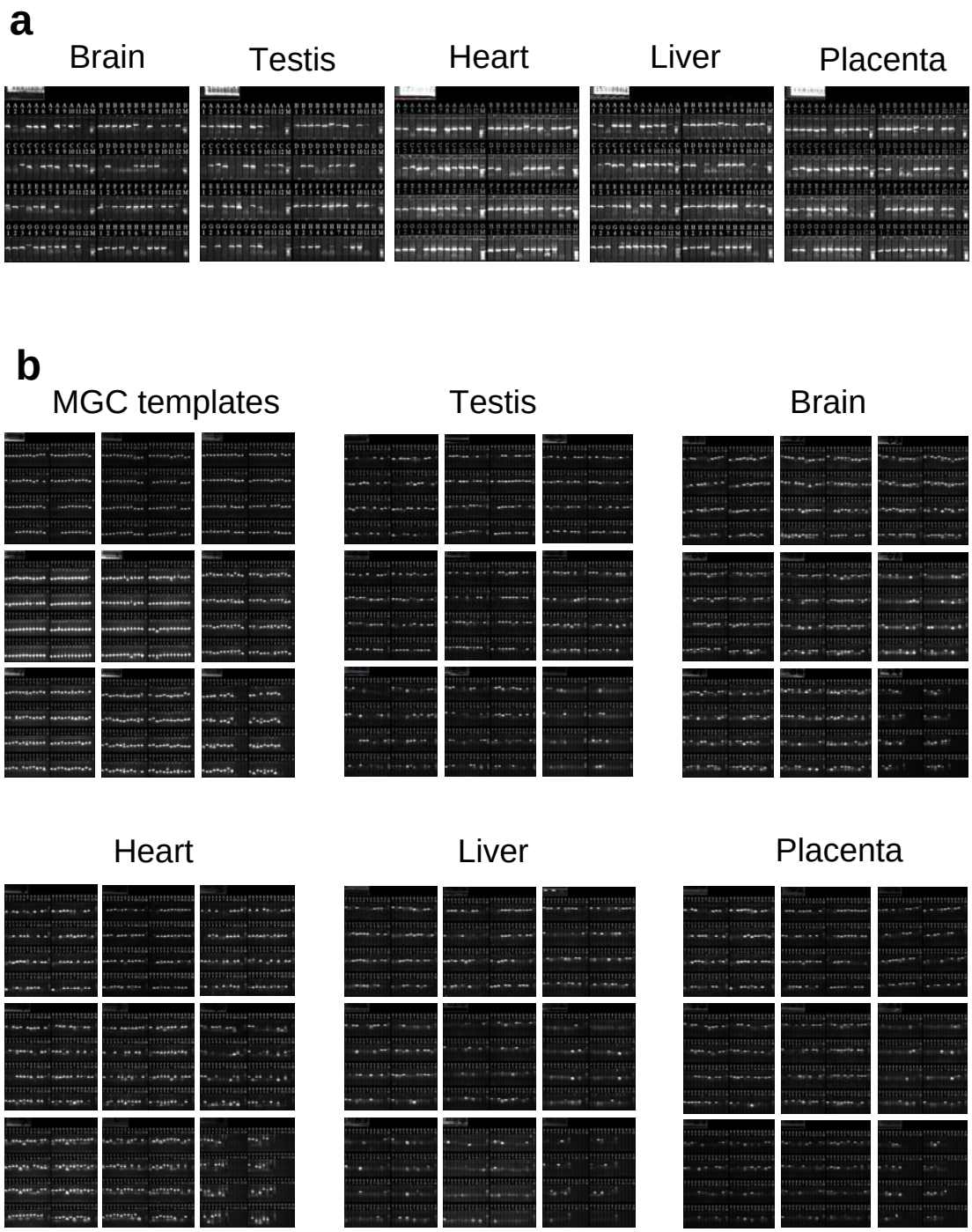
Isoform discovery by targeted cloning, ‘deep-well’ pooling and parallel sequencing

Kourosh Salehi-Ashtiani, Xinping Yang, Adnan Derti, Weidong Tian, Tong Hao, Chenwei Lin, Kathryn Makowski, Lei Shen, Ryan R Murray, David Szeto, Nadeem Tusneem, Douglas R Smith, Michael E Cusick, David E Hill, Frederick P Roth & Marc Vidal

Supplementary figures and text:

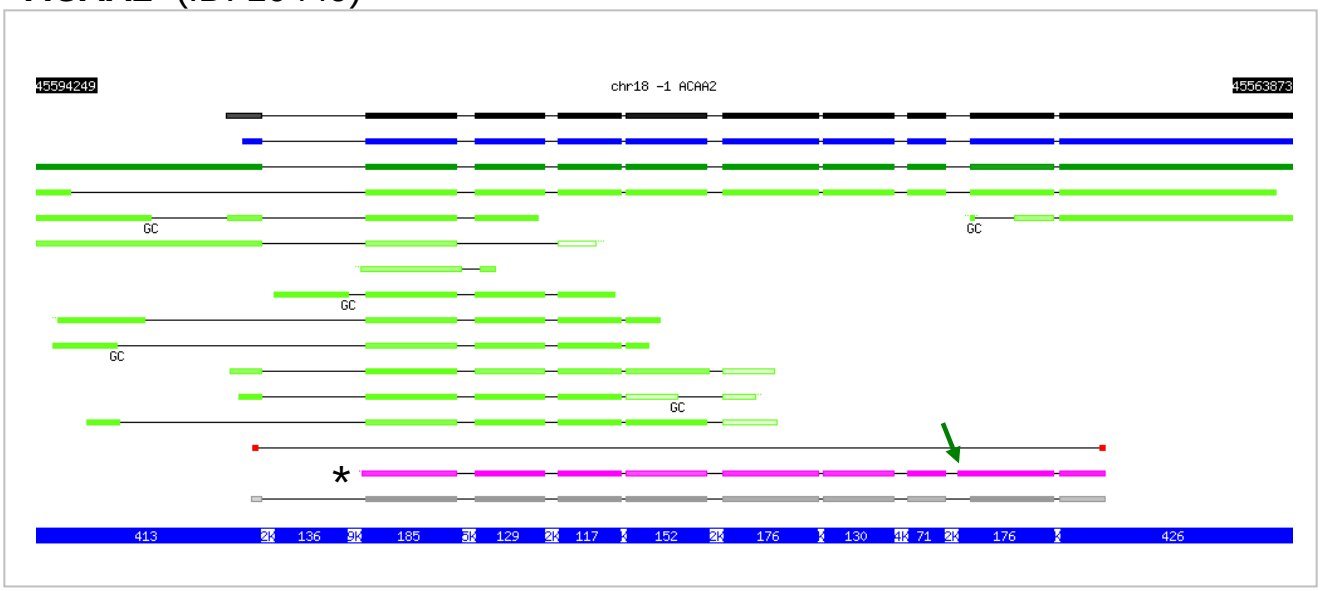
Supplementary figure 1



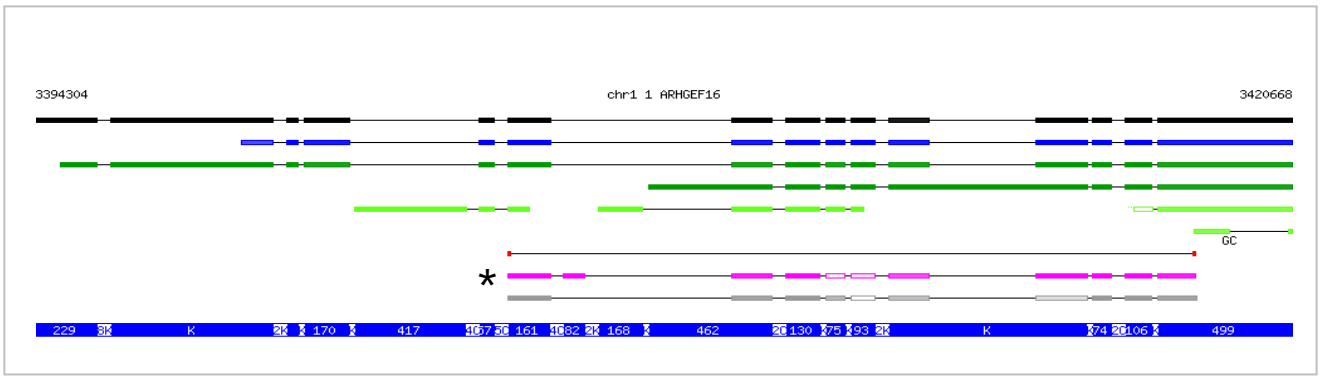
Large scale amplification of human ORFs. (**a**) A set of 94 ORFs (with expected sizes between ~1.1 and 1.4 kb) were amplified from the five tissues indicated. The RNA preparations were reverse transcribed, and the RT products used as template for PCR amplification using human ORFeome primers. PCR products for 22 of these reactions were pooled from the five tissues then recombinationally cloned. (**b**) Disease-related ORFs, as defined in OMIM10, were PCR amplified from reverse transcribed RNA or from cloned ORFs (indicated as “MGC template”). The latter were subsequently pooled as a single “deep well” and sequenced by the 454 FLX platform. ORF sizes of this set ranged from ~0.15 kb to 5.1 kb.

Supplemental figure 2a

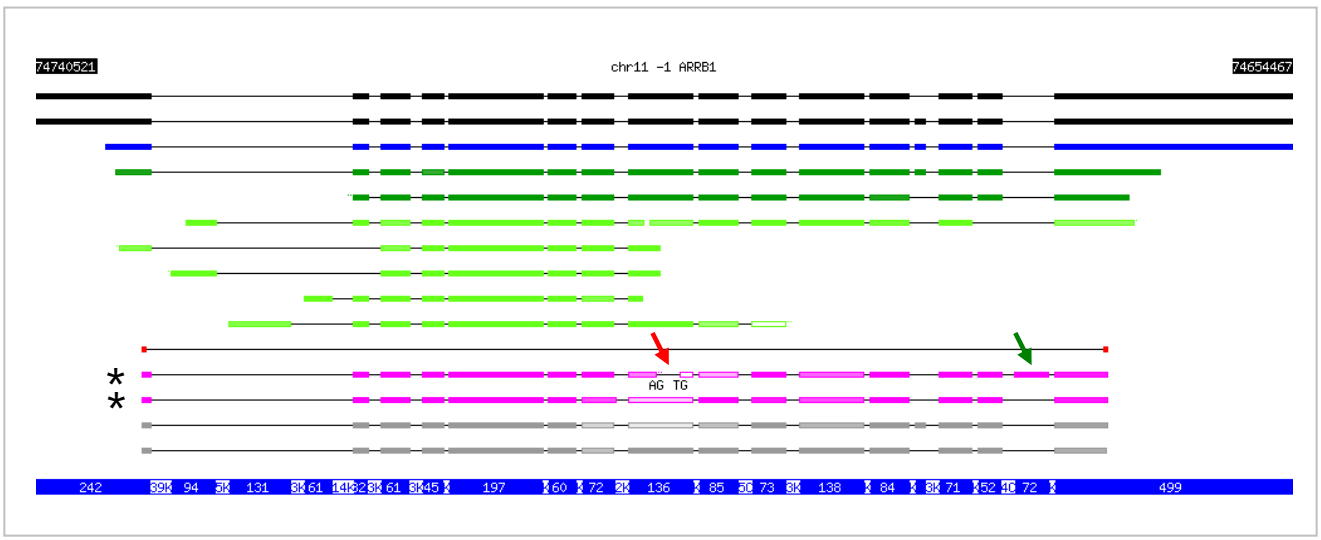
ACAA2 (ID: 10449)



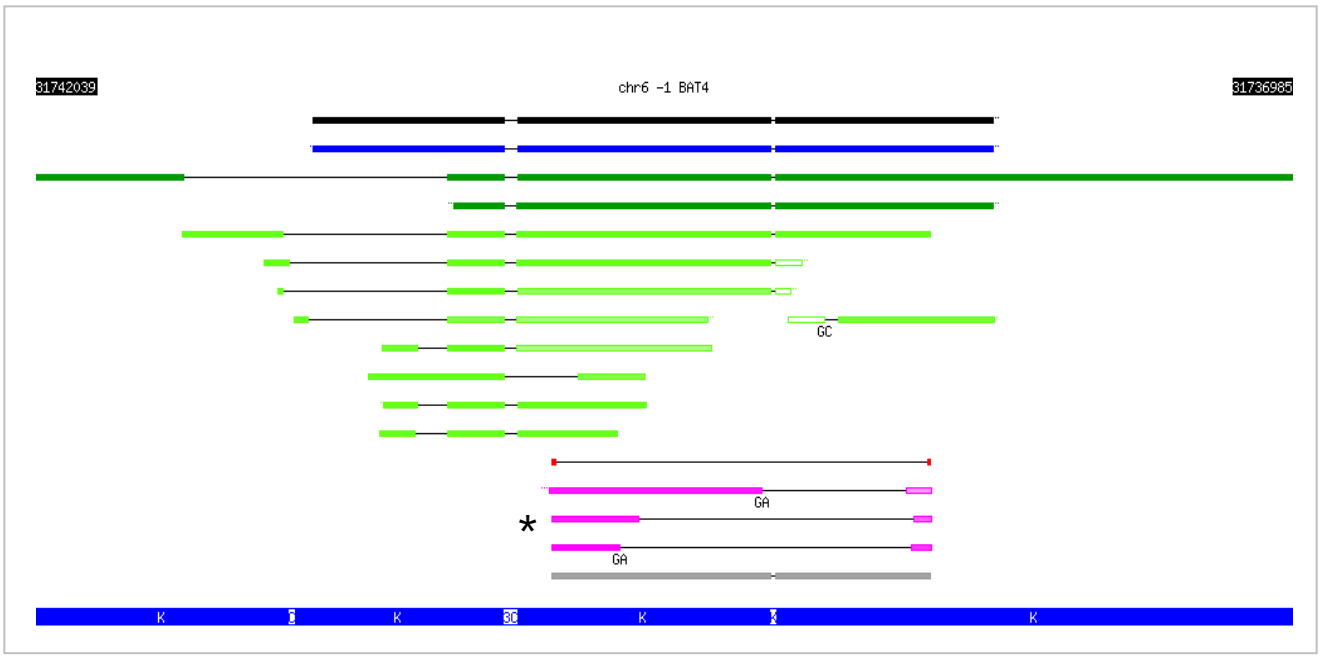
ARHGEF16 (ID: 27237)



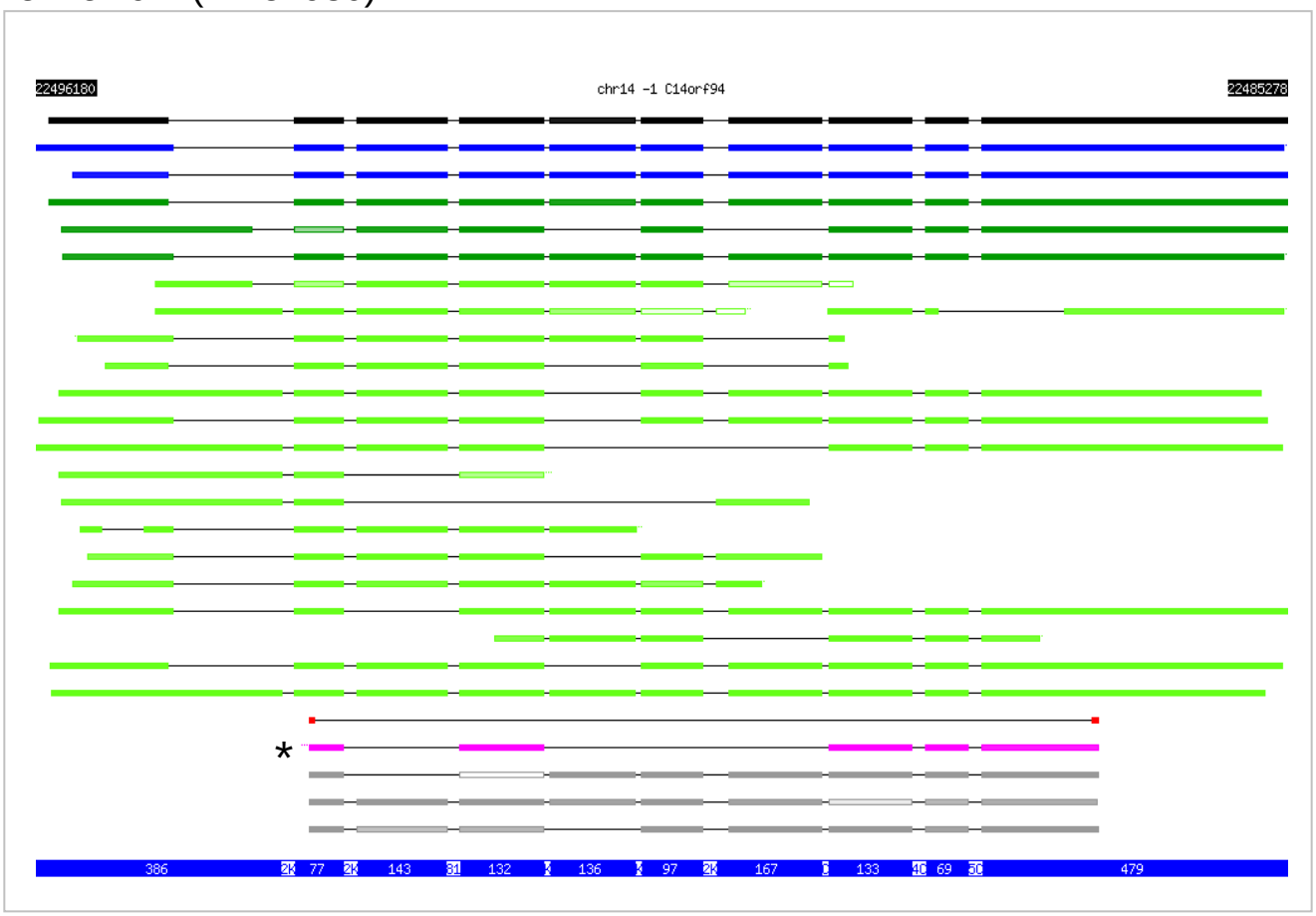
ARRB1 (ID: 408)



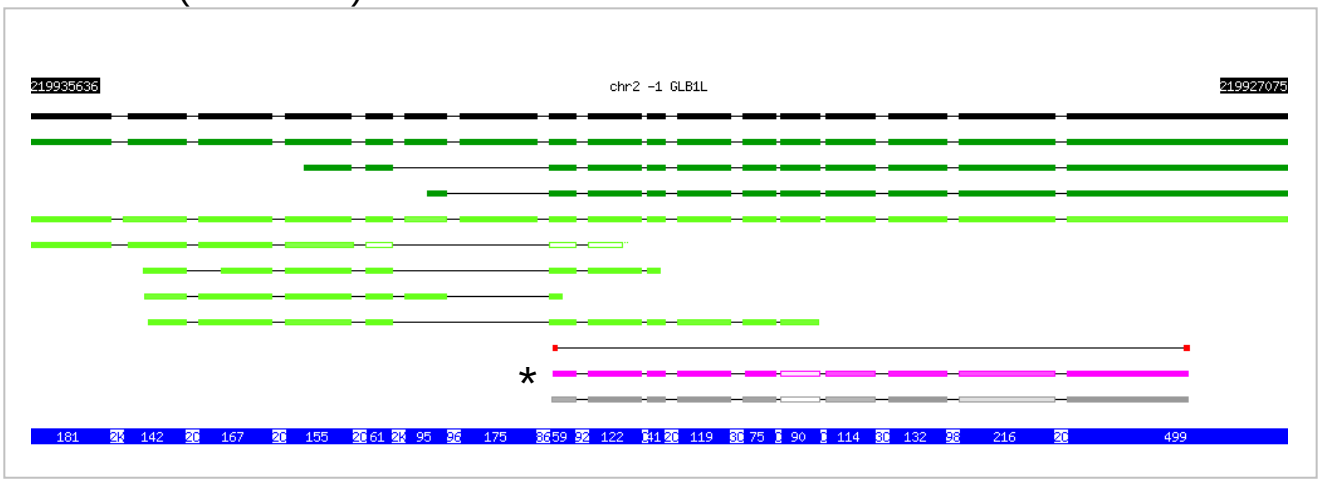
BAT4 (ID: 7918)



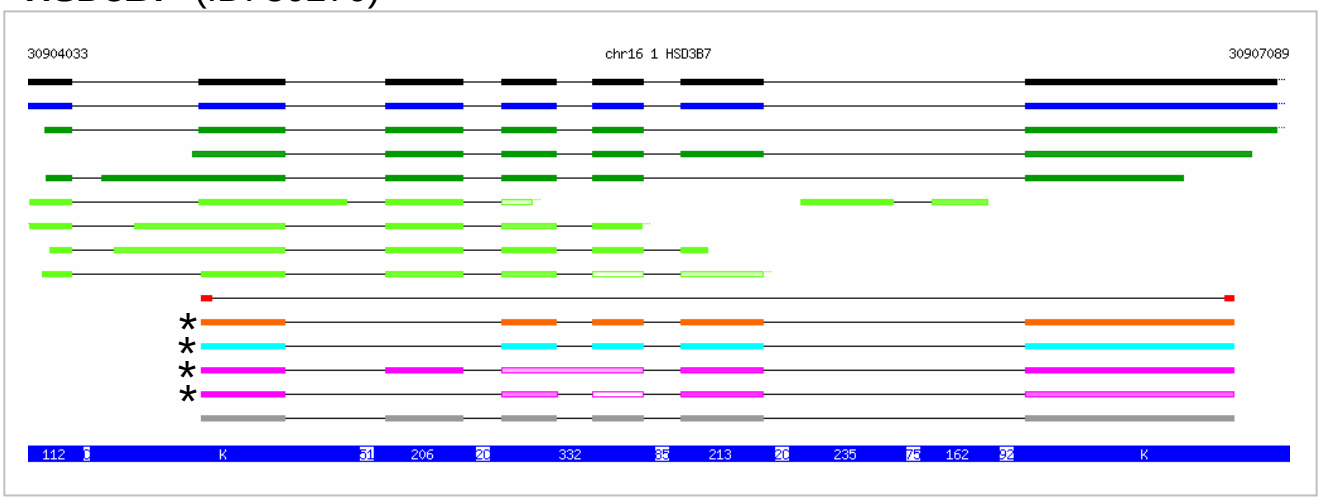
c14orf94 (ID: 54930)



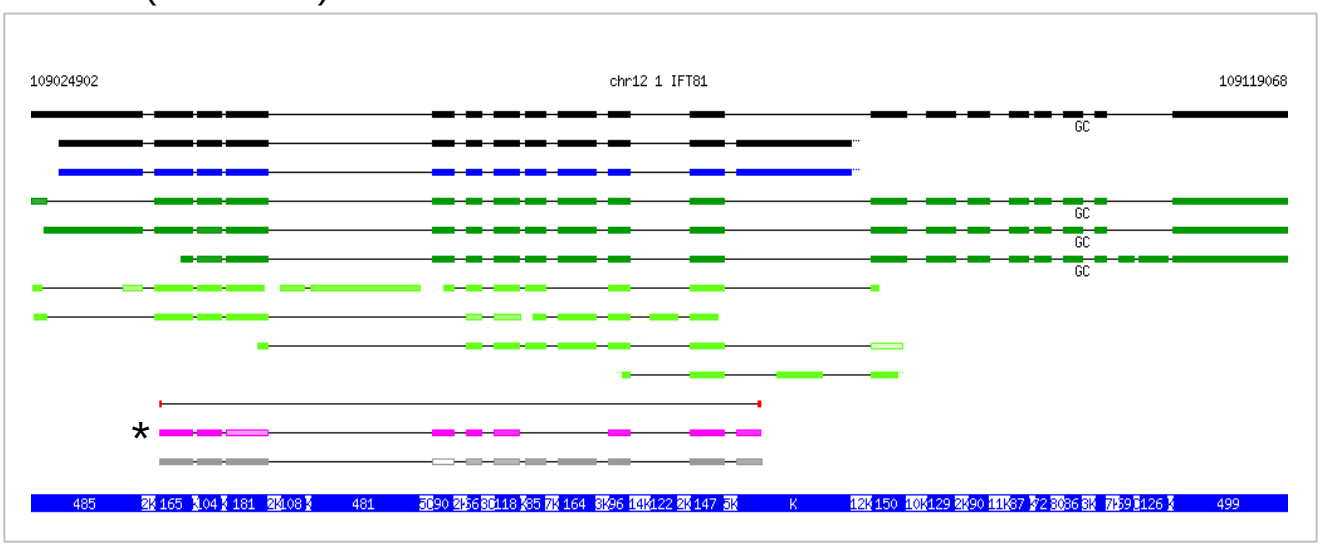
GLB1L (ID: 79411)



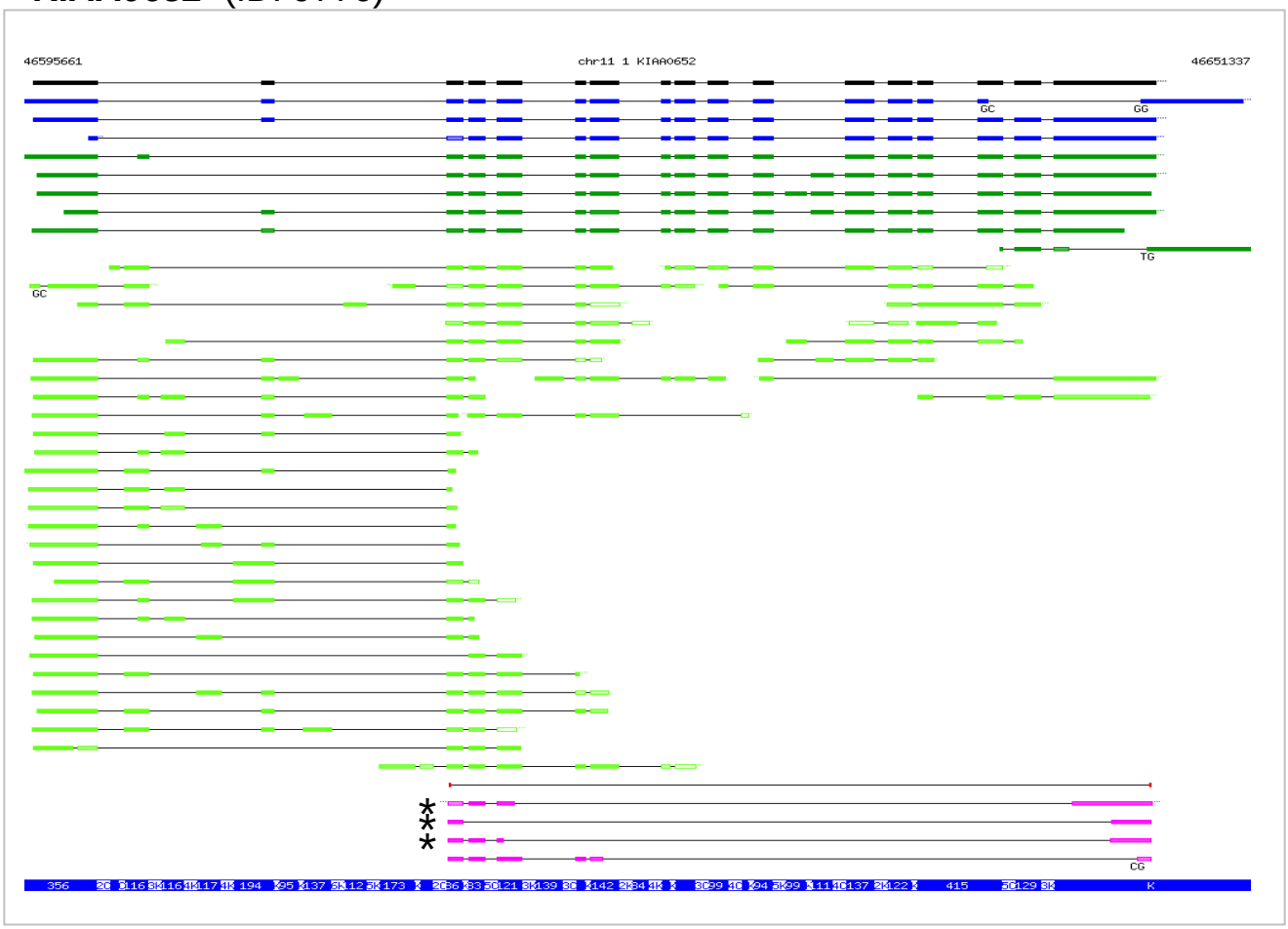
HSD3B7 (ID: 80270)



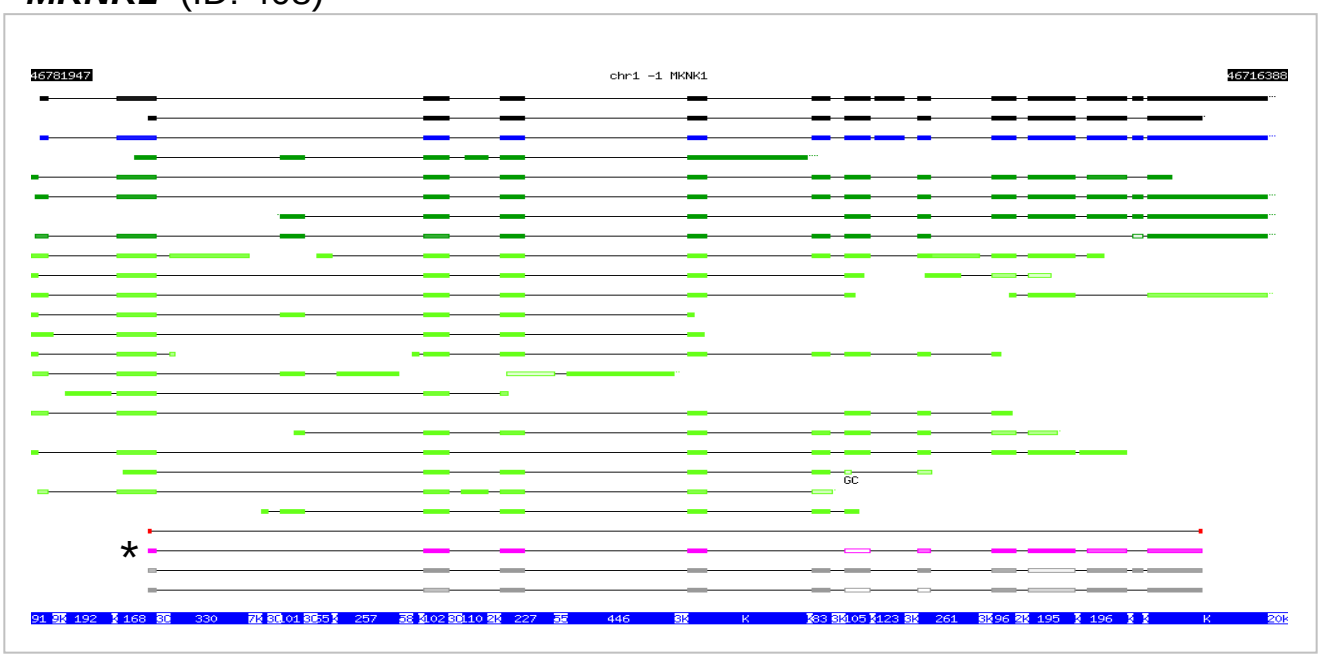
IFT81 (ID: 28981)



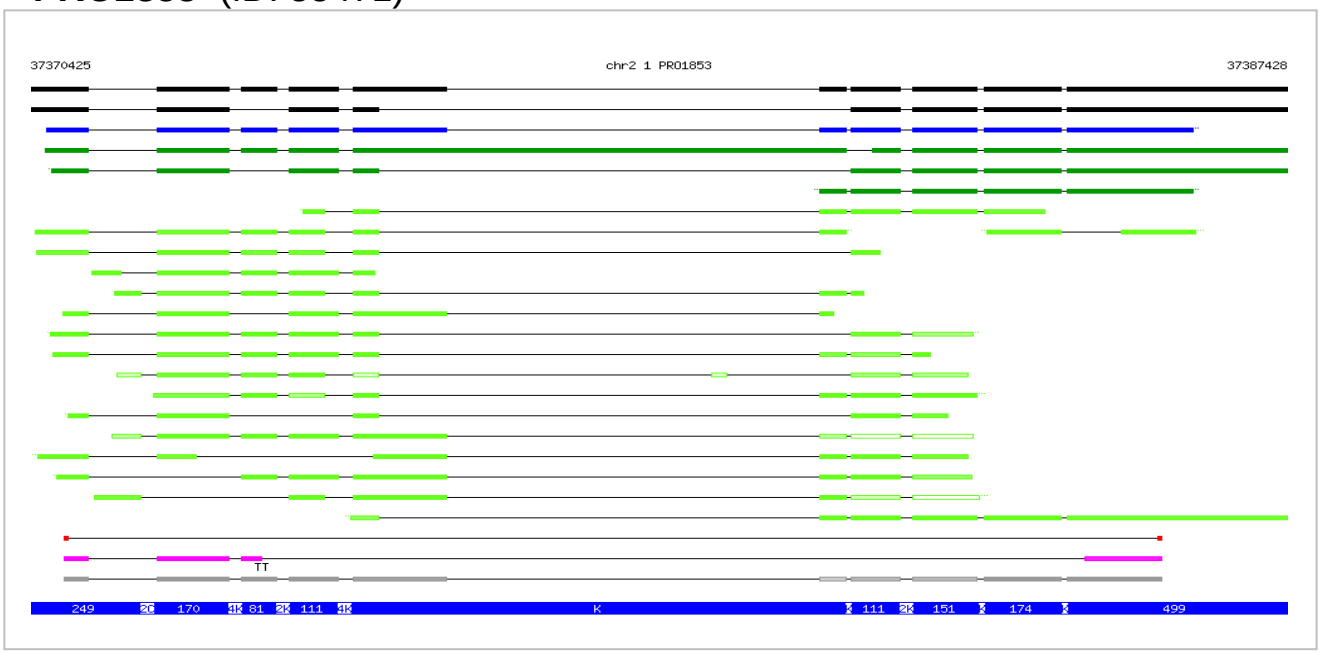
KIAA0652 (ID: 9776)



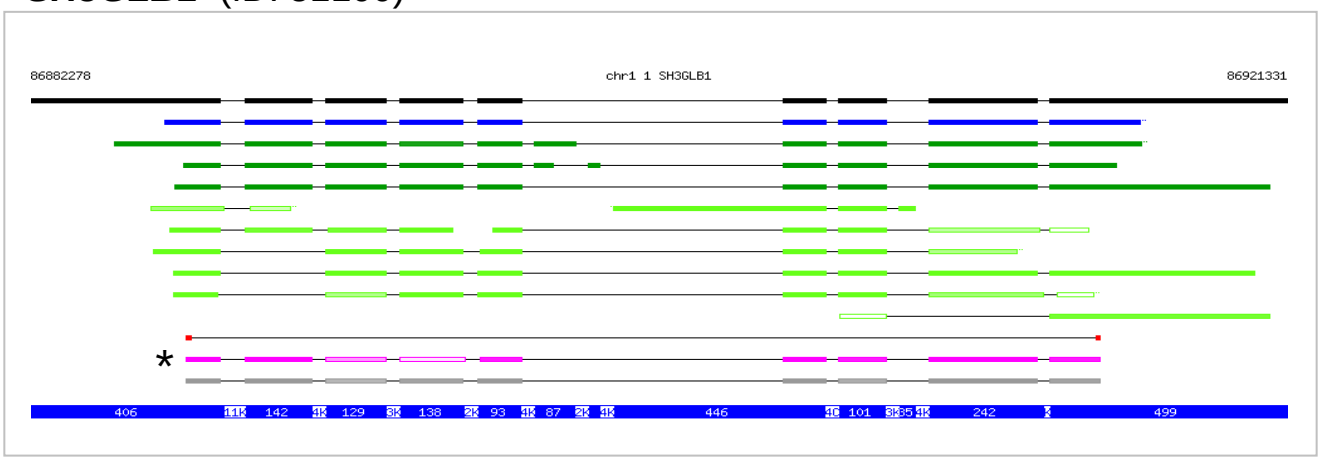
MKNK1 (ID: 408)



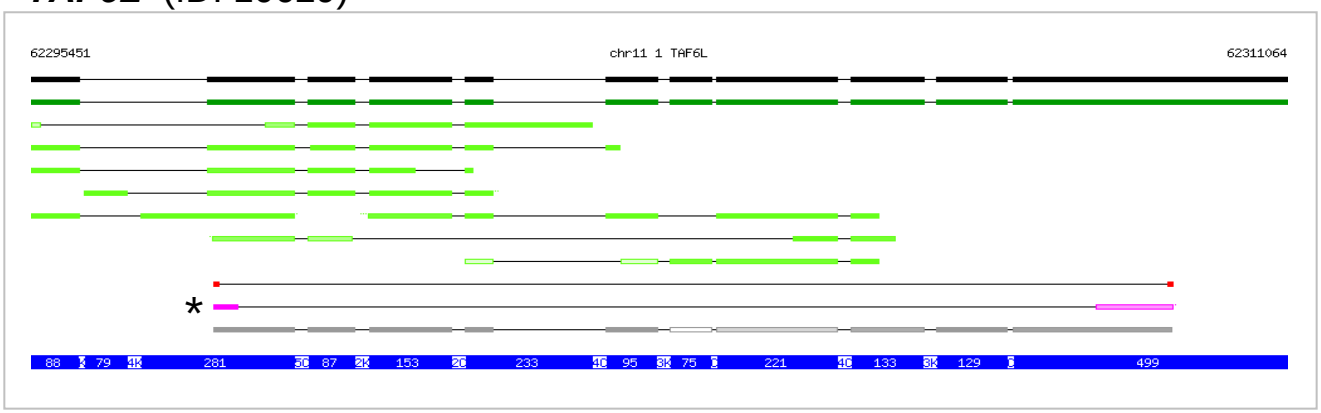
PRO1853 (ID: 55471)



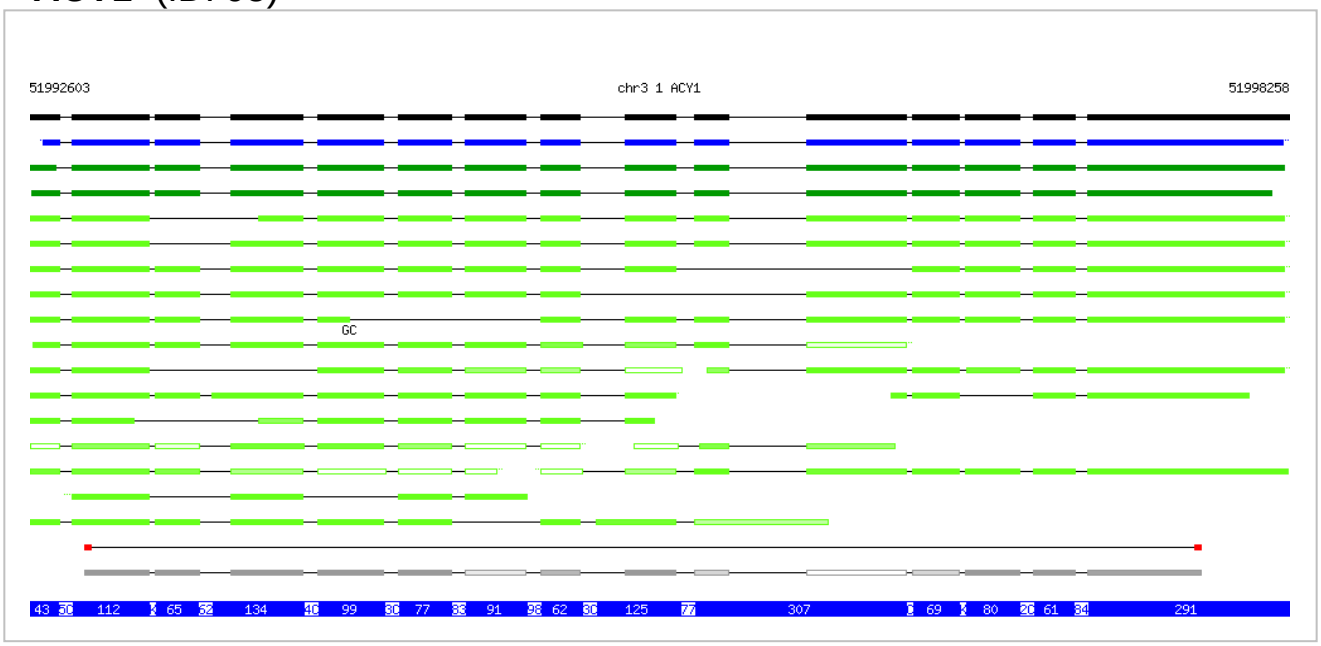
SH3GLB1 (ID: 51100)



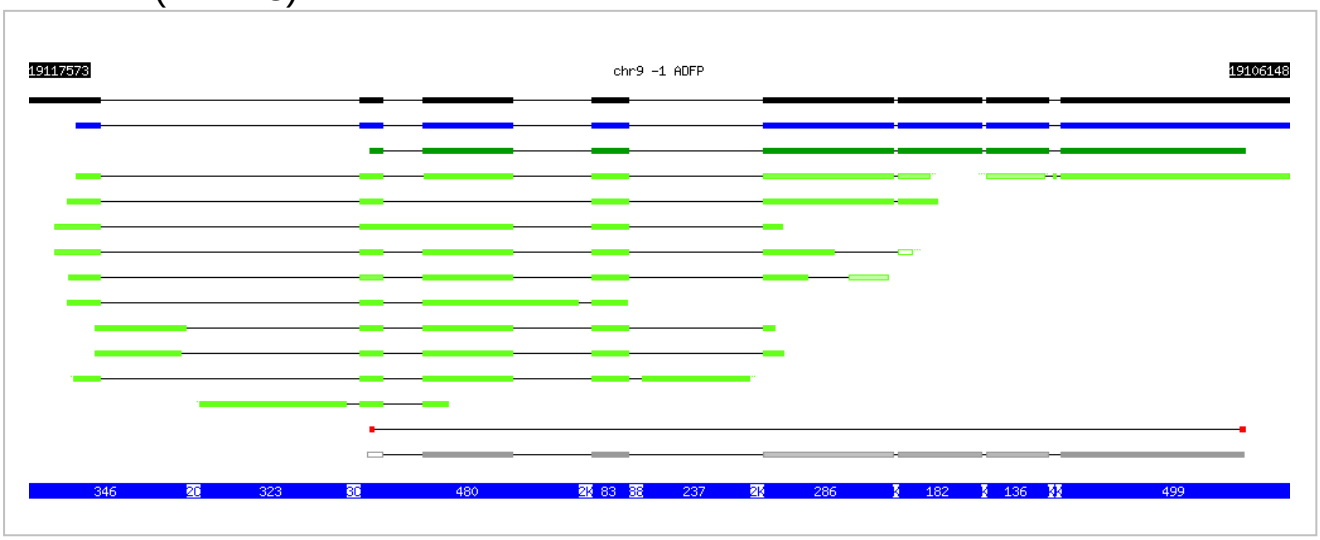
TAF6L (ID: 10629)



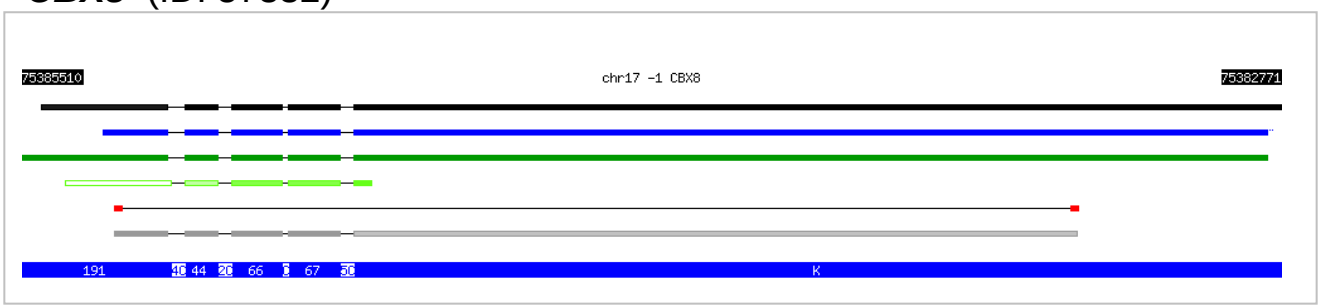
ACY1 (ID: 95)



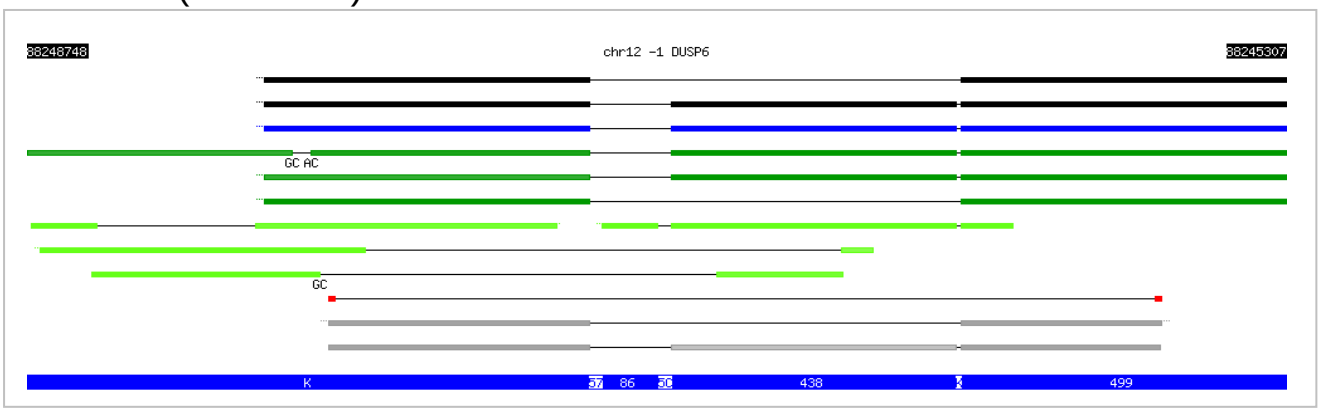
ADFP (ID: 123)



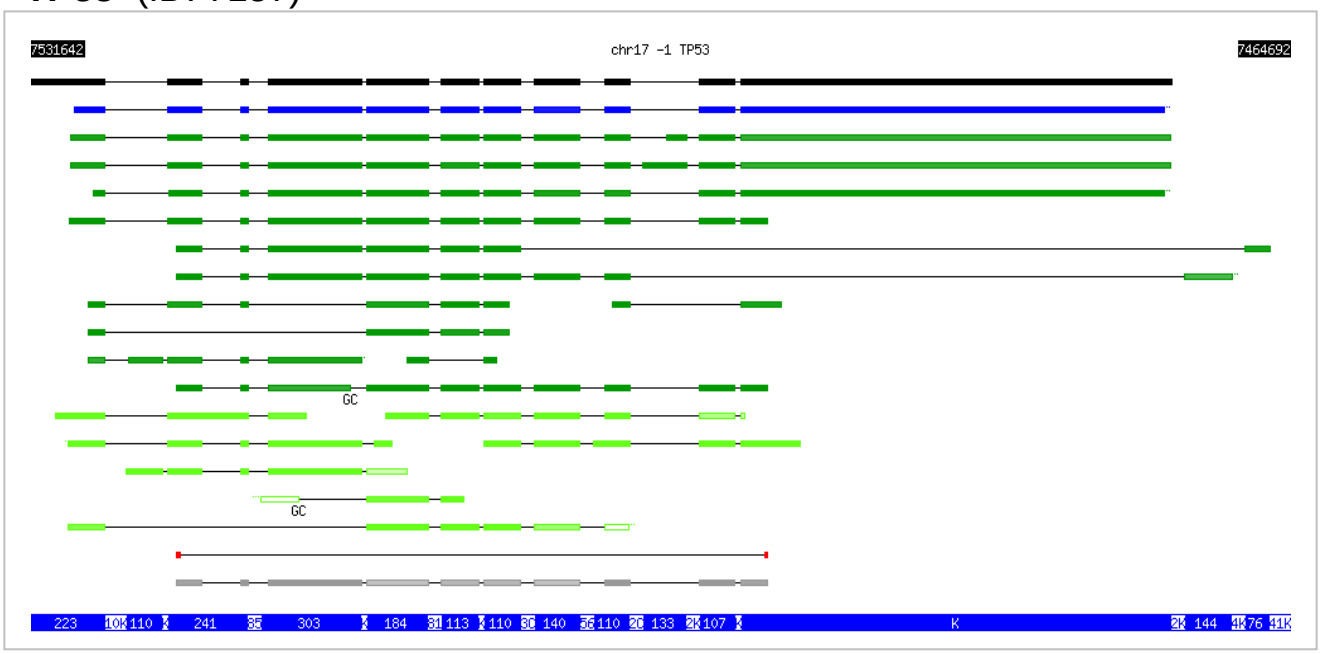
CBX8 (ID: 57332)



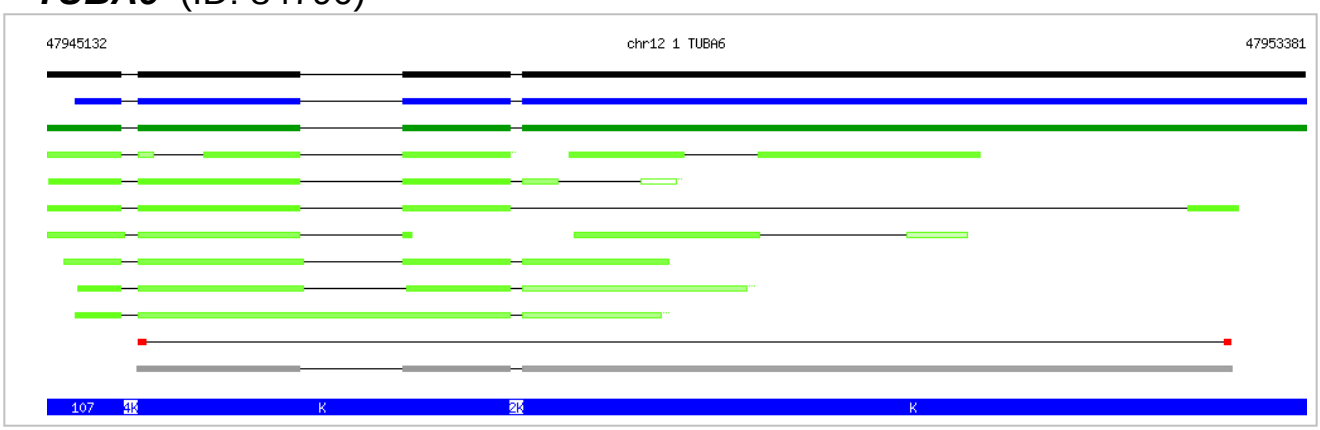
DUSP6 (ID: 11848)



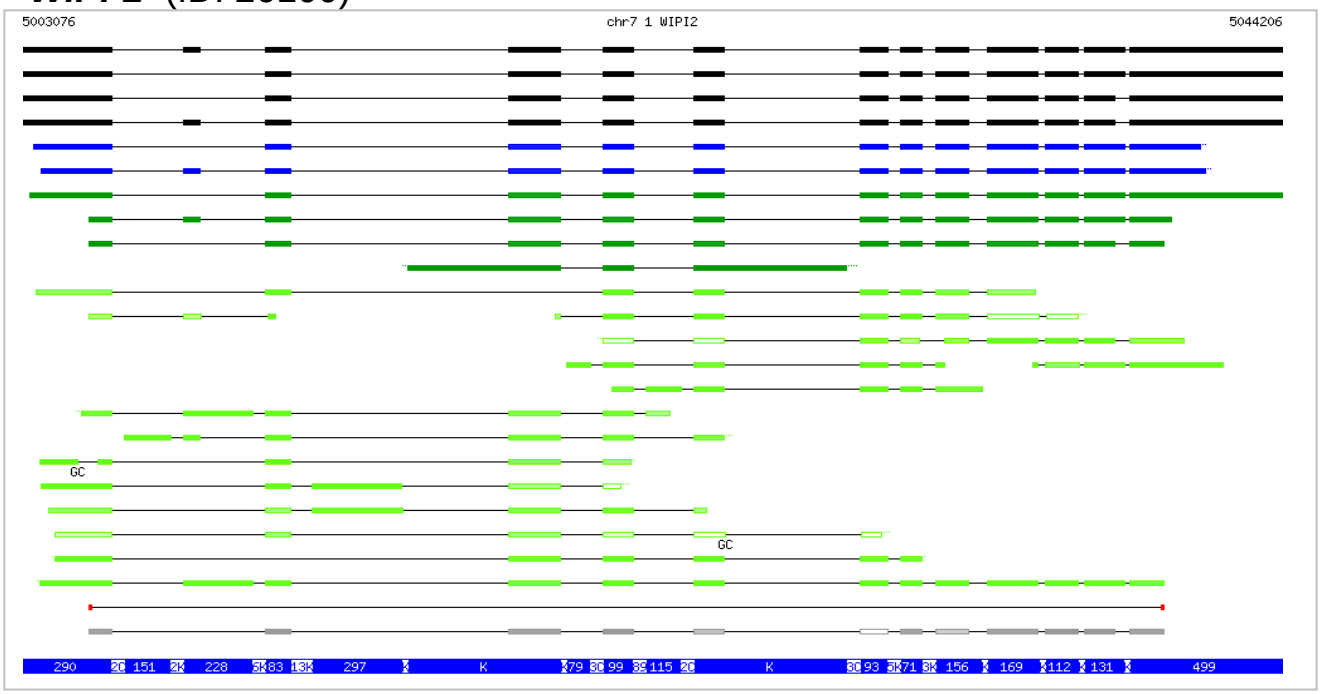
TP53 (ID: 7157)



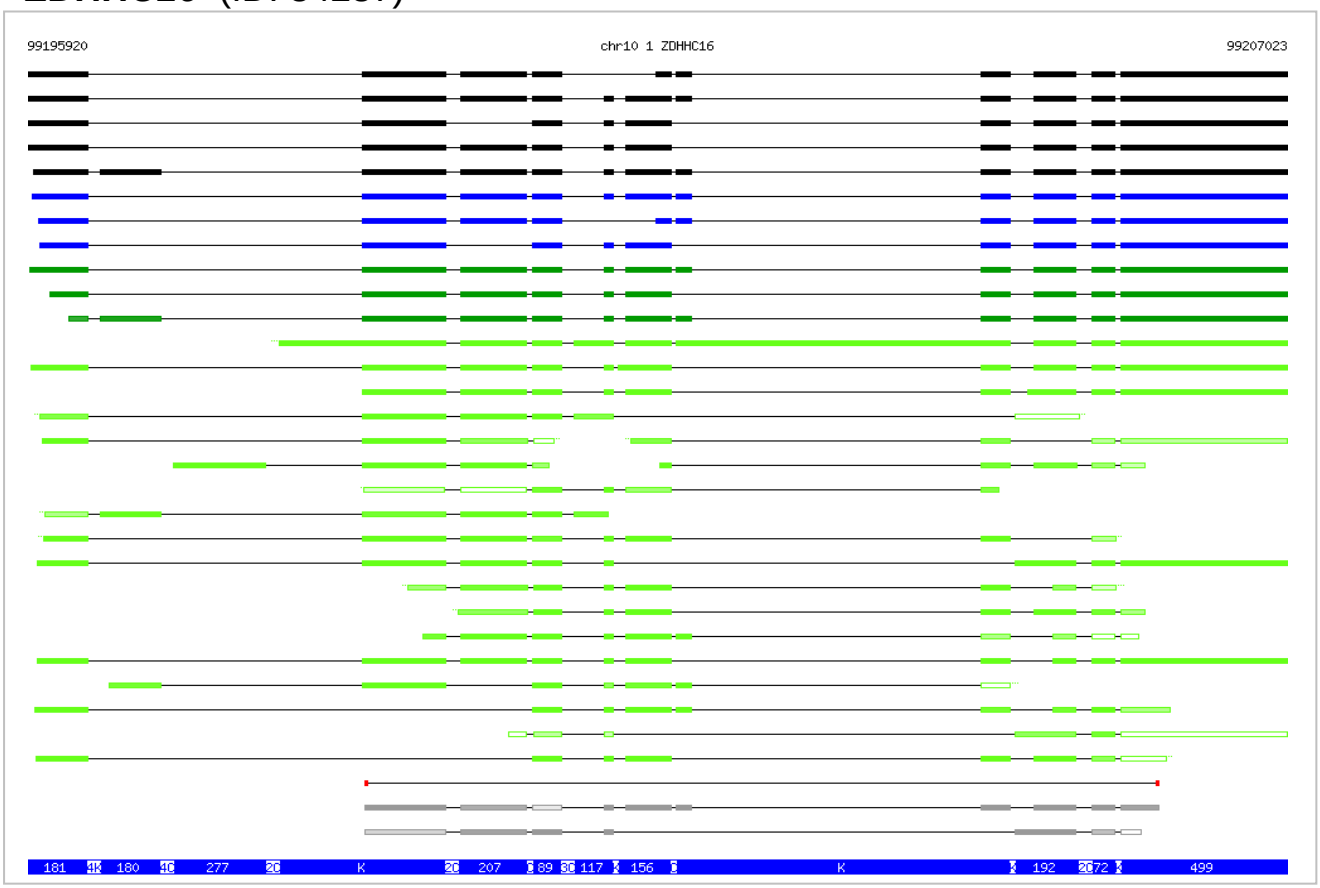
TUBA6 (ID: 84790)



WIPI-2 (ID: 26100)

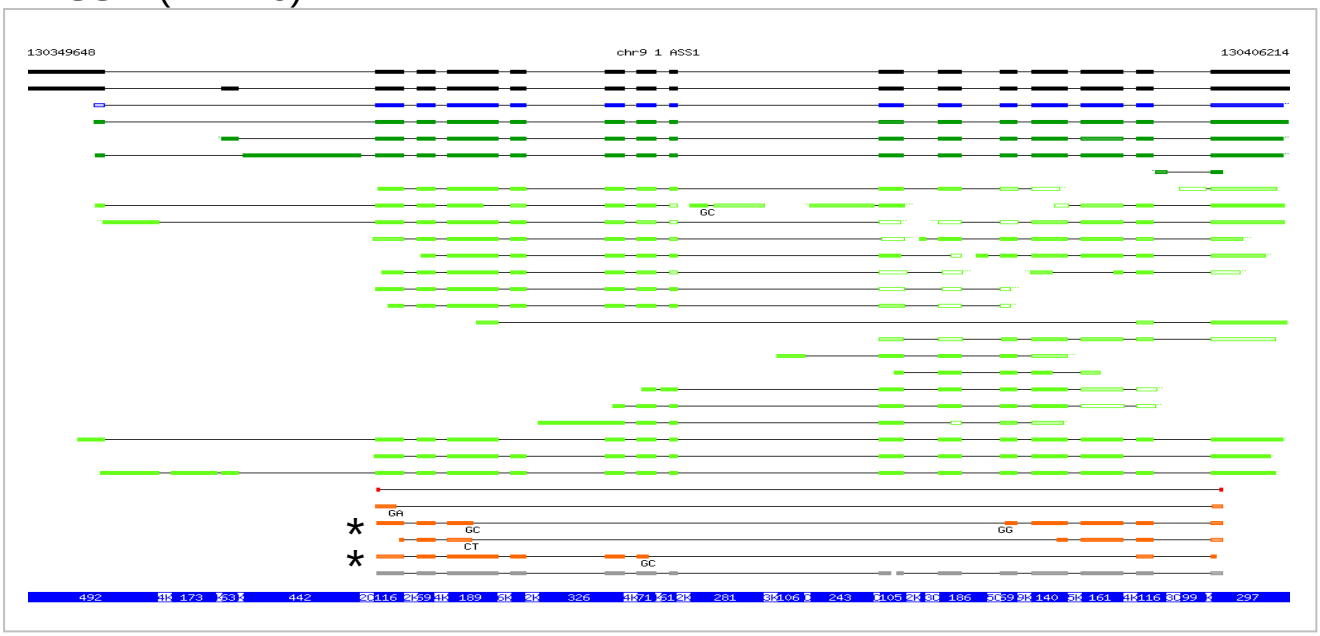


ZDHHC16 (ID: 84287)

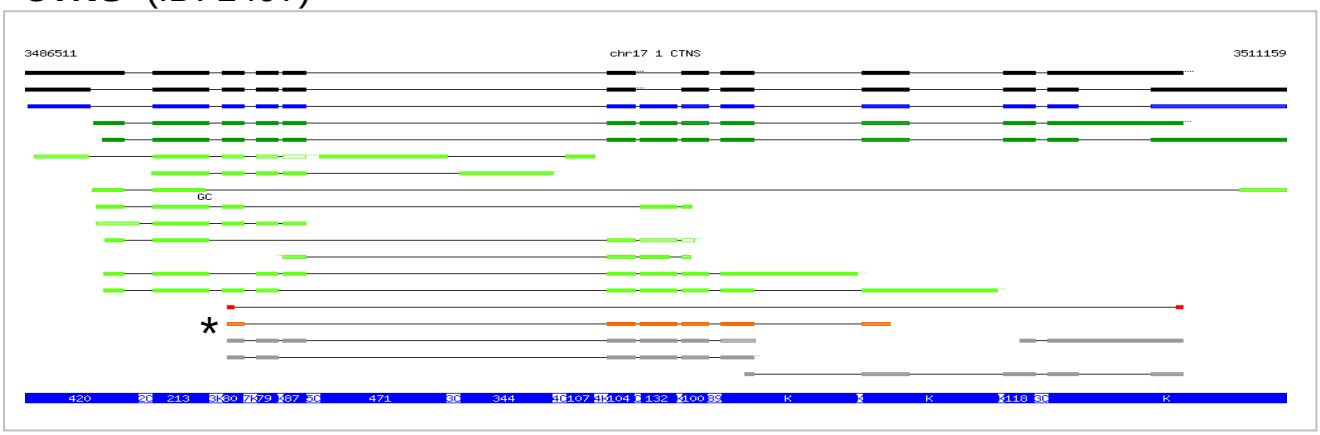


Supplemental figure 2b

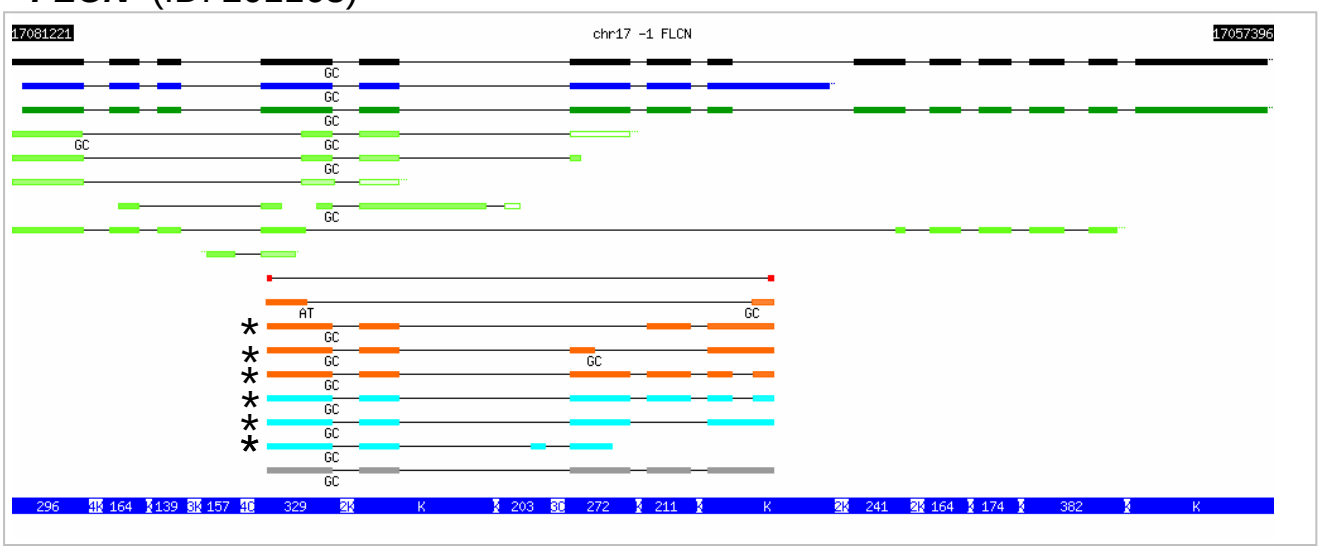
ASS1 (ID: 445)



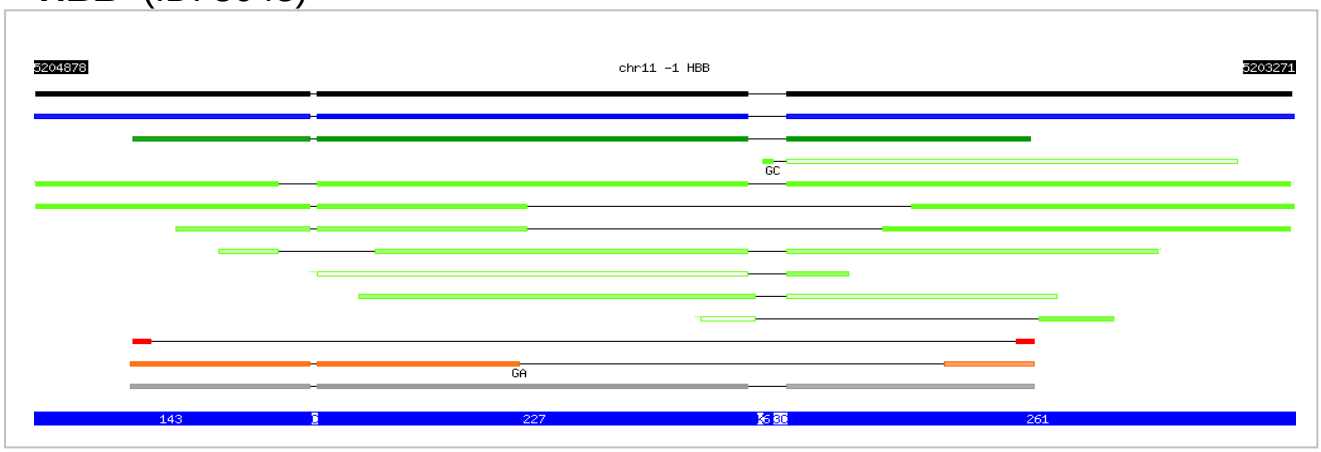
CTNS (ID: 1497)



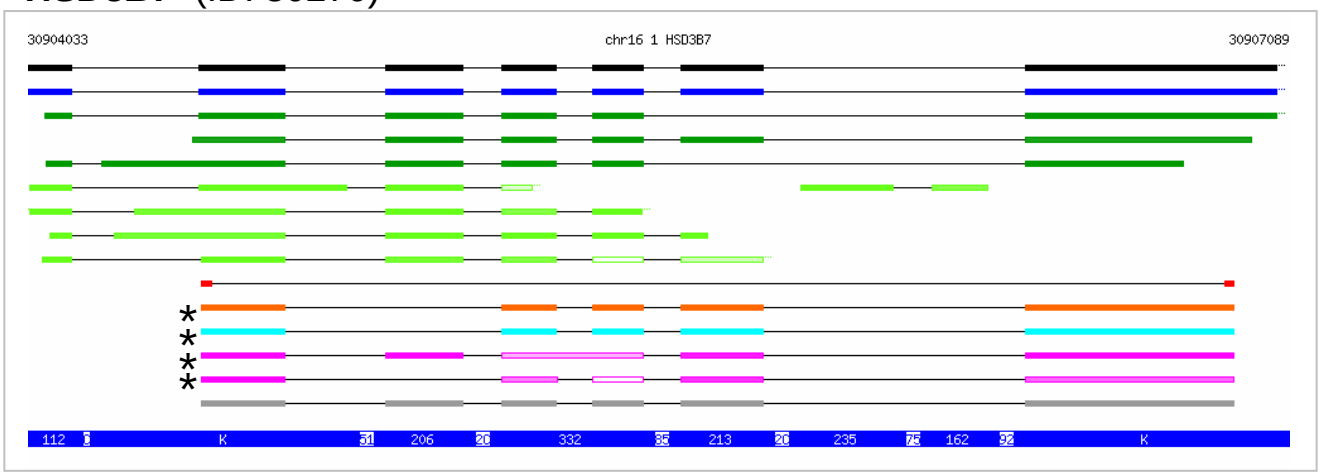
FLCN (ID: 201163)



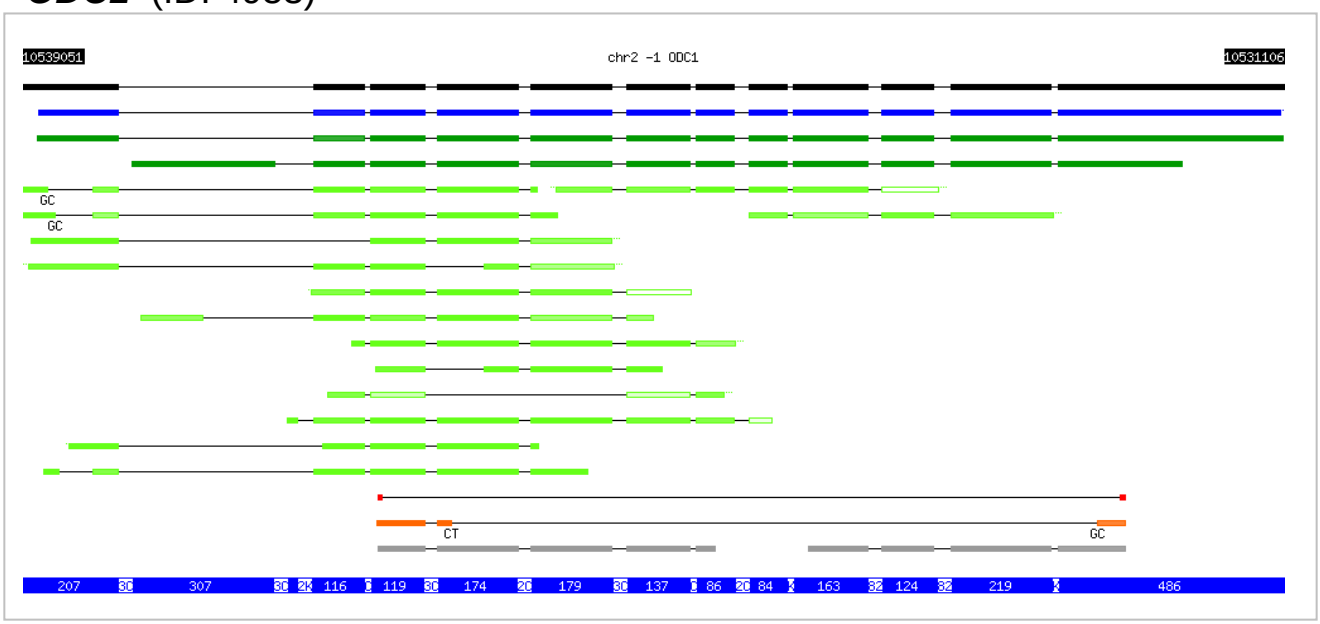
HBB (ID: 3043)



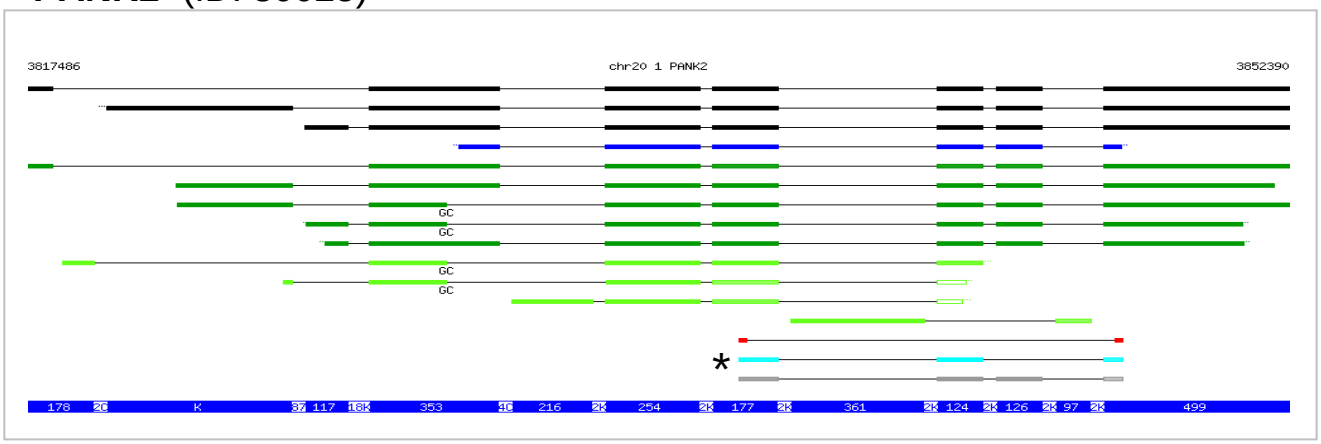
HSD3B7 (ID: 80270)



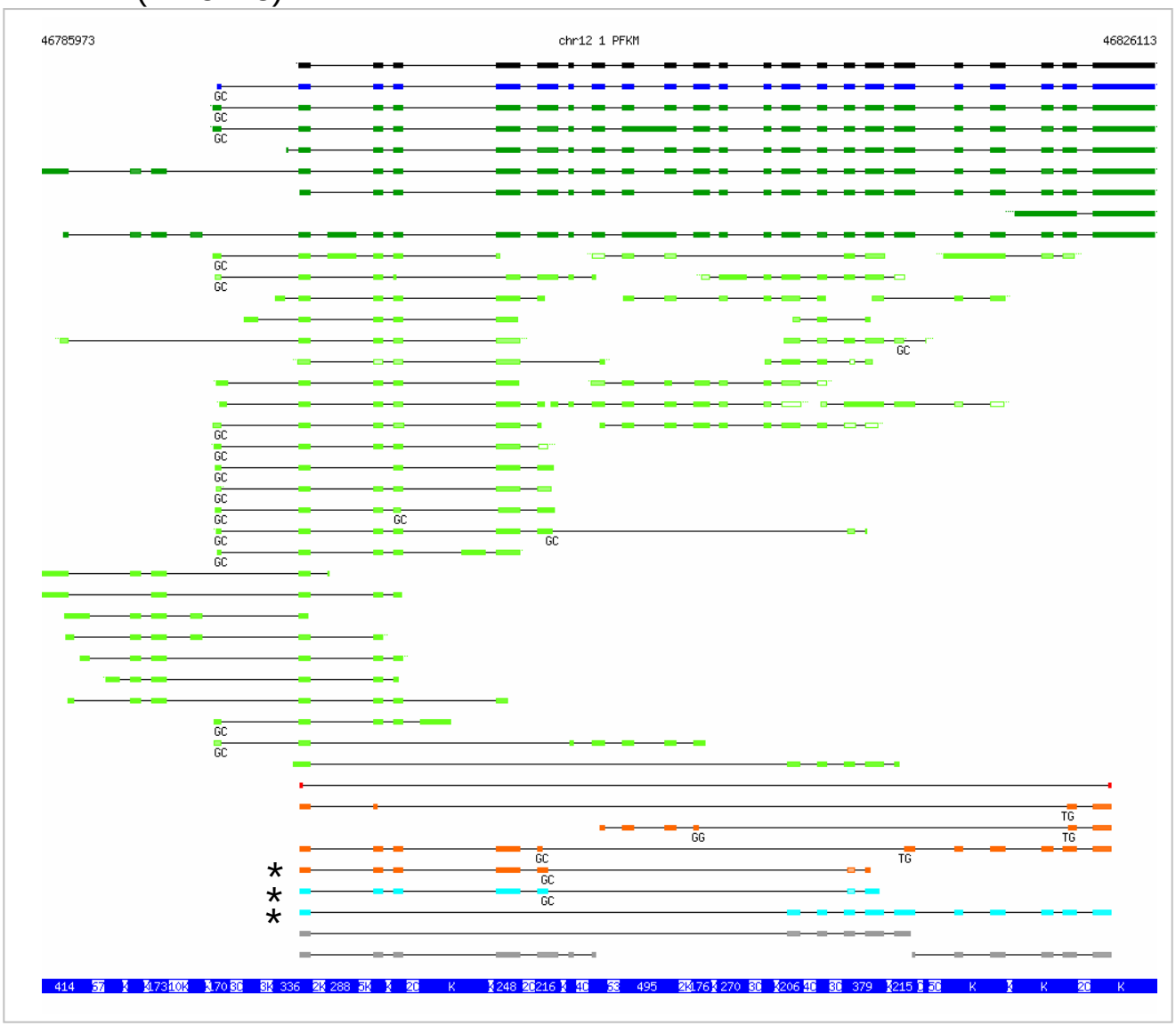
ODC1 (ID: 4953)



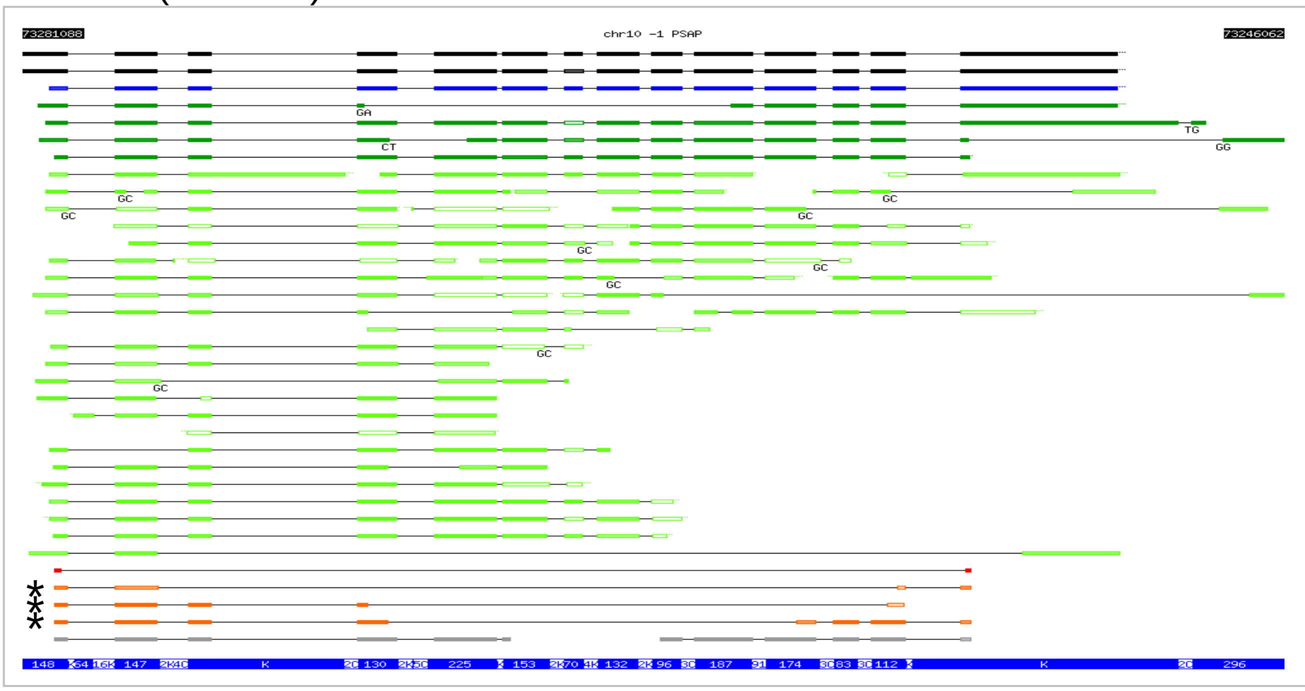
PANK2 (ID: 80025)



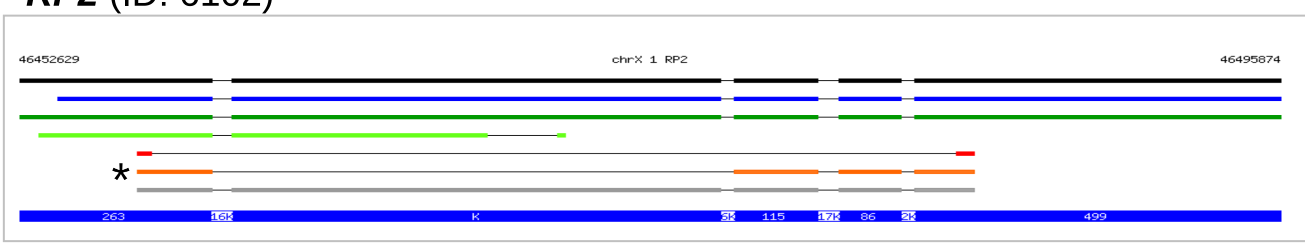
PFKM (ID: 5213)



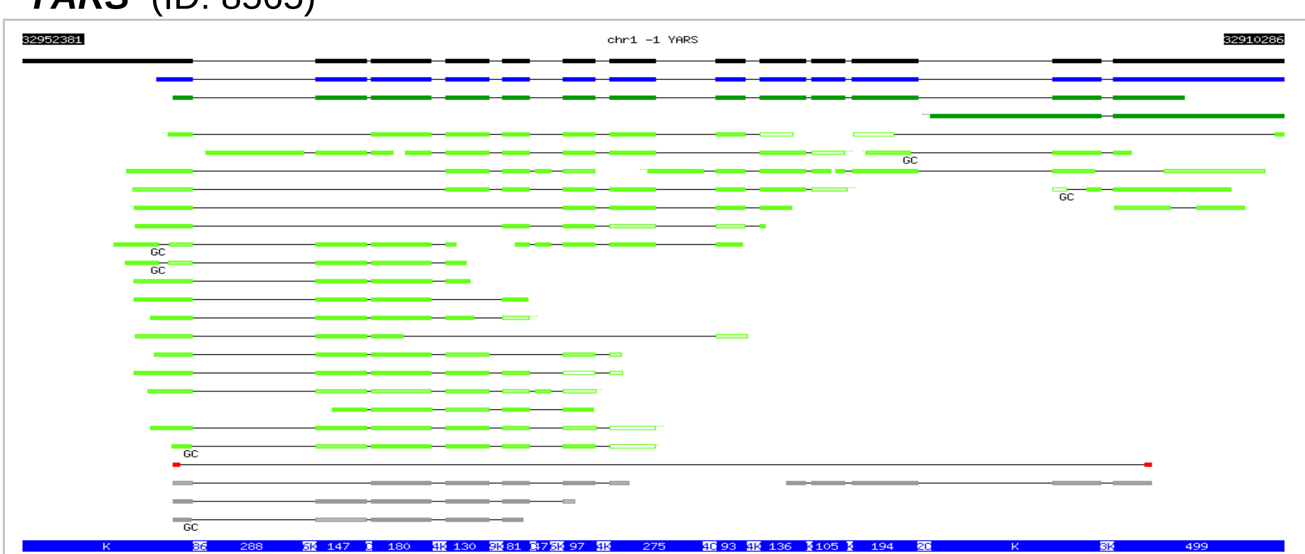
PSAP (ID: 5660)



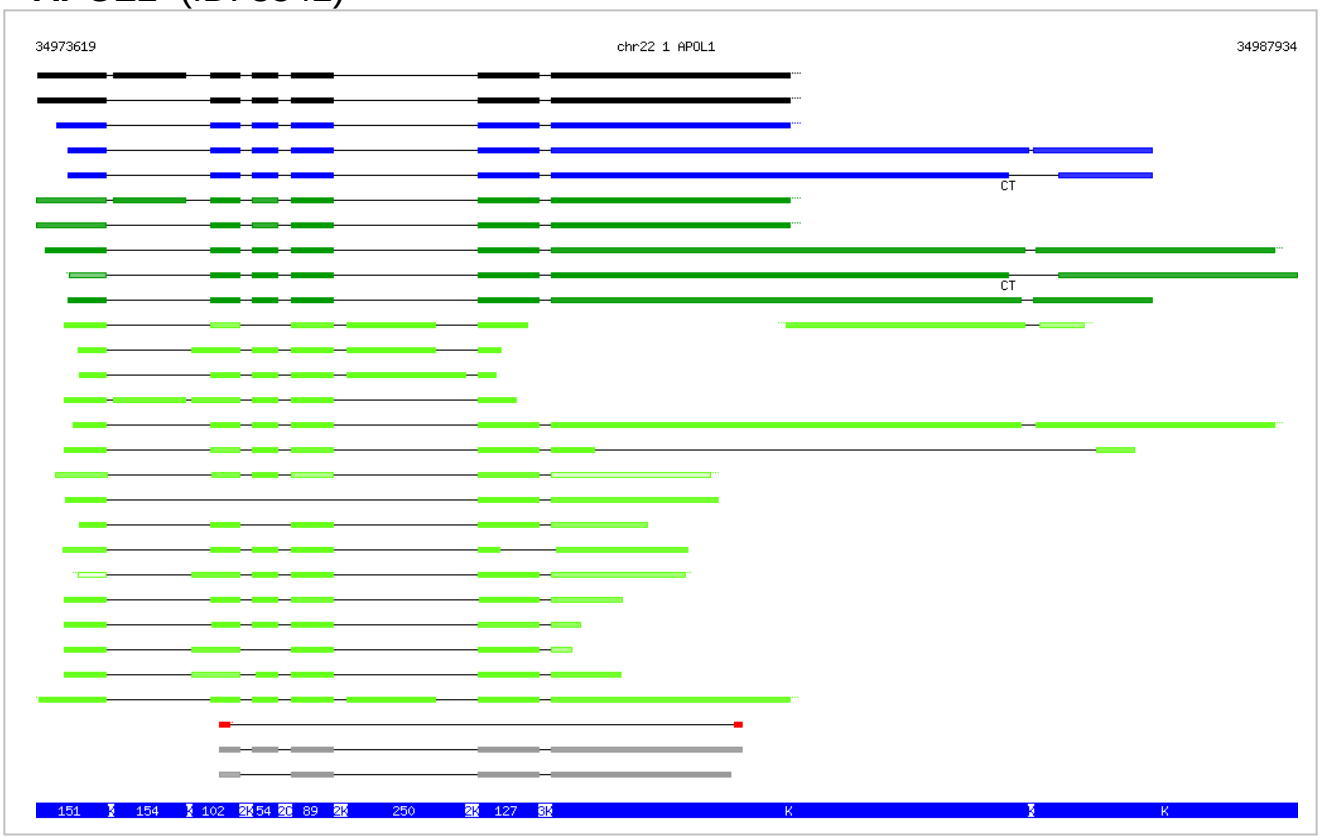
RP2 (ID: 6102)



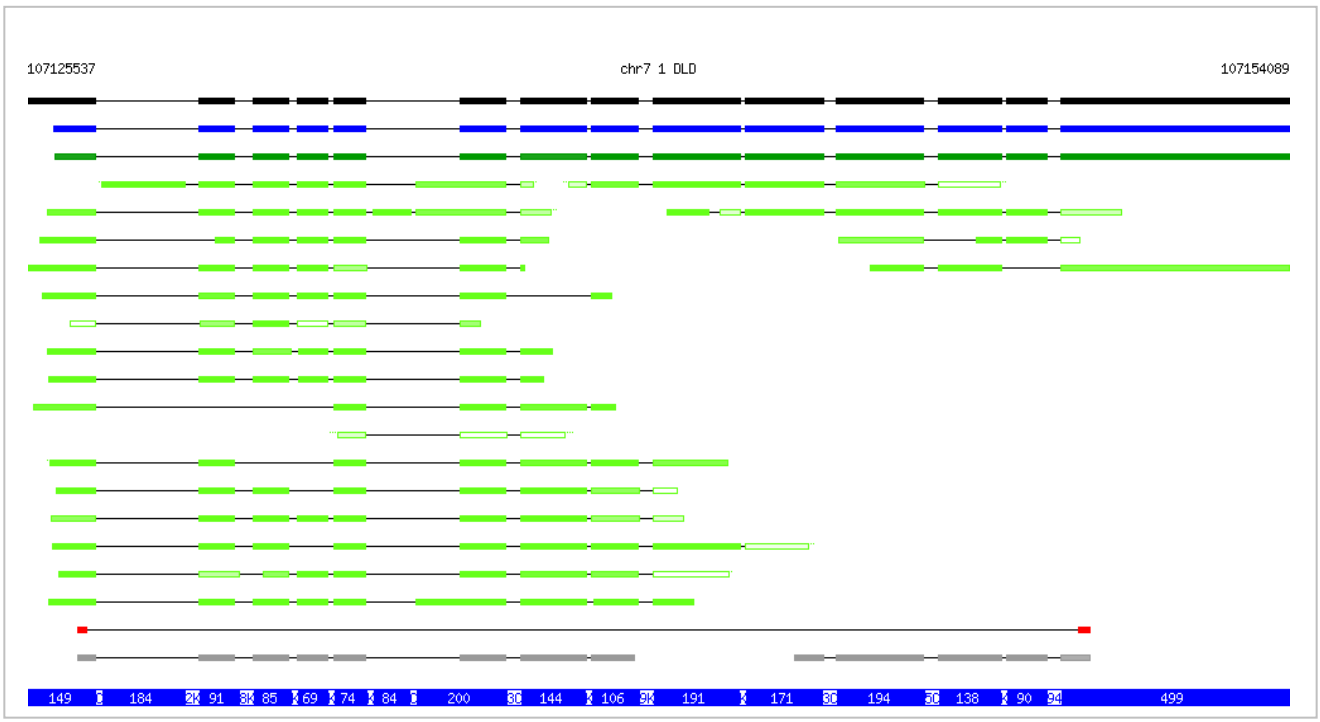
YARS (ID: 8565)



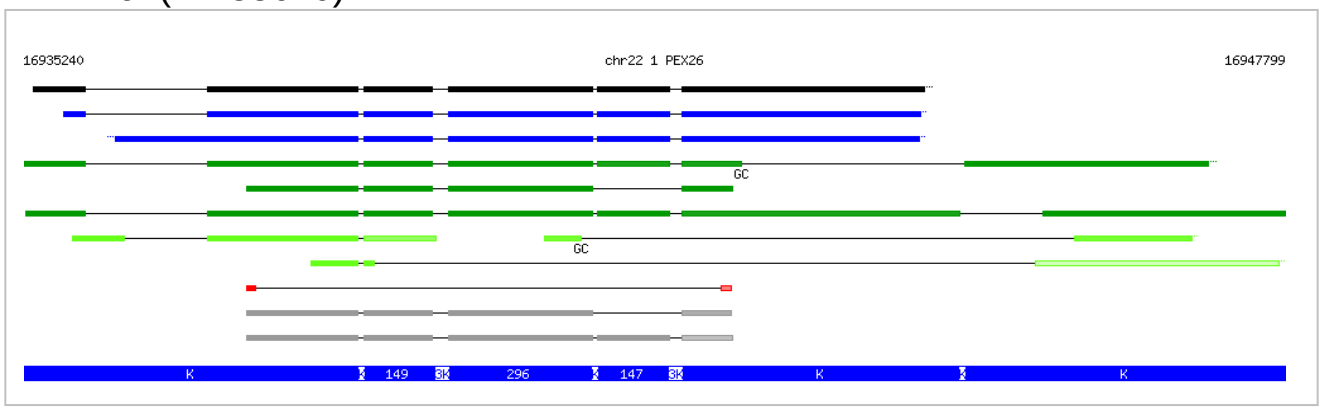
APOL1 (ID: 8542)



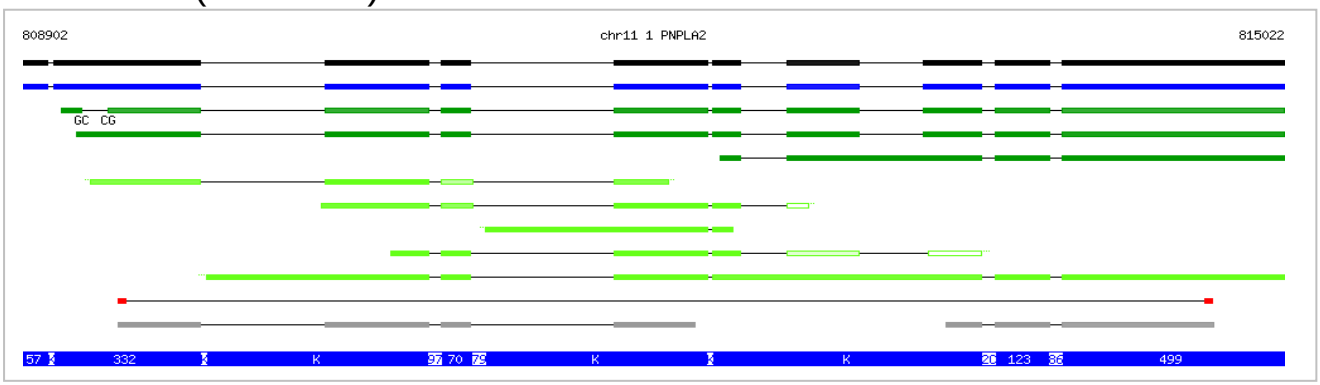
DLD (ID: 1738)



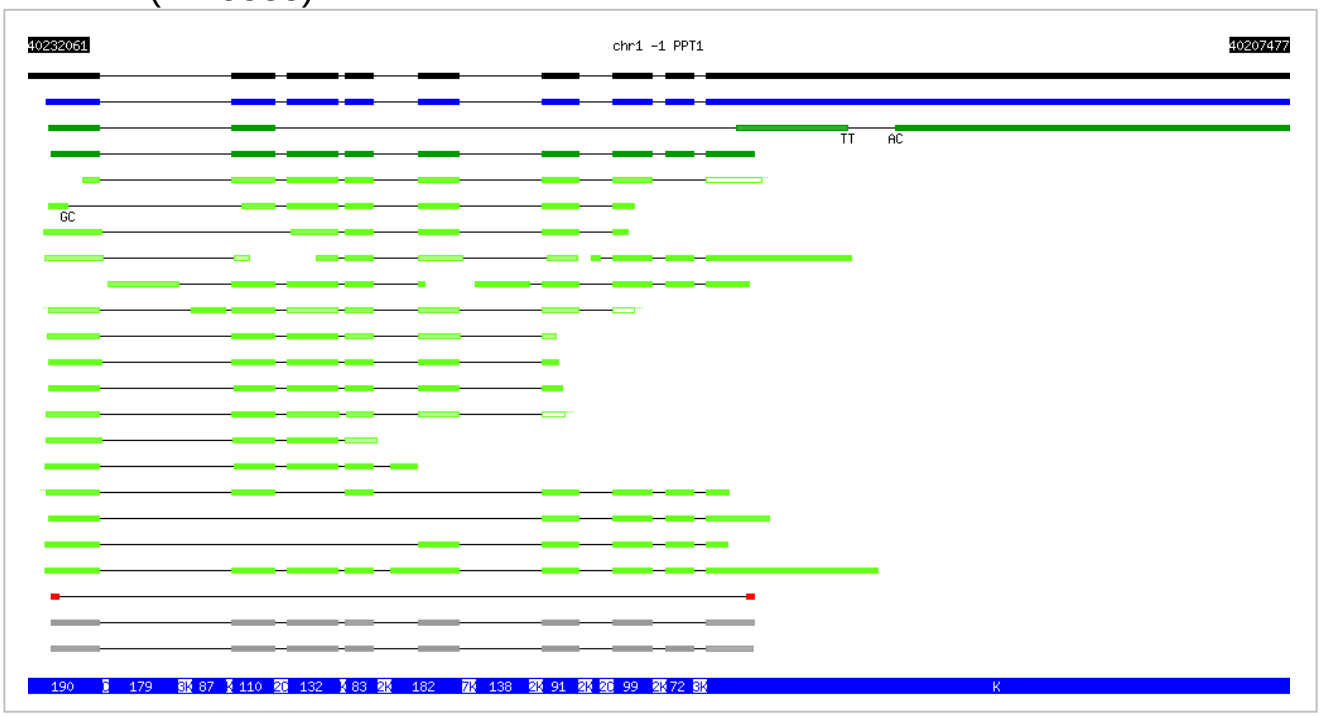
PEX26 (ID: 55670)



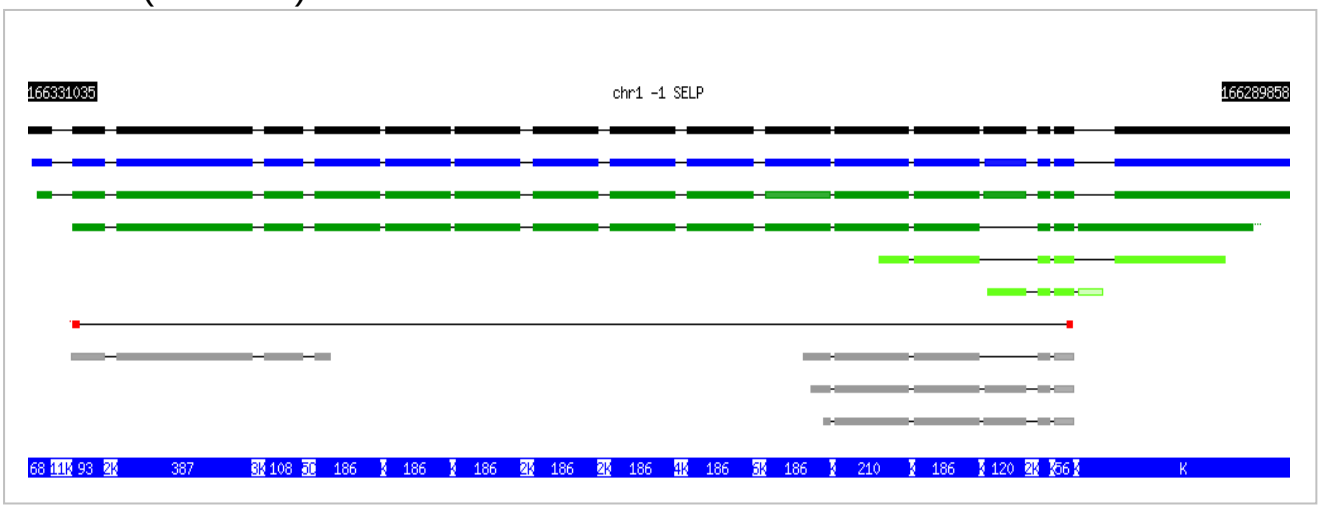
PNPLA2 (ID: 57104)



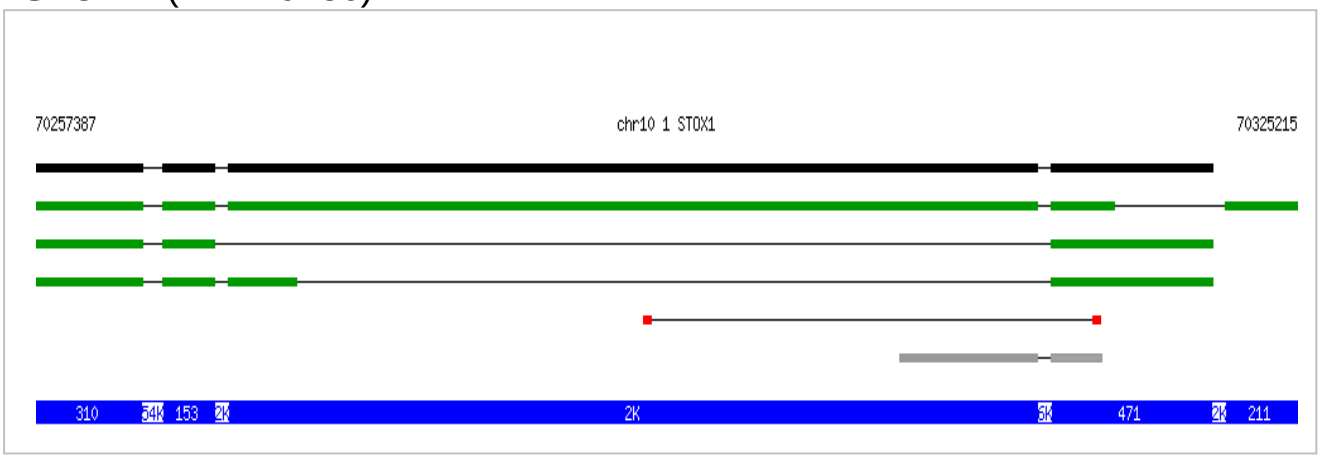
PPT1 (ID: 5538)



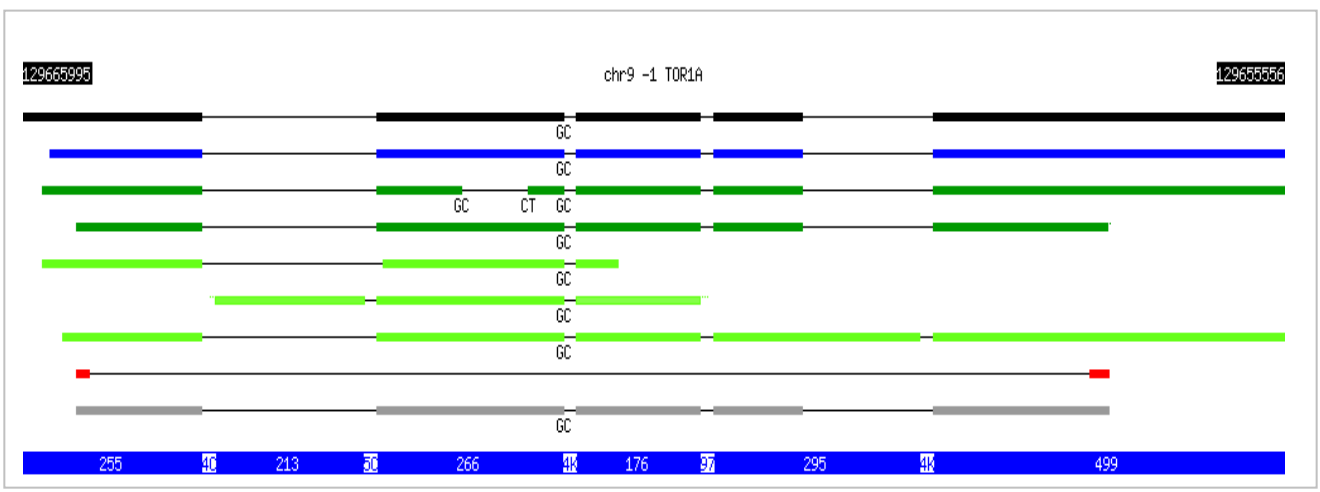
SELP (ID: 6403)



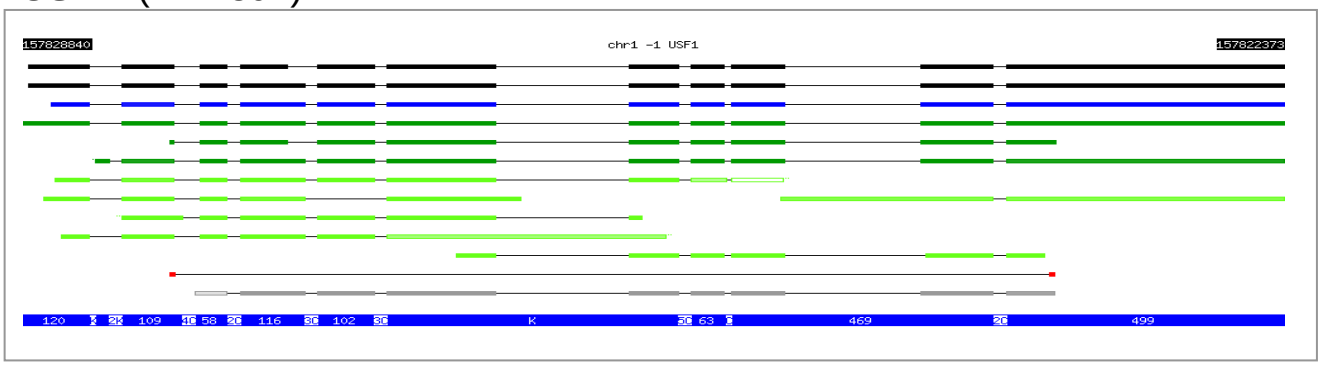
STOX1 (ID: 219736)



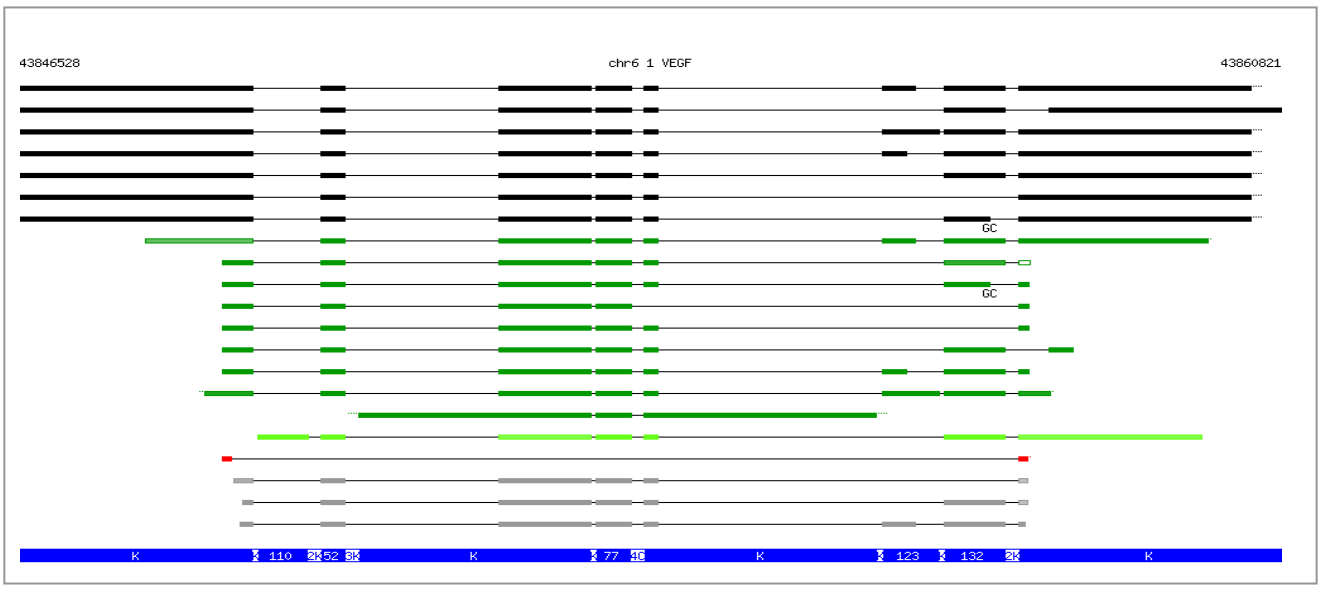
TOR1A (ID: 1861)



USF1 (ID: 7391)



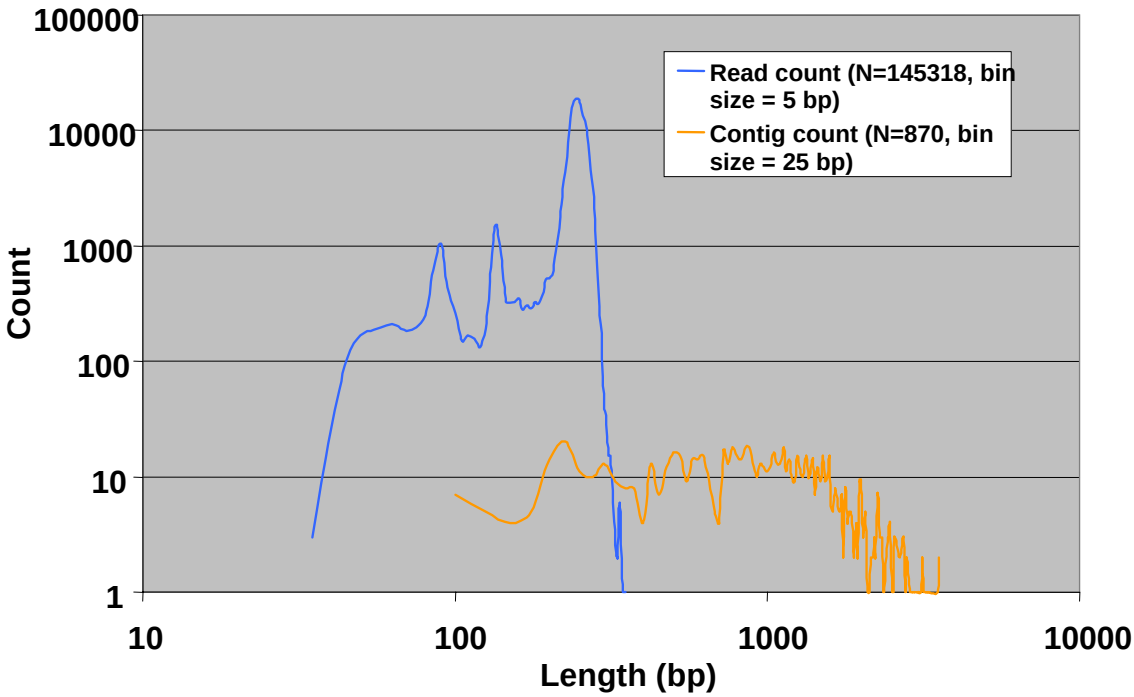
VEGF (ID: 7422)



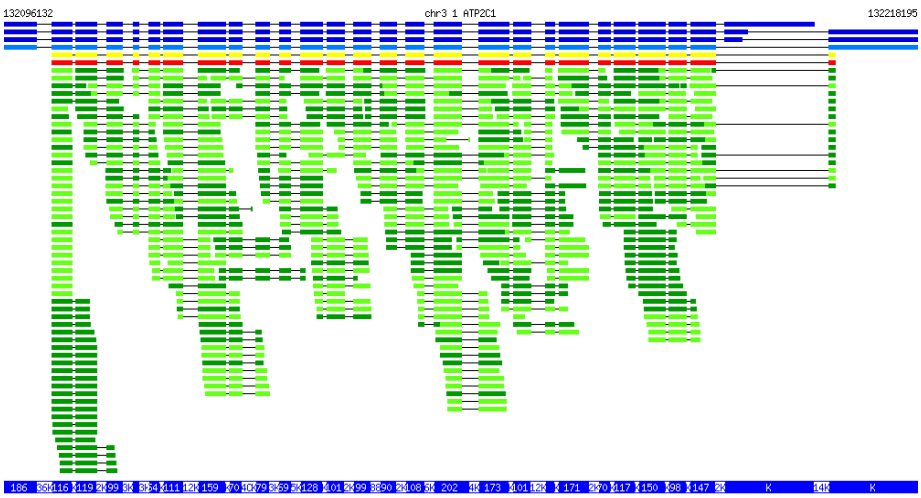
Alignments of sequences obtained from cloning of RT-PCR products. RT-PCR products from a pool of five tissues (Set 1) or from brain or testis PCR products (Sets 2 and 3) were cloned. From each set 12 colonies corresponding to each gene were picked and sequenced by conventional Sanger capillary sequencing. (**a**) Alignments for Set 1. (**b**) Alignments for Sets 2 and 3. Obtained sequences were aligned to the genome and compared with RefSeq (black), MGC (blue), GenBank (green) and dbEST (light green). Redundant alignments and those not spanning introns were removed. Results are shown for all the genes from which ORFs were cloned. Transcripts with exon/intron structures that are exactly recapitulated by previously known MGC, Refseq, GenBank or dbEST transcripts are shown in gray, while novel transcript variants are shown in purple for the pooled cloning experiment, orange for brain, and cyan for testis. The positions of primers used for the RT-PCR are shown in red. Color saturation indicates % identity, ranging from light (≤ 90% identity) to dark (≥ 99% identity). Non-canonical splice donor/acceptor dinucleotides are indicated. Asterisks (*) indicate novel variants with canonical (GT…AG) or GC…AG splice signals; green arrows in the alignment figures of *ACAA2* and *ARRB1* point to subtle alternative splicing in canonical exons; and a red arrow in the alignment figure of *ARRB1* points to an aberrant alignment due to poor sequence. In instances where full length sequences were not available, novelty was determined based on the available portion. In the interest of clarity, introns with signals other than GY-AG in ESTs were split, and remaining fragments not spanning introns were discarded. However, these introns were used in the assessment of novelty. Introns are compressed in order to highlight exon structures. Chromosomal coordinates are indicated at the top of each panel. Lengths of exonic (white on blue) and intronic (reversed) segments are shown in bp at the bottom of each panel (C = 100; K = 1000). With a few exceptions, untranslated regions longer than 500 bp are clipped.

Supplementary figure 3

a

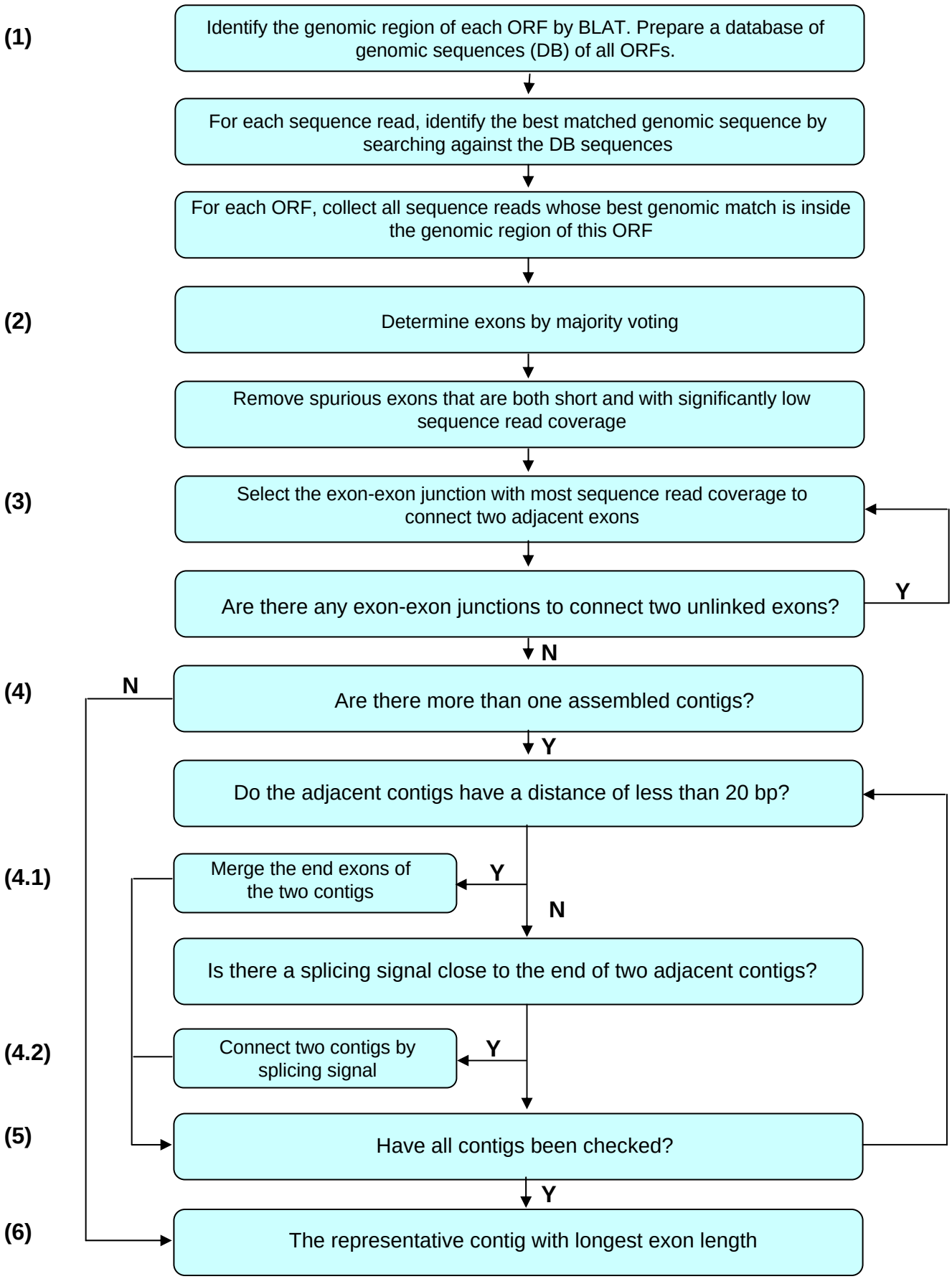


b



Size distribution of the obtained 454 reads and an example of genomic alignment of the assembled contigs. In (**a**), size distribution of the reads (blue) and ORF contigs (orange) assembled using ORF reference sequences is shown. The main peak occurs at ~240 bases. The two smaller peaks likely correspond to shearing “hotspots” in the flanking sequences. In (**b**) genomic alignment of reads and corresponding mRNAs for one ORF (*ATP2C1*) is shown. Boxes represent exons, with introns (thin lines between exons) compressed to emphasize exon structure. Refseq/MGC full-length mRNAs are shown in blue, with light blue indicating mRNA structures identical to the ORFs used in the experiment (UTRs are not distinguished from coding ORFs). The ORF reference sequence used as template for PCR amplification of the ORFs is in yellow. Aligned sequencing reads are shown in light green, and those that aligned to the opposite strand in dark green. The final contig assembled using the ORF reference sequence is shown in red. The chromosome, gene name, strand and genomic boundaries of the alignment are indicated at the top. The span of each exonic fragment and intron is indicated on the blue bar at the bottom. This gene (*ATP2C1*) encodes multiple alternatively transcribed variants, and our alignment algorithm was able to correctly assemble the 454 reads to the corresponding sequence.

Supplementary figure 4



Smart Bridging Assembly (SBA). The flow chart describes the main steps in the assembly pipeline. (**1**) collecting sequence reads whose corresponding best match genomic sequence is inside the genomic region of a given ORF; (**2**) determining exons by majority vote, and removing spurious exons with low coverage and short sequence length (for each exon, we compute read coverage, the average number of reads at each nucleotide position inside the exon). We then compute its average and standard deviation. If an exon has a sequence length < 15bp, a corresponding coverage < 10 reads and a coverage z-score of < -1, and is not linked to other exons with more than one sequence read, we consider it a spurious exon; (**3**) connecting adjacent exons when there are sequence reads linking two exons and choosing the corresponding exon-exon junctions by majority vote (the exon-exon junction found in the best match genomic sequence of most sequence reads); (**4**) bridging assembled contigs by either merging the end-exons of the two adjacent contigs if they have a distance of less than 20 bp, or linking two adjacent contigs if there are splicing signals close to the end of two contigs and sufficient bridging sequence. Specifically, we apply the following procedures to bridge adjacent contigs:

- If the distance between the end of these two contigs is less than 20 bp, we fill in the gap with genomic sequences to merge two contigs

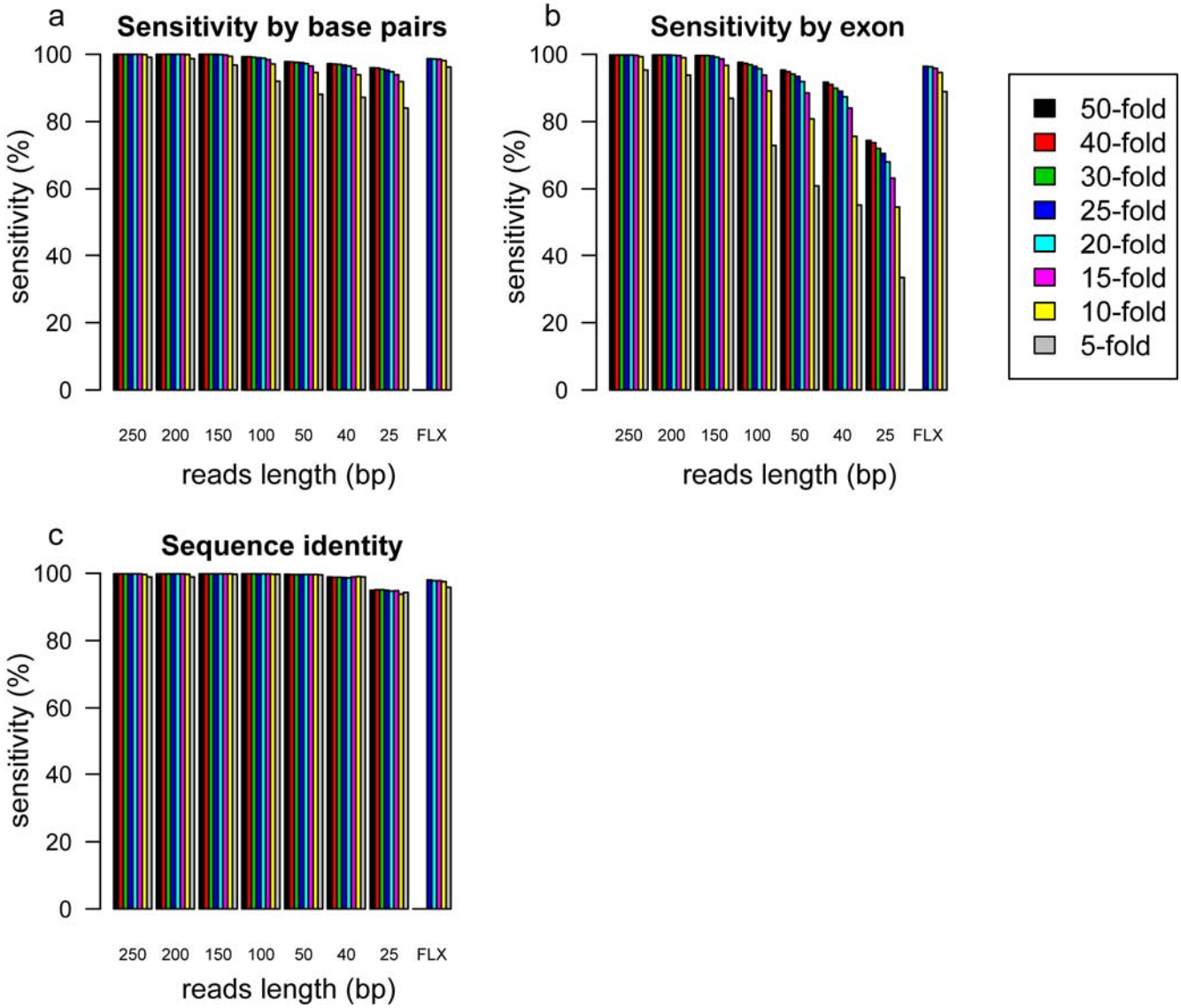
- If the distance is greater than 20bp, we search for appropriate splicing near the end nucleotide of these two contigs (- 50 bp, + 50 bp). For convenience, we have reversed some sequence reads to make them consistent with the orientation of the genomic sequence of the ORF. Consequently, the splicing signal can be either GT…AG or CT…AC. After identifying a candidate splicing site, we collect all sequence reads that have sequences outside this position and obtain the consensus unaligned sequence (≥ 4 bp). We then compare the unaligned sequence of one contig to the corresponding genomic sequences with same length and located inside the splicing site of the other contig, and compute a sequence identity. For each combination of candidate splicing sites connecting the two contigs, we obtain a joint sequence identity based on the unaligned sequence (at least one sequence identity needs to be above 90%). Finally, we select the splicing site that corresponds to the best joint sequence identity to bridge the end exons of these two contigs

- If using the unaligned sequence to find the “bridging” splicing site fails, we search for splicing signals close to the internal ends of the two contigs within a certain range (e.g., (- 10 bp, + 200 bp) for left contig, and (- 200 bp, + 10 bp) for right contig). We choose the splicing site closest to the end nucleotide to bridge two end exons of the two contigs

- if all above fails, we leave the two contigs unbridged.

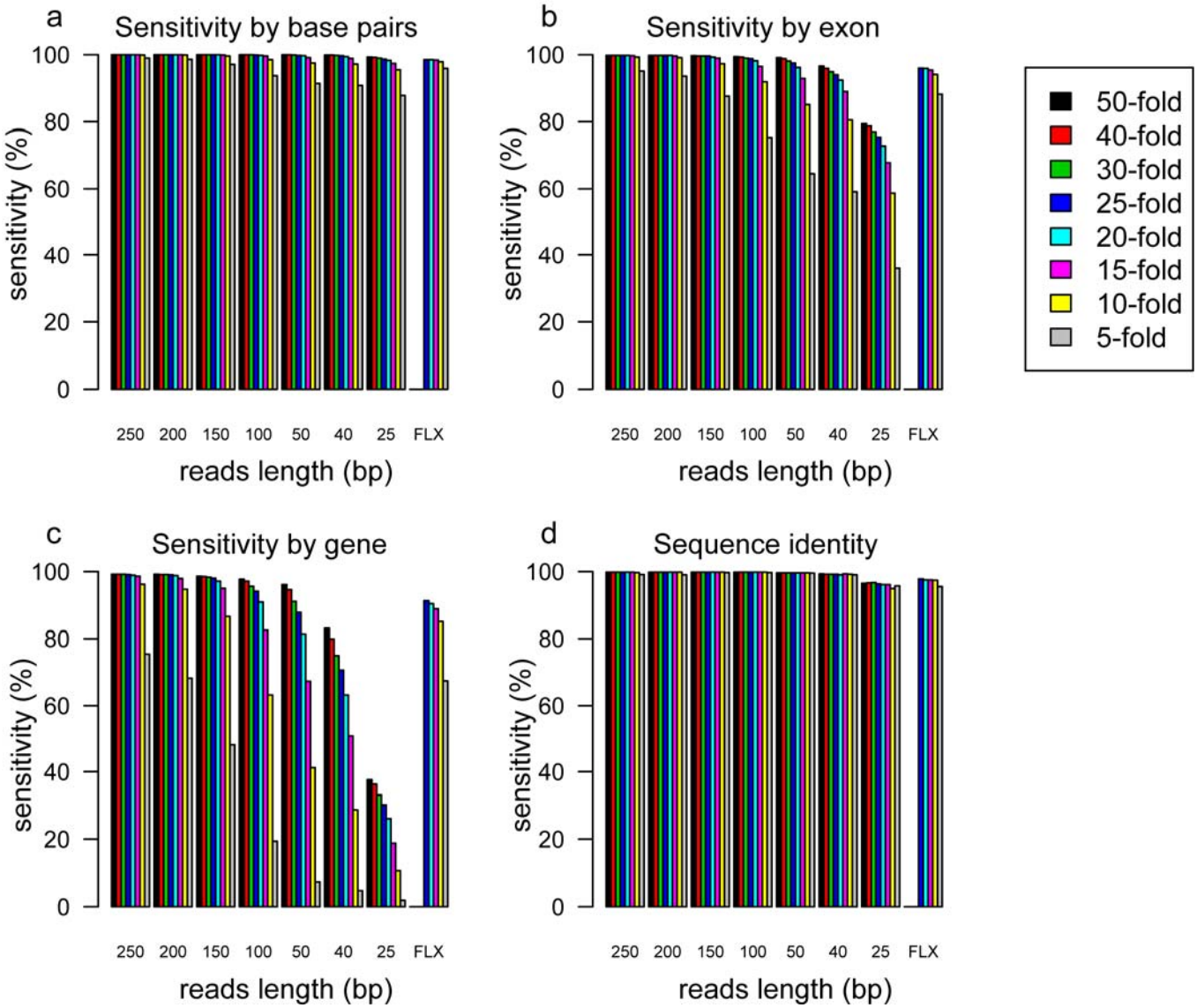
(**5**) repeating steps (**4**) until no more contigs can be bridged; (6) selecting the contig with the longest exon as the representative contig of an ORF.

Supplementary figure 5



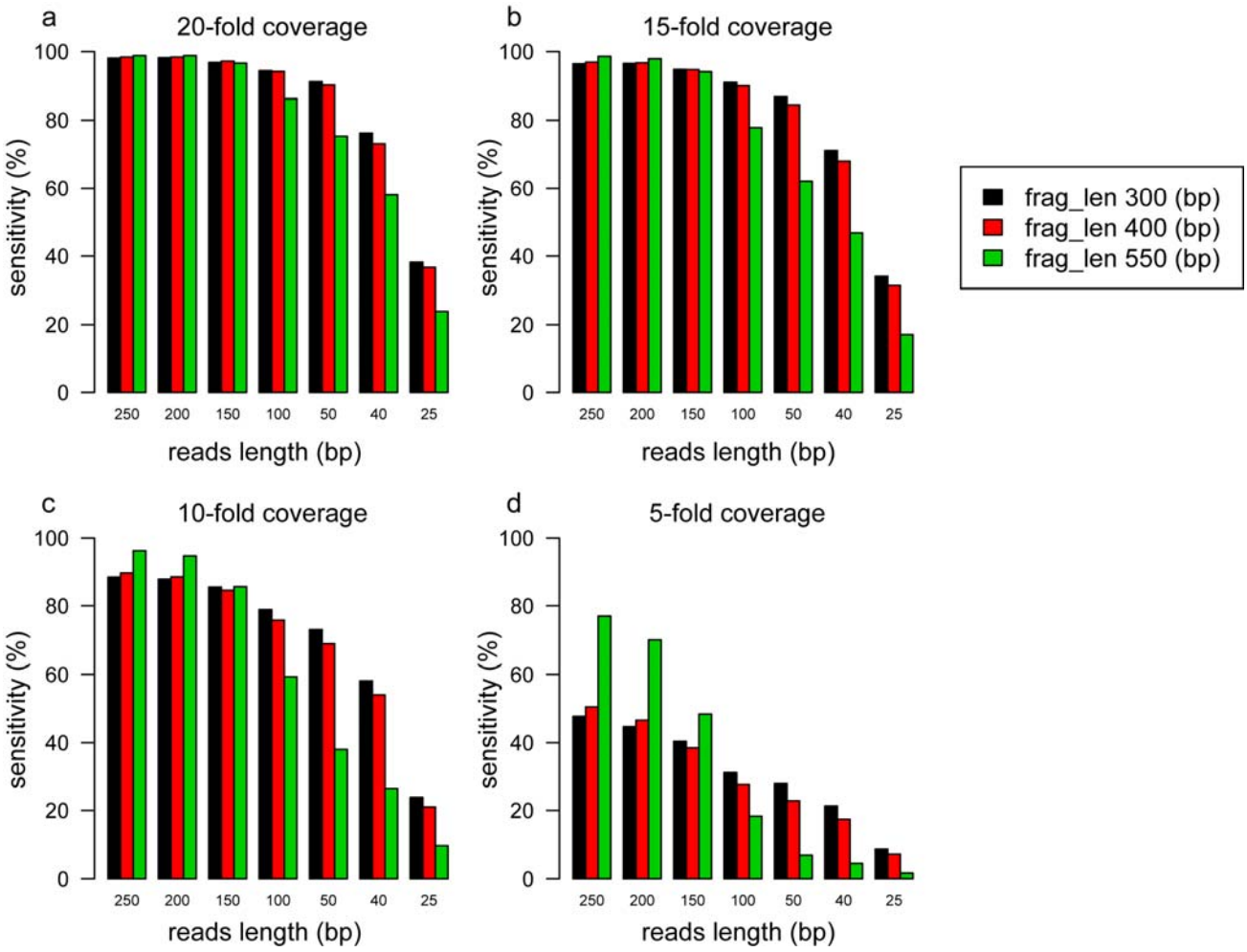
In silico simulation of contig assembly for different read lengths. The same set of ORF sequences used in the 454 FLX run were randomly fragmented *in silico* with average fragment size of 550 base pairs and range of 300-800 bp. Different sequence read lengths (250, 200, 150, 100, 50, 40, and 25 bp), and different fold coverages (50, 40, 30, 25, 20, 15, 10, and 5 fold) were simulated. For each ORF, we assembled contigs based on all available sequence reads that have a corresponding best match in the genomic region of the ORF. Four parameters were evaluated: (**a**) sensitivity by base pair = average percentage of base pairs of an ORF found in the assembled contig; (**b**) sensitivity by exon = average percentage of exons of an ORF found in the assembled contig; and (**c**) sequence identity = average percentage of sequence identity between the assembled contig consensus sequence and the ORF sequence.

Supplementary figure 6



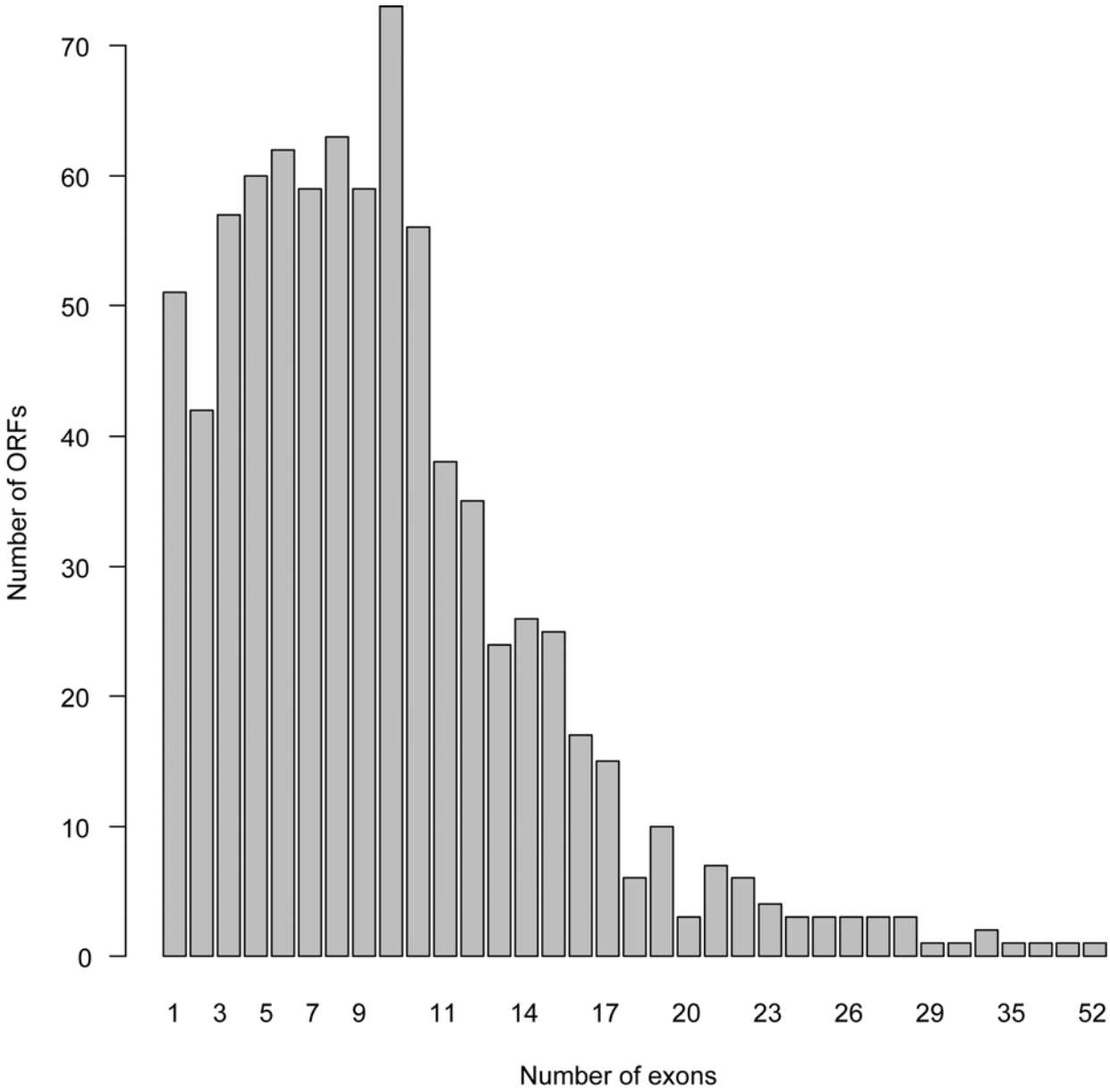
Computer simulation of contig assembly. Simulations were carried out as described in the main text (**Fig. 3b**), except that ORFs with sequence length ≤ 500 bp were excluded from the set. In the simulation, the average size of randomly generated fragments was set at 550 bp to mimic the actual 454 sequencing. Because only the ends of the fragments are sequenced, this setting may result in unsuccessful assembly of short ORFs. After removing ORFs with sequence length < 500 bp, a better assembly rate is obtained. For example, for sequence read length of 50 bp, to have more than 90% and 80% ORFs with correctly assembled gene structure, 30 and 20 fold coverage is needed for ORFs with sequence length > 500 bp, respectively. In contrast, if all ORFs are included, the corresponding fold coverage is 50 and 25, respectively.

Supplementary figure 7



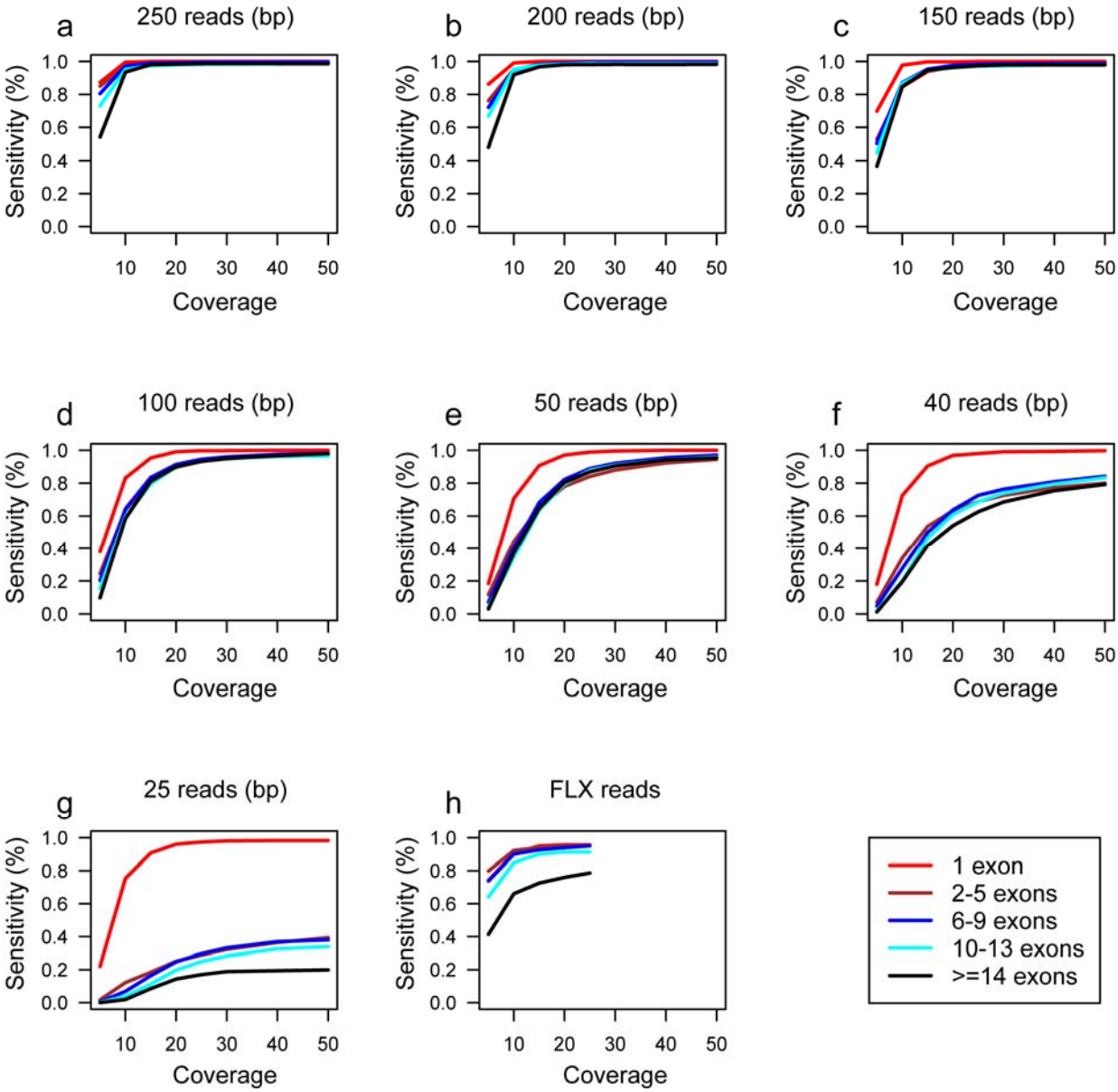
Computer simulation of the effect of fragment size on contig assembly. The average fragment size was set at either 300 or 400 bp (in contrast to 550 bp in **Fig. 3b, Supplementary Fig. 5,** and **Supplementary Fig. 6** online). The assembly sensitivity is compared by gene structure at 20, 15, 10 and 5 fold coverage in (**a**), (**b**), (**c**), and (**d**), respectively. For long sequence reads (≥ 150 bp), reducing fragment size has no effect on assembly rate when fold coverage is greater than 10, and has negative effects when fold coverage is 5. For short sequence reads (≤ 100 bp), reducing fragment size has positive effect on assembly rate at any fold coverage. While the observed assembly rate was 75.3% when the average fragment size was 550 bp, sequence read lengths of 50 bp at 20-fold coverage achieved a correct assembly rate of 91.2% and 90.2% for average fragment sizes of 300 bp and 400 bp, respectively.

Supplementary figure 8



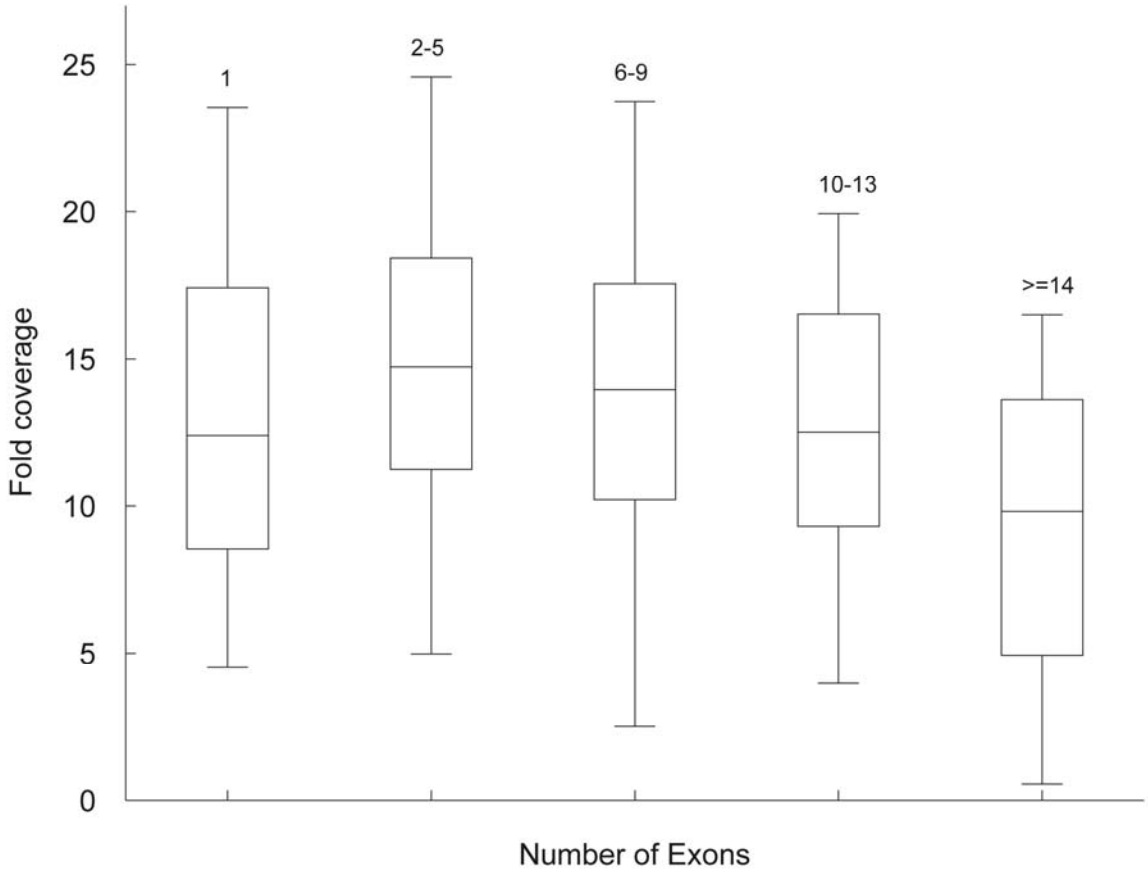
The distribution of genes with different number of exons. Genes were grouped by their corresponding number of exons, then the number of genes in each group was counted.

Supplementary figure 9



The effect of exon number on transcript assembly. Because the transcript length has a major effect on transcript assembly, we removed those transcripts with a length of less than 500 bp from this analysis. We divided the remaining transcripts into five groups according to their exon number: those with one exon (red), 2-5 exons (brown), 6-9 exons (blue), 10-13 exons (cyan), and 14 or more exons (black). For each group, we then measured the sensitivity of the assembled transcripts in terms of gene structure (exon-intron) at different levels of read coverage. Panels (**a-h**) were organized according to different sequence read length.

Supplementary figure 10



The distribution of sequence read coverage of the FLX reads. The read coverage of the FLX reads is estimated to be 25. However, the real read coverage of the FLX reads may be different for different transcripts. As in **Supplementary Fig. 9** online, the transcripts were grouped by their exon number. For each transcript group, we computed the read coverage of the FLX reads for each transcript, and plotted the 5th, 25th, median, 75th, and 95th percentiles of read coverage for these transcripts.

Supplementary Note

ORF analysis

Analyses of the obtained isoforms can reveal novel biological insights. For instance, two different types of alternative splicing events were found to occur in *HSD3B7* (**Fig. 2a**). One type involved skipping of exon 2 while the other involved retention of the intron between exons 3 and 4. With the testis and brain variants that skip exon 2, the coding frame is maintained throughout the length of the isoform, while for the variant retaining the intron the coding frame is shifted resulting in a premature stop codon. Interestingly, these alternative splicing events span a region encoding catalytic domains of the encoded enzyme. Two exons skipped in a novel coding variant of *IFT81* (**Supplementary Fig. 2a** online) are similarly predicted to encode a signal peptide and a transmembrane region. Finally, with *FLCN* (**Fig. 2c**) the coding frame is maintained in two of the variants, while the other variants are frameshifted with or without a premature stop codon.

While some of the non-canonical, non-(GC…AG) variants found are likely to result from PCR mutations or sequencing errors, several of the non-canonical splice junctions recovered in *FLCN* splice variants (**Fig. 2c**) are also captured in the RefSeq, MGC and GenBank databases, and as such likely represent genuine splicing events (**Fig. 2c**). Other instances of non-canonical splicing have been noted[1,2]. These non-canonical variants are best evaluated in light of other biological and experimental evidence and should not be categorically dismissed, as they may represent transcripts of biologically relevant isoforms. For instance, analysis of the resulting transcript as to whether or not the reading frame is preserved or frame-shifted may further help to evaluate the identified sequence. Follow-up assays using junction-specific probes in hybridization experiments can also be carried out to verify the presence of the identified non-canonical junctions.

In silico studies

Simulation of contig assembly for different read lengths

To evaluate the effects of read length on assembly success rates, we examined the rates of successful assembly *in silico* with a simulation of fragmentation and sequencing that tested different read lengths at different depths of coverage. The simulation shows that the fraction of bases determined accurately is not significantly affected by read length (**Supplementary**. **Fig. 5a** online**)**. Even for the most challenging scenario, 25 base read lengths and 5-fold coverage, the assembled contigs still covered more than 80% of the ORF sequence. When examined on a per-exon basis, read lengths of 25 bases recover the complete and accurate sequence for only ~30% of exons at 5-fold coverage, ~50% of exons at 10-fold coverage, and not more than 80% of exons even at 50-fold coverage (**Supplementary Fig. 5b** online). However, with read lengths of 40 and 50 bases the algorithm could correctly and completely identify

more than 90% of exons of a given ORF with fold coverage of 30 and 20, respectively.

Although sequence reads with different length and fold coverage may result in different assembly rates (**Supplementary Fig. 5c** online), the average sequence identity of the assembled contigs to the ORF reference sequences is high. Notably, 25 base read lengths produced an average sequence identity of only ~95%, while all other read lengths examined yielded an average sequence identity of 100%. In summary, these results show that the fraction of genes with completely correct assembly of all exons is more strongly affected by read length than the fraction of correctly-assembled exons or bases.

We also examined the dependence of successful assembly on ORF length and fragmentation procedure by simulating the fragmentation and sequencing process (**Supplementary Fig. 6, 7** online). The average length of the OMIM ORFs investigated was 1,283 bp (compared to 1,369 bp for the average size of human RefSeq ORFs). Excluding ORFs of less than 500 bp improved assembly success rates (the simulation shows that these ORFs do not fragment well, so that internal regions are less touched by sequencing). Decreasing the shearing size of DNA from 500 to 400 or 300 bases *in silico* improves the success rates by providing better access to these internal regions.

The effect of exon number on transcript assembly

Analysis of the ORFs sequenced in this study by exon numbers shows that single exon ORFs only constituted ~6% of the set (**Supplementary Fig. 8** online). To examine the effect of exon number on assembly, we first removed those transcripts with a length of less than 500 bp from this analysis, since transcript length strongly affects transcript assembly (as described in the previous section) and can mask the effect of exon number. We divided the remaining transcripts into five groups according to their exon number. For each group, we then measured the sensitivity of the assembled transcripts in terms of gene structure (exon-intron) at different levels of read coverage. Intronless genes (those with only one exon) have a significantly better sensitivity than those with two or more exons (**Supplementary Fig. 9** online). This is especially obvious for short sequence reads. Even for sequence reads as short as 25 bp, the rate of correct assembly of intronless genes is above 90% when the fold coverage is above 15. The success in assembly of intronless genes is due to the "bridging" mechanism of the Smart Bridging Assembly (SBA) method, *i.e.*, where two contigs aligned to the genome are separated by a gap that is too small to contain an intron, these contigs are "bridged" by filling the gap with the known genomic sequence. For genes with two or more exons, when sequence read length is above 50 bp, the number of exons has no significant effect on transcript assembly except when the read coverage is low (i.e., 5). For short sequence reads, such as 40 and 25 bp, there is a significant effect on transcript assembly only when the number of exons is more than 14.

For FLX sequence reads, the transcripts with many exons (>= 14) have a significantly worse sensitivity, mainly because the real read coverage of the FLX reads in this category is lower than estimated. To address this point, we examined the distribution of read coverage of the FLX reads in relationship to the number of exons. The transcripts with more than 13 exons have lower read coverage than other groups of transcripts (**Supplementary Figure 10** online). The median read coverage of the transcripts in this group is less than 10, while more than 25% of the transcripts have a read coverage of less than 5, which significantly affected the sensitivity of transcript assembly for this group of transcripts. The lower coverage of large ORFs may be due to non-equimolar mixing of amplimers. Such unequal mixing may be due to: 1) large fragments are more difficult to amplify, 2) even when the *amounts* of amplified DNAs are equal, large fragments would have a lower molar concentration as a consequence of their mass. To correct this, a 96-well format DNA quantitation step can be introduced before mixing to normalize molar concentration of amplimers.

In general, not all clones in a deep well are expected to be sequenced successfully in the first attempt. For instance, ORFs with many exons will have lower success rates than the average ORF. For any ORF whose assembly has failed, a new deep well with fewer ORFs can be generated, allowing sequencing at a higher coverage. The use of mate-pair reads, if supported by the sequencing platform in use, can further enhance the coverage in such deep wells by providing additional reads. We expect that nearly all ORFs can be sequenced successfully using this iterative approach.

Supplementary Methods

RT-PCR

The primers used for RT-PCR are those previously used to generate human ORFeome versions 1.1 (ref. 3) and 3.1 (ref. 4). The forward primers begin with the ATG start codon then extend into the ORF; the reverse primers terminate just before the stop codon. The forward and reverse primers were designed to have a T_m (for the genome-complementary segment) of 60 ± 5°C. Both forward and reverse primers carry the Gateway tails for recombinational cloning. Human ORFeome primer sequences can be found at http://horfdb.dfci.harvard.edu.

RT-PCR was carried out using testis, brain, heart, liver and placenta RNAs purchased from Ambion Corp. Total RNA isolated from these tissues was reverse-transcribed using dT_{16} and random hexamer primers. Primed RNAs were then incubated with Superscript III RT (Invitrogen) using the recommended conditions, such that the final RNA concentration was 0.3 µg / µl. Following heat inactivation at 94^{o}C, the reverse transcribed material was added to a “master PCR mix” and dispensed into 96-well plates containing all PCR reagents, including the hot-start KOD polymerase and reverse-transcribed template (10 µl

per 1 ml of PCR mix). Gene-specific primers were added last from PCR-ready 96-well plates. PCR was performed for 40 cycles, with each cycle consisting of denaturation at 94°C and annealing at 60°C for 30 seconds followed by extension at 70°C for 1 to 5 minutes, depending on the expected length of the amplicons.

Gateway cloning of amplicons

PCR products comprising cloning Sets 1, 2, and 3 were recombinationally cloned into a Gateway compatible plasmid (pDONR223) using the BP reaction[3]. To ensure scalability of the procedure, no gel purification was performed. The products from the BP reactions were used to transform chemically competent DH5α *E. coli* in 96-well format plates containing spectinomycin for growth and selection of entry clone-bearing cells. Following growth in liquid media, the transformant minipools were plated on solid media plate. Twelve isolated colonies were picked, grown in liquid culture, then used as template to generate final DNA template for sequencing. PCR products were sequenced using conventional automated cycle sequencing to generate ORF sequence tags (or OSTs[5]).

Alignment of Sanger-sequenced ORFs (Sets 1-3), mRNAs and ESTs

The Sanger-sequenced ORFs were aligned to the genome as follows. Forward and reverse sequences were vector-trimmed with cross_match (version 0.990329, default options) and assembled using Phrap[6]. Sequences were aligned to the genome (UCSC version hg17) using Blat (version 31x1, option oneOff=1)[7]. Fragments with identity lower than 90% were discarded. For each sequence, aligned fragments were joined if they were immediately contiguous or separated by one base pair, and if the corresponding genomic fragments were in concordant order and orientation no more than 300 Kbp apart. The SIM4 program[8] was used to realign the sequences at all loci reported by BLAT, and those results were added to BLAT alignments. To identify transcripts reported in incorrect orientation, we then examined internal flanking positions, presumed to be splicing signals. We determined the strand of the alignment by assigning a score of +1 for each GT...AG, GC...AG and AT...AC pair of splice donor and acceptor dinucleotides, -1 for their reverse complements, otherwise + 0.5 for a GT donor and - 0.5 for an AC acceptor (which is likely to be a reverse-complemented GT donor). We summed these scores for each alignment, and reverse-complemented alignments with negative sums. The best alignment was chosen as that containing the greatest number of matching bases in one or more putative exons not separated by a splice junction other than GT…AG. Subsequently, only intron-spanning alignments were retained.

mRNAs and ESTs overlapping the Sanger-sequenced ORFs were obtained from two sources. First, coordinates of UCSC alignments of Genbank and Refseq mRNAs (09-Apr-2008) and ESTs (25-Mar-2008) were scanned for any overlaps with the ORFs; ESTs with multiple reported genomic loci were excluded. Second, these mRNAs and ESTs were matched to their UniGene

clusters (Build 210), and their coordinates were extended to any unaligned sequences within the clusters. All sequences were then realigned to the genome at the reported loci using BLAT (options oneOff=1, trimT). Adjacent segments were joined, and alignment orientations were corrected if necessary. The best alignment was chosen as the one with the highest number of matching bases, with low-complexity stretches detected by Dust (NCBI toolkit) excluded to avoid matches to retro-pseudogenes. Alignments containing any splice signal pair other than GT…AG were repeated with SIM4 and replaced if successful; any remaining non-canonical alignments were retained. Membership of GenBank mRNAs (Release 162) in the Mammalian Gene Collection (MGC) was determined from GenBank records. Alignments of PCR sequences were then mined for splice junctions and retained introns not present in overlapping mRNAs, and candidates were examined visually for valid splicing signals and unambiguous alignments. The presence of a segment of poor quality in a Sanger-sequenced ORF did not preclude its novelty if the novel splicing event occurred elsewhere in the sequence.

454 Sequencing

PCR products generated from ~820 cDNA templates were pooled in equimolar ratios into one well, partially purified using a Qiagen PCR Cleanup column and sheared by nebulization to a size range of 300-800 bases. The sheared DNA was end-repaired and adapters were ligated to the ends of the DNA. The products were purified and used to set up an emulsion PCR reaction. The emulsions were processed, broken and the clonally amplified beads were enriched. Approximately 600,000 DNA carrying beads were loaded onto two of the four regions of a Picotitre plate along with enzyme beads and packing beads, and inserted into the cartridge on the 454 FLX instrument. The FLX was run for 7.5 hrs during which 100 flow cycles were completed.

Alignment and assembly of short reads (fOSTs)

Genomic sequences of each ORF were obtained from human reference genome sequence (UCSC version hg17) according to its genomic location. We then developed two methods of assembly using genomic sequences as reference, 1) “Conventional assembly” employing existing computational tools, 2) “Smart Bridging Assembly” or SBA for which we developed specialized software to enhance assembly success rates (**Supplementary Fig. 3** online)**.** For the first method, reads were vector-trimmed using cross_match[6], and then grouped based on their BLAT results against reference sequences. Reads in each group were assembled into one or more contigs using CAP3[9] with options “-f 10 -o 21 -c 12”. Since CAP3 requires a minimum overlap length of 20 nucleotides, contigs and singlets from CAP3 in each group were further assembled using Phrap with options “-minscore 8 and -minmatch 8” to catch sequences sharing 8 nucleotides and more. To determine quality of assembled contigs, we compared them with their corresponding original ORF sequences using BLAST (bl2seq).

For the SBA method (**Supplementary Fig. 3** online) every read was aligned by BLAT against the reference ORF sequences. We collected all sequence hits corresponding to a given sequence read and assembled its best matching genomic alignment. The algorithm looks for a continuous stretch of genomic sequence that not only covers most of the sequence read, but also has the best sequence identity to the sequence read. For some sequence reads, the corresponding best match may include one or more introns. These best matches are used to determine the exon-exon structure when assembling contigs.

Once the best matching genomic alignment is identified for each sequence read, all corresponding best matches are pooled and assembled into contigs. To do this, the algorithm first determines the exons by "majority vote": a) a nucleotide position is considered to be inside an exon if this is indicated by the majority of best matches; b) an exon is a continuous fragment in which all nucleotides are said to be inside the "exon". After determining the exons, a filtering procedure removes spurious exons that are both short (e.g., < 15bp) and have significantly lower coverage than other exons in this genomic sequence. Next, all exon-exon junctions are determined from those best matches that span introns, and coverage is computed by counting the number of best matches including these structures. Again, employing majority vote, the program selects the exon-exon structure with the best coverage to connect two exons. We repeat this step to connect more exons until there are no available exon-exon structures. This may result in one or more contigs for a given gene/ORF. When there are multiple contigs, we attempt to merge them successively until no contigs can be merged. In our simulation, the contig with the longest exon is chosen as the representative contig.

The SBA method uses two additional approaches to merge additional contigs. First, two consecutive contigs separated by less than 20 bp (a gap too small for an intron) are merged and the missing exonic fragment sequence is inferred from the corresponding genomic sequence. Second, contigs separated by more than 20 bp are assessed for the possibility that an intron occurs between them but that sequence "bridging" the exon-exon junction was not of sufficient length for BLAT to connect the two exons. For this, we examined the proximal ends of each contig for appropriate candidate splice sites. For each combination of candidate splice sites, and each sequence read mapping to the 5' contig with a region 3' to the candidate splicing signal (including sequence unaligned by BLAT), we compared this region with the genomic sequence downstream of the candidate splice site in the 3' contig. This process was repeated, with regions of 3' contig reads that extend 5' to the downstream candidate splice site being compared to the genomic sequence upstream of the upstream candidate splicing signal. We used the combination of candidate splicing sites yielding the alignments with the best sequence identity to bridge the two exon ends. If this fails, we search for candidate splice sites closest to the last aligned base on both contigs within a certain distance to bridge the two exon ends. Failing this, we

leave the two contigs unbridged. Further details on the "bridging" procedures can be found in **Supplementary Fig 3** online.

Simulation of sequencing and assembly to assess effects of read length and coverage

Simulation was based on random fragmentation of the known cDNA sequences of the 820 human "disease" ORFs. In each random fragmentation, a given cDNA sequence is randomly sheared into N fragments, with N following a Poisson distribution ($\lambda = \frac{cDNA length}{Average\ fragment size}$). Real fragments in the 454 sequencing experiment range from 300 to 800 bp, so the average fragment size was roughly estimated to be 550 bp. We repeated the random fragmentation many times to obtain many fragments for each cDNA sequence. Then, we determined the number of random sequence reads needed for a given experiment with specific sequence read length and fold coverage as $M = \frac{fold\,coverage \times \sum_{n=820} cDNA length}{read\,length}$. Because ORFs with longer cDNAs yield more fragments, a given fragment has a probability of $P_i = \frac{cDNA\ length_i}{\sum_i cDNA\ length_i}$ from a specific cDNA. Finally, we "sequenced" each fragment from either the 5' or 3' end, using read lengths sampled from $r = read\ length \pm (\frac{read\ length}{10})$, and modeling sequencing error to be 2% with the erroneous base chosen uniformly at random from the remaining three possible bases. To explore the impact of fold coverage and sequence read length on final transcript assembly, we generate random sequence reads with read length ranging from 25, 40, 50, 100, 150, 200, to 250 bp, and fold coverage ranging from 5, 10, 15, 20, 25, 30, 40, to 50 fold.

Supplementary references

1. Ng, B. *et al.* Increased noncanonical splicing of autoantigen transcripts provides the structural basis for expression of untolerized epitopes. *J. Allergy Clin. Immunol.* **114**, 1463-70 (2004).
2. Burset, M., Seledtsov, I. A. & Solovyev, V. V. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.* **28**, 4364-75 (2000).
3. Rual, J. F., Hill, D. E. & Vidal, M. ORFeome projects: gateway between genomics and omics. *Curr. Opin. Chem. Biol.* **8**, 20-25 (2004).
4. Lamesch, P. *et al.* hORFeome v3.1: a resource of human open reading frames representing over 10,000 human genes. *Genomics* **89**, 307-315 (2007).

5. Reboul, J. *et al.* Open-reading-frame sequence tags (OSTs) support the existence of at least 17,300 genes in *C. elegans*. *Nat. Genet.* **27**, 332-336 (2001).
6. Gordon, D., Abajian, C. & Green, P. Consed: a graphical tool for sequence finishing. *Genome Res.* **8**, 195-202 (1998).
7. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656-64 (2002).
8. Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M. & Miller, W. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* **8**, 967-74 (1998).
9. Huang, X. & Madan, A. CAP3: A DNA sequence assembly program. *Genome Res.* **9**, 868-77 (1999).
10. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514-517 (2005).