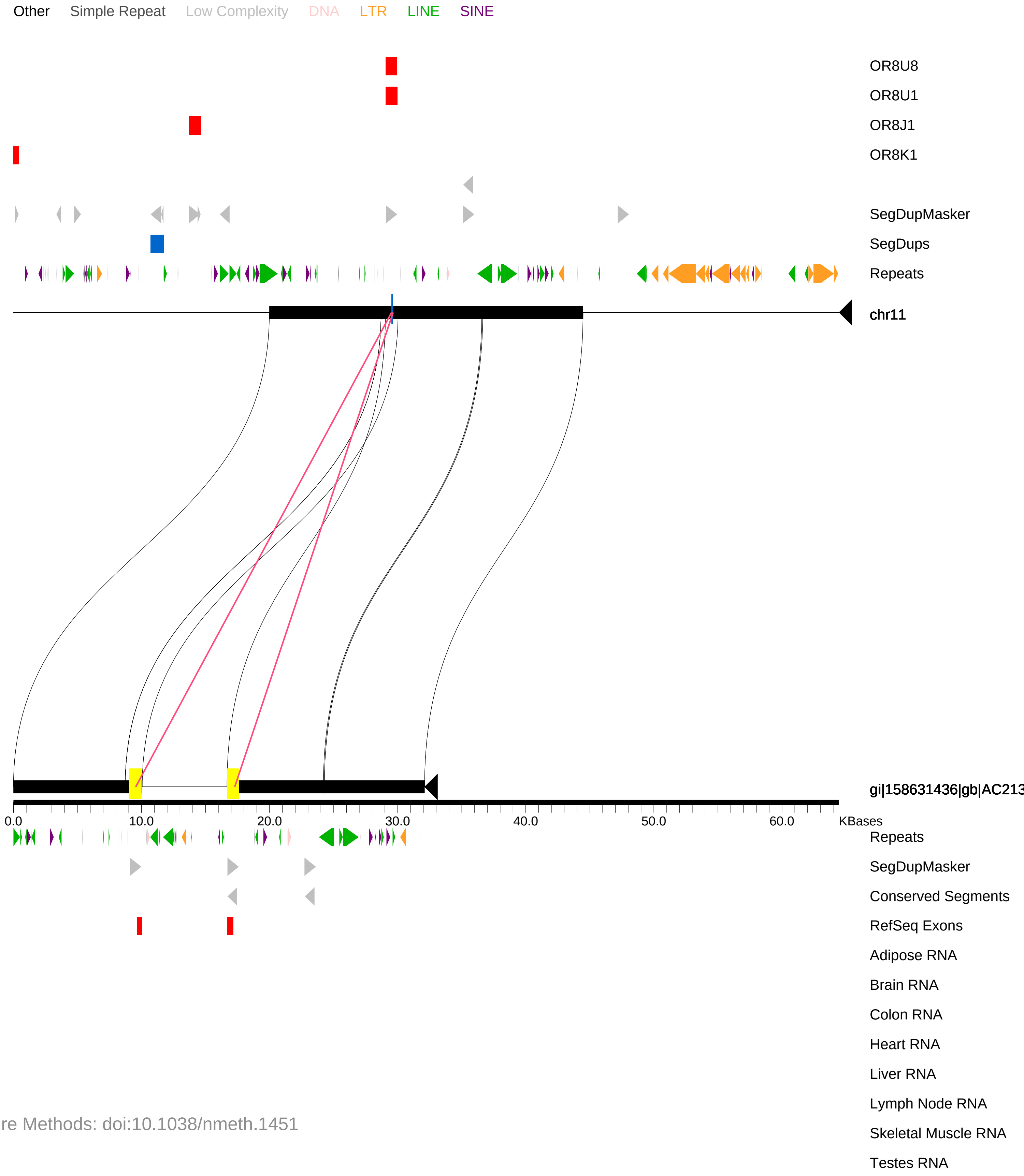
Insertion Size: 11105



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC213240.fa

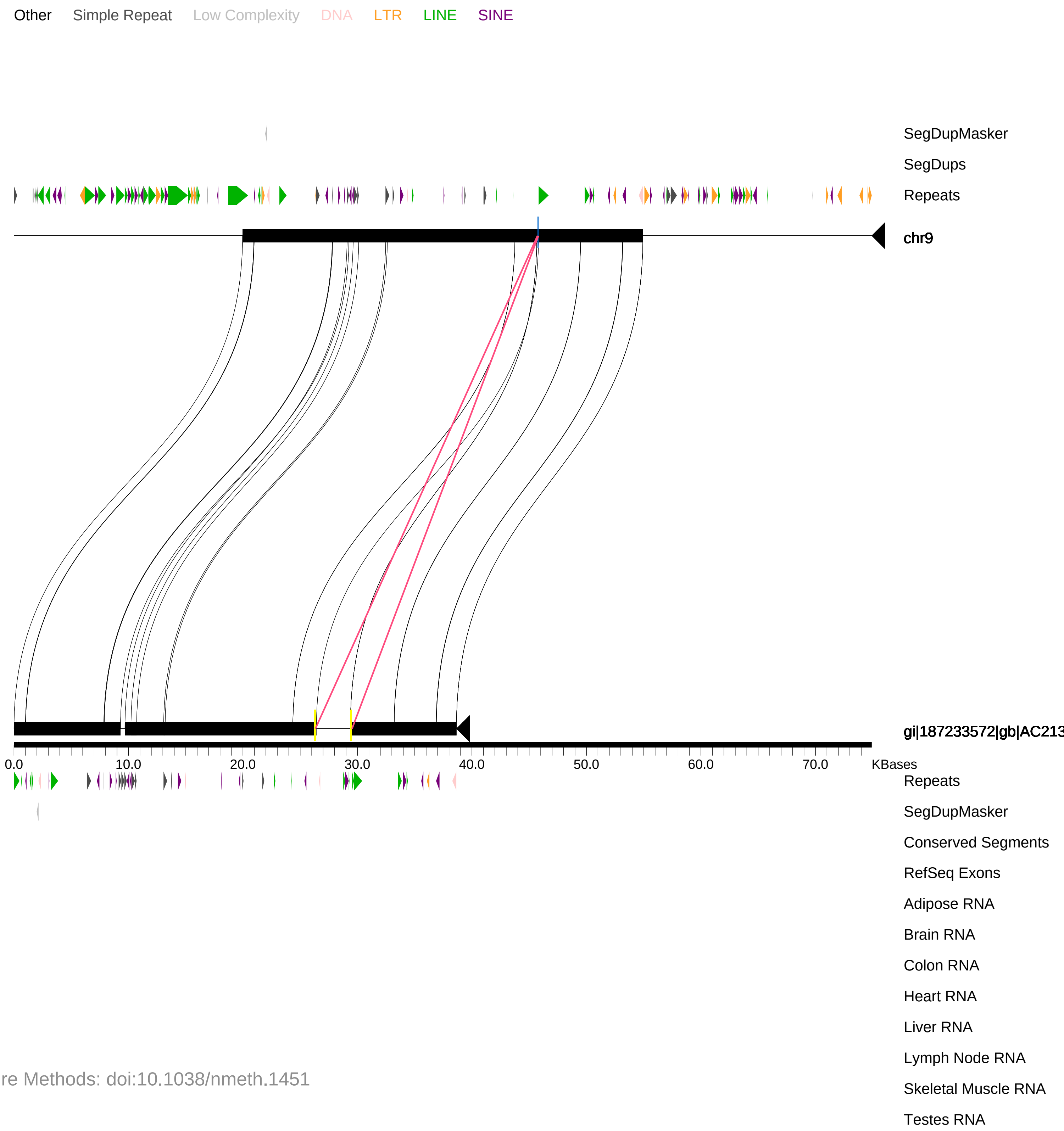
Insertion Size: 7731



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC213440.fa

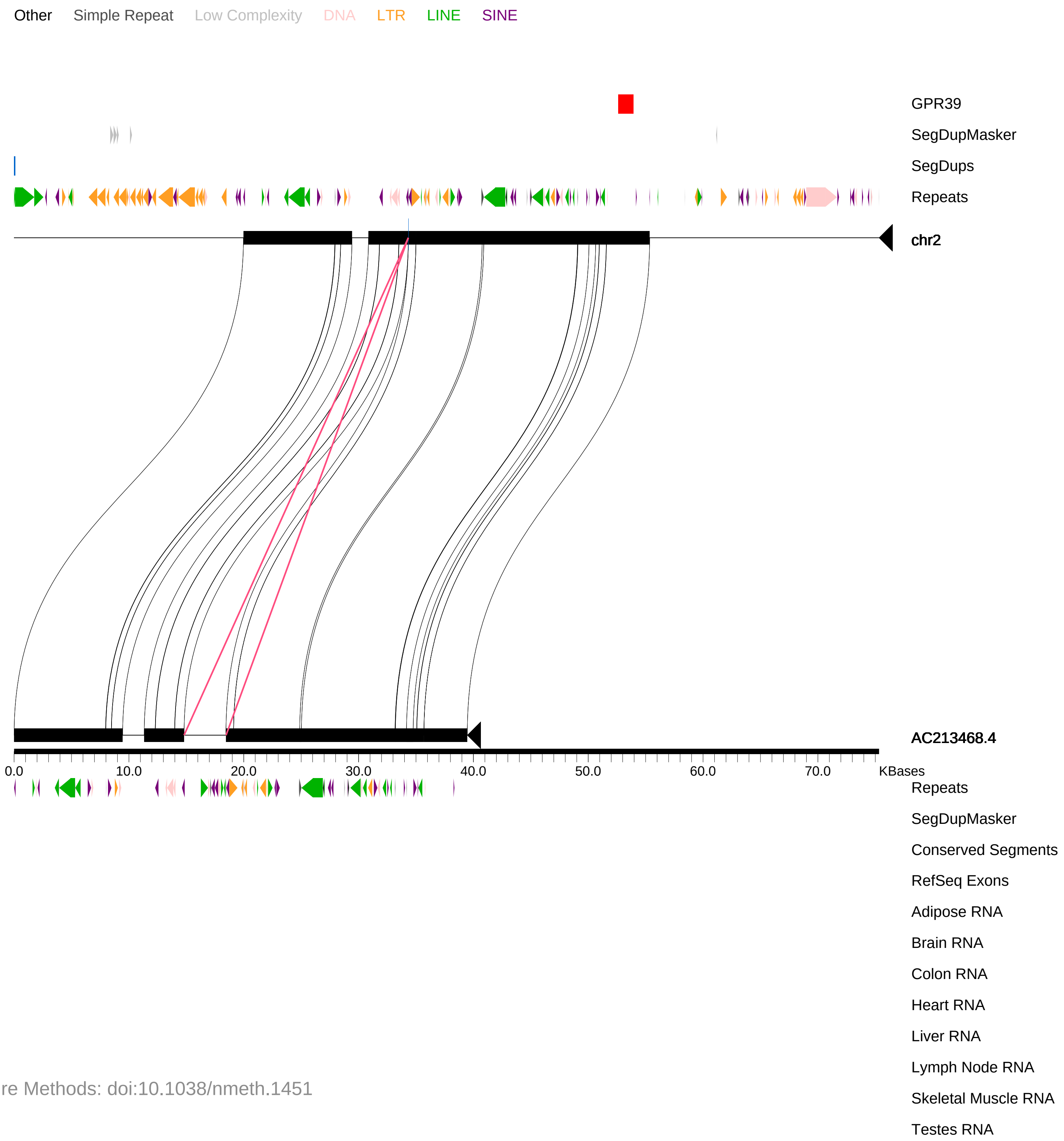
Insertion Size: 3196



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC213468.rc.fa

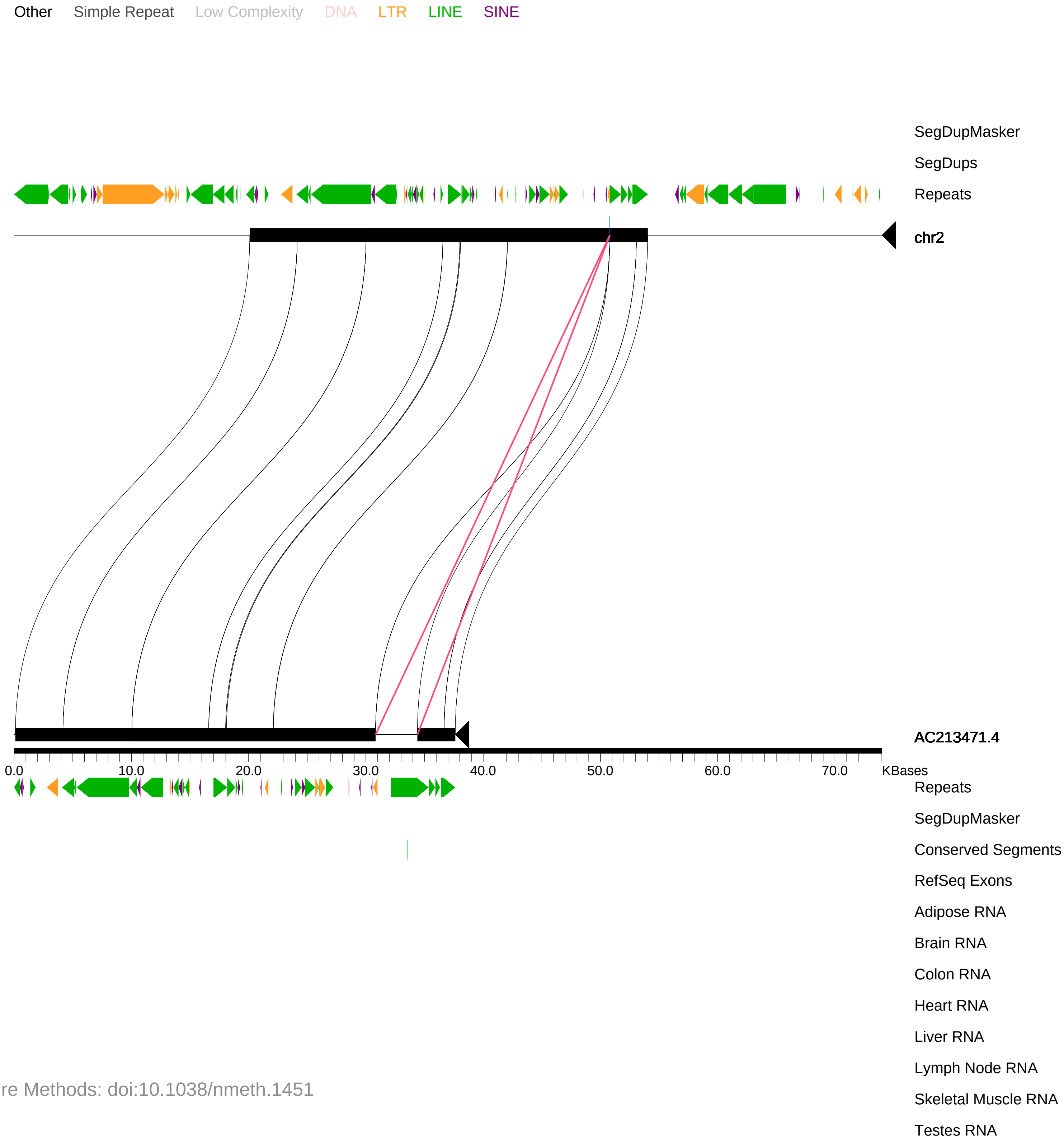
Insertion Size: 3657



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC213471.rc.fa

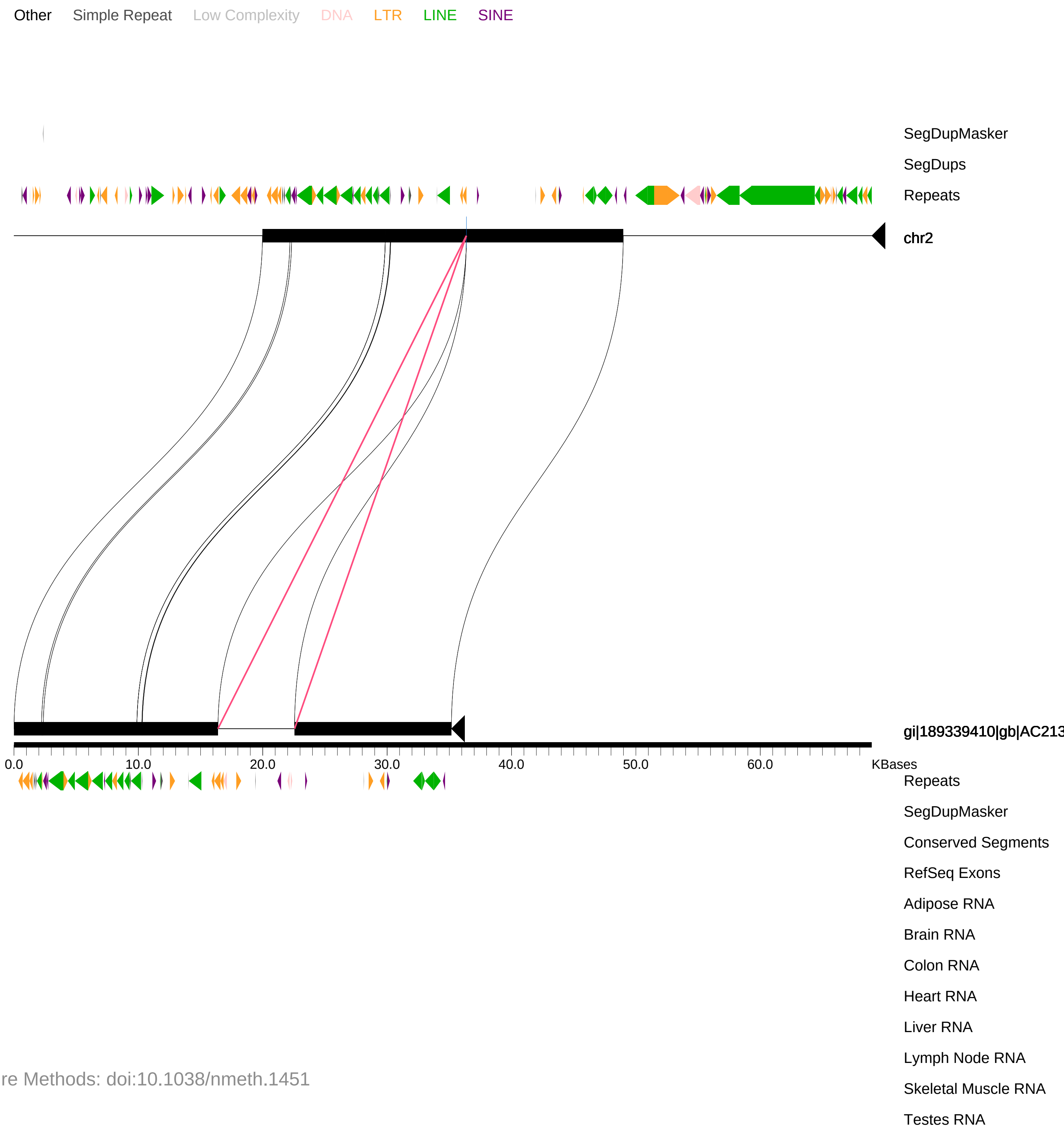
Insertion Size: 3571



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC213472.fa

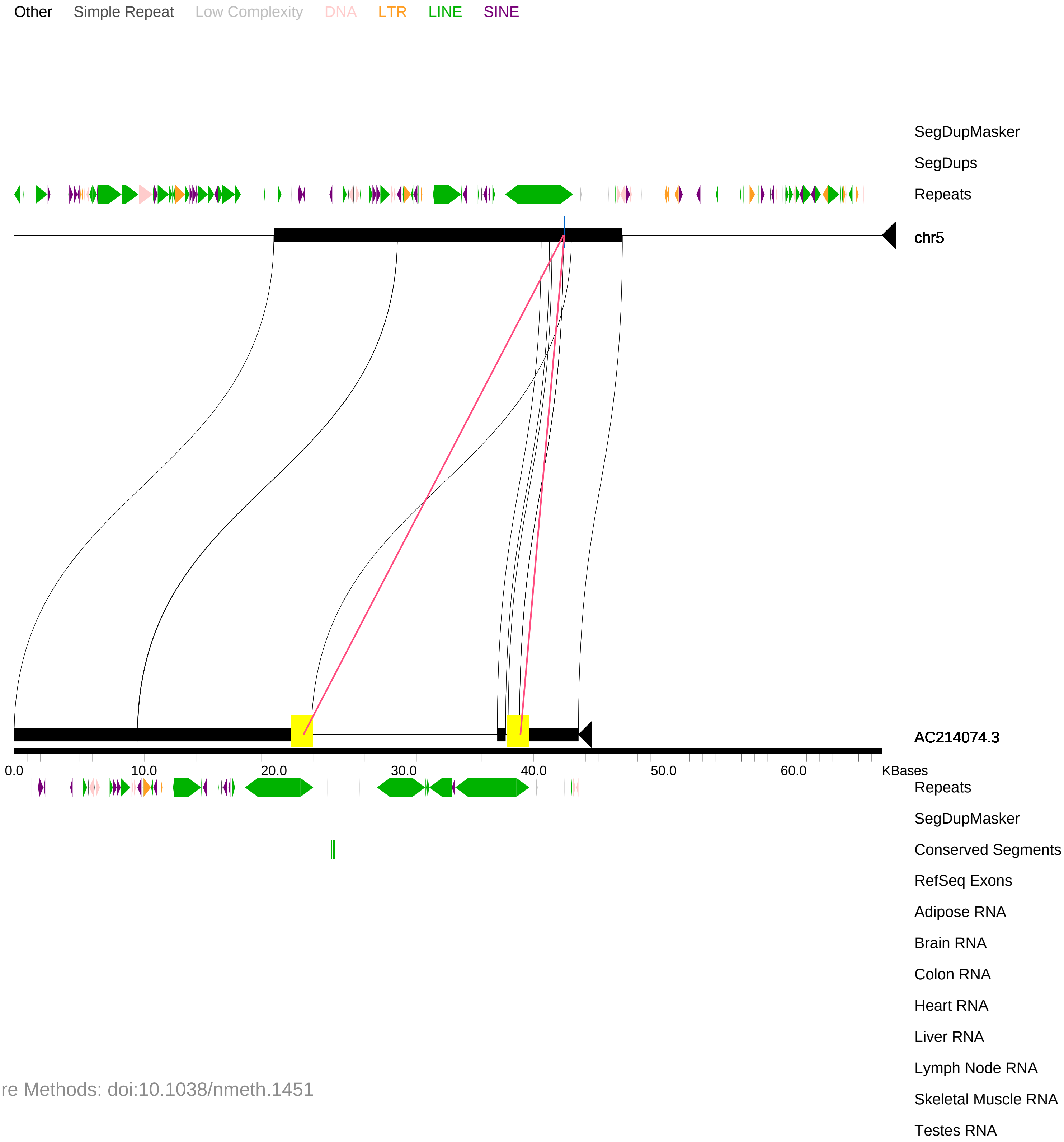
Insertion Size: 6163



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC214074.rc.fa

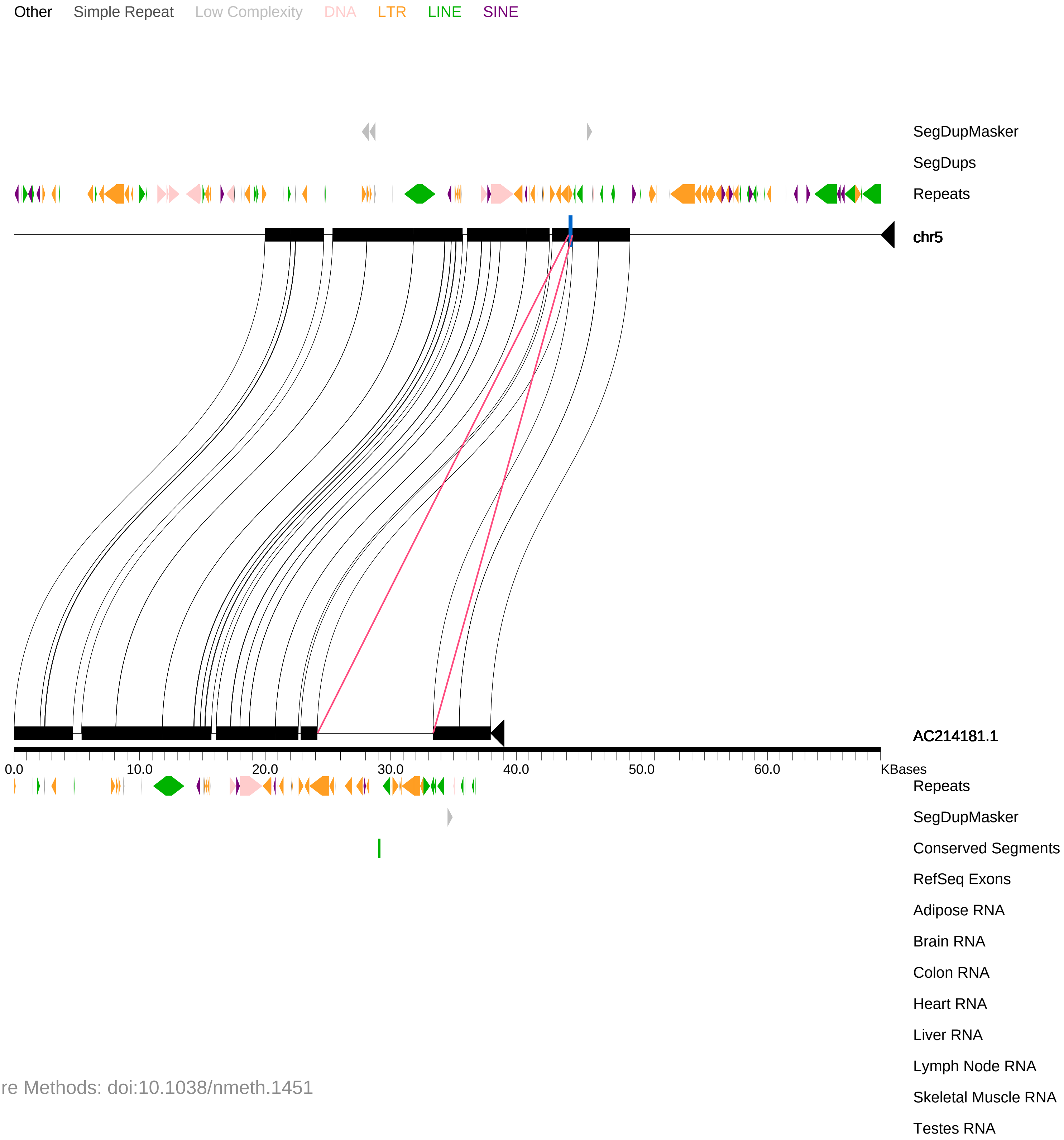
Insertion Size: 16695



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC214181.rc.fa

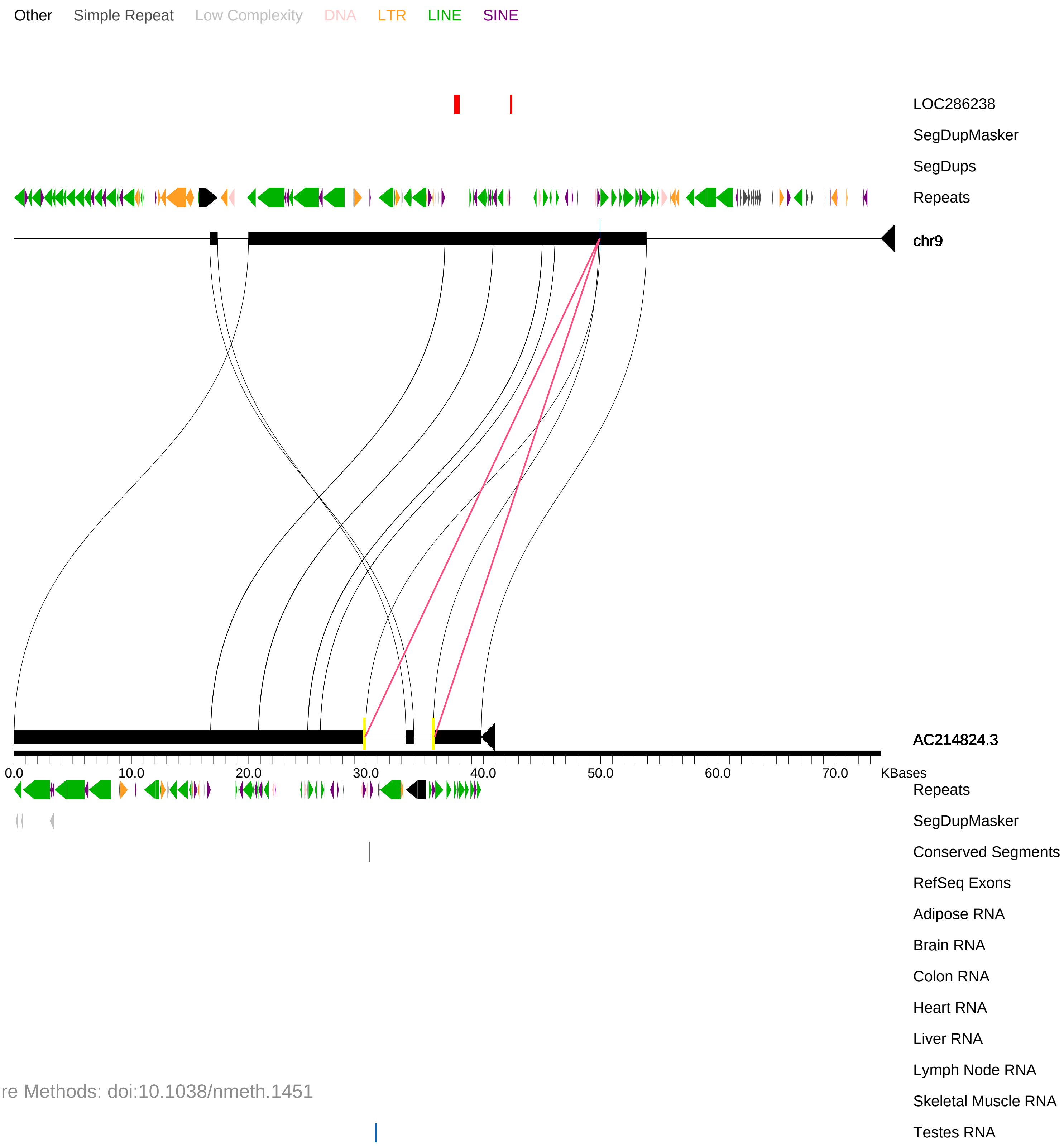
Insertion Size: 9216



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC214824.rc.fa

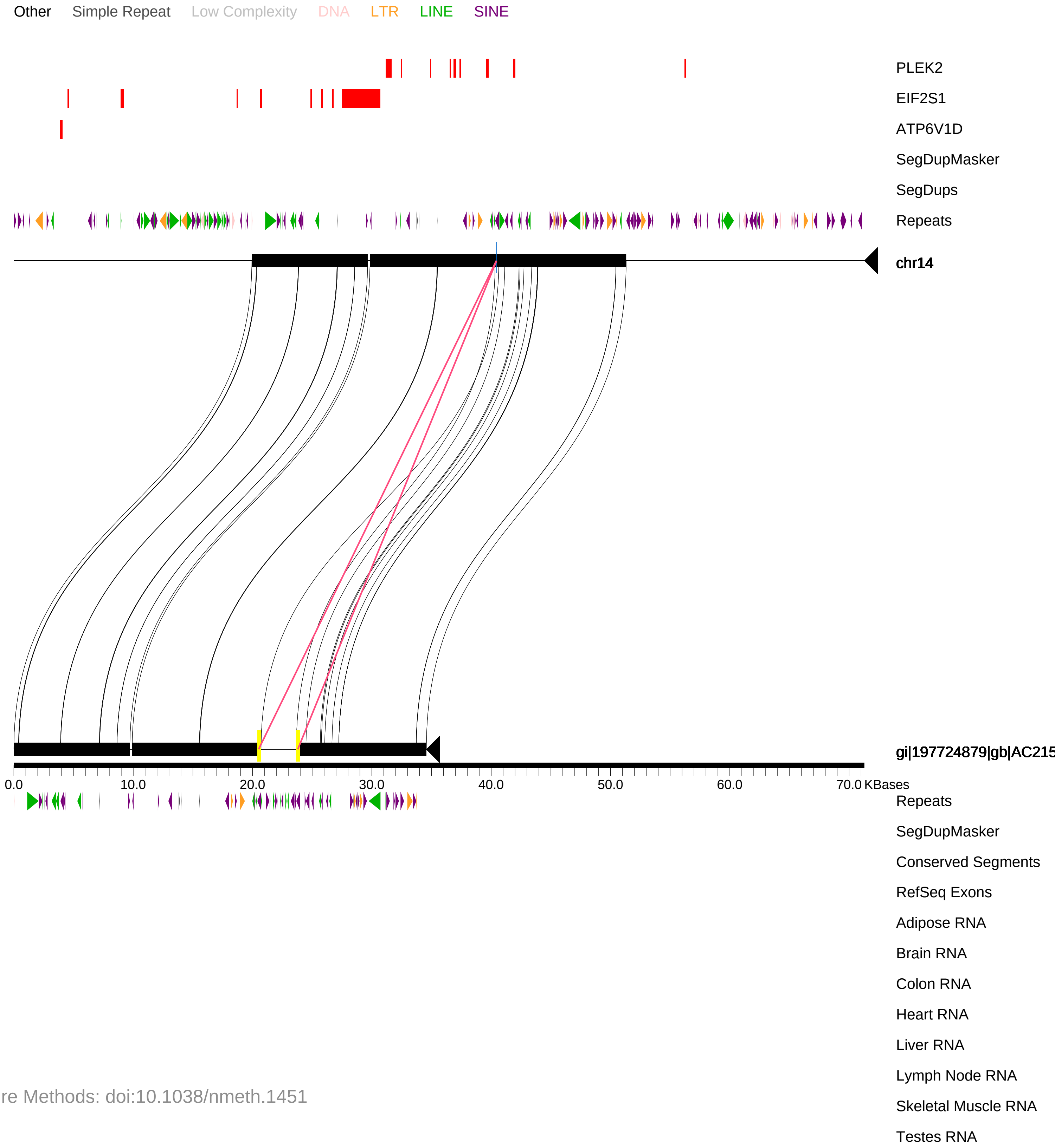
Insertion Size: 5929



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC215288.fa

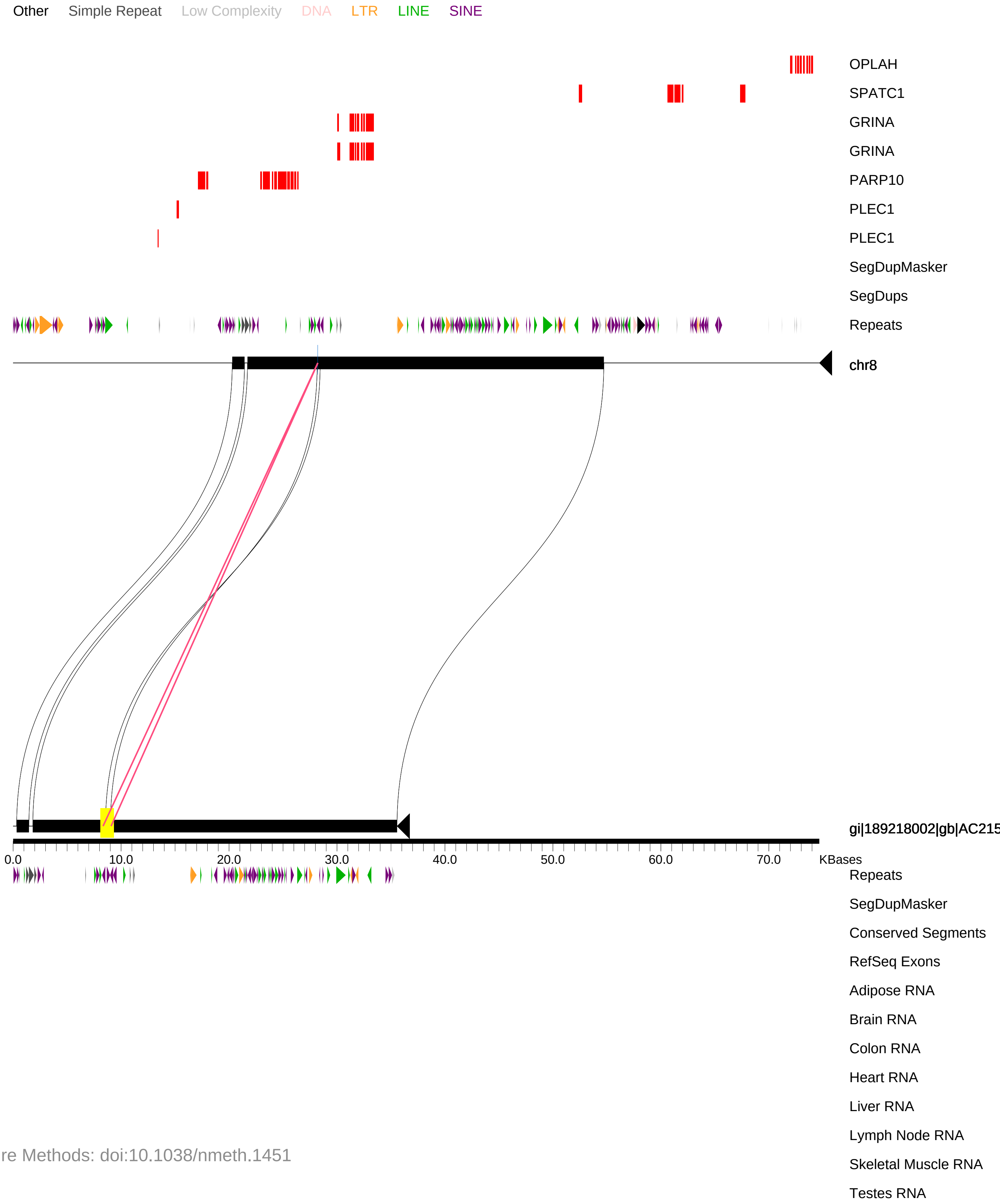
Insertion Size: 3283



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC215339.fa

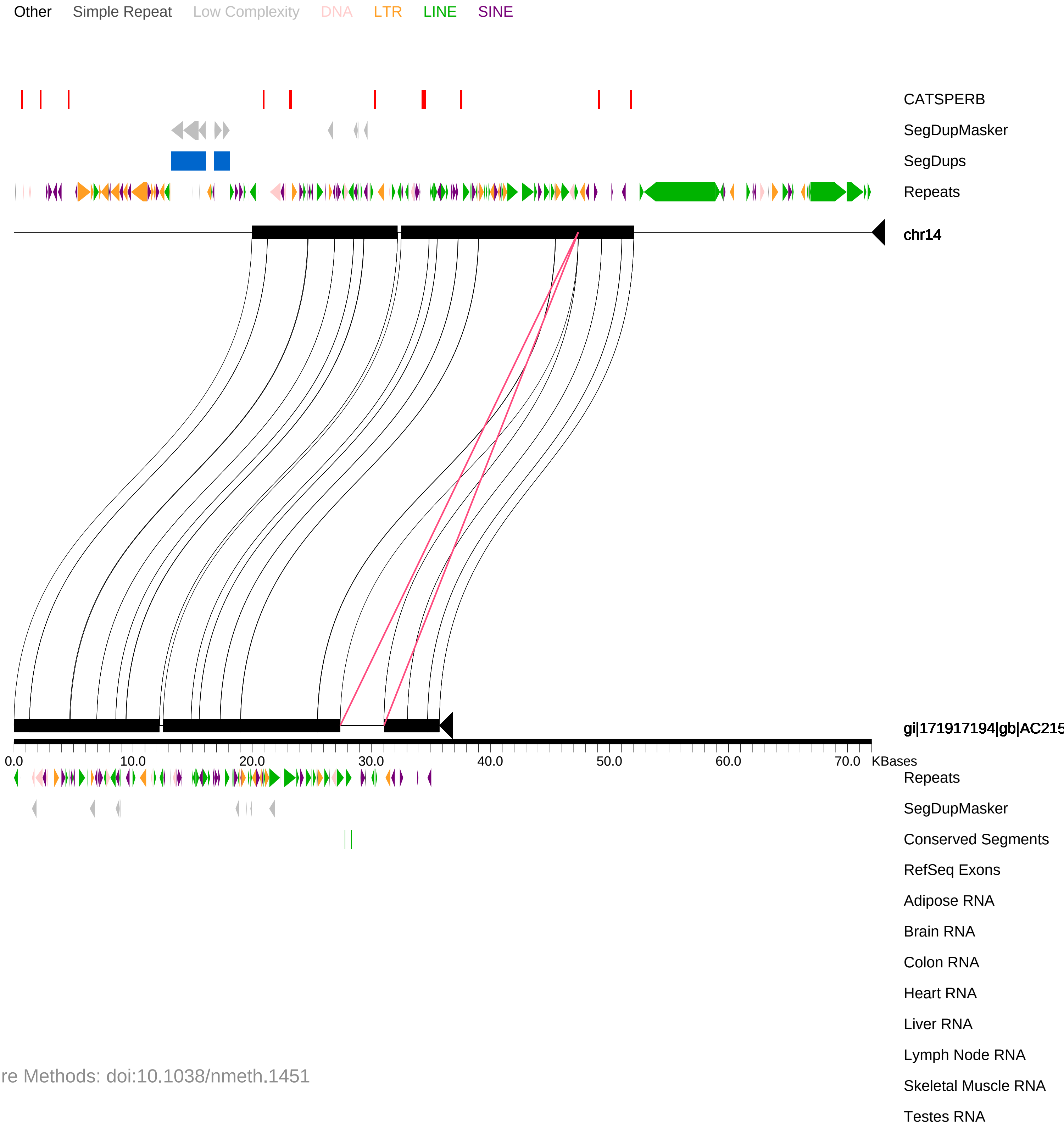
Insertion Size: 743



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC215700.fa

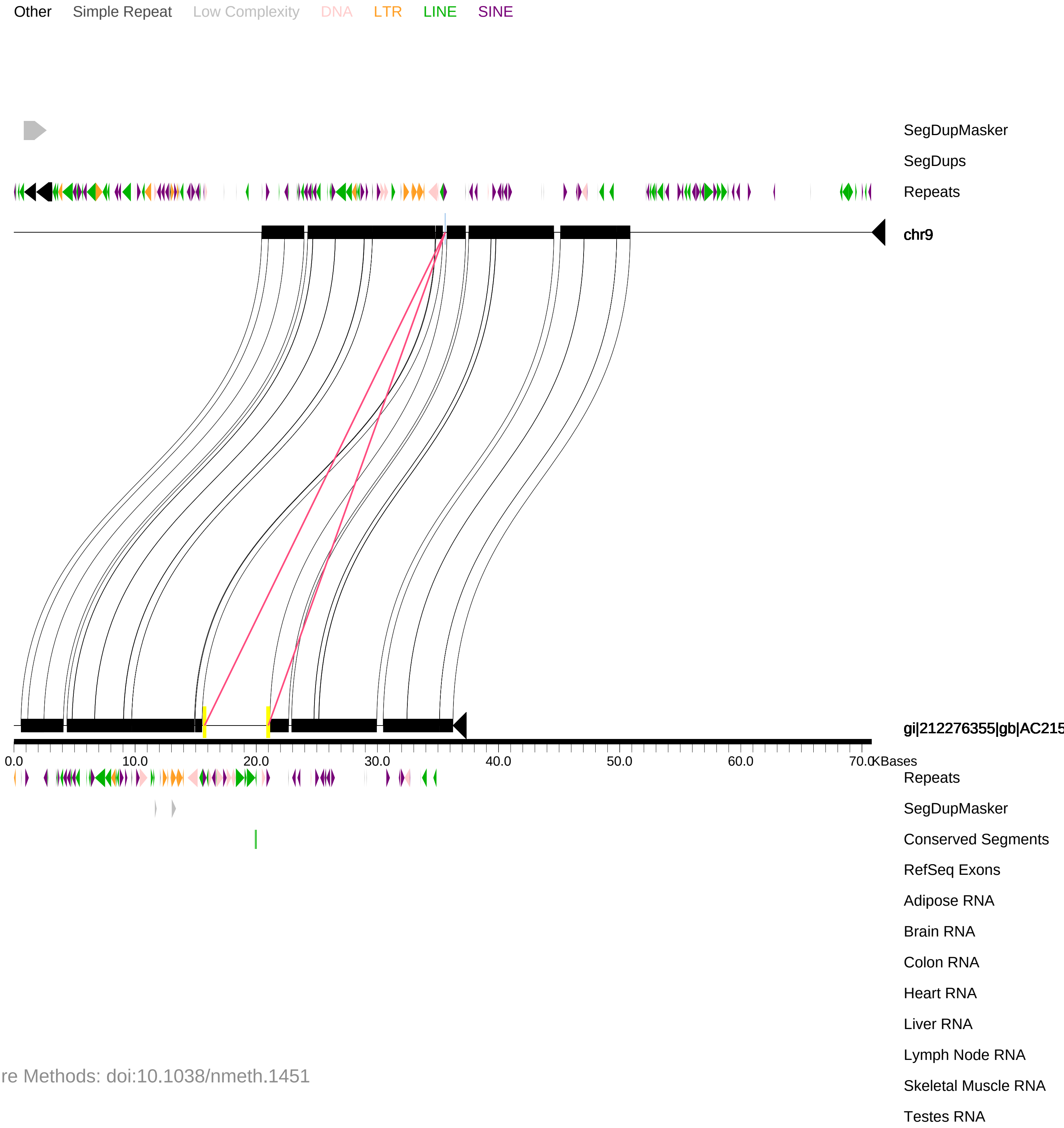
Insertion Size: 3673



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC215710.fa

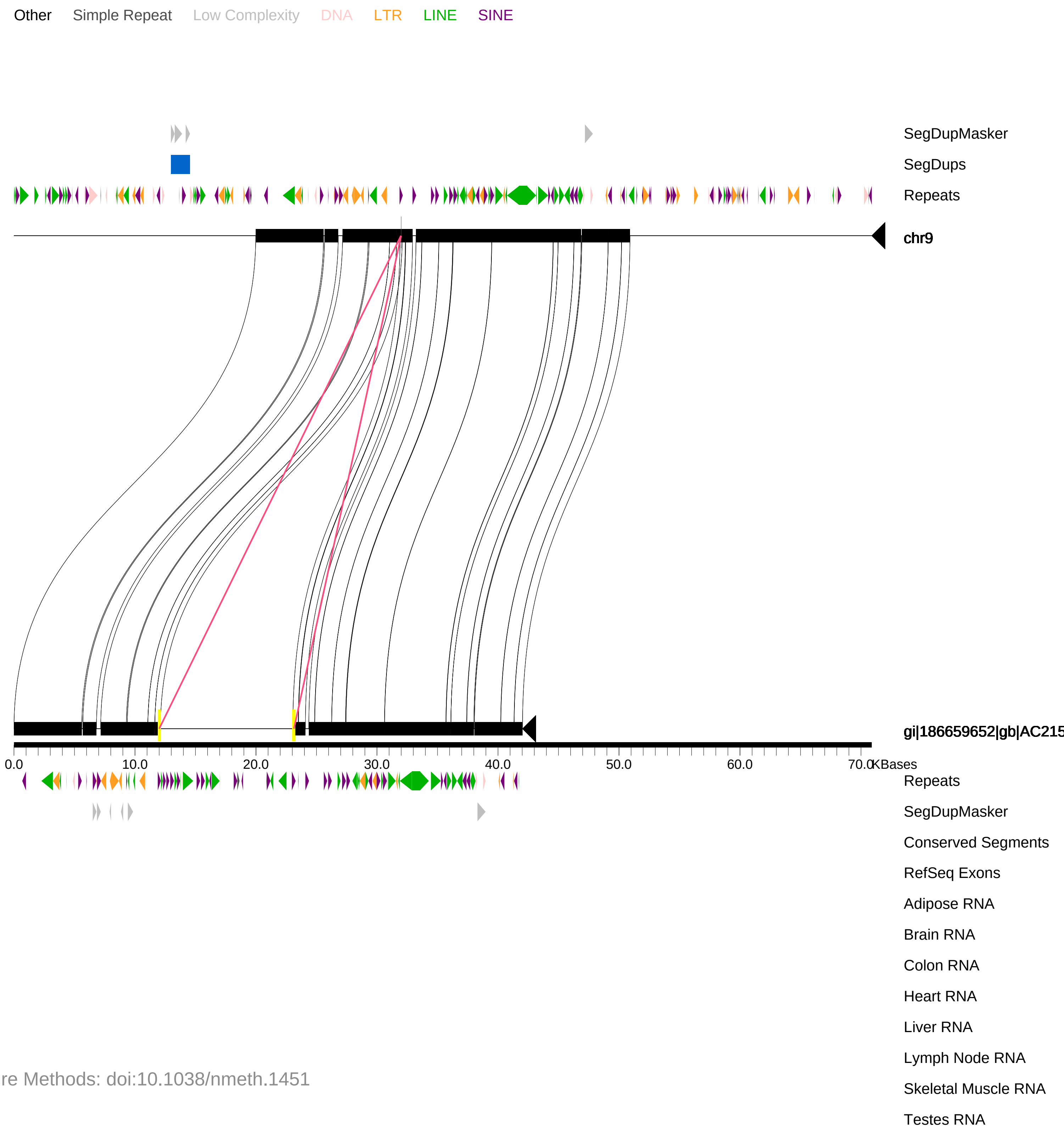
Insertion Size: 5260



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC215799.fa

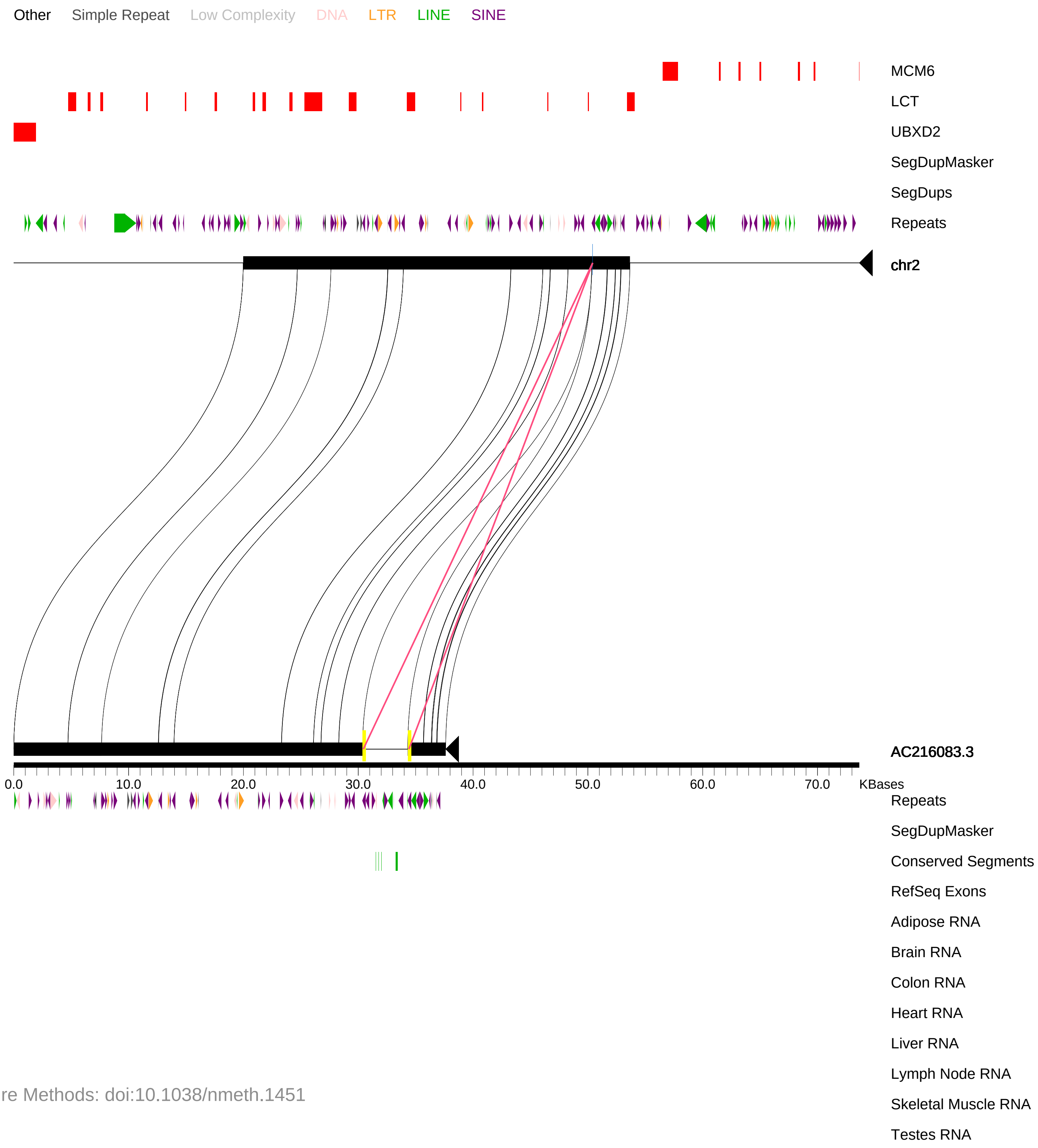
Insertion Size: 11140



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC216083.rc.fa

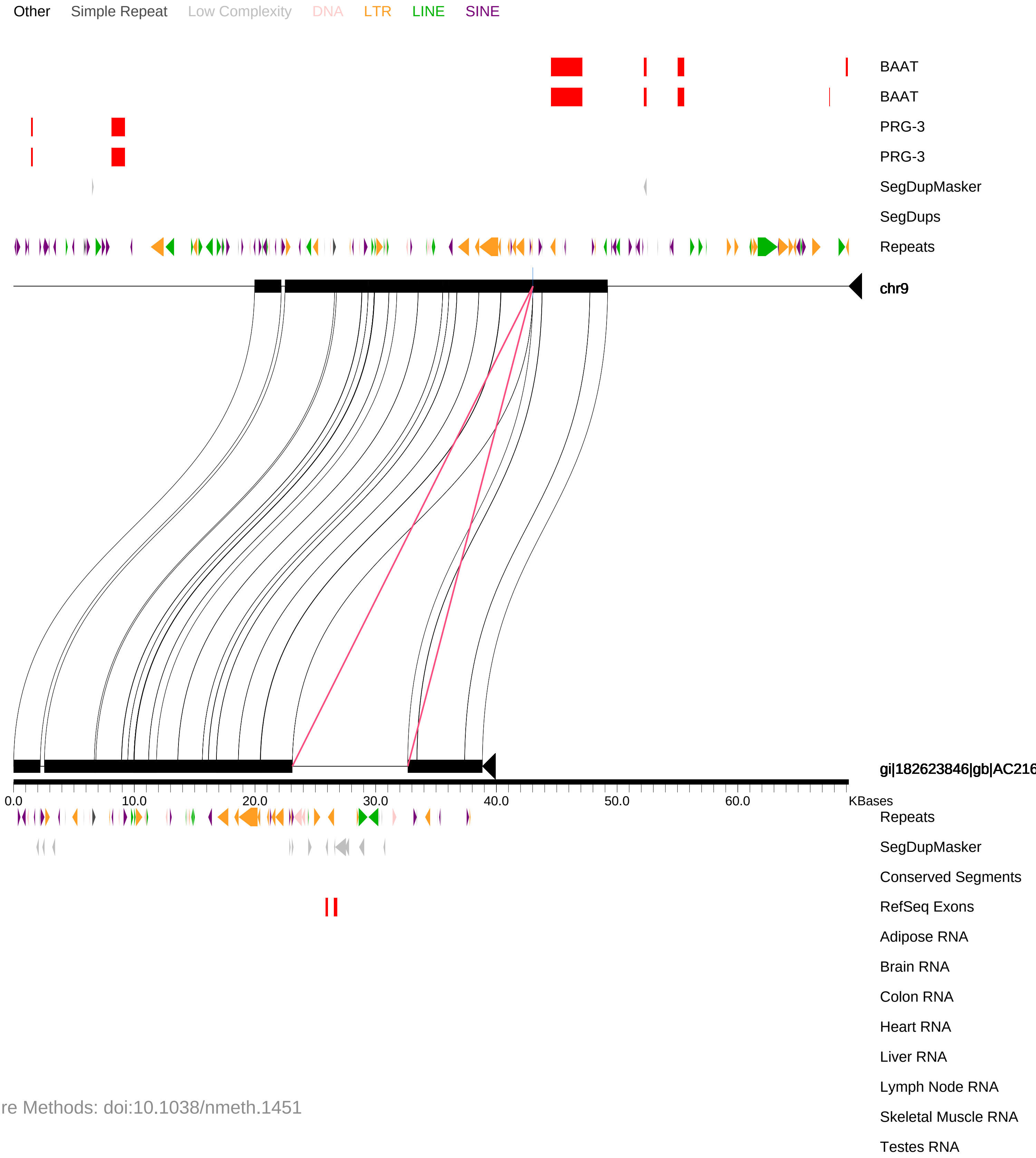
Insertion Size: 3970



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC216089.fa

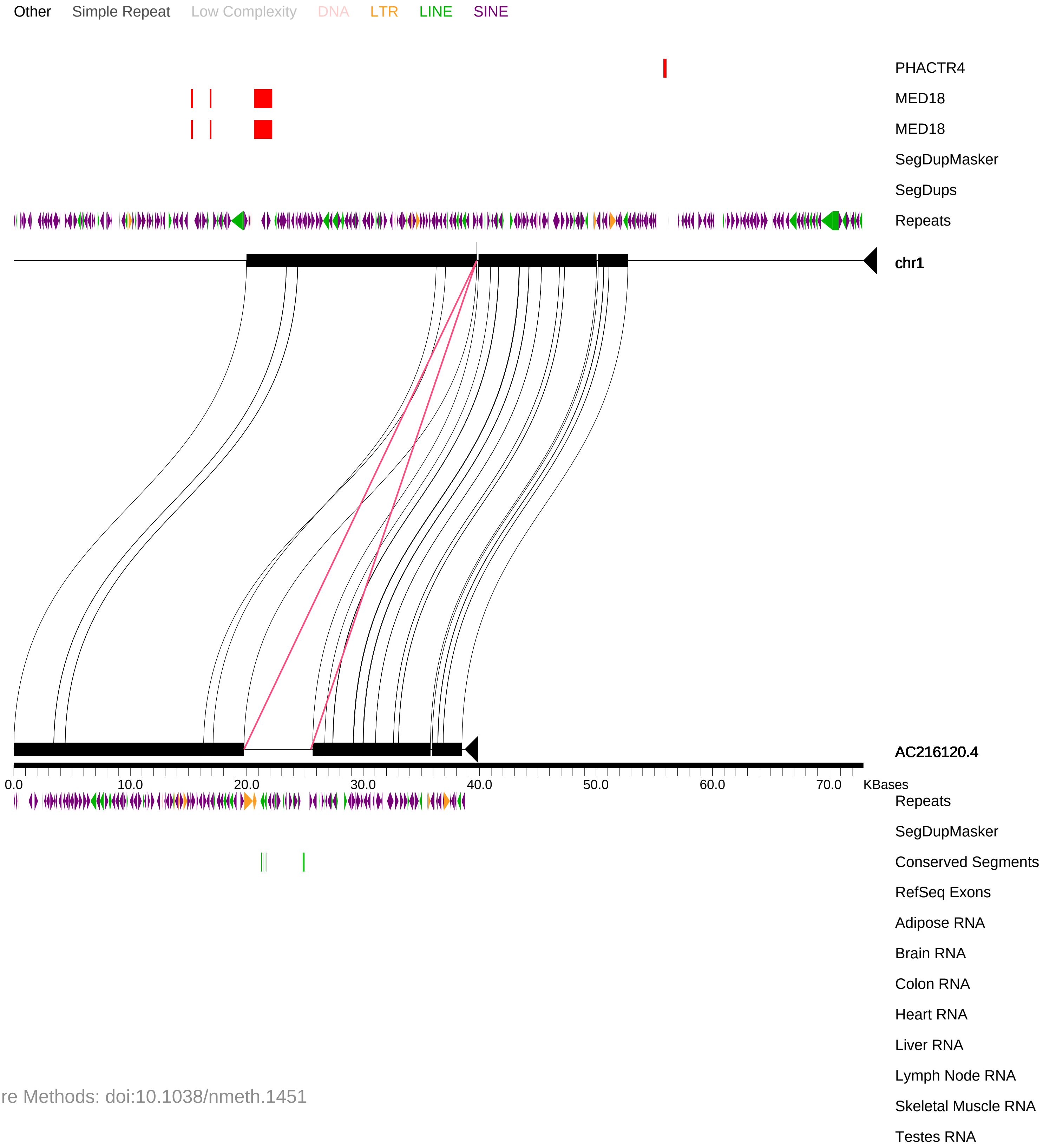
Insertion Size: 9568



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC216120.rc.fa

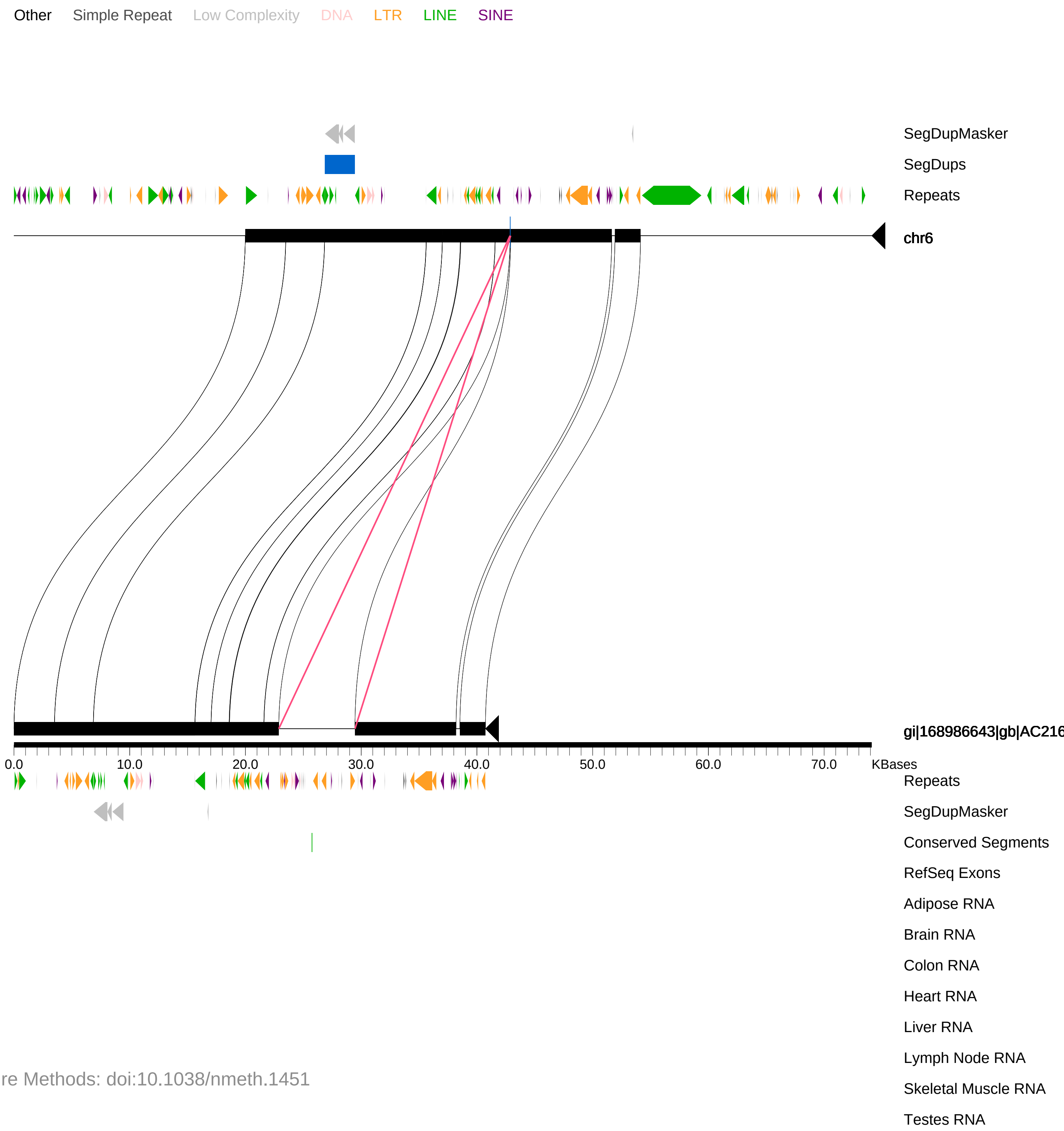
Insertion Size: 5736



Nature Methods: doi:10.1038/nmeth.1451

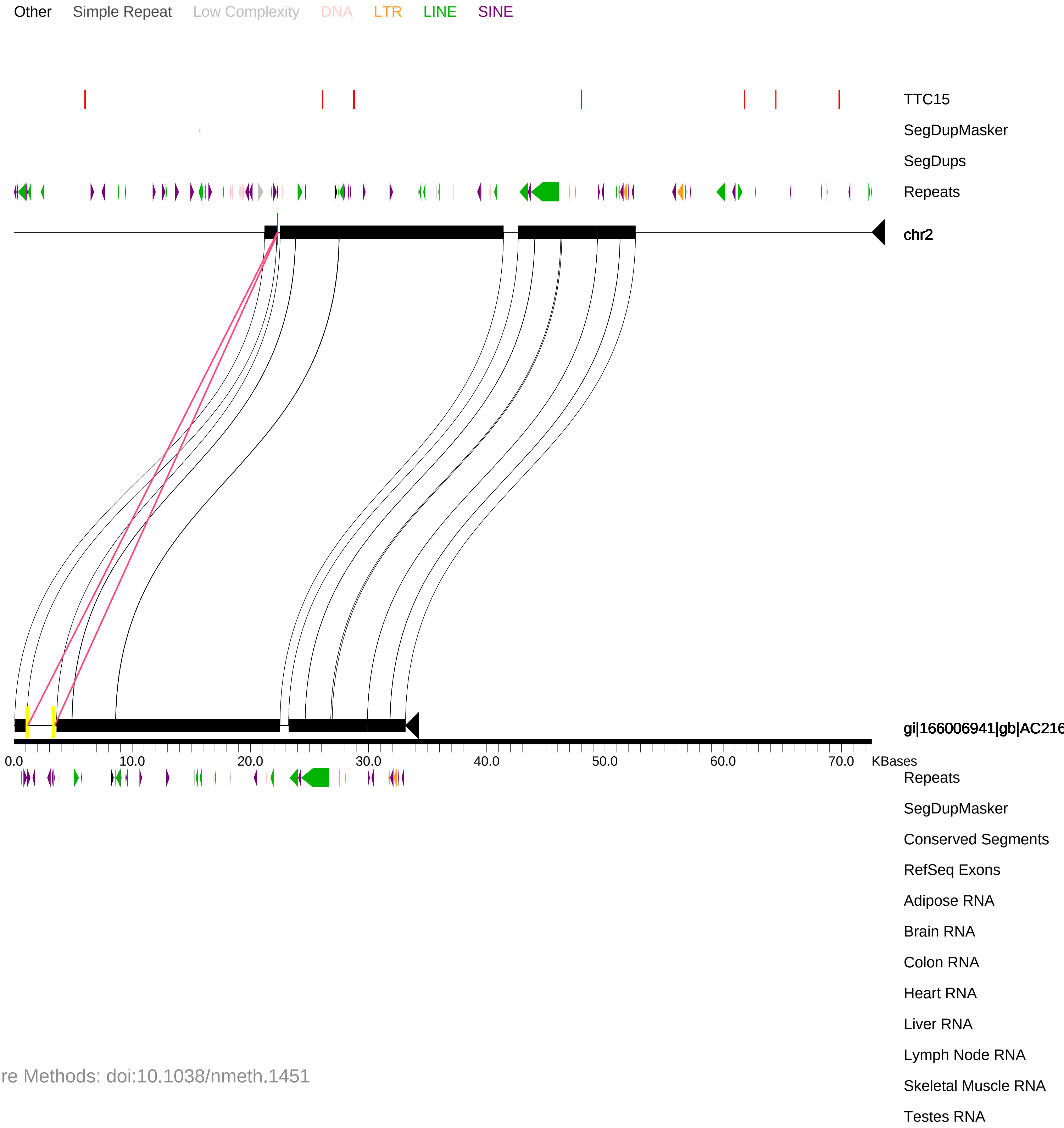
Clone file = AC216138.fa

Insertion Size: 6583



Clone file = AC216281.fa

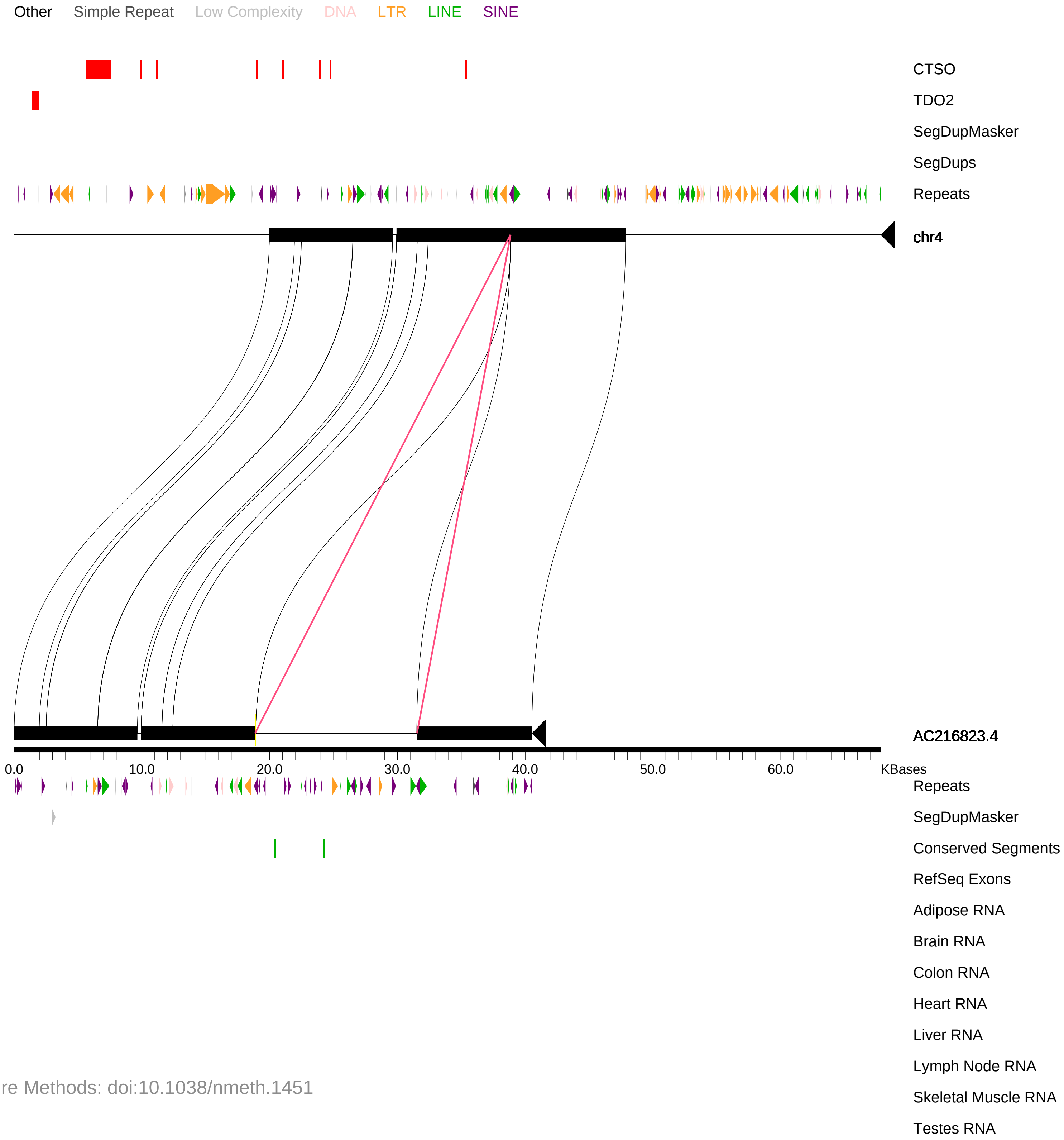
Insertion Size: 2274



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC216823.rc.fa

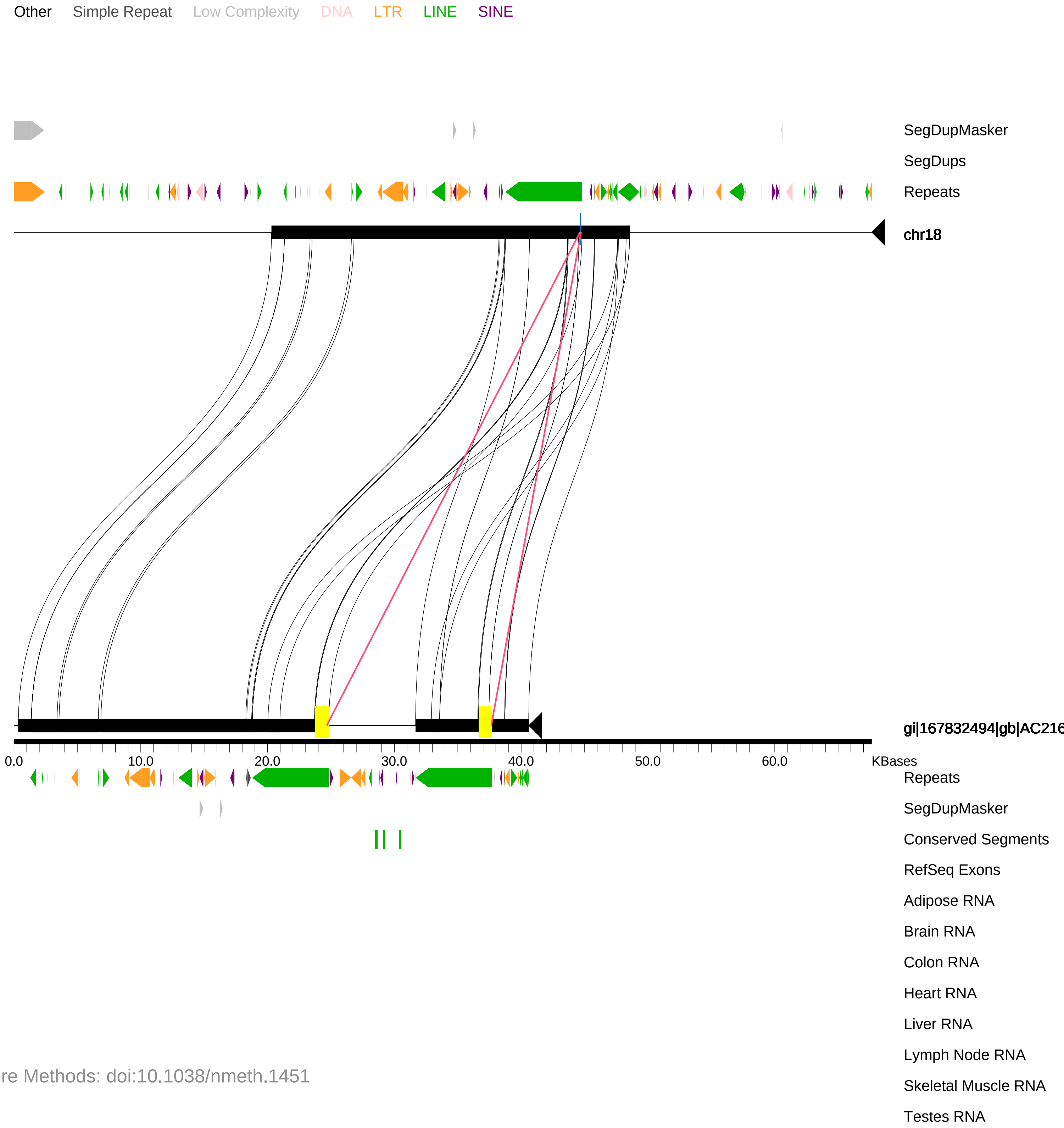
Insertion Size: 12675



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC216971.fa

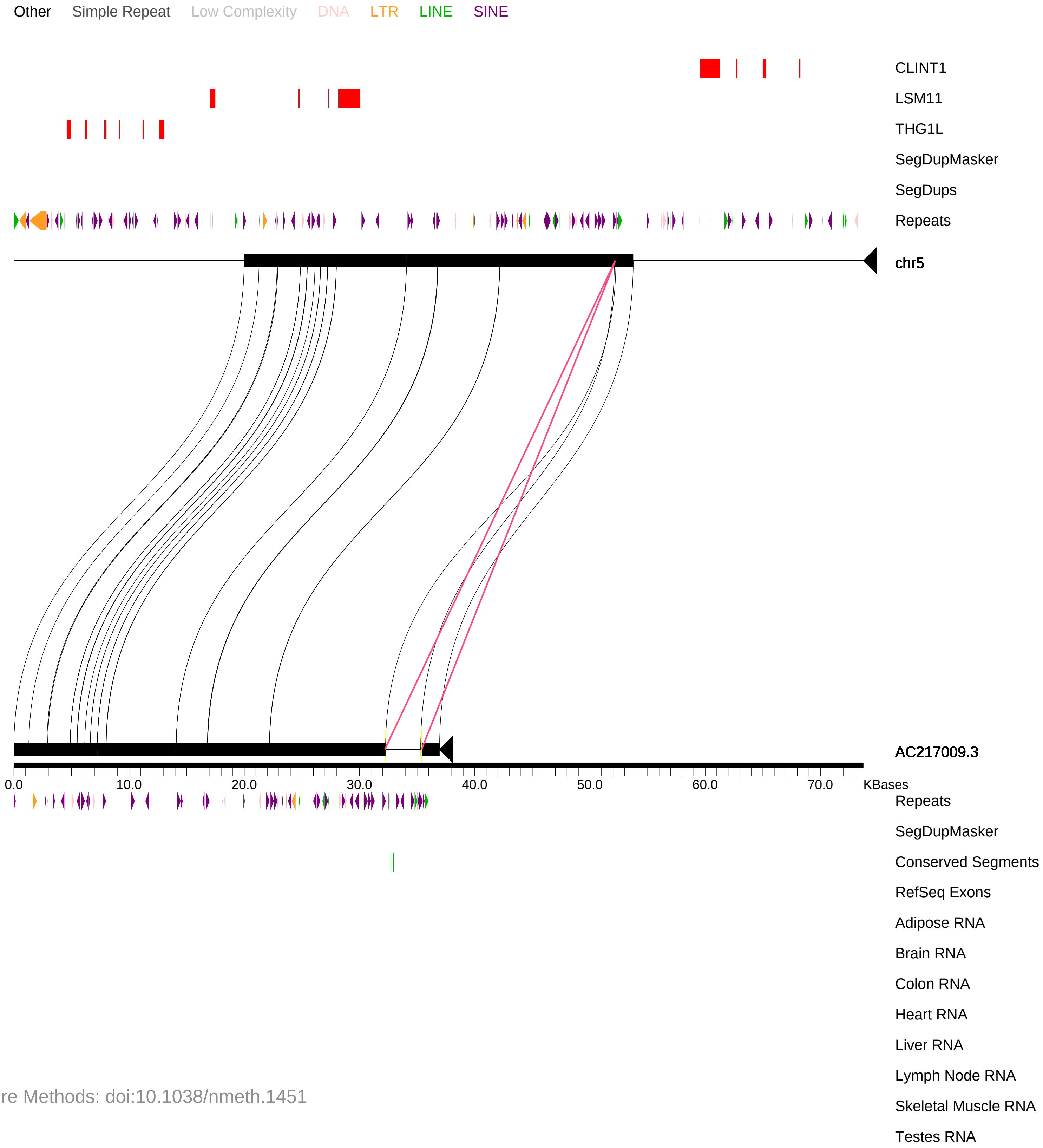
Insertion Size: 12976



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC217009.rc.fa

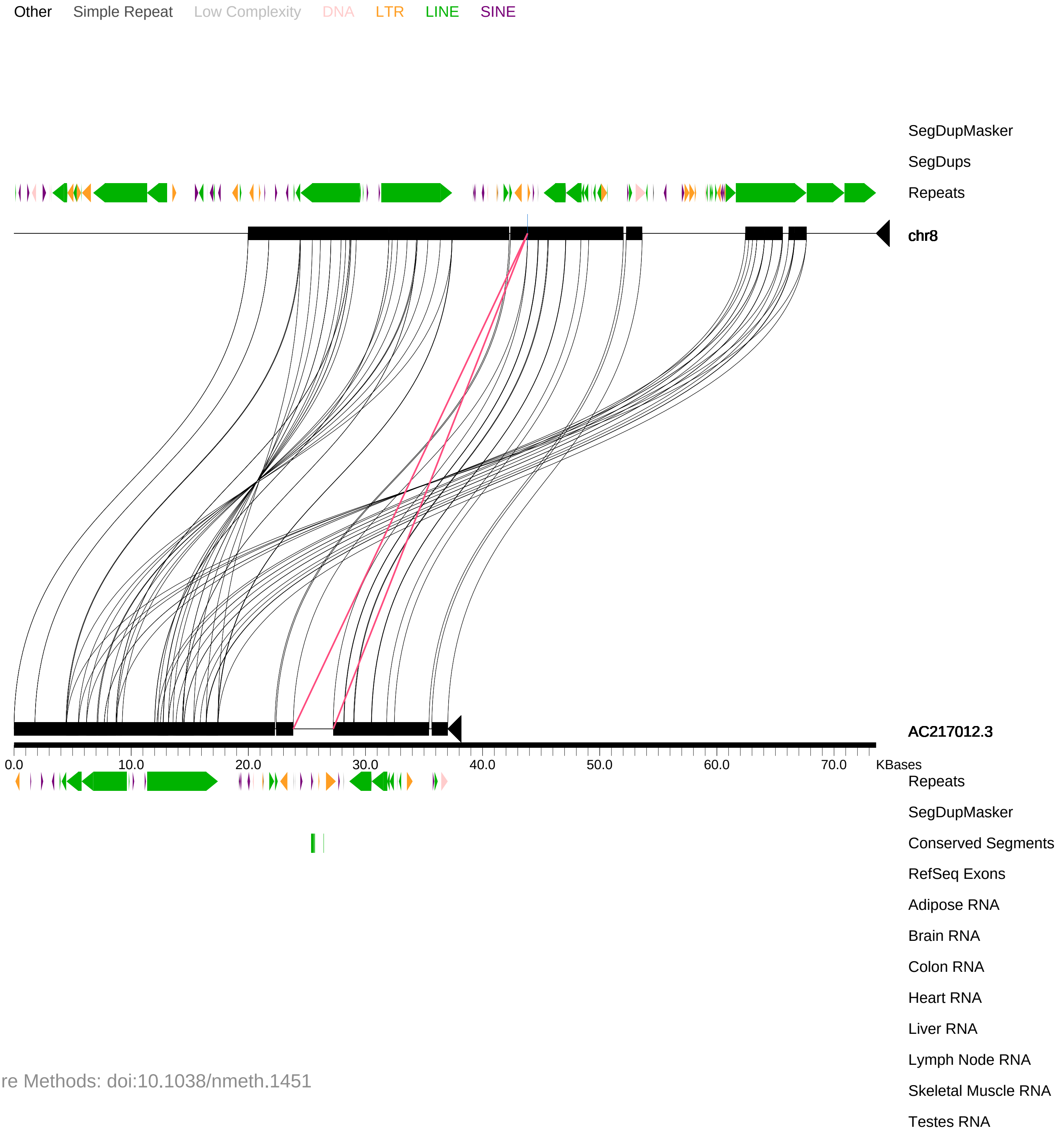
Insertion Size: 3192



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC217012.rc.fa

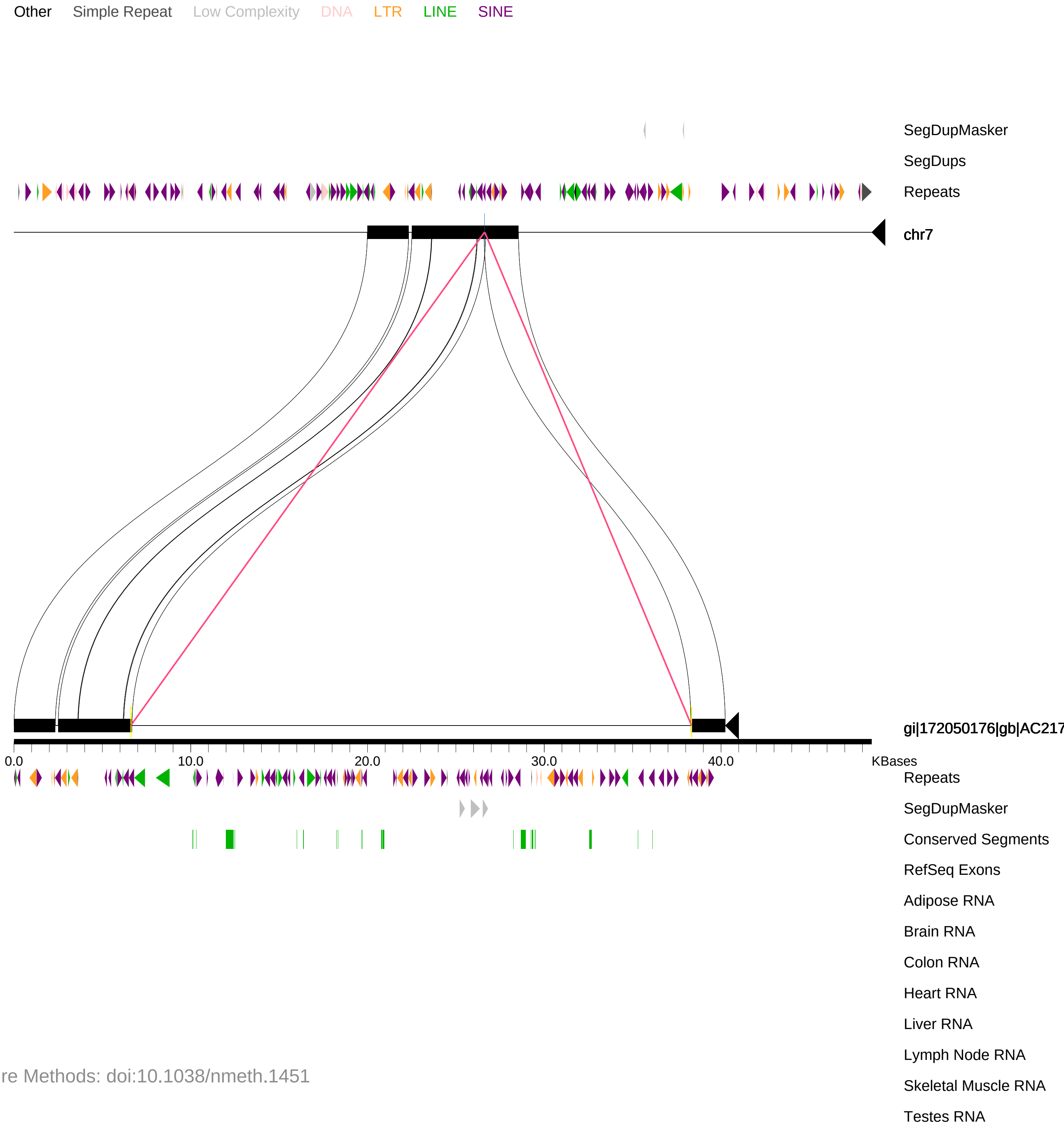
Insertion Size: 3412



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC217018.fa

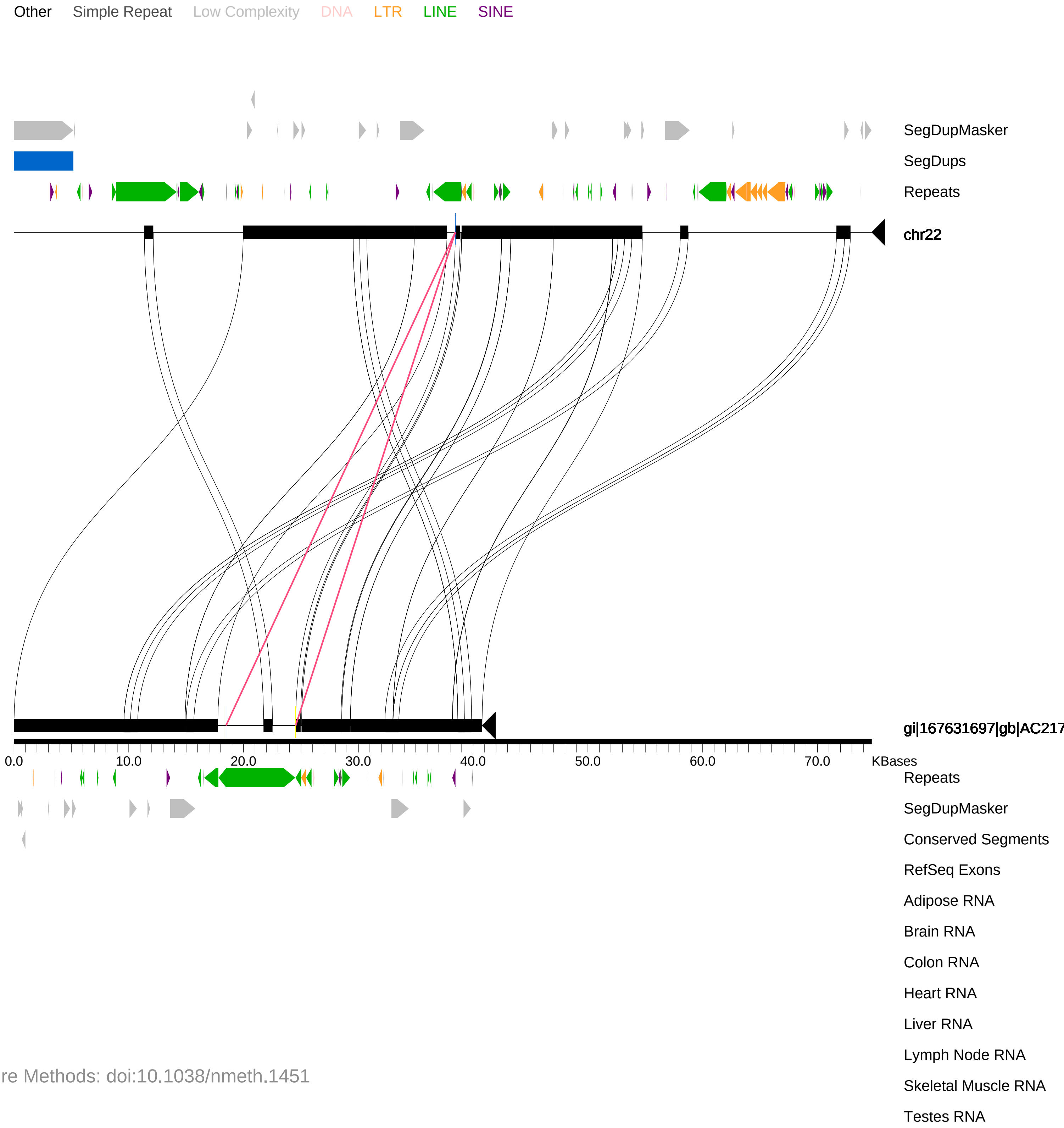
Insertion Size: 31727



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC217064.fa

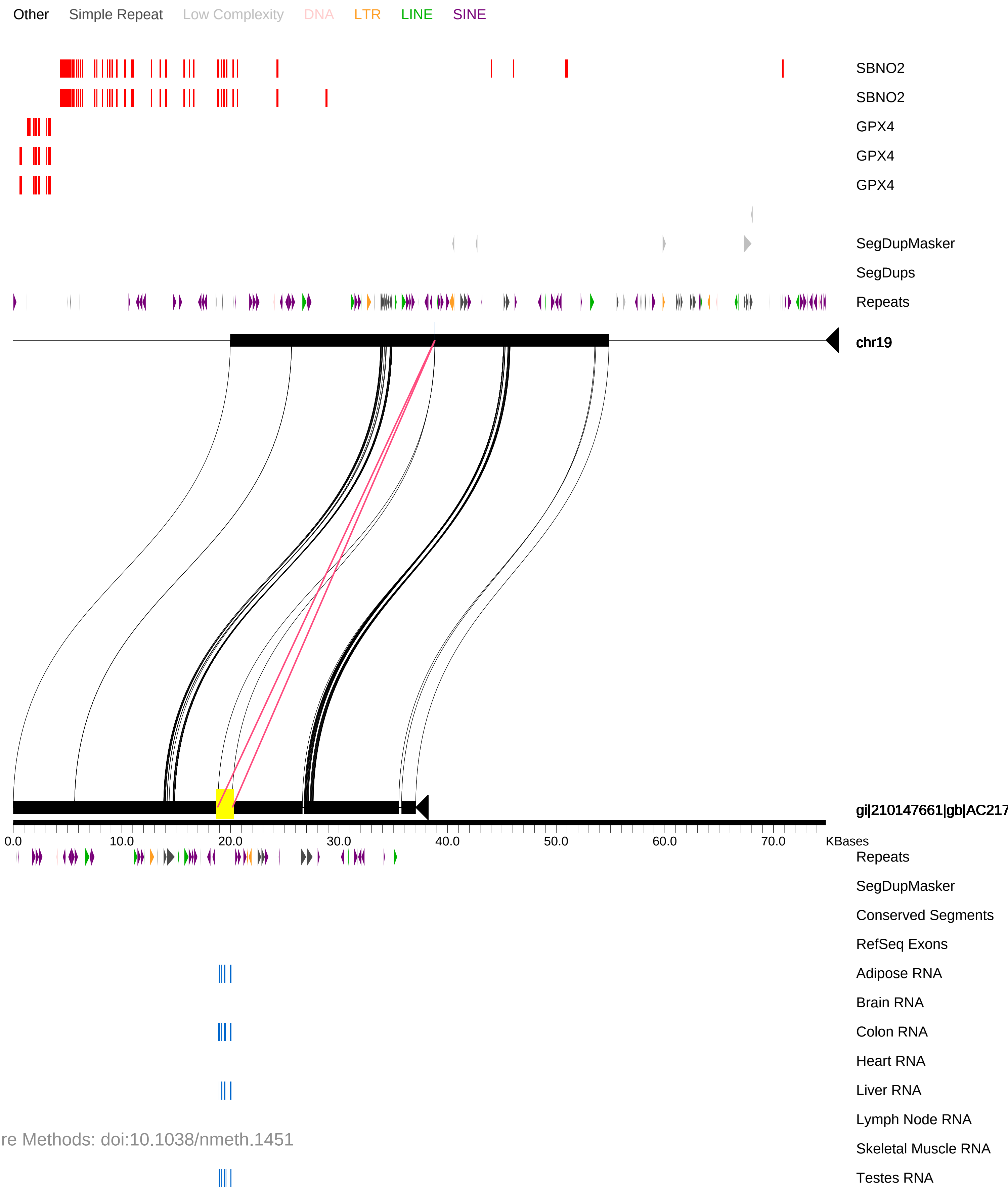
Insertion Size: 6067



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC217140.fa

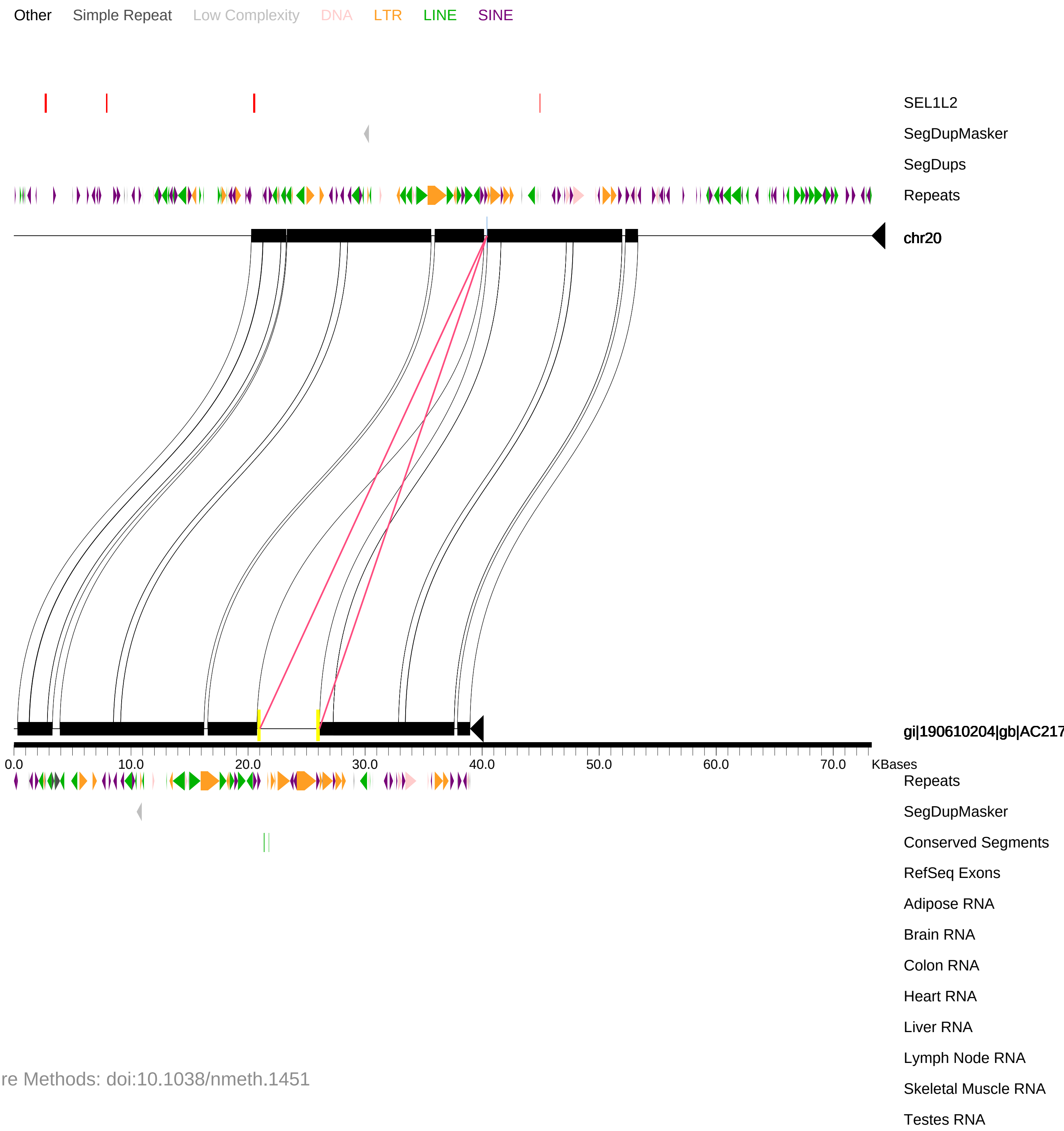
Insertion Size: 1370



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC217326.fa

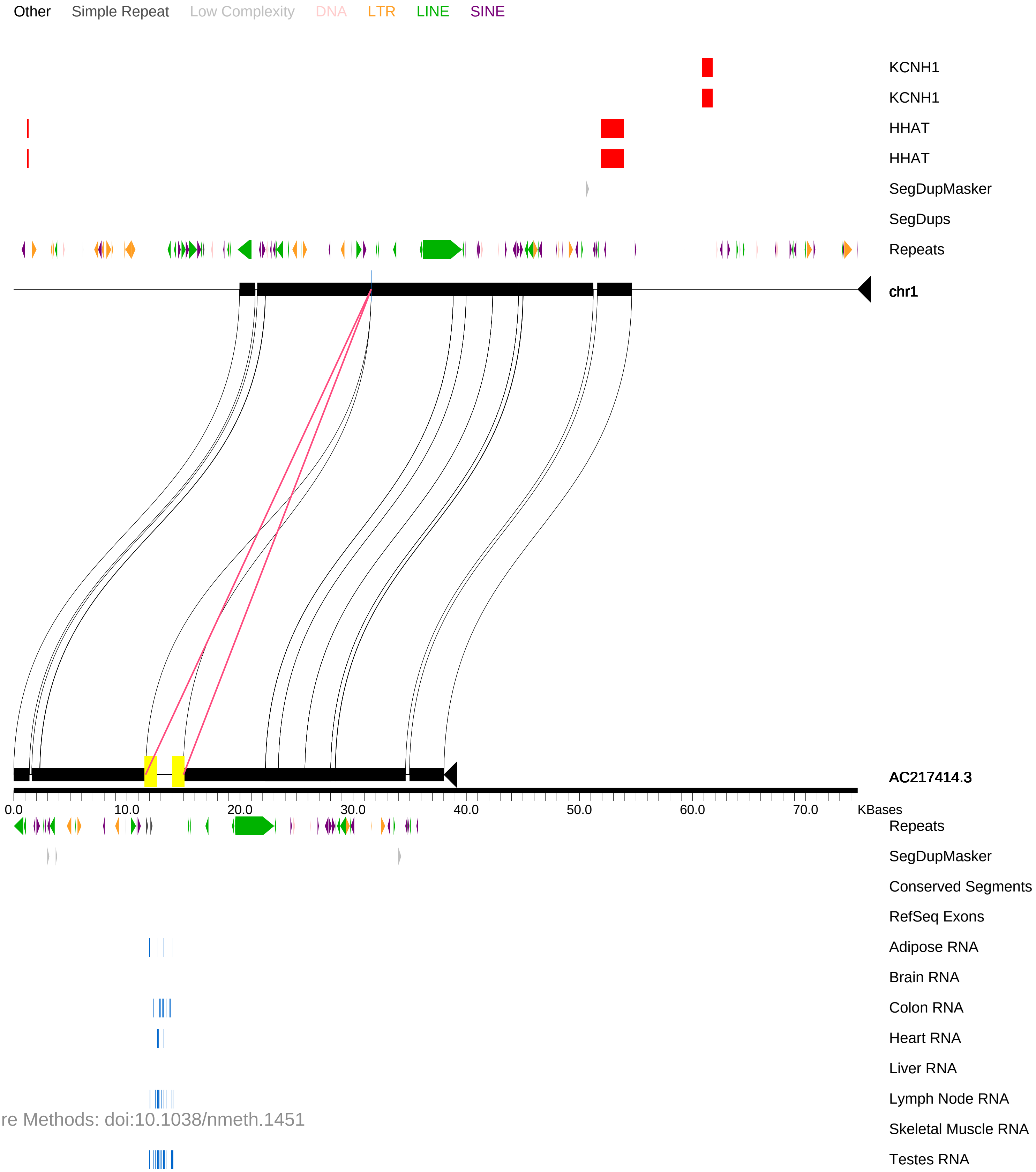
Insertion Size: 5083



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC217414.rc.fa

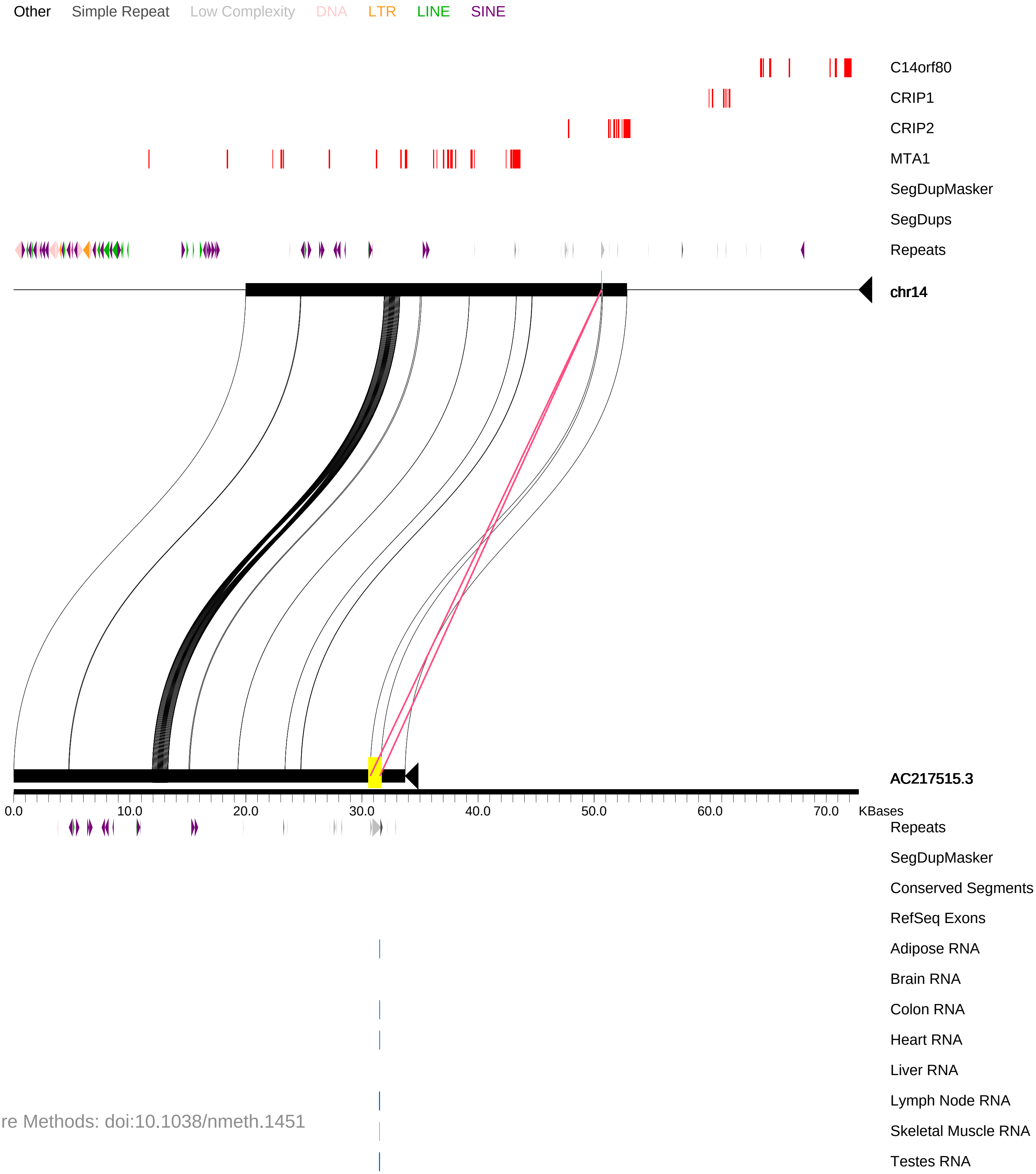
Insertion Size: 3348



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC217515.rc.fa

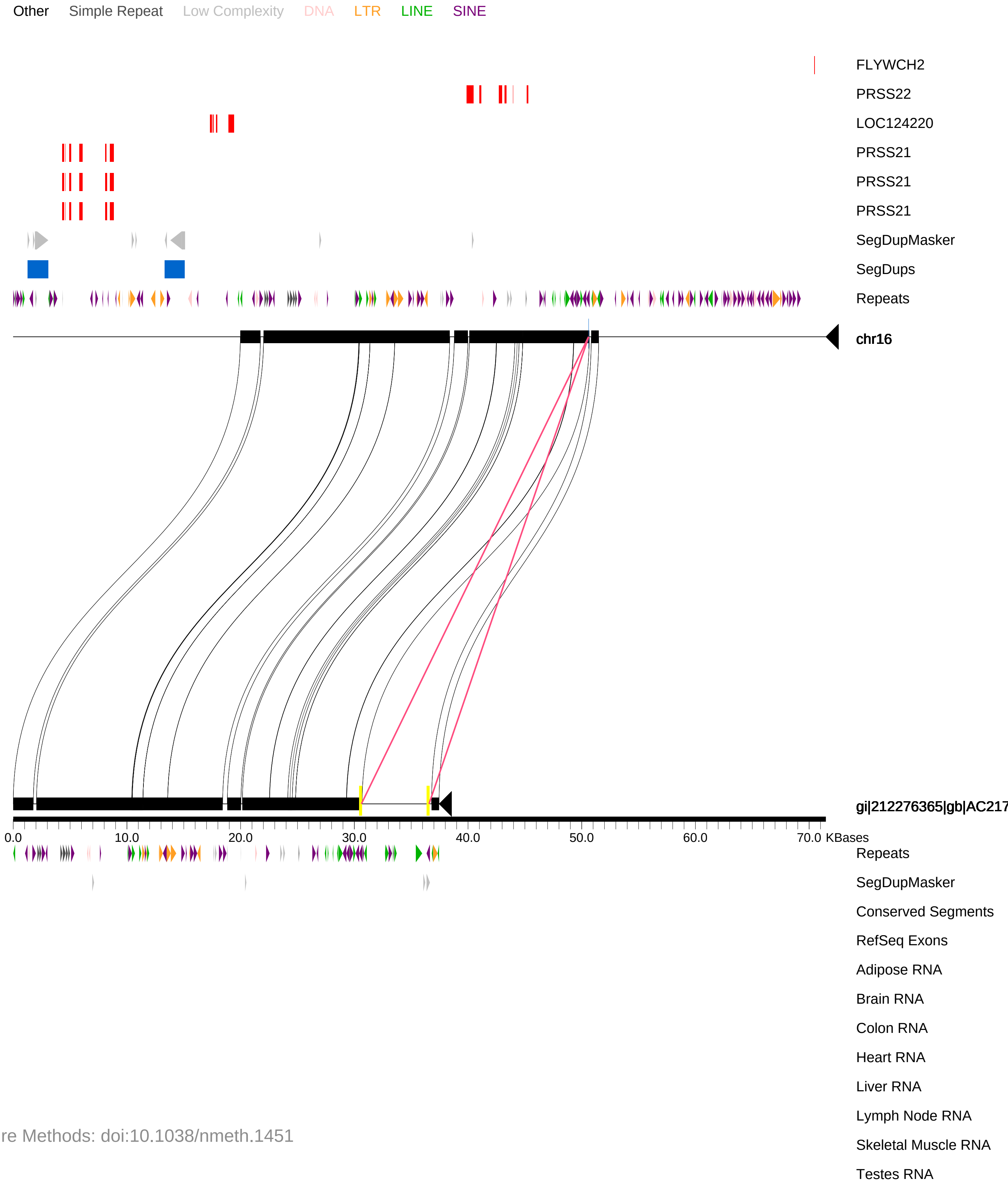
Insertion Size: 839



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC217628.fa

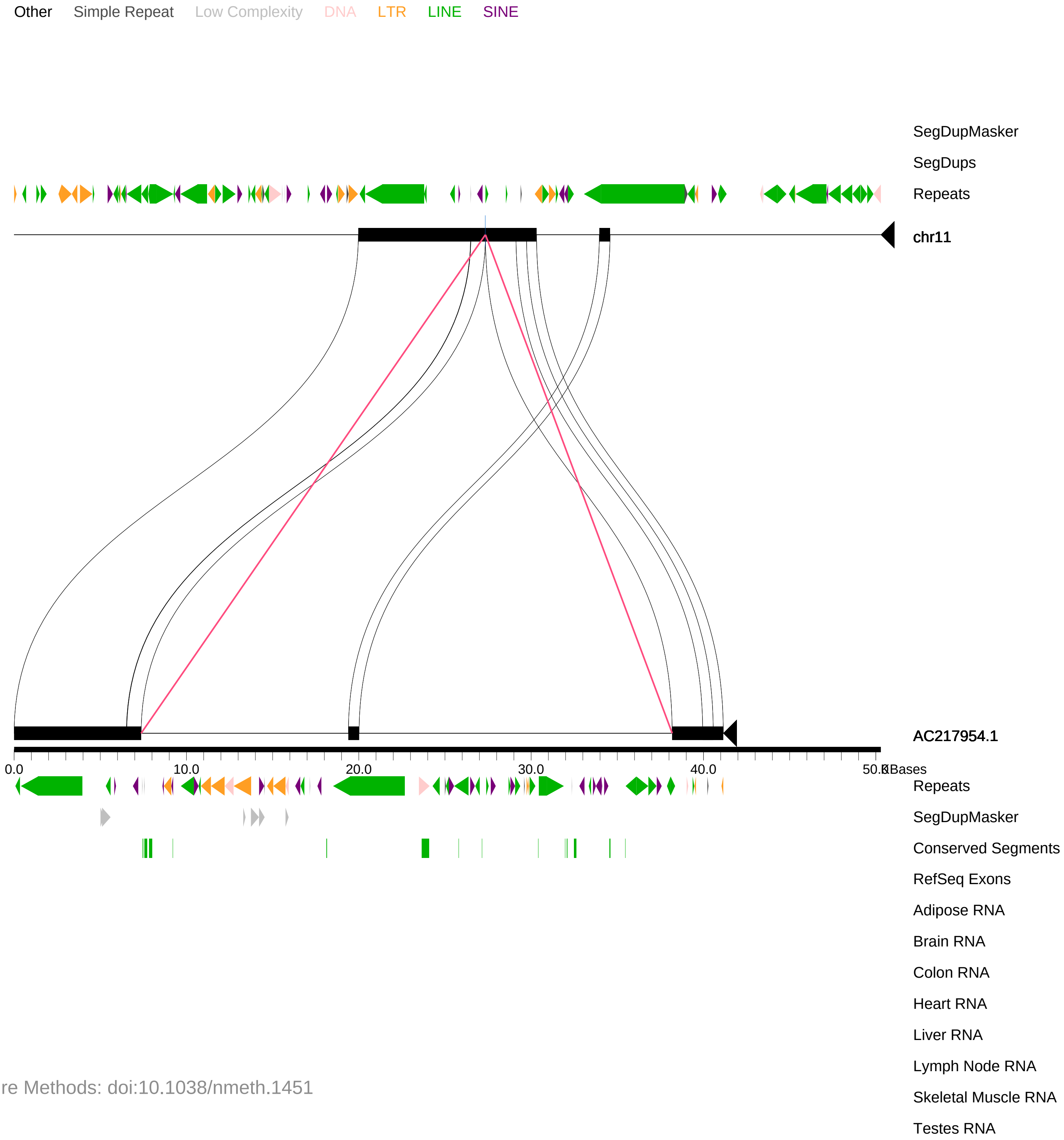
Insertion Size: 5948



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC217954.rc.fa

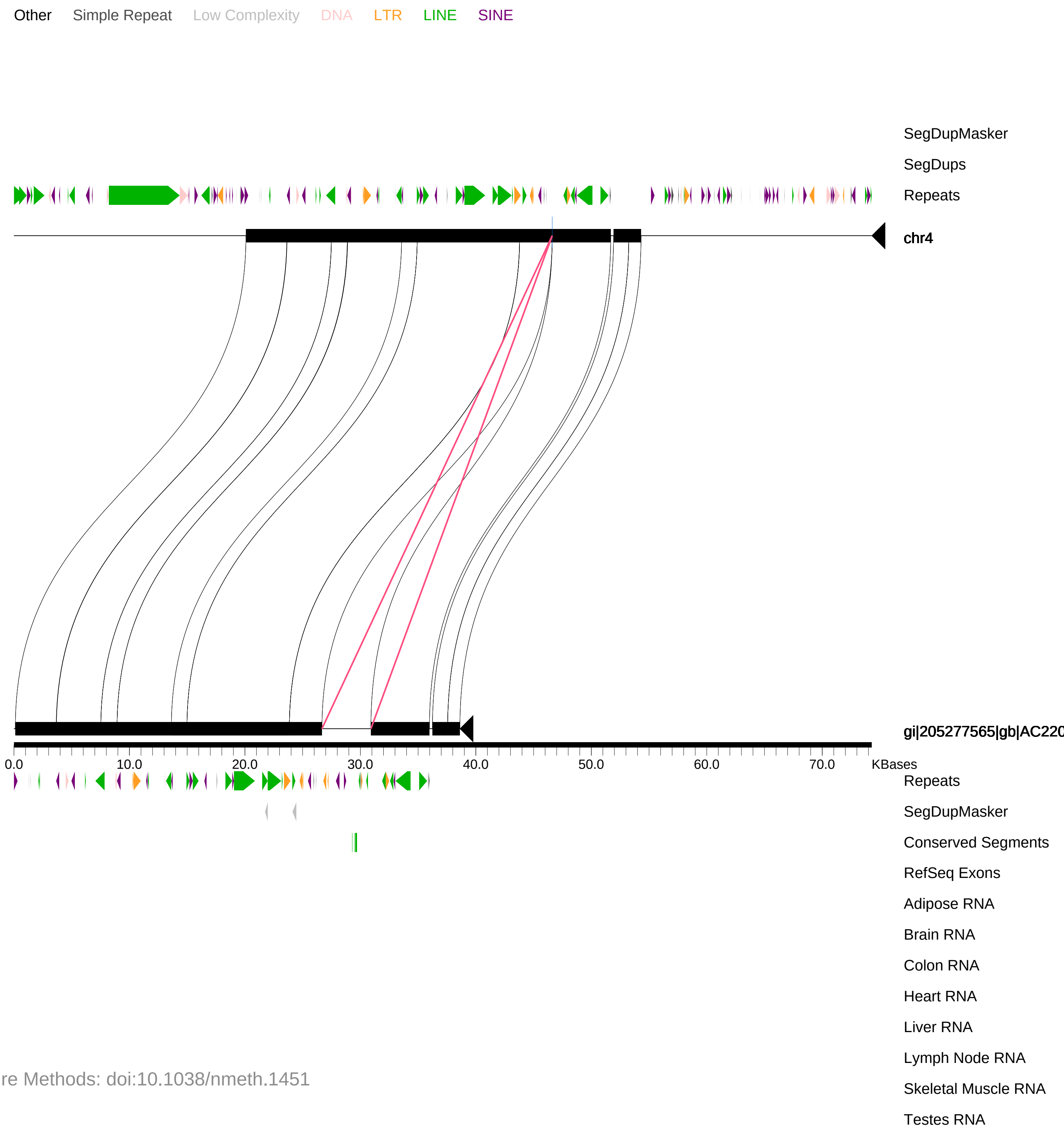
Insertion Size: 30820



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC220966.fa

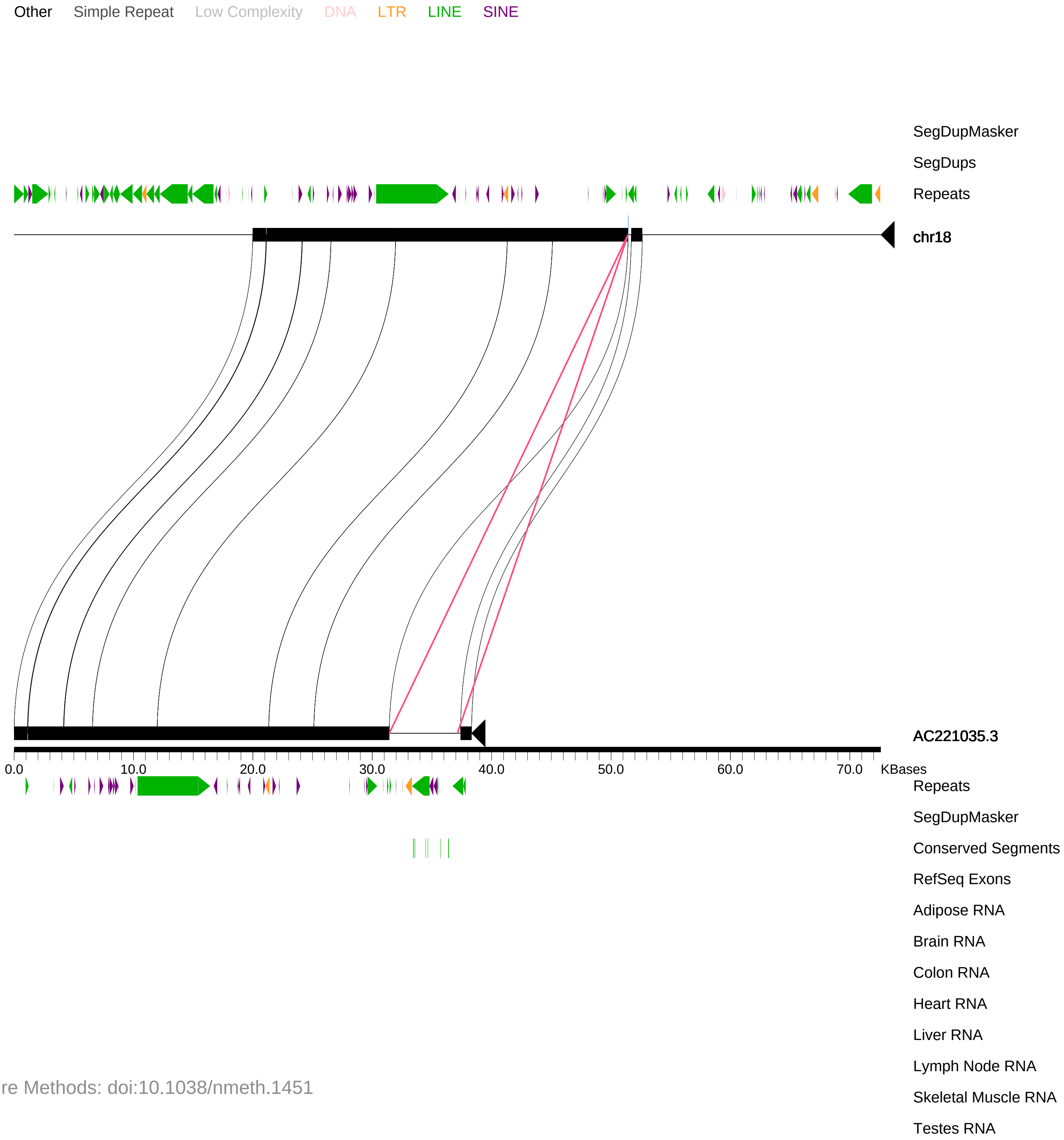
Insertion Size: 4243



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC221035.rc.fa

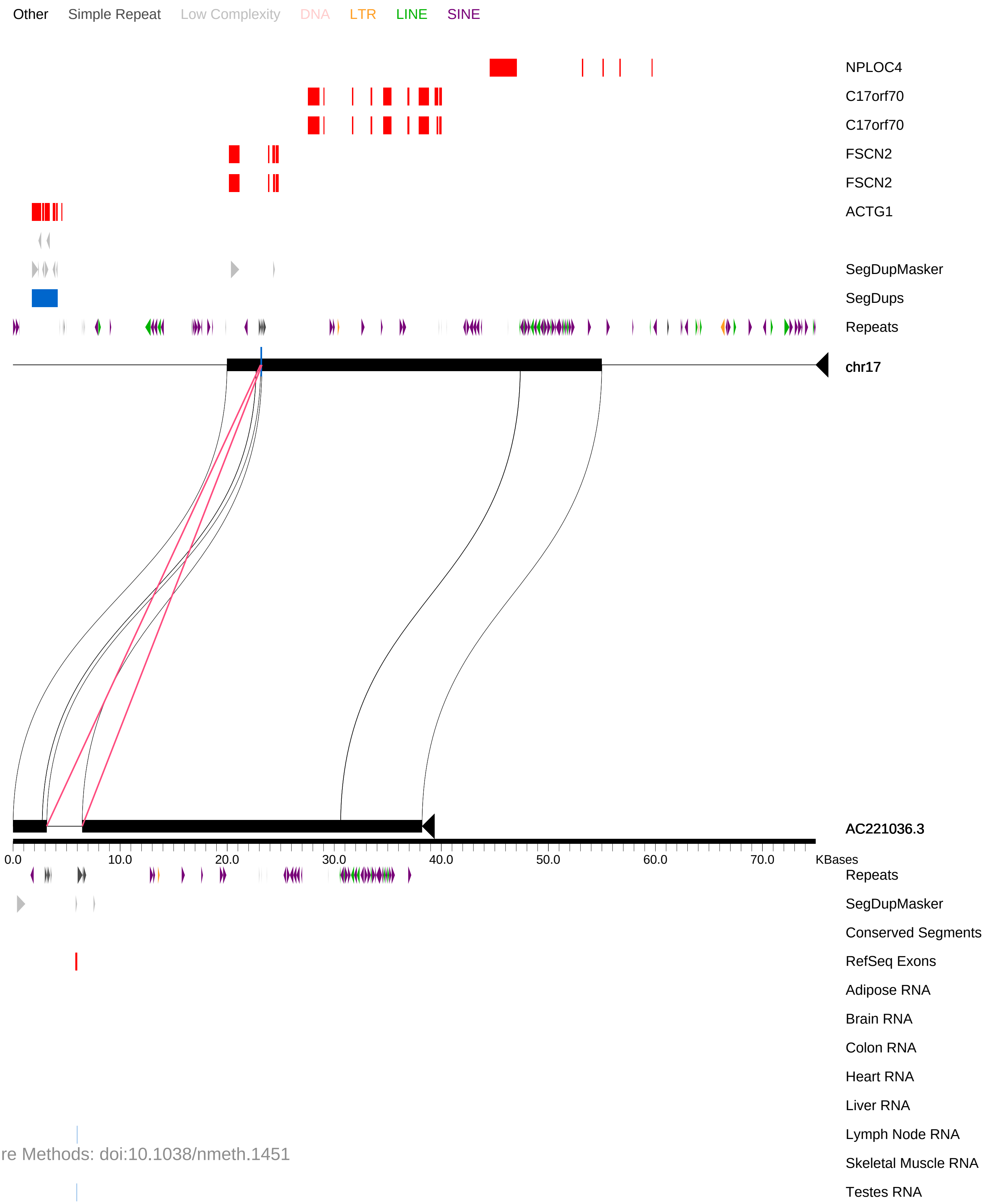
Insertion Size: 5699



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC221036.rc.fa

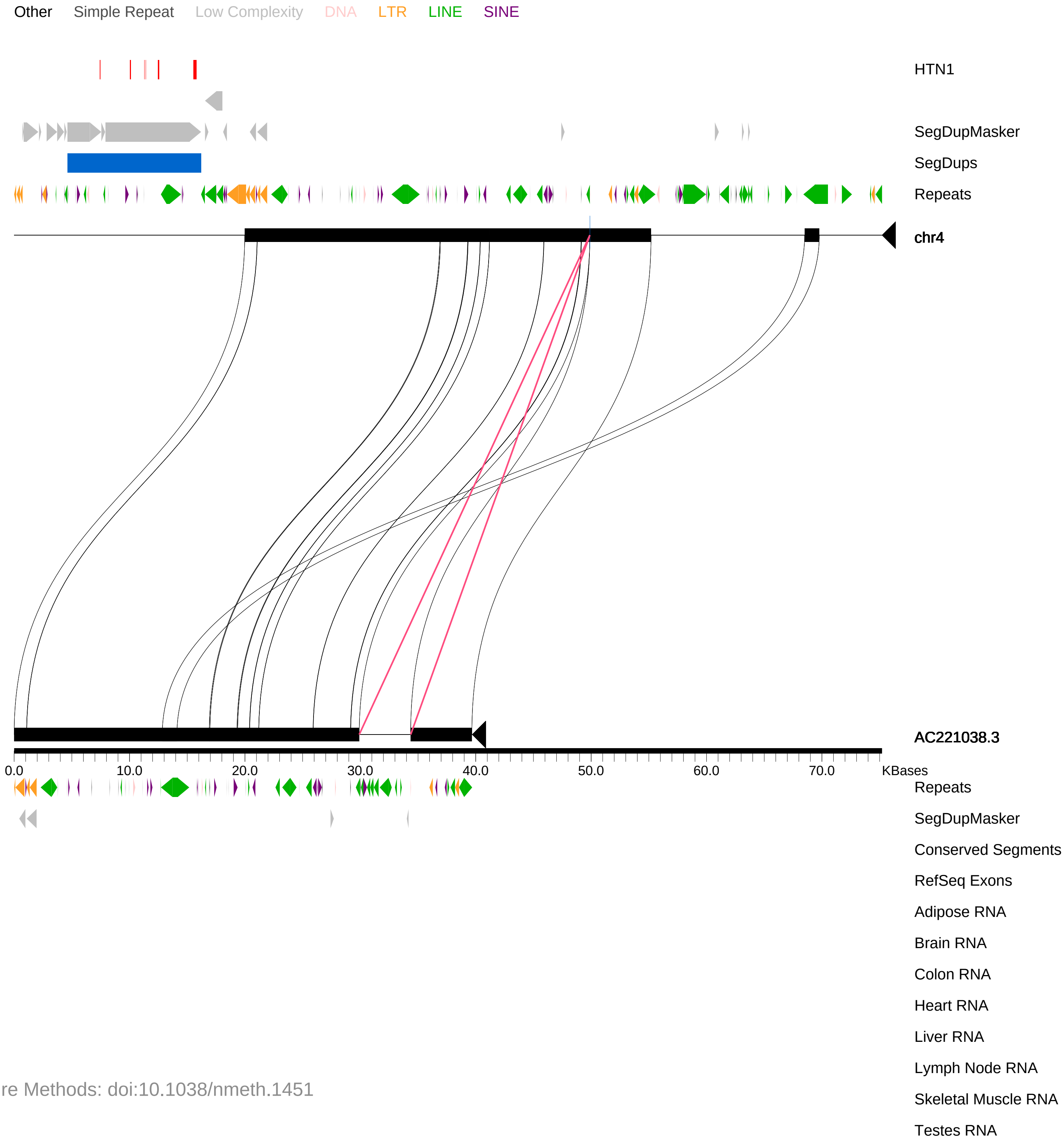
Insertion Size: 3326



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC221038.rc.fa

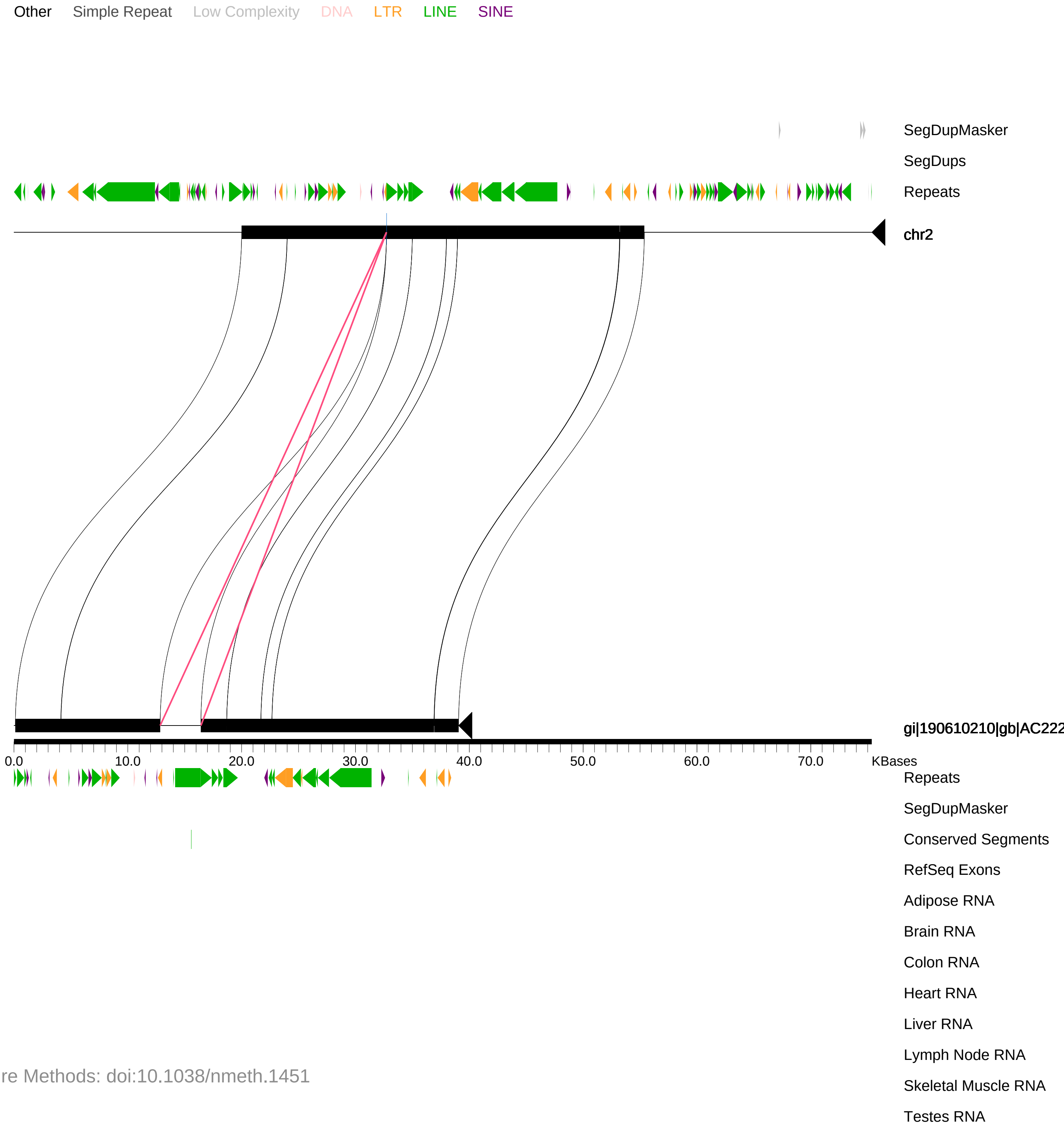
Insertion Size: 4460



Nature Methods: doi:10.1038/nmeth.1451

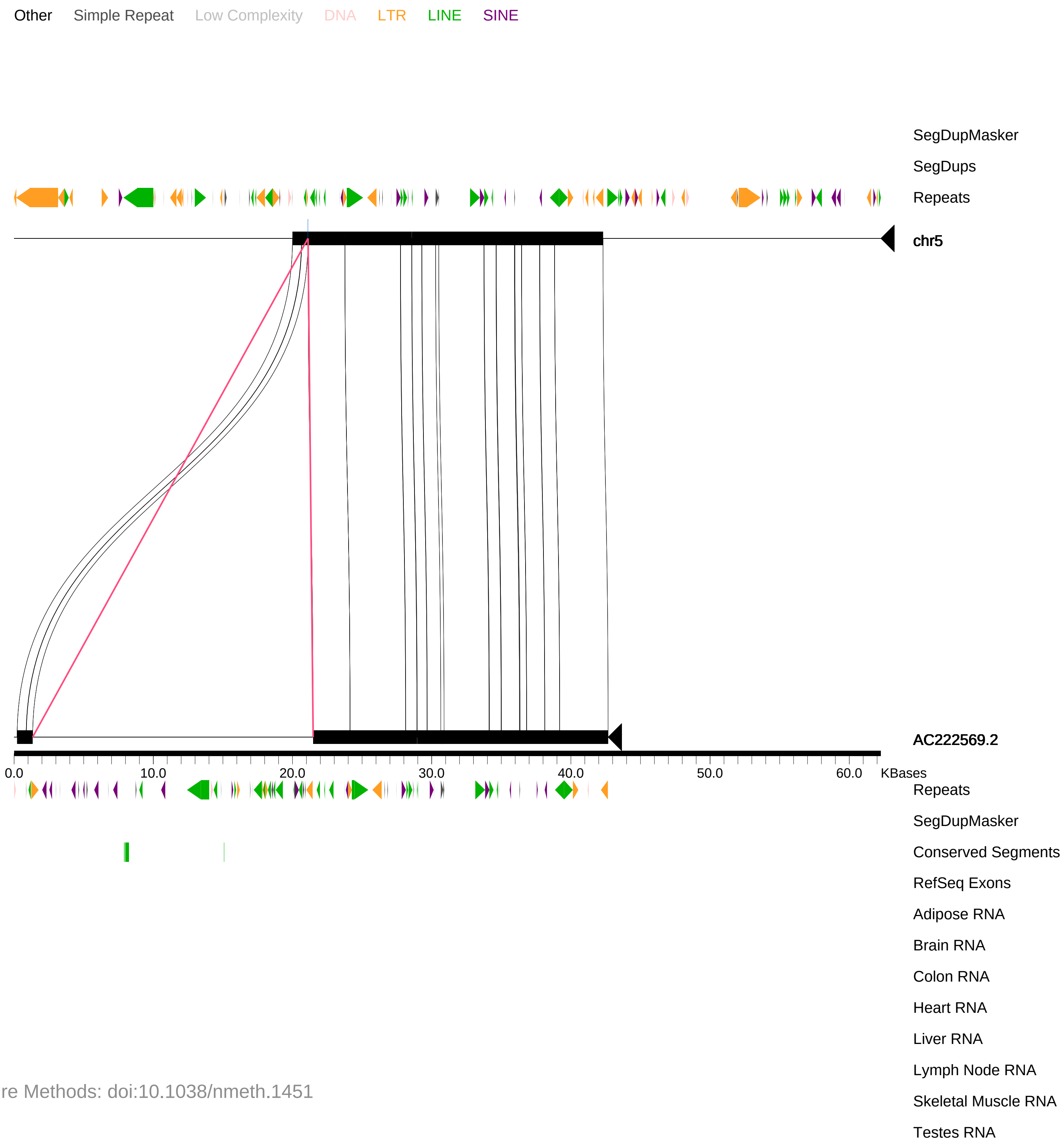
Clone file = AC222568.fa

Insertion Size: 3579



Clone file = AC222569.rc.fa

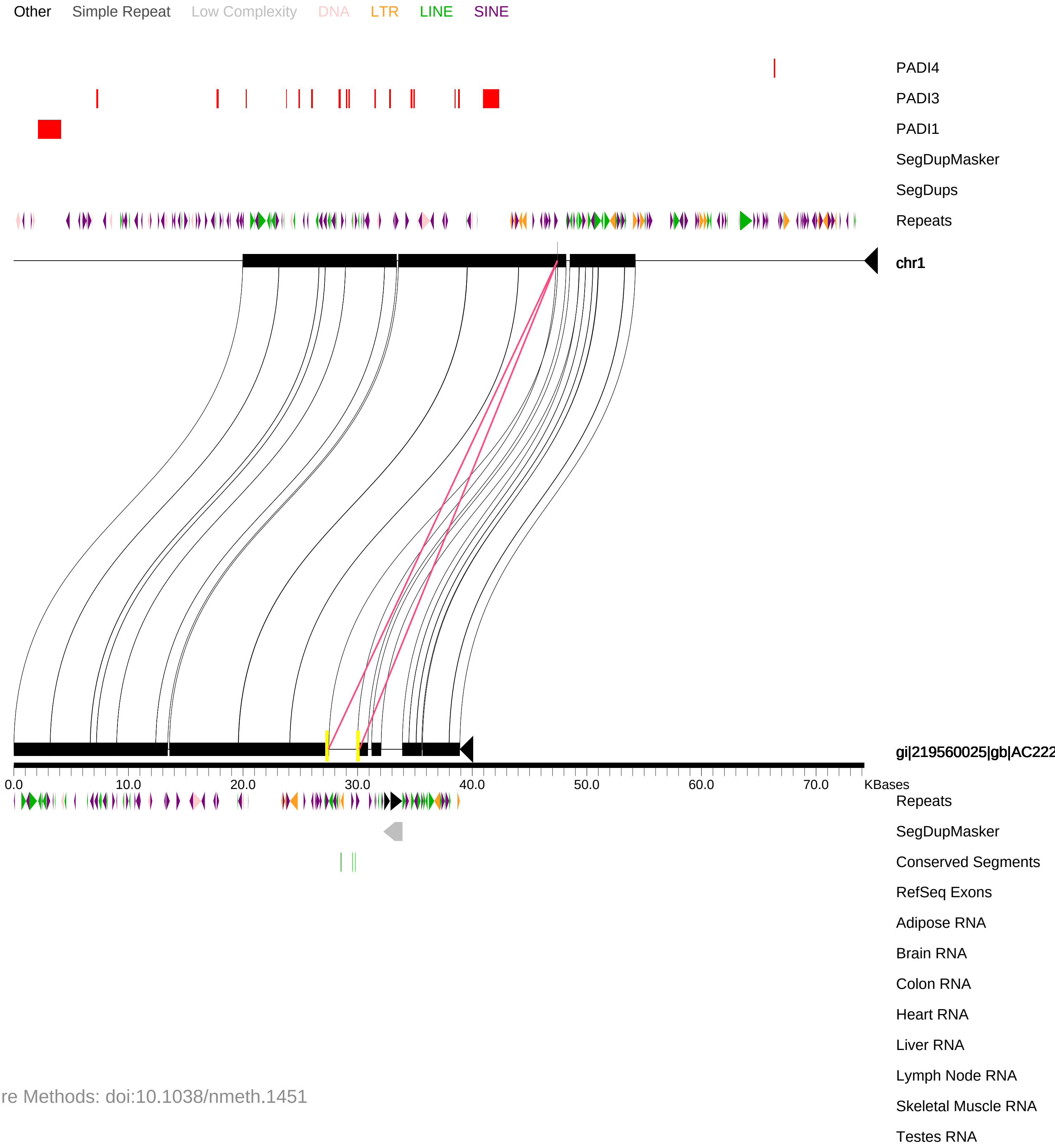
Insertion Size: 20156



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC222570.fa

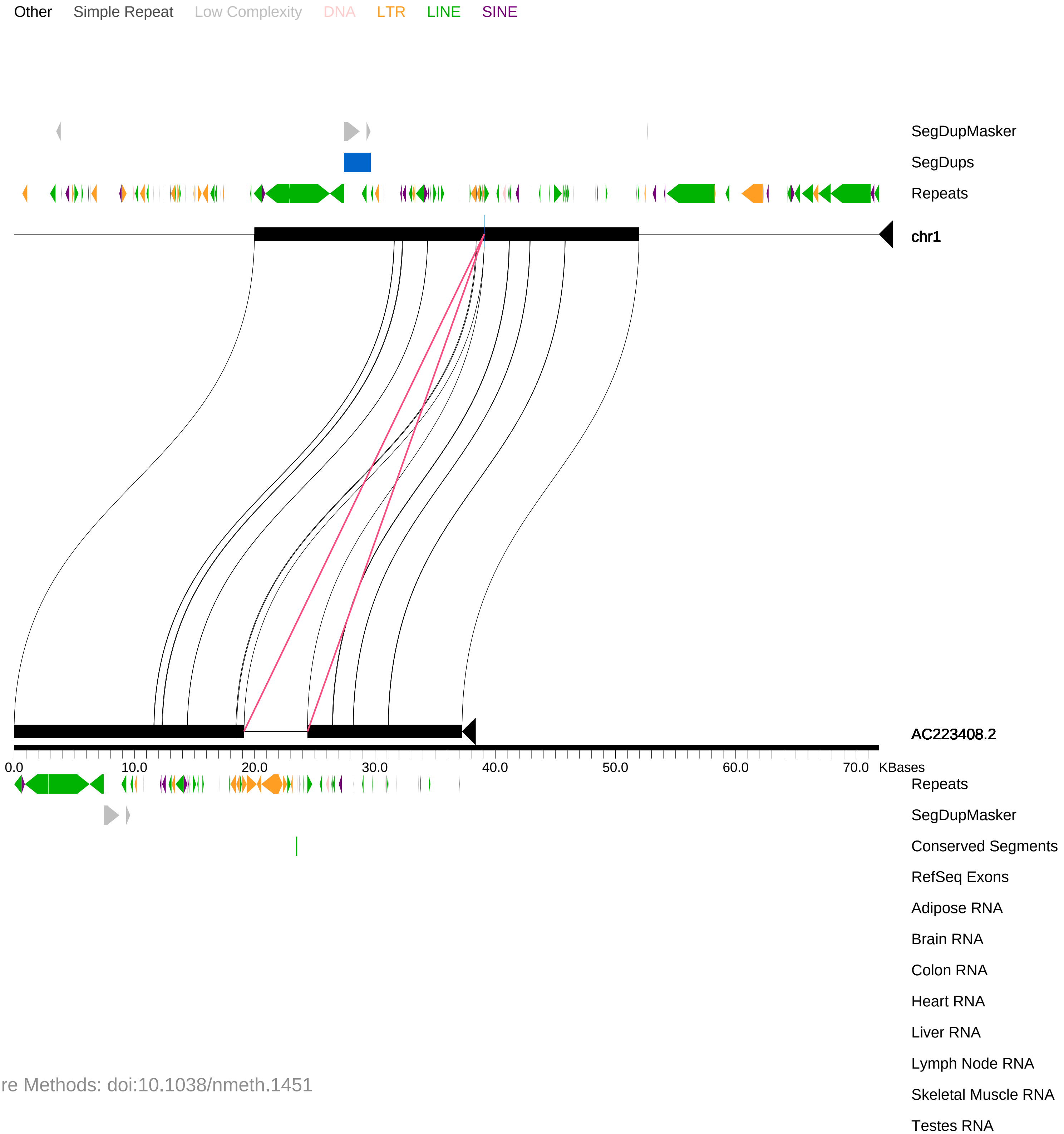
Insertion Size: 2694



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC223408.rc.fa

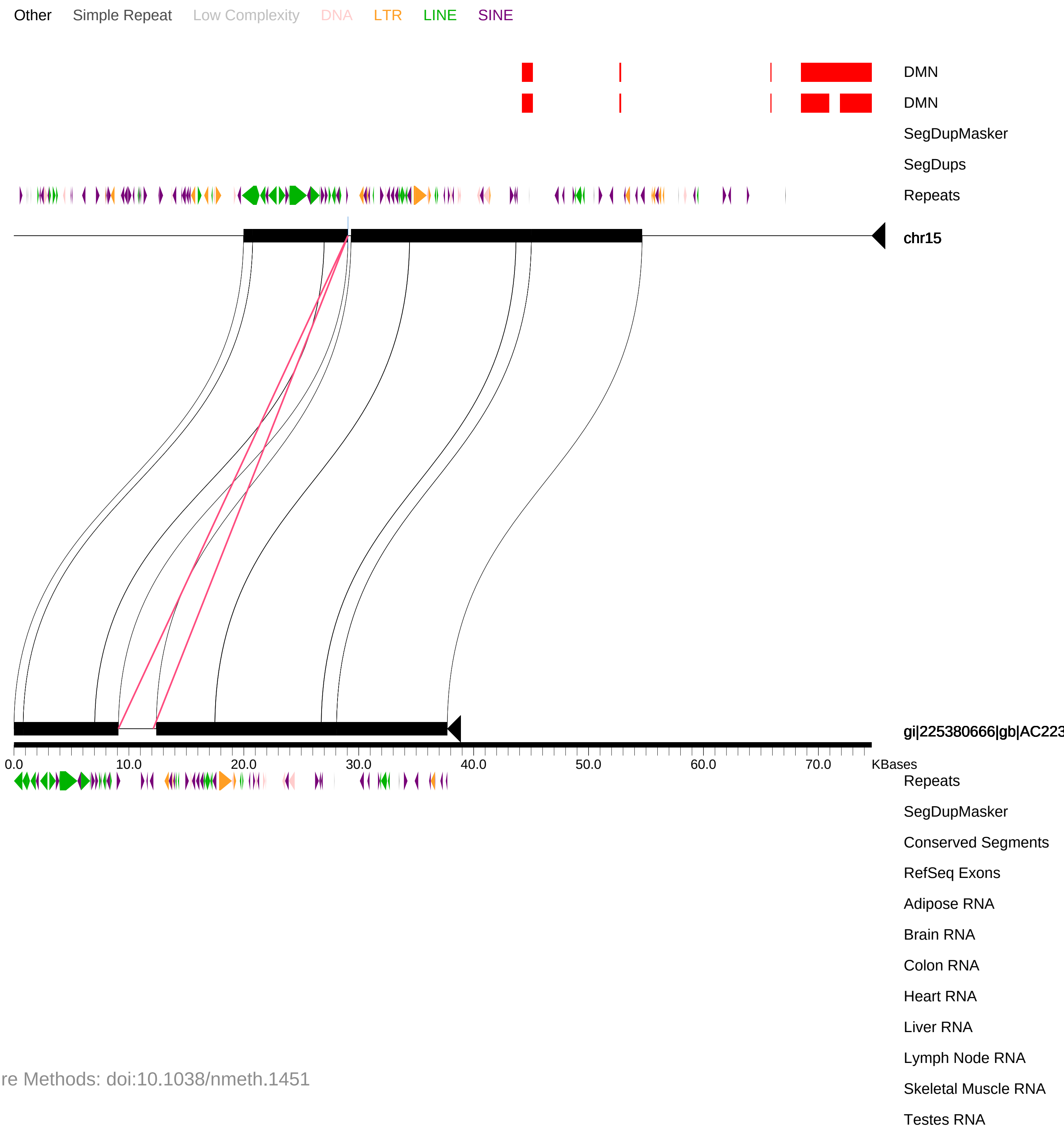
Insertion Size: 5283



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC223423.fa

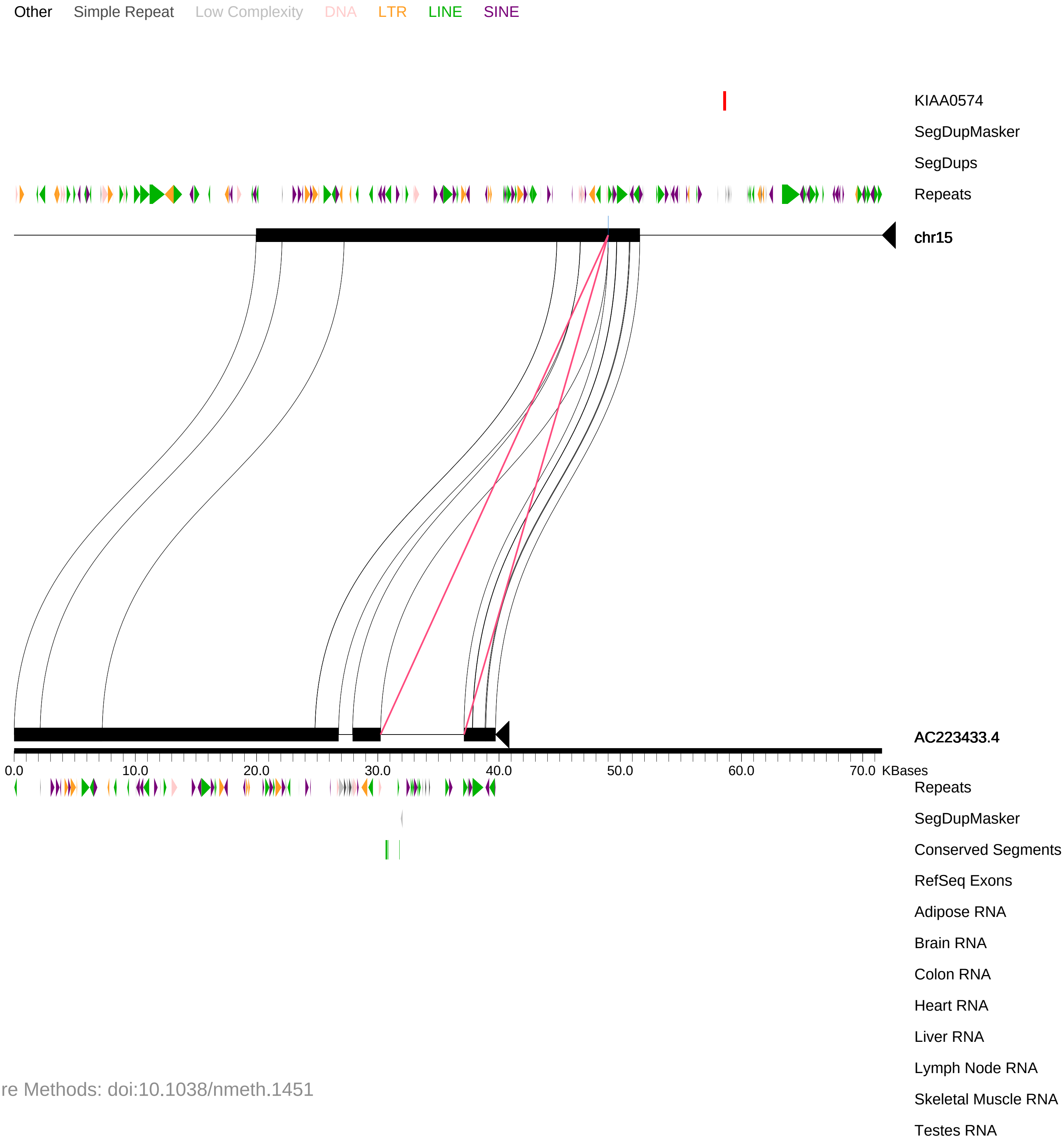
Insertion Size: 3037



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC223433.rc.fa

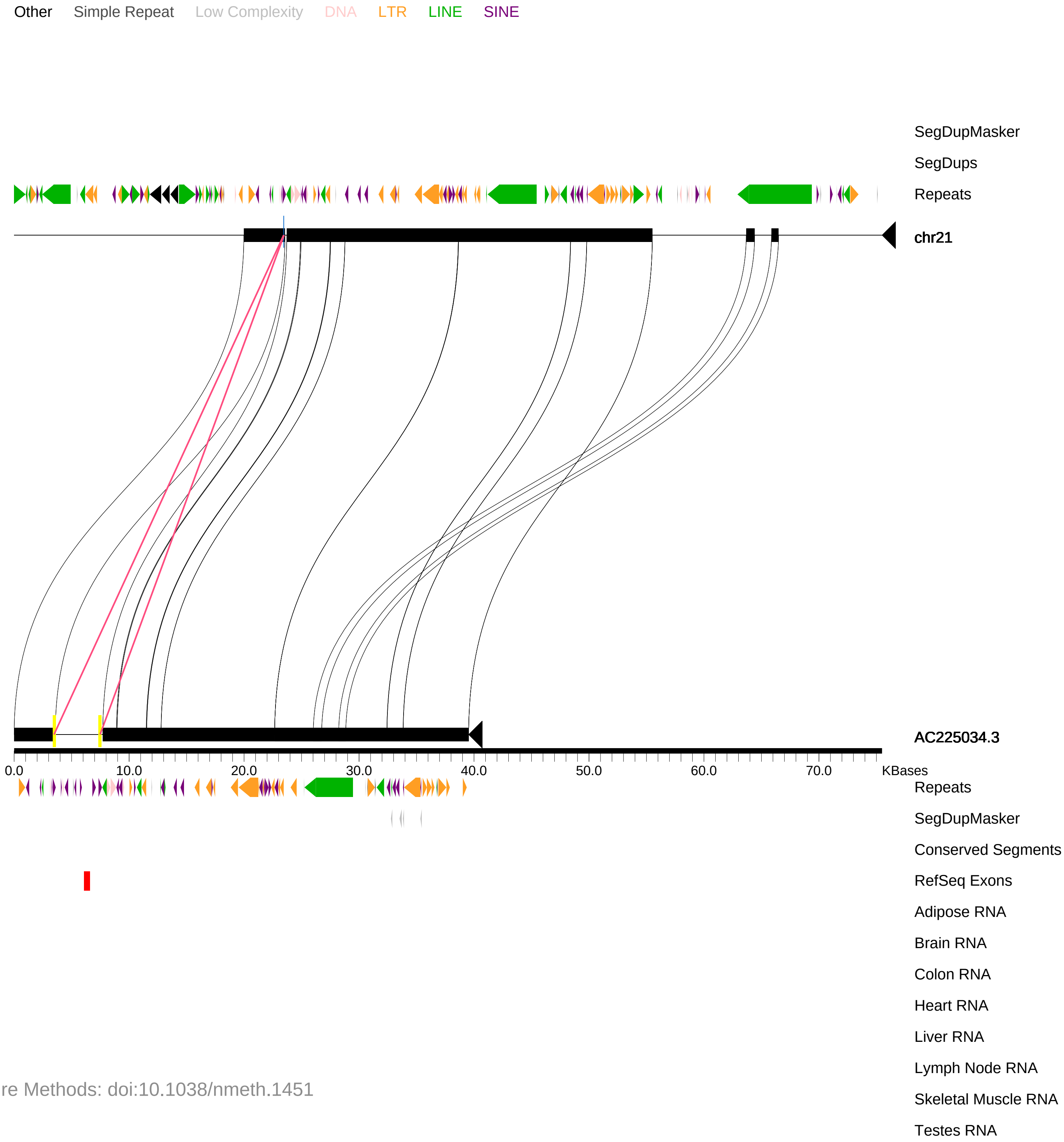
Insertion Size: 6887



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC225034.rc.fa

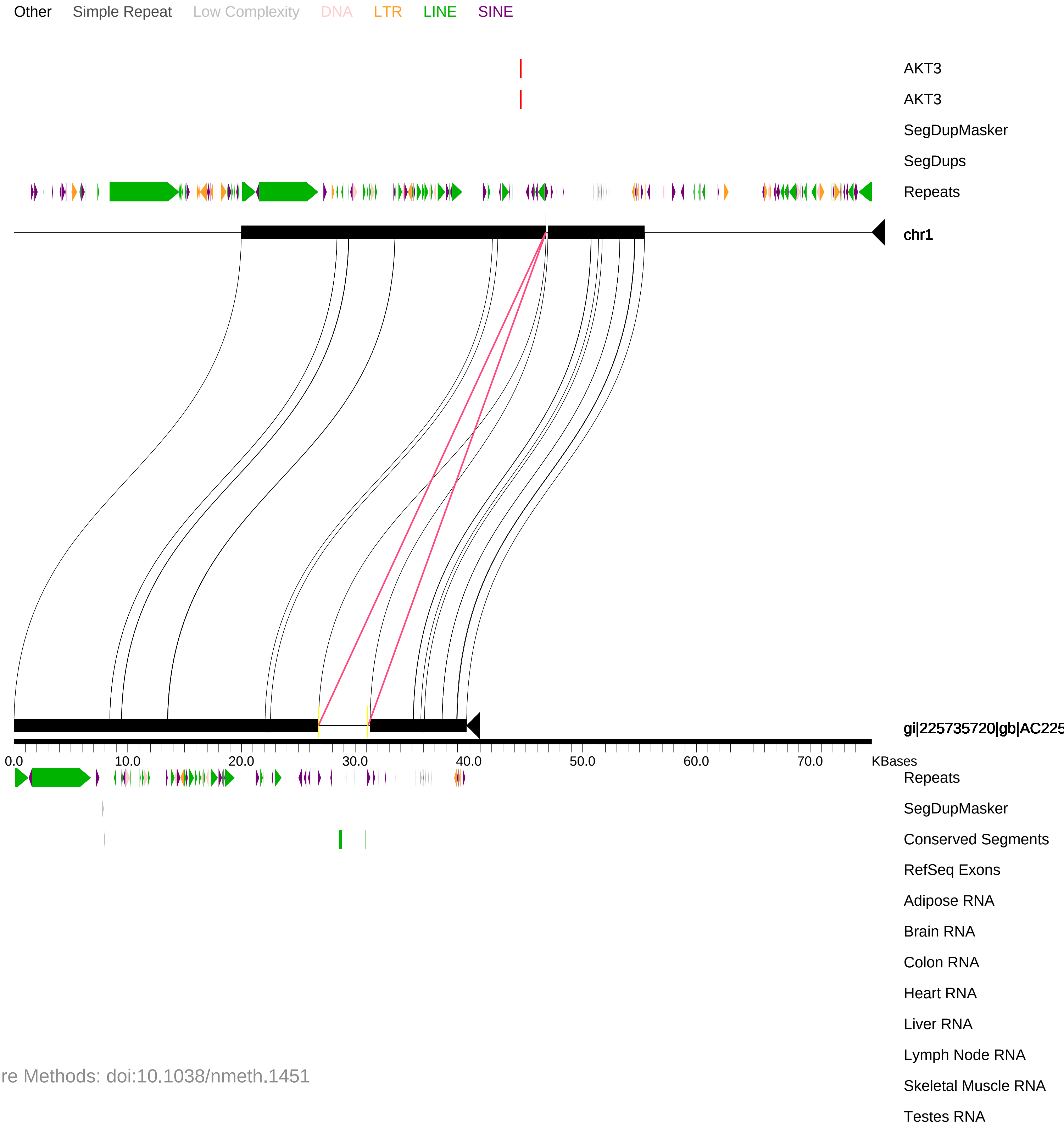
Insertion Size: 4010



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC225099.fa

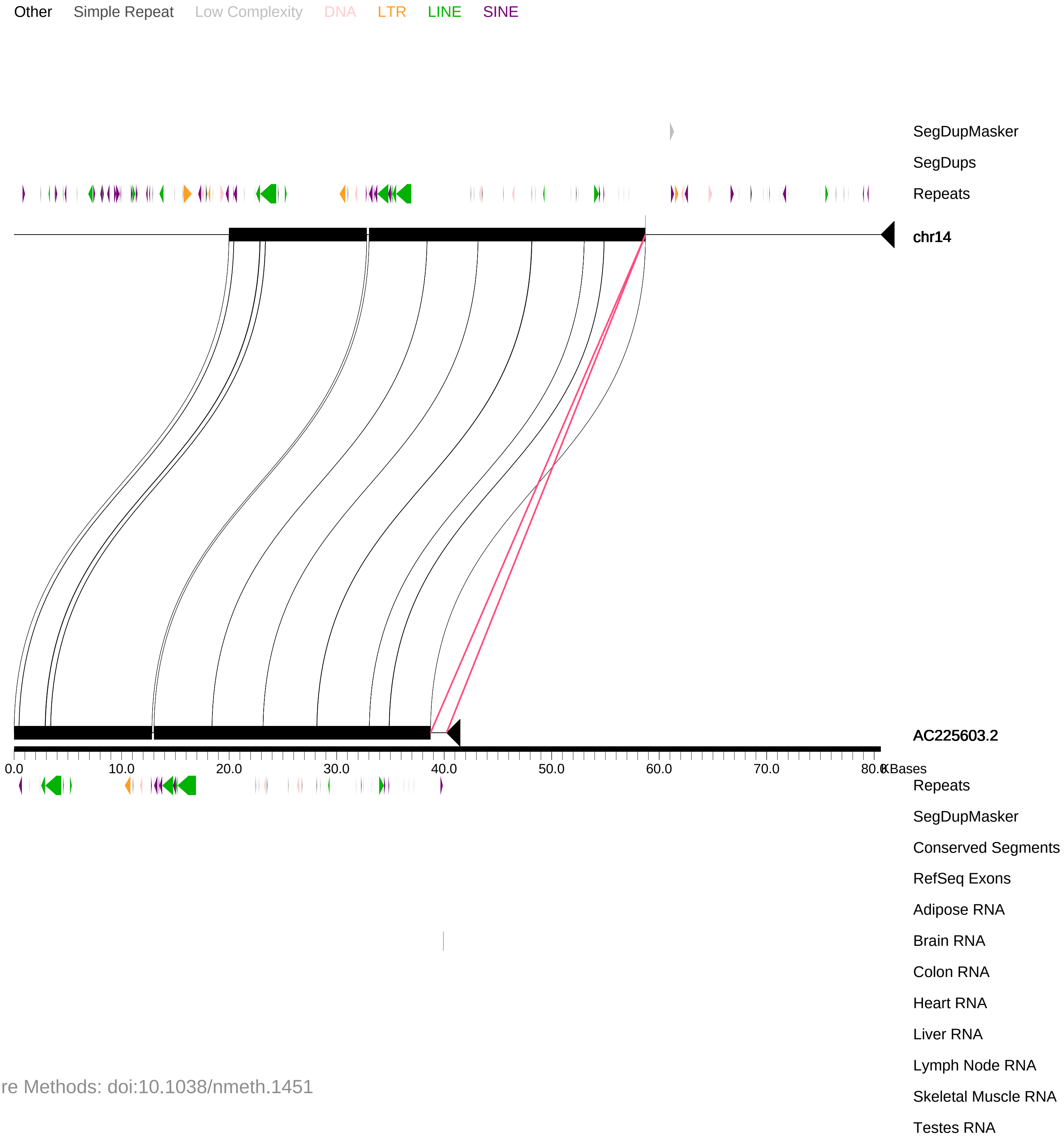
Insertion Size: 4375



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC225603.rc.fa

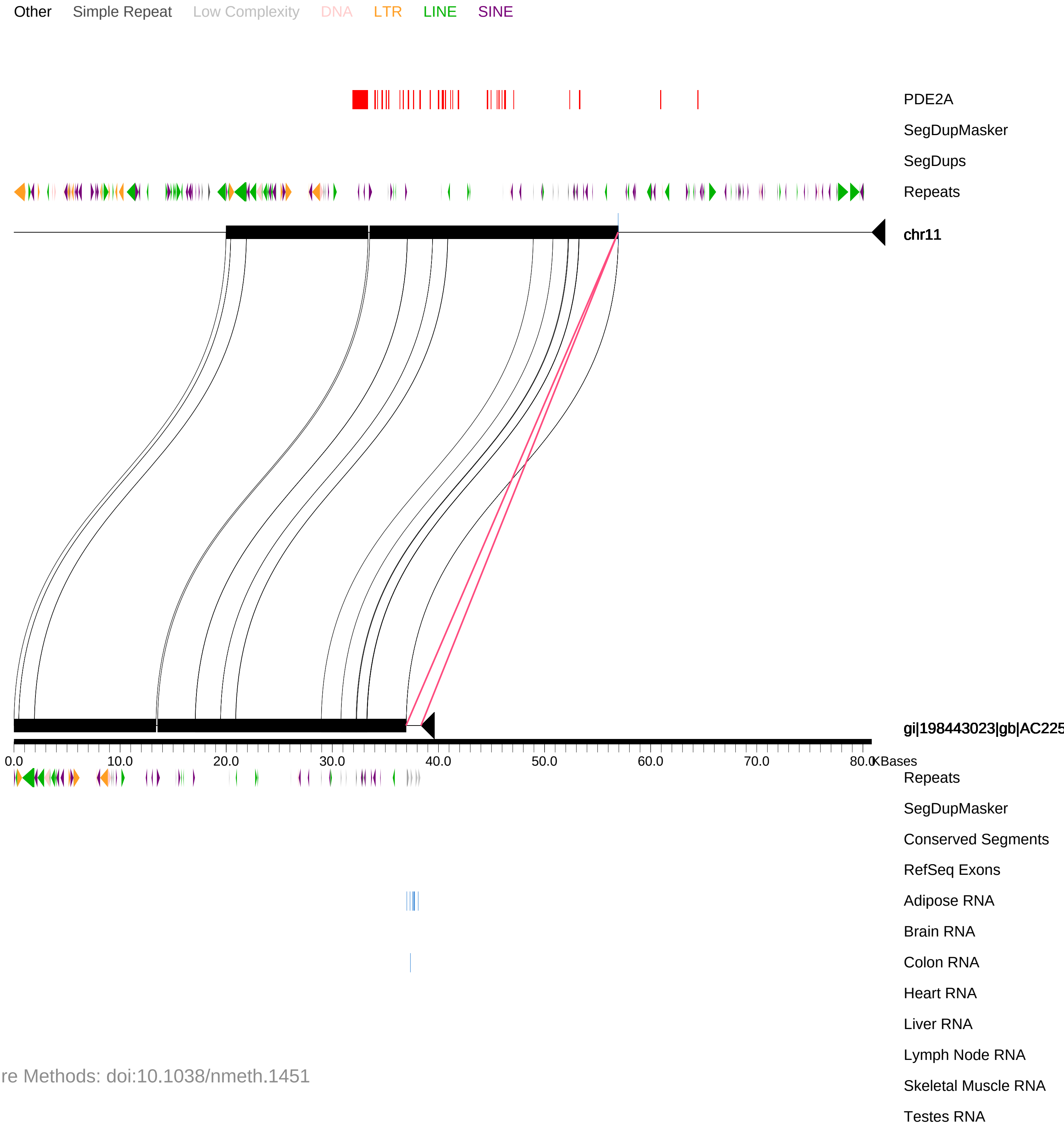
Insertion Size: 1490



Nature Methods: doi:10.1038/nmeth.1451

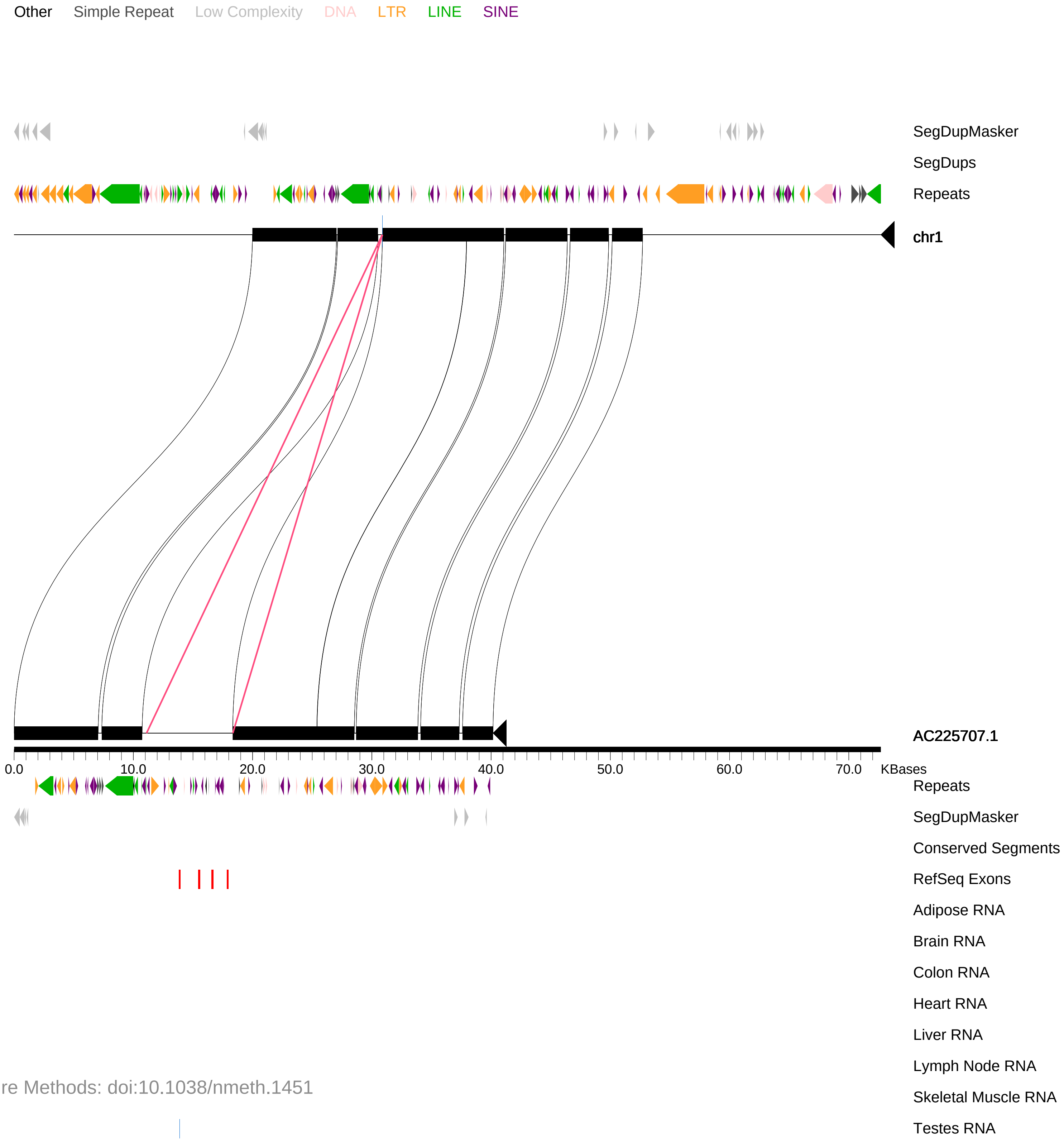
Clone file = AC225617.fa

Insertion Size: 1413



Clone file = AC225707.rc.fa

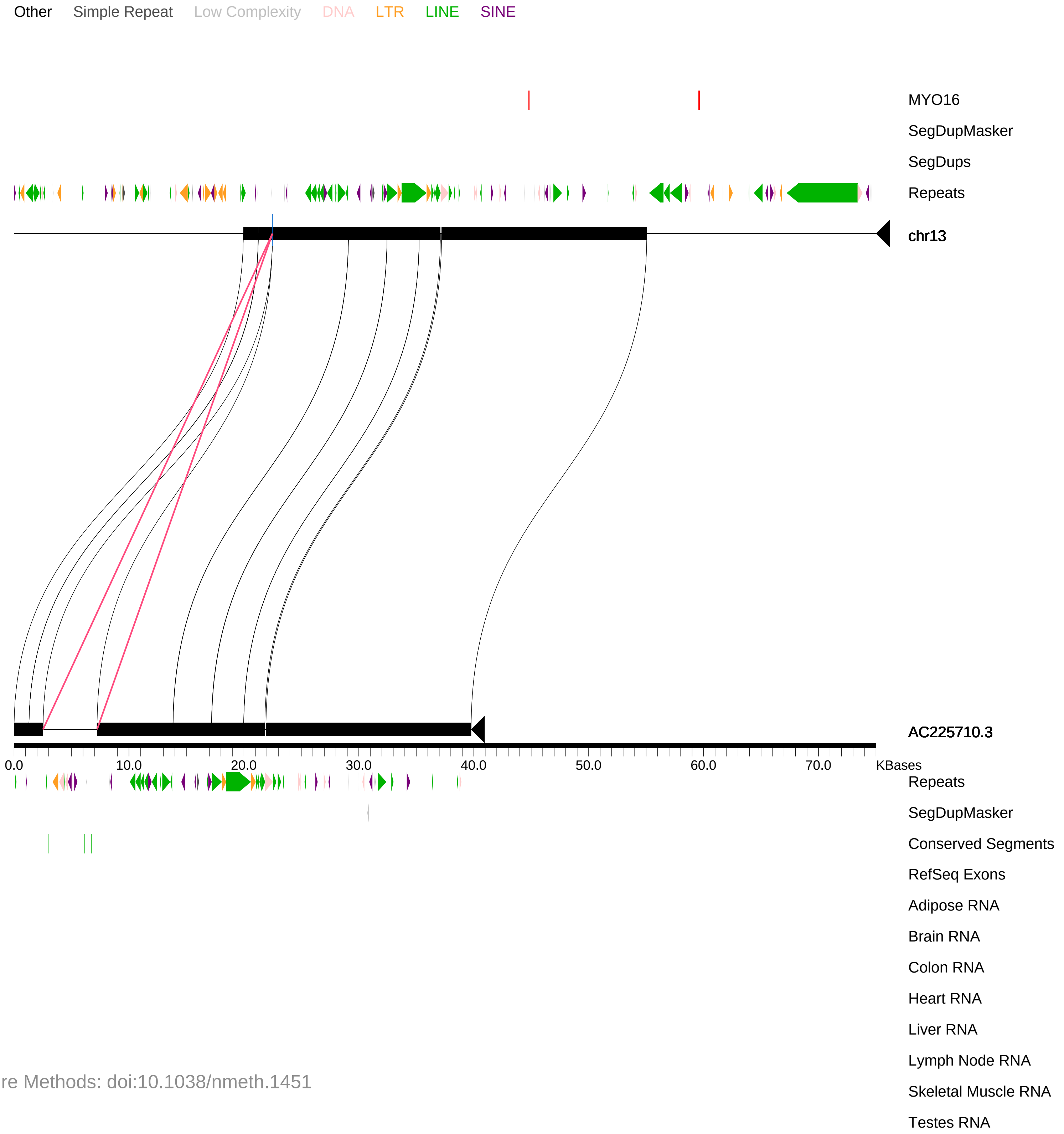
Insertion Size: 7234



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC225710.rc.fa

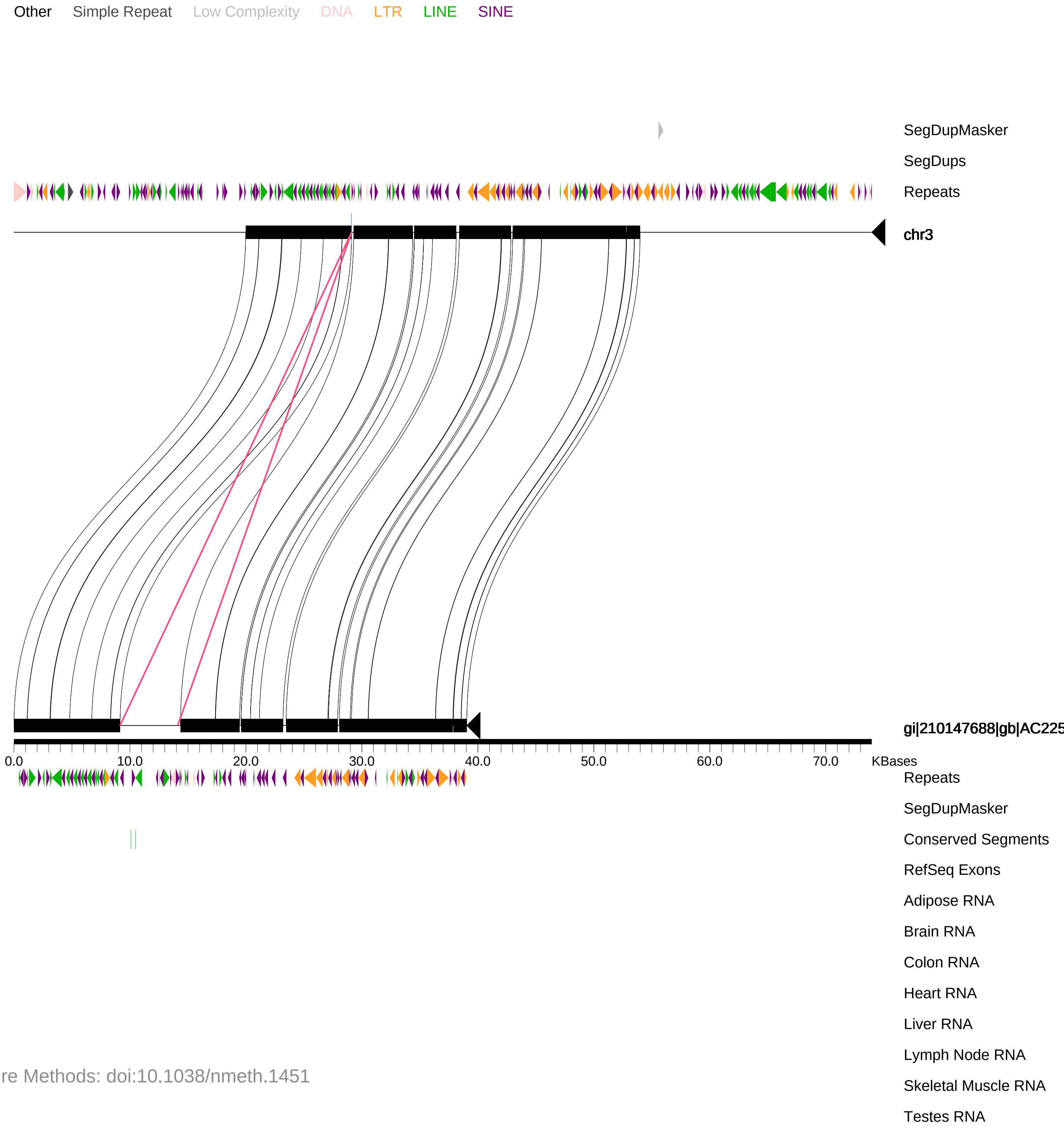
Insertion Size: 4705



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC225712.fa

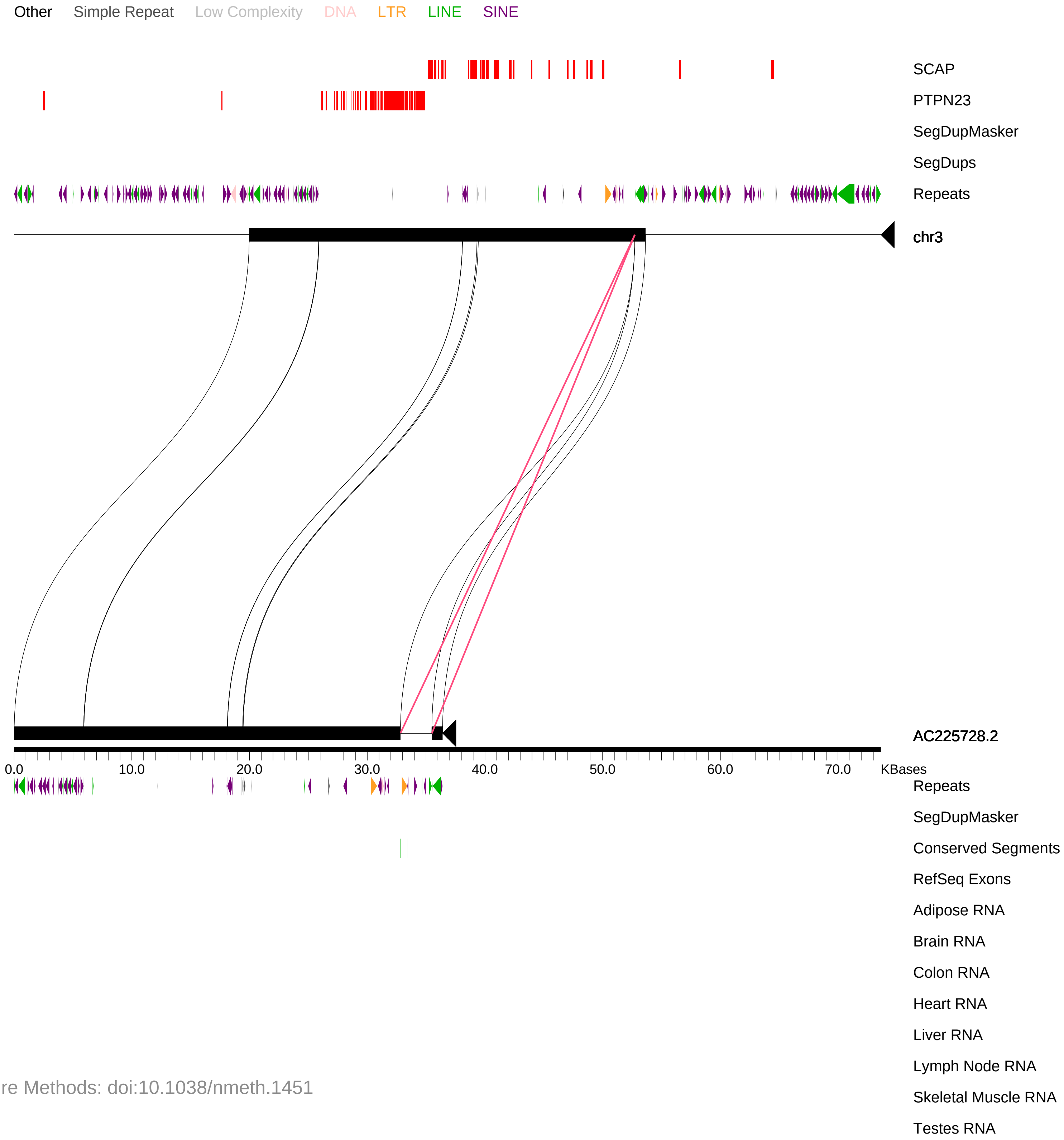
Insertion Size: 4975



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC225728.rc.fa

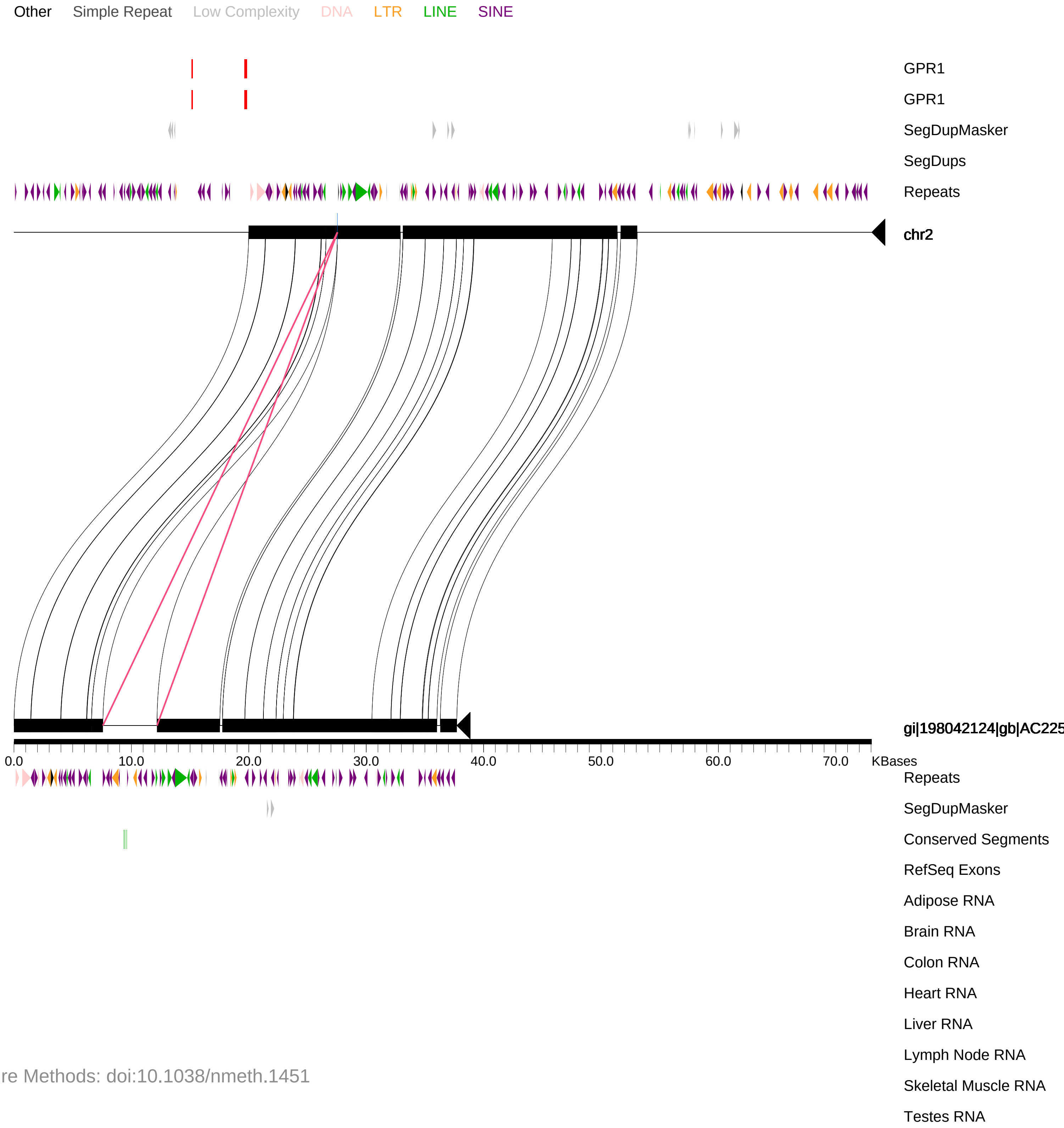
Insertion Size: 2674



Nature Methods: doi:10.1038/nmeth.1451

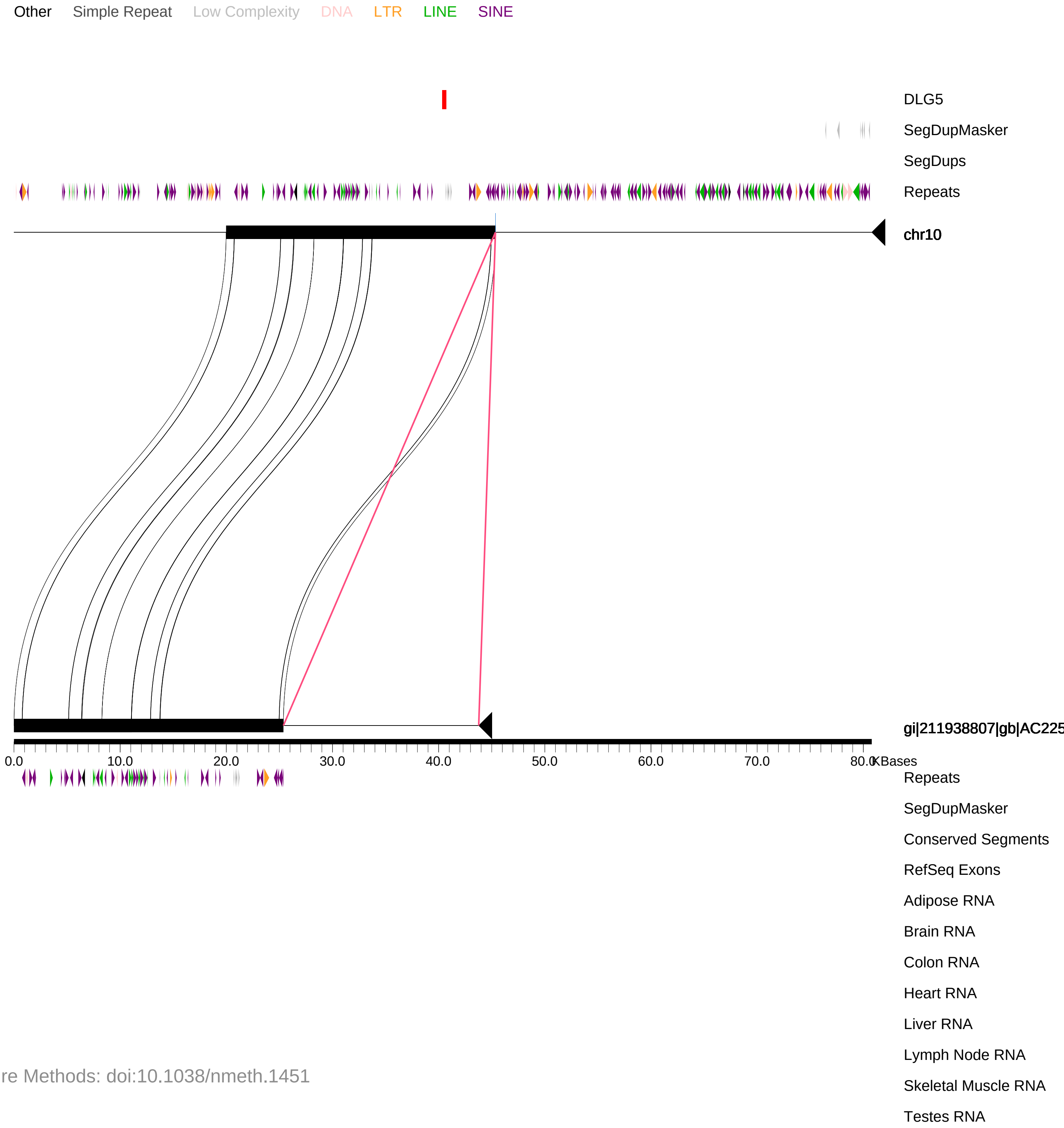
Clone file = AC225768.fa

Insertion Size: 4610



Clone file = AC225822.fa

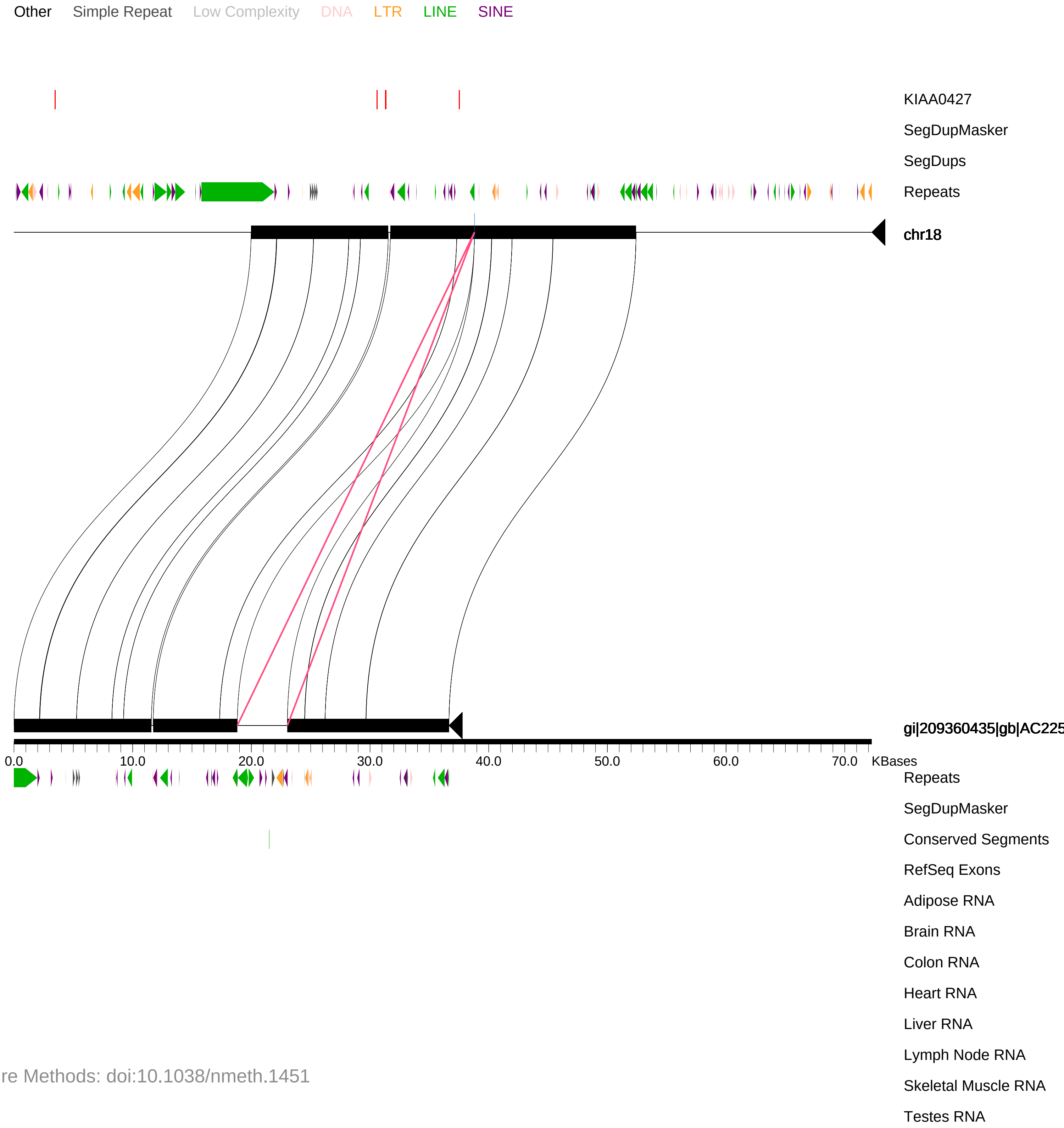
Insertion Size: 18398



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC225829.fa

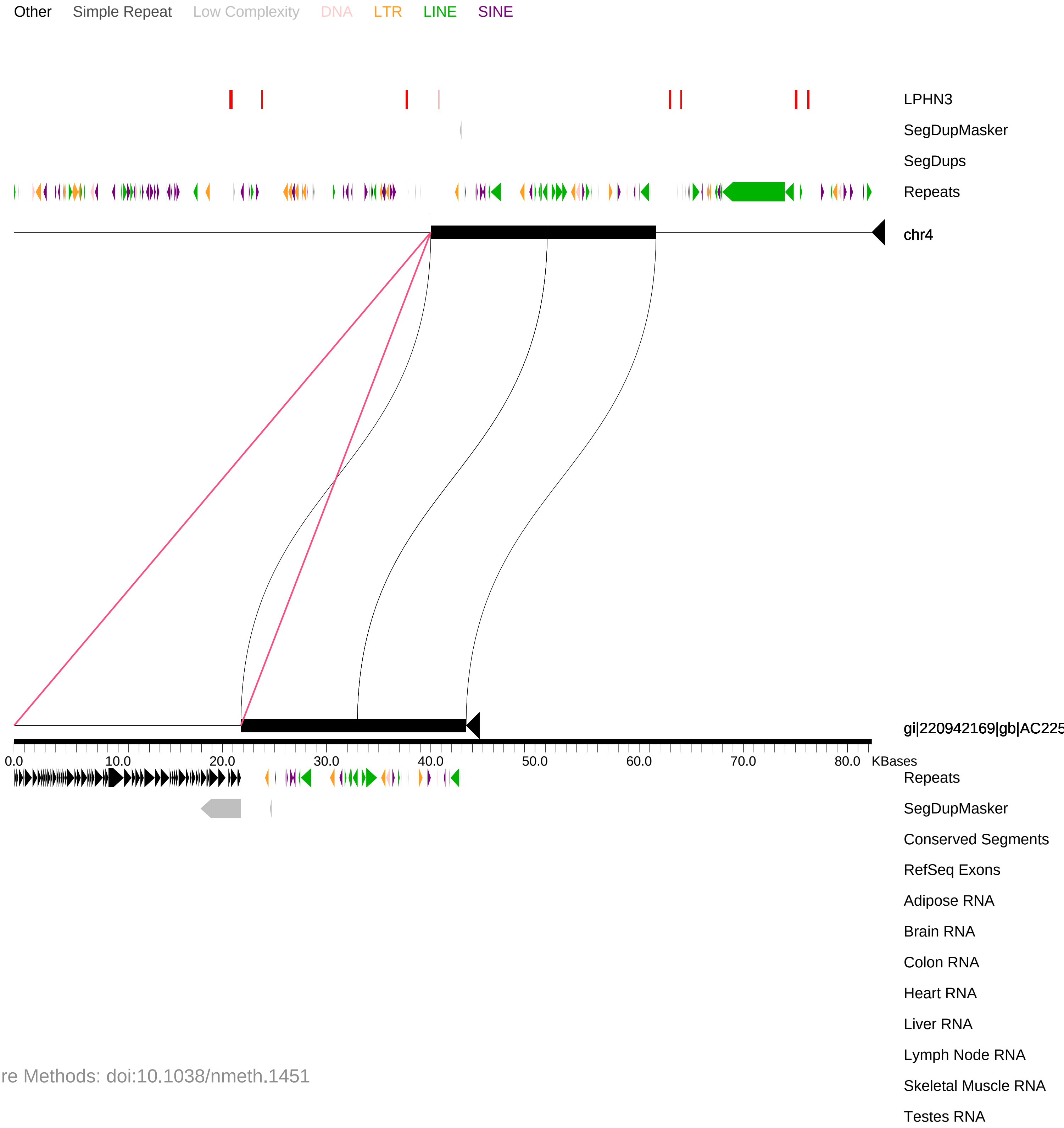
Insertion Size: 4225



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC225889.fa

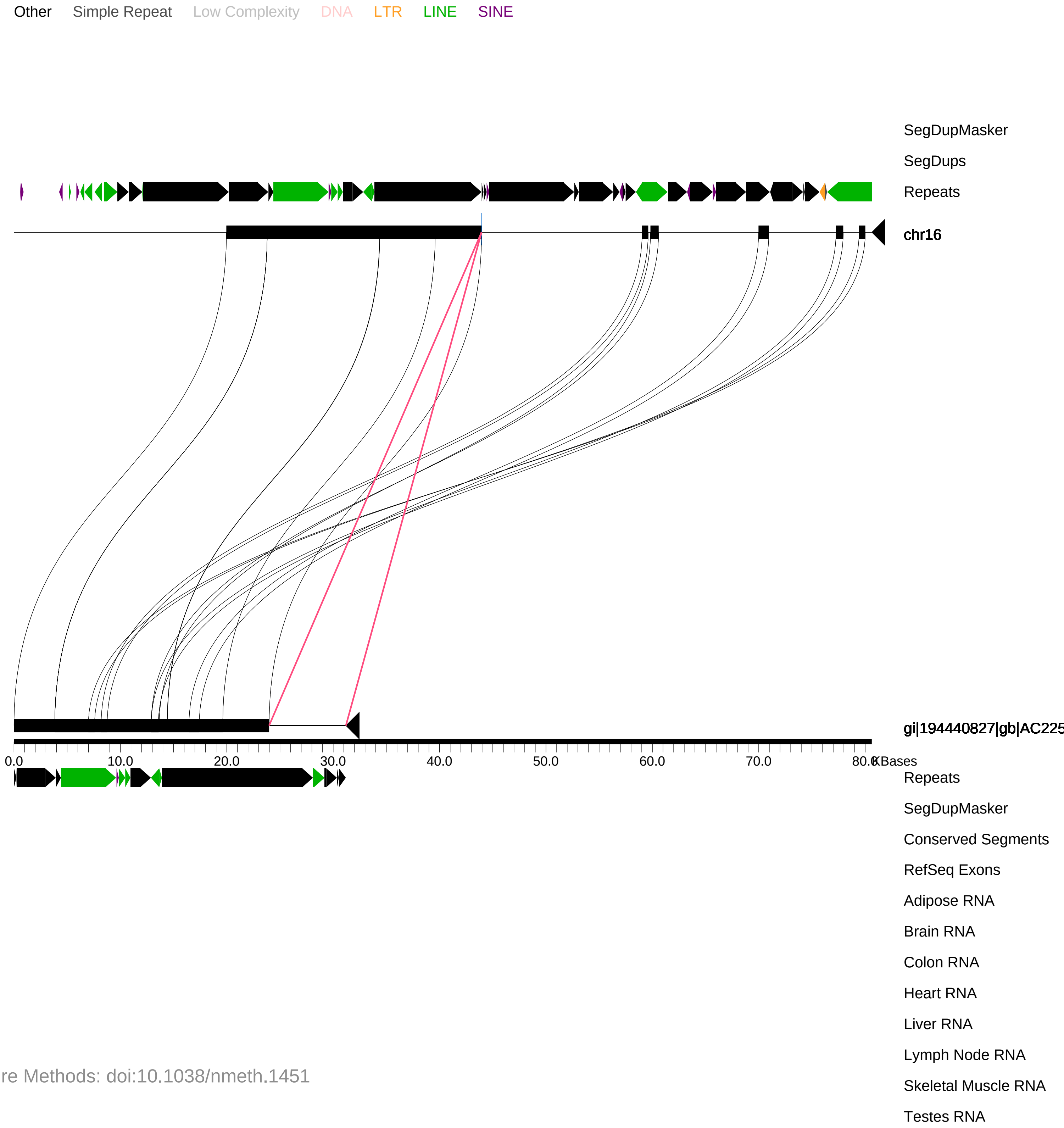
Insertion Size: 21777



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC225984.fa

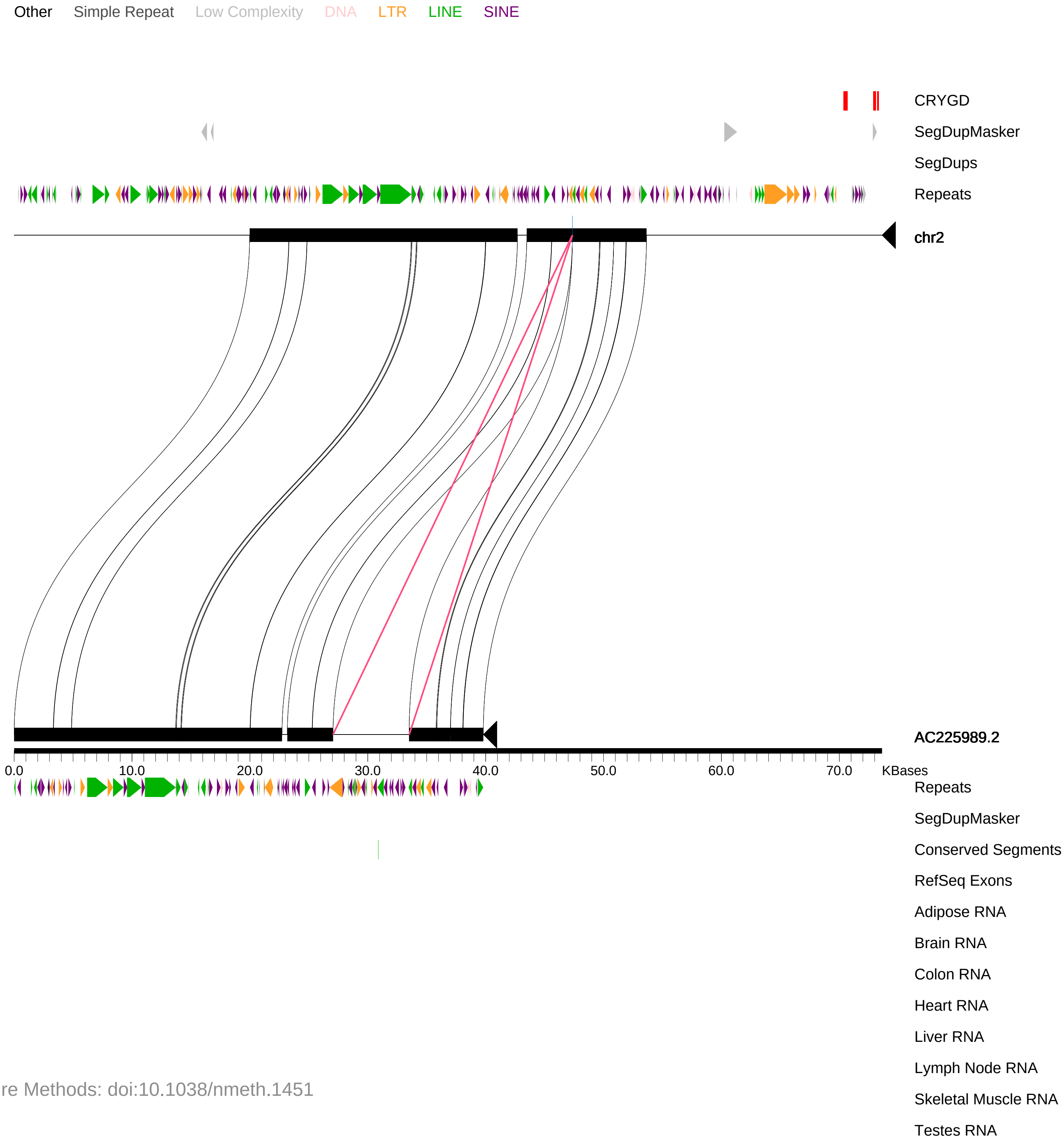
Insertion Size: 7220



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC225989.rc.fa

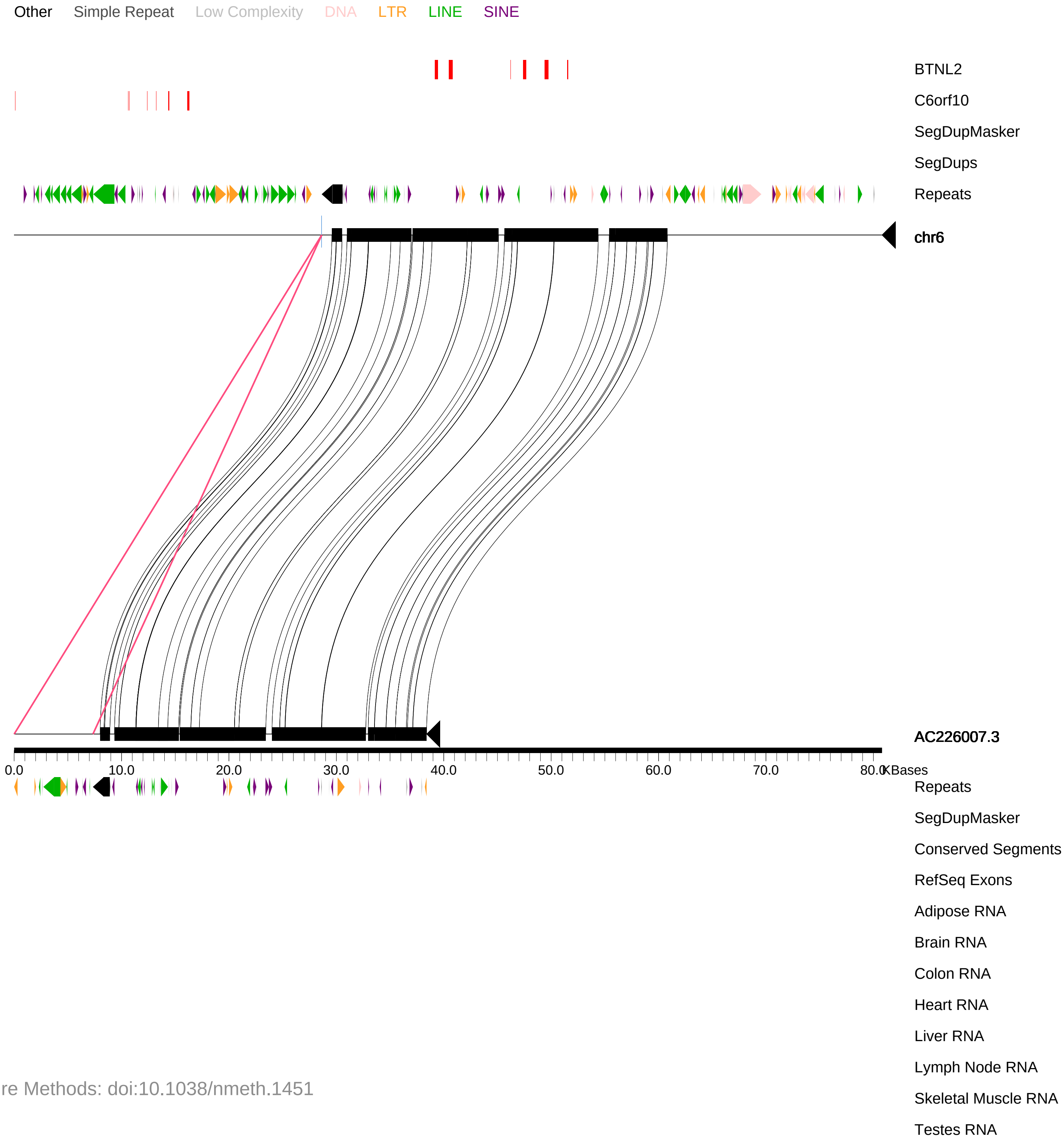
Insertion Size: 6465



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC226007.rc.fa

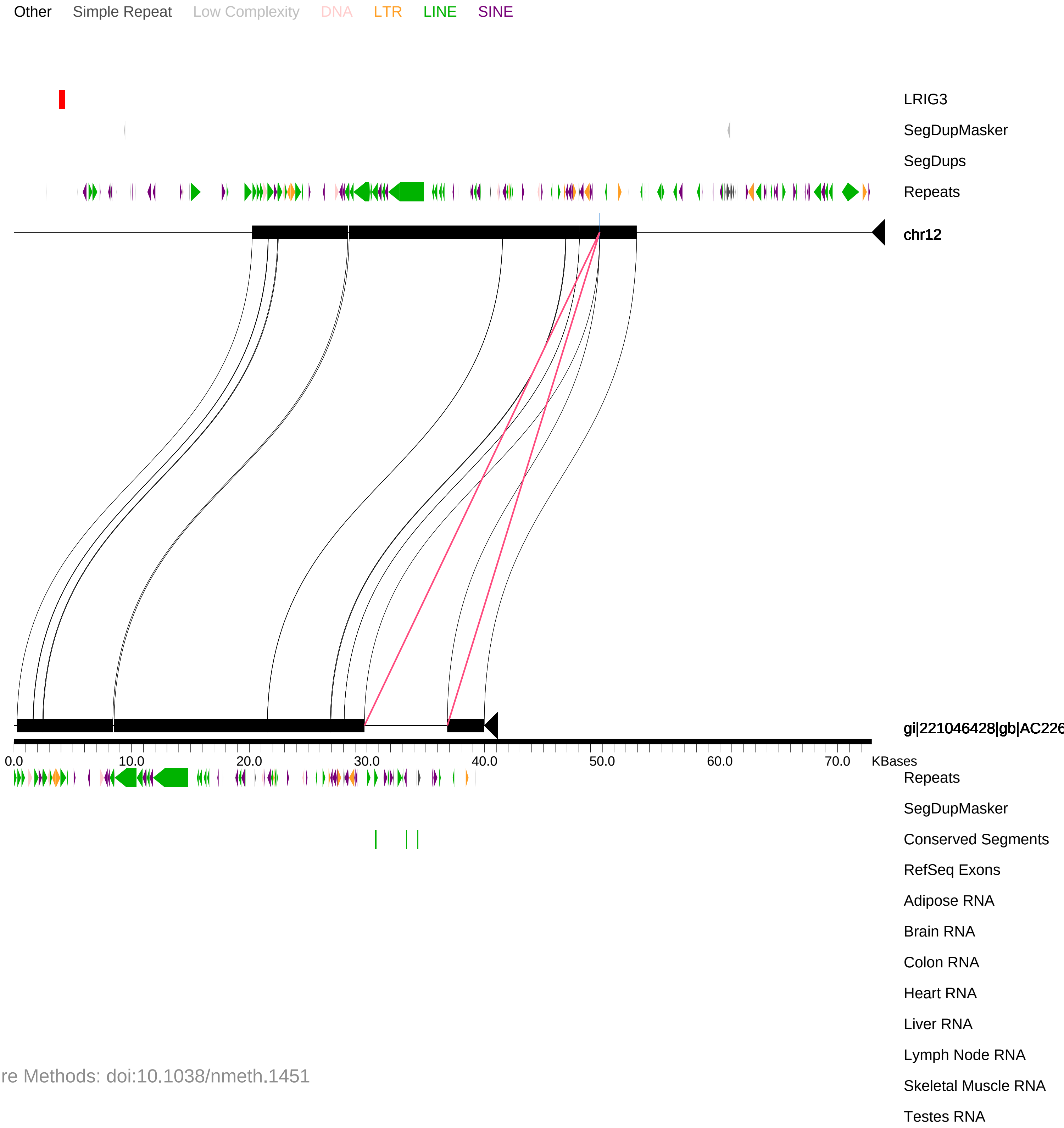
Insertion Size: 7338



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC226108.fa

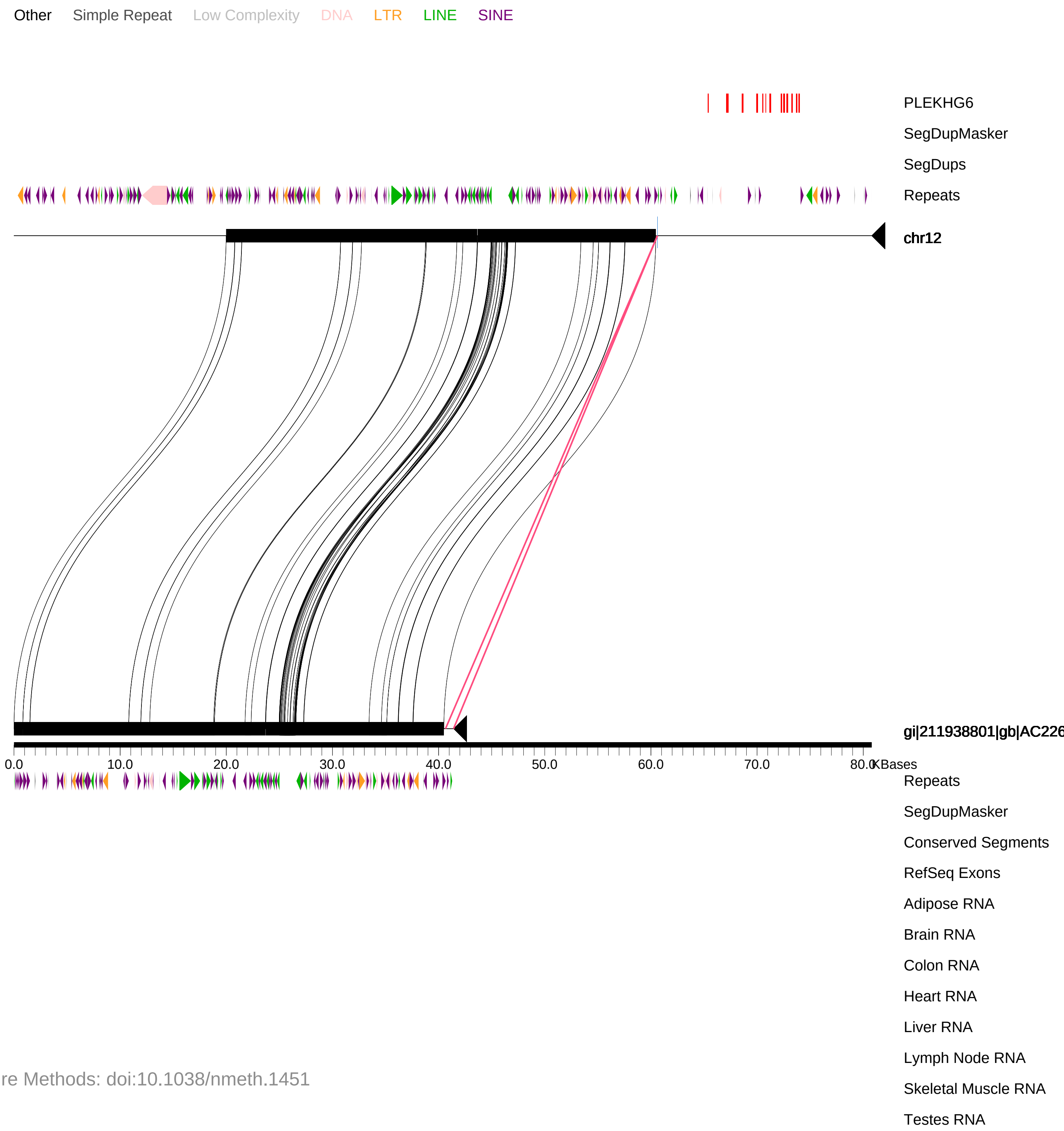
Insertion Size: 7045



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC226116.fa

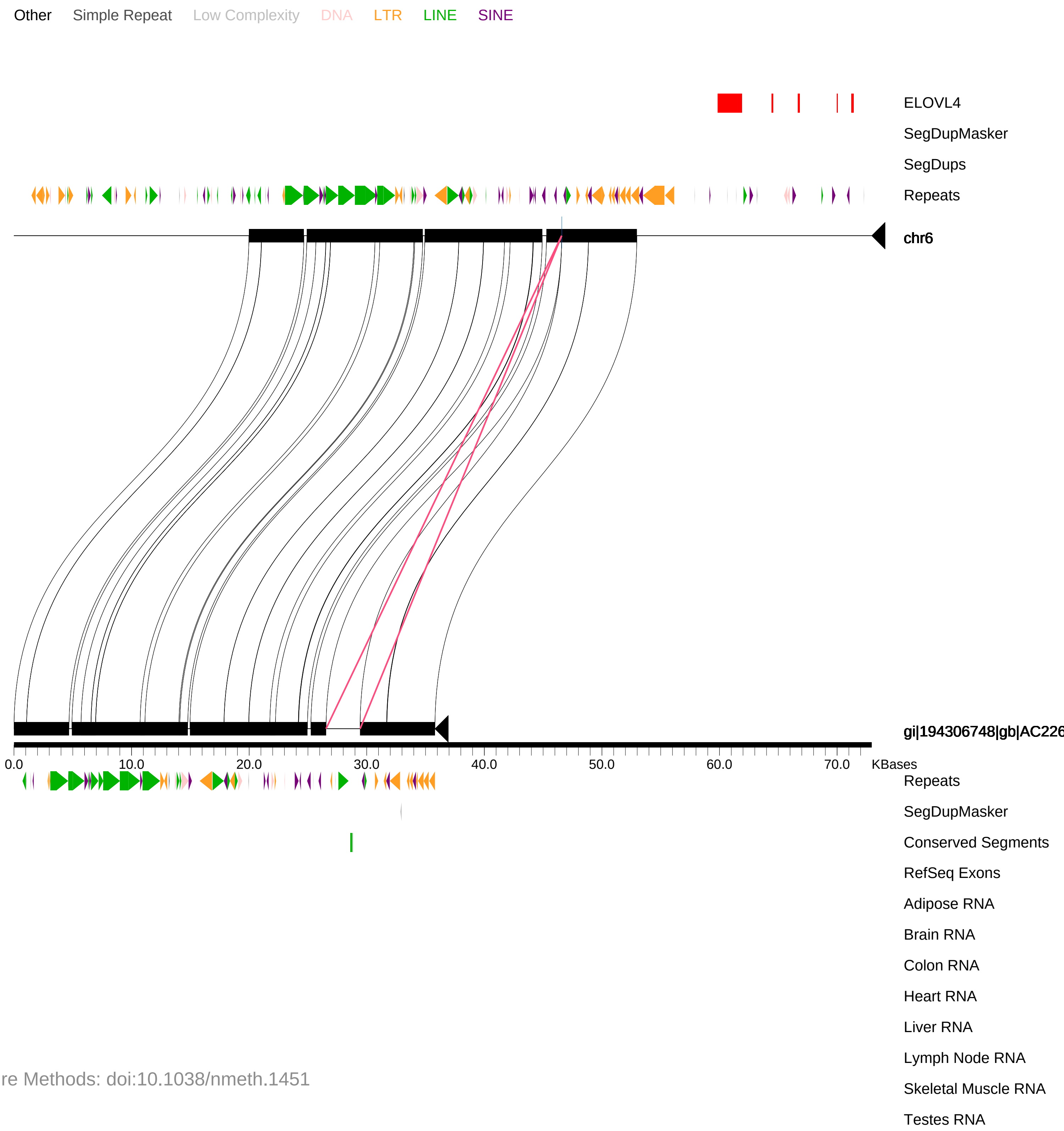
Insertion Size: 758



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC226139.fa

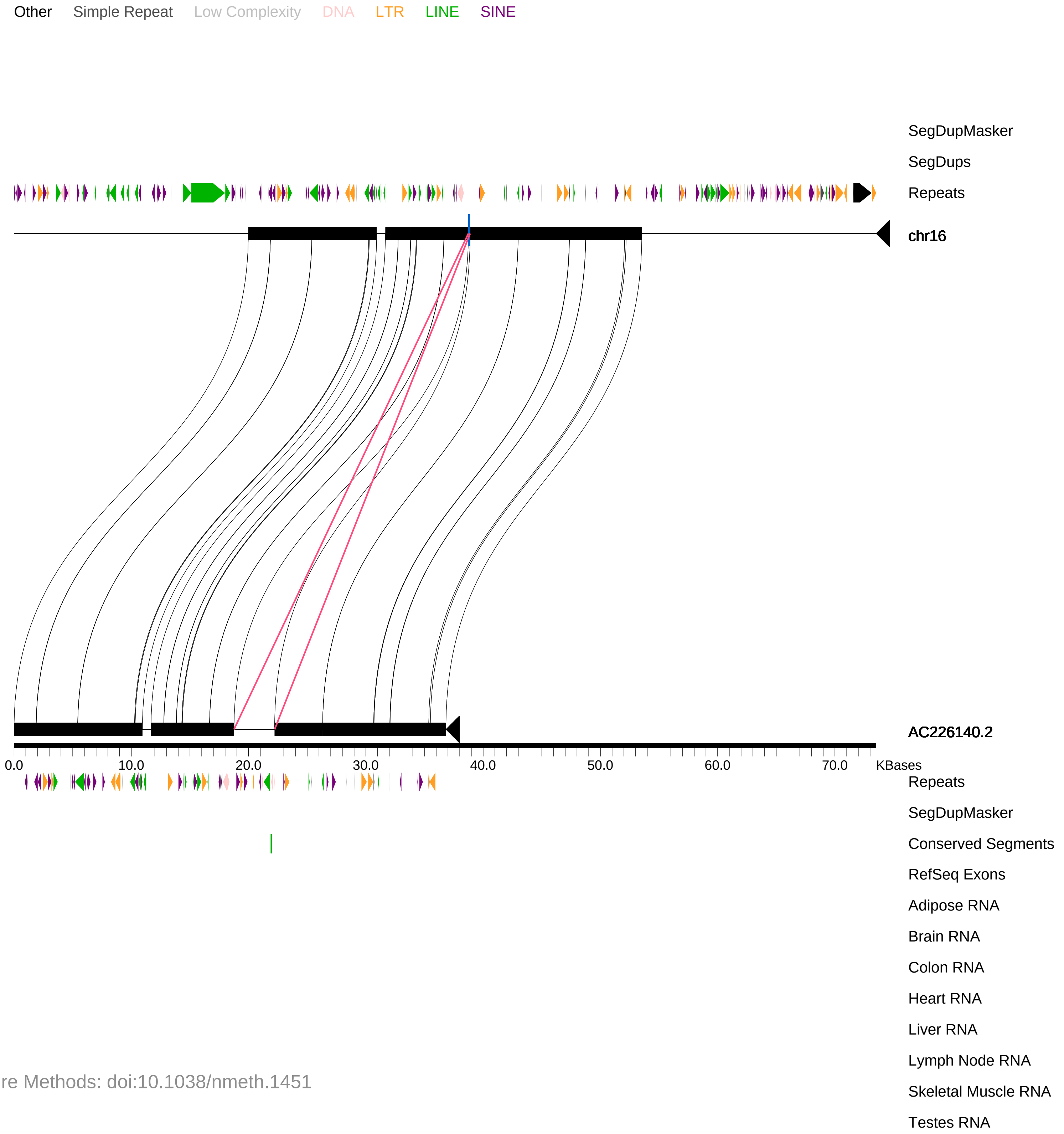
Insertion Size: 2898



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC226140.rc.fa

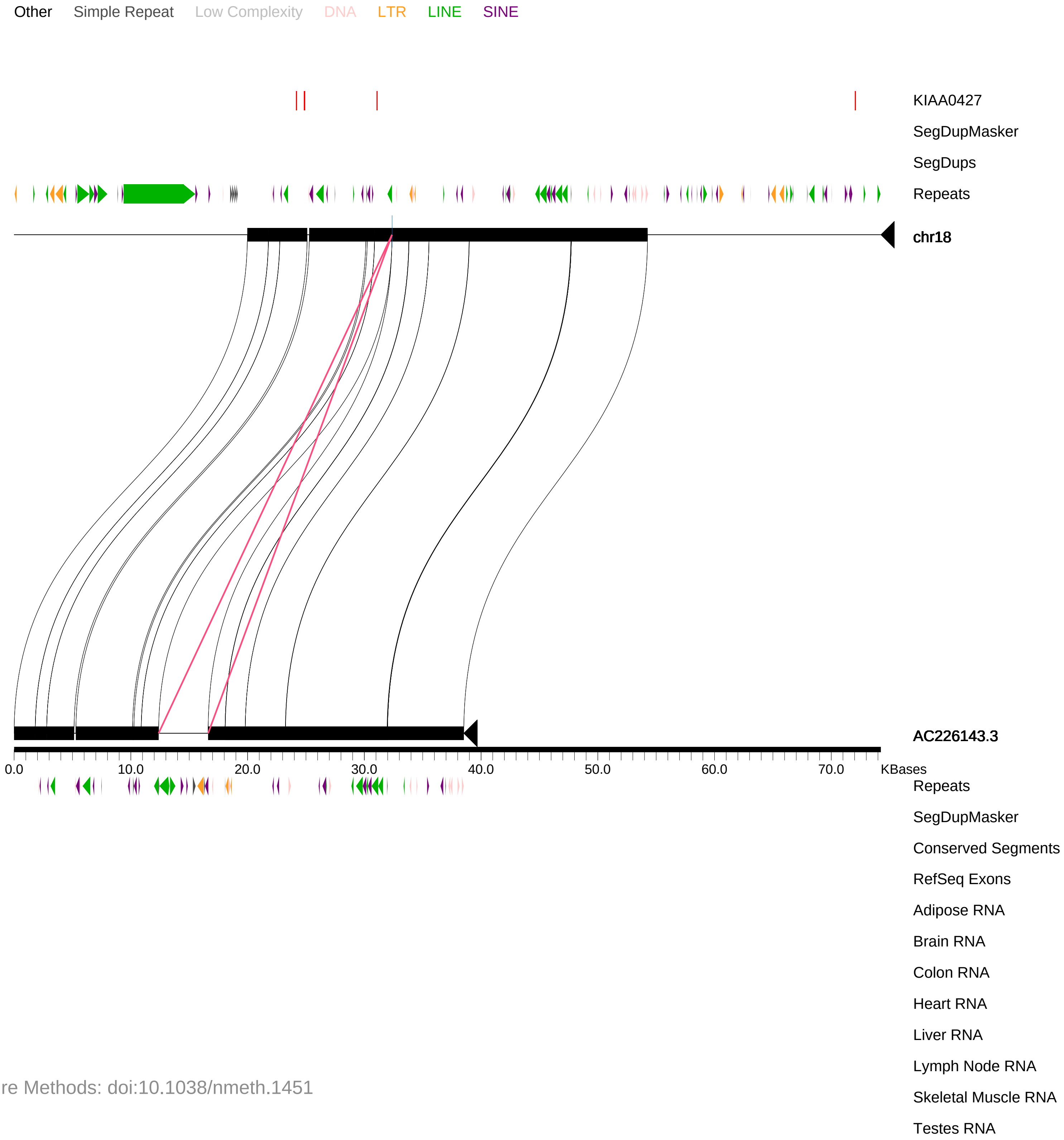
Insertion Size: 3464



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC226143.rc.fa

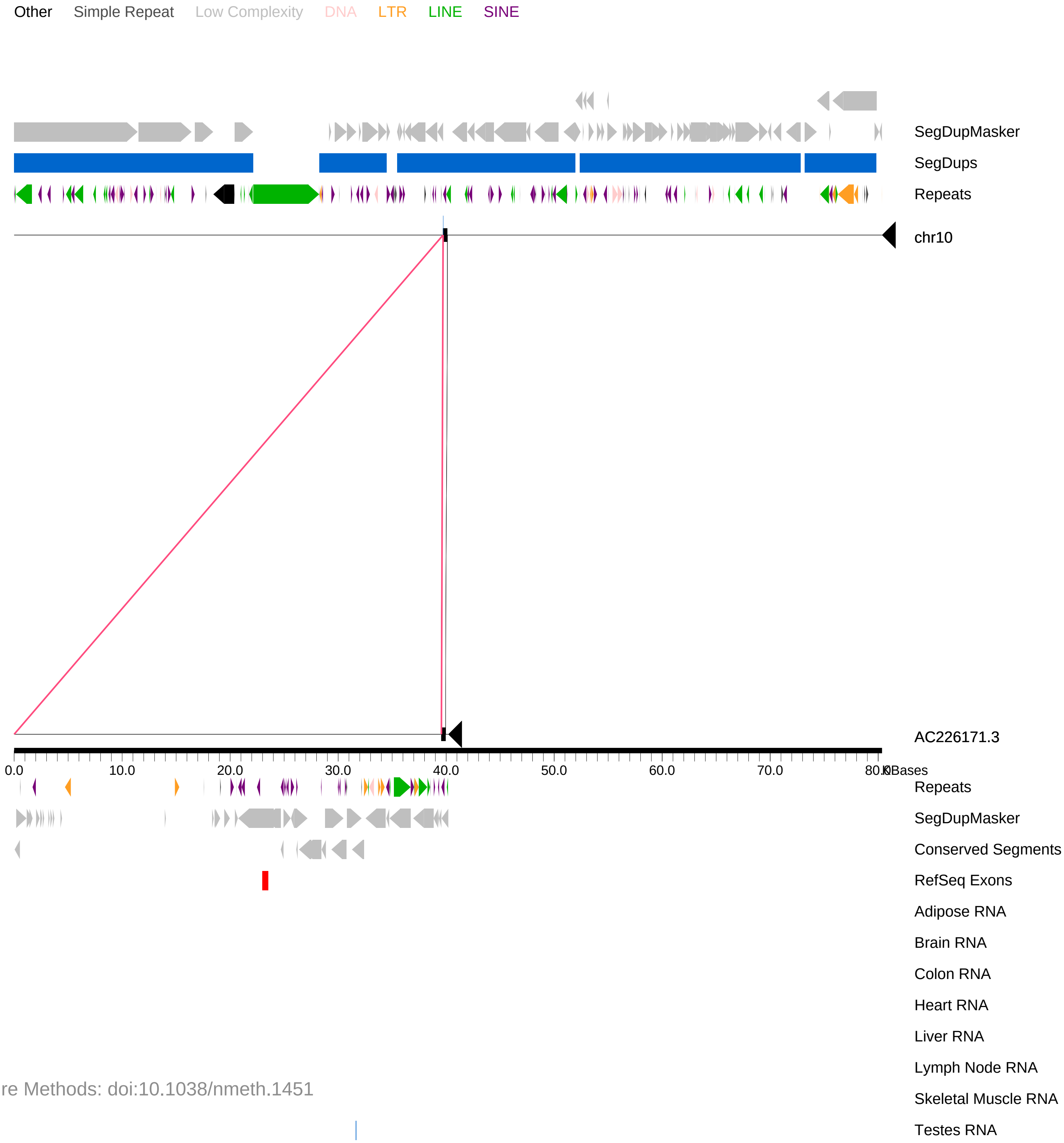
Insertion Size: 4247



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC226171.rc.fa

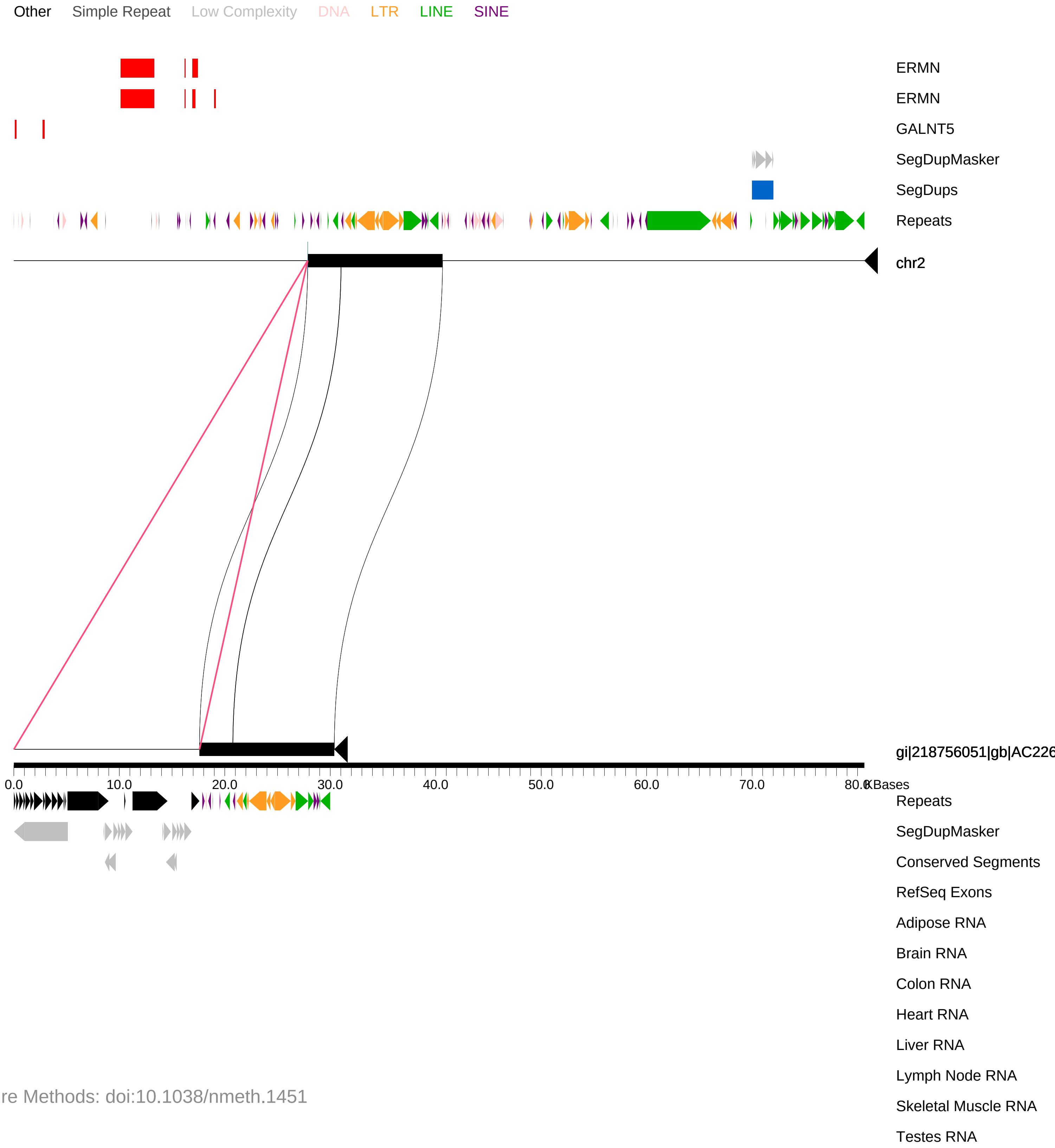
Insertion Size: 39559



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC226495.fa

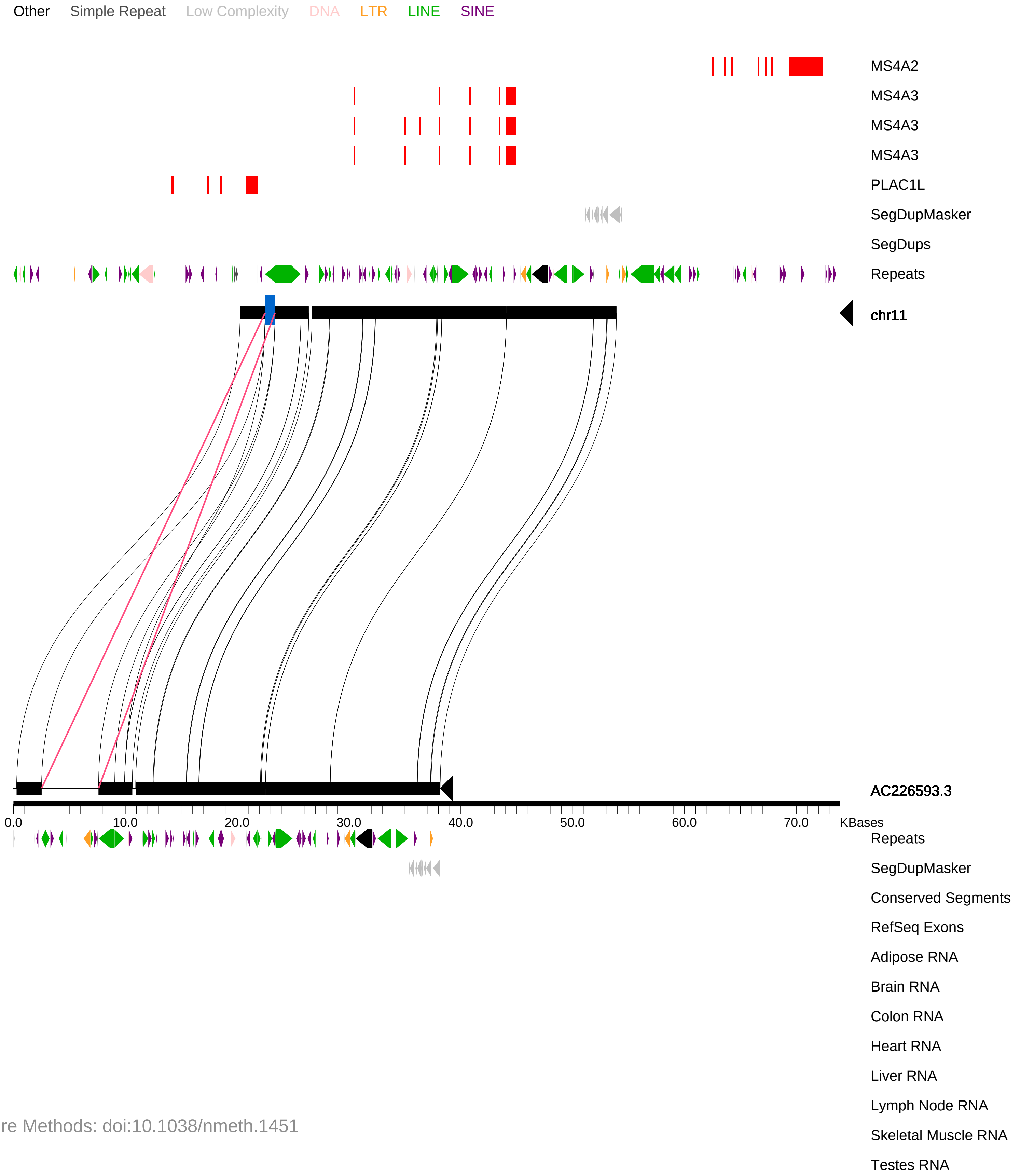
Insertion Size: 17600



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC226593.rc.fa

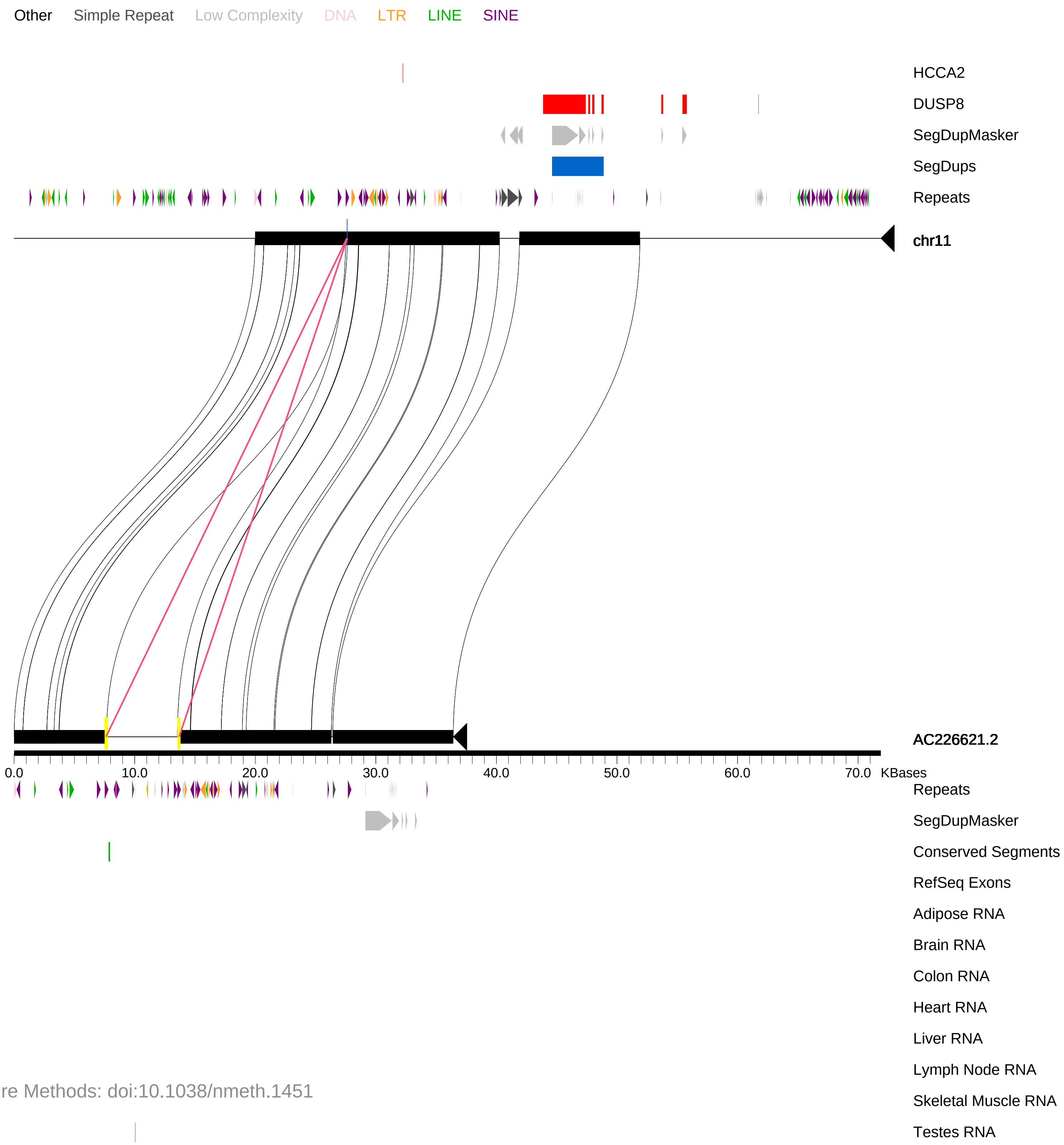
Insertion Size: 5095



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC226621.rc.fa

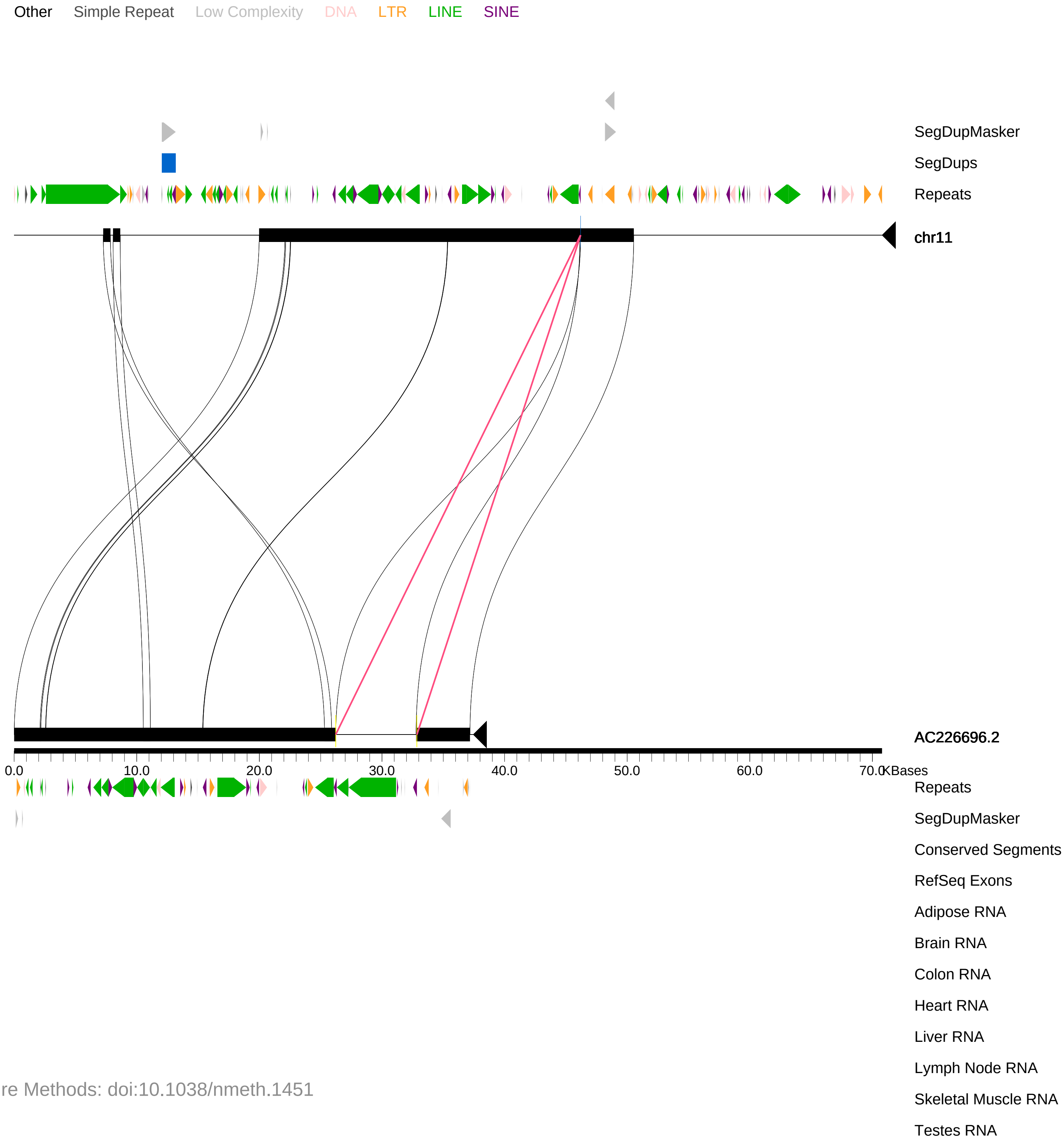
Insertion Size: 6066



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC226696.rc.fa

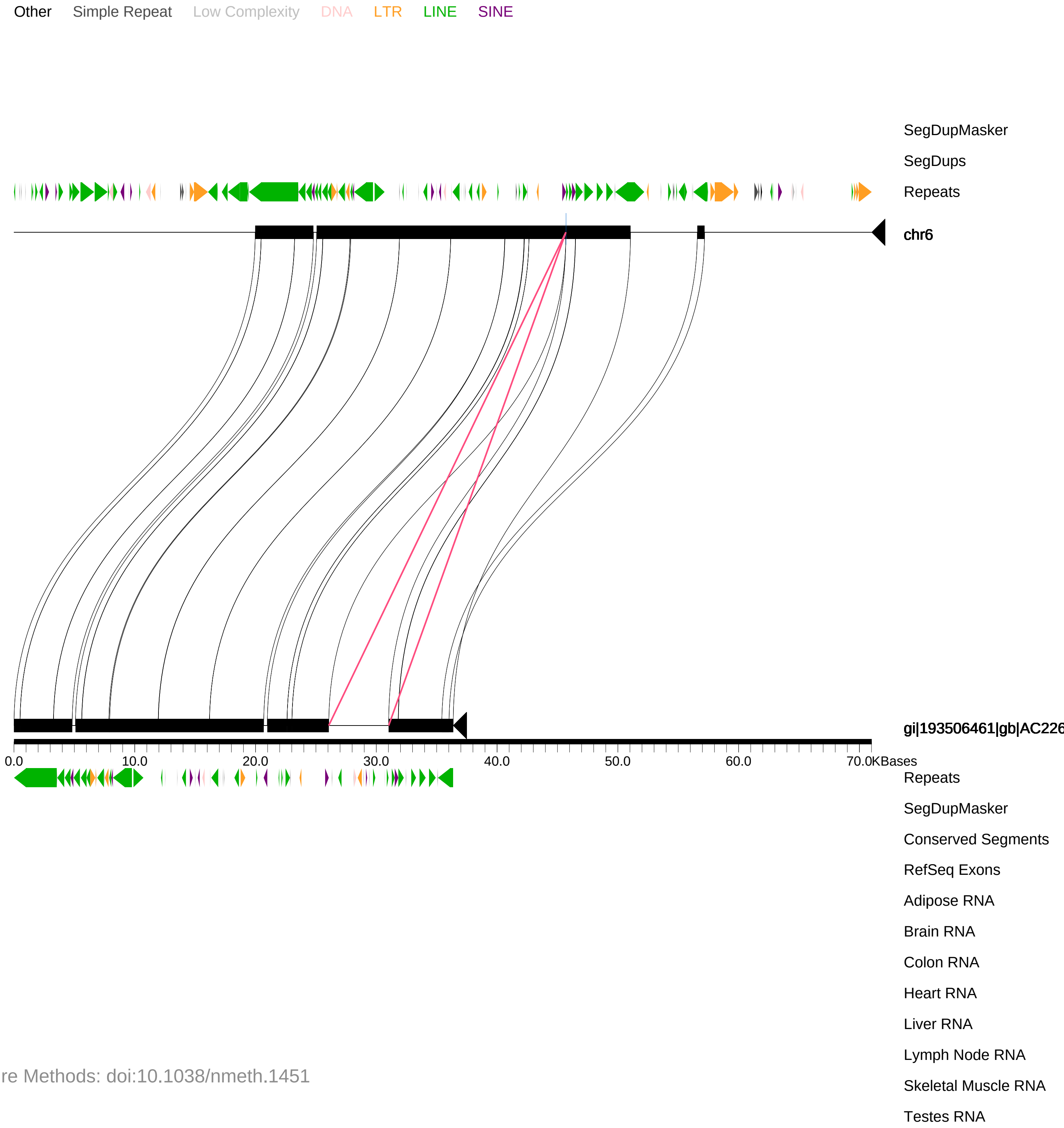
Insertion Size: 6624



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC226697.fa

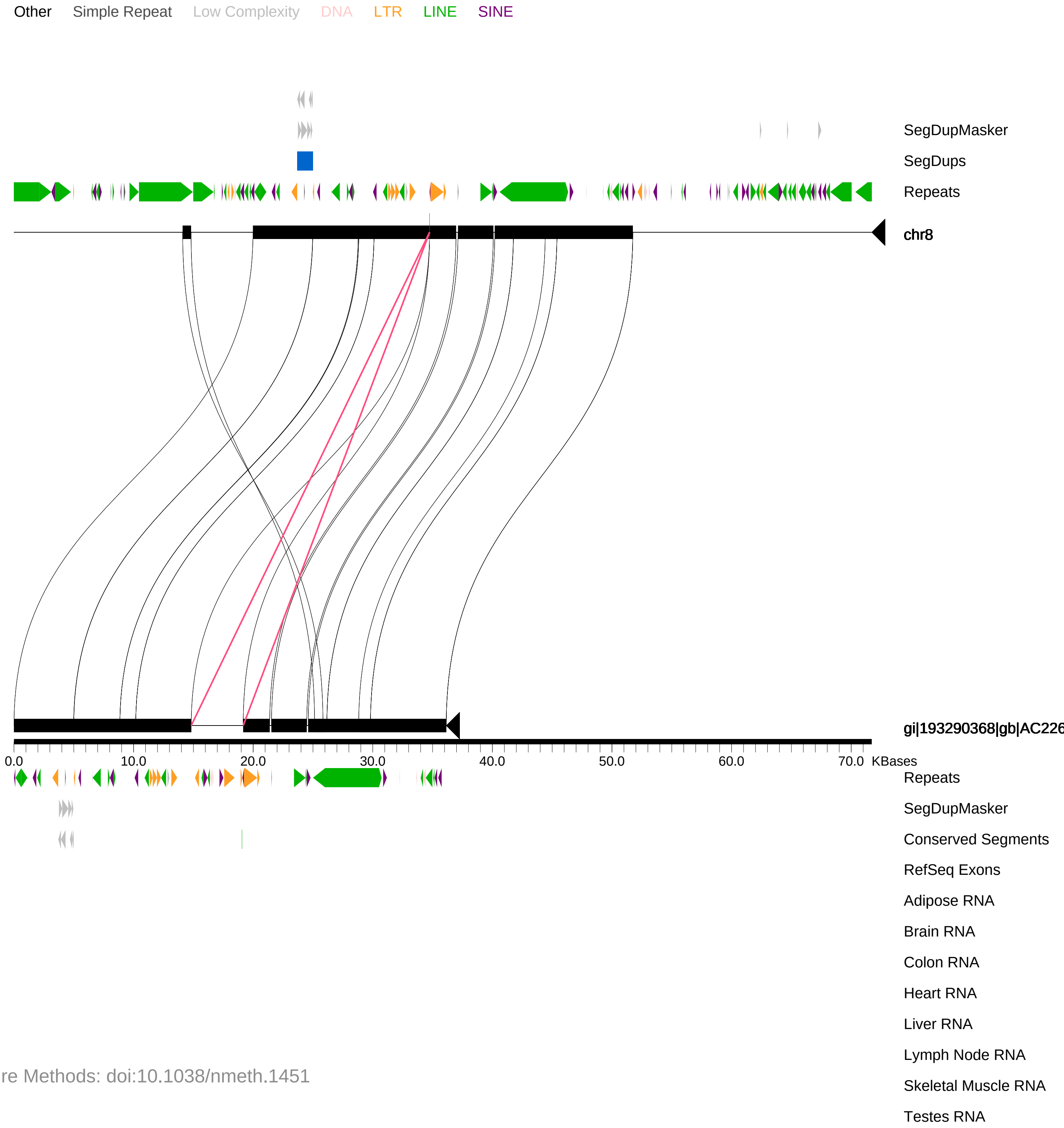
Insertion Size: 4963



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC226699.fa

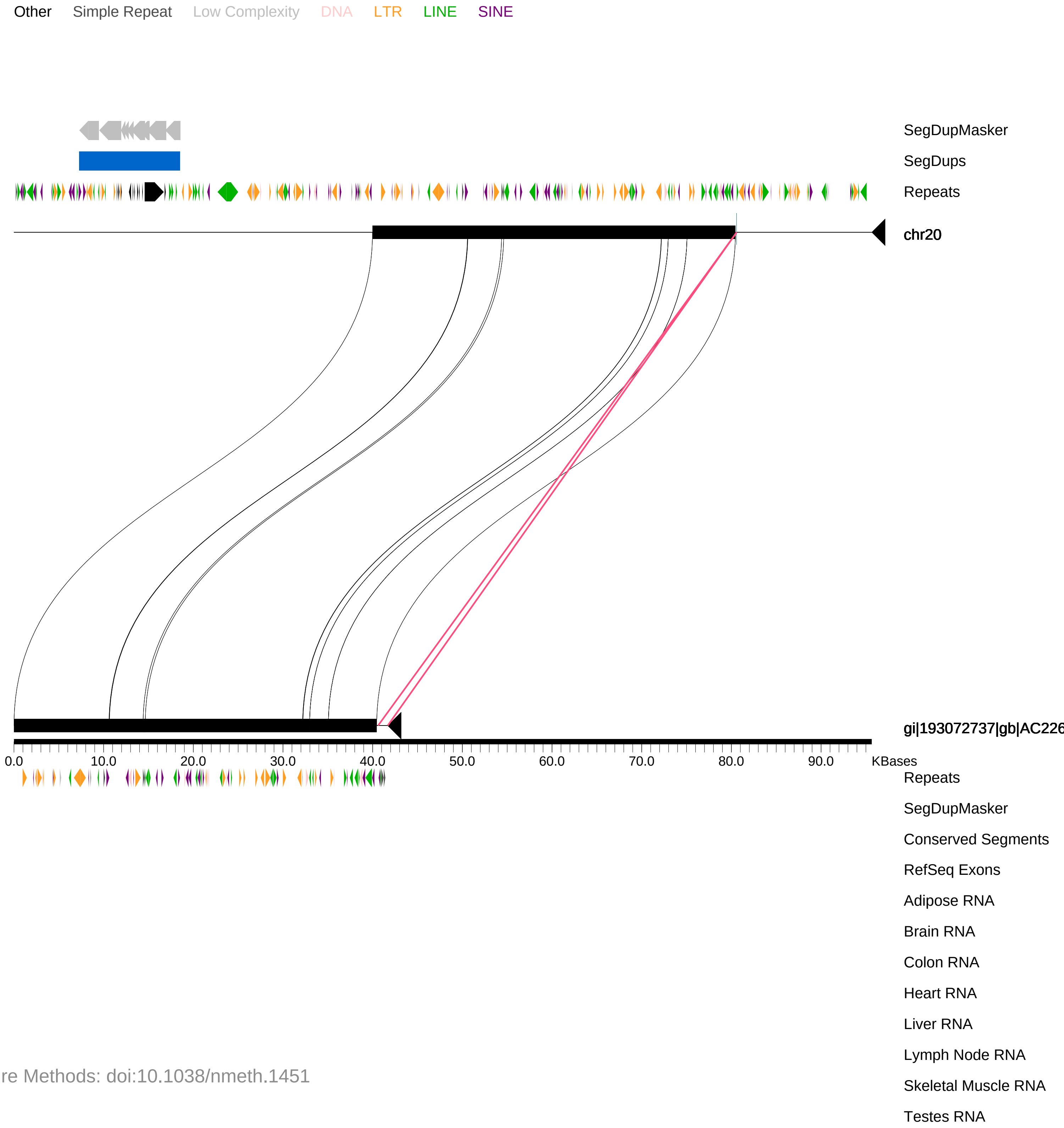
Insertion Size: 4349



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC226724.fa

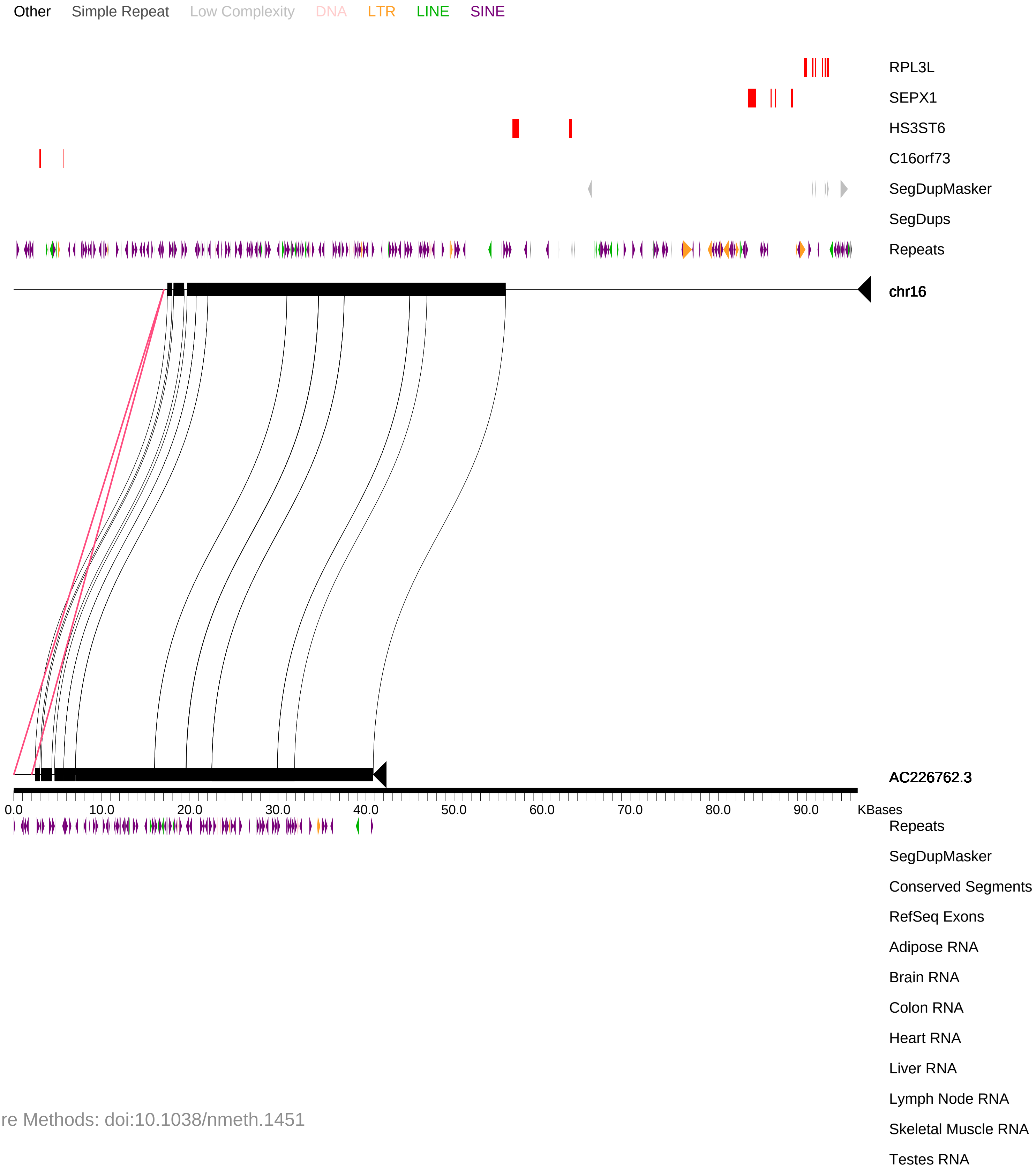
Insertion Size: 1077



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC226762.rc.fa

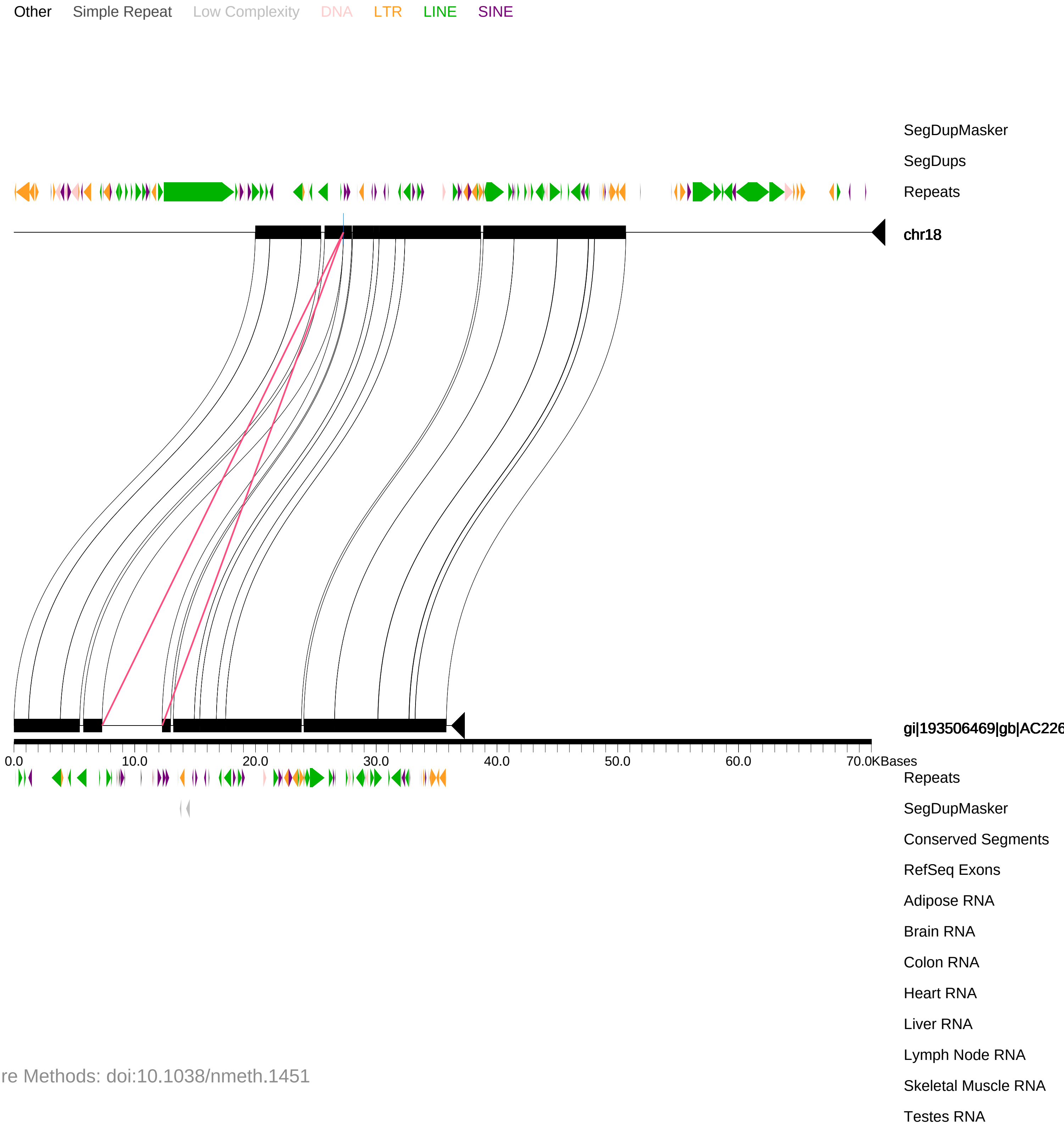
Insertion Size: 2033



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC226767.fa

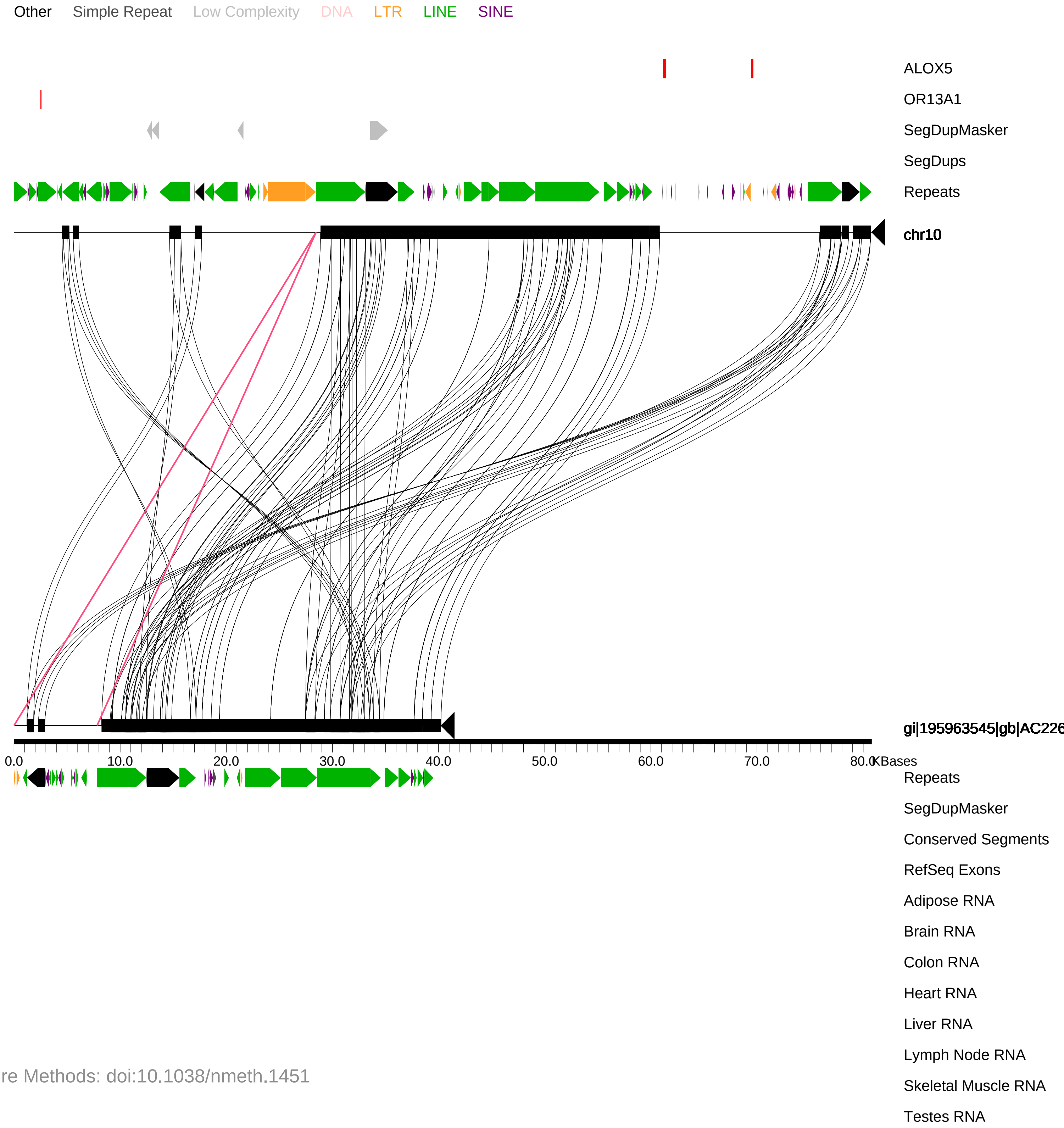
Insertion Size: 4981



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC226804.fa

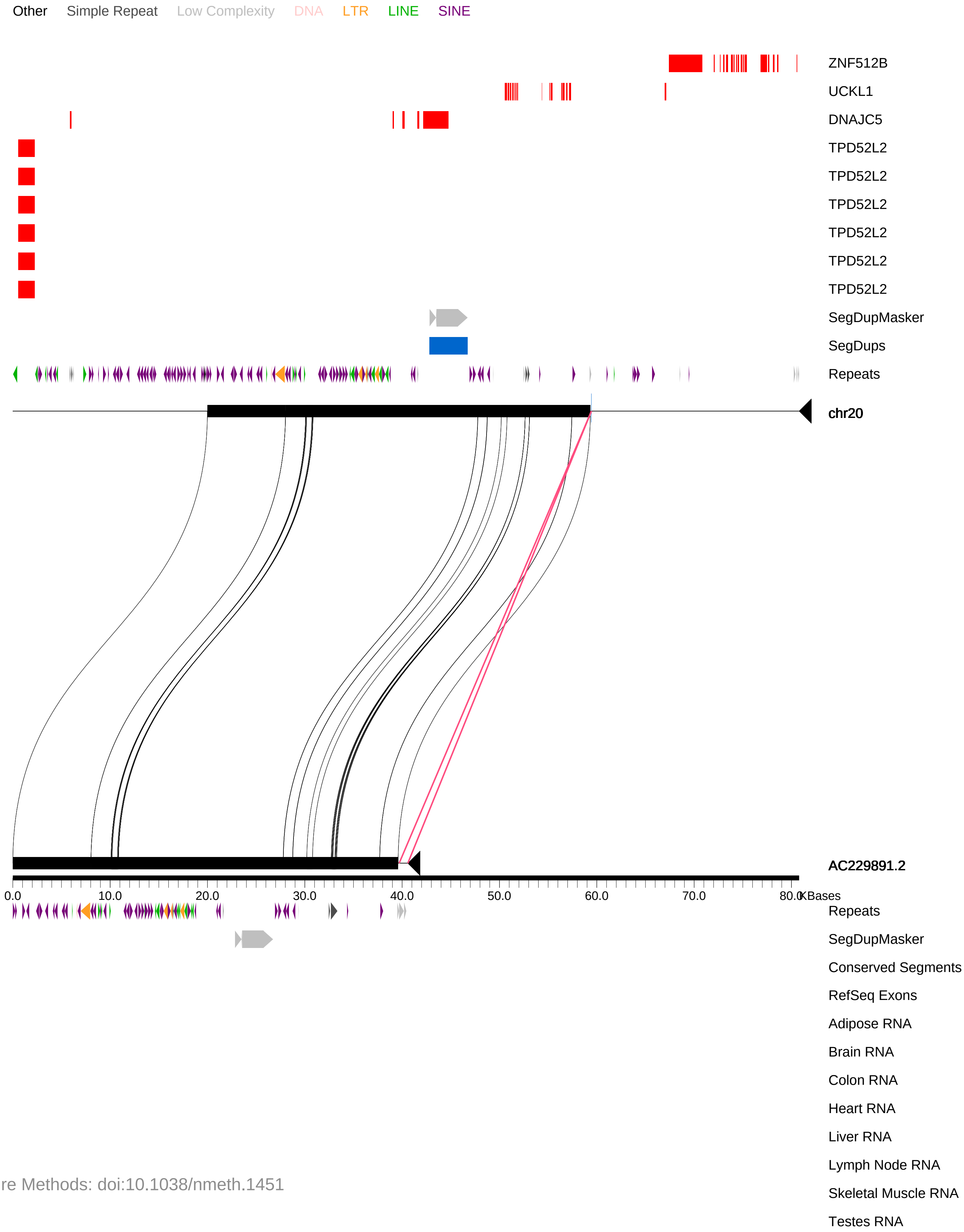
Insertion Size: 7832



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC229891.rc.fa

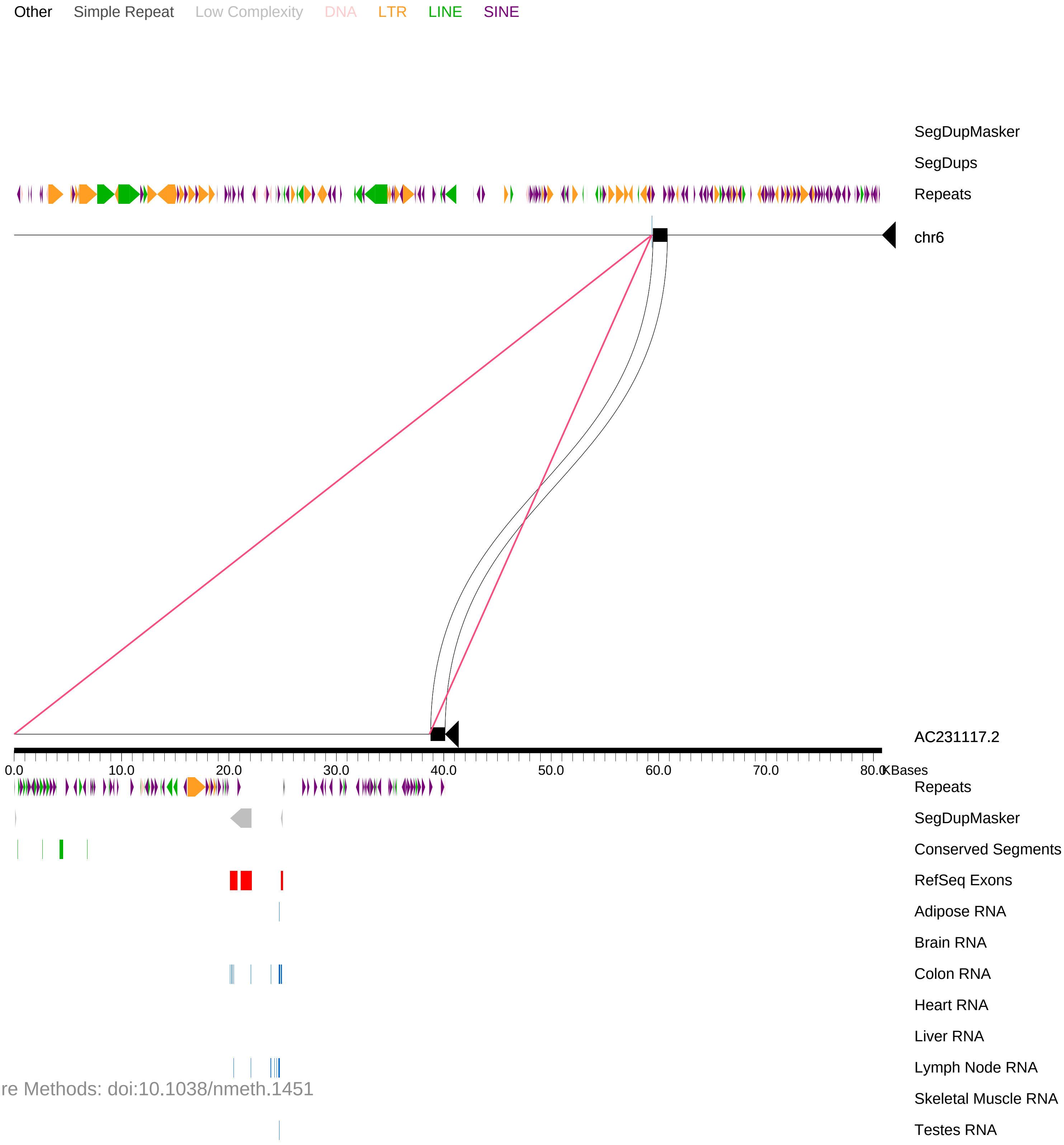
Insertion Size: 892



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC231117.rc.fa

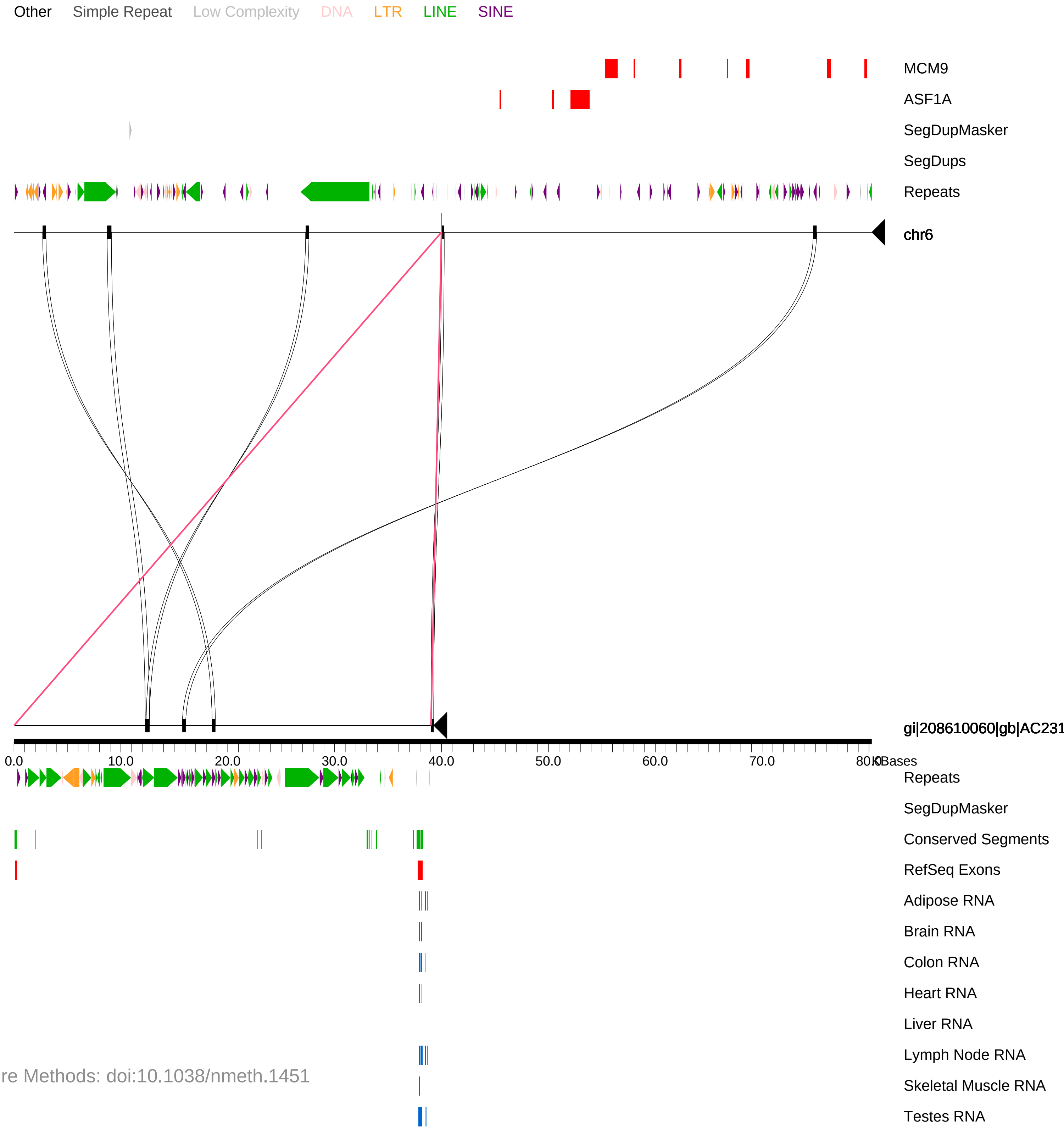
Insertion Size: 38667



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC231118.fa

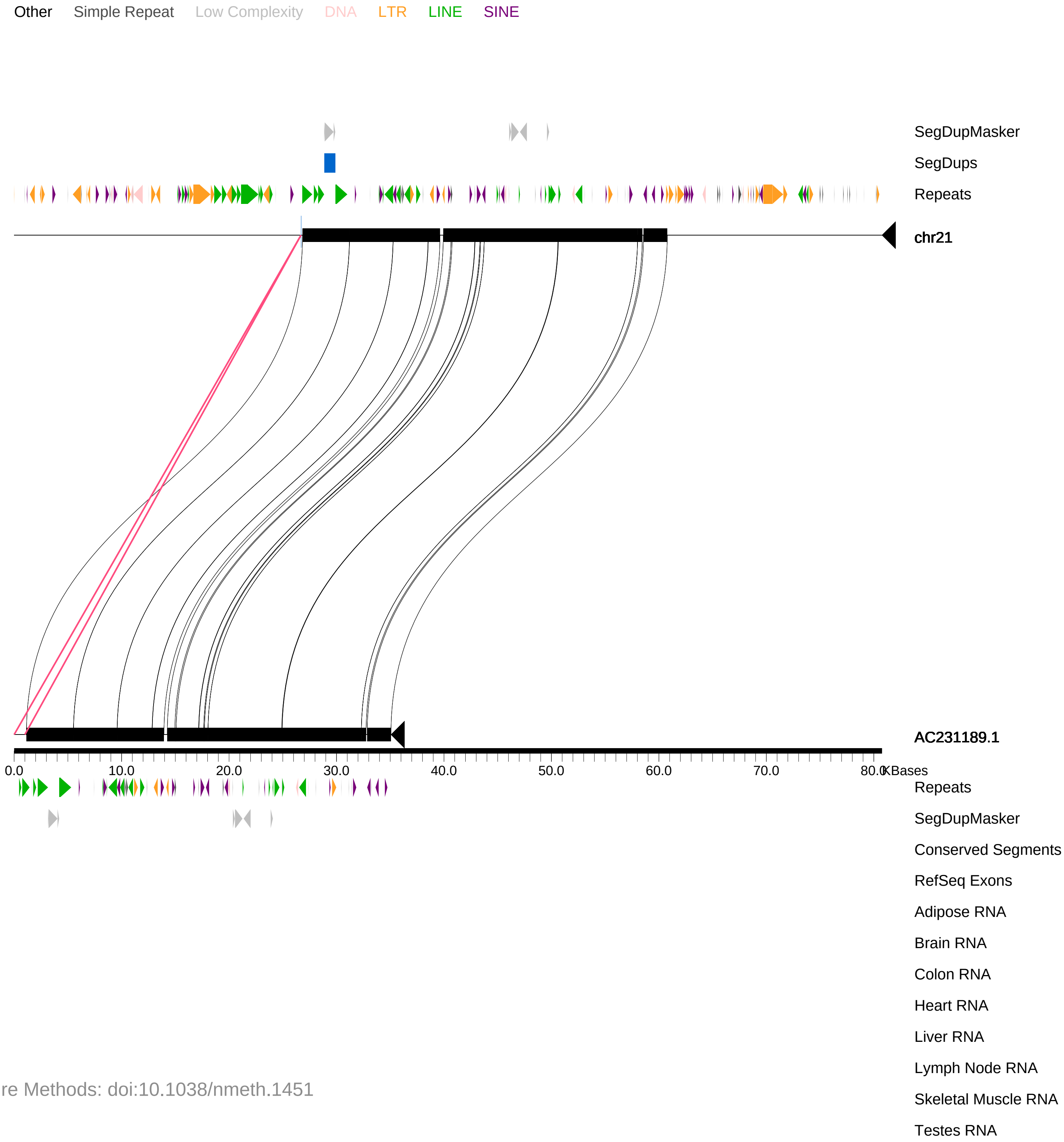
Insertion Size: 39015



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC231189.rc.fa

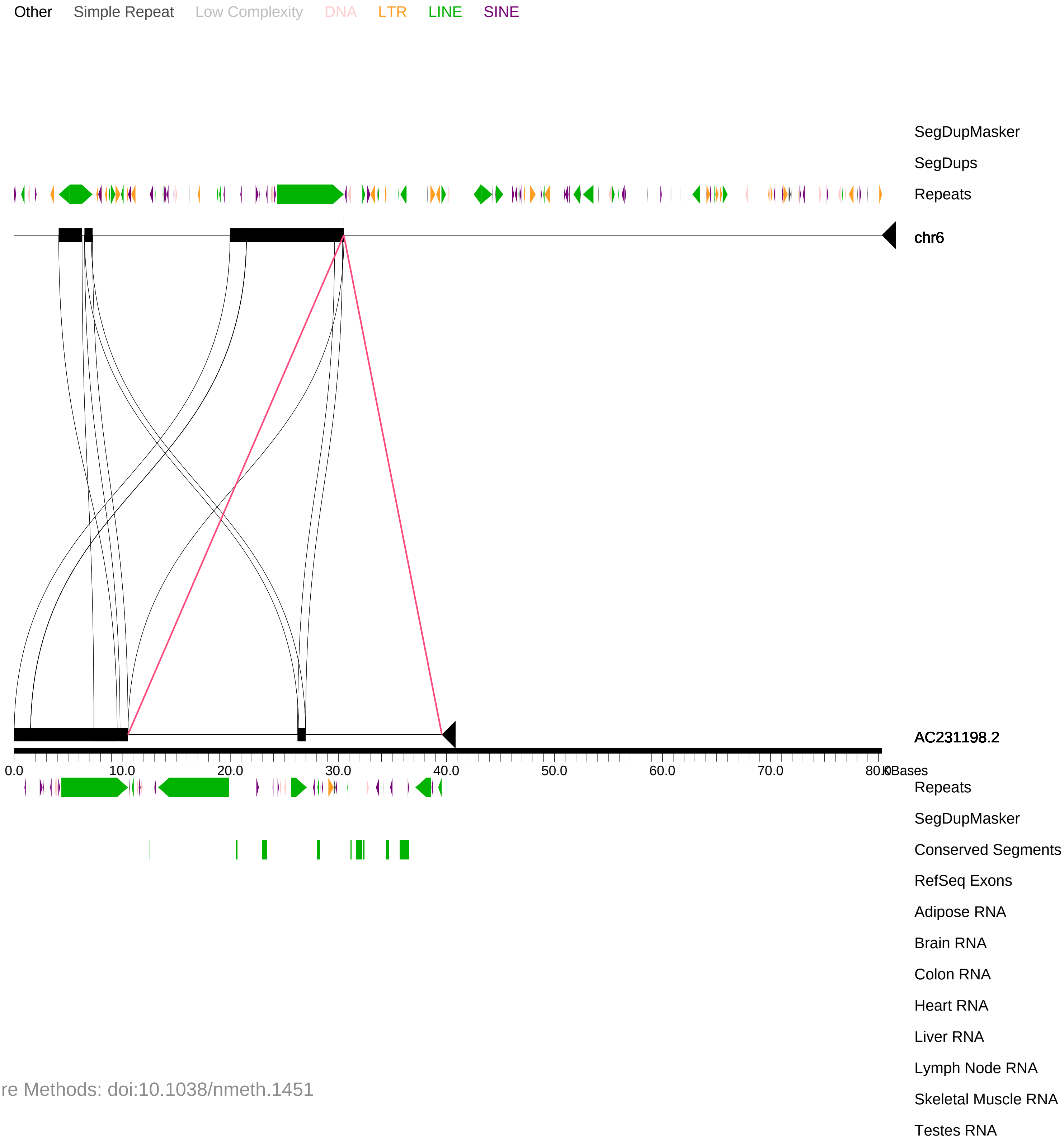
Insertion Size: 1002



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC231198.rc.fa

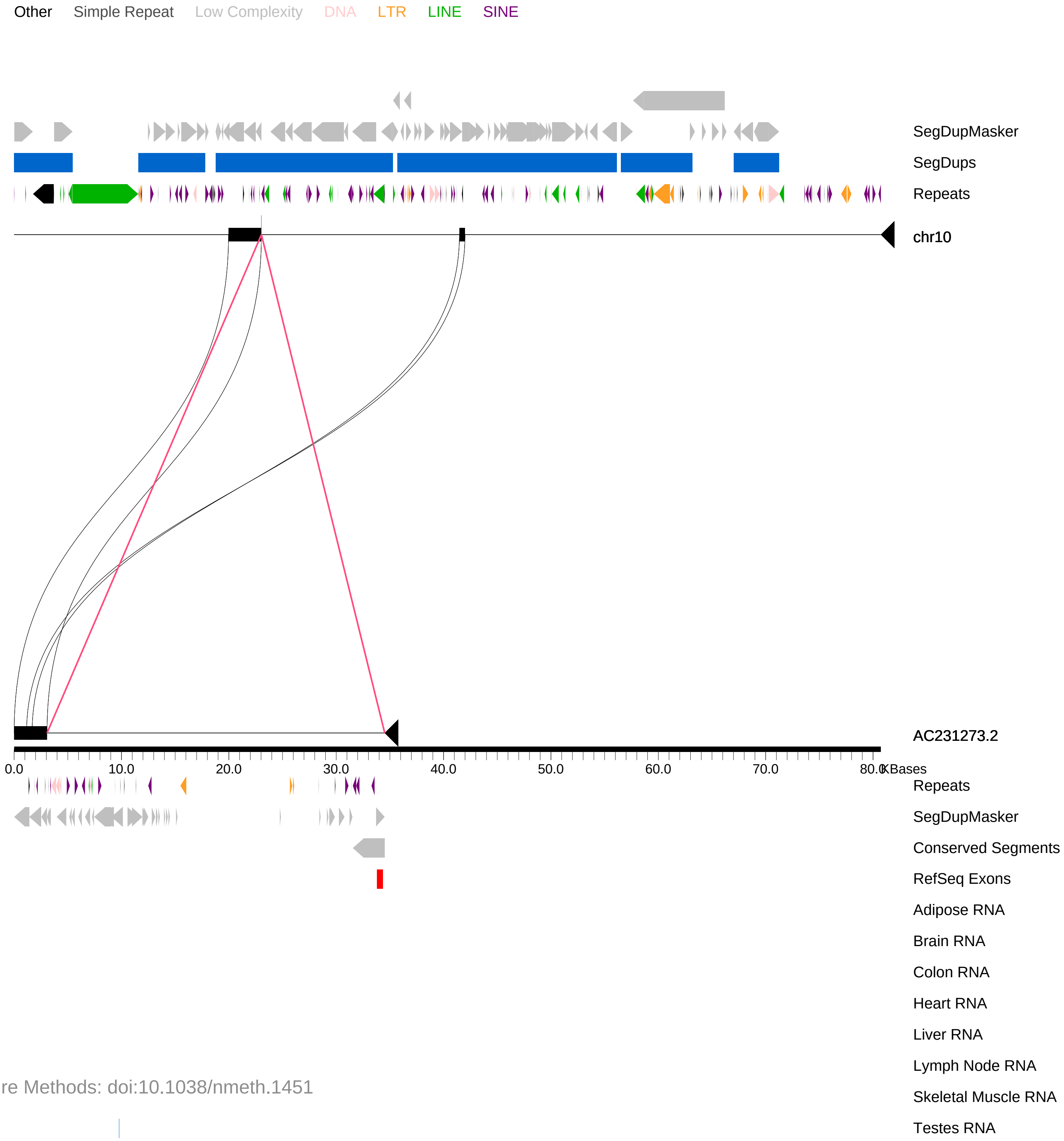
Insertion Size: 29057



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC231273.rc.fa

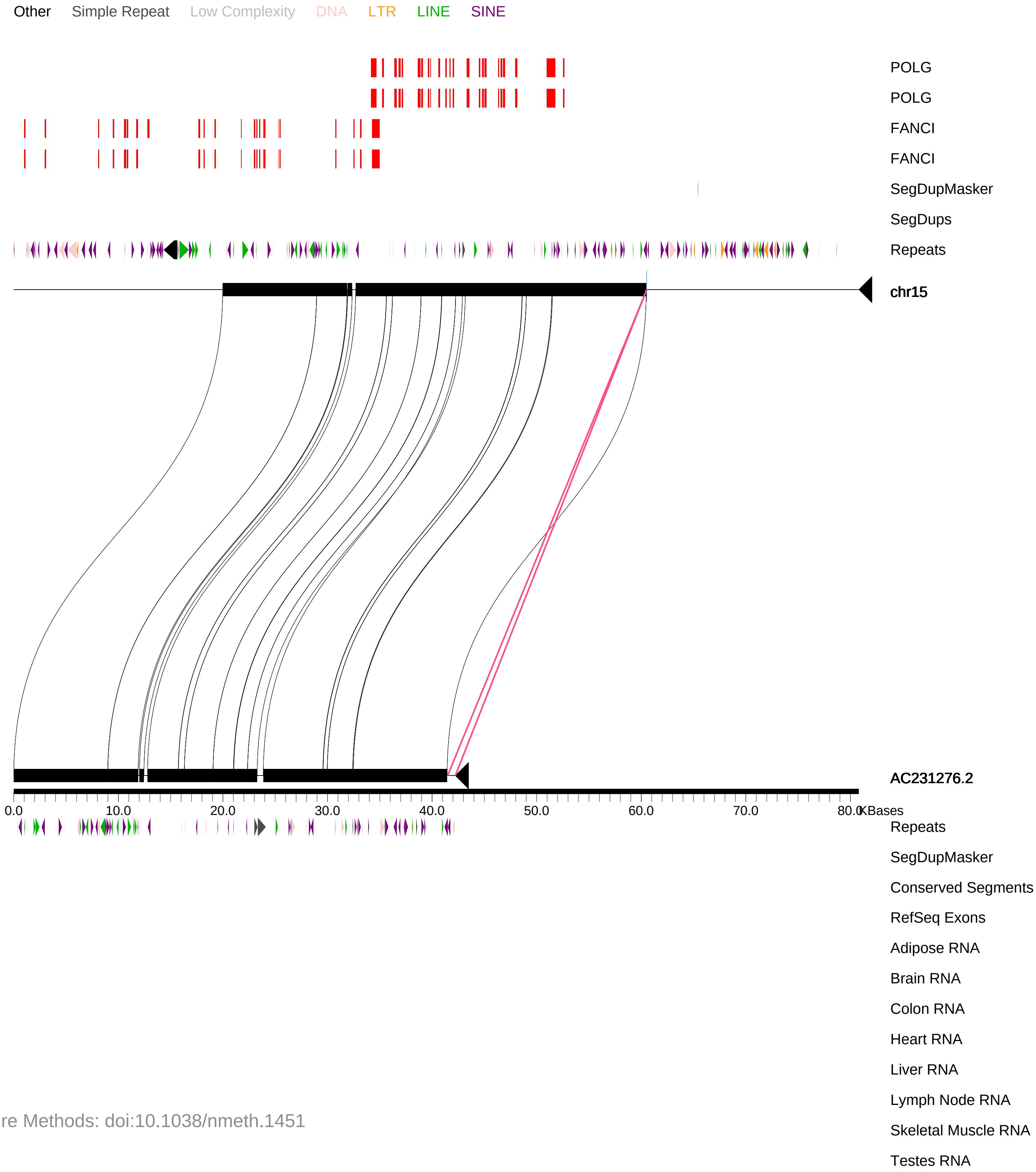
Insertion Size: 31455



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC231276.rc.fa

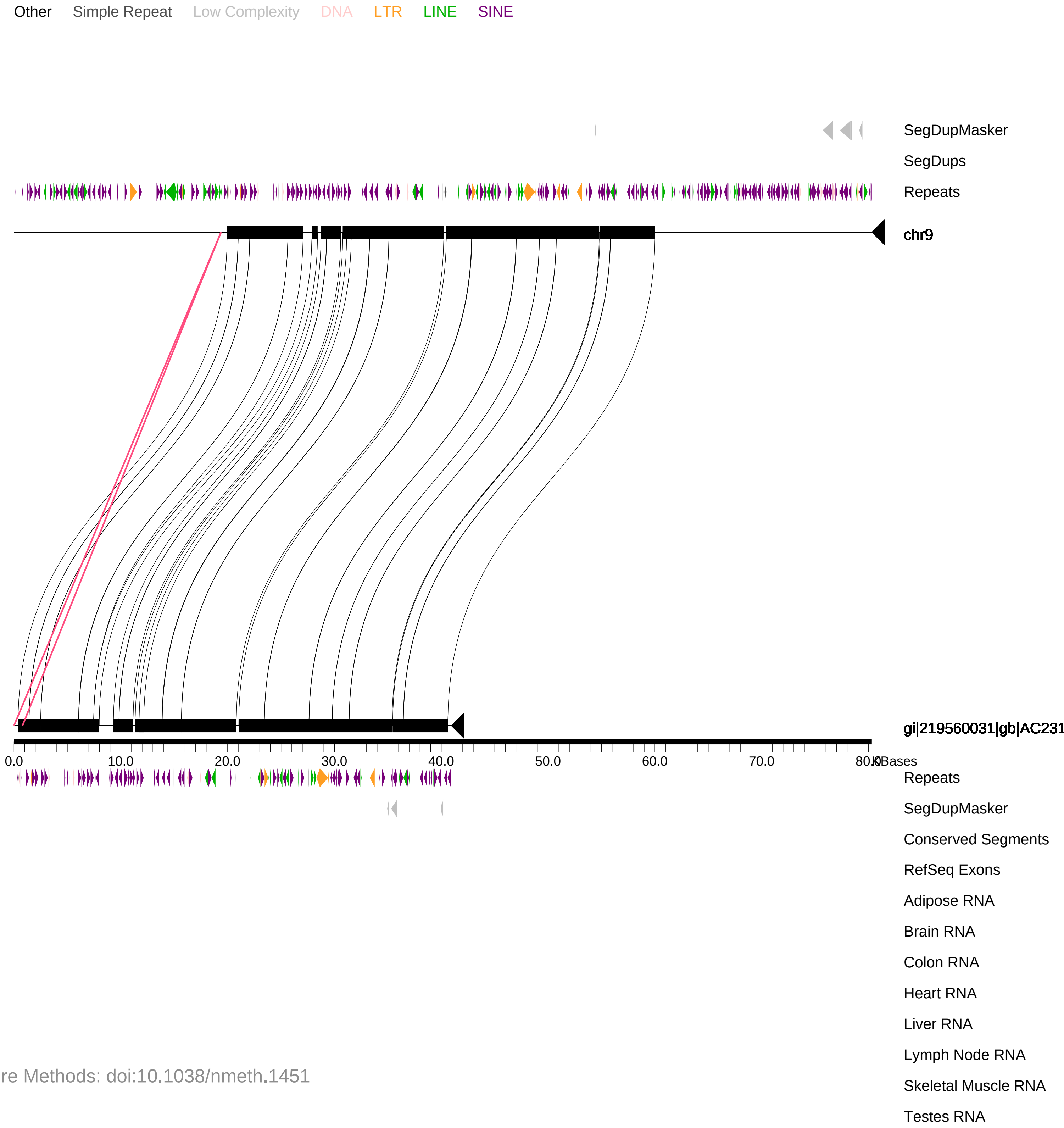
Insertion Size: 735



Nature Methods: doi:10.1038/nmeth.1451

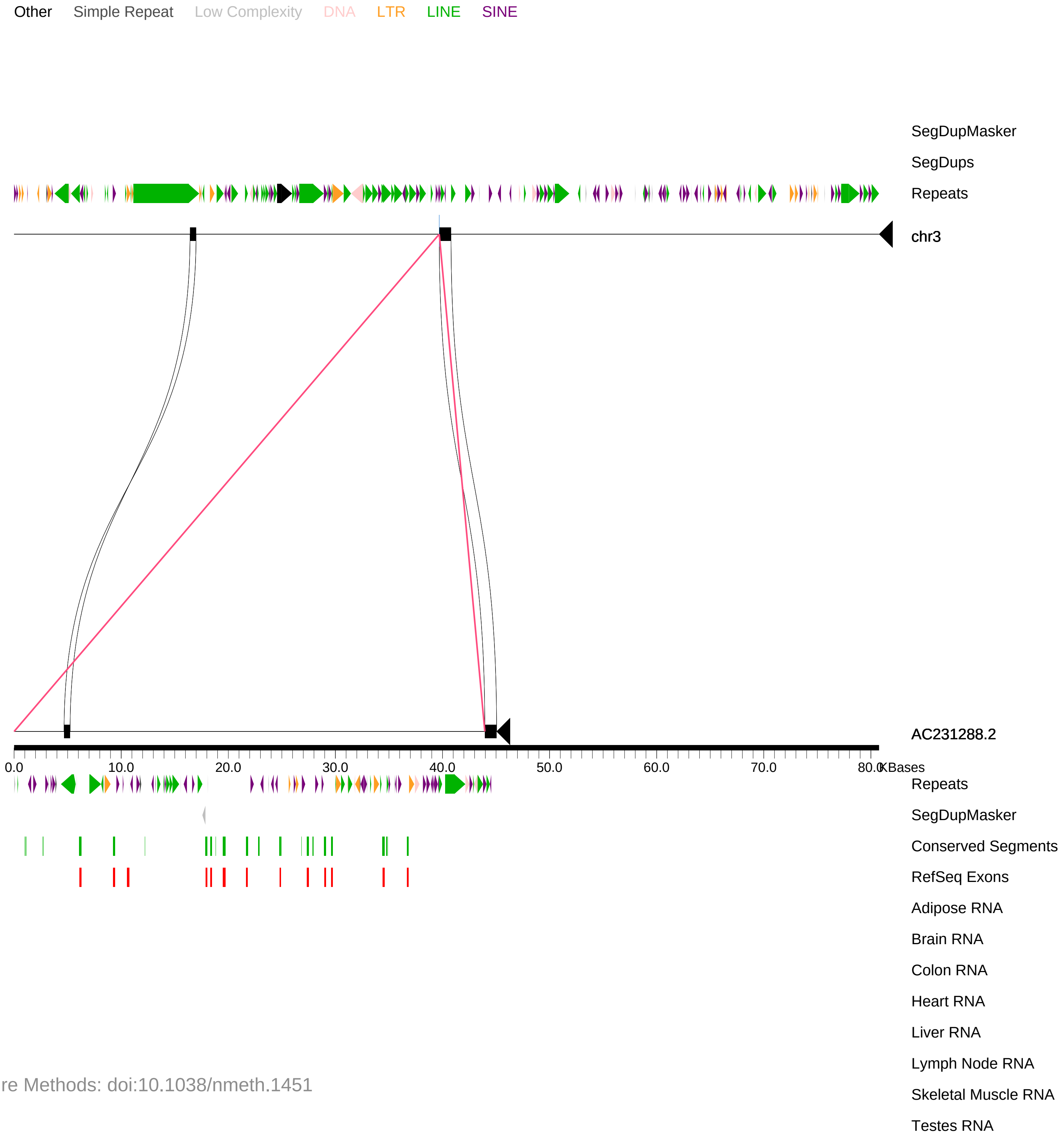
Clone file = AC231287.fa

Insertion Size: 798



Clone file = AC231288.rc.fa

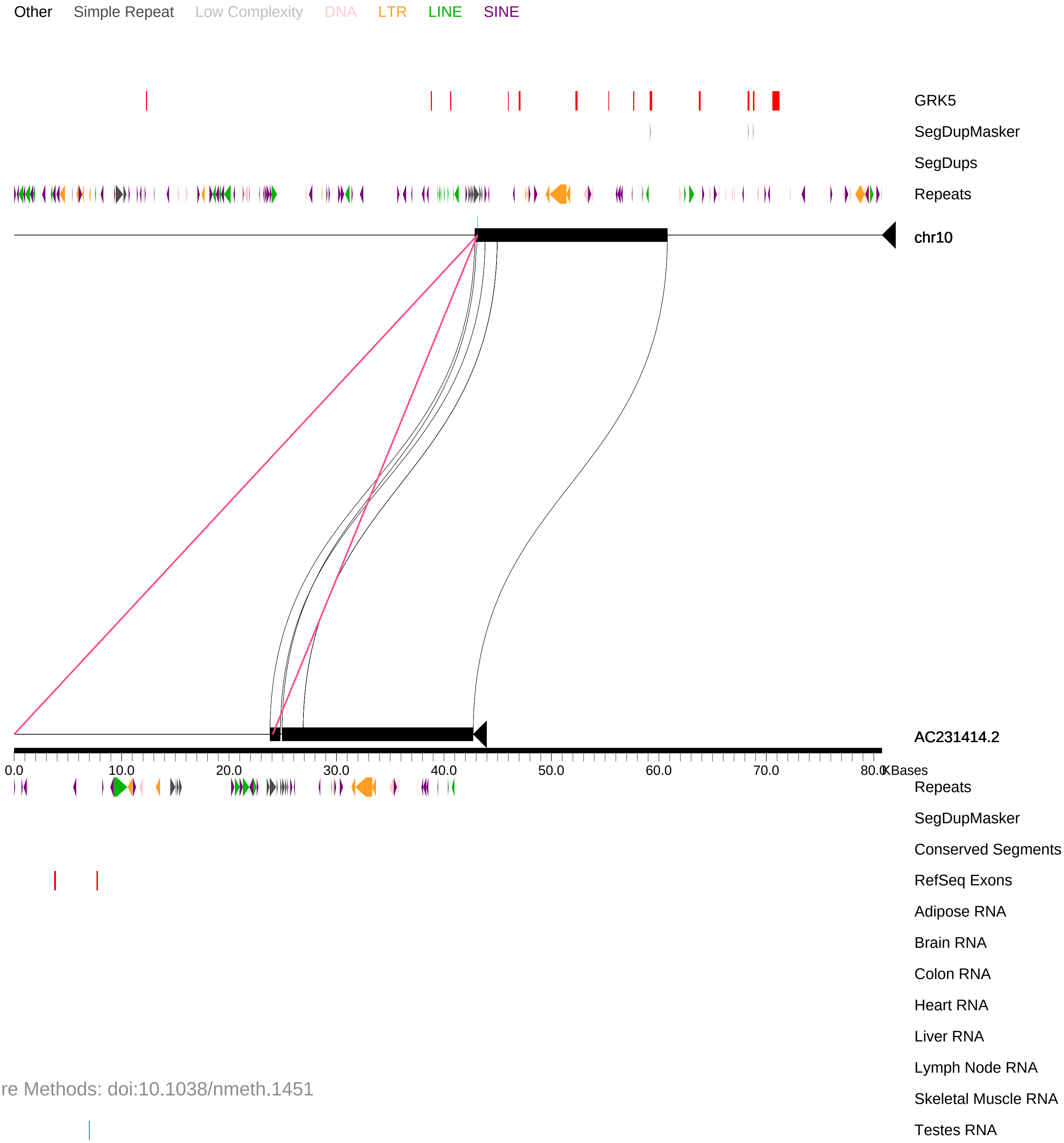
Insertion Size: 43954



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC231414.rc.fa

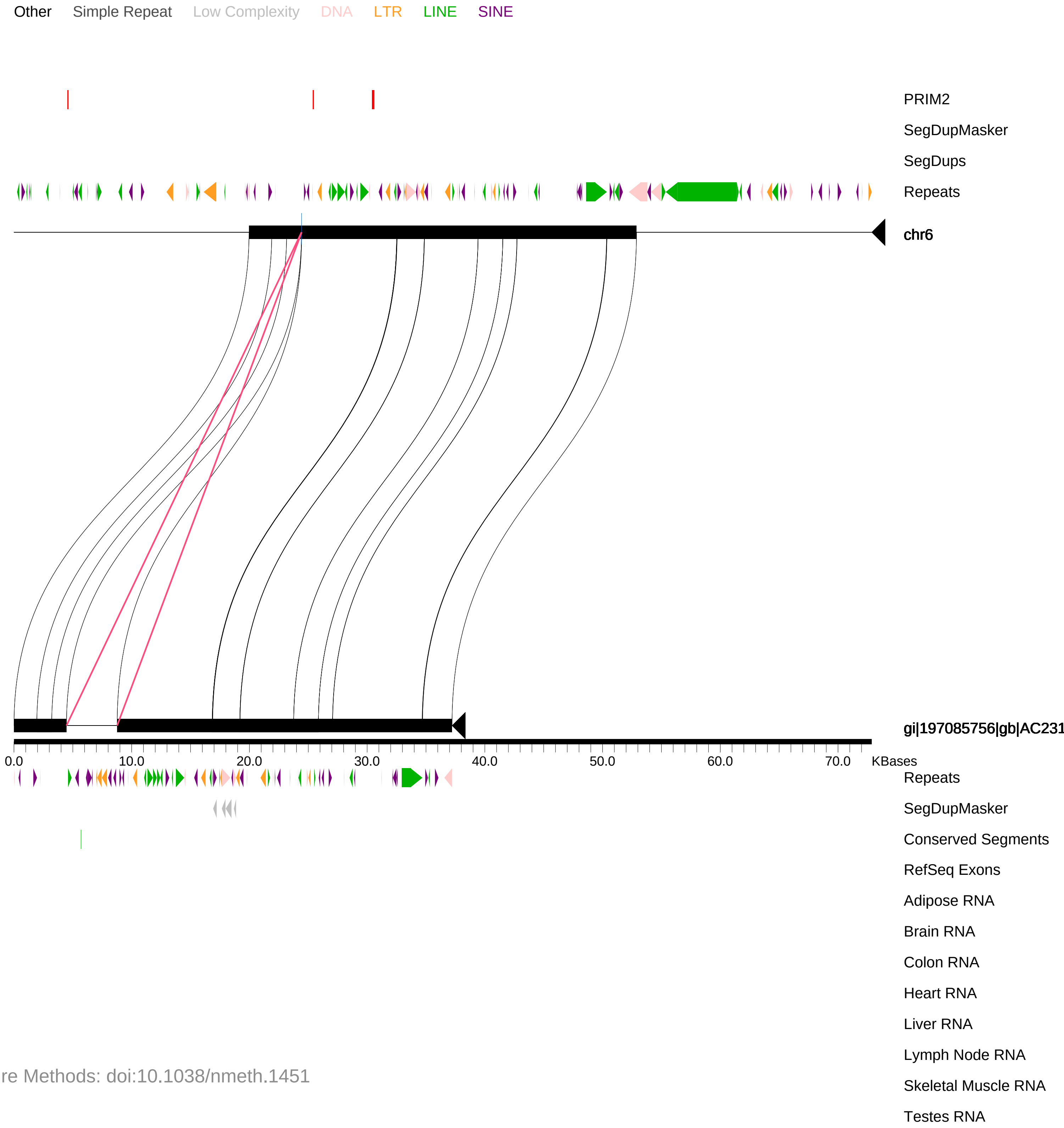
Insertion Size: 24063



Nature Methods: doi:10.1038/nmeth.1451

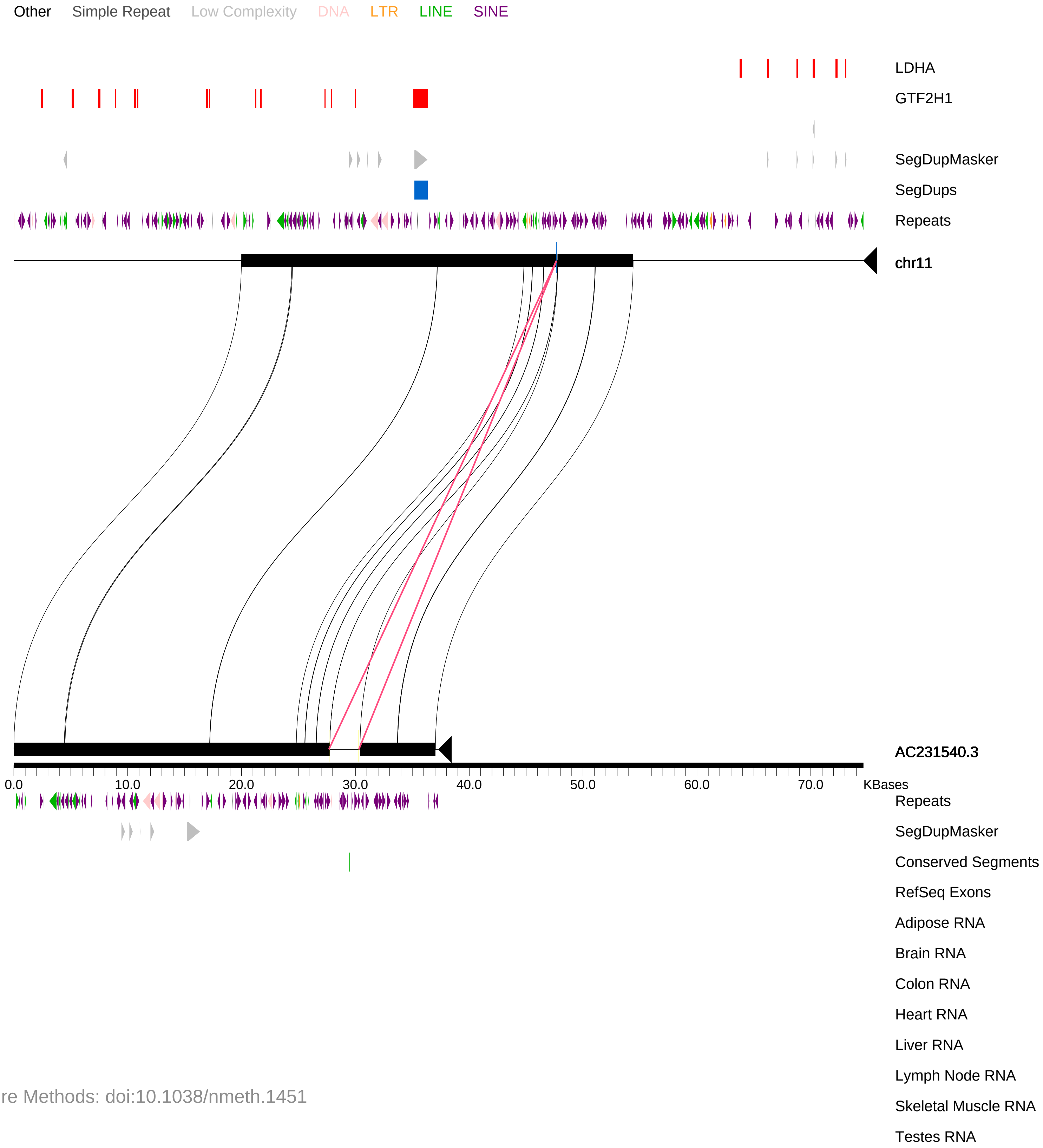
Clone file = AC231536.fa

Insertion Size: 4306



Clone file = AC231540.rc.fa

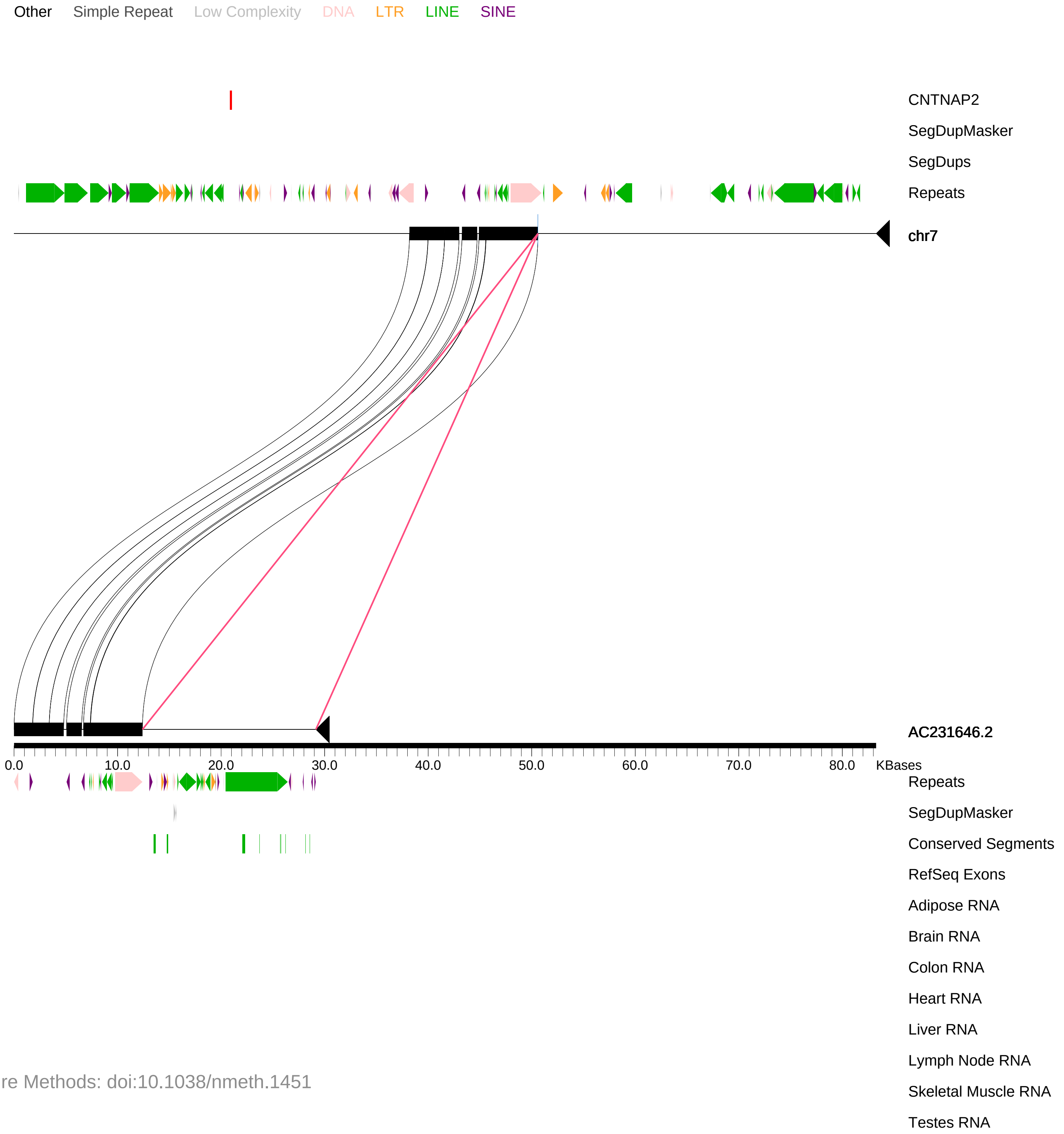
Insertion Size: 2633



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC231646.rc.fa

Insertion Size: 16754

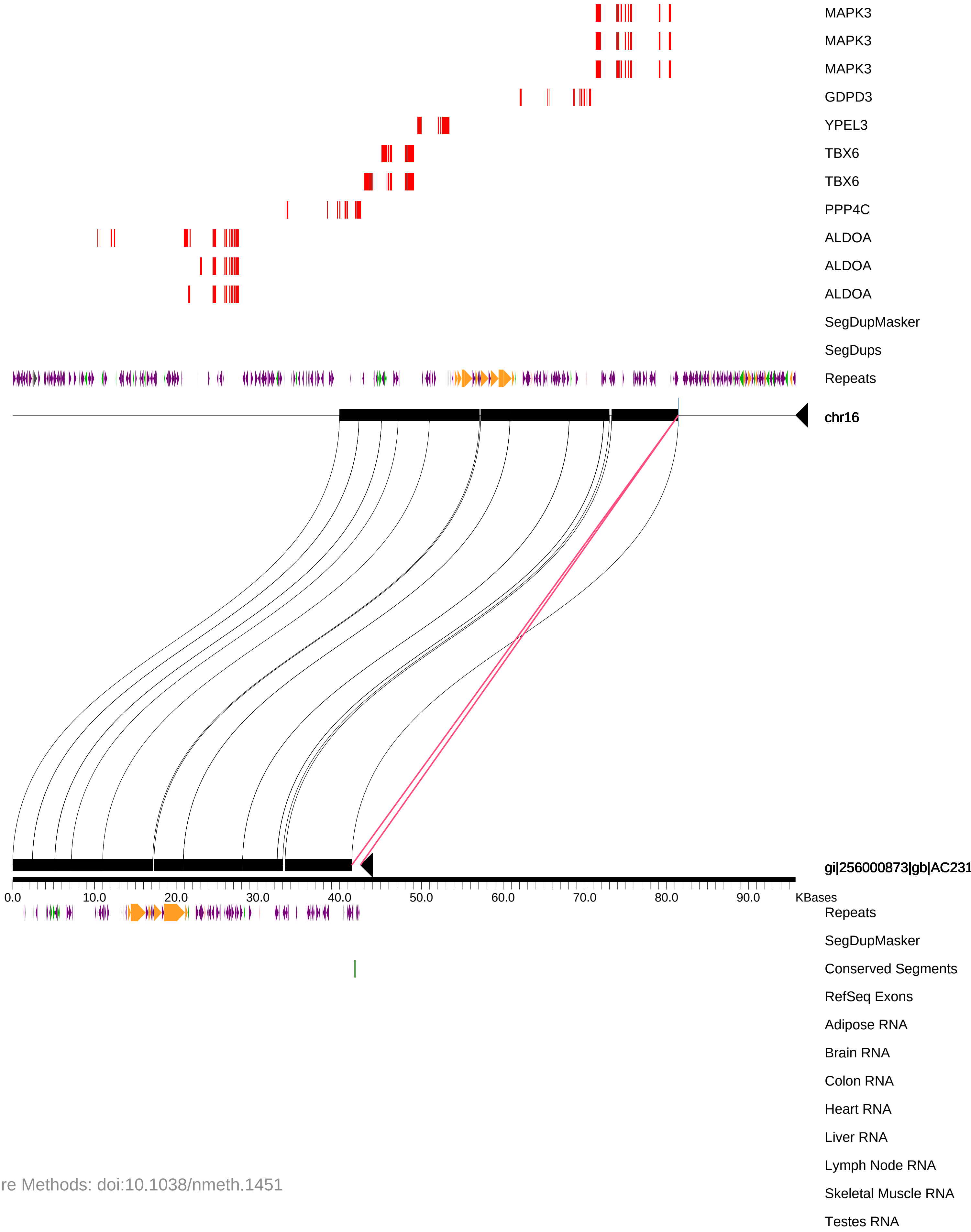


Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC231649.fa

Other Simple Repeat Low Complexity DNA LTR LINE SINE

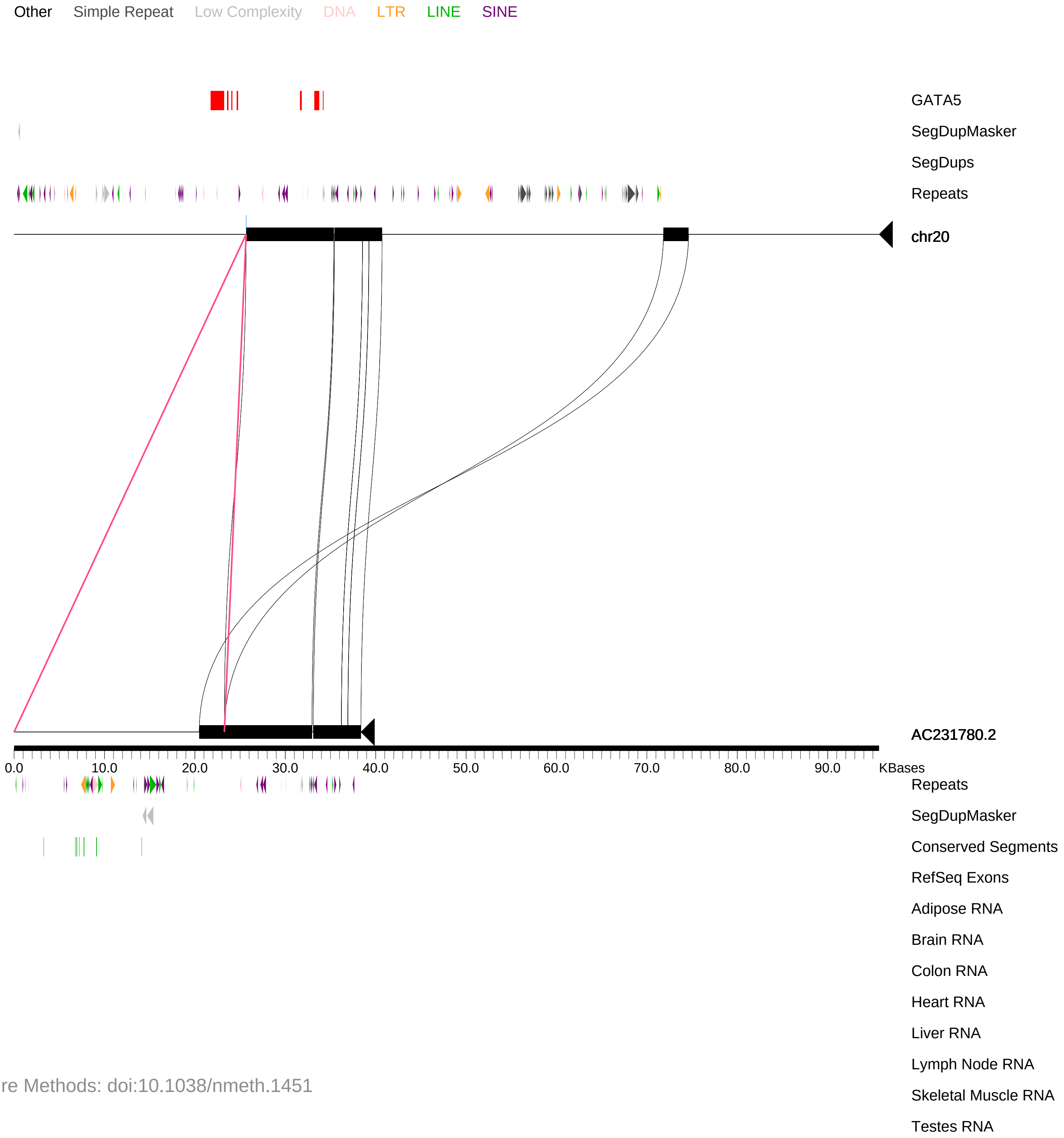
Insertion Size: 1052



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC231780.rc.fa

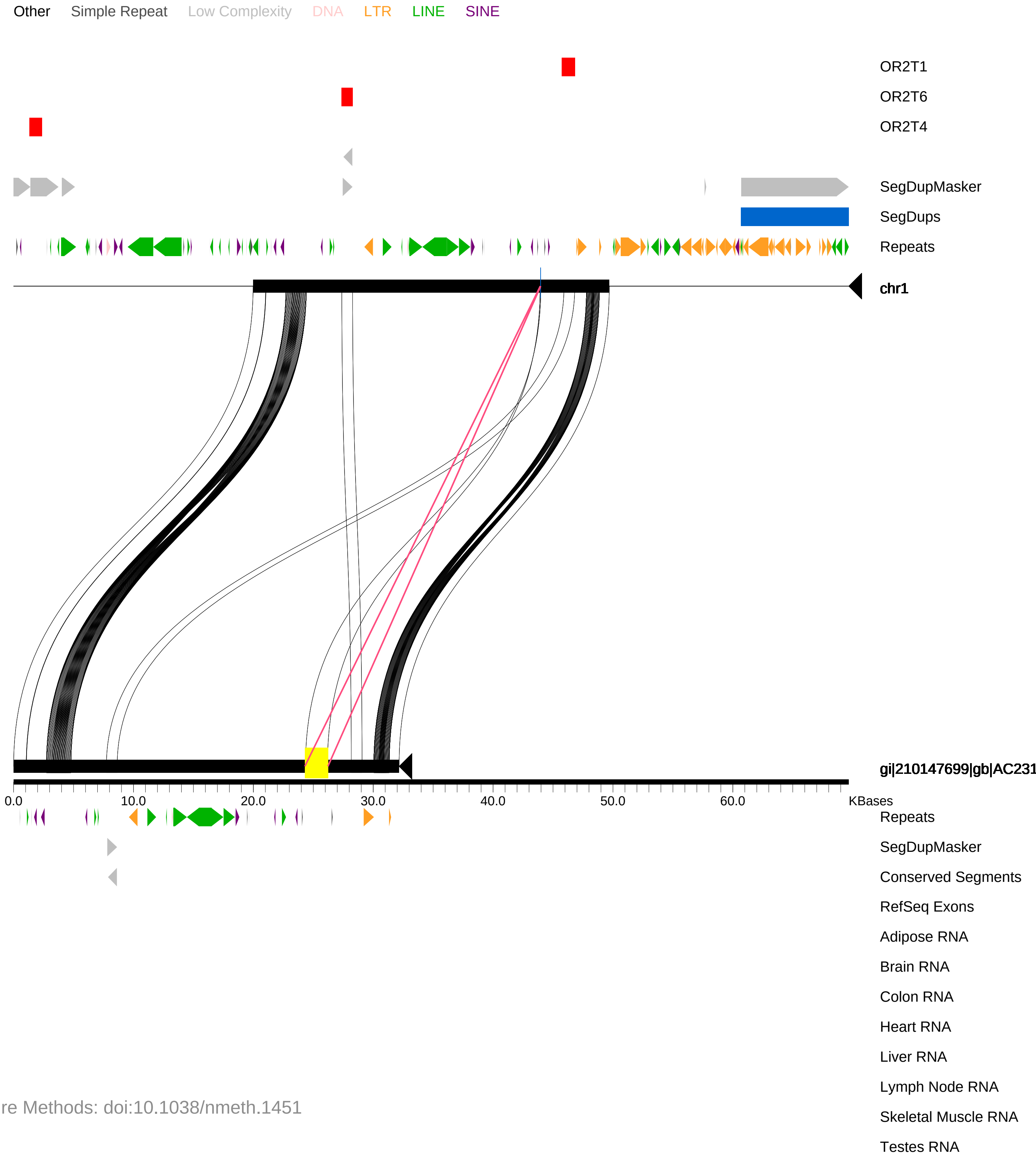
Insertion Size: 23257



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC231953.fa

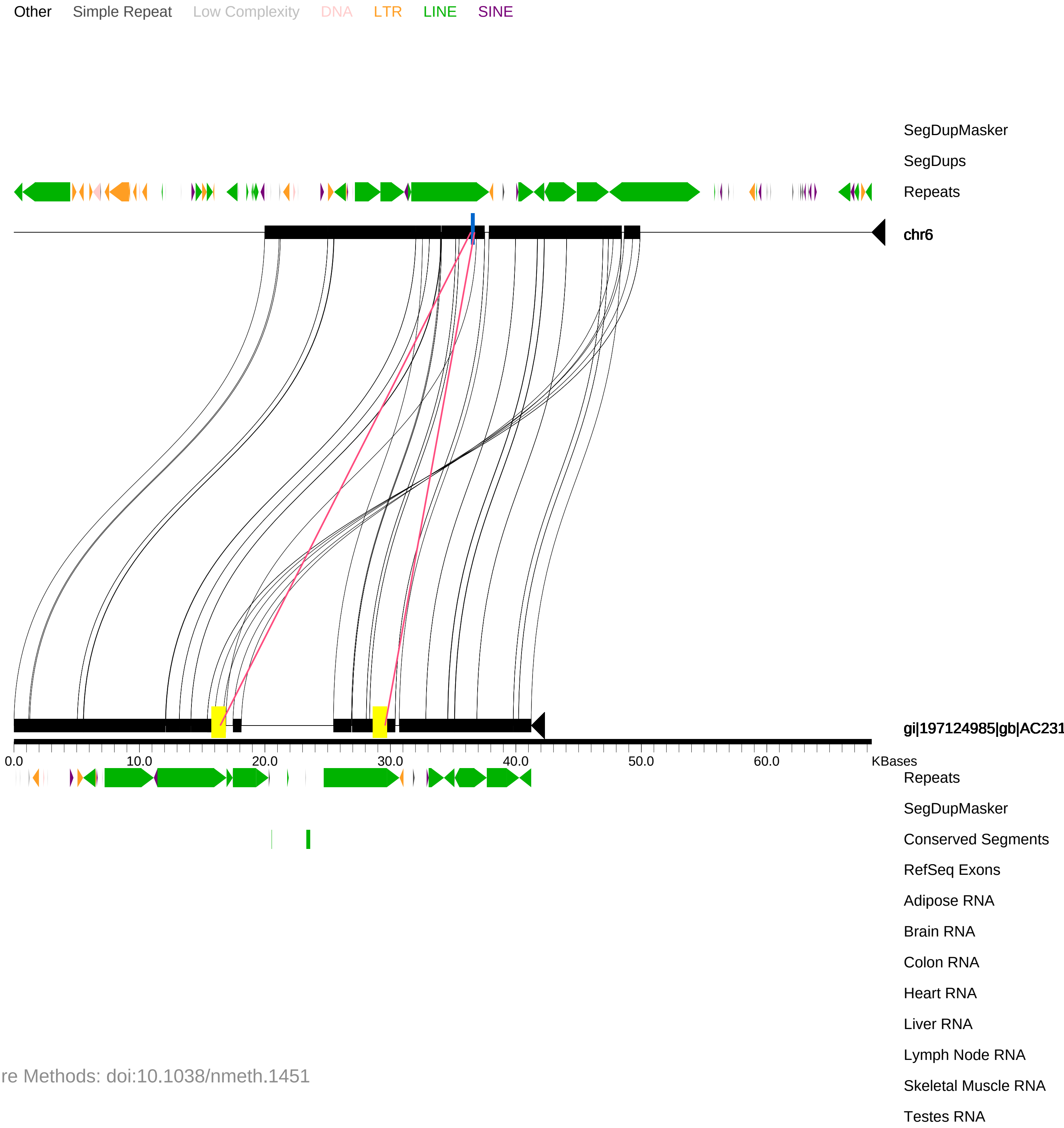
Insertion Size: 1925



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC231958.fa

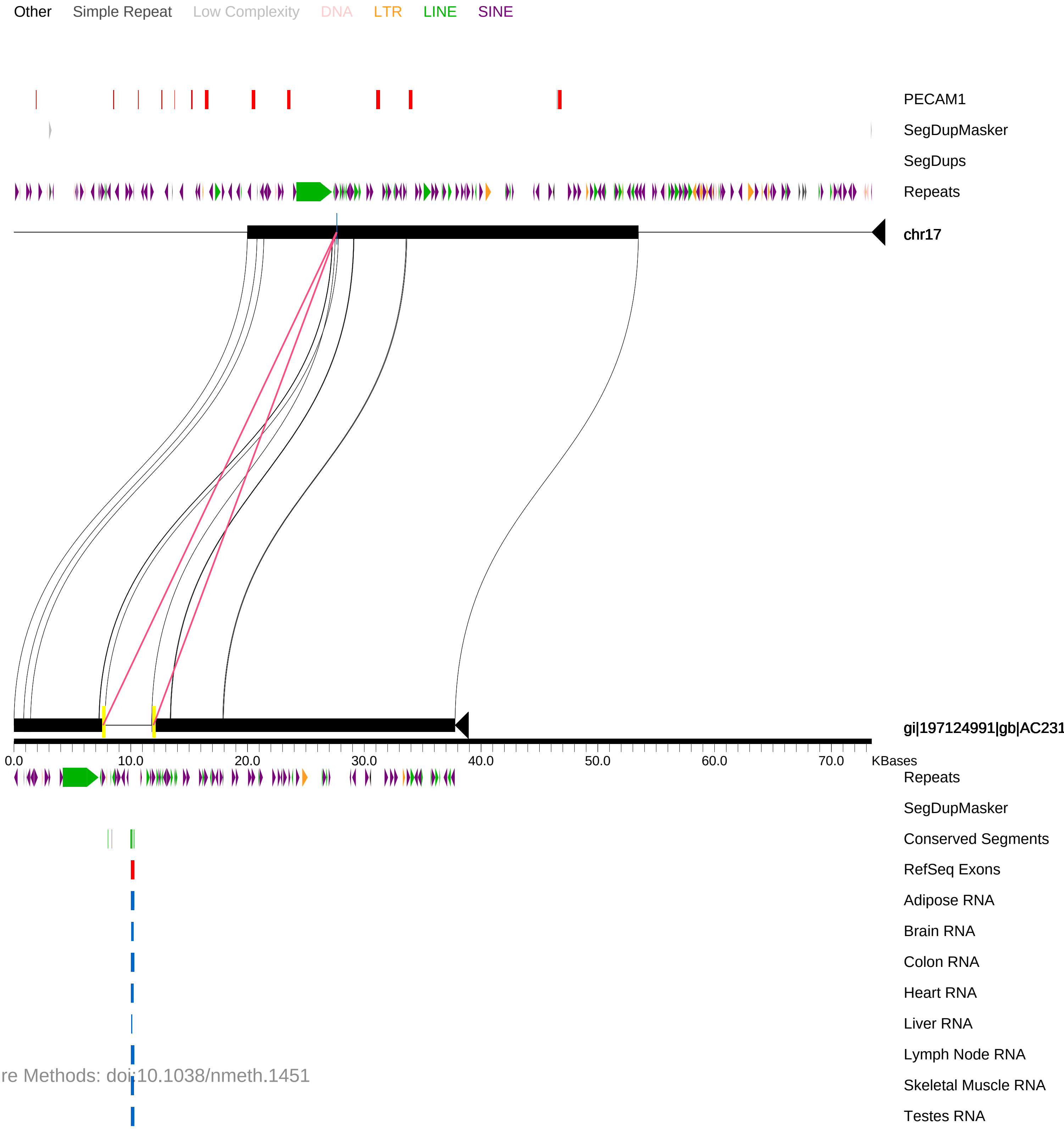
Insertion Size: 13141



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC231962.fa

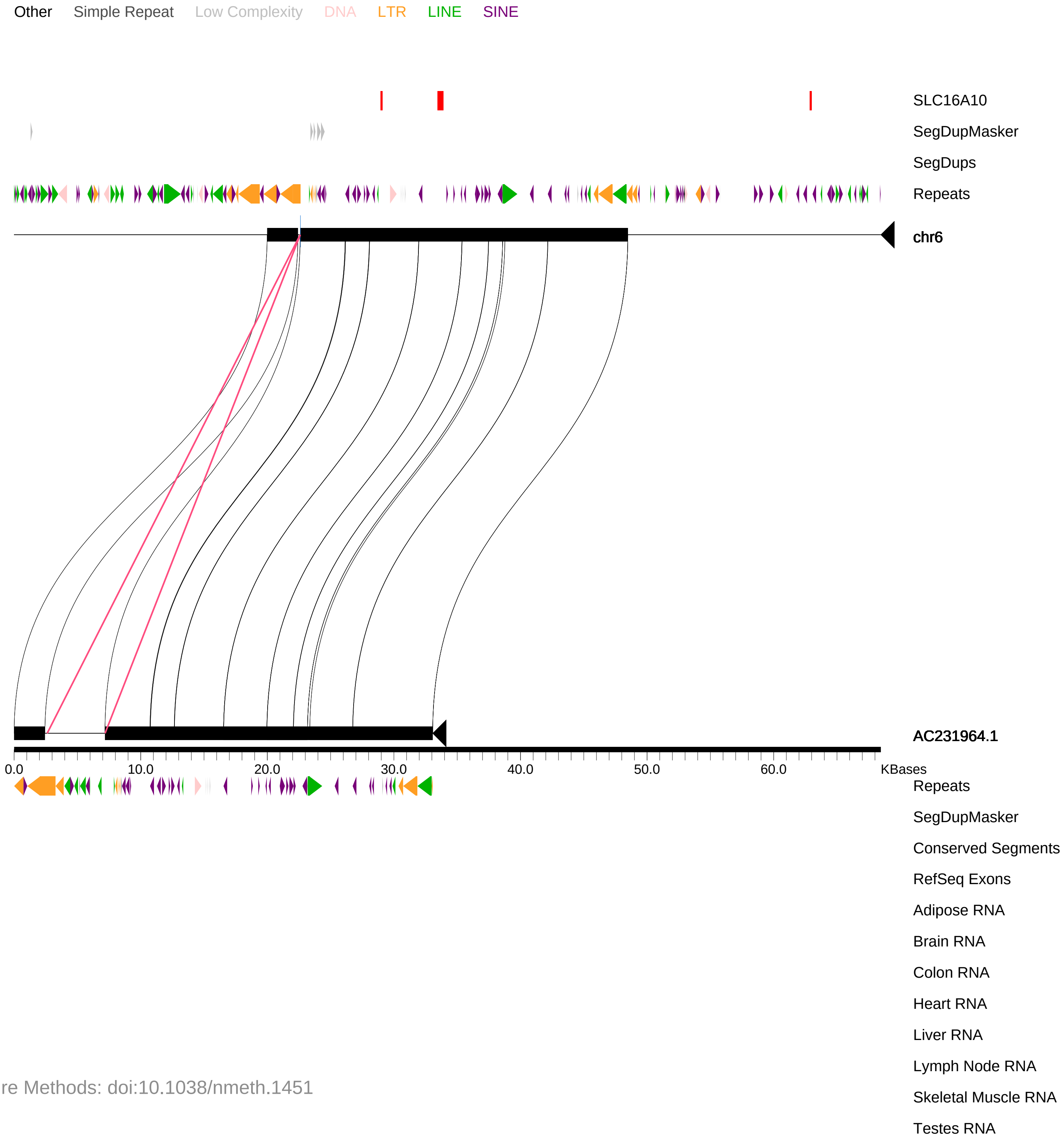
Insertion Size: 4316



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC231964.rc.fa

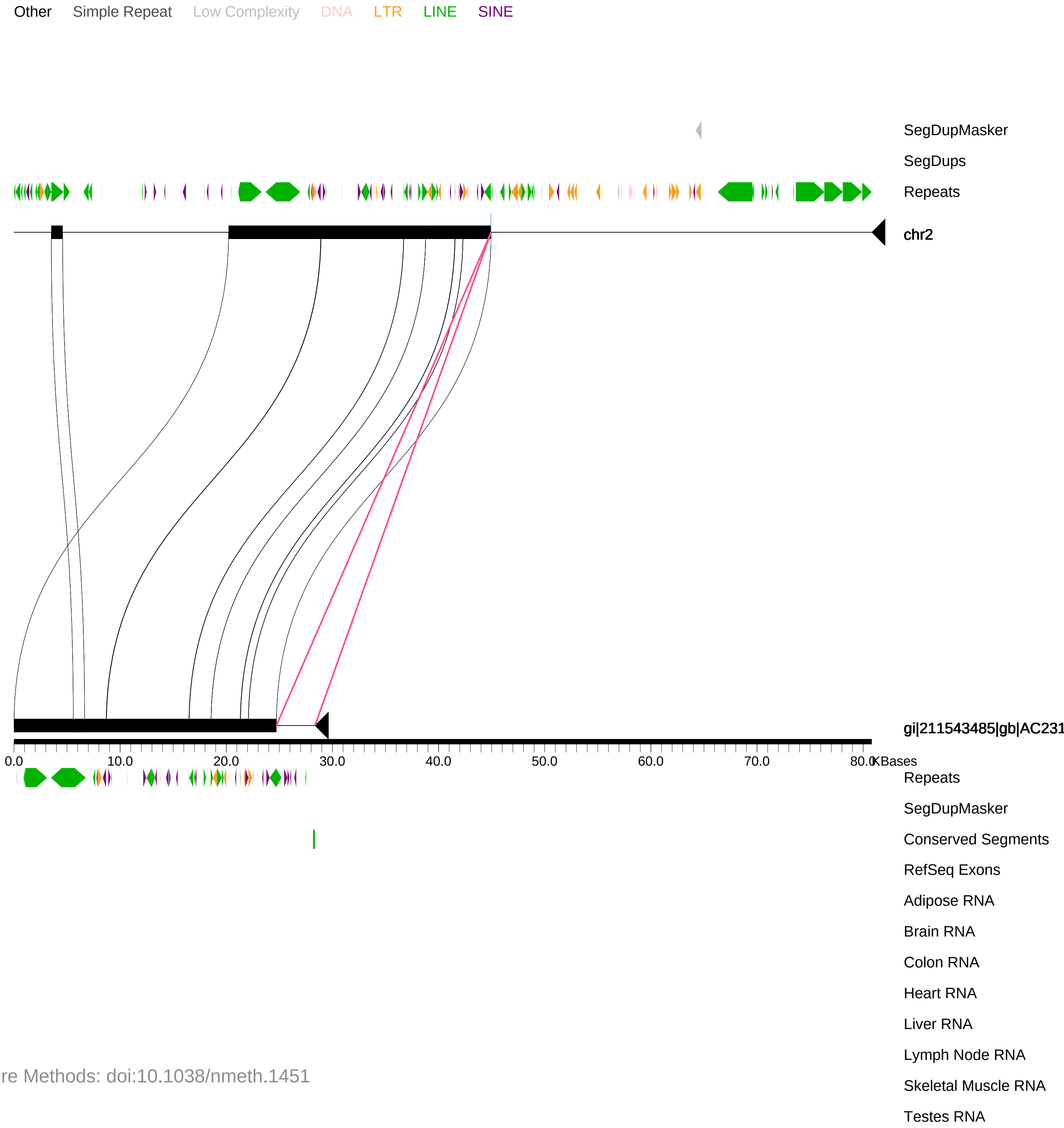
Insertion Size: 4571



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC231980.fa

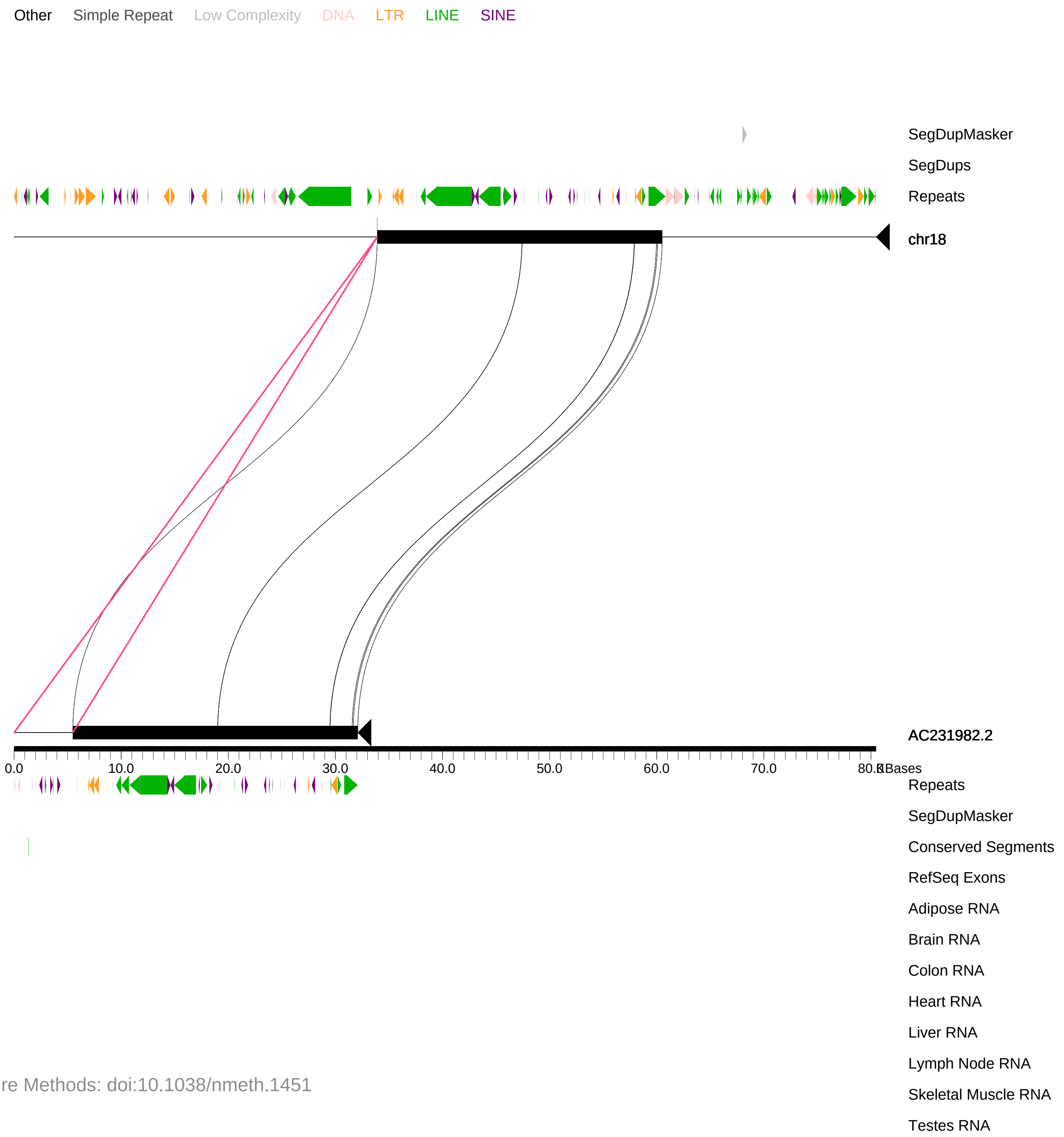
Insertion Size: 3638



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC231982.rc.fa

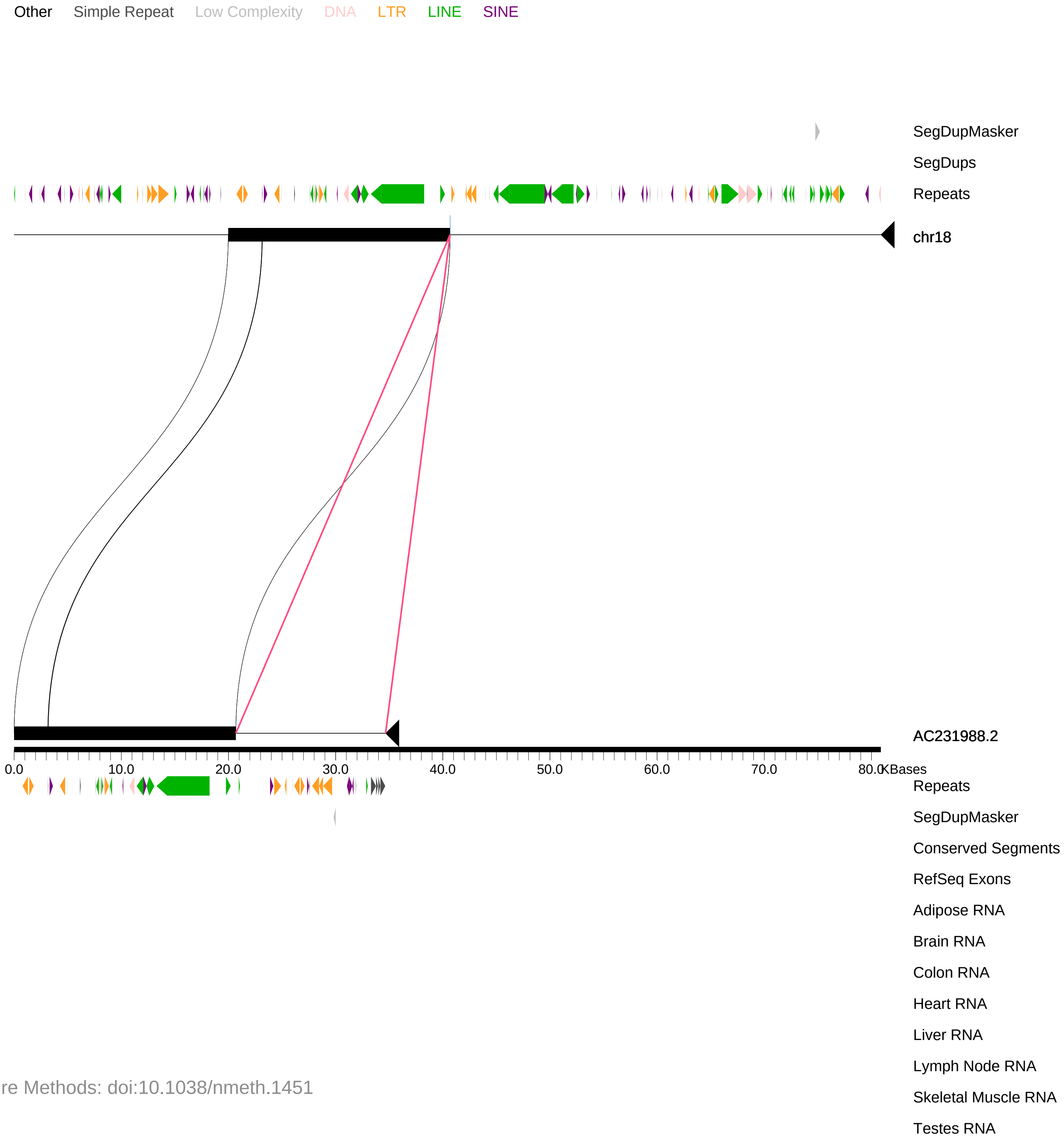
Insertion Size: 5487



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC231988.rc.fa

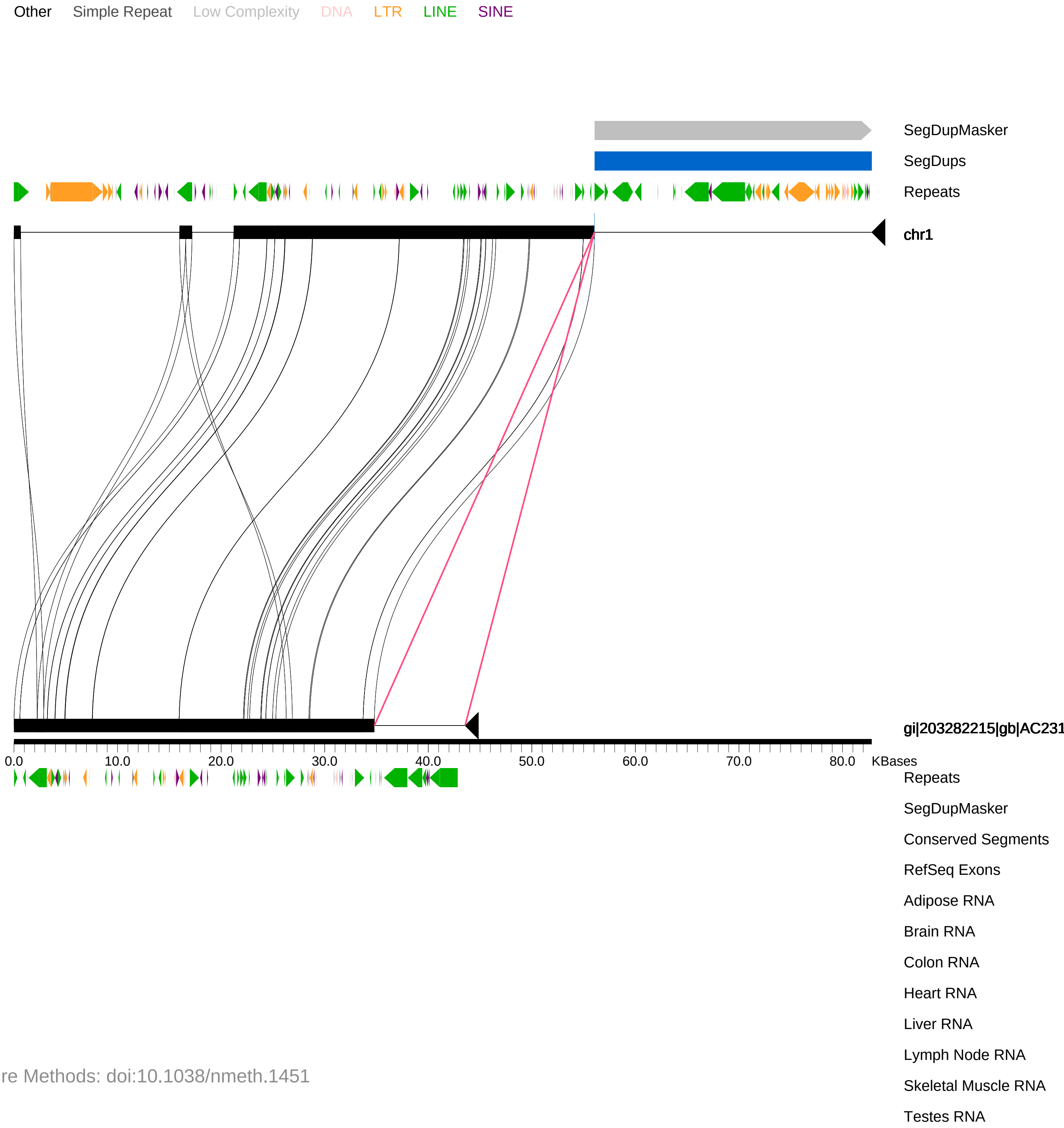
Insertion Size: 13975



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC231989.fa

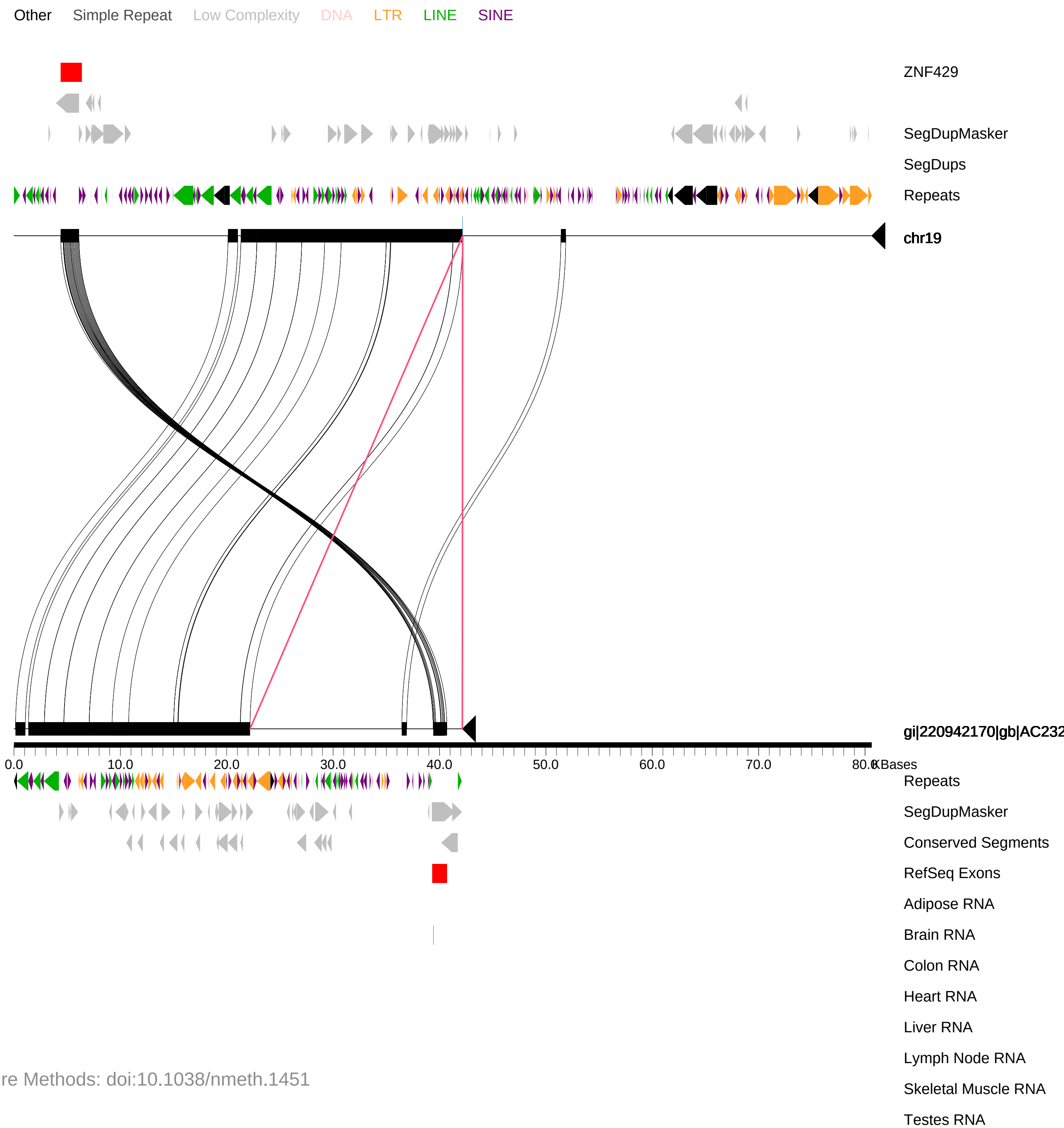
Insertion Size: 8761



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC232224.fa

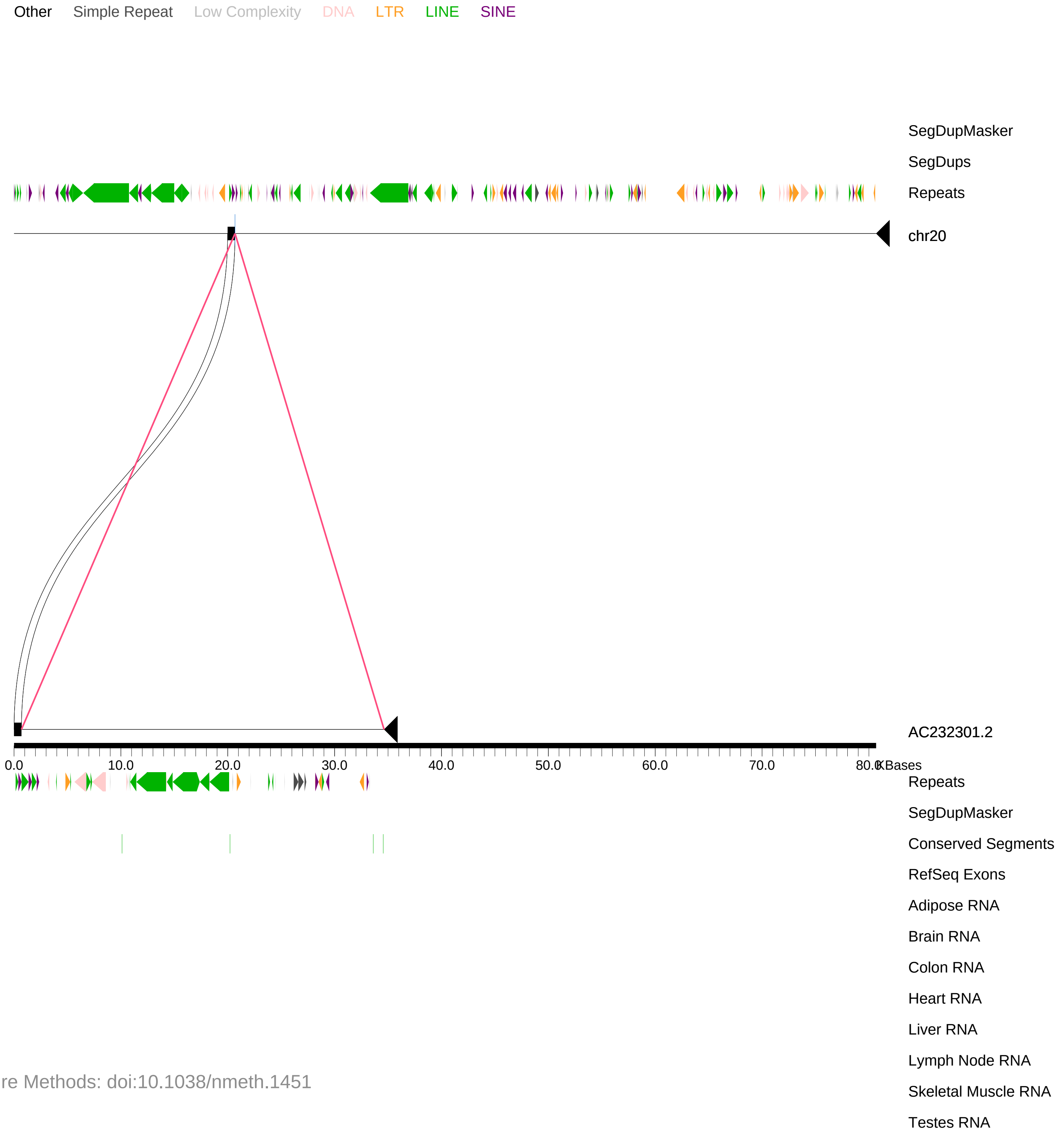
Insertion Size: 19960



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC232301.rc.fa

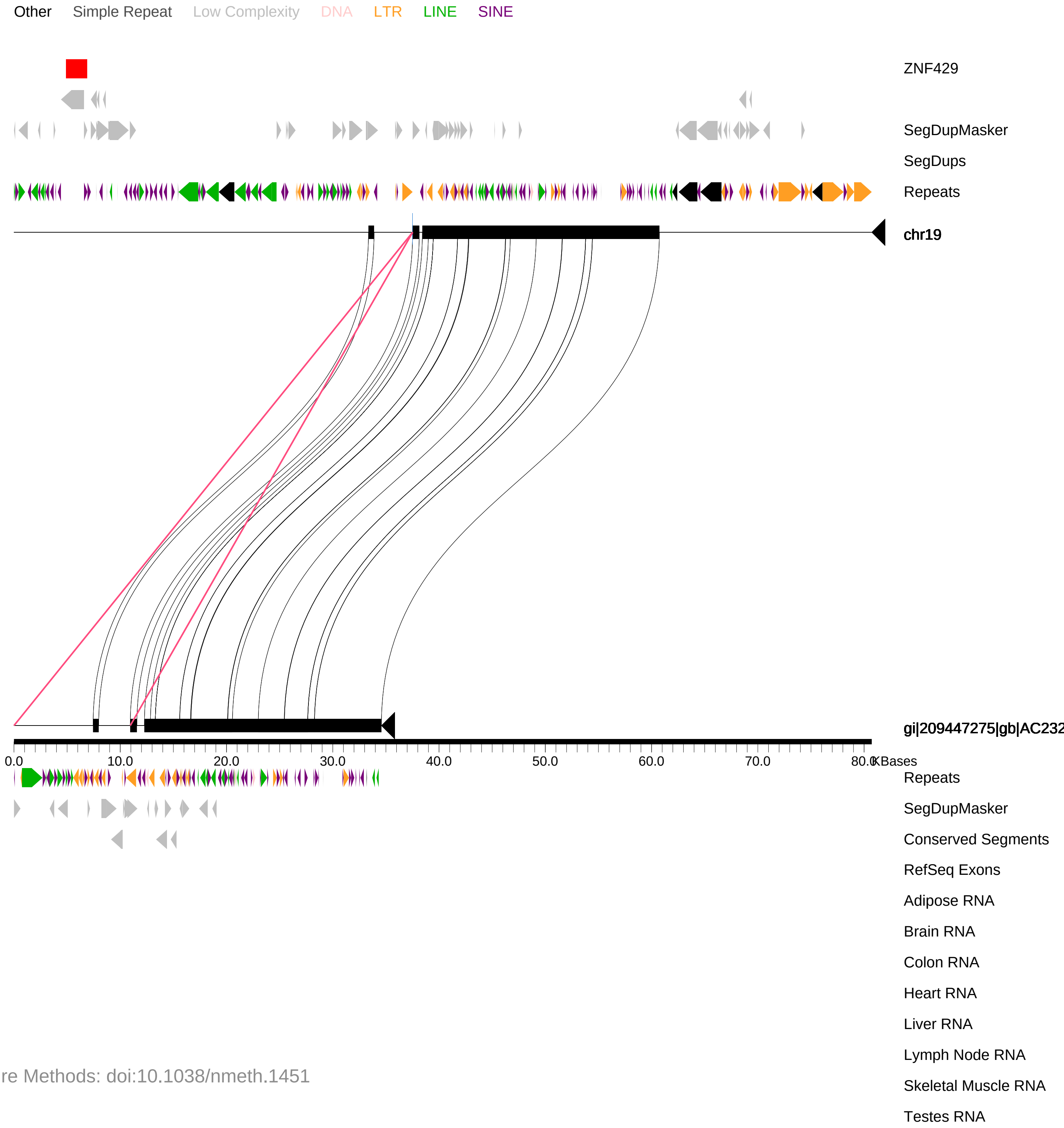
Insertion Size: 33937



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC232302.fa

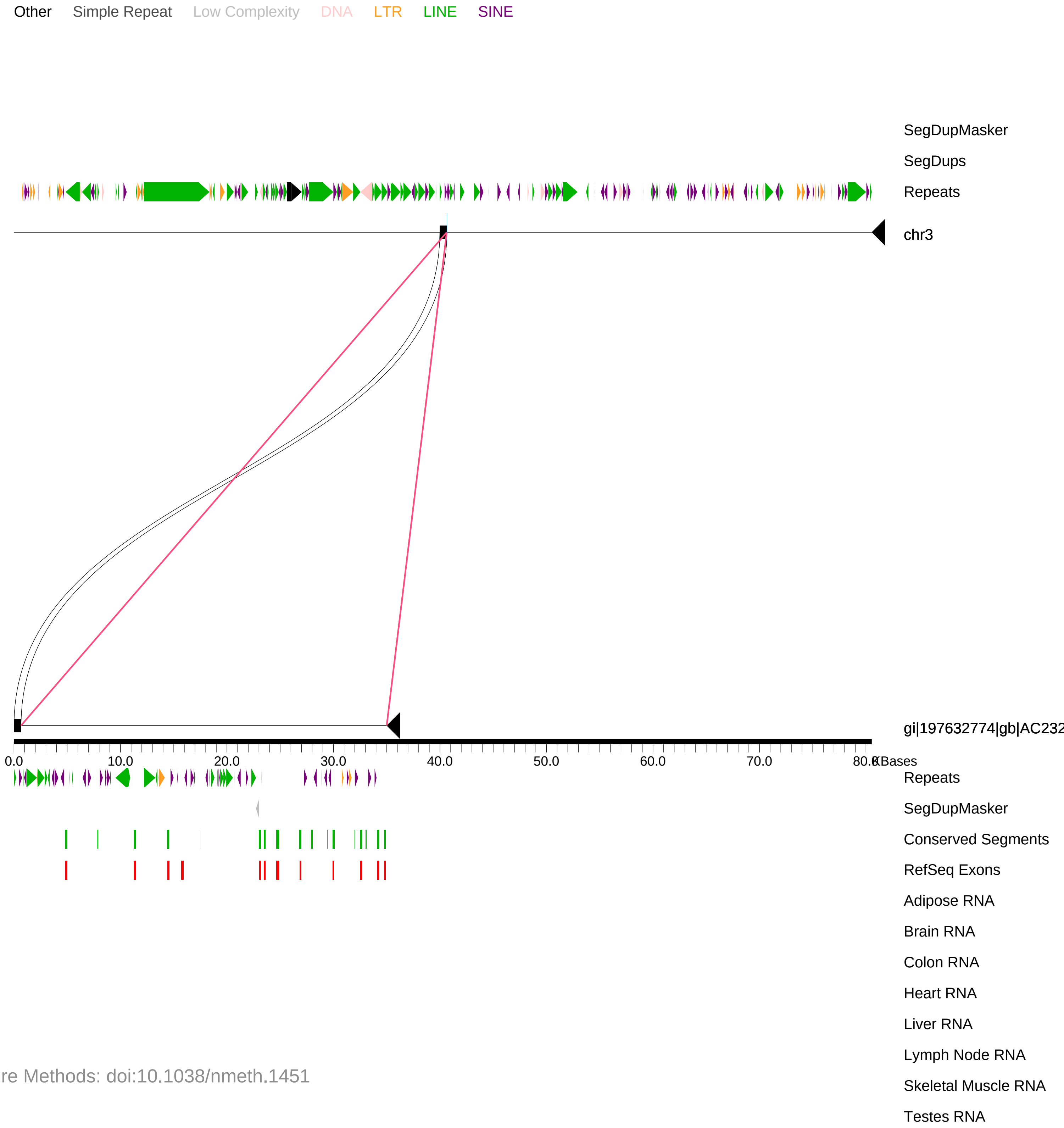
Insertion Size: 10943



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC232304.fa

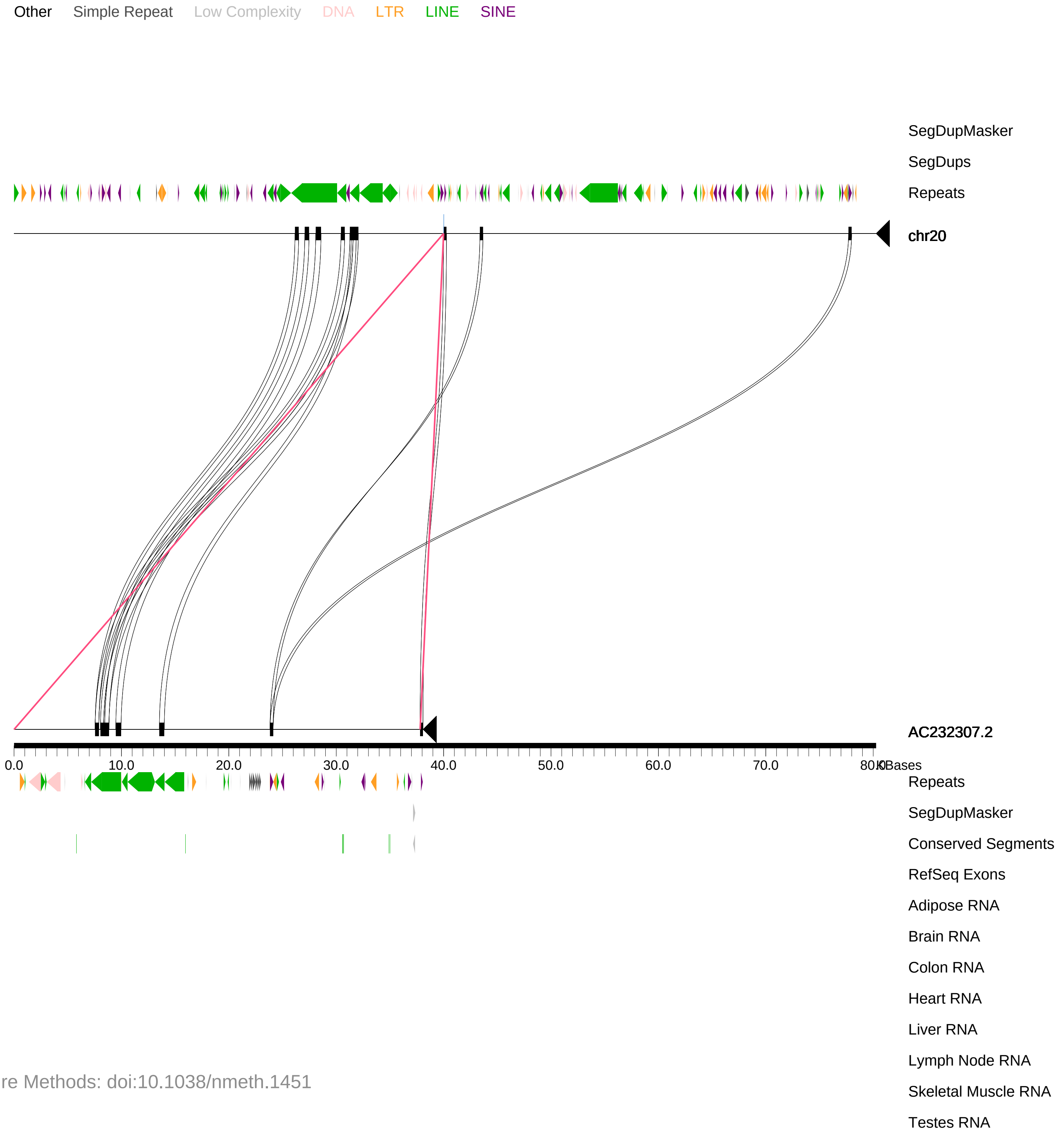
Insertion Size: 34338



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC232307.rc.fa

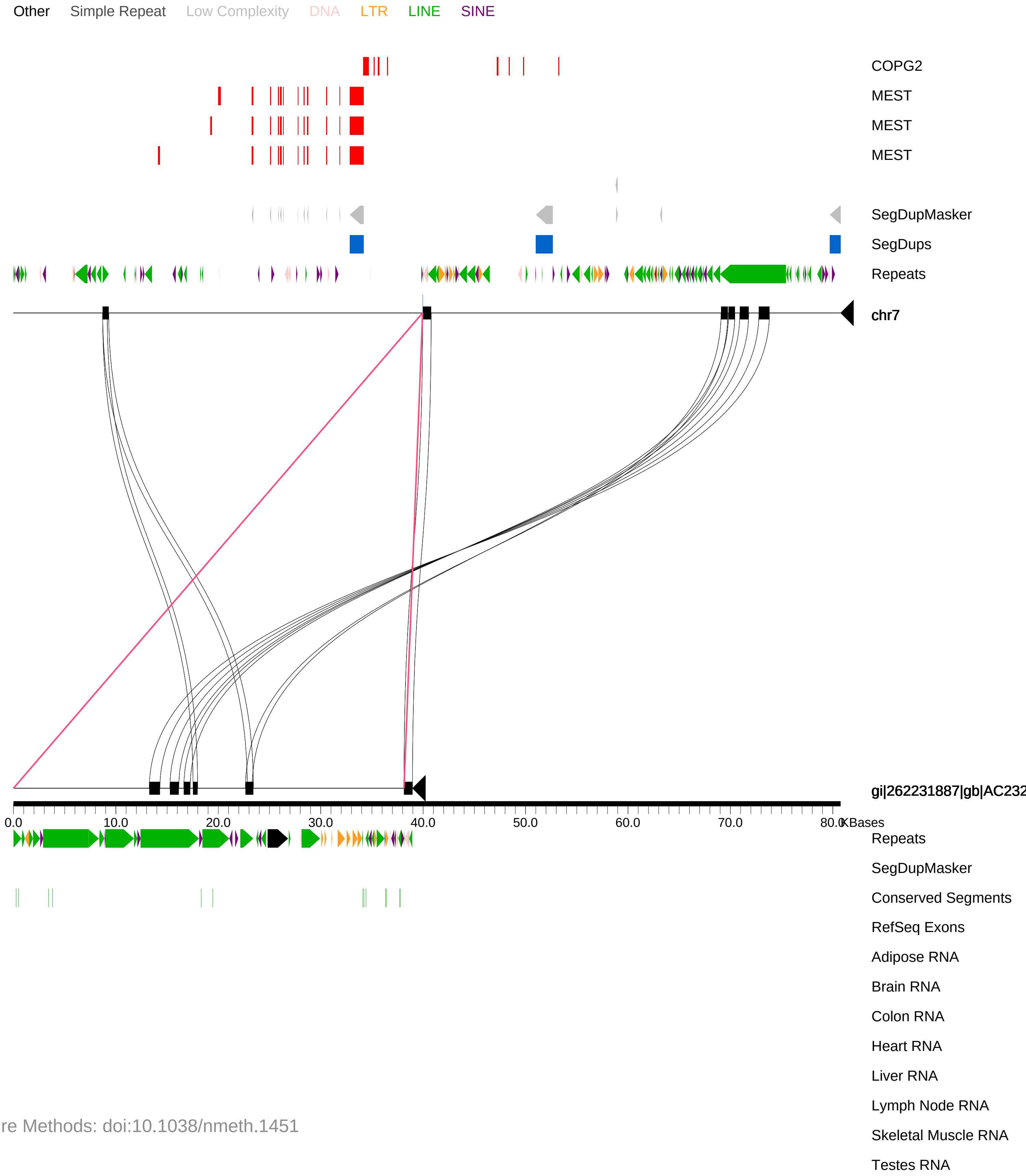
Insertion Size: 37813



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC232309.fa

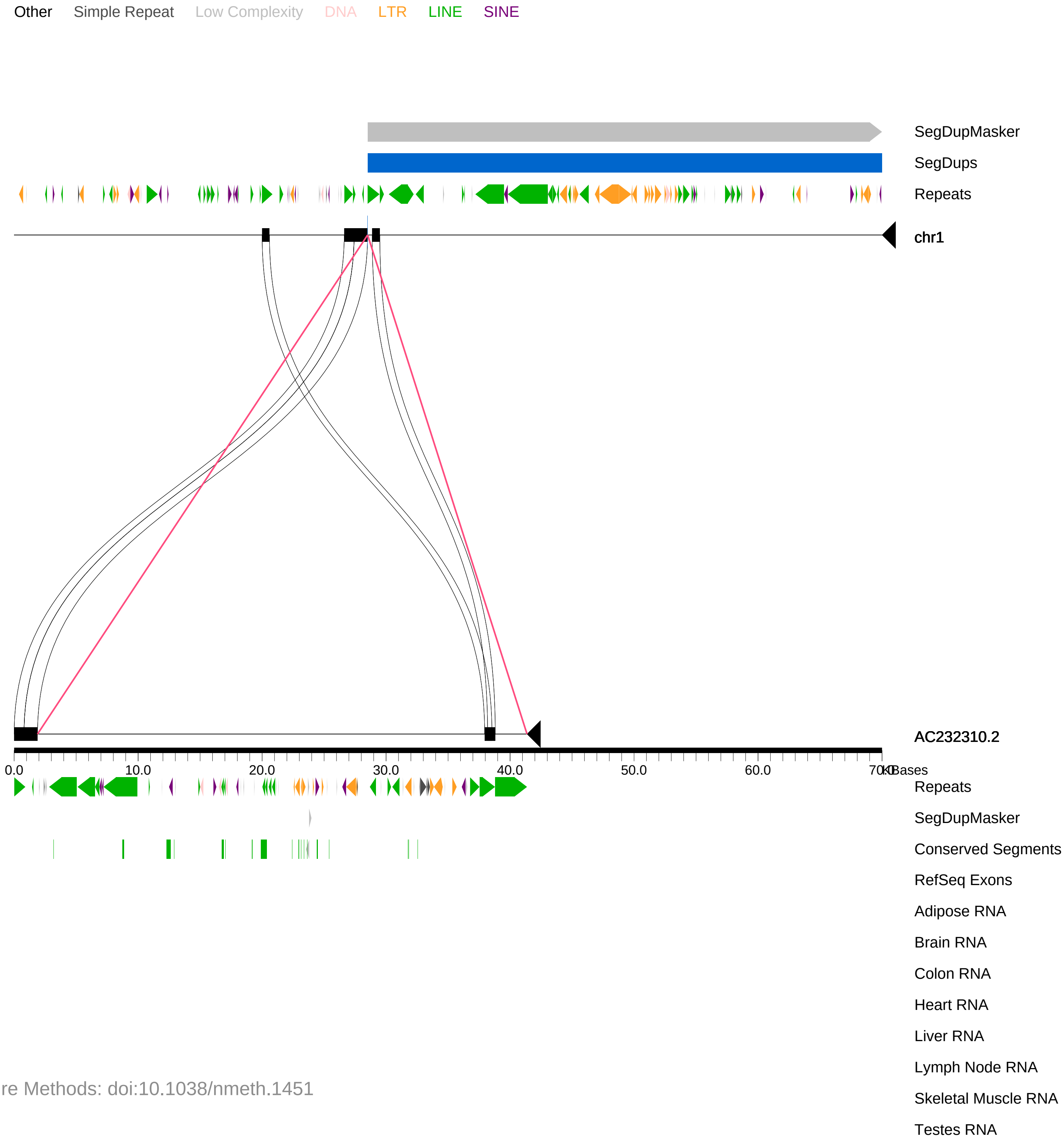
Insertion Size: 38130



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC232310.rc.fa

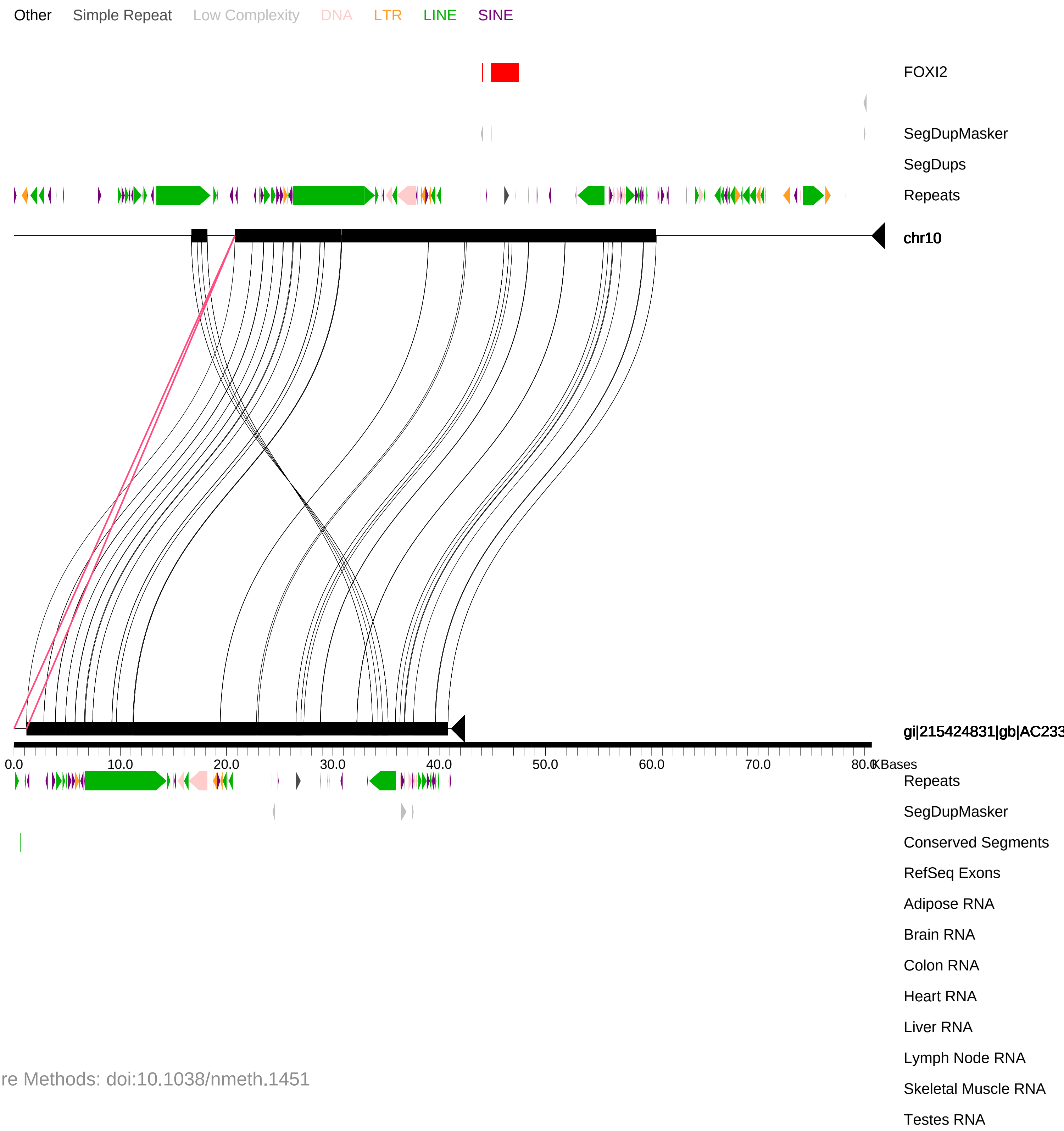
Insertion Size: 39484



Nature Methods: doi:10.1038/nmeth.1451

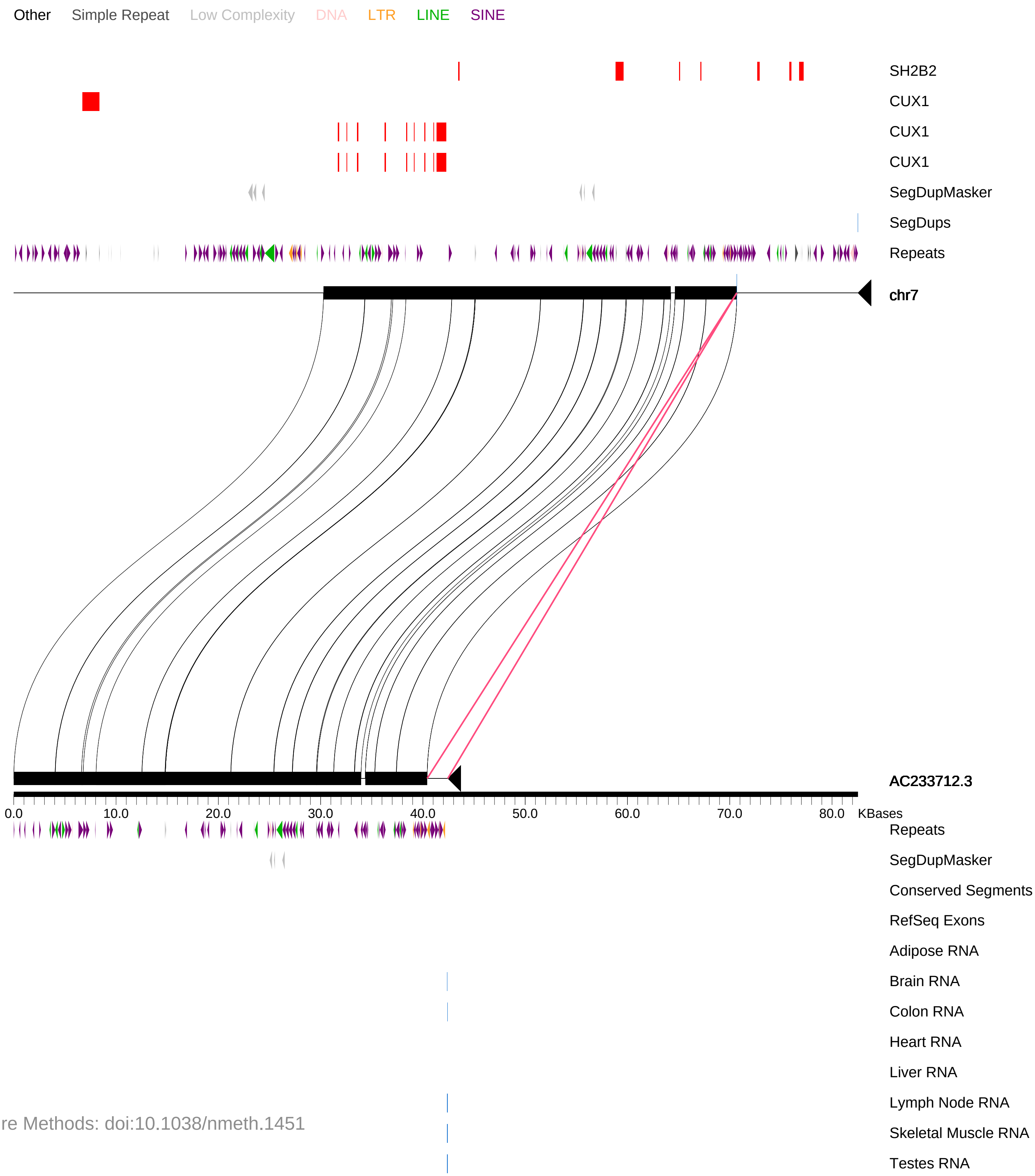
Clone file = AC233314.fa

Insertion Size: 1179



Clone file = AC233712.rc.fa

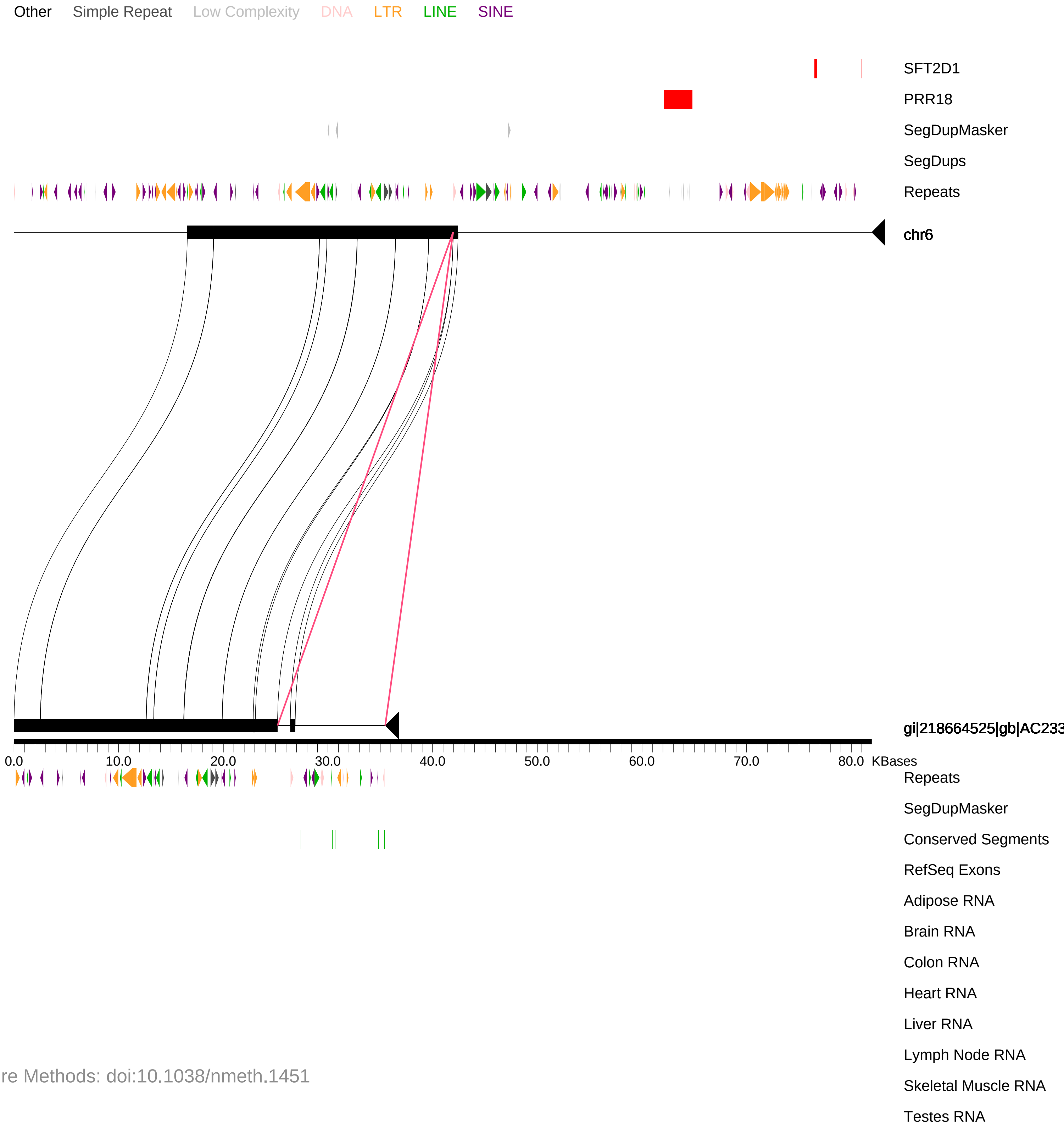
Insertion Size: 2017



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC233714.fa

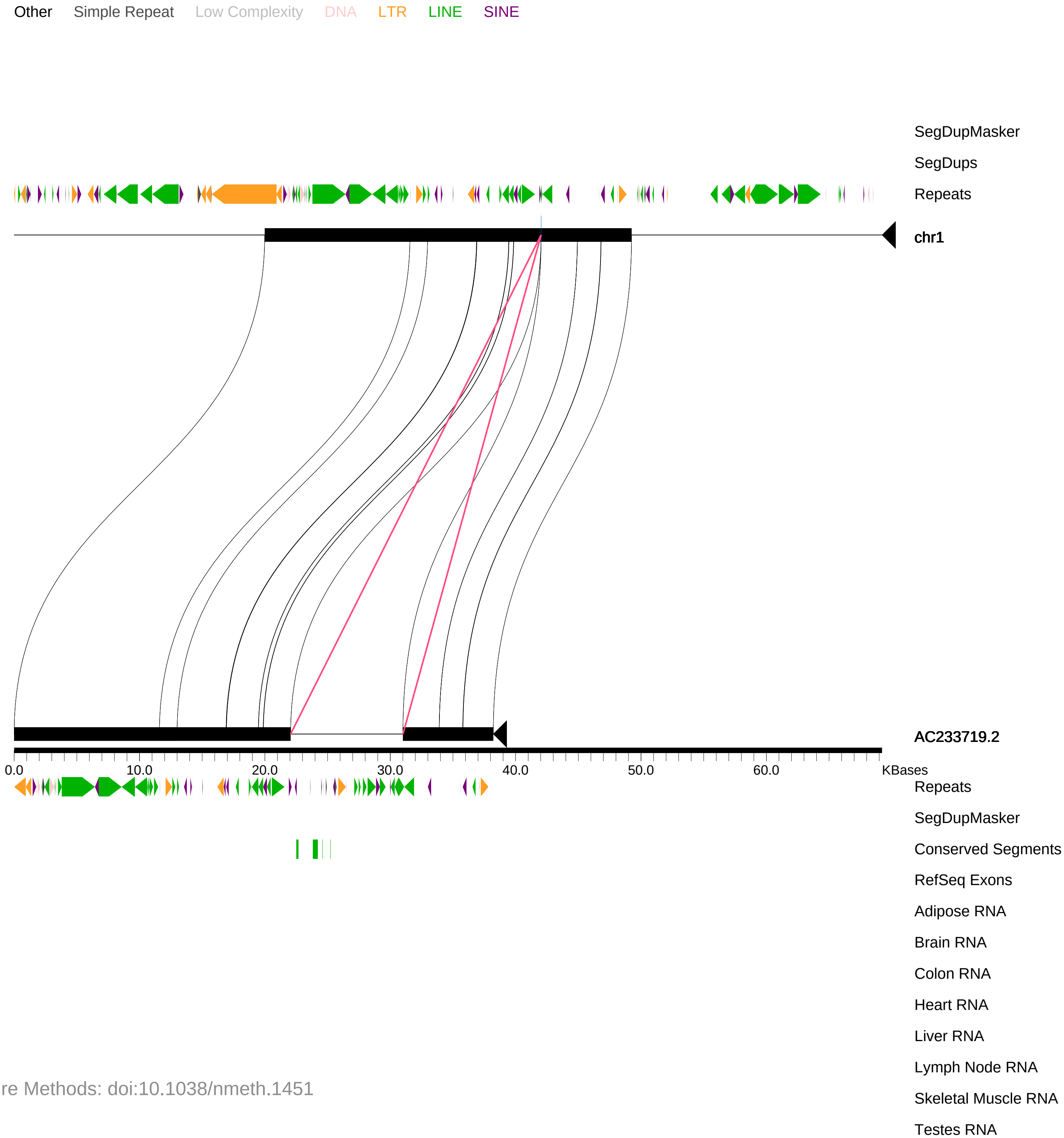
Insertion Size: 10288



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC233719.rc.fa

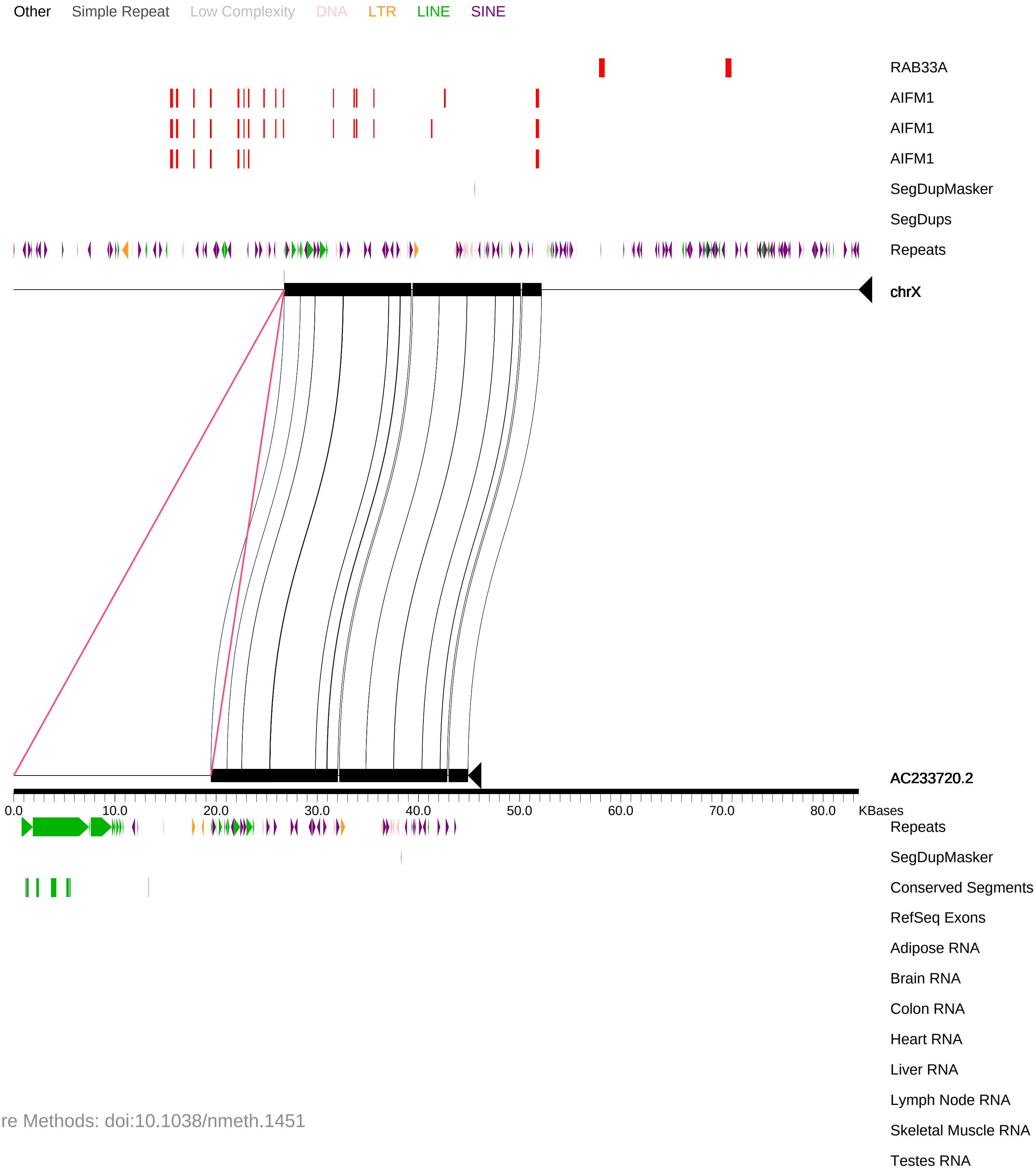
Insertion Size: 8961



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC233720.rc.fa

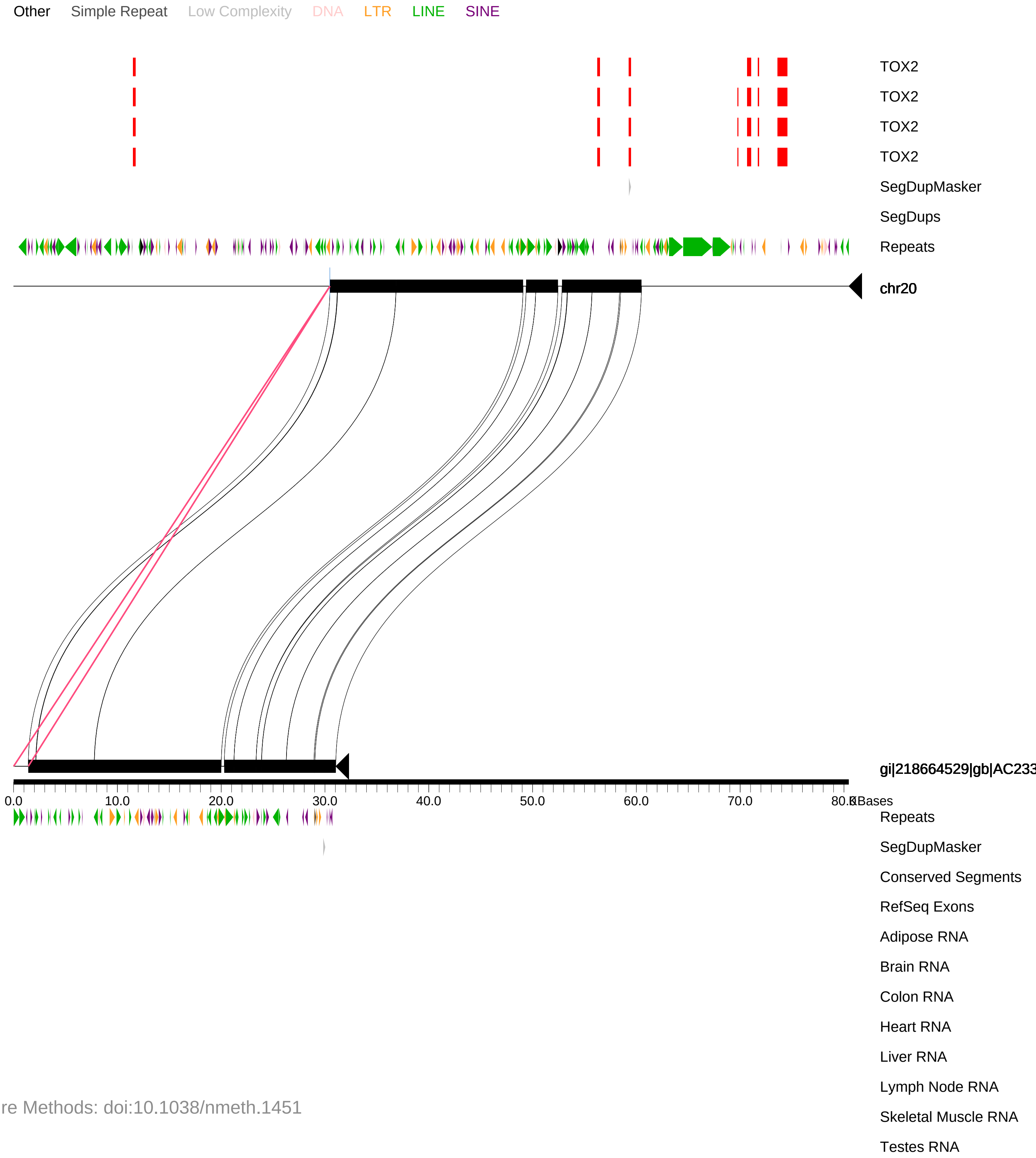
Insertion Size: 19486



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC233721.fa

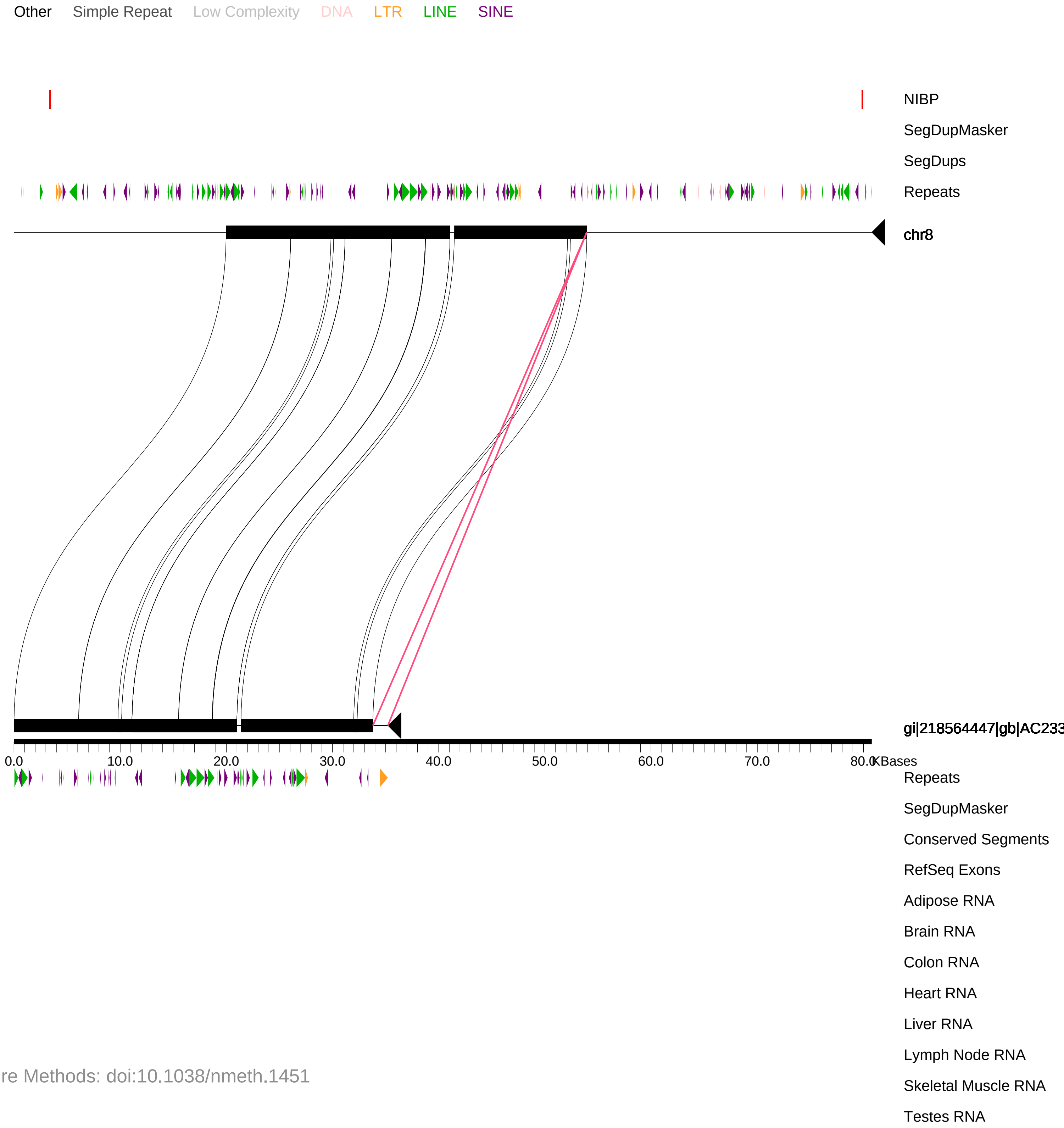
Insertion Size: 1415



Nature Methods: doi:10.1038/nmeth.1451

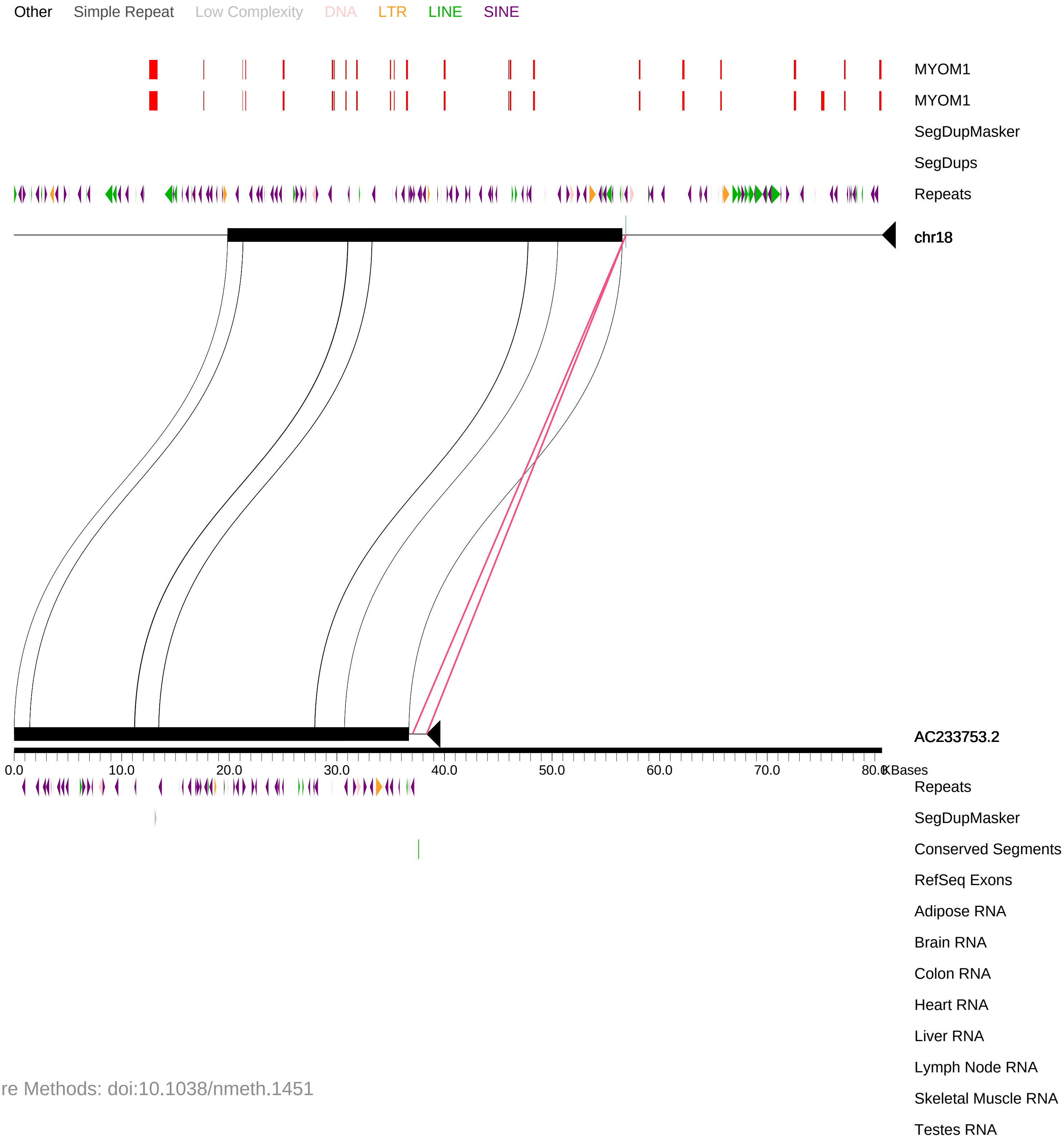
Clone file = AC233722.fa

Insertion Size: 1411



Clone file = AC233753.rc.fa

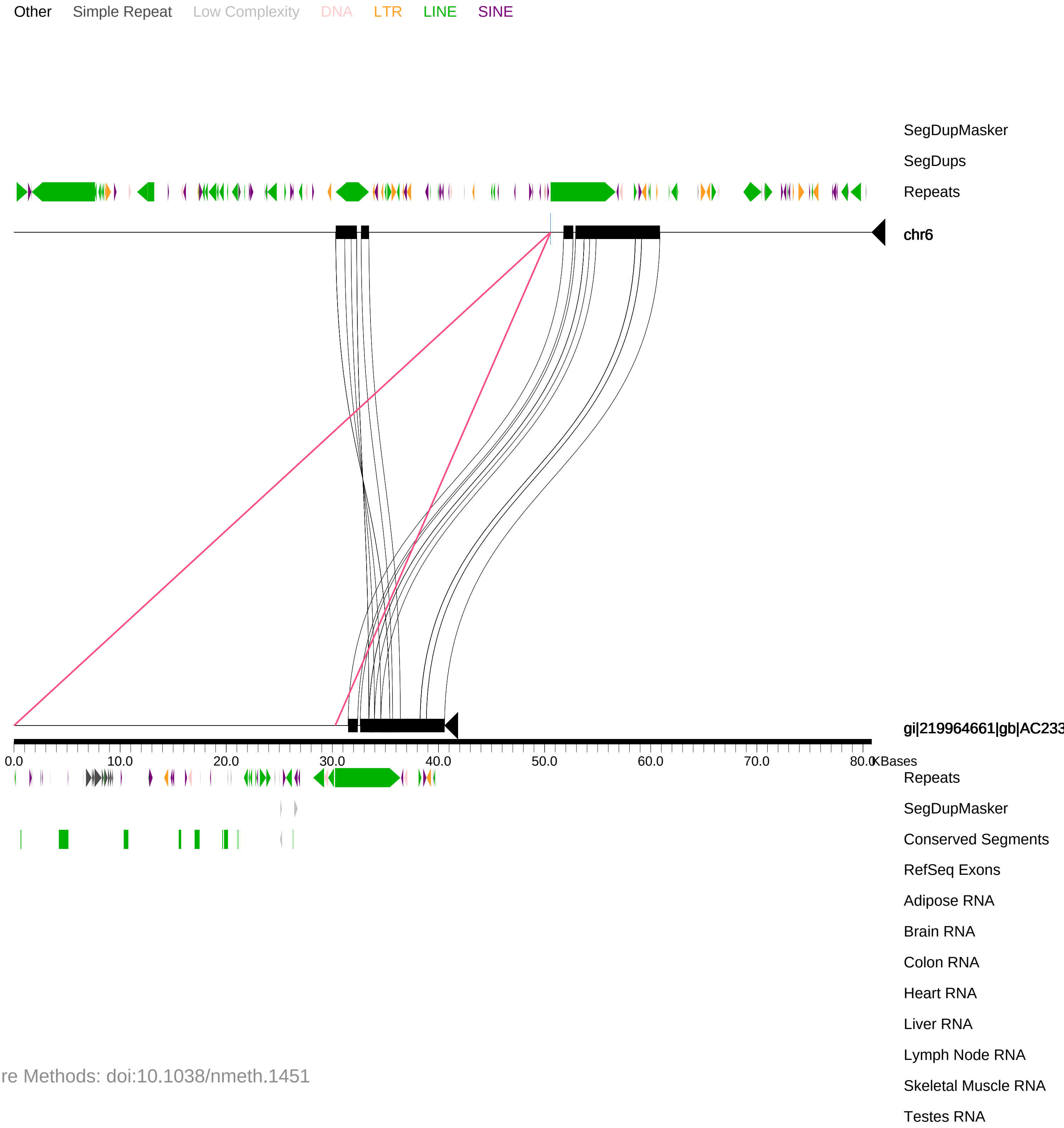
Insertion Size: 1311



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC233754.fa

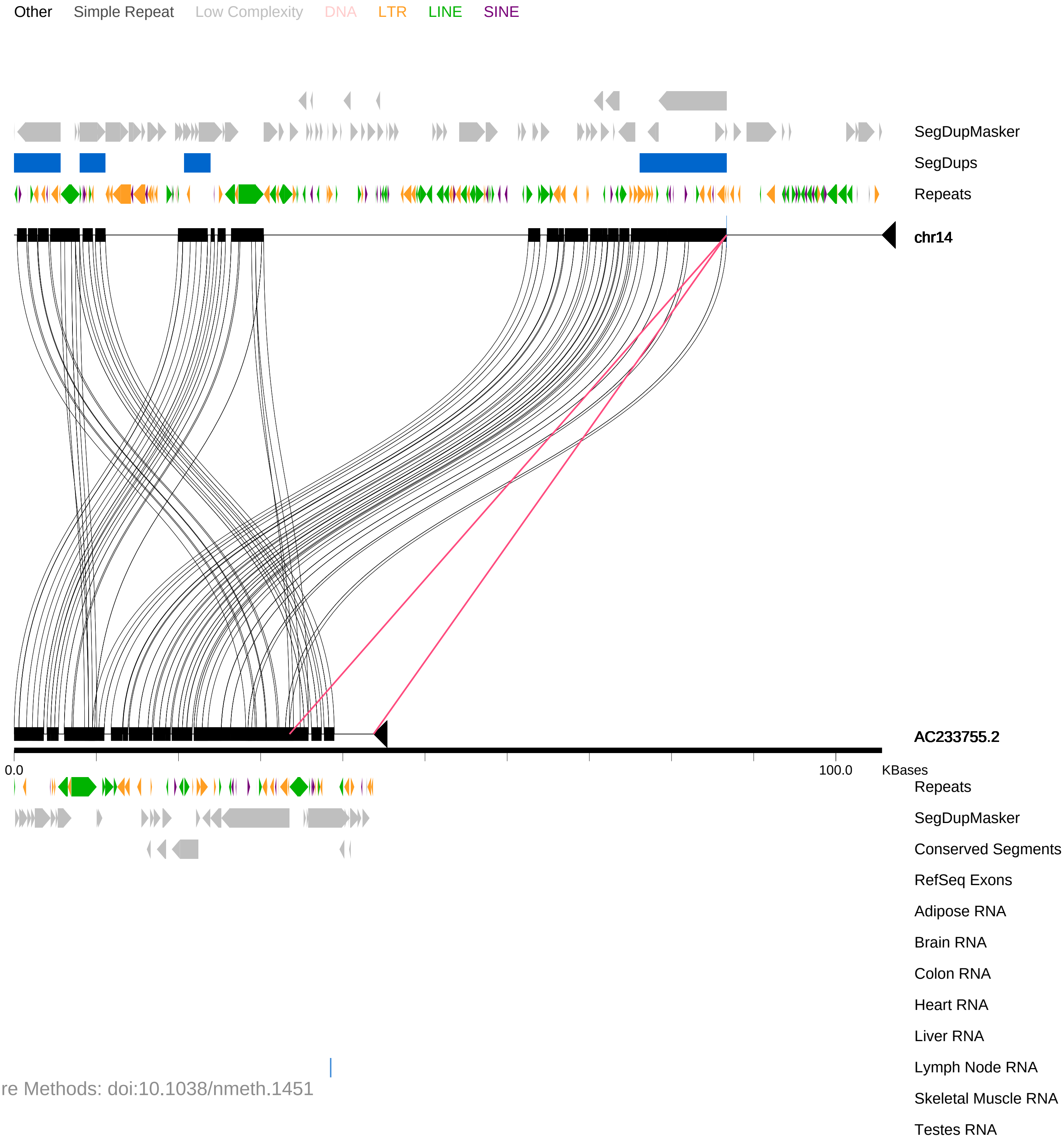
Insertion Size: 30268



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC233755.rc.fa

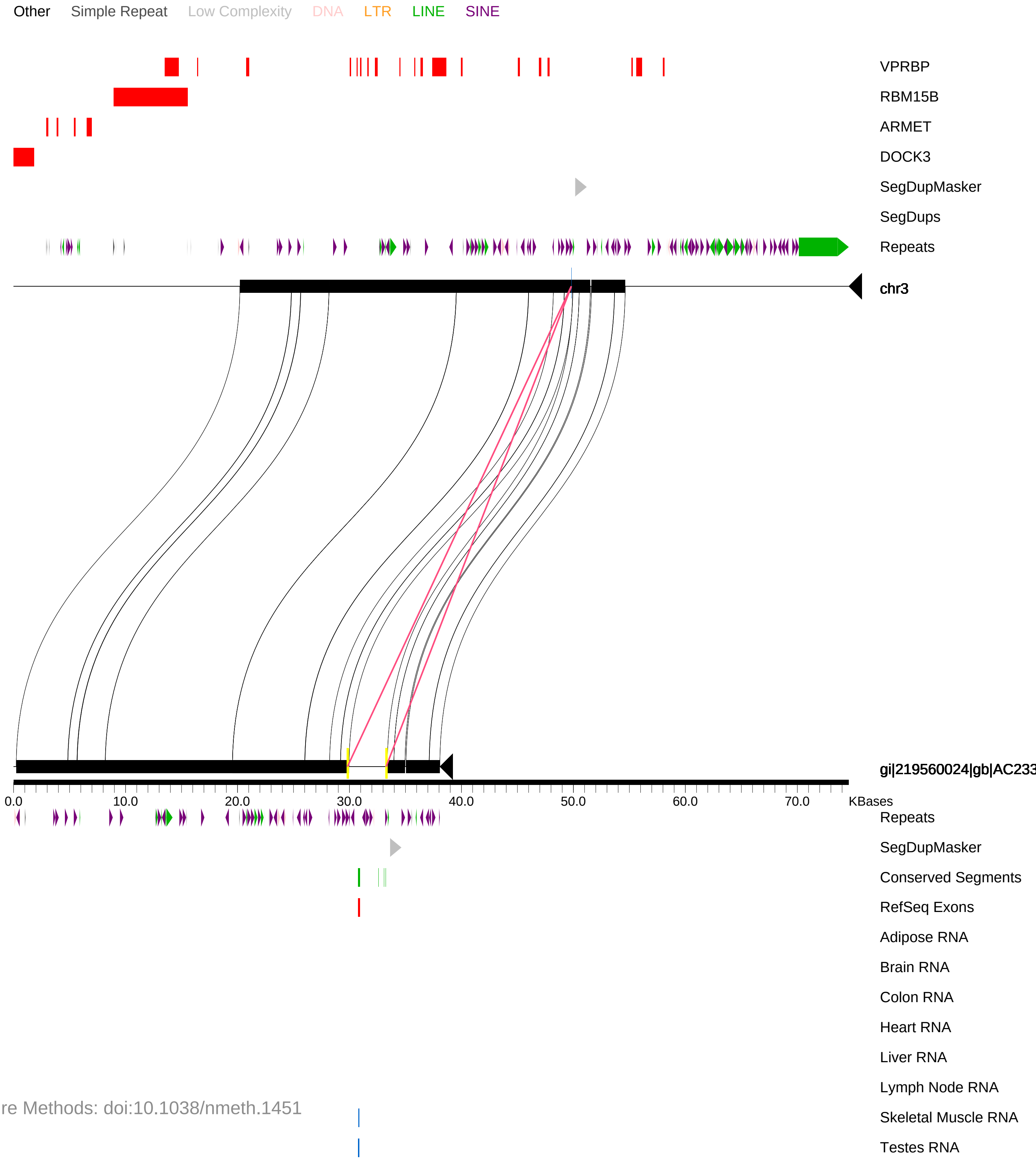
Insertion Size: 10229



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC233756.fa

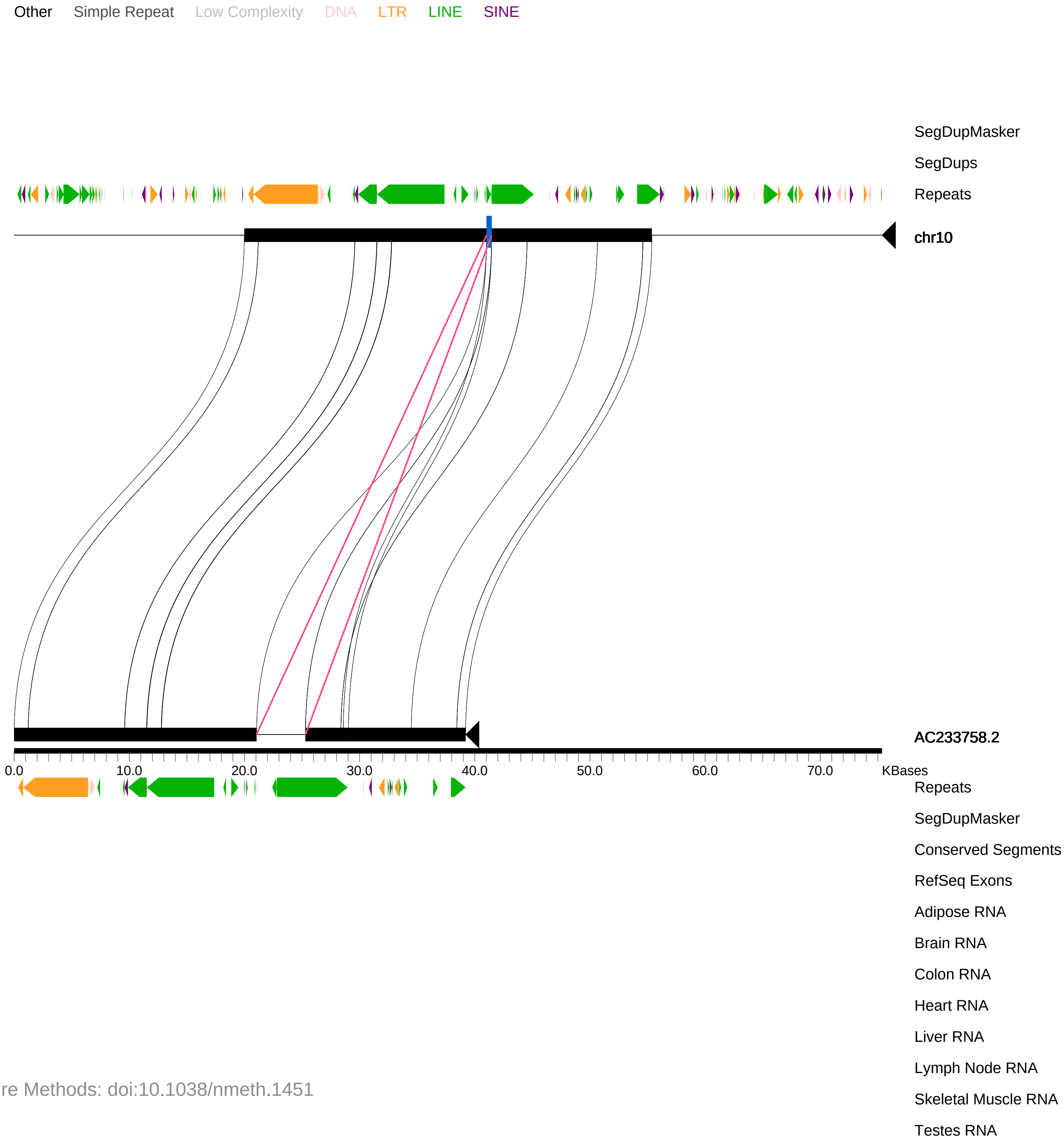
Insertion Size: 3456



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC233758.rc.fa

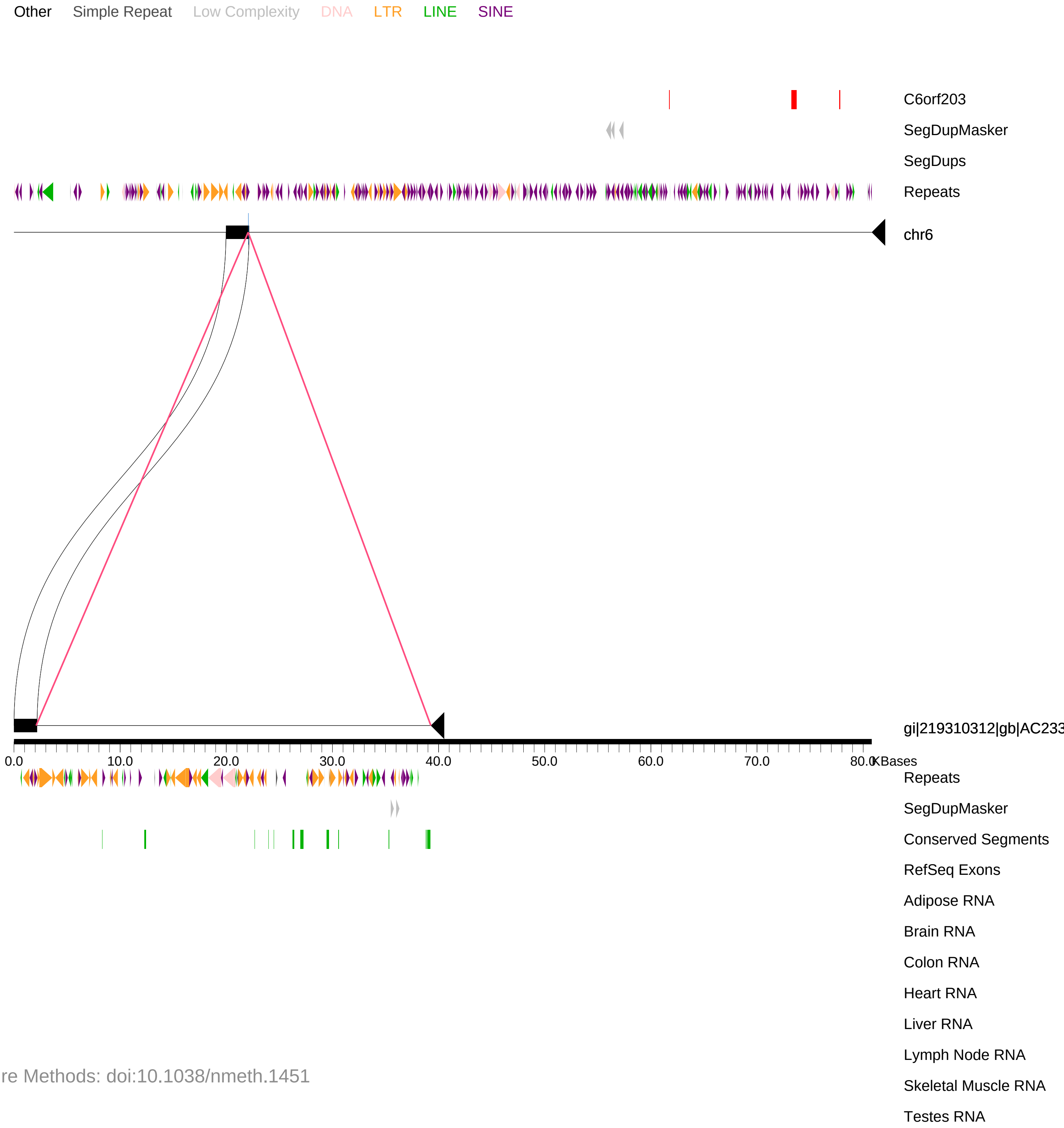
Insertion Size: 4258



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC233764.fa

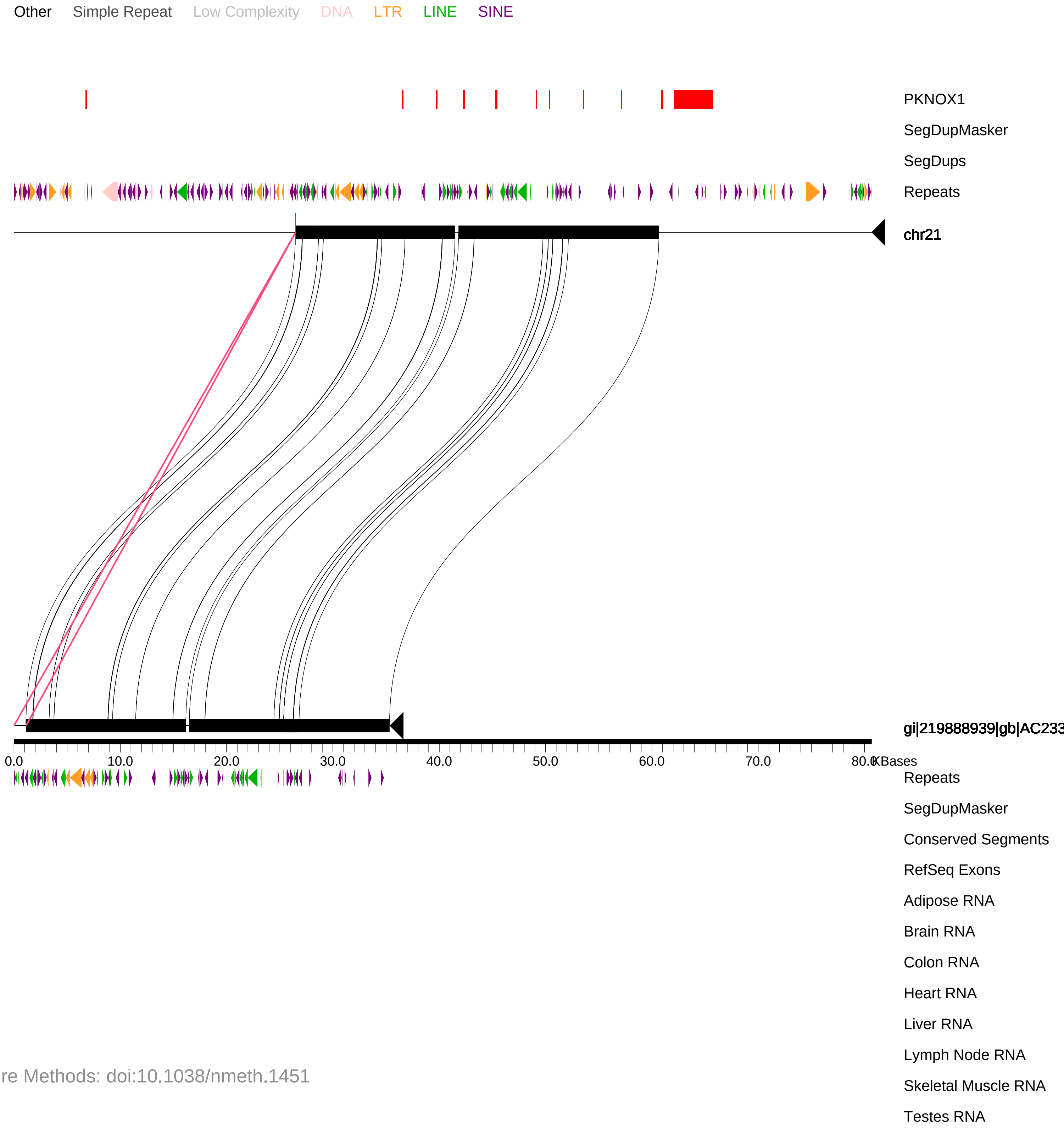
Insertion Size: 37184



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC233768.fa

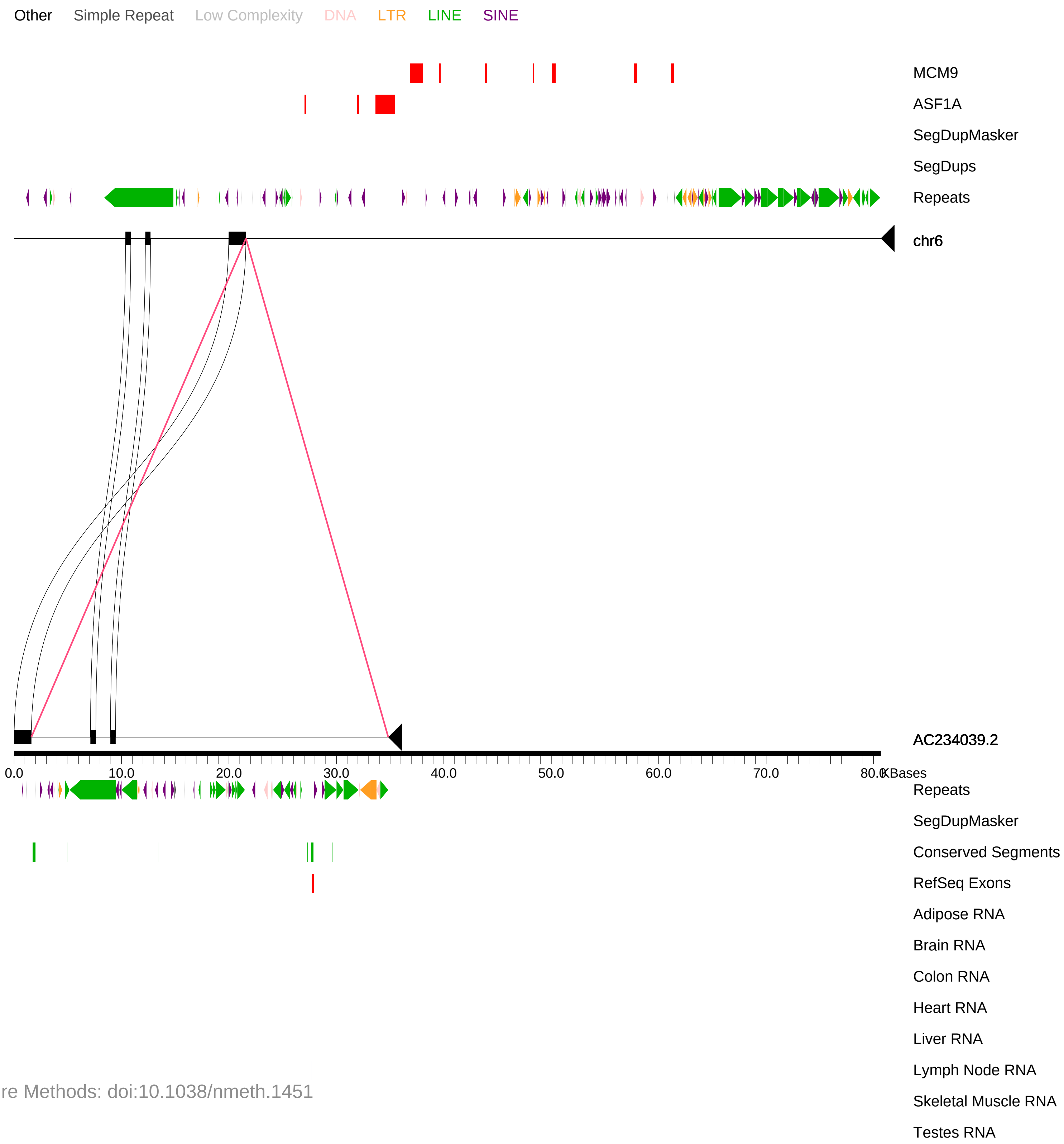
Insertion Size: 1147



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC234039.rc.fa

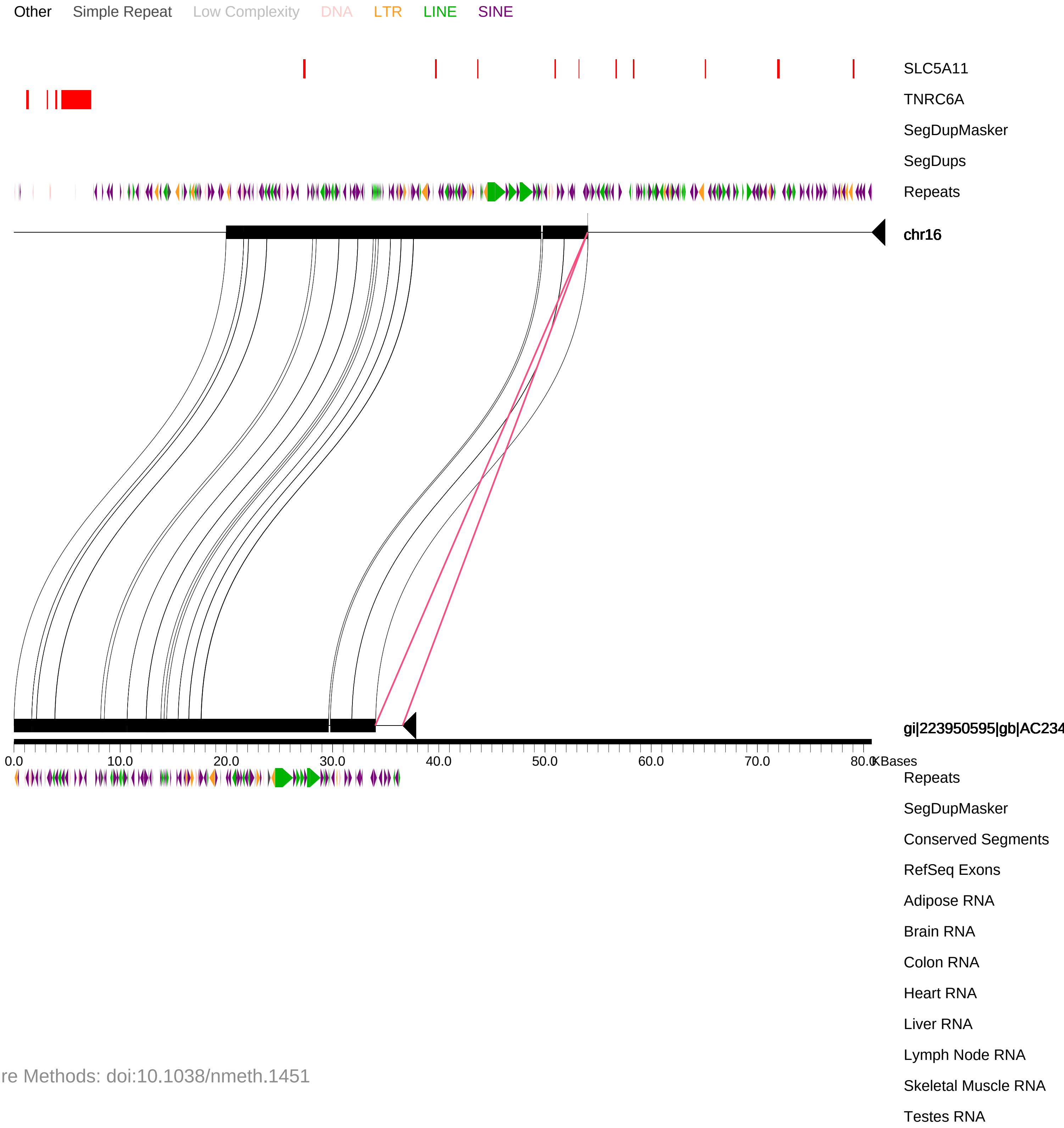
Insertion Size: 33226



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC234142.fa

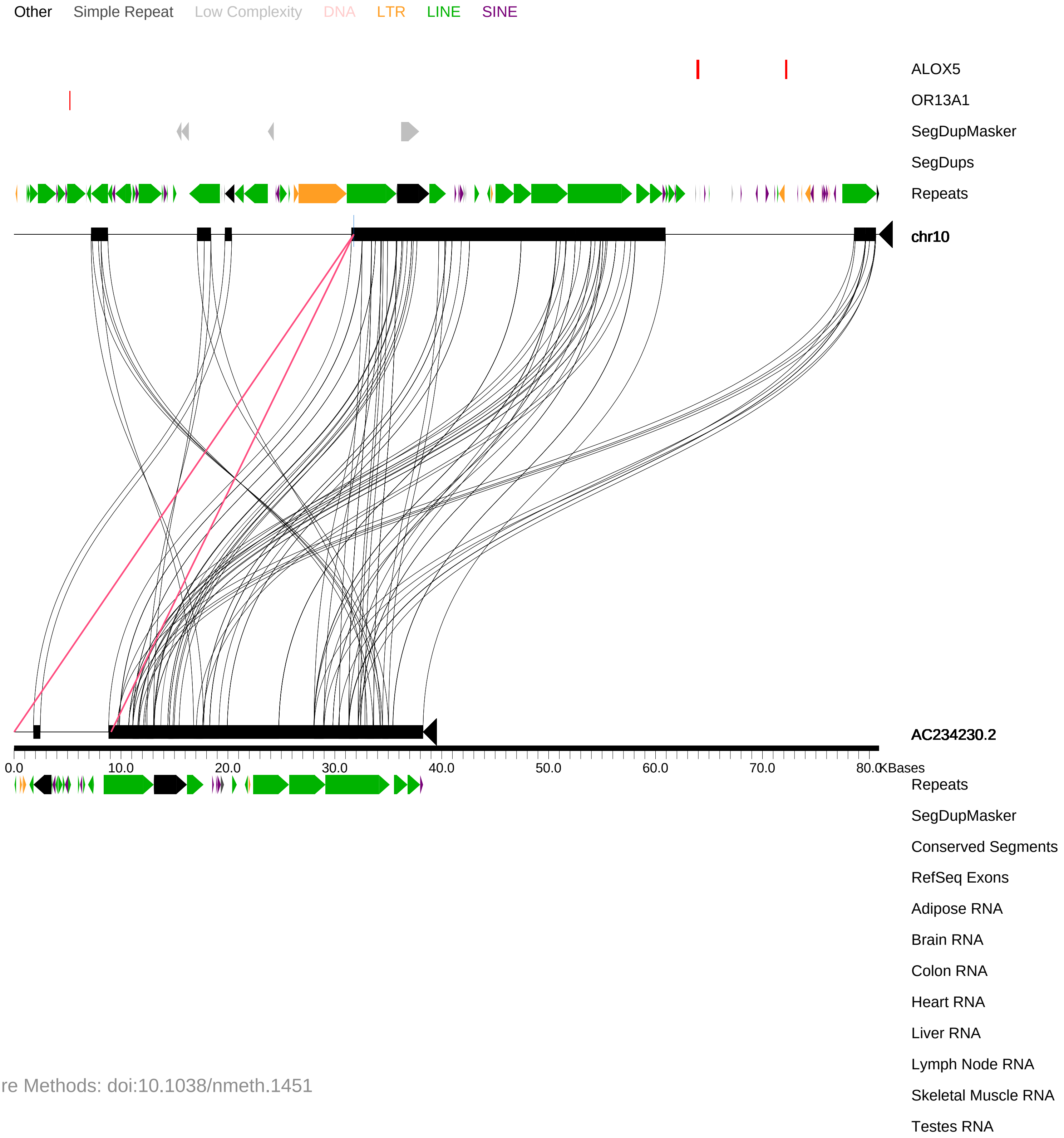
Insertion Size: 2576



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC234230.rc.fa

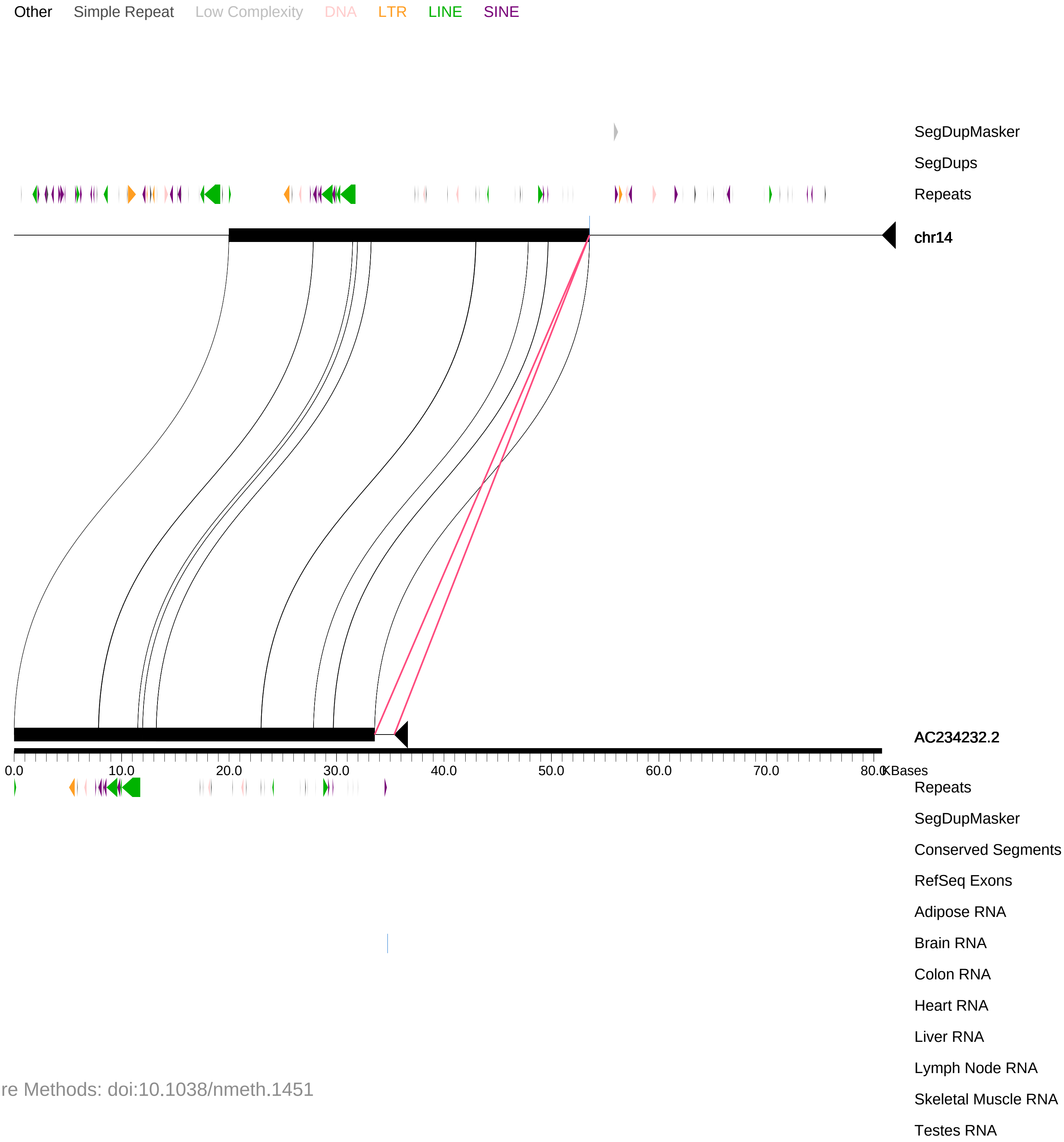
Insertion Size: 9071



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC234232.rc.fa

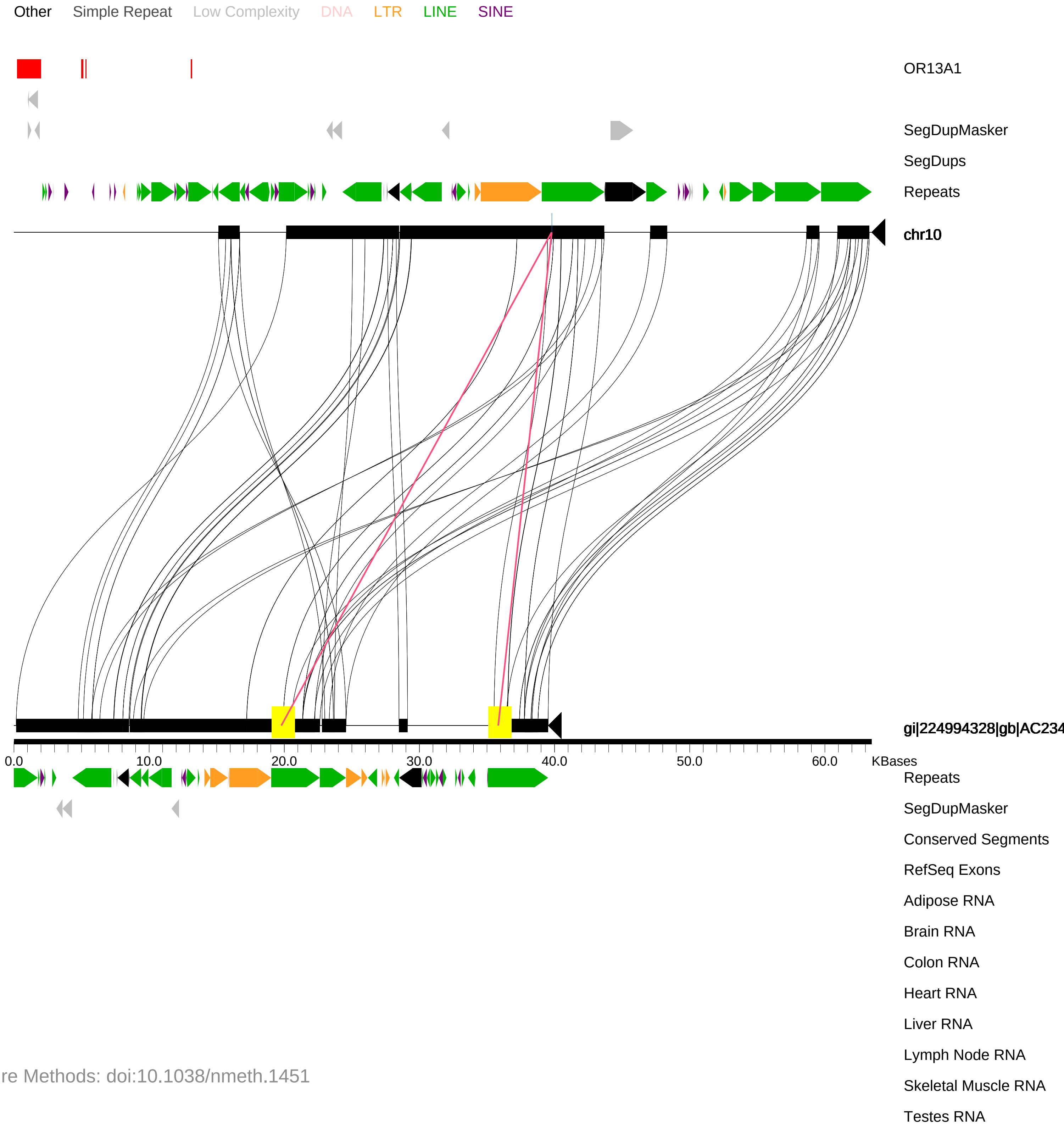
Insertion Size: 1828



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC234305.fa

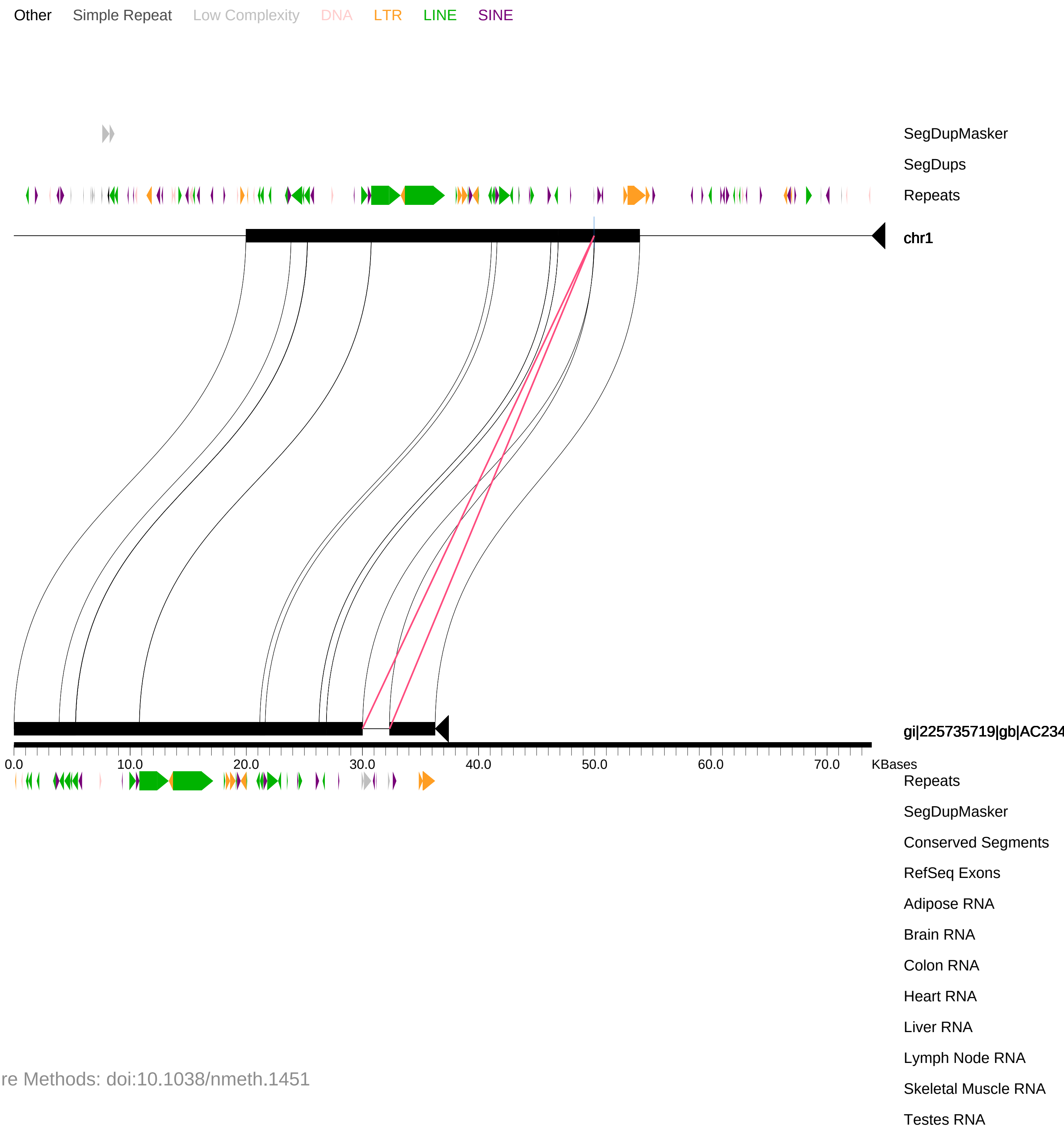
Insertion Size: 16052



Nature Methods: doi:10.1038/nmeth.1451

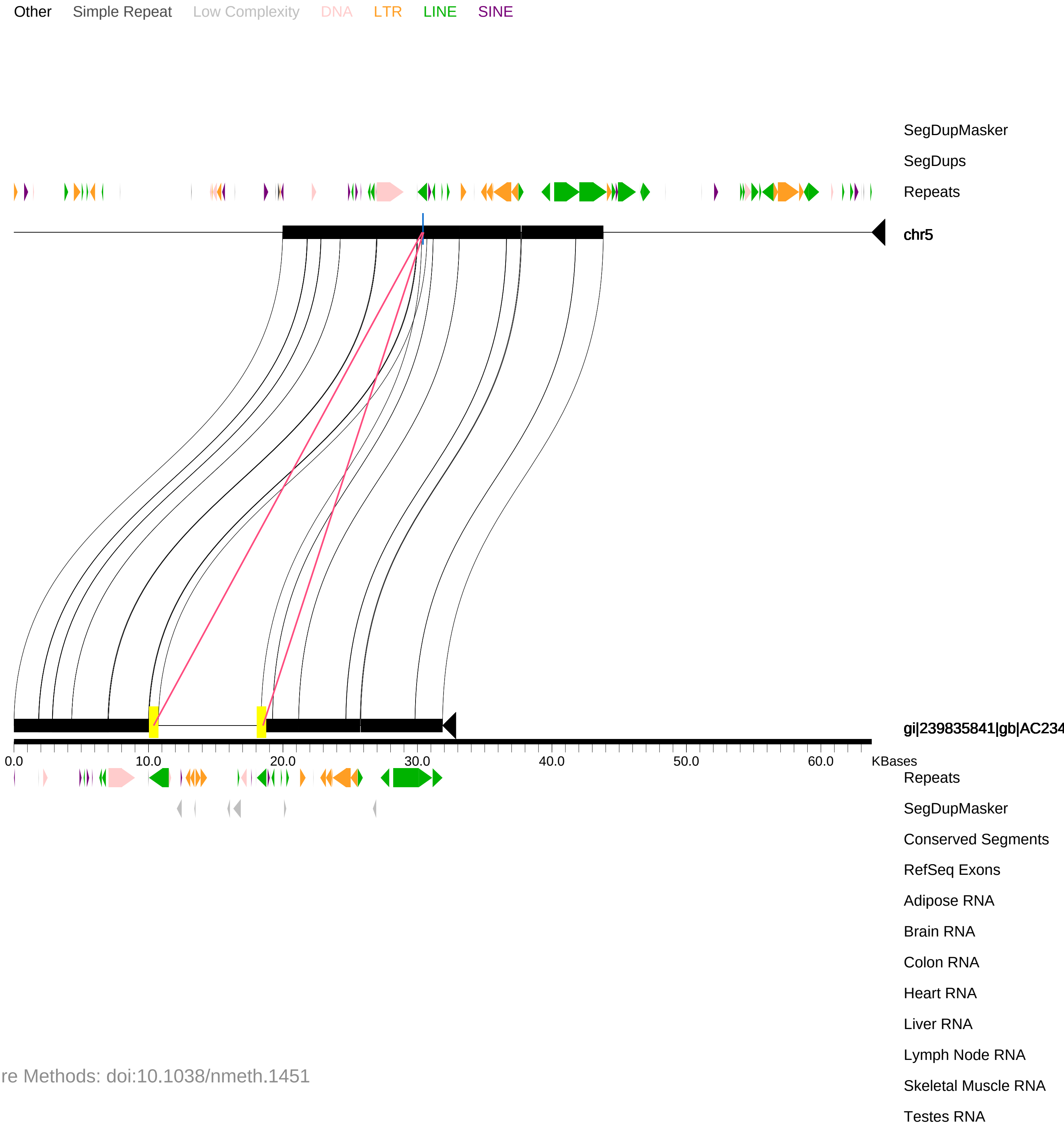
Clone file = AC234425.fa

Insertion Size: 2340



Clone file = AC234851.fa

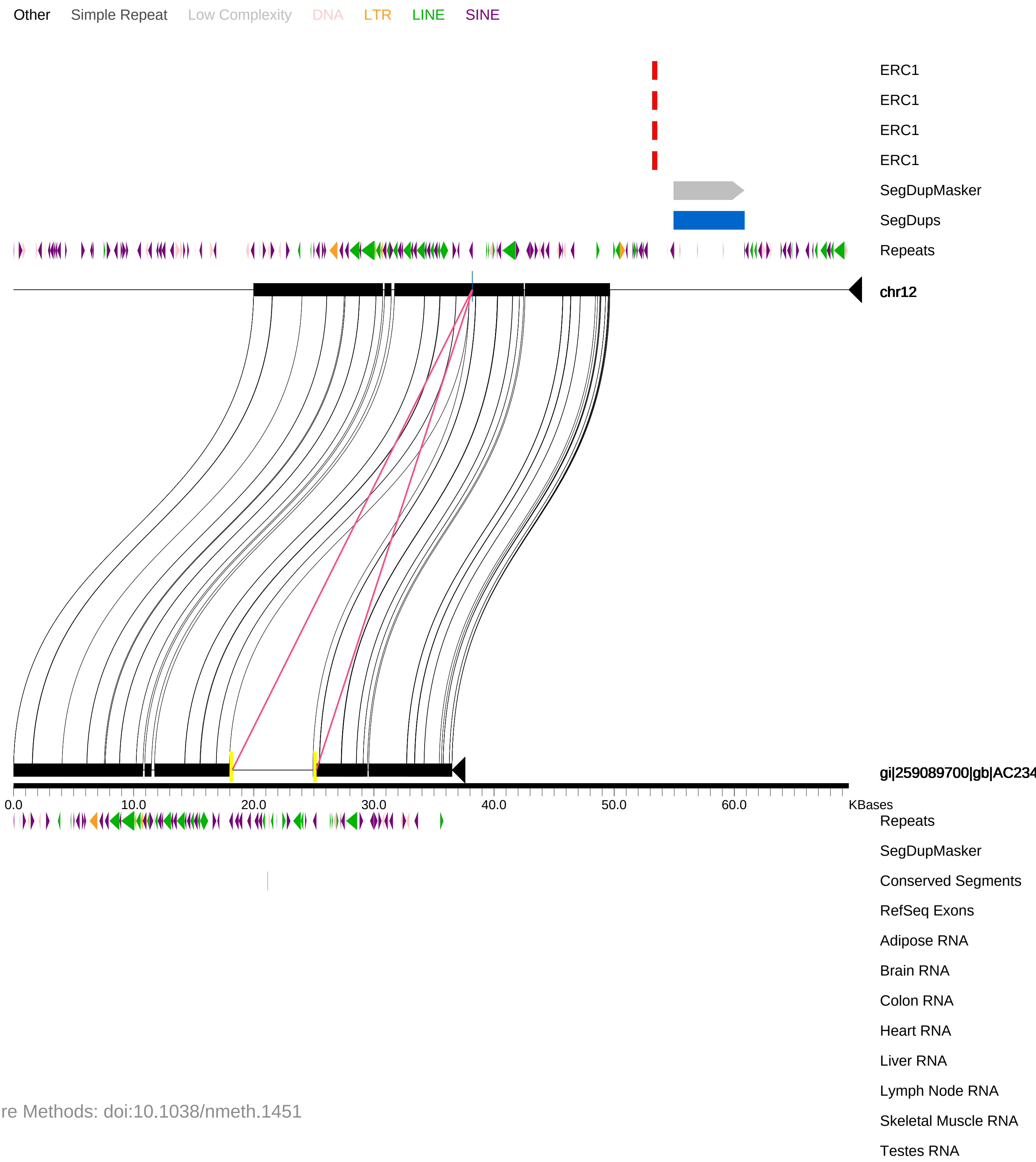
Insertion Size: 8133



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC234852.fa

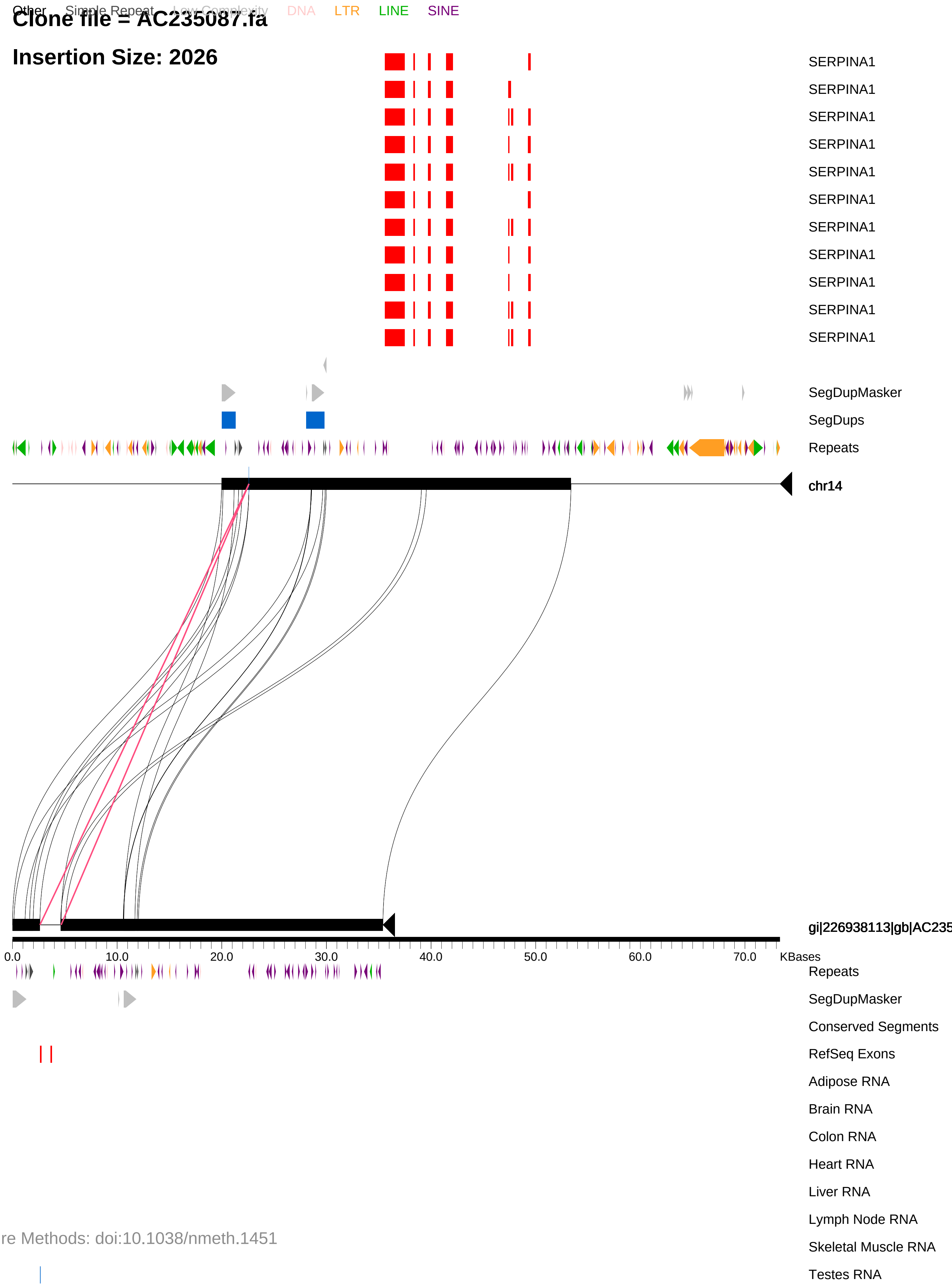
Insertion Size: 7000



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC235087.fa

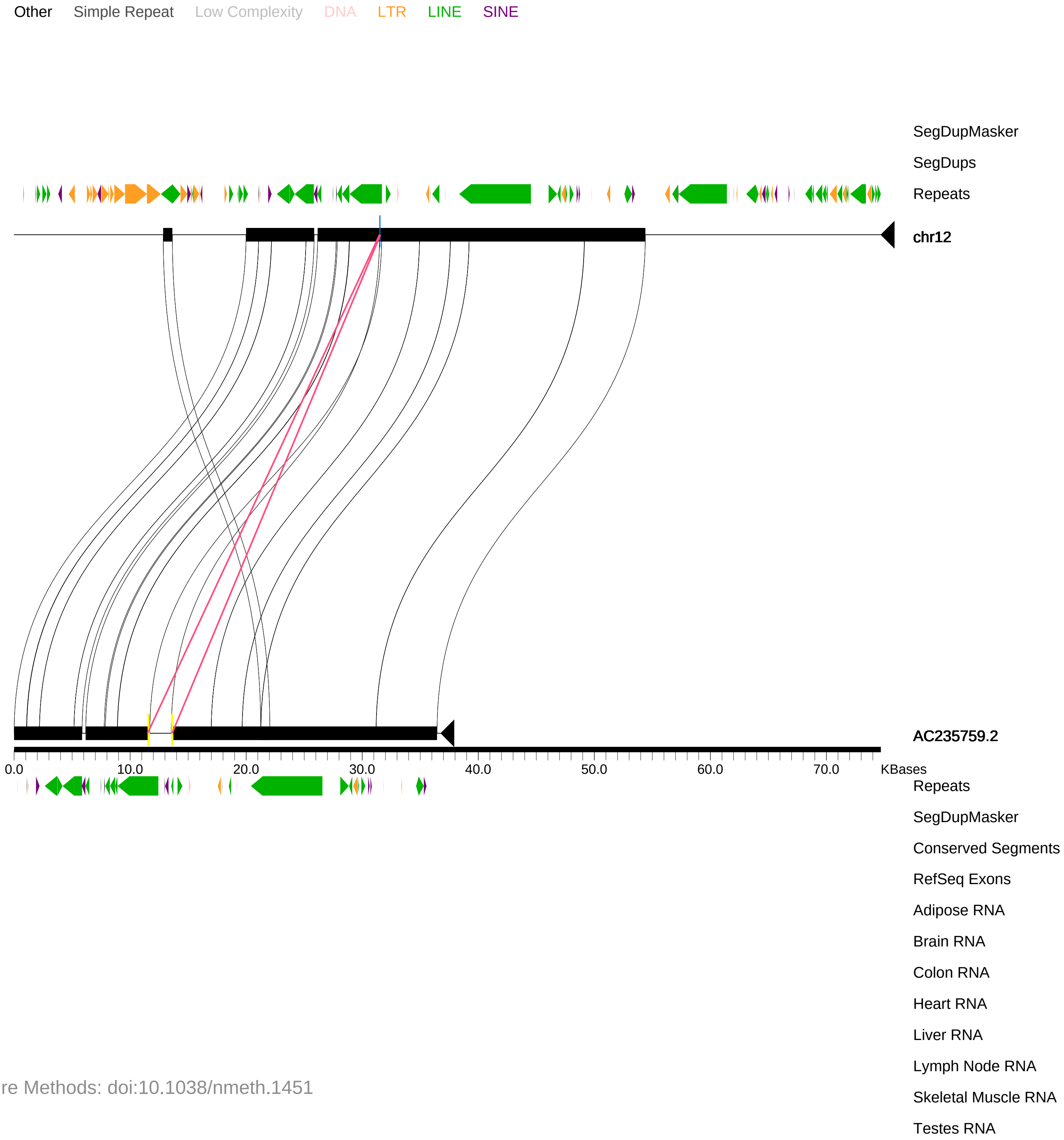
Insertion Size: 2026



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC235759.rc.fa

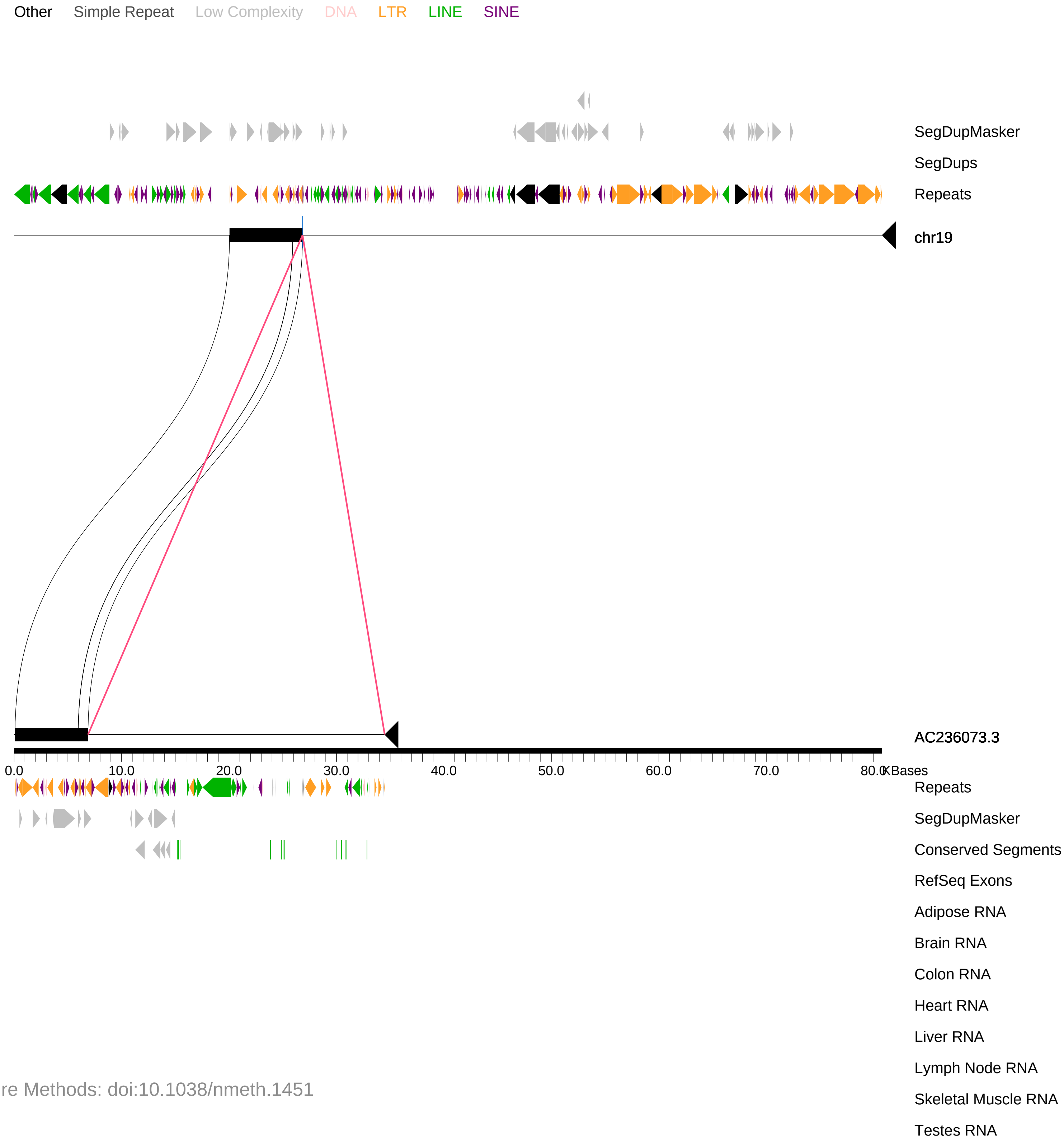
Insertion Size: 2113



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC236073.rc.fa

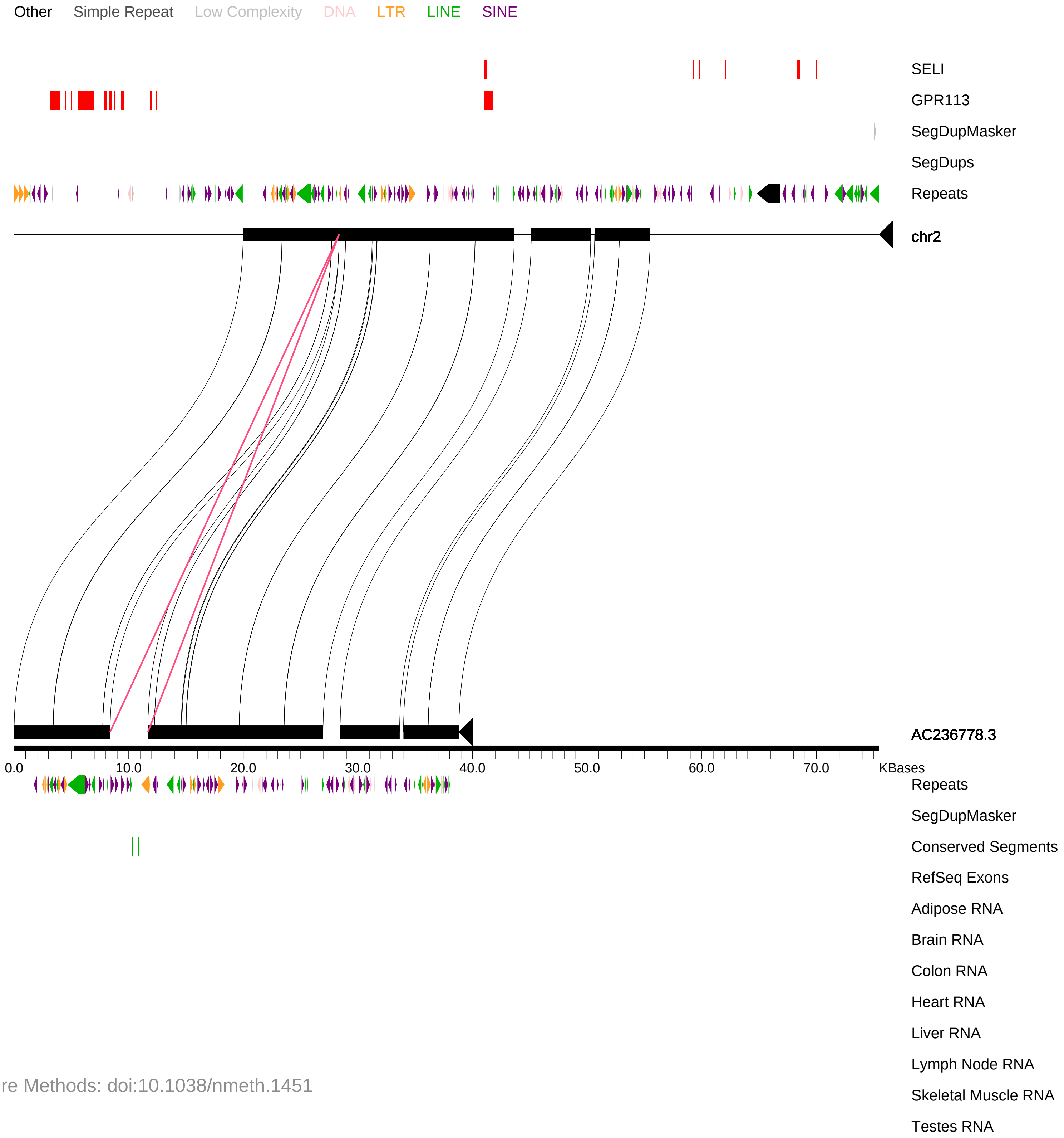
Insertion Size: 27613



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC236778.rc.fa

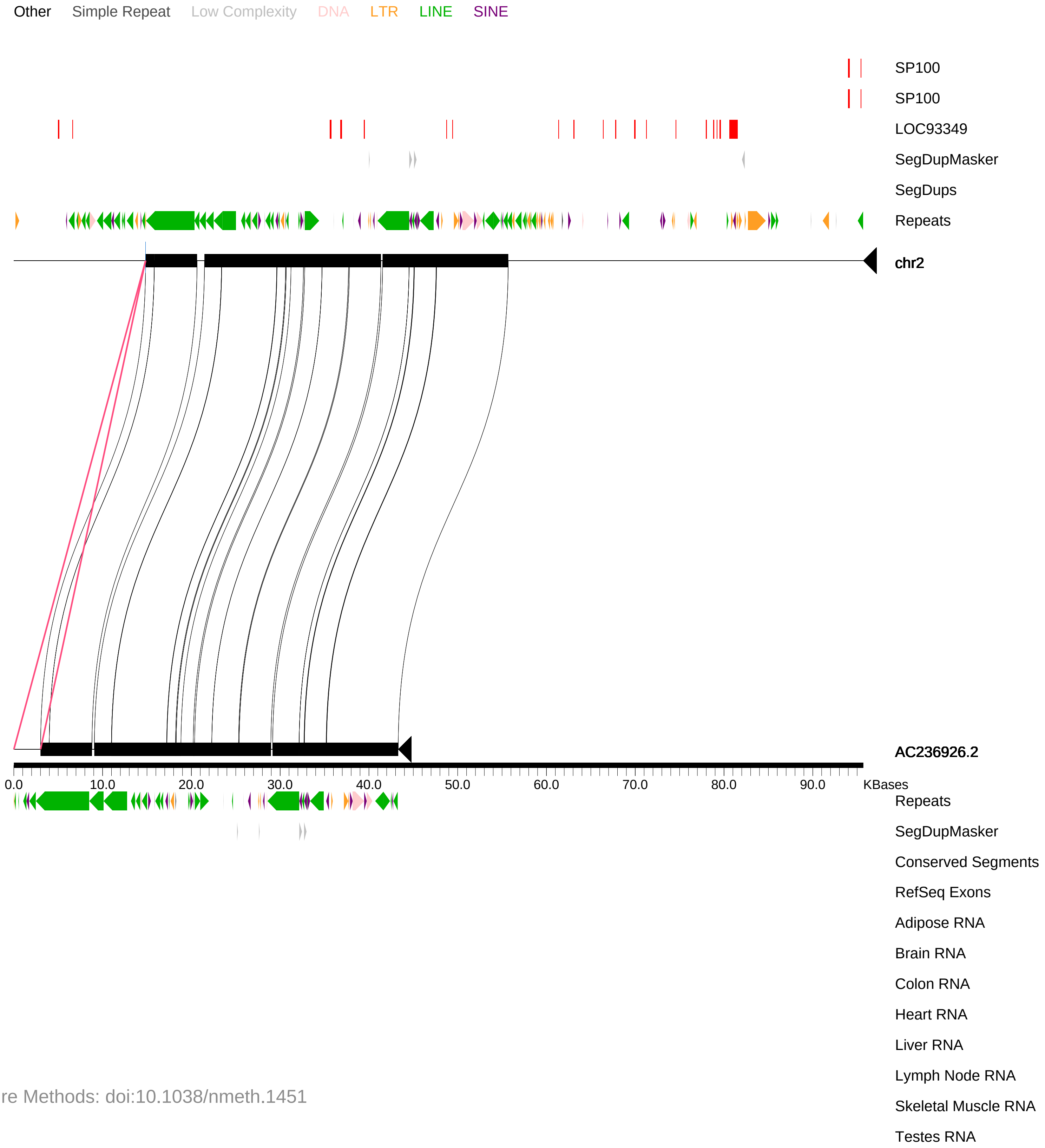
Insertion Size: 3305



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC236926.rc.fa

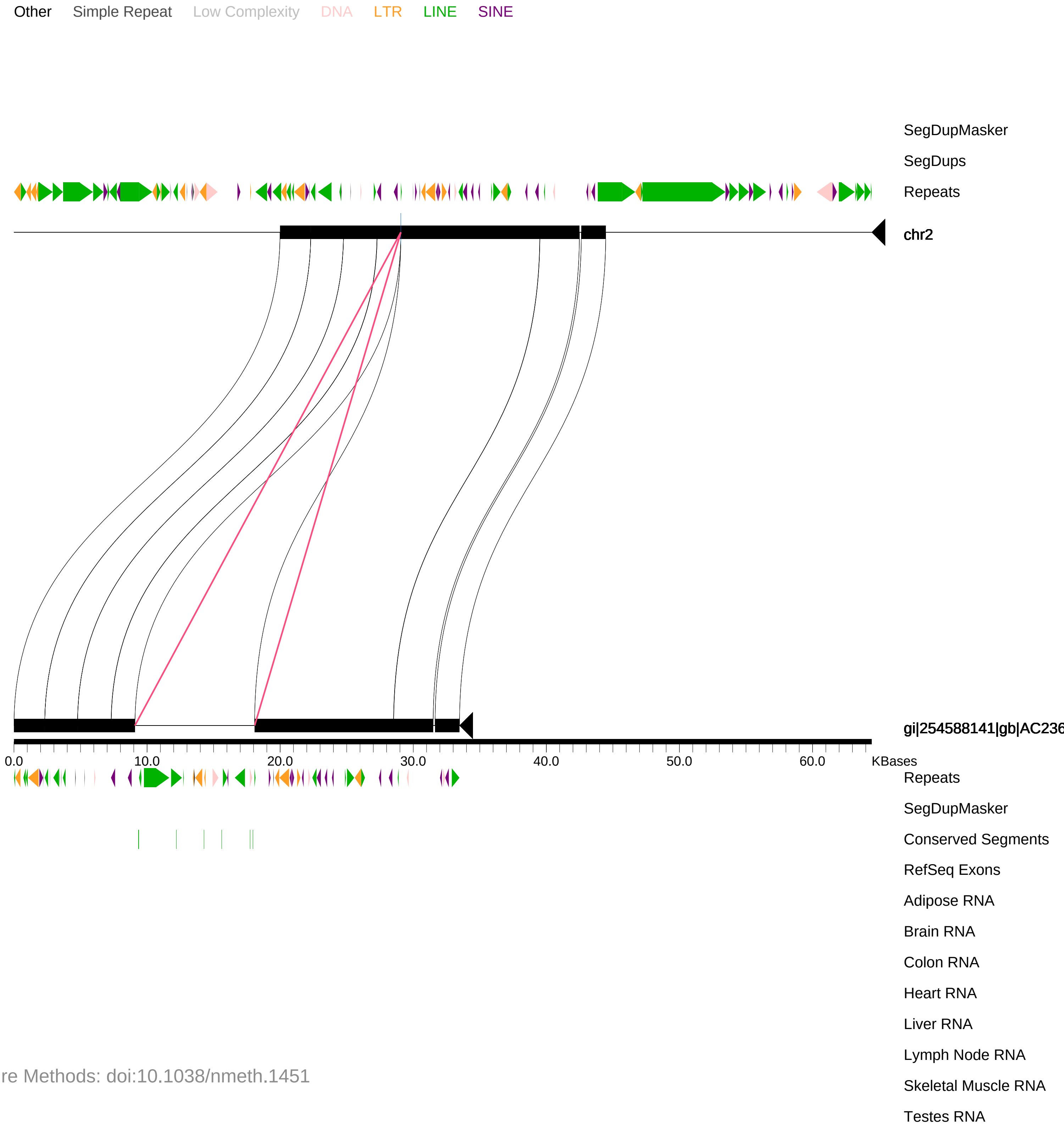
Insertion Size: 3008



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC236964.fa

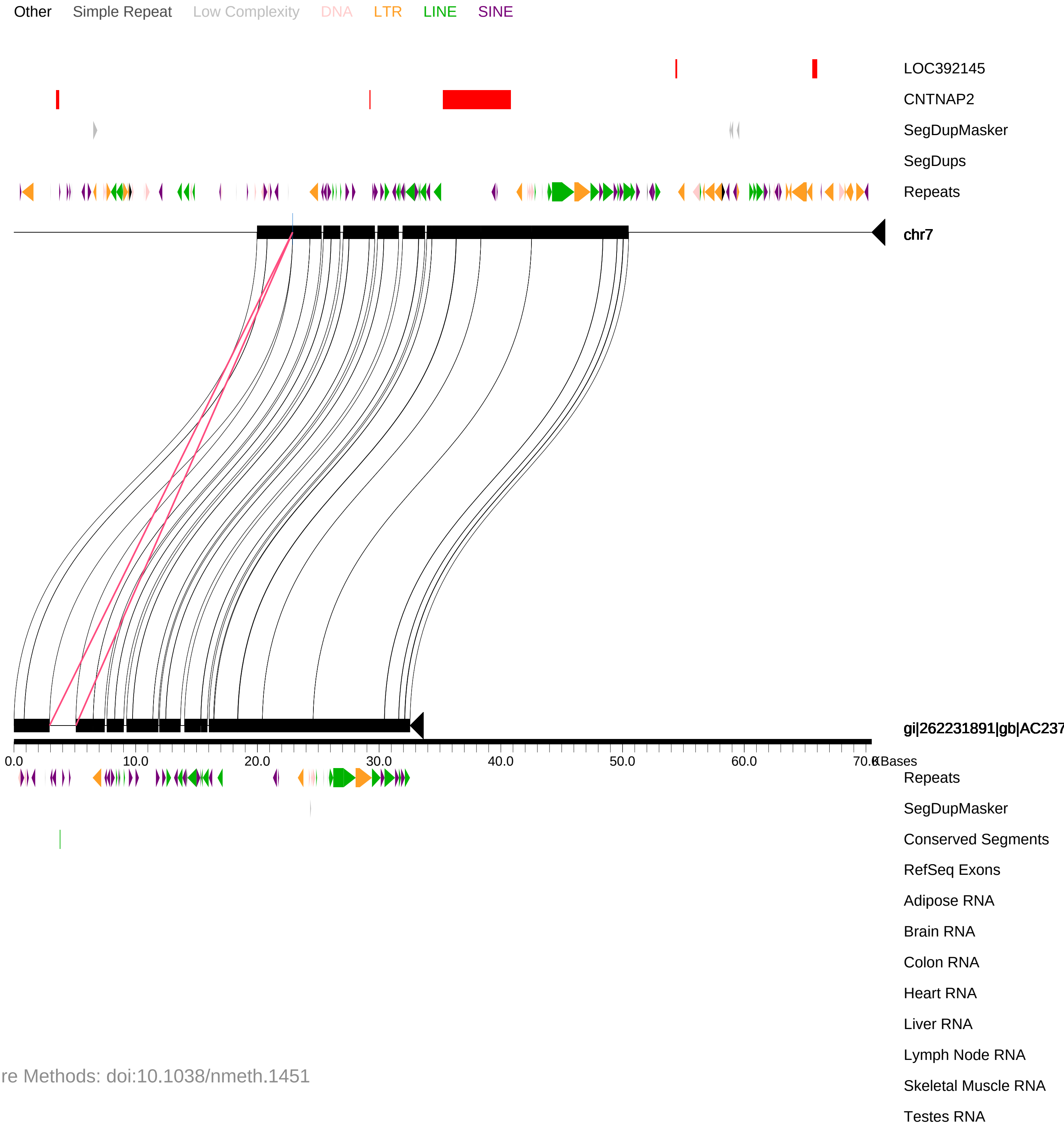
Insertion Size: 8998



Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC237106.fa

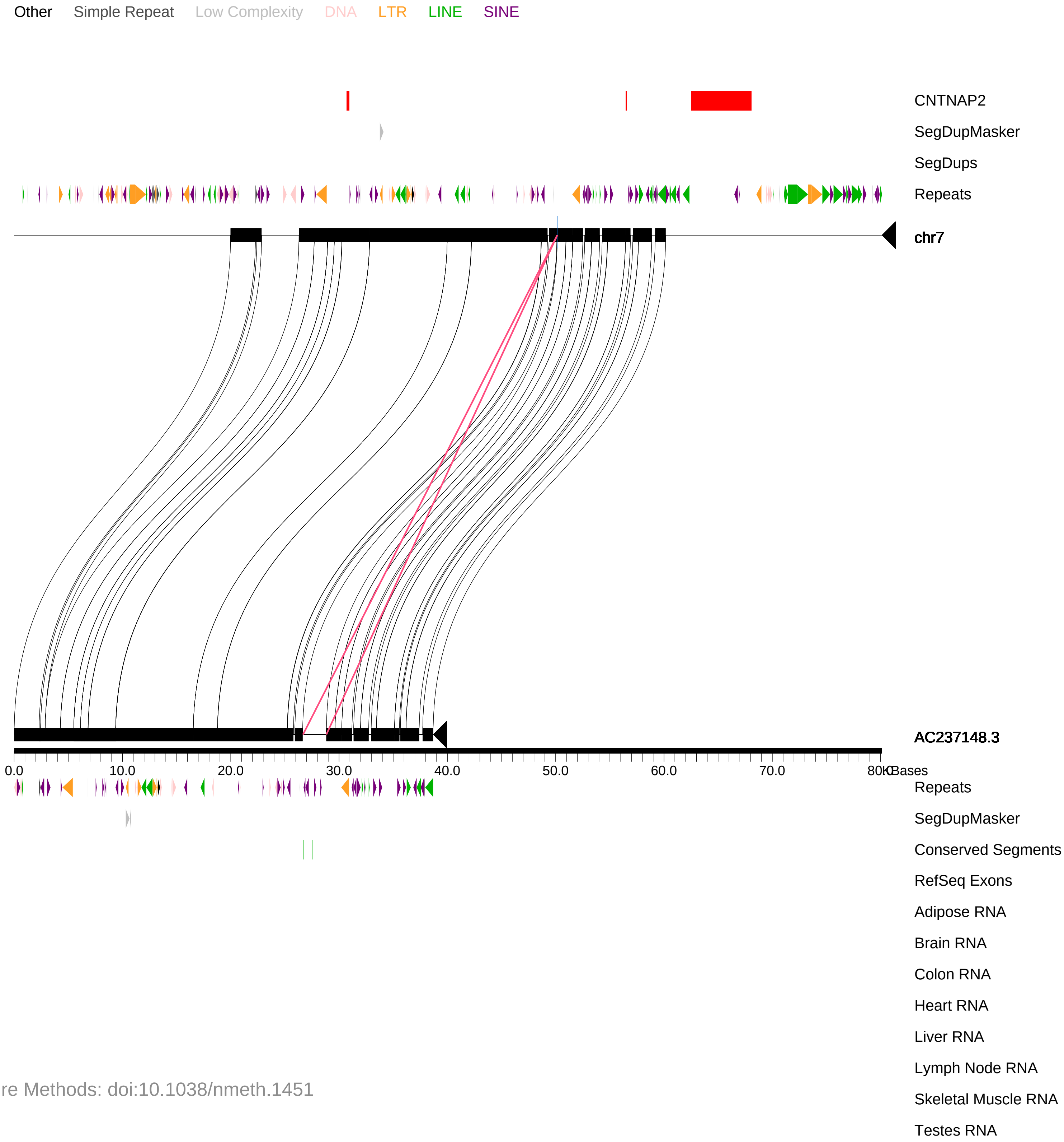
Insertion Size: 2160



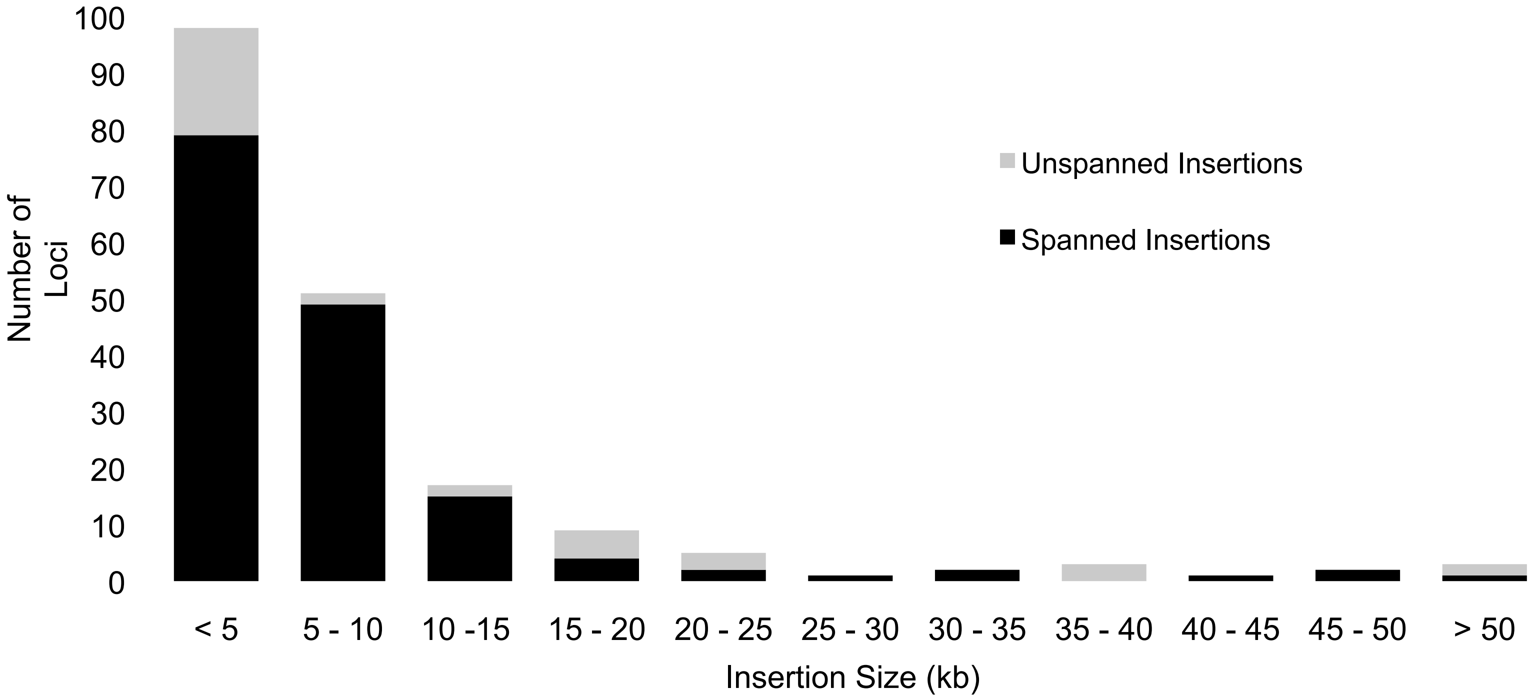
Nature Methods: doi:10.1038/nmeth.1451

Clone file = AC237148.rc.fa

Insertion Size: 2155

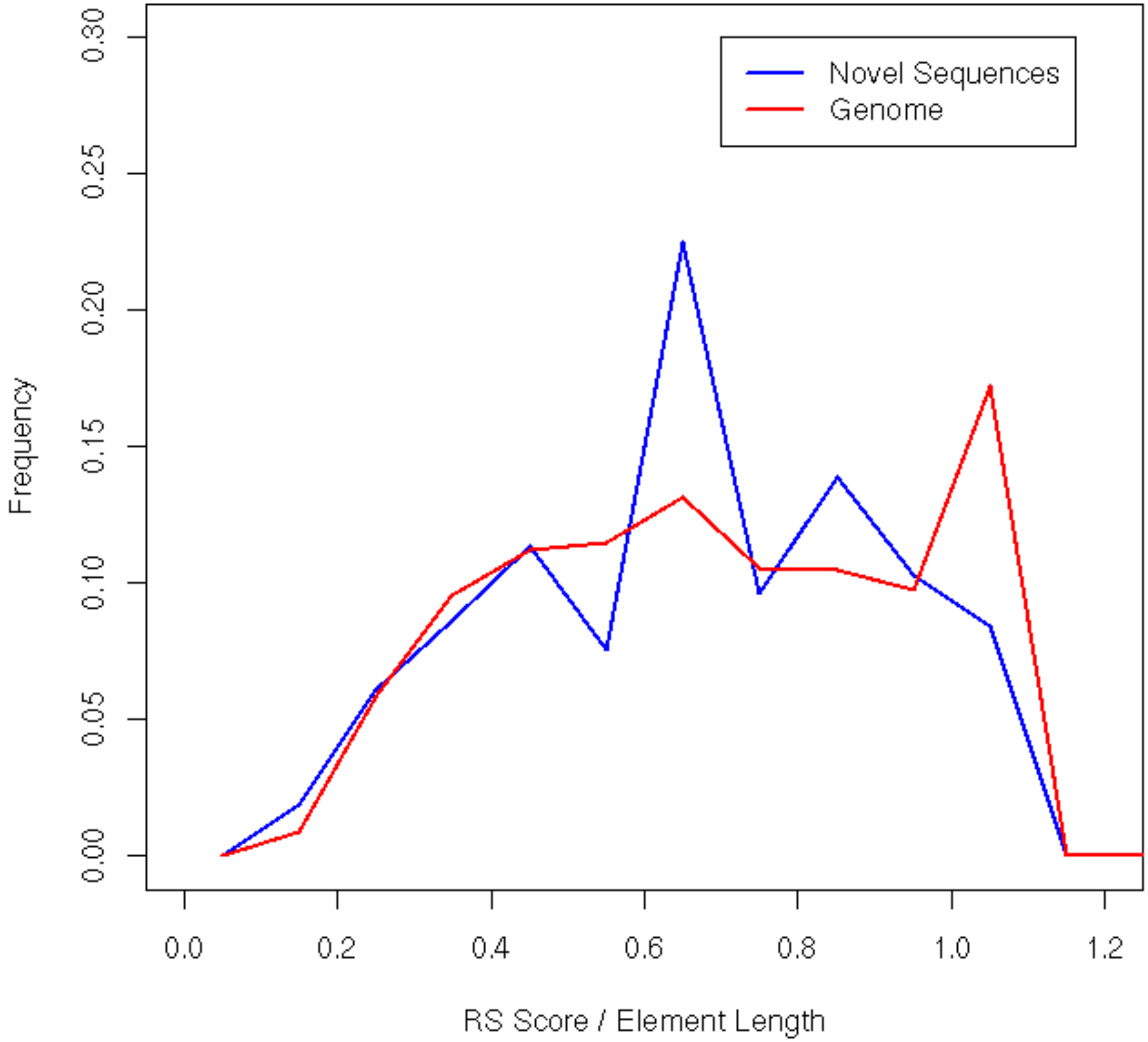


Nature Methods: doi:10.1038/nmeth.1451



Supplementary Figure 6 Size distribution for sequenced insertions

Insertion sizes were determined by complete sequencing of 192 novel insertions corresponding to 1.67 Mb of sequence. The filled black bars correspond to 156 completely spanned insertions for which the entire insertion sequence could be determined (including 4 sites >40 kb spanned using multiple OEA clones). The grey segment represents the amount of sequence captured by OEA clones at 36 loci that have not been traversed. These sequences therefore represent a minimum estimate of the true insertion size.



Supplementary Figure 7 Distribution of constraint for conserved elements
The distribution of total RS score divided by element length is shown for 477 conserved elements identified in the insertion sequences and for 1,133,900 constrained elements identified in the Ensembl Compara 51 9-species alignments. In both data-sets the human sequence was omitted prior to calculating levels of constraint.

Supplementary Table 2 Map locations of anchored loci
The 400 novel sequence loci having a defined genome anchor are shown. Contigs were required to have consistent anchors located within 100 kb of each other. Coordinates are given for the NCBI build35 genome assembly.

locus name	chrm	begin	end
novel-locus_1	chr2	1481419	1628043
novel-locus_2	chr4	167892757	167996626
novel-locus_7	chr12	131442013	131461790
novel-locus_10	chr20	60490150	60603931
novel-locus_14	chr6	95705659	95737264
novel-locus_15	chr6	119218754	119296390
novel-locus_16	chr11	69396933	69495348
novel-locus_18	chr20	53517037	53593400
novel-locus_19	chr12	132350817	132383947
novel-locus_21	chr3	66375963	66459324
novel-locus_22	chr6	69020036	69021785
novel-locus_23	chr11	70439901	70518289
novel-locus_24	chrX	76390972	76408442
novel-locus_25	chr13	111577720	111592136
novel-locus_26	chr8	21623864	21630103
novel-locus_31	chr2	89435851	89468868
novel-locus_43	chr13	113620894	113648269
novel-locus_45	chr7	98423577	98501730
novel-locus_47	chr22	47339640	47347031
novel-locus_48	chr8	58105700	58187102
novel-locus_50	chr9	134297576	134305963
novel-locus_51	chr9	130113403	130150824
novel-locus_53	chr2	149476515	149486750
novel-locus_54	chr3	32653619	32728200
novel-locus_55	chr12	107796063	107931938
novel-locus_57	chr3	195477493	195576652
novel-locus_59	chr6	80099674	80181401
novel-locus_65	chr17	260347	263109
novel-locus_78	chr8	21371450	21452967
novel-locus_83	chr7	47942254	48053088
novel-locus_88	chr1	204774710	204775525
novel-locus_92	chr5	94538753	94614515
novel-locus_93	chr18	13938431	14011851
novel-locus_96	chr12	126124874	126192633
novel-locus_98	chr2	16289006	16319752
novel-locus_100	chrX	147018894	147019681
novel-locus_102	chr20	4336438	4416508
novel-locus_103	chr8	142453189	142533856
novel-locus_107	chr1	3935662	3945471
novel-locus_109	chr2	4459296	4479518

novel-locus_111	chr7	129707195	129786385
novel-locus_112	chr11	56189696	56263803
novel-locus_115	chr17	47995161	48070195
novel-locus_116	chr19	21566038	21574238
novel-locus_118	chr1	231435637	231472816
novel-locus_119	chr13	25245195	25325741
novel-locus_122	chr10	103638825	103647016
novel-locus_123	chr1	220040658	220170044
novel-locus_126	chr4	31449331	31457033
novel-locus_129	chr9	88005411	88083297
novel-locus_130	chr6	13194747	13276790
novel-locus_132	chr15	75746052	75811857
novel-locus_133	chr3	146433392	146512713
novel-locus_136	chr10	82925050	82999452
novel-locus_138	chr5	97632155	97650512
novel-locus_139	chr21	14696133	14766367
novel-locus_140	chr19	39466456	39616620
novel-locus_141	chr9	101420958	101424685
novel-locus_142	chr5	29067224	29072091
novel-locus_143	chr17	68331759	68429248
novel-locus_145	chr1	202614909	202636666
novel-locus_147	chr1	207175638	207258541
novel-locus_148	chr4	121454593	121536385
novel-locus_150	chr2	117393057	117400457
novel-locus_151	chr13	113484022	113513392
novel-locus_152	chr10	37865567	37866229
novel-locus_153	chr17	18818166	18822889
novel-locus_154	chr12	121139614	121154611
novel-locus_156	chr6	104783168	104858561
novel-locus_157	chr18	66245599	66318778
novel-locus_158	chr2	114386235	114462833
novel-locus_160	chrX	36818399	36824805
novel-locus_167	chr3	84994700	85070040
novel-locus_168	chr14	100452761	100530750
novel-locus_170	chr7	286375	293725
novel-locus_173	chr2	233790004	233891008
novel-locus_174	chr7	97995601	98065163
novel-locus_175	chr21	29373076	29452594
novel-locus_178	chrX	151884818	152093528
novel-locus_181	chr4	157196240	157272185
novel-locus_182	chr20	31669	48843
novel-locus_184	chr2	1214417	1220376
novel-locus_186	chr1	103457801	103495814
novel-locus_188	chr1	31602684	31677453
novel-locus_189	chr3	47406235	47491479
novel-locus_190	chr22	42993199	43095758
novel-locus_192	chr4	190292101	190373677
novel-locus_193	chr13	85621905	85637284
novel-locus_194	chr14	63299763	63375958

novel-locus_195	chr1	5294323	5439512
novel-locus_198	chr2	208747685	208828456
novel-locus_199	chr14	93861171	93943187
novel-locus_200	chr20	46211899	46212576
novel-locus_201	chr15	64142118	64220468
novel-locus_203	chr15	25137673	25146319
novel-locus_204	chr8	49209535	49288171
novel-locus_205	chr11	42740780	42742233
novel-locus_206	chr3	57321452	57396643
novel-locus_207	chr21	21355906	21428601
novel-locus_209	chr4	1382182	1498091
novel-locus_210	chr19	1077800	1155279
novel-locus_213	chr1	3806192	3829997
novel-locus_214	chrX	148648267	148663207
novel-locus_215	chr7	50104635	50221793
novel-locus_216	chr12	120925161	120959528
novel-locus_218	chr11	59535950	59609789
novel-locus_225	chr18	3058670	3064899
novel-locus_227	chr4	32702762	32712352
novel-locus_228	chr7	67635135	67711458
novel-locus_229	chr6	32426346	32497046
novel-locus_234	chr18	70494293	70500353
novel-locus_235	chr10	128571311	128698498
novel-locus_236	chr9	130935212	130936037
novel-locus_237	chr2	21032948	21135012
novel-locus_241	chr6	169965084	170097542
novel-locus_242	chr11	8011976	8101216
novel-locus_247	chr11	89594565	89669563
novel-locus_248	chr11	55497249	55497929
novel-locus_255	chr2	237539180	237617954
novel-locus_257	chr22	44645603	44750046
novel-locus_259	chr1	245177563	245304384
novel-locus_260	chr7	154652949	154662015
novel-locus_261	chr6	82985969	83065812
novel-locus_262	chr12	37495639	37570779
novel-locus_263	chr17	76902168	77154294
novel-locus_264	chr11	91668716	91742501
novel-locus_265	chr18	68751197	68823968
novel-locus_266	chr15	18783377	18839127
novel-locus_267	chr19	8831130	8901539
novel-locus_268	chr22	17769882	17770654
novel-locus_274	chr11	1116549	1128154
novel-locus_277	chrX	4785906	4866299
novel-locus_279	chr15	27602915	27681706
novel-locus_280	chr20	11269943	11283252
novel-locus_282	chr5	26875049	26951181
novel-locus_284	chr15	96394881	96406914
novel-locus_285	chr2	169514440	169588018
novel-locus_288	chr6	80628948	80708997

novel-locus_290	chr2	129601211	129684038
novel-locus_292	chr5	124757014	124828339
novel-locus_293	chr12	85117092	85191426
novel-locus_294	chr5	157176629	157177532
novel-locus_296	chr16	30001584	30078443
novel-locus_297	chr8	130787977	130865169
novel-locus_299	chr4	97739760	97821211
novel-locus_301	chr2	126754568	126825276
novel-locus_302	chr5	12701053	12771819
novel-locus_303	chr18	44413428	44493375
novel-locus_305	chr18	63234122	63311901
novel-locus_307	chr3	185596339	185671910
novel-locus_312	chr10	88215267	88295610
novel-locus_316	chr10	76287791	76288387
novel-locus_319	chr12	107438437	107516755
novel-locus_321	chr12	6157100	6322814
novel-locus_322	chr1	61525878	61613498
novel-locus_323	chr6	165125903	165206642
novel-locus_324	chr10	27613358	27678648
novel-locus_330	chr20	56384826	56465692
novel-locus_331	chr8	144976824	145203821
novel-locus_334	chr7	138794105	138903690
novel-locus_335	chr1	195173079	195240427
novel-locus_336	chr16	2811479	2894586
novel-locus_340	chr12	74743546	74762735
novel-locus_341	chr19	16066314	16068963
novel-locus_345	chr3	173930601	173931380
novel-locus_347	chrX	153144348	153148144
novel-locus_348	chr16	74580032	74654377
novel-locus_351	chr6	66578098	66657455
novel-locus_352	chr3	37764227	37765373
novel-locus_353	chr16	24749347	24754513
novel-locus_354	chr12	57606424	57684731
novel-locus_357	chr22	21041188	21234037
novel-locus_358	chr8	11679871	11680547
novel-locus_360	chr19	52577941	52653812
novel-locus_361	chr15	81659620	81665855
novel-locus_363	chr18	18345758	18425174
novel-locus_367	chrX	46984237	46985100
novel-locus_369	chr10	57970607	58038405
novel-locus_370	chr14	68367956	68442260
novel-locus_371	chr14	66894193	66971670
novel-locus_372	chr1	169953072	169953840
novel-locus_373	chr10	133230019	133314858
novel-locus_376	chr19	23907633	24097614
novel-locus_378	chr3	66074582	66105441
novel-locus_382	chr1	118610995	118677369
novel-locus_391	chr9	136517708	136532700
novel-locus_392	chr13	108100745	108177118

novel-locus_393	chr9	134409143	134425396
novel-locus_396	chr5	9984065	10063672
novel-locus_397	chr8	140201079	140282038
novel-locus_398	chr6	47503109	47574626
novel-locus_400	chr21	33726140	33805201
novel-locus_403	chr2	238308435	238378520
novel-locus_404	chr22	49006330	49030376
novel-locus_405	chr12	71835647	71907496
novel-locus_406	chr17	75631905	75711696
novel-locus_409	chr11	7755871	7861303
novel-locus_411	chrX	49637575	49658722
novel-locus_412	chr10	35898429	35980340
novel-locus_413	chr7	147496806	147577214
novel-locus_419	chr3	42323395	42400728
novel-locus_420	chr19	23257373	23341795
novel-locus_422	chr1	72480921	72555312
novel-locus_424	chr2	206876550	206955276
novel-locus_430	chr4	75888432	75897435
novel-locus_431	chr8	103466241	103540446
novel-locus_432	chr11	1479130	1556727
novel-locus_434	chr1	202835136	202840150
novel-locus_438	chr21	22141928	22142829
novel-locus_441	chr12	7163775	7172788
novel-locus_443	chr6	26113395	26190934
novel-locus_445	chr22	39980738	39985767
novel-locus_452	chr14	45853679	45928667
novel-locus_464	chr19	3412333	3413231
novel-locus_471	chr6	166627704	166644926
novel-locus_486	chr1	244931381	245064453
novel-locus_487	chr1	23309076	23387660
novel-locus_492	chr1	45795821	45814176
novel-locus_498	chr2	41606020	41611819
novel-locus_499	chr17	51355578	51356317
novel-locus_506	chr16	66267237	66340089
novel-locus_508	chr9	131223632	131303084
novel-locus_514	chr20	42049424	42126436
novel-locus_515	chr6	167343700	167421425
novel-locus_517	chr19	3900225	3978555
novel-locus_525	chr1	165056626	165061354
novel-locus_527	chr6	168498247	168498830
novel-locus_528	chr4	175578544	175652233
novel-locus_534	chr22	35545798	35623542
novel-locus_539	chr4	54590542	54671223
novel-locus_541	chr20	60765616	60774959
novel-locus_542	chr1	22611594	22866272
novel-locus_545	chr2	26430855	26508234
novel-locus_547	chr1	163868131	163944393
novel-locus_548	chr11	55860598	55941107
novel-locus_549	chr10	45185815	45189290

novel-locus_550	chr10	66913778	66991268
novel-locus_552	chr13	28015984	28096159
novel-locus_559	chr4	77064193	77069909
novel-locus_561	chr12	131333046	131334407
novel-locus_566	chr1	87186345	87255943
novel-locus_568	chr3	155340238	155422266
novel-locus_569	chr20	34320237	34500535
novel-locus_571	chr22	47737162	47741990
novel-locus_572	chr13	83780623	83859087
novel-locus_579	chr11	23288572	23360655
novel-locus_585	chr9	425236	495590
novel-locus_587	chr8	143963816	143978558
novel-locus_588	chr19	4657087	4729098
novel-locus_593	chr1	17277403	17403684
novel-locus_595	chr14	91130504	91213247
novel-locus_596	chr18	73906706	73911436
novel-locus_598	chr14	81147136	81224222
novel-locus_600	chr9	135426316	135508740
novel-locus_605	chr4	21238070	21238981
novel-locus_607	chr2	3105907	3111199
novel-locus_615	chr5	68515572	68589020
novel-locus_616	chr7	8835046	8903291
novel-locus_618	chrX	153269284	153272694
novel-locus_627	chr19	7240485	7242010
novel-locus_628	chr12	1010794	1090060
novel-locus_632	chr4	31670240	31674586
novel-locus_634	chr18	39597622	39598350
novel-locus_636	chr1	242110099	242188954
novel-locus_641	chr17	59758532	59838455
novel-locus_647	chr2	4919965	4997319
novel-locus_648	chr9	76826895	76827659
novel-locus_650	chr6	90601640	90607960
novel-locus_654	chr16	67668976	67820339
novel-locus_656	chr12	6890387	7026836
novel-locus_659	chr7	146791242	146791880
novel-locus_667	chr17	4690715	4768384
novel-locus_673	chr8	96344941	96345746
novel-locus_676	chr2	239290394	239363341
novel-locus_686	chr4	56482404	56492204
novel-locus_687	chr2	136385564	136464616
novel-locus_692	chr11	47581644	47655493
novel-locus_693	chr6	8631505	8632341
novel-locus_695	chr11	23036298	23036527
novel-locus_701	chr6	47779504	47780315
novel-locus_704	chr3	156469963	156547188
novel-locus_707	chr21	23713599	23716366
novel-locus_711	chr11	101259654	101263034
novel-locus_721	chr3	113329103	113401614
novel-locus_723	chr8	28217714	28218595

novel-locus_724	chr4	155402315	155475287
novel-locus_727	chr19	46695220	46770921
novel-locus_729	chr14	89004365	89005271
novel-locus_733	chr10	79335632	79336433
novel-locus_739	chr16	85539372	85615300
novel-locus_742	chr15	97405595	97487762
novel-locus_743	chr11	11215533	11284026
novel-locus_745	chr8	117792667	117794529
novel-locus_749	chr13	80273212	80273970
novel-locus_751	chr1	240296368	240371490
novel-locus_753	chr3	20274500	20358068
novel-locus_758	chr2	759002	835377
novel-locus_760	chr3	126251164	126333743
novel-locus_763	chr17	7105901	7185216
novel-locus_764	chrX	46071093	46078716
novel-locus_774	chr1	55455355	55460119
novel-locus_776	chr2	109648566	109652701
novel-locus_784	chr8	141007450	141079160
novel-locus_786	chr14	48519669	48598232
novel-locus_790	chrX	77300162	77300827
novel-locus_848	chr2	26806201	26888743
novel-locus_857	chr8	52825664	52829427
novel-locus_860	chr1	33278871	33284501
novel-locus_865	chr9	125614213	125696264
novel-locus_867	chr2	231067405	231068105
novel-locus_879	chr2	116372059	116449844
novel-locus_895	chrX	115454488	115455730
novel-locus_898	chr4	59535726	59544386
novel-locus_909	chr4	30174882	30251778
novel-locus_910	chr3	176541606	176542272
novel-locus_911	chr7	4590890	4591651
novel-locus_914	chr3	152827893	152899795
novel-locus_917	chr11	30869774	30947788
novel-locus_921	chr12	56720429	56726359
novel-locus_922	chr3	51405432	51481083
novel-locus_923	chr16	26050677	26130094
novel-locus_924	chr13	102106237	102183846
novel-locus_927	chr7	154505696	154515878
novel-locus_929	chr8	41842601	41919355
novel-locus_930	chr1	113216157	113292540
novel-locus_932	chr8	142820821	142900760
novel-locus_934	chr1	216431885	216505446
novel-locus_940	chr1	105788134	105867015
novel-locus_941	chr16	1899983	1900735
novel-locus_944	chr5	36559945	36633964
novel-locus_948	chr14	24771450	24851269
novel-locus_949	chr19	341285	427658
novel-locus_952	chr4	32582916	32585541
novel-locus_955	chr1	231856703	231936127

novel-locus_959	chr12	122879703	122958700
novel-locus_960	chr6	57461337	57538663
novel-locus_966	chr3	177467790	177473967
novel-locus_968	chr8	129769005	129849253
novel-locus_969	chr1	104156660	104232417
novel-locus_976	chr17	8661213	8739724
novel-locus_977	chr18	1790427	1870817
novel-locus_979	chr11	69858071	69933114
novel-locus_980	chr6	111556822	111632489
novel-locus_981	chr6	28208187	28289091
novel-locus_989	chr4	14671100	14673808
novel-locus_991	chr9	36497236	36497555
novel-locus_996	chr16	8034648	8114328
novel-locus_1001	chr13	48344731	48420797
novel-locus_1003	chr21	43249158	43328009
novel-locus_1004	chr16	71906893	71907441
novel-locus_1011	chr17	76557923	76559017
novel-locus_1013	chr22	19144260	19220761
novel-locus_1015	chr10	47162731	47163520
novel-locus_1016	chr14	103390061	103469589
novel-locus_1019	chr3	104565597	104638695
novel-locus_1021	chrX	148792028	148795766
novel-locus_1023	chr22	48030441	48031854
novel-locus_1024	chr7	153837847	153857864
novel-locus_1027	chr10	121193888	121194691
novel-locus_1042	chr2	133020493	133025391
novel-locus_1043	chr11	18316057	18391475
novel-locus_1045	chr15	87646370	87726367
novel-locus_1046	chr13	57772531	57853916
novel-locus_1050	chr21	18054857	18063247
novel-locus_1055	chrX	4660017	4660693
novel-locus_1058	chr12	52844765	52927758
novel-locus_1062	chr2	94778948	94779774
novel-locus_1064	chr8	137067558	137144659
novel-locus_1072	chr10	129362351	129442606
novel-locus_1073	chr1	230612928	230688944
novel-locus_1075	chr15	24648685	24653323
novel-locus_1076	chr12	53736210	53737044
novel-locus_1077	chr6	68339493	68413803
novel-locus_1078	chr10	84996773	85007367
novel-locus_1079	chr4	66595153	66596642
novel-locus_1081	chr17	18175017	18180681
novel-locus_1085	chr9	8962753	8963633
novel-locus_1090	chr1	82395292	82471368
novel-locus_1093	chr9	24931453	25012018
novel-locus_1099	chr21	33087575	33166499
novel-locus_1111	chr16	69849507	69922166
novel-locus_1116	chr6	65117910	65118698
novel-locus_1117	chr8	62251035	62323984

novel-locus_1118	chr9	101165070	101239260
novel-locus_1127	chr7	8182765	8186116
novel-locus_1128	chr1	29652644	29653551
novel-locus_1131	chr10	28332956	28411469
novel-locus_1132	chr3	3167290	3174264
novel-locus_1133	chr3	96094684	96170632
novel-locus_1134	chr14	57443692	57520089
novel-locus_1138	chr2	109524077	109533725
novel-locus_1145	chr8	50749998	50827856
novel-locus_1151	chr5	155155412	155163193
novel-locus_1152	chr13	89080921	89153453
novel-locus_1160	chr9	89532348	89534743
novel-locus_1164	chr14	82618348	82695497
novel-locus_1173	chr15	72252005	72333336
novel-locus_1176	chr21	46004959	46077013
novel-locus_1179	chr11	125971515	126047969
novel-locus_1181	chr20	5707232	5712082

Supplementary Table 3 FISH analysis of orphan clones
Summary of FISH results from fosmids corresponding to 68 orphan contigs established based on fingerprinting from genomic library, G248 (WIBR2, NA15510).

Classification	Number of Contigs
Assembly Gap	31 (45%)
Interstitial	15 (22%)
Telomeric	10 (15%)
Acrocentric	8 (12%)
Pericentromeric	4 (6%)

Supplementary Table 5 Noise-multiplier results

The 890 contigs identified as polymorphic using the noise-multiplier approach are listed.

Contig Name
freeze2_10009
freeze2_10056
freeze2_10062
freeze2_10068
freeze2_10077
freeze2_10083
freeze2_10116
freeze2_10139
freeze2_10200
freeze2_103
freeze2_10302
freeze2_10363
freeze2_10397
freeze2_10443
freeze2_10459
freeze2_10478
freeze2_10585
freeze2_1060
freeze2_10726
freeze2_10750
freeze2_10806
freeze2_109
freeze2_10901
freeze2_1093
freeze2_10952
freeze2_1097
freeze2_10972
freeze2_10988
freeze2_10991
freeze2_11001
freeze2_11030
freeze2_11104
freeze2_11227
freeze2_11258
freeze2_11285
freeze2_11303
freeze2_11323
freeze2_11329
freeze2_11344
freeze2_11358
freeze2_11371
freeze2_114
freeze2_11402
freeze2_11438
freeze2_11445
freeze2_11473

freeze2_11487
freeze2_11506
freeze2_11513
freeze2_11535
freeze2_11537
freeze2_11545
freeze2_11546
freeze2_1155
freeze2_11592
freeze2_11637
freeze2_11667
freeze2_11674
freeze2_11675
freeze2_11677
freeze2_11729
freeze2_11812
freeze2_11862
freeze2_11903
freeze2_11957
freeze2_11963
freeze2_11991
freeze2_12
freeze2_12001
freeze2_12005
freeze2_12043
freeze2_12072
freeze2_12123
freeze2_12167
freeze2_12205
freeze2_1226
freeze2_12320
freeze2_12368
freeze2_12370
freeze2_12378
freeze2_1243
freeze2_12437
freeze2_12485
freeze2_125
freeze2_12556
freeze2_12568
freeze2_1260
freeze2_1264
freeze2_12644
freeze2_12693
freeze2_12696
freeze2_127
freeze2_12727
freeze2_12735
freeze2_12750

freeze2_12760
freeze2_12778
freeze2_12823
freeze2_12833
freeze2_12874
freeze2_12957
freeze2_12989
freeze2_13048
freeze2_13063
freeze2_13072
freeze2_13082
freeze2_13086
freeze2_13089
freeze2_13099
freeze2_13142
freeze2_13156
freeze2_13185
freeze2_13215
freeze2_1332
freeze2_13343
freeze2_13357
freeze2_13368
freeze2_13373
freeze2_13377
freeze2_1338
freeze2_13401
freeze2_13424
freeze2_13434
freeze2_13438
freeze2_13472
freeze2_13491
freeze2_13606
freeze2_13627
freeze2_13649
freeze2_1365
freeze2_13702
freeze2_13757
freeze2_13829
freeze2_1388
freeze2_13902
freeze2_13917
freeze2_13947
freeze2_13953
freeze2_13993
freeze2_140
freeze2_14037
freeze2_14073
freeze2_14082
freeze2_14102

freeze2_14103_altContig2
freeze2_14131
freeze2_14139
freeze2_14155
freeze2_14166
freeze2_14202
freeze2_14204
freeze2_14213
freeze2_14226
freeze2_14232
freeze2_1424
freeze2_14272
freeze2_14293
freeze2_14296
freeze2_14380
freeze2_14384
freeze2_14455
freeze2_14641
freeze2_14655
freeze2_14667
freeze2_14737
freeze2_14757
freeze2_14761
freeze2_14770
freeze2_14777
freeze2_14778
freeze2_14793
freeze2_14800
freeze2_14805
freeze2_14818
freeze2_14832
freeze2_14842
freeze2_14850
freeze2_14871
freeze2_14898
freeze2_1491
freeze2_14913
freeze2_14942
freeze2_14949
freeze2_14950
freeze2_14951
freeze2_14956
freeze2_14983
freeze2_14987
freeze2_14992
freeze2_14993
freeze2_15012
freeze2_15031
freeze2_15037

freeze2_15080
freeze2_15092
freeze2_15093
freeze2_15118
freeze2_15121
freeze2_15143
freeze2_15175
freeze2_1530
freeze2_15555
freeze2_15578
freeze2_15587
freeze2_15604
freeze2_1562
freeze2_15663
freeze2_15783
freeze2_15808
freeze2_16014
freeze2_16018
freeze2_1605
freeze2_16063
freeze2_16064_altContig2
freeze2_16084
freeze2_16100
freeze2_1612
freeze2_16138
freeze2_16232
freeze2_1633
freeze2_1645
freeze2_16469
freeze2_1652
freeze2_16642
freeze2_16670
freeze2_16673
freeze2_16686
freeze2_16744
freeze2_16758
freeze2_1677
freeze2_1680
freeze2_16801
freeze2_16829
freeze2_169
freeze2_16913
freeze2_16944
freeze2_16955
freeze2_1698
freeze2_16997
freeze2_17012
freeze2_17086
freeze2_17255

freeze2_17303
freeze2_1735
freeze2_17375
freeze2_17461
freeze2_17629
freeze2_1765
freeze2_17764
freeze2_178
freeze2_17820
freeze2_17847
freeze2_17872
freeze2_179
freeze2_1794
freeze2_18067
freeze2_18168
freeze2_1817
freeze2_18229
freeze2_18239
freeze2_18283
freeze2_1832
freeze2_18430
freeze2_1849
freeze2_1852
freeze2_1856
freeze2_18567
freeze2_18585
freeze2_1883
freeze2_1901
freeze2_19089
freeze2_1912
freeze2_1916
freeze2_192
freeze2_193
freeze2_19364
freeze2_19478
freeze2_1953
freeze2_197
freeze2_1980
freeze2_19865
freeze2_1990
freeze2_19958
freeze2_2009
freeze2_2012
freeze2_2021
freeze2_2022
freeze2_2027
freeze2_2049
freeze2_20507
freeze2_206

freeze2_2065
freeze2_20760
freeze2_2087
freeze2_2095
freeze2_21084
freeze2_2129
freeze2_213
freeze2_2140
freeze2_21411
freeze2_21425
freeze2_21451
freeze2_21461
freeze2_21475
freeze2_21545
freeze2_216
freeze2_21606
freeze2_21617
freeze2_21640
freeze2_21707
freeze2_21737
freeze2_21764
freeze2_21793
freeze2_21846
freeze2_21884
freeze2_21967
freeze2_22039
freeze2_22136
freeze2_2216
freeze2_22235
freeze2_22334
freeze2_22358
freeze2_22361
freeze2_22404
freeze2_22421
freeze2_22485
freeze2_22530
freeze2_22587
freeze2_22708
freeze2_2275
freeze2_22846
freeze2_22974
freeze2_22990
freeze2_23044
freeze2_23166
freeze2_23236
freeze2_23335
freeze2_23398
freeze2_23429
freeze2_2345

freeze2_23489
freeze2_23569
freeze2_23611
freeze2_2372
freeze2_23740
freeze2_23888
freeze2_2394
freeze2_24019
freeze2_24255
freeze2_2441
freeze2_2444
freeze2_24444
freeze2_2448
freeze2_24679
freeze2_24687
freeze2_2470
freeze2_24756
freeze2_2484
freeze2_24857
freeze2_2506
freeze2_2512
freeze2_25394
freeze2_2541
freeze2_2556
freeze2_25577
freeze2_25833
freeze2_2602
freeze2_2606
freeze2_261
freeze2_26162
freeze2_26185
freeze2_26317
freeze2_2634
freeze2_26401
freeze2_26654
freeze2_2667
freeze2_26743
freeze2_2676
freeze2_26974
freeze2_27056
freeze2_27156
freeze2_27232
freeze2_27382
freeze2_27393
freeze2_27459
freeze2_27464
freeze2_27469
freeze2_2749
freeze2_27549

freeze2_27555
freeze2_27613
freeze2_27662
freeze2_27721
freeze2_27818
freeze2_2796
freeze2_27968
freeze2_27989
freeze2_2833
freeze2_2836
freeze2_2838
freeze2_28529
freeze2_28534
freeze2_28582
freeze2_2866
freeze2_2872
freeze2_2876
freeze2_28819
freeze2_2905
freeze2_29125
freeze2_2914
freeze2_29188
freeze2_2920
freeze2_2931
freeze2_29354
freeze2_29400
freeze2_2951
freeze2_29516
freeze2_29558
freeze2_29606
freeze2_29620
freeze2_29698
freeze2_29724
freeze2_2979
freeze2_29818
freeze2_29836
freeze2_29892
freeze2_29985
freeze2_29998
freeze2_30072
freeze2_3010
freeze2_30177
freeze2_3018
freeze2_30187
freeze2_3019
freeze2_30191
freeze2_30380
freeze2_30403
freeze2_3047

freeze2_30540
freeze2_30556
freeze2_3060
freeze2_30602
freeze2_30603
freeze2_30701
freeze2_30722
freeze2_30798
freeze2_308
freeze2_30900
freeze2_3092
freeze2_31
freeze2_31075
freeze2_31141
freeze2_31171
freeze2_315
freeze2_31540
freeze2_3155
freeze2_31622
freeze2_3168
freeze2_3172
freeze2_31723
freeze2_31747
freeze2_3176
freeze2_31780
freeze2_3179
freeze2_31863
freeze2_31954
freeze2_3198
freeze2_3199
freeze2_32
freeze2_3220
freeze2_32208
freeze2_3236
freeze2_3248
freeze2_325
freeze2_3250
freeze2_3259
freeze2_3277
freeze2_328
freeze2_3280
freeze2_3308
freeze2_3322
freeze2_335
freeze2_3363
freeze2_3364
freeze2_3366
freeze2_3381
freeze2_3392

freeze2_3398
freeze2_3399
freeze2_34028
freeze2_3417
freeze2_34205
freeze2_3427
freeze2_34340
freeze2_3436
freeze2_34468
freeze2_3447
freeze2_3450
freeze2_3455
freeze2_3464
freeze2_34648
freeze2_3473
freeze2_3474
freeze2_34944
freeze2_3496
freeze2_3515
freeze2_35164
freeze2_3519
freeze2_3531
freeze2_3558
freeze2_35607
freeze2_3565
freeze2_3583
freeze2_35865
freeze2_3595
freeze2_3597
freeze2_3598
freeze2_3604
freeze2_3614
freeze2_3617
freeze2_362
freeze2_368
freeze2_3686
freeze2_3697
freeze2_3734
freeze2_3739
freeze2_3755
freeze2_3783
freeze2_381
freeze2_3839
freeze2_3854
freeze2_389
freeze2_3897
freeze2_3915
freeze2_3944
freeze2_3948

freeze2_3967
freeze2_4067
freeze2_415
freeze2_4250
freeze2_436
freeze2_4387
freeze2_4422
freeze2_4434
freeze2_4446
freeze2_449
freeze2_4585
freeze2_4589
freeze2_4605
freeze2_4607
freeze2_4613
freeze2_4658
freeze2_470
freeze2_4708
freeze2_478
freeze2_4797
freeze2_481
freeze2_489
freeze2_500
freeze2_5022
freeze2_5031
freeze2_5056
freeze2_507
freeze2_51
freeze2_5183
freeze2_5201
freeze2_5207
freeze2_5211
freeze2_5213
freeze2_5267
freeze2_527
freeze2_5288
freeze2_53
freeze2_5308
freeze2_5317
freeze2_5340
freeze2_5385
freeze2_5427
freeze2_544
freeze2_545
freeze2_5461
freeze2_5463
freeze2_5475
freeze2_5498
freeze2_5500

freeze2_5513
freeze2_5533
freeze2_5553
freeze2_5555
freeze2_5578
freeze2_5619
freeze2_5624
freeze2_5630
freeze2_5667
freeze2_5672
freeze2_5693
freeze2_5712
freeze2_5719
freeze2_5733
freeze2_5742
freeze2_5756
freeze2_5769
freeze2_5770
freeze2_5780
freeze2_5785
freeze2_5791
freeze2_5817
freeze2_5820
freeze2_5829
freeze2_5833
freeze2_5857
freeze2_5869
freeze2_5870
freeze2_5872
freeze2_5919
freeze2_5920
freeze2_5930
freeze2_5953
freeze2_5958
freeze2_5964
freeze2_5995
freeze2_6019
freeze2_6081
freeze2_6084
freeze2_6115
freeze2_6139
freeze2_615
freeze2_6156
freeze2_6157
freeze2_6166
freeze2_6167
freeze2_6193
freeze2_6219
freeze2_6224

freeze2_6237
freeze2_6262
freeze2_6264
freeze2_6269
freeze2_6276
freeze2_6286
freeze2_6314
freeze2_6316
freeze2_6321
freeze2_6329
freeze2_6340
freeze2_6348
freeze2_6351
freeze2_6394
freeze2_6414
freeze2_6450
freeze2_646
freeze2_6479
freeze2_6503
freeze2_6510
freeze2_6516
freeze2_652
freeze2_6538
freeze2_6543
freeze2_6563
freeze2_6604
freeze2_6641
freeze2_6647
freeze2_6649
freeze2_6650
freeze2_6709
freeze2_6729
freeze2_6766
freeze2_6784
freeze2_6805
freeze2_6824
freeze2_688
freeze2_6915
freeze2_6921
freeze2_6924
freeze2_6925
freeze2_6965
freeze2_6982
freeze2_6985
freeze2_6987
freeze2_6997
freeze2_7027
freeze2_7055
freeze2_7058

freeze2_706
freeze2_7075
freeze2_7115
freeze2_7117
freeze2_7129
freeze2_7161
freeze2_7169
freeze2_7175
freeze2_7223
freeze2_7241
freeze2_7272
freeze2_7276
freeze2_7280
freeze2_7391
freeze2_7397
freeze2_7400
freeze2_7405
freeze2_741
freeze2_7433
freeze2_7436
freeze2_7438
freeze2_7440
freeze2_7461
freeze2_7463
freeze2_7468
freeze2_7486
freeze2_7494
freeze2_7495
freeze2_75
freeze2_7504
freeze2_7514
freeze2_753
freeze2_7553
freeze2_7562
freeze2_7563
freeze2_7566
freeze2_7567
freeze2_7571
freeze2_7580
freeze2_7595
freeze2_76
freeze2_7632
freeze2_7671
freeze2_7683
freeze2_7690
freeze2_7694
freeze2_7707
freeze2_7722
freeze2_7743

freeze2_7754
freeze2_7763
freeze2_7773
freeze2_7783
freeze2_7795
freeze2_7805
freeze2_7808
freeze2_7812
freeze2_7814
freeze2_7818
freeze2_7822
freeze2_7841
freeze2_7844
freeze2_7848
freeze2_7851
freeze2_7877
freeze2_7906
freeze2_7924
freeze2_7934
freeze2_7935
freeze2_795
freeze2_7998
freeze2_8
freeze2_8011
freeze2_8012
freeze2_8018
freeze2_8028
freeze2_8030
freeze2_8059
freeze2_8076
freeze2_8111
freeze2_8141
freeze2_8146
freeze2_818
freeze2_8191
freeze2_8197
freeze2_8209
freeze2_8214
freeze2_8244
freeze2_8251
freeze2_8263
freeze2_8269
freeze2_8291
freeze2_8309
freeze2_8336
freeze2_8365
freeze2_8369
freeze2_8379
freeze2_8416

freeze2_8479
freeze2_8486
freeze2_85
freeze2_856
freeze2_857
freeze2_8586
freeze2_8589
freeze2_8628
freeze2_8635
freeze2_8653
freeze2_8662
freeze2_8673
freeze2_8698
freeze2_8701
freeze2_8722
freeze2_8729
freeze2_8755
freeze2_8765
freeze2_8766
freeze2_8772
freeze2_8780
freeze2_8784
freeze2_8785
freeze2_8800
freeze2_8802
freeze2_8808
freeze2_8814
freeze2_8817
freeze2_8824
freeze2_8825
freeze2_8827
freeze2_8852
freeze2_8866
freeze2_8877
freeze2_8882
freeze2_8888
freeze2_8890
freeze2_8895
freeze2_8899
freeze2_89
freeze2_8901
freeze2_8918
freeze2_8924
freeze2_8930
freeze2_8935
freeze2_8955
freeze2_8958
freeze2_8959
freeze2_8975

freeze2_8977
freeze2_8982
freeze2_8990
freeze2_8998
freeze2_9005
freeze2_9014
freeze2_9028
freeze2_9038
freeze2_9050
freeze2_9053
freeze2_9069
freeze2_909
freeze2_9113
freeze2_9137
freeze2_9163
freeze2_9190
freeze2_9193
freeze2_9196
freeze2_9254
freeze2_9254_altContig2
freeze2_9310
freeze2_9310_altContig2
freeze2_9332
freeze2_9383
freeze2_9388
freeze2_9403
freeze2_9453
freeze2_9476
freeze2_9494
freeze2_9529
freeze2_9542
freeze2_9556
freeze2_9568
freeze2_9596
freeze2_9600
freeze2_9608
freeze2_9613
freeze2_9656
freeze2_9695
freeze2_9716
freeze2_9720
freeze2_9759
freeze2_9773
freeze2_9775
freeze2_9789
freeze2_9790
freeze2_9806
freeze2_9827
freeze2_9843

freeze2_9863
freeze2_9869
freeze2_9877
freeze2_9897
freeze2_9913
freeze2_9931
freeze2_9938
freeze2_9974
freeze2_9975
freeze2_9980
freeze2_9995

Supplementary Table 9 Novel insertions with high V_{ST}

Loci with a V_{ST} value greater than 0.5 are listed.

Approximate Position		Discovery Populations	Insertion Size (kb)	Mean V_{ST}	F_{ST}	Distance to Nearest Gene (kb)	Gene Name
chr8	52,825,664	ASN,CEU,G248	~1.7	0.82	0.60	59	*PXDNL*
chr20	11,241,310	ASN,YRI	4.8	0.73	0.70	578	*BTBD3*
chr16	66,267,237	YRI	~0.9	0.71	0.58	1	*GFOD2*
Unknown		YRI	~1.1	0.68	0.31	--	--
chr20	5,707,232	CEU,YRI	~0.8	0.63	0.53	5	*C20orf196*
chr2	136,424,834	ASN,YRI	3.9	0.61	0.49	0	*LCT*
chr2	132,989,594	YRI	3.7	0.58	0.42	13	*GPR39*
chr10	84,968,888	ASN,YRI	5.1	0.57	0.36	234	*NRG3*
chr22	39,980,738	G248,YRI	~0.9	0.54	0.42	1	*RANGAP1*
chr8	130,827,234	ASN,CEU,YRI	9.1	0.54	0.37	42	*MLZE*

Supplementary Table 11 Composition of sequenced insertions

The repeat and duplication content of the sequenced insertions was determined using RepeatMasker and DupMasker. Results were compared to a data set of the same size randomly sampled from the genome. The 95% confidence interval based on 40 trials is indicated in parentheses.

	Sequenced Insertions	**NCBI build36** **(95% Confidence Interval)**
G+C	40.8%	40.9% (39.4% - 42.5%)
RepeatMasked	54.9%	47.2% (44.7% - 50.9%)
DupMasked	6.4%	6.4% (3.4% - 10.8%)
SINEs	12.6%	13.8% (11.9% - 16.3%)
LINEs	25.9%	20.3% (16.9% - 23.8%)
LTR elements	9.3%	8.3% (6.2% - 10.0%)

Supplementary Table 12 Comparison of sequenced insertions with GRCh37
Information is given for five regions which have been altered between the build36 and GRCh37 assemblies.

Fosmid Accession	Build36 Position	Gene	GRCh37 Status	GRCh37 Change
AC208058, AC210765	chr17:4733350	*MINK1*	Contains new sequence	BAC AC233723
AC231962	chr17:59798472	*PECAM1*	New gap, additional sequence missing	Removed BAC AC138744
AC221036	chr17:77112939	*FSCN2*	Contains new sequence	BAC AC137896
AC234039,AC231118	chr6:119258178	*ASF1A*	Contains new sequence	BAC AL359634 and HuRef ABBA01026024
AC232309	chr7:129939130	*COPG2*	Contains new sequence plus a gap	Fosmids AC144863, AC145656 and AC145213 plus new gap

Supplementary Note

1. Focused analysis of orphan clones from a single individual 2
FISH mapping G248 orphan contigs 2
2. Assembling novel sequence contigs from nine individuals 3
Novel sequence loci 3
Additional filters and removal of mapping artifacts 4
3. Custom oligonucleotide array targeting novel sequence contigs.......... 6
Filtering by hybridization intensity... 6
Filtering additional artifactual probes 7
4. Analysis of other human genomes.. 9
5. Analysis of non-human primates.. 10
6. Polymorphism analysis of novel sequences 11
Noise-multiplier approach 12
Contig genotyping based on cluster fitting 13
Comparison of alternative polymorphism calling schemes 14
7. High copy-number contigs 14
8. OligoFISH experiments 15
9. Analysis of sequenced clones 16
Variant genotyping using unique breakpoint k-mers 19
Capturing larger insertions using OEA clones... 21
10. Number of insertions represented in each sample............................. 21
11. Comparison with Illumina SOAP *de novo* assembly 23
Comparison with individual OEA end sequences 23
Comparison with assembled novel insertion contigs..................................... 23
Comparison with sequenced NA18507 insertions 25
12. References.. 29

1. Focused analysis of orphan clones from a single individual

We conducted an in-depth analysis of orphan clones from a single individual (G248 library, sample NA15510). We identified 4,773 clones where neither end maps against the build35 genome reference despite the presence of high-quality sequence at both ends and where both ends contained at least 100 bp of non-repeatmasked sequence. This requirement removed ~55,000 clones corresponding to alpha-satellite sequence from consideration. We selected 1,499 fosmids for restriction fingerprint analysis based on comparisons with chimpanzee WGS data. Useable data from four restriction enzymes was obtained for 1,378 of these clones. We used the Contig Builder program to link together clones into larger contigs based on the restriction map [1]. This resulted in 13 contigs that were formed from 10 or more clones (Table 1.1). These 13 largest contigs account for 82% of the analyzed clones and have a total spanned size of 4.292 Mb. An additional 125 fosmids (9%) are in contigs with three or more clones.

Contig	Clones/Contig	Contig Size (bp)	Clone Depth
1	277	836,359	13.8X
2	209	653,199	11.9X
3	157	519,126	12.5X
4	134	508,651	10.9X
5	117	478,474	10.1X
6	49	237,043	8.0X
7	45	162,090	10.7X
8	42	208,759	8.5X
9	36	179,769	8.4X
10	19	140,018	5.8X
11	15	135,634	4.6X
12	13	106,989	4.9X
13	11	126,681	3.8X
Total	1124	4,292,792	

Table 1.1 Large contigs built from orphan fosmids from the G248 library.

We had been concerned that many orphan clones would come from large blocks of uncharacterized highly repetitive DNA that is not present in the human reference sequence. A reassuring feature of these data is that most of the orphan fosmids assembled into a small number of contigs with reasonable values for depth-of-coverage (i.e., the larger contigs have depths comparable to the 10X depth of the G248 fosmid library). After further correcting for likely contaminants (such as Epstein-Barr virus and bacterial sequences), we identified a total of 72 physical contigs encompassing 479 clones. These contigs are estimated to encompass 3.9 Mb of sequence.

FISH mapping G248 orphan contigs

Using individual orphan clones as probes, we determined the position of 68 of these contigs by FISH. We found that 22% (15/68) of the contigs mapped interstitially, 46% (31/68) were associated with genome assembly gaps (based on a targeted study of existing assembly gaps and overlapping clone sequences[1]), and the remainder mapped to telomeric, pericentromeric, or acrocentric positions (Supplementary Table 3). This analysis indicates that a substantial amount of uncharacterized euchromatic sequence may be recoverable using the fosmid clone library resource, and demonstrates the ability to link together large elements by considering orphan clones.

2. Assembling novel sequence contigs from nine individuals

We identified 44,415 high-quality fosmid end sequences from nine individuals that do not map onto the human genome reference sequence (NCBI build35)[2]. This set includes individual sequences from 26,001 one-end anchored clones (OEA) and 9,207 orphan clones. Combined analysis of orphan and OEA clones permits the capture of new insertions larger than the clone insert size (40 kb) as well as anchoring information that can be used to place sequences within the reference assembly. Using phrap (http://phrap.org), we assembled these 44,415 sequences into an initial set of 3,963 sequence contigs (total size=4,465,116 bp; max=4,647 bp; N_{50}=1,148 bp). Over half of the contigs (2,034/3,963) contain sequence contributed by at least one orphan clone, suggesting that they represent short segments of unrepresented sequences that are longer than 40 kb.

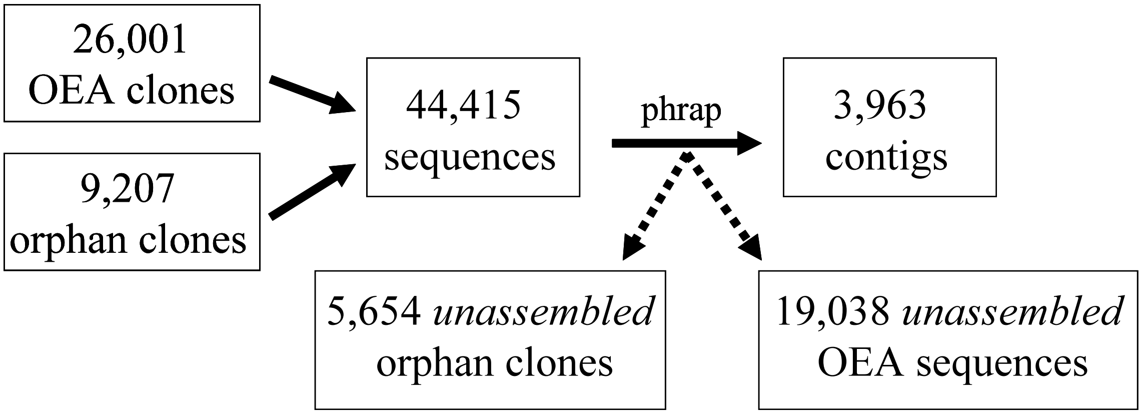

26,001
OEA clones
9,207
orphan clones
44,415
sequences
phrap
3,963
contigs
5,654 *unassembled* orphan clones
19,038 *unassembled* OEA sequences

Figure 2.1 Flow chart of sequence assembly procedure. The 3,963 sequence contigs include contributions from 6,963 OEA and 3,553 orphan clones.

Novel sequence loci

Based on genomic position and orphan clone contributions, we reduced the 3,963 sequence contigs into a set of non-redundant insertion loci. The 3,963 sequence contigs include sequence from 3,553 orphan clones. 51% (2,034/3,963) of the contigs contain sequence contributed by at least one orphan clone, with 1,888 orphan clones contributing to multiple contigs. Mate-pair information from these orphan clones can link 1,677 of the contigs into scaffolds. In total, this information reduces the 3,963 contigs to 2,626 scaffolds, of which 13% of (340/2,626) encompass more than one contig. 59% of the contigs (2,324/3,963) have anchoring information from OEA clones (contigs formed by multiple OEA clones are required to have consistent anchoring within 100 kb). Additionally, we defined 2,354 clusters of anchored contigs by merging contigs anchored within 50 kb of each other (approximately 3 standard deviations above the mean fosmid insert size). Of these, 281 include multiple contigs. The contig scaffolds make use of paired information from orphan clones and the clustered positions make use of anchoring information from OEA clones. Combining these two methods, the 3,963 contigs are reduced to a set of 1,182 loci.

Additional filters and removal of mapping artifacts

We applied several additional filters to identify potential artifacts among the 3,963 assembled contigs. First, we conducted an additional computation search comparing the sequences against other sequence databases from GenBank (the nt and HTGS databases). We identified and removed contigs having high identity hits against non-primate genome sequences (such as yeast, mouse, cat and pig) as well as bacterial contaminants. Furthermore, we used the results of array comparative genomic hybridizations to remove additional contaminants as well as sequences that were recalcitrant to effective CGH probe design (described in Section 3 below). Following these steps a total of 2,363 contigs remained (Table 2.2).

Criteria	Number of Contigs	Total Contig Bp	Number of Contigs With Anchor	Number of Loci	Number of Loci With Anchor
Assembled	3,963	4,465,116	2,283	1,182	587
Apply Computational Filter	3,702	4,197,374	2,270	1,063	578
Apply Experimental Filter	2,363	2,834,149	1,551	720	418
Apply NA18507 Artifact Filter	2,363	2,834,149	1,490	720	400

Table 2.2 Results of additional filtering steps. The number of contigs, the number of contigs having a clear genomic anchoring, and the associated number of loci and anchored loci are shown after the successive application of each filter. The NA18507 artifact filter applies only to inferred genomic positions and does not remove any contig sequences.

Several lines of evidence indicate that there is a high rate of clone chimerism in the NA18507 (ABC8) clone library. We observe an 8-fold greater fraction of clones with ends mapping to different chromosomes for this library (Table 2.3). Although a small fraction of these trans-chromosomal mapping clones may represent real rearrangements, the majority are likely the result of rearranged clones.

Library	Sample	Number of Clones with Unique Mapping	Number of Uniquely Mapped Trans-chromosomal Clones	Percent Trans-chromosomal
G248	NA15510	594,609	4,135	0.70%
ABC7	NA18517	616,947	15,811	2.56%
ABC8	NA18507	1,050,579	75,268	7.16%
ABC9	NA18956	738,786	4,366	0.59%
ABC10	NA19240	741,949	5,305	0.72%
ABC11	NA18555	724,998	7,049	0.97%
ABC12	NA12878	755,087	3,728	0.49%
ABC13	NA19129	757,837	4,889	0.65%
ABC14	NA12156	782,310	3,055	0.39%

Table 2.3 Fraction of clones from each library with ends mapping to different chromosomes.

We FISH mapped 24 contigs with a predicted location from a single NA18507 anchor. Only one of these contigs mapped to the predicted location while five mapped to a different chromosome and 18 mapped to the p-arms of the acrocentric chromosomes despite a predicted anchoring elsewhere. The intensity of the signals on the acrocentric chromosomes precludes a clear assessment of other hybridization signals that may be present elsewhere in the genome. BLAST searches show that 13 of these contigs match sequenced BACs not included in the genome assembly that have been assigned to 22p [3]. Seven of these contigs also correspond to a group of contigs that show a correlated pattern of drastically increased copy-number based on arrayCGH.

We also examined the complete sequence of an OEA clone from NA18507 (ABC8-43024000F4, AC226835). This clone contains sequence that matches chromosome 6 as well as sequence from a BAC (AL592188) assigned to 22p. PCR primers that amplify DNA isolated from the clone fail to amplify NA18507 genomic DNA, indicating that this clone represents a rearranged structure.

Since assembly requires the presence of overlapping sequence reads, clone artifacts may be enriched among assembled insertions involving sequences from the p-arms of acrocentric chromosomes. This may occur if rearranged clones containing sequence from rDNA repeat arrays are more likely to overlap by chance with true orphan clones representing those sequences. Future studies involving assembly of unmapped sequences derived from high-throughput sequencing should be aware of such mapping and cloning artifacts.

Based on these results we conclude that clone chimerism in the NA18507 library could result in a large fraction of mismapped loci. We note that nearly 50% more clones were obtained from this library than from the others. It is possible that the efforts to increase clone yield from this library led to the increased rate of artifacts. We have therefore removed from analysis all mapping positions for contigs anchored by a single NA18507 OEA clone.

3. Custom oligonucleotide array targeting novel sequence contigs

We designed a custom oligonucleotide array with 54,931 probes targeting the novel sequence contigs using relaxed probe design criteria in order to maximize coverage of contigs. We performed a series of arrayCGH experiments comparing individuals from the HapMap [4] project using sample NA15510 as a common reference. This comparison included all nine individuals that were used for sequence discovery. We filtered probes based on two metrics to eliminate contaminants and produced a high-performing array.

Filtering by hybridization intensity

We assessed the fluorescence hybridization intensity for each probe relative to the background noise level. We examined the distribution of processed fluorescence signals for X-linked probes in male samples and found that a cutoff on the processed signal of 256 'counts' provided clear separation between single copy probes and the background noise as indicated by a set of x-chromosome catalog CGH probes (Figure 3.1). Any probe that did not have a maximum processed signal count above this threshold for any of the nine samples used for the sequence discovery phase was removed from further analysis (Figure 3.2). We employed a similar procedure for a separate array targeting the unassembled single OEA sequences. Since processed signals on that array were slightly lower, a threshold of 181 counts was used.

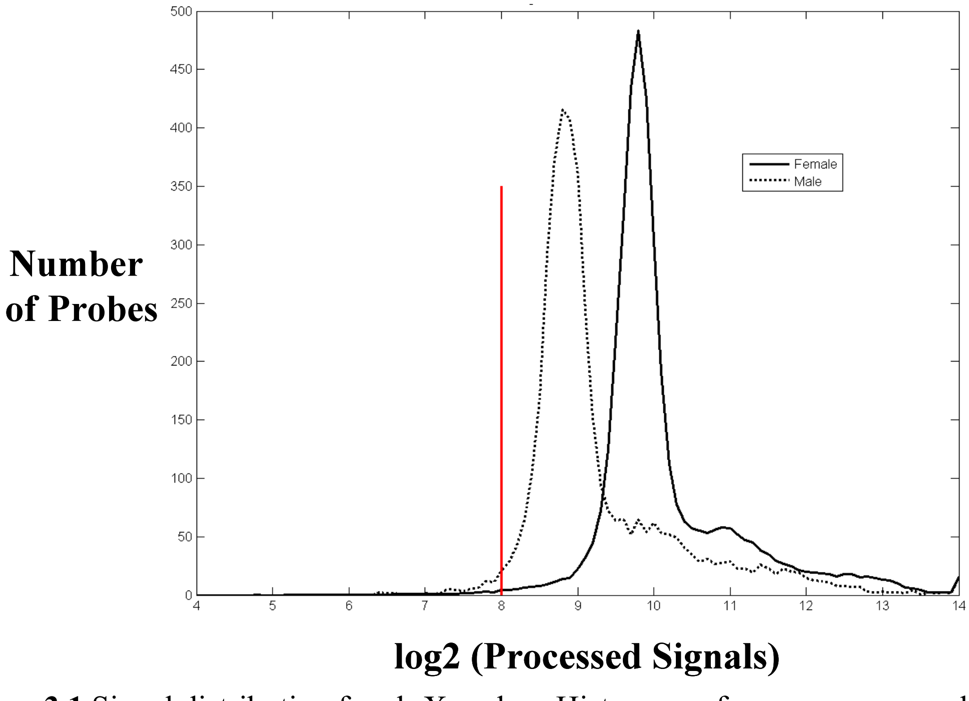


Figure 3.1 Signal distribution for chrX probes. Histogram of average processed signal counts are shown separately for females (solid line) and males (dashed line). For the novel sequence contigs a value of 256 signal counts (red line, log2(256)=8) was chosen as a threshold.

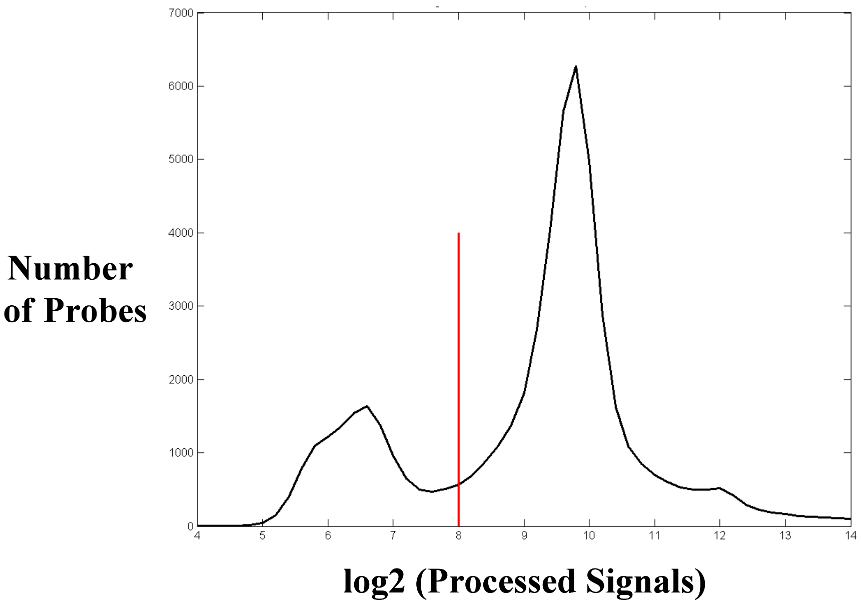


Figure 3.2 Signal distribution of novel sequence probes. Approximately 16% of the probes did not have a signal above the threshold for any of the nine samples used for discovery and were therefore removed from further analysis.

Filtering additional artifactual probes

An examination of the intensity plots across individual array experiments identified a subset of probes having a variable pattern of intensities that is consistent across all samples but independent of assigned genomic position. We hypothesized that these probes may be enriched for artifactual performance due to abnormal hybridization characteristics. In order to investigate this possibility, we defined clusters of similarly performing probes based on the Pearson correlation of log2 ratios across experiments using CAST [5] (Figure 3.3). The three largest clusters contained probes from many different contigs that behaved in a correlated manner independent of contig or chromosome assignment (Figure 3.4).

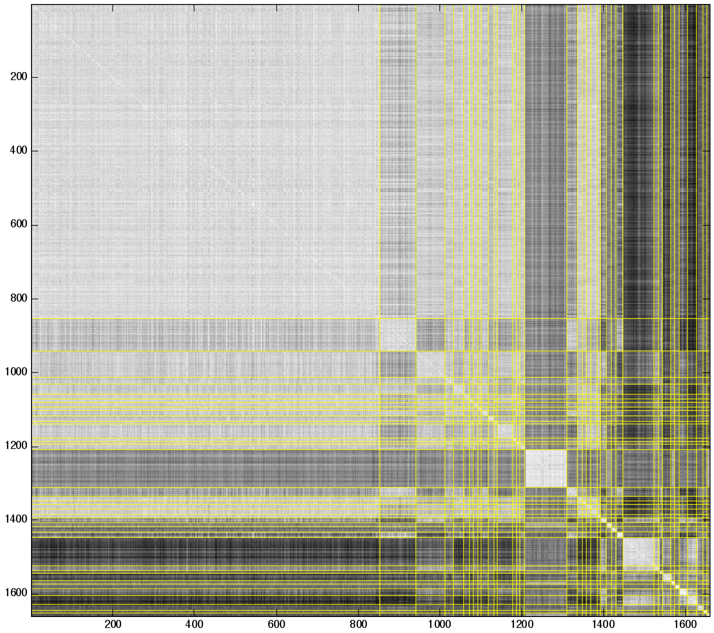


Figure 3.3 Correlation matrix of log ratios for a subset novel sequence probes. Stronger correlations are indicated by a lighter color. Clusters were defined by a minimum correlation threshold of 0.70 and a minimum cluster size of eight probes. The three largest clusters contain correlated probes from contigs that mapped to multiple different chromosomes (Figure 3.4).

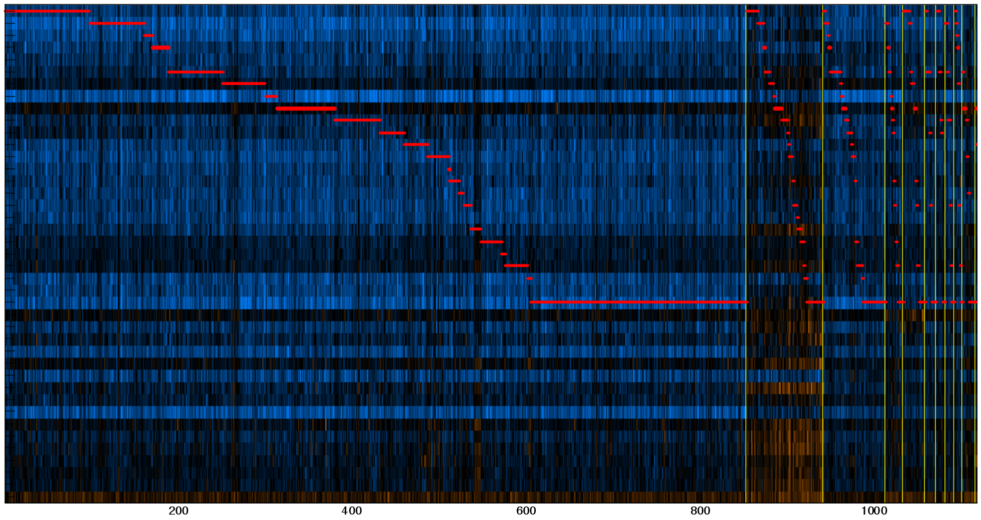


Figure 3.4 Heat map of log2 ratios for the 10 largest clusters. Each row represents a different hybridization experiment and each column is an individual probe. Blue color indicates lower hybridization intensity relative to the reference sample, black indicates the same hybridization intensity, and orange indicates increased hybridization. Vertical yellow lines separate probes assigned to different clusters. Within each cluster, probes are ordered based on their genomic position (based on mapping of OEA clones). The horizontal red lines indicate the chromosomal assignments, with the lowest line indicating unassigned probes. Note that probes mapping throughout the genome show

similar intensity patterns across samples. The artifactual signal we attempt to filter out is represented by the three largest clusters.

We concluded that these clusters represented probes having a reproducible, but artifactual, behavior. We removed all probes that had a correlation with one of the artifactual signatures of 0.7 or greater. Many of these artifactual probes do not pass Agilent's commercial probe design criteria.

Only contigs that were represented by at least three probes that passed all of these criteria were considered for further analysis. Although this strict quality control removes real human sequences from consideration, the applied metrics permit an assessment of copy-number differences among the individuals while reducing false classifications due to sequence and array artifacts.

4. Analysis of other human genomes

We compared the novel sequence contigs against other genome sequences using several approaches. First, we compared the contigs with additional human genome assemblies using megaBLAST (blastall version 2.2.11 with the following options: -e 1e-50 -F F -n T -b 100). We found that a substantial number of contigs that did not pass the arrayCGH probe filtering criteria nonetheless have hits against other human genome assemblies (Table 4.1) thus suggesting that our contaminant and validation filters are conservative and that there are real human sequences that we have excluded.

	All Contigs (n=3,963)	**Pass All Filters (n=2,363)**
NCBI build 35 (>=100 bp, >= 98%)	4	3
NCBI build 36 (>=100 bp, >= 98%)	336	221
GRCh37 (>=100 bp, >= 98%)	995	600
HuRef (>=100 bp, >= 98%)	2,084	1,467

Table 4.1 Comparison of assembled contigs with other human genome assemblies. The four contigs that map against build35 reflect the difference in mapping individual sequence traces as opposed to larger assembled contigs.

We assessed additional mapping information provided by other human assemblies by examining the 320 loci without an assigned build35 position. 175 of these 320 loci consist solely of 'orphan-only' contigs; the remainder are unassigned as a result of artifacts in the ABC8 library (see Supplementary Note section 2) or inconsistent mapping positions among clone-ends, such as occurs when traces have a best match to different copies of repeated or duplicated sequences. We searched end sequences from the individual orphan clones corresponding to these 175 loci against additional human genome assemblies. We found that 21 of these loci have a mapping against the build36 genome assemblies, with 14 of the 21 corresponding to the pseudo-autosomal region of the X and Y chromosomes. Similarly, we observe hits for 54 of the loci against the HuRef assembly. Clones from 35 loci match chromosomal segments from the GRCh37 assembly, with another 36 loci having matches to unplaced sequence contigs included in GRCh37.

Since over 60% of the contigs mapped to the HuRef genome assembly [6], we explored the presence of these sequences in WGS data from additional human genomes. Using megaBLAST we searched against 74.2 million 454 pyro-sequencing reads from the JDW genome [7] and found matches for 2,001/2,363 hits (>=100 bp, >=98% identity). We also analyzed Illumina sequence data from the YH and NA18507 genomes using mrFAST [8-10]. We considered as present any contig having an estimated median copy number of at least 0.5 based on mapped read depth. By these criteria, 1,716/2,363 contigs are present in YH and 1,698/2,363 in NA18507.

5. Analysis of non-human primates

We searched for the presence of the 2,363 contigs in other primate species using two approaches: (1) a bioinformatics search of genome sequence data from chimpanzee and orangutan and (2) an arrayCGH experiment comparing a single chimpanzee with sample NA15510.

We searched the 2,363 contigs against three data sets: the most recent chimpanzee genome assembly (pantro2), chimpanzee whole-genome shotgun (WGS) sequence reads (31.3 million reads, 29.3 Gbp), and orangutan WGS reads (25.5 million reads, 21.7 Gbp). All searches were performed using megaBLAST (blastall version 2.2.11) with the following options: -e 1e-50 -F F -n T -b 100. RepeatMasked contigs were used to search the WGS databases. For chimpanzee, we required that hits be at least 100 bp in length with 97% sequence identity. For orangutan, we used a reduced threshold of 100 bp and 95% sequence identity. We found sequence matches for 68% (1,599/2,363) of the contigs in chimpanzee and for 52% (1,217/2,363) in orangutan, with 45% (1,071/2,363) found in both species. Interestingly, only 62% (989/1,599) of the sequences with matches against PTR sequence data were found by searches against both the genome assembly and the WGS reads (Figure 5.1).

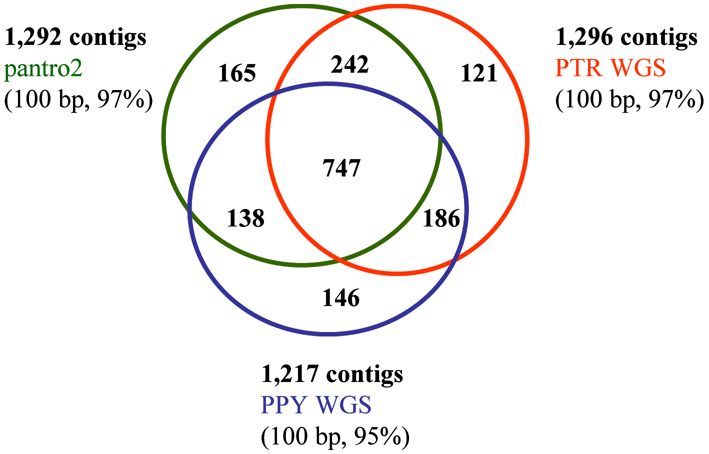


Figure 5.1 BLAST searches of 2,363 novel sequence contigs. The contigs were searched against the chimpanzee genome assembly (pantro2) and individual WGS reads from chimpanzee (PTR) and orangutan (PPY). Contigs were RepeatMasked before searching against WGS databases.

As a further test we performed an arrayCGH experiment using DNA from a single chimpanzee. Based on the single-channel intensity data, 84% (1,985/2,363) of the contigs have an estimated copy number of at least 1.0 in chimpanzee. This includes 85% (1,361/1,599) of the contigs with matches to chimpanzee genome sequence data, as well as an additional 624 contigs. Combining array results with the sequence searches we find evidence for 94% (2,223/2,363) of the contigs in chimpanzee and 96% (2,266/2,363) in either chimpanzee or orangutan.

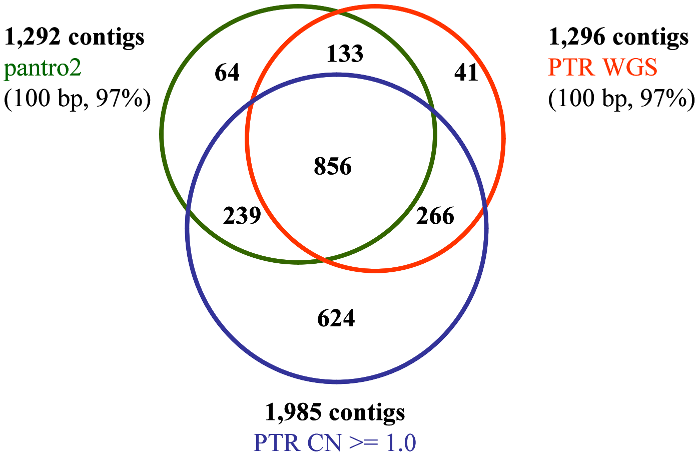


Figure 5.2 Comparison of 2,363 contigs with chimpanzee sequence data and array intensity.

6. Polymorphism analysis of novel sequences

We assessed the polymorphism of the 2,363 contigs that passed all arrayCGH and contaminant criteria among 28 unrelated HapMap samples hybridized against a common female reference (sample NA15510, the source of the G248 fosmid library). We identified polymorphic contigs using two different approaches.

Sample ID	Population	Sex
NA10847	CEU	Female
NA10851	CEU	Male
NA11832	CEU	Female
NA11840	CEU	Female
NA11993	CEU	Female
NA12004	CEU	Female
NA12156	CEU	Female
NA12813	CEU	Female
NA12878	CEU	Female
NA18552	JPT+CHB	Female
NA18555	JPT+CHB	Female
NA18564	JPT+CHB	Female
NA18573	JPT+CHB	Female
NA18942	JPT+CHB	Female
NA18947	JPT+CHB	Female
NA18956	JPT+CHB	Female
NA18980	JPT+CHB	Female
NA18502	YRI	Female
NA18507	YRI	Male
NA18517	YRI	Female
NA18523	YRI	Female
NA18861	YRI	Female
NA19102	YRI	Female
NA19116	YRI	Female
NA19129	YRI	Female
NA19132	YRI	Female
NA19172	YRI	Female
NA19240	YRI	Female
NA15510	Unknown	Female (reference sample)

Table 6.1 Human samples used in arrayCGH analysis.

Noise-multiplier approach

For this approach we used the median probe log2 ratio for each contig for each sample. For each sample we compared the median contig log2 value with the average of the median contig log2 values of the self-self hybridizations. If the difference between the sample log2 and the self-self log2 is at least N times greater than the root-square sum of the standard errors of the self-self and the sample hybridizations, we labeled a sample as being a gain or loss, as appropriate. Specifically, for each contig *c* and sample *i* we determined:

$M_{i,c}$: the median log2 ratio of the probes in contig *c* for sample *i*
$E_{i,c}$: the standard error of the log2 ratios measured for each probe in contig *c* in sample *i*

S_c : the mean of the median log2 ratios calculated for contig *c* from all self-self experiments
$E_{r,c}$: the standard error of the log2 ratios measured for each probe in contig *c* in the self-self experiments
N : the noise-multiplier threshold

Then, if $M_{i,c} > S_c + N^*(E_{i,c}^2 + E_{r,c}^2)^{0.5}$ we labeled sample *i* as being a 'gain' for contig *c* and if $M_{i,c} < S_c - N^*(E_{i,c}^2 + E_{r,c}^2)^{0.5}$ we labeled sample *i* as being a 'loss' for contig *c*. For this analysis, we used a noise multiple value of N=3, a threshold determined based on comparisons with intervals that clustered into distinct copy-number classes (described below).

This method produces a matrix of trinary values, each of which corresponds to the direction of each sample with respect to the reference state. These are interpreted as higher copy number, lower copy number or same copy number as the reference. Such data can be used to crudely estimate the polymorphism for the sample if it is assumed that the reference is in a well-defined common state. However, the noise across samples may appear as different copy number states. An important limitation of this approach is the absence of the assignment of individual genotypes. Also, this approach cannot discriminate between different copy-number states when all samples are higher or all are lower than the reference state. Nevertheless, using this method, 38% (890/2,363) of the contigs are identified as polymorphic among the 28 unrelated HapMap samples (Supplementary Table 5). Limiting analysis to the 26 unrelated females, this approach identifies 35% (834/2,363) as polymorphic.

Contig genotyping based on cluster fitting

Before summarizing the measurements for each contig, the individual probes with similar profiles across the samples are clustered using the CAST algorithm[5] using the Pearson correlation to compute similarity between probes. The CAST algorithm identifies clusters that have an average similarity between probes above a given threshold. An iterative approach is used to find the largest cluster with the highest clustering threshold. For each contig, an initial similarity threshold of 0.95 is applied. If no cluster containing more than 40% of the probes is found, the threshold is relaxed in increments of 0.05 until the largest cluster that has at least 40% of the probes is found, or until a threshold of 0.5 is reached, whichever comes first. If a large cluster is not found with at least 40% of the probes or a minimum of three probes, then all the probes are used for subsequent quantification over the contig. Using this approach, 840 (36%) of the 2,363 contigs were clustered, with an average of 64% of the probes clustered per clustered interval. The number of probes quantified is 30,469 (89%) of the total of 34,276 probes that pass the previous filters.

Using the probe sets identified by the above clustering procedure, median log2 ratios and signals are calculated for each sample and each contig. Median log2 ratios are then clustered across the collection of samples [11] and absolute copy-numbers are assigned to those contigs that cluster into distinct log-ratio clusters. This is done by fitting the median cluster values to the log2 ratios of distinct small copy numbers corresponding to states for both the samples and the reference (see Figure 2 in the main manuscript). This procedure is applied to contigs where the reference sample is not homozygously deleted. In cases where it is determined that the reference sample represents a homozygous deletion, based on an analysis of the reference channel signal level and the log ratios, copy numbers are assigned using a single-color approach. This approach replaces sample log2 ratios with the logs of the ratios of the sample signals to the median of those signals that are significantly above zero and then applies the fitting strategy described above.

Using this approach, we identified 518 contigs that are fitted to a copy-number state. Of these, 461 contigs (20% of the total 2,363 contigs) are fitted to two or more distinct states and are considered to be polymorphic. The copy numbers for these 461 cluster-fitted contigs are provided in Supplementary Table 6. Limiting analysis to the 26 unrelated females, this approach identifies 404 contigs as polymorphic.

Comparison of alternative polymorphism calling schemes

Comparing the results of these two approaches shows that only 49% (443/908) of the contigs identified as polymorphic were labeled as such by both methods (Figure 6.1). This is to be expected since only 22% (518/2,363) of the contigs could be fitted to a defined copy-number state. We note that 96% (443/461) of the contigs fitted to distinct copy-number states were also identified as polymorphic by the noise-multiplier method.

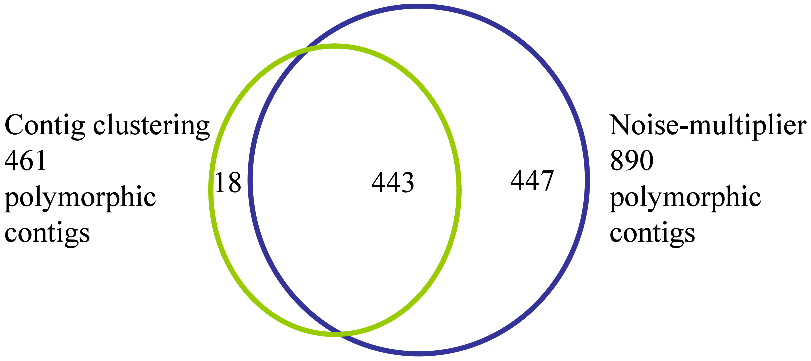


Figure 6.1 Comparison of contigs identified as polymorphic by both calling strategies.

7. High copy-number contigs

Cluster analysis identified 48 high copy-number contigs that showed a correlated ratiometric pattern across samples consistent with a high copy number.

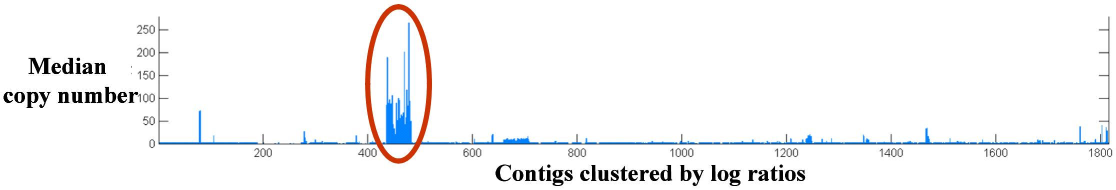


Figure 7.1 Identification of high copy-number cluster. Contigs were clustered based on the pattern of log2 ratios across samples (only clusters with two or more contigs are shown). The height of each bar corresponds to the median copy number estimated for each contig. The cluster consisting of 48 high copy contigs is circled in red.

Clones corresponding to 10 of these 48 high copy-number contigs have been mapped by FISH to the p-arms of the acrocentric chromosomes, with one mapping to the subtelomeric region on 10p. BLAST searches indicate that 20/48 contigs match the 43-kb rDNA repeat unit (U13369; >98% identity), and 28/48 contigs match sequenced BAC clones assigned to 22p. These results suggest that the high copy-number contigs largely correspond to sequences that are present in multiple copies on the p-arms of acrocentric chromosomes; with 20 of the contigs corresponding to sequences not represented in existing 22p BAC sequences [3].

8. OligoFISH experiments

We performed FISH using probes created from libraries of synthesized oligonucleotides targeted against the sequence of three insertion loci (AC217954, AC222569 and AC208058). The use of oligonucleotide-based probes as opposed to labeled fosmids permits the targeting of just the insertion sequence. Metaphase FISH confirmed the predicted genomic location for all three sequences. Each insertion sequence was interrogated for copy-number polymorphism by multiple individual sequence contigs. Two of the insertions (AC217954 and AC222569, Tables 8.1 and 8.2) showed consistent predicted copy numbers using the modal-clustering approach across all contigs and were confirmed by oligoFISH in four samples. In contrast, the six contigs assigned to AC208058 were not fitted into distinct copy-number classes. These contigs have an inconsistent status with three of the six contigs identified as polymorphic using the noise-multiplier approach. OligoFISH targeting this 5.4-kb insertion indicates a diploid copy number of 2 for three analyzed individuals (NA18507, NA18573 and NA18523).

Sample	Predicted CN (3 contigs)	oligoFISH CN
NA19240	0	0
NA12878	1	1
NA15510	1	1
NA18555	2	2

Table 8.1 Results of oligoFISH analysis for insertion AC217954. This insertion is represented by three sequence contigs, each of which has consistent called copy numbers.

Sample	Predicted CN (3 contigs)	oligoFISH CN
NA12156	0	0
NA18942	2	2
NA15510	1	1
NA12878	1	1

Table 8.2 Results of oligoFISH analysis for insertion AC222569. This insertion is represented by three sequence contigs, each of which has consistently called copy numbers.

9. Analysis of sequenced clones

Breakpoints were identified for the 222 sequenced fosmid clones based on a comparison with the build36 genome reference (Supplementary Table 10). First, the program miropeats[12] was used to identify approximate breakpoint positions (Figure 9.1).

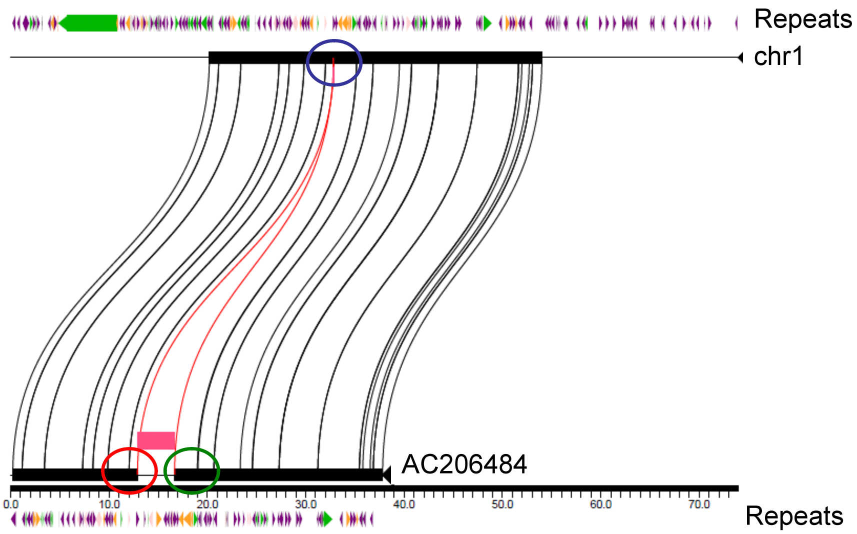
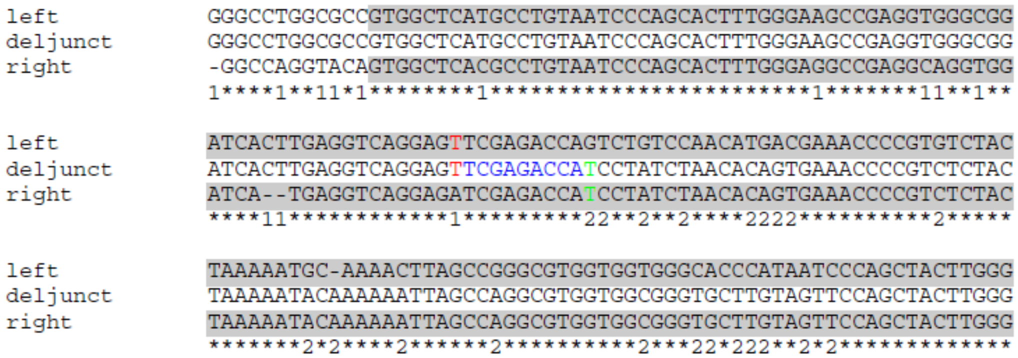


Figure 9.1 Initial breakpoint identification. The sequence of clone AC206484 is compared with chr1 using the program miropeats. Black lines indicate segments of matching sequence between the clone and the chromosome. An insertion in the clone, relative to the chromosome sequence, is identified by the magenta box. The curved magenta lines depict the approximate breakpoints. Sequences matching the left insertion breakpoint (red circle), the right insertion breakpoint (green circle), and the corresponding segment from the deletion haplotype (blue circle) are indicated.

The segments from the two edges of the insertion were extracted and aligned in turn with the corresponding sequence from the deletion haplotype. These alignments were then combined to form a three-way alignment. Using this alignment, the innermost positions at which the deletion fragment is a better match to the 'left' or 'right' side of the insertion region (Figure 9.2) are identified. In this manner, breakpoints are determined at nucleotide-level resolution. Additionally, comparisons of the sequence around the breakpoints are performed to identify additional segments of sequence homology encompassing the identified breakpoints.



```
left          GGGCCTGGCGCCGTGGCTCATGCCTGTAATCCCAGCACTTTGGGAAGCCGAGGTGGGCGG
deljunct      GGGCCTGGCGCCGTGGCTCATGCCTGTAATCCCAGCACTTTGGGAAGCCGAGGTGGGCGG
right         -GGCCAGGTACAGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCAGGTGG
              1****1**11*1********1************************1*******11**1**

left          ATCACTTGAGGTCAGGAGTTCGAGACCAGTCTGTCCAACATGACGAAACCCCGTGTCTAC
deljunct      ATCACTTGAGGTCAGGAGTTCGAGACCATCCTATCTAACACAGTGAAACCCCGTCTCTAC
right         ATCA--TGAGGTCAGGAGATCGAGACCATCCTATCTAACACAGTGAAACCCCGTCTCTAC
              ****11************1*********22**2**2****2222**********2*****

left          TAAAAATGC-AAAACTTAGCCGGGCGTGGTGGTGGGCACCCATAATCCCAGCTACTTGGG
deljunct      TAAAAATACAAAAAATTAGCCAGGCGTGGTGGCGGGTGCTTGTAGTTCCAGCTACTTGGG
right         TAAAAATACAAAAAATTAGCCAGGCGTGGTGGCGGGTGCTTGTAGTTCCAGCTACTTGGG
              *******2*2****2******2**********2***22*222**2*2*************
```

Figure 9.2 Breakpoint alignment. The resulting alignment for the sequences identified in Figure 9.1 is shown. A ‘1’ indicates a match between the sequence from the deletion fragment (‘deljunct’) and the left-insertion sequence. A ‘2’ indicates a match between the sequence from the deletion fragment and the right-insertion sequence. The variant breakpoints are defined by the innermost positions of clear match to the left or right segments (red and green aligned nucleotides, reported in Supplementary Table 10). In this example, these two positions are separated by 9 bp that perfectly matches at both insertion breakpoints (blue text). In Supplementary Table 10 this variant is reported as being class ‘c1’, indicating 9 bp of perfect identity. In contrast, a ‘c3’ variant contains 0 bp of identity and a ‘c2’ variant contains unmatched sequence on the deletion haplotype that is not present at either edge of the insertion. The dark grey shading indicates the extent of additional matching sequence at the two breakpoints. In this example, the additional homology extends for 284 bp and has a sequence identity of 84.15%.

The breakpoint coordinates identified from the alignment are used to define the size and extent of each variant. In the case of OEA clones, the matching segment at the other end of the insertion is not captured. Therefore, only one coordinate could be identified at the sequence level. The images shown in Figure S4 depict the bp-level resolved breakpoints as well as the additional extent of breakpoint homology and other annotations. An example for clone AC206484 is shown in Figure 9.3.

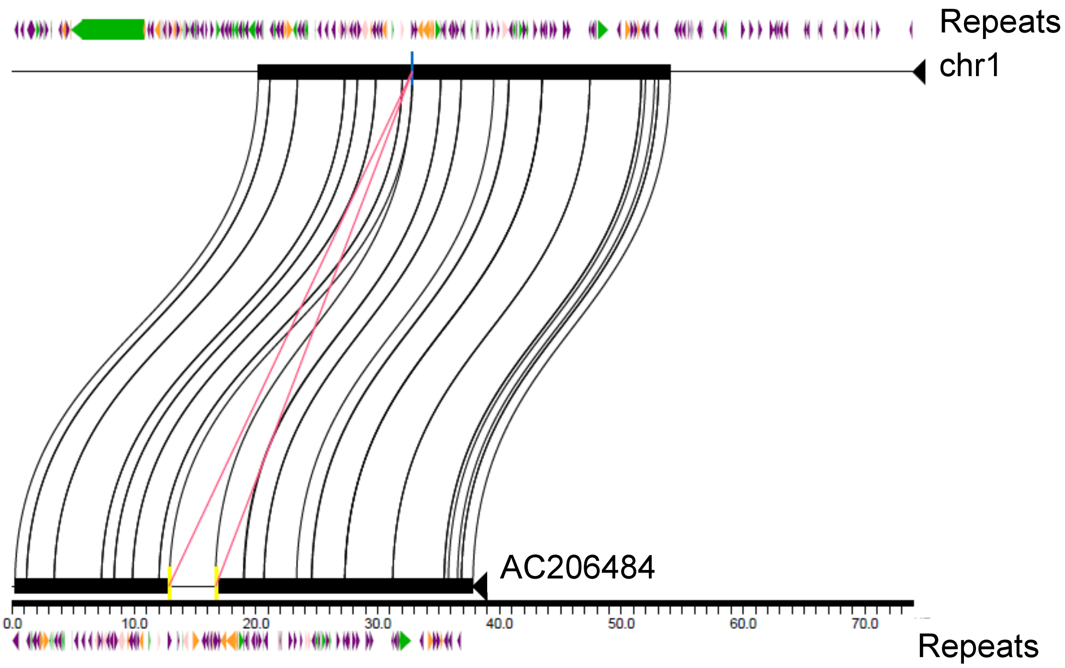


Figure 9.3 Final annotated breakpoint image. The final annotated breakpoint image for AC206484 is shown. The image is similar to that shown in Figure 9.1. However, the straight magenta lines correspond to the breakpoint positions identified by sequence alignment (the red and green position in Figure 9.2). The yellow boxes at the two edges of the insertion correspond to the 284 bp of matching (84.15% identity) sequence found between the two insertion edges. The thin blue box at the breakpoint on chr1 represents the 9 bp of perfectly matching sequence found on each side of the insertion and present on the deletion haplotype.

Variant genotyping using unique breakpoint k-mers

The sequence-resolved breakpoints from 152 insertions sequenced in individual fosmid clones were used to identify diagnostic k-mers specific to each variant. Comparison of the sequenced clones identifies three breakpoint segments: one on the build36 chromosome sequence (the 'deletion' allele) and two on the sequenced fosmid (the 'insertion' allele). A set of diagnostic k-mers was defined by searching all overlapping k-mers from each breakpoint against sequence data from the build36 assembly and the collection of insertion-containing fosmids. For this analysis, a k-mer size of 36 and one substitution was permitted in the searching. In order to be a considered diagnostic, a deletion k-mer must have a single match (including up to one substitution) to the build36 sequence and no matches against the fosmid sequences. Insertion k-mers were required to have a single match against the fosmid sequences and no matches against the build36 genome sequence. Using these criteria, 71% (108/152 loci) of the loci were represented by at least one deletion k-mer and one insertion k-mer (Figure 6B).

Next, Illumina sequence data[9] from NA18507 was searched against this collection of k-mers using mrsFAST (http://mrfast.sourceforge.net). Both the Illumina reads and targeted k-mers had a length of 36, and only perfect matches were recorded. The normalized number of reads supporting each allele is first determined since the deletion and insertion alleles may have a different number of diagnostic k-mers:

$$I = \frac{R_I}{T_I}$$

$$D = \frac{R_D}{T_D}$$

where T_I and T_D are the number of diagnostic k-mers for the insertion and deletion alleles of a given variant and R_I and R_D are the number of reads that match the diagnostic insertion or deletion k-mers. A breakpoint search score is then calculated using these normalized support counts:

$$\text{breakpoint search score} = 2\left(\frac{I}{I + D}\right)$$

A score of 2.0 will be calculated if there are no reads that match the deletion k-mers. Similarly, a score of 0.0 will result if there are no reads that match the insertion k-mers. If there are no reads that match either insertion or deletion k-mers then the breakpoint score is undefined and no genotype is determined. To define an integer genotype, the breakpoint search score is simply rounded. That is, variants having a score >= 0.5 and <=1.5 are assigned to the heterozygous (copy number=1) class.

Genotypes could be determined for 106 of the 108 variants that had diagnostic k-mers using Illumina sequence data from NA18507. The scores are reported in Supplementary Table 14. A histogram of these scores is shown in Figure 9.4.

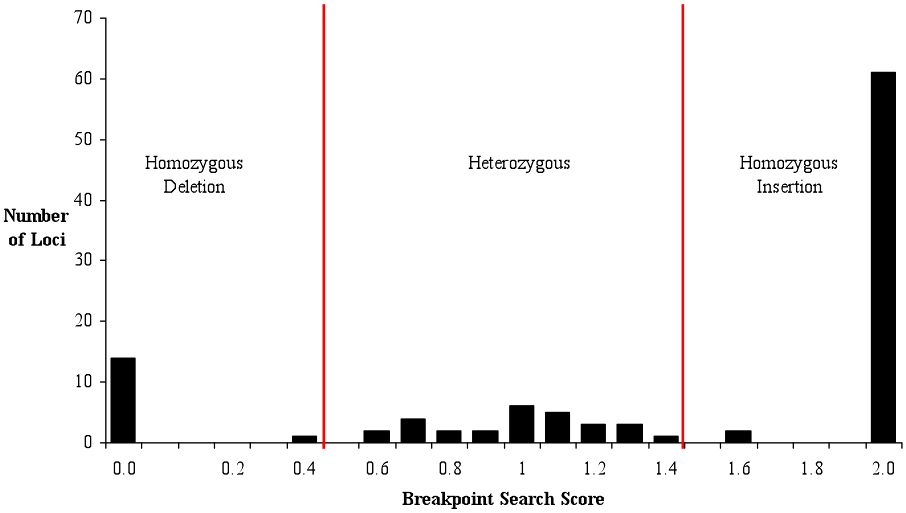


Figure 9.4 Breakpoint search score distribution for sample NA18507. A score was determined for sample sequenced variants in sample NA18507. Genotypes can be assigned by applying a score threshold of 0.5 and 1.5 (red lines). 53 of these variants were also assigned a genotype by arrayCGH. Applying these score thresholds results in 94.3% (50/53 variants) genotype agreement.

Capturing larger insertions using OEA clones

OEA clones that extend into an insertion can be used to capture the sequence of insertions that are greater than the 40-kb clone size. The analyzed sequences include eight loci flanked by sequenced OEA clones. OEA clone sequence overlaps indicate that the complete insertion sequence has been captured for four of these loci (Table 9.1).

Position (Mb)	Clones	Entire Insertion Captured	Insertion Size (bp)
chr3:57.3	AC232304, AC231288	Yes	48,436
chr6:51.2	AC233754, AC231198	No	> 59,325
chr6:107.4	AC233764, AC231117	No	> 75,851
chr6:119.2	AC234039, AC231118	Yes	65,026
chr10:27.6	AC231273, AC226171	Yes	47,298
chr18:63.2	AC231988, AC231982	No	>19,462
chr19:21.5	AC232224, AC236073, AC232302	No	> 38,556
chr20:53.5	AC232301, AC232307	Yes	41,476

Table 9.1 Summary of large insertion flanked by sequenced OEA clones. Coordinates are given relative to the build36 genome assembly.

10. Number of insertions represented in each sample

We estimated the yield of new insertion sequence likely to be discovered in additional genomes by considering how many of the 720 insertion loci were found in only a one of the nine individuals we used for sequence discovery. Because of the comparatively low sequence coverage of each genome (approximately 0.3X), not all individuals containing an insertion actually contributed unmapped end-sequences towards its discovery. We therefore combined the library source information of the individual clones used to discover each locus with arrayCGH genotyping results for the sequence contigs to determine which individuals contain each insertion. 56% of the total loci (401/720) were present in all nine individuals. This includes 240 loci that are not polymorphic among the 28 individuals analyzed by arrayCGH. 69 of the loci were present in just one of the nine analyzed individuals (Figure 10.1). If analysis is limited to the 400 loci with anchored positions in the euchromatin, we find that only 11 loci are present in only one of the individuals used for discovery (Figure 10.2). Thus, although additional genome projects are likely to uncover a large number of new insertions (as in the 7,240 single unassembled anchored sequence traces we identified, Supplementary Table 4), our results indicate that the majority of large novel insertion sequences have been captured using these nine individuals.

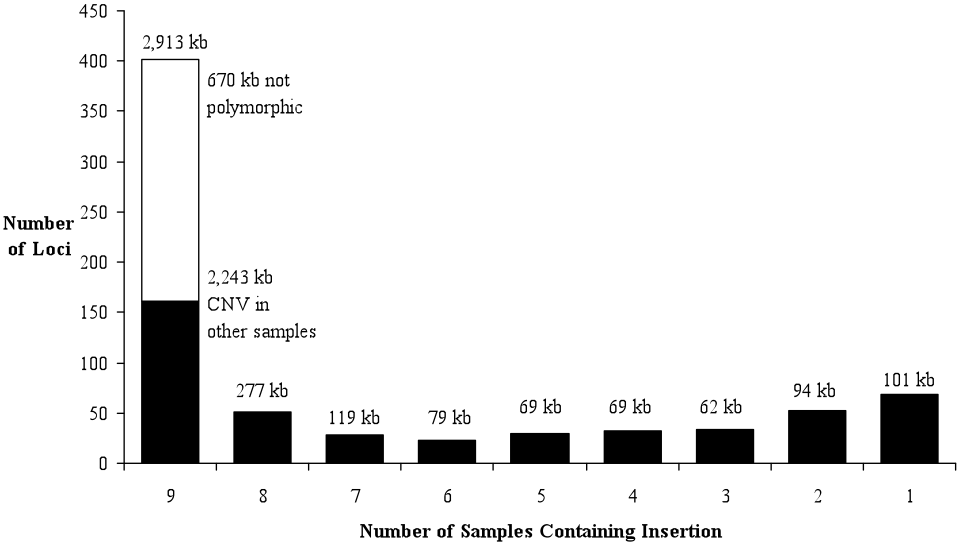


Figure 10.1 Distribution of 720 loci among nine individuals used for sequence discovery. The height of each bar indicates the number of loci found in exactly 1, 2, 3, etc. of the 9 individuals used for sequence discovery. There were 161 loci present (in at least one copy) in all nine individuals that were also identified as polymorphic. The white bar corresponds to the 240 loci that were not found to be polymorphic based on analyses of 28 individuals. An estimate of the total insertion size is given above each bar. This estimate is derived by summing the sizes of the individual contigs contributing to each locus and therefore should be considered to be a lower-bound estimate of the true insertion size.

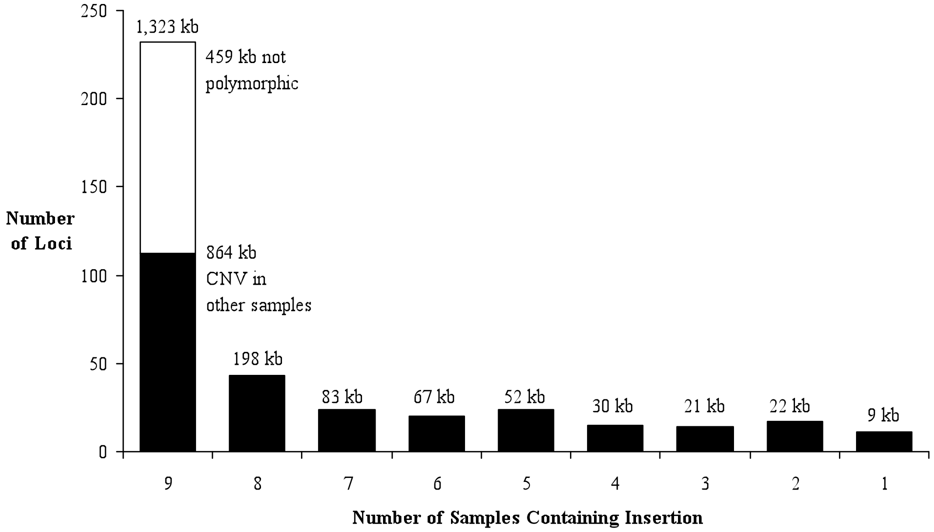


Figure 10.2 Distribution of 400 loci having an anchored map position among nine individuals used for sequence discovery. 11 loci were found in only a single individual. Insertion sizes were calculated as in Figure 10.1

11. Comparison with Illumina SOAP *de novo* assembly

We downloaded the novel insertion data reported in Li et al. [13] from http://yh.genomics.org.cn/download.jsp and made several comparisons with the data sets described in the manuscript.

Comparison with individual OEA end sequences

1,126 of the 7,240 OEA end sequences that passed our filters were derived from sample NA18507 (ABC8 clone library). We searched these 1,126 sequences against the novel-sequence contigs assembled from the Illumina data. Requiring a match of 100 bp with at least 98% identity, we find that 80 of these sequences have at least a partial match with novel sequences assembled from the YH genome and 108 have a match to the NA18507 genome. The Li et al. next-gen data set consists of contigs >=100 bp that do not match against the build36 assembly. Our analysis has been largely focused on the bd35 assembly, however we note that only 19 of the 1,126 ABC8 OEA sequences map onto bd36, indicating that this does not account for the discrepancy. We therefore conclude that we have identified a substantial amount of additional sequence in NA18507, although we recognize that Li et al. have identified many shorter sequences that we could not detect based on our 0.3X sequence coverage of this genome.

Comparison with assembled novel insertion contigs

We also compared the 2,363 assembled contigs, which include contributions from nine individuals, with the sequences reported in Li et al. We find that 68% of our contigs (1,602/2,363) have at least a partial match to the Li NA18507 data set.

Criteria	*Number of Contigs*
Match to NA18507	1,602
Match to YH	1,528
Match to NA18507 or YH	1,680

Table 11.1 Comparison of 2,363 assembled contigs with the Li et al. NA18507 and YH data sets.

Often, the Li sequences only represented a portion of the fosmid ESP contigs. We quantified this by calculating the fraction of each ESP-assembled contig that matched sequence in the Li data set.

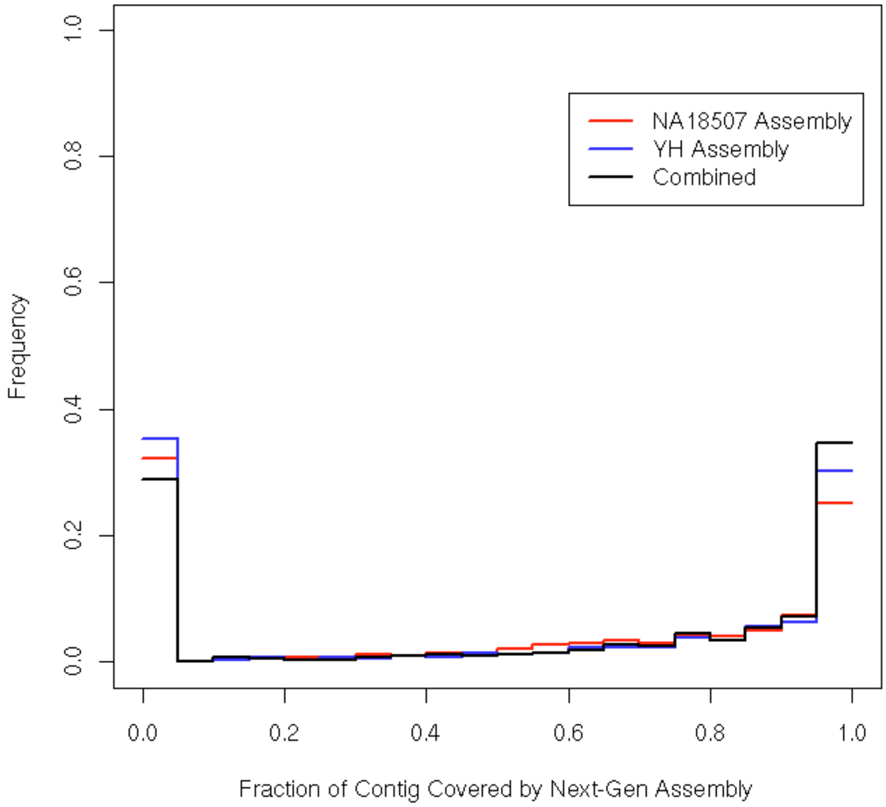


Figure 11.1 Histogram of the fraction of assembled OEA contigs that match sequences assembled by Li et al.

For the Li NA18507 data (the red line in Figure 11.1) we find that 32% of fosmid ESP contigs are not covered at all (761/2,363), and that 25% (591/2,363) have coverage of at least 95%. If this calculation is limited to only the 1,602 contigs with at least some coverage in the Li et al. NA18507 data set, we find that 37% (591/1,602) have a coverage of at least 95% and that the median fraction covered is 89%.

From these findings, we conclude that assembly of next-generation sequencing reads identifies a fraction of the sequences detected by fosmid ESP assembly but does not recapitulate the complete length of the insertions.

Comparison with sequenced NA18507 insertions

Additionally, we compared the Li et al. NA18507 sequences with insertion sequences obtained from 21 fully sequenced NA18507 fosmid clones. 17 of the 21 NA18507 insertions have at least some representation in the Li et al. NA18507 data set. Three of the four missing insertions involve GC-rich low-complexity sequence or satellite sequences. The fourth (clone AC233720) consists of LINE derived sequences as well as other elements.

The mean covered fraction of the 17 matching insertion sequences is 65%. We observe that 7 of the 17 fosmid insertions have matches to sequence from Li et al. that are assigned to more than one scaffold. The representation of contiguous sequence across multiple scaffolds would severely limit an understanding of the long-range continuity and structural organization of these sequences.

Clone Accession	Chrm	Position (bd36)	Clone Type	Length of Cloned Insertion Sequence	NA18507 Illumina coverage	NA18507 Contigs	NA18507 Scaffolds
AC225822	chr10	79,360,990	Not spanned	18,398	0.0%	0	0
AC231414	chr10	121,177,030	Not spanned	24,063	71.0%	26	5
AC213240	chr11	55,900,146	Spanned	7,491	84.9%	3	1
AC225617	chr11	71,989,867	Not spanned	1,413	0.0%	0	0
AC234852	chr12	1,047,608	Spanned	6,926	61.0%	9	2
AC234232	chr14	100,493,647	Not spanned	1,828	87.6%	2	1
AC234142	chr16	24,791,820	Not spanned	2,576	44.8%	4	2
AC225984	chr16	34,775,185	Not spanned	7,220	64.0%	5	1
AC231982	chr18	63,272,202	Not spanned	5,487	89.4%	6	1
AC232302	chr19	21,543,554	Not spanned	10,943	32.0%	7	4
AC236073	chr19	21,548,698	Not spanned	27,613	40.9%	9	4
AC231980	chr2	117,424,590	Not spanned	3,638	69.3%	2	1
AC226495	chr2	157,901,100	Not spanned	17,600	0.0%	0	0
AC233721	chr20	42,087,581	Not spanned	1,415	100.0%	1	1
AC232301	chr20	53,556,647	Not spanned	33,937	76.0%	10	2
AC231189	chr21	23,682,295	Not spanned	1,002	95.1%	2	1
AC233768	chr21	43,287,469	Not spanned	1,147	39.4%	1	1
AC225889	chr4	62,460,282	Not spanned	21,777	6.0%	6	5
AC234851	chr5	124,794,281	Not spanned	7,929	61.9%	9	1
AC233722	chr8	141,042,221	Not spanned	1,411	83.3%	2	1
AC233720	chrX	129,102,314	Not spanned	19,486	0.0%	0	0

Table 11.2 Comparison of sequenced NA18507 insertions with the Li et al. contigs.

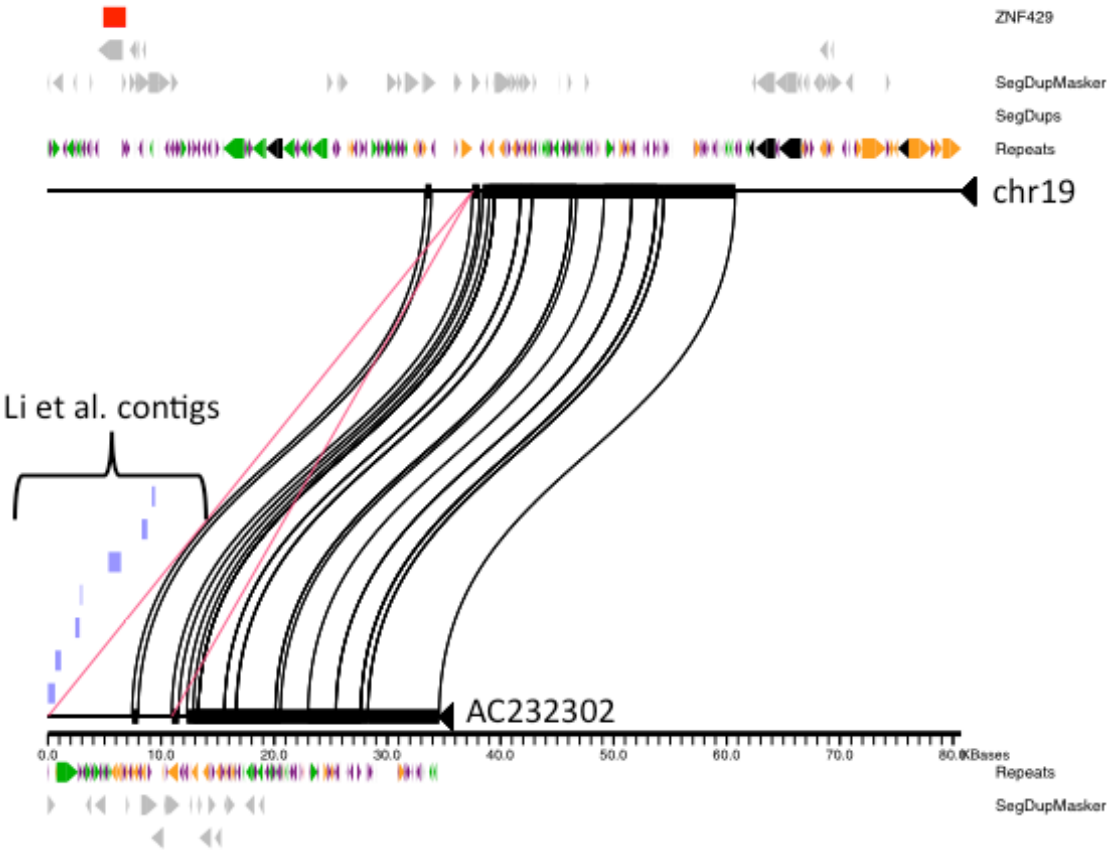


Figure 11.2 Comparison of clone AC232302 with Li et al. NA18507 assembled contigs The depicted clone extends 10.9 kb into an unspanned insertion on chr19. The purple rectangles represent the positions of NA18507 contigs from Li et al. mapped against the clone sequence. The seven mapped contigs from Li et al. represent 32% of the insertion sequence captured in the fosmid clone. These seven Li et al. contigs are assigned to four different scaffolds. This example illustrates a common problem in obtaining contiguity and correctly assigning location for novel insertions that are rich in repetitive or duplicated DNA.

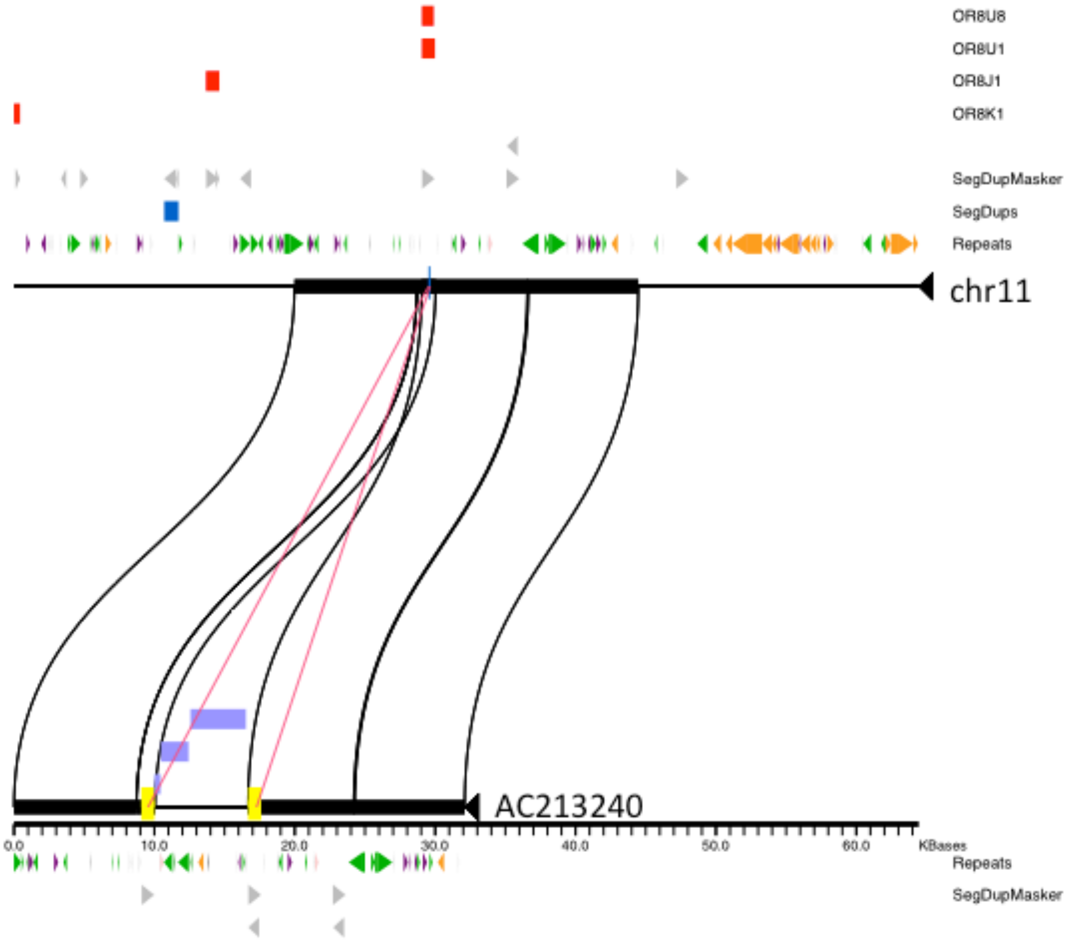


Figure 11.3 Comparison of clone AC213240 with Li et al. NA18507 assembled contigs This clone spans a 7.4 kb insertion on chr11. Three Li et al. contigs, assigned to a single scaffold, represent 85% of this insertion. This example illustrates the utility of next-gen *de novo* assembly for relatively unique regions.

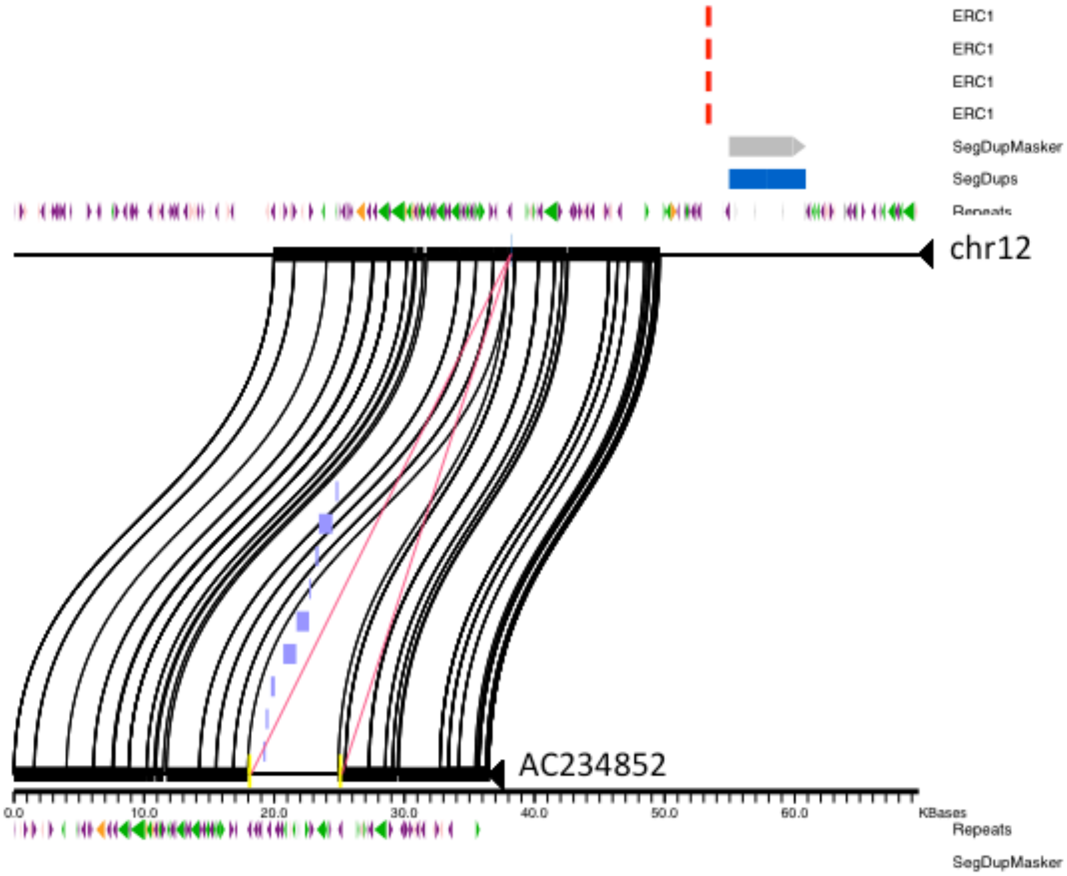


Figure 11.4 Comparison of clone AC234852 with Li et al. NA18507 assembled contigs This clone spans a 6.9-kb insertion on chr12. There are nine contigs from Li et al. that match this insertion. The nine contigs cover 61% of the insertion sequence and are assigned to two different scaffolds.

12. References

1. Bovee, D. et al. Closing gaps in the human genome with fosmid resources generated from multiple individuals. *Nat Genet* **40**, 96-101 (2008).
2. Kidd, J.M. et al. Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56-64 (2008).
3. Cole, C.G. et al. Finishing the finished human chromosome 22 sequence. *Genome Biol* **9**, R78 (2008).
4. IHMC A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-861 (2007).
5. Ben-Dor, A., Shamir, R. & Yakhini, Z. Clustering gene expression patterns. *J Comput Biol* **6**, 281-297 (1999).
6. Levy, S. et al. The Diploid Genome Sequence of an Individual Human. *PLoS Biol* **5**, e254 (2007).
7. Wheeler, D.A. et al. The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872-876 (2008).
8. Wang, J. et al. The diploid genome sequence of an Asian individual. *Nature* **456**, 60-65 (2008).
9. Bentley, D.R. et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-59 (2008).
10. Alkan, C. et al. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet* **41**, 1061-1067 (2009).
11. Perry, G.H. et al. The fine-scale and complex architecture of human copy-number variation. *Am J Hum Genet* **82**, 685-695 (2008).
12. Parsons, J. Miropeats: graphical DNA sequence comparisons. *Comput Appl Biosci* **11**, 615-619 (1995).
13. Li, R. et al. Building the sequence map of the human pan-genome. *Nat Biotechnol* **28**, 57-63.