Supplemental Material for:

RefSeq Functional Elements as experimentally assayed non-genic reference standards and functional interactions in human and mouse

Catherine M. Farrell*, Tamara Goldfarb, Sanjida H. Rangwala, Alexander Astashyn, Olga D. Ermolaeva, Vichet Hem, Kenneth S. Katz, Vamsi K. Kodali, Frank Ludwig, Craig L. Wallin, Kim D. Pruitt, and Terence D. Murphy

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

*Corresponding author, e-mail: farrelca@ncbi.nlm.nih.gov

Table of Contents

Supplemental Methods

Record curation

RefSeq sequence and Gene database records were curated simultaneously. For sequence records, experimentally validated features were curated from the literature according to criteria and scope described in the Results section and Supplemental Table S2B. Functional elements were selected for curation initially based on manually scanning the literature for review articles on element types in scope, e.g., gene regulatory elements, recombination regions or replication origins, and then identifying specific well-characterized elements described therein, searching for specific publications or citations in PubMed, and following links to full text articles in PubMed Central or via journal websites. Other strategies for identifying elements for curation included targeted searches for elements associated with genes of high biomedical interest (e.g., *ACE2*, *BRCA1*, *CFTR*, *HBB* and other frequently accessed genes in the NCBI Gene database), searches for publications that employ large-scale screening techniques (e.g., MPRA or CRISPRi screens), searches in PubMed or PubMed Central for specific terms like "validated enhancer" or "experimental validation," and through outreach efforts such as identification of published studies from presentations and informal discussions at research conferences or based on requests from users and colleagues in the research community. Following identification of initial publications with experimental evidence for each functional element, additional publications were identified by searching the literature for related publications from the same laboratory, by reviewing cited publications, by searching for publications that use names or aliases specific for each element, by finding publications associated with aligning sequence submissions, and/or by finding later publications that cite the initially identified publications with experimental evidence. Experimental data were processed from publication methods sections, figures or supplemental material, where genomic regions used for feature ranges were based on raw sequences supplied directly in publications, associated download files or public database accessions, or based on genomic fragments determined from published PCR primers, restriction enzyme fragments or genomic coordinates. Coordinates were remapped to the current human or mouse reference assemblies using the NCBI Remap (Kitts et al. 2016) or UCSC LiftOver (Rosenbloom et al. 2015) tools when necessary. If required, corresponding authors were contacted for data clarifications or to obtain missing data elements needed for curation, e.g., primer sequences, genomic coordinates or assembly version information. For manual curation, a Genome Workbench (https://www.ncbi.nlm.nih.gov/tools/gbench/) in-house custom alignment tool was used to align experimentally validated sequences from publications to the genome, RefSeq biological region ranges were determined following grouping of closely located and functionally related features (if applicable), and preliminary genomic RefSeqs were uploaded to an NCBI backend database, as described previously for RefSeq transcript and protein accession creation

(Rajput et al. 2015). A Genome Workbench editing interface (Sayers et al. 2020; https://www.ncbi.nlm.nih.gov/tools/gbench/tutorial12/) was used to annotate features on preliminary RefSeq accessions according to INSDC specifications and as described in the Results section and Figure 2. An additional 100 bps of sequence was added to each end of all RefSeqs for additional genomic context, though it was not included in the overall parental 'biological_region' feature range nor in underlying functional feature ranges. Annotated feature ranges were mostly based on the exact experimentally tested fragments, which frequently include extensions to regions denoted as core sequences in publications, e.g., protein binding oligonucleotides used in gel shift assays typically include sequences flanking core binding motifs. Descriptions of the logic and conditions used for the annotation of specific feature types are available in Supplemental Table S2B and the feature annotation glossary on our webpage (www.ncbi.nlm.nih.gov/refseq/functionalelements/#Feature_Annotation_Glossary), which also includes links to feature terms and definitions in the INSDC and SO databases. Overlapping feature annotation was allowed for distinct in-scope features, where each feature was treated as a separate entity and annotated separately based on the exact range shown to exhibit the particular experimental activity, including features with different SO_types (e.g., a regulatory element that functions as an enhancer in one cell type and a silencer in another, or a protein binding site that overlaps a regulatory features, as in the Figures 2 and 3 examples) or even features with the same SO_type but with distinct ranges and activities from different or sometimes the same publication (e.g., overlapping enhancers that are active in different cell types). Internal quality assurance (QA) was employed to ensure the presence of valid biological region features on all RefSeqs and the use of appropriate controlled vocabularies and mandatory field provisions according to INSDC and RefSeqFE specifications, as described in the Results section, Supplemental Table S2B and on our webpage (www.ncbi.nlm.nih.gov/refseq/functionalelements/#Functional_Element_RefSeq_Feature_Annotation), followed by final NCBI processing to produce the public genomic RefSeq (NG_ accession) in flat file, ASN.1, FASTA and other format options available through the Nucleotide database (https://www.ncbi.nlm.nih.gov/nucleotide/).

Automated or semi-automated processing was used to produce RefSeqs for large-scale experimentally validated features. For VISTA enhancers, sequences for the human and mouse positive subsets were downloaded from the VISTA database (https://enhancer.lbl.gov/; Visel et al. 2007), aligned to the current human or mouse reference genomes by BLAST, followed by automated RefSeq creation and feature annotation in an in-house database, with subsequent QA testing prior to public RefSeq release in the Nucleotide database. For other large-scale datasets including bulk-screened, MPRA- or CRISPRi-validated functional elements (e.g., Wang et al. 2006; Roh et al. 2007; Petrykowska et al. 2008; Narlikar et al. 2010; Kheradpour et al. 2013; Andersson et al. 2014; Ernst et al. 2016; Gasperini et al. 2019; Fulco et al. 2019), data were wrangled prior to RefSeq curation using UNIX command line tools to select higher confidence data with appropriate filtration when necessary (e.g., to select positively validated elements or those meeting a threshold considered significant by the authors), to determine overlaps with pre-existing records, to remap genomic ranges to the current reference assembly and to retrieve feature sequences in bulk, followed by manual RefSeq creation with customized markup. For both automated and manual processing, if a functionally related feature from a new data source overlapped a pre-existing biological region, the new feature was manually annotated on the pre-existing RefSeq alongside feature annotation from other sources, with RefSeq end adjustment and version incrementation if necessary.

For creation of biological region records in the Gene database, an in-house CGI interface was used to curate and store associated publications (identified as described for sequence records above, also including related publications not used for experimental evidence, e.g., review articles), nomenclature, sequence associations with ranges, a curator-provided summary derived from information in the literature, manually curated orthology links (human and mouse only), chromosome, organism and biological region locus type designations, and links to external databases if applicable. NCBI automatic processing was used to convert to public records and to populate other data fields, including graphical displays, GDV browser links (following genome annotation), GeneRIFs, phenotypes, variation and Gene LinkOut data if applicable, as described in Gene documentation (www.ncbi.nlm.nih.gov/books/NBK3841/#EntrezGene.Quick_Start). Listings in the ‘Feature type(s)’ field (Supplemental Fig. S1, green tab) were populated from feature types annotated on the associated RefSeq accession, with groupings based on feature classes. As is the case for NCBI conventional gene records, all information associated with biological regions may be dynamically updated, as indicated by timestamps on Gene records, RefSeq summaries, genome annotation releases and associated RefSeq accessions.

Genome annotation

RefSeqFEs were treated as curated RefSeq genomic sequences in the NCBI Eukaryotic Genome Annotation Pipeline workflow (Supplemental Table S1 documentation links with alignment algorithms provided in www.ncbi.nlm.nih.gov/books/NBK169439/; Pruitt et al. 2014; McGarvey et al. 2015). RefSeqFE genomic records were aligned to the human GRCh38.p13 (accession GCF_000001405.39) or mouse GRCm39 (accession GCF_000001635.27) reference genome assemblies using BLAST (Altshul et al. 1990), and filtered to confirm their intended location based on coverage, identity, and matching genome assembly components. Placements on alternative assembly components were retained if the location was considered allelic based on alignments between the primary and alternate assembly sequences. The RefSeqFE features were then projected onto the reference genome coordinates, and identified with ‘RefSeqFE’ source markup. Features were named based on their SO_types (feature types in Supplemental Table S2A,B and the feature table at www.ncbi.nlm.nih.gov/refseq/functionalelements/#Feature_table) and assigned to biological region GeneIDs based on ‘/db_xref’ qualifier values in feature metadata (see Fig. 2). Annotation locations for parental ‘biological_region’ features were propagated to relevant Gene database records, data were formatted to include feature metadata, and then made publicly available on FTP (multiple links in Supplemental Table S1) and in ‘Biological regions, aggregate’ tracks in graphical displays, as described in the Results section.

Interaction data

Interactions having the experimental support criteria described in the Results section were tracked along with supporting publications in an in-house database during curation. All interactions were tracked at the level of full-length biological regions or target genes. Following relevant genome annotation releases, a combination of SQL and UNIX command line was used to determine pairwise interactions and genome annotation locations of interacting biological regions and target genes. These were compiled in bigInteract format (https://genome.ucsc.edu/goldenPath/help/interact.html; Haeussler et al. 2019) with a custom column listing supporting publications, where the `bedToBigBed` utility from the UCSC binary utilities directory (http://hgdownload.soe.ucsc.edu/admin/exe/; Kent et al. 2010) was used to convert to binary format along with a custom autoSql file containing field descriptions, which can be viewed from bigInteract files using the UCSC `bigBedInfo` utility with the `-as` option. Interaction types were labeled in the name field (column 4) along with the symbols of both loci involved in the relevant interaction, where the locus with the 5’-most start position on the chromosome is listed first and is represented in columns 9-13 (see name formatting details in HTML documents at https://ftp.ncbi.nlm.nih.gov/refseq/FunctionalElements/trackhub/hg38/). The files were publicly released on the RefSeq FTP site (download paths in Supplemental Table S1). Interaction data were also displayed as pipe-separated lists of interacting loci symbols in a custom column in biological region bigBed files (described for Track hub provision below), in ‘/function’ qualifiers on biological region features on RefSeq flat files, and as attributes in column 9 of GFF3 files for ‘biological_region’ features, when applicable. Additional details are available on our webpage (www.ncbi.nlm.nih.gov/refseq/functionalelements/#Interactions).

Track hub provision

The RefSeqFE Hub was created according to UCSC track hub specifications (Raney et al. 2014) and hosted on the NCBI RefSeq FTP site (https://ftp.ncbi.nlm.nih.gov/refseq/FunctionalElements/trackhub/hub.txt). For all tracks, genomic coordinates were derived from annotation locations in the indicated human or mouse annotation release (GFF3 file download paths indicated in Supplemental Table S1), and chromosome notation was converted from RefSeq- to UCSC-style format based on mapping provided in genome assembly reports from the NCBI Assembly database (https://www.ncbi.nlm.nih.gov/assembly/; Kitts et al. 2016). For feature tracks, features were extracted from GFF3 files based on the 'RefSeqFE' source (column 2) but excluding parental 'biological_region' features (command for extraction of features from GFF3 files available at www.ncbi.nlm.nih.gov/refseq/functionalelements/#Feat_extraction). Various metadata fields including select attributes in column 9 were extracted and split among appropriate columns to produce feature files in bigBed (binary indexed BED 9+4) format, also joining on the associated RefSeq accessions from the relevant biological record (available in the `gene2refseq.gz` file at https://ftp.ncbi.nih.gov/gene/DATA/), and with conversion from non-binary to binary format using the `bedToBigBed` utility from the UCSC binary utilities directory along with a custom autoSql file containing field descriptions. Features were named (column 4) by concatenating the SO_types (column 3 in GFF3 file) with the relevant biological region GeneID (extracted from 'Dbxref' attributes in column 9), and with inclusion of the bound moiety (column 9 attribute) for 'protein_bind' features. Parental biological region bigBed files were prepared in binary indexed BED 9+5 format by extracting various metadata fields from Gene records, with addition of interacting loci strings as displayed in RefSeq flat files (described above). Regulatory interaction and recombination partner tracks were prepared in bigInteract format as described for interaction data above. Further content descriptions and the FTP locations of our bigBed and bigInteract files are available on our webpage (www.ncbi.nlm.nih.gov/refseq/functionalelements/#bigBed), with additional details in track-specific HTML documents (multiple URLs in Supplemental Table S1) and in table schema provided by compatible genome browsers. Schema for our bigBed and bigInteract files can also be obtained by using the `bigBedInfo` utility with the `-as` option to display field descriptions in the output, and these binary indexed files can be converted back to non-binary format using the `bigBedToBed` program from the UCSC binary utilities directory.

Publication statistics

Publications (listed as PMIDs) were extracted from column 12 of `bigBedToBed`-converted human `FEfeats_AR109.20201120.bb` and mouse `FEfeats_AR109.bb` files (available at https://ftp.ncbi.nlm.nih.gov/refseq/FunctionalElements/trackhub/data/). Any comma-separated lists of PMIDs were de-collapsed in separate lines, where each line represents the use of each PMID as evidence for a feature. Human and mouse extracted PMIDs were combined, sorted and counted using the `uniq –`

c command to determine the number of features each PMID was used as evidence for, as shown in columns A-B of Supplemental Table S2C. The distribution of feature counts derived from publications was further determined, again using the uniq -c command, e.g., to determine how many publications were used as evidence for just one feature and so on, as shown in columns D-E of Supplemental Table S2C and the embedded chart.

Biological region statistics

GeneIDs were extracted from the name field (column 4) of bigBedToBed-converted human FEfeats_AR109.20201120.bb and mouse FEfeats_AR109.bb files, as used for publication statistics above. The GeneIDs were sorted, uniquified and counted to determine how many distinct biological regions were annotated for human AR 109.20201120 and mouse AR 109. These numbers were independently verified by sorting, uniquifying and counting GeneIDs extracted from 'Dbxref' attributes (column 9) of parental 'biological_region' features in equivalent AR GFF3 files (FTP paths in Supplemental Table S1). The biological region counts are displayed in Figure 4E (Locus count row) and Supplemental Table S2D. To determine the number of features per biological region, GeneIDs were extracted from bigBedToBed-converted FEfeats_AR##.bb files for primary annotation locations only (e.g., to omit secondary annotations of the same biological region on alternative genome assembly components), then sorted and counted using the uniq -c command. Counts were further determined for biological regions with single features, multiple features and >=10 features, as indicated in Supplemental Table S2D. Biological regions with at least one feature from a large-scale study were determined from bigBedToBed-converted FEfeats_AR##.bb files by selecting features based on publications resulting in >50 features as shown in Supplemental Table S2C (PMID:27701403, PMID:17130149, PMID:30612741, PMID:23512712, PMID:19165926, PMID:24670763, PMID:25395542, PMID:20981099, PMID:31784727, PMID:26614388 and PMID:17135569), then sorting, uniquifying and counting distinct GeneIDs extracted from the name field. Single-feature biological regions based on a large-scale study were similarly counted but with selection of GeneIDs with single features, as determined above.

Feature statistics

Feature counts were based on the extraction of 'RefSeqFE' source features from the human AR 109.20201120 or mouse AR 109 GFF3 files (FTP paths in Supplemental Table S1), where SO_types (column 3) were extracted and counted on the UNIX command line, with exclusion of parental 'biological_region' features (see command for extraction of features from GFF3 files at https://www.ncbi.nlm.nih.gov/refseq/functionalelements/#Feat_extraction). Groupings by feature class or type are indicated in Supplemental Table S2A. Average feature lengths were calculated by extracting individual feature types, totaling their lengths and dividing by the number of features. Average lengths were also determined for features grouped by class, all features combined and parental biological regions. Feature length minimums, maximums and standard deviations from the mean were determined for individual feature types, feature groupings and biological regions using UNIX command line calculations. Boxplots for feature length distributions, including for individual and grouped features shown in Figures 4 and 6 and Supplemental Figures S3, S4 and S6, were created using the BoxPlotR web tool (http://shiny.chemgrid.org/boxplotr/; Spitzer et al. 2014). Genome coverage was calculated following conversion of non-parental functional features to BED format (https://genome.ucsc.edu/FAQ/FAQformat#format1; chromosome, start converted to 0-base, stop, concatenated SO_types and GeneIDs for name) and with collapse of all overlapping features using the `bedtools merge` command from the BEDTools software package (Quinlan 2014).

Gene-relative location determinations

RefSeqFE features were converted to BED format as above, where the bound moiety was additionally concatenated with the SO_type and GeneID in the name field (column 4) for 'protein_bind' features. All gene type features ('gene', 'pseudogene' and '*_gene_segment' SO_types) were extracted from the same human and mouse GFF3 files, with conversion to BED format using a concatenation of gene biotype, GeneID and gene symbol for the name column. Clinically relevant gene feature subsets were subsequently extracted from the same BED file based on representation in the RefSeqGene (RSG), Locus Reference Genomic (LRG) and ClinVar submissions datasets. The RSG gene list was extracted from https://ftp.ncbi.nih.gov/refseq/H_sapiens/RefSeqGene/gene_RefSeqGene on February 9, 2021, and the LRG gene list from https://www.lrg-sequence.org/data/ on March 8, 2021. ClinVar submissions genes were selected from the gene summary file dated March 2, 2021 at https://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab_delimited/gene_specific_summary.txt, with filtration for submissions that reported the gene (column 5 >=1) and for genes with pathogenic or likely pathogenic variants (column 6 >=1) to assert a higher probability of clinical relevance. Exon and CDS features were extracted from relevant genome annotation GFF3 files and converted to BED format, with capture of the SO_type ('exon' or 'CDS'), GeneID and gene symbol in the name field, and with strand-specific collapse of all overlapping exon or CDS features using the `bedtools merge` command (Quinlan 2014). Gene 5'-proximal 2 kb regions were determined by first extracting annotated start positions for transcript type features in GFF3 files (features with a `transcript_id` attribute in column 9, NM_ or NR_ accession prefixes only and excluding exon features), converting to BED format with a concatenation of the GeneID and gene symbol in the name field, uniquifying according to strand, and then modifying the coordinates to represent 2 kb upstream of plus strand starts or 2 kb downstream of minus strand starts. BED files were additionally prepared for intronic regions by using the `bedtools subtract` command (Quinlan 2014) with relevant BED files to subtract exons from gene ranges, for UTR regions by subtracting CDS regions from exons, for whole intergenic regions by subtracting gene ranges from GRCh38.p13 and GRCm39 whole chromosomes/scaffolds/contigs, and for gene 5'-distal intergenic regions by subtracting gene 5'-proximal intergenic regions from whole intergenic regions, where gene ranges had additionally been subtracted from 2 kb upstream of transcript start sites to remove any residual genic regions resulting from nearby genes or splice variants that start internally within a gene.

For the called overlaps shown in Figure 5A,B and Supplemental Table S3, the above BED files were used to perform serial intersections of RefSeqFE features with gene, exon, CDS and gene 5'-proximal features using the `bedtools intersect` command (Quinlan 2014), where RefSeqFE features or subsets were used in the A file and at least 10% RefSeqFE feature overlap was required (`-f 0.10` option). RefSeqFE features were first intersected with gene features where the positive subset was scored as gene-overlapping, while non-overlapping features were scored as intergenic. Gene-overlapping RefSeqFE features were then intersected with exon features, where non-overlapping features were scored as intronic. Exon-overlapping RefSeqFE features were subsequently intersected with CDS features, where non-overlapping features were scored as UTR-overlapping. In addition, intergenic RefSeqFE features were intersected with gene 5'-proximal features, where non-overlapping features were scored as intergenic, gene 5'-distal. For all intersections, RefSeqFE features in relevant overlapping or non-overlapping subsets were uniquified and counted, and the percentage of features in each gene-relative location was calculated and charted (Fig. 5A,B). Following determination of all RefSeqFE feature classifications (gene-overlapping, exon-overlapping, intronic, CDS-overlapping, UTR-overlapping, intergenic gene 5'-proximal and intergenic gene 5'-distal), multiway intersections of all RefSeqFE features (A file) were performed with multiple B files containing the subsets (reciprocal 100% overlap conditions, `-f 1.0 -r` options) to prepare a summary file with gene-relative locations of each feature, and with collapse using the `bedtools groupby` command (Quinlan 2014) to display one line per feature for multiple classifications (Supplemental Table S3B,C).

For the unrestricted overlaps shown in Figure 5C,D, Supplemental Figure S5 and Supplemental Table S3A, intersections were performed at default conditions (no `-f` option) to determine all possible RefSeqFE feature (A file) intersections with collapsed genes, exons, introns, CDS, UTR, whole intergenic, gene 5'-proximal intergenic and gene 5'-proximal intergenic regions (B file). Intersecting RefSeqFE features were uniquified to determine unrestricted overlap feature counts. The `bedtools fisher` (https://bedtools.readthedocs.io/en/latest/content/tools/fisher.html) and `bedtools jaccard` (https://bedtools.readthedocs.io/en/latest/content/tools/jaccard.html) tools were used to determine two-tailed Fisher p-values and Jaccard statistics for each intersection, as shown in Supplemental Table S3A. Lengths of overlap (column 9 in `-wo` intersection output from BED4-formatted A and B files) were extracted for each intersecting feature, overlap lengths were summed when different segments of the same feature overlapped multiple features in the B file (e.g., a RefSeqFE feature that overlaps two exons), the length of the RefSeqFE feature was determined and the total overlap length was divided by the feature length to determine the degree of overlap. Violin plots (Fig. 5C,D; Supplemental Fig. S5) to display degree of overlap distributions were created using the BoxPlotR web tool (http://shiny.chemgrid.org/boxplotr/; Spitzer et al. 2014) for intersections at each gene-relative subregion, the cumulative results for all regions combined (intersections with each subregion carried out separately, then results combined) and for mined out features belonging to regulatory, recombination, protein_bind or other feature type classes. Minimum, maximum, average and standard deviation from the mean statistics were also calculated for degree of overlap data, as shown in Supplemental Table S3A.

Overlapping gene statistics

For statistics on overlapping genes (Fig. 5E), their biotypes were extracted from the B file name field (column 8 in intersection `-wo` output) following intersection of RefSeqFE features with gene features, then sorted, uniquified and counted with percentage determination. Additionally, RefSeqFE feature intersections were performed with B files containing subsets of genes from the RSG, LRG or ClinVar datasets (described above), and overlapping gene symbols were similarly extracted from B file name fields, uniquified and counted. Extracted gene symbols from those datasets were also combined and uniquified to determine the total number of clinically relevant genes that overlap RefSeqFE features. For all RefSeqFE features showing gene overlap, gene biotypes, symbols and clinically relevant dataset indications (if applicable) were added to a summary file with feature classifications, with further addition of the associated RefSeq accession and organism for each RefSeqFE feature (Supplemental Table S3B,C).

Regulatory interaction target gene statistics

Human target gene symbols were extracted from pairwise regulatory interactions represented in the `FEregintxns_AR109.20201120.inter.bb` file (https://ftp.ncbi.nih.gov/refseq/FunctionalElements/trackhub/data/human/AR109.20201120/), where target gene symbols and 'target_gene' tags are included in the name field for relevant interactions. The gene symbols were uniquified and counted to determine the total number of human target genes. The target gene list was also compared to RSG, LRG and ClinVar gene lists (described above) to determine status, counts and percentages in each clinically relevant gene dataset. The target gene list was additionally compared to a cumulative list of genes from all three clinically relevant datasets, where the target gene was scored as clinically relevant if found in at least one of those datasets (Supplemental Table S4, column E).

Comparative dataset analyses

Human AR 109.20201120 and mouse AR 109 features were converted to BED format as described for location determinations above. They were compared to gene regulatory features from the ENCODE cCRE (The ENCODE Project Consortium et al. 2020), Ensembl Regulation (Zerbino et al. 2016), FANTOM5 (Andersson et al. 2014), VISTA (Visel et al. 2007) and dbSUPER (Khan and Zhang, 2016) resources. Data were obtained and prepared from those resources as follows:

- ENCODE cCRE data: Human and mouse bigBed files were downloaded from https://hgdownload.soe.ucsc.edu/gbdb/hg38/encode3/ccre/ (`encodeCcreCombined.bb` file dated 2020-05-20) and https://hgdownload.soe.ucsc.edu/gbdb/mm10/encode3/ccre/ (`encodeCcreCombined.bb` file dated 2021-05-26). They were converted to non-binary format using the `bigBedToBed` UCSC utility, followed by conversion to 4-column BED format with capture of the ENCODE cCRE feature ID and feature type in the name field, and with chromosome notation converted from UCSC- to RefSeq-style based on mappings provided in columns 7 and 10 of the assembly report files for human GRCh38.p13 (accession GCF_000001405.39) and mouse GRCm38.p6 (accession GCF_000001635.26). The mouse BED file was further remapped to the GRCm39 assembly (accession GCF_000001635.27) using the NCBI Remap API (Kitts et al. 2016; https://www.ncbi.nlm.nih.gov/genome/tools/remap/docs/api).
- Ensembl Regulation data: Human and mouse GFF3 files were downloaded from http://ftp.ensembl.org/pub/release-104/regulation/homo_sapiens/ (`homo_sapiens.GRCh38.Regulatory_Build.regulatory_features.20210107.gff.gz` file) and http://ftp.ensembl.org/pub/release-104/regulation/mus_musculus/ (`mus_musculus.GRCm39.Regulatory_Build.regulatory_features.20201021.gff.gz` file). They were converted to 4-column BED format with capture of the Ensembl Regulation ID and feature type in the name field, and with chromosome notation converted from Ensembl- to RefSeq-style based on mappings provided in the assembly report files for human GRCh38.p13 and mouse GRCm39 (columns 1 and 7 for main chromosomes or 5 and 7 for unlocalized and unplaced scaffolds).

- FANTOM5 enhancers: Human and mouse BED files were downloaded from https://fantom.gsc.riken.jp/5/datafiles/reprocessed/hg38_latest/extra/enhancer/ (`F5.hg38.enhancers.bed.gz` file dated 2016-03-30) and https://fantom.gsc.riken.jp/5/datafiles/reprocessed/mm10_latest/extra/enhancer/ (`F5.mm10.enhancers.bed.gz` file dated 2018-09-07). They were converted to 4-column BED format with inclusion of the FANTOM5 ID (essentially the GRCh38/hg38 or GRCm38/mm10 genomic coordinates) in the name field, and with chromosome notation converted to RefSeq-style as described for ENCODE cCRE conversions above, also with remapping of the mouse BED file to the GRCm39 assembly.
- VISTA enhancers: All human and mouse enhancers (both positives and negatives) that were current on 2021-08-26 were retrieved and downloaded from https://enhancer.lbl.gov/cgi-bin/imagedb3.pl?form=ext_search&show=1 in FASTA format. Header lines were extracted and coordinate and identifier information within them was split to produce a 4-column BED format, with capture of the VISTA enhancer ID and its positive or negative status in the name field. Human enhancer features were remapped from GRCh37 (accession GCF_000001405.13) to GRCh38.p13 (accession GCF_000001405.39), and mouse features from MGSCv37 (accession GCF_000001635.18) to GRCm39 (accession GCF_000001635.27) using the NCBI Remap API described above. Chromosome notation was also converted from UCSC- to RefSeq-style as described for ENCODE cCRE features above.
- dbSUPER super-enhancers: Zipped BED files from 102 human and 25 mouse cell/tissue types were downloaded from http://asntech.org/dbsuper/download.php on 2019-07-19 (human `all_hg19_bed` file on 2019-07-19 and mouse `all_mm9_bed` file on 2019-08-16). For each extracted BED file representing a specific cell/tissue type, the cell/tissue name was concatenated with the dbSUPER ID in the name field, and then all features from all files were combined in a single 4-column BED file. Name fields for any super-enhancers with identical genomic coordinates were collapsed using the `bedtools groupby` command (Quinlan 2014; was applicable for human only), followed by remapping of the human file from GRCh37 to GRCh38.p13 and the mouse file from MGSCv37 to GRCm39 as described for VISTA enhancers above, also with conversion of the chromosome notation to RefSeq-style.

Feature counts were determined from each prepared BED file (Fig. 6A; Supplemental Table S5A), genome coverage was calculated for all collapsed features from each dataset (collapses via the `bedtools merge` command; Fig. 6A; Supplemental Table S5A), genome coverage percentages versus the GRCh38.p13 or GRCm39 reference assemblies were determined and plotted in Supplemental Figure

S6B, and feature length minimums, maximums, averages and standard deviations from the mean were calculated as described for RefSeqFE feature statistics above. The same statistics were also determined for all features from the ENCODE cCRE, Ensembl Regulation, FANTOM5, VISTA and dbSUPER datasets combined. All statistics are shown in Supplemental Table S5A. Additionally, boxplots displaying feature length distributions from each dataset and the RefSeqFE dataset were created as described for RefSeqFEs above (Fig. 6C,E; Supplemental Fig. S6A). Feature length distributions were also plotted for individual feature types from the ENCODE cCRE and Ensembl Regulation datasets (Supplemental Fig. S6C,D).

To determine overlaps between RefSeqFE features and gene regulatory features from the other datasets, intersections were performed with the `bedtools intersect` command (Quinlan 2014) at default conditions (no `-f` option) using RefSeqFE features in the A file and non-NCBI dataset features in the B file. All features in comparative datasets were intersected with either all RefSeqFE features, RefSeqFE regulatory class features only or RefSeqFE enhancer features only. Extracted enhancer features from each dataset were additionally intersected with RefSeqFE enhancer features, where just the positive subset of VISTA enhancers were used in that enhancer:enhancer intersection. Intersections were also carried out using a B file containing all features from the ENCODE cCRE, Ensembl Regulation, FANTOM5, VISTA and dbSUPER datasets combined. Intersecting RefSeqFE and comparative dataset features were uniquified to determine overlapping feature counts and percentage overlap with respect to each dataset, and all input, overlapping and percentage overlap data were documented in Supplemental Table S5B,C. Percentage overlaps with respect to RefSeqFE features were plotted in Figure 6B,D. Fisher p-values and Jaccard statistics were also determined (results shown in Supplemental Table S5B,C) for all intersections as described for genomic subregion intersections above.

For each pairwise overlap resulting from the intersection of all RefSeqFE features with all combined non-NCBI dataset features, the lengths of both the RefSeqFE and comparative dataset features were determined, and the overlap length (column 9 in `-wo` intersection output) was divided by each feature length to determine the degree of overlap with respect to each dataset feature. Supplemental Table S5D,E was prepared to display all pairwise overlaps along with the feature lengths and degrees of overlap with respect to each feature (see column descriptions for Supplemental Table S5D,E below). RefSeqFE features showing overlap with features in any of the comparative datasets were also uniquified with collapse of the overlapping dataset names (using the `bedtools groupby` command), where the collapsed list of datasets with an overlapping feature was added to the corresponding feature in Supplemental Table S3B,C, column G. In addition, Supplemental Table S5F was prepared to display

RefSeqFE features that lack overlap with features in the non-NCBI datasets (output from intersections performed with the −v option), where non-intersecting features from the human and mouse RefSeqFE datasets were combined and the organism was indicated. The same features also lack an overlapping dataset name in column G of Supplemental Table S3B,C.

Supplemental graphical displays information

To visualize the non-genic biological regions and features, multiple graphical displays were provided for standalone RefSeqs and their genome-annotated contexts. As is the case for all sequence records in NCBI's Nucleotide database, each standalone RefSeq can be viewed in graphical format (Supplemental Fig. S2) via a 'Graphics' link at the top of each flat file (Rangwala et al. 2021). In that display, annotated features are shown in separate configurable subtracks depending on their feature class (Supplemental Table S2A), and mouseover on any feature activates a pop-up box with functional and descriptive metadata, including links to GeneIDs, supporting publications, sequence downloads and BLAST options. In Gene database records, we provide an overview graphic showing the biological region genomic location in the context of neighboring loci (Supplemental Fig. S1, grey tab), in addition to a graphical view embed showing feature annotation on the reference genome assembly. Gene records also include a link to NCBI's Genome Data Viewer (GDV) (www.ncbi.nlm.nih.gov/genome/gdv/; Rangwala et al. 2021), which provides more complete genome browser options (Fig. 3A). For newly created biological regions that have not yet been annotated on the reference genome, the graphical view in Gene defaults to an 'NG_' RefSeq display, as described above (Supplemental Fig. S2).

For genome-annotated features, all functional features except parental biological region features were aggregated with color coding according to feature class, and were displayed in a 'Biological regions, aggregate' track for the indicated NCBI AR (Fig. 3A), again with mouseover-activated metadata boxes as described above. The same 'Biological regions, aggregate' track can also be viewed in NCBI's Variation Viewer (www.ncbi.nlm.nih.gov/variation/view/; NCBI Resource Coordinators, 2015). Since GDV and Variation Viewer are full genome browsers, other tracks of interest such as variation data, user-specific data from custom tracks, remotely connected files or track hubs can be configured for viewing alongside RefSeqFE annotations. Biological regions are also searchable in the NCBI genome browsers by their symbols, GeneIDs or genomic locations, where mouseover on an item in the search results table displays a metadata box, including the locus summary, nomenclature and a Gene database record link.

To expand the range of genome browsers our non-genic annotations can be viewed in and to graphically display the interaction data, we also created a RefSeq Functional Elements track hub, abbreviated

RefSeqFE Hub. It was created in UCSC track hub format (Raney et al. 2014) and serves as a gateway for data visualization, extraction, download and interoperability. It is hosted from the RefSeq FTP site (connection URL: https://ftp.ncbi.nlm.nih.gov/refseq/FunctionalElements/trackhub/hub.txt), registered in the Track Hub Registry (Aken et al. 2017) and is a Public Hub in the UCSC Genome browser. Compatible genome browsers include the UCSC Genome Browser (all tracks, metadata and display options; Fig. 3B), NCBI's GDV and Variation Viewer (biological region and feature tracks only, though features are best viewed in the existing 'Biological regions, aggregate' track described above, Fig. 3A), the Ensembl genome browser (Howe et al. 2021; biological region and feature tracks only), among others. The RefSeqFE Hub provides a hierarchical view of the parental biological regions versus underlying functional features, where parental biological regions are represented in a separate track (bigBed format) along with key metadata including locus symbols, descriptions, biological region summaries, interacting loci and associated RefSeq accessions. Color-coded underlying features are represented in a features track (bigBed format), with custom metadata including feature descriptions, experimental evidence types, publication support and more. The RefSeqFE Hub also includes separate tracks for the regulatory and recombination interactions in bigInteract format, which currently has limited genome browser compatibility but is fully viewable with configurable settings in the UCSC Genome Browser (Fig. 3B). The track hub's compatibility with multiple genome browsers enables users to take advantage of additional capabilities in specific genome browsers. For instance, NCBI's GDV provides options to link to BLAST or to extract specific feature sequences (Rangwala et al. 2021), while the UCSC Genome Browser provides options to download data using the Table Browser (Karolchik et al. 2004) or Data Integrator (Hinrichs et al. 2016), to export to the Galaxy platform via the Table Browser (Mangan et al. 2014), or to convert to JSON output via a RESTful API (Lee et al. 2020). Further details on the RefSeqFE Hub can be found in hub- or track-specific documentation and on our webpage (Supplemental Table S1, track hub links).

Supplemental Table descriptions

All Supplemental Tables are provided as separate Microsoft Excel files, some with multiple tabs as indicated in the descriptions below. For tables listing features with genomic coordinates (Supplemental Tables S3B,C and S5D,E,F), all coordinates are based on the human GRCh38.p13 or mouse GRCm39 reference assemblies, chromosome notation is in RefSeq format, all start positions are 0-based and all stop positions are 1-based.

Supplemental Table S1. Website links for RefSeqFE information and data access.

Supplemental Table S2. RefSeqFE feature types, statistics, definitions, annotation policies, publication-to-feature and biological region-to-feature metrics. Multiple tabs are provided as follows:

- **Supplemental Table S2A.** RefSeqFE feature types and statistics from human AR 109.20201120 and mouse AR 109. Additional details are provided in column-specific footnotes.
- **Supplemental Table S2B.** Current RefSeqFE feature types with SO IDs, INSDC feature keys and classes, definitions from INSDC or SO and RefSeq Functional Element annotation policies. Additional details are provided in column-specific footnotes.
- **Supplemental Table S2C**. Publications used for feature evidence in human AR 109.20201120 and mouse AR 109 combined (columns A-B), and the distribution of feature numbers derived from them (columns D-E and embedded chart).
- **Supplemental Table S2D.** Human AR 109.20201120 and mouse AR 109 biological regions with categorizations based on feature counts and/or large-scale study derivation. Large-scale studies are defined by the publications supporting >50 features in Supplemental Table S2C.

Supplemental Table S3. Data from RefSeqFE feature intersections with various genomic subregions, with summary data per feature including locations, overlapping gene and comparative dataset information, and associated RefSeq accessions. Multiple tabs are provided as follows:

- **Supplemental Table S3A.** Statistics for RefSeqFE feature overlaps with genomic subregions from human AR 109.20201120 and mouse AR 109. Additional details are provided in footnotes.
- **Supplemental Table S3B.** Human AR 109.20201120 RefSeqFE features (columns 1-4 in BED format), gene-relative locations, overlapping genes and datasets (if applicable), and associated RefSeq accessions. Gene-relative locations represent the called overlaps in Supplemental Table S3A and are charted in Figure 5A.
- **Supplemental Table S3C.** Mouse AR 109 RefSeqFE features (columns 1-4 in BED format), gene-relative locations, overlapping genes and datasets (if applicable), and associated RefSeq accessions. Gene-relative locations represent the called overlaps in Supplemental Table S3A and are charted in Figure 5B.

Supplemental Table S4. Human target genes represented in AR109.20201120 regulatory interactions and their occurrence in clinically relevant gene datasets. The ‘Clinically relevant’ column indicates ‘Yes’ when the target gene is present in at least one of the three clinically relevant gene datasets. Total ‘Yes’ counts are indicated at the end of each relevant column.

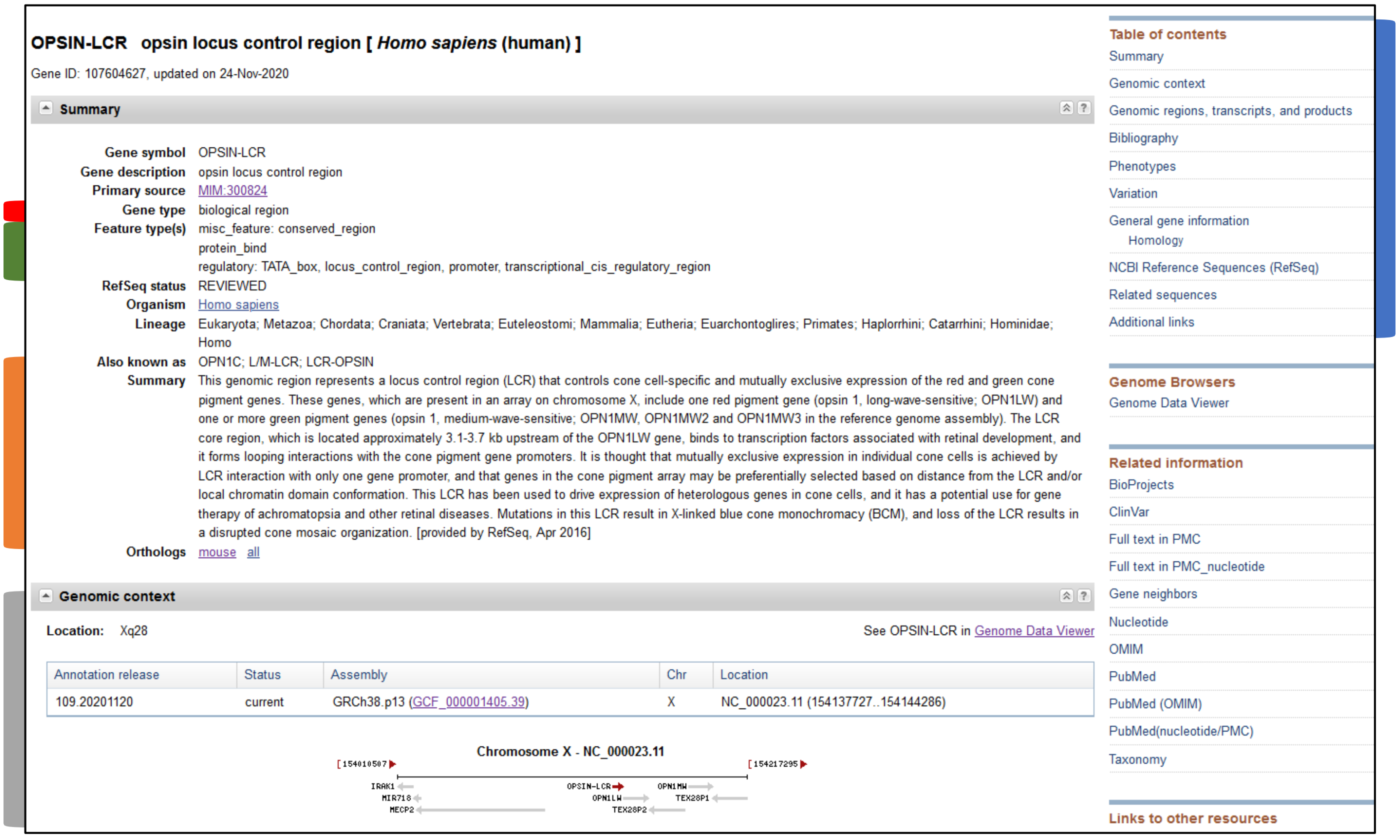
Supplemental Table S5. Statistics, content information and feature intersection output for RefSeqFEs compared to features in the ENCODE cCRE, Ensembl Regulation, FANTOM5 enhancer, VISTA enhancer and dbSUPER super-enhancer resources. Multiple tabs are provided as follows:

- **Supplemental Table S5A.** Statistics and general content information for RefSeqFEs and other comparative gene regulatory datasets. Additional details are provided in column-specific footnotes.
- **Supplemental Table S5B. S**tatistics for human AR 109.20201120 RefSeqFE intersections with other gene regulatory datasets. Additional details are provided in column-specific footnotes.
- **Supplemental Table S5C.** Statistics for mouse AR 109 RefSeqFE intersections with other gene regulatory datasets. Additional details are provided in column-specific footnotes.
- **Supplemental Table S5D.** Output for human AR 109.20201120 RefSeqFE intersections with other gene regulatory datasets, including feature and overlap lengths and degrees of overlap with respect to each dataset feature. Represented columns:
    A. **RefSeqFE chr** – the GRCh38.p13 chromosome/scaffold/contig the RefSeqFE feature is annotated on, shown in RefSeq notation.
    B. **RefSeqFE start** – the RefSeqFE feature start position (0-based).
    C. **RefSeqFE stop** – the RefSeqFE feature end position (1-based).
    D. **RefSeqFE feature** – the RefSeqFE feature name indicated by a concatenation of the feature SO_type and associated GeneID, with the bound moiety also indicated for protein binding site features.
    E. **Other dataset chr** – the GRCh38.p13 chromosome/scaffold/contig the comparative dataset feature is annotated on, shown in RefSeq notation.
    F. **Other dataset start** – the comparative dataset feature start position (0-based).
    G. **Other dataset stop** – the comparative dataset feature end position (1-based).
    H. **Other dataset feature** – the comparative dataset feature name including a concatenation of the dataset name, the feature type (if applicable) and IDs or other feature-specific information provided in download files from the particular resource.
    I. **Overlap length** – length of overlap between the RefSeqFE and comparative dataset features.
    J. **RefSeqFE feat_length** – length of the RefSeqFE feature.
    K. **Other feat_length** – length of the comparative dataset feature.
    L. **RefSeqFE overlap degree** – degree of overlap (overlap length/RefSeqFE feature length) for the RefSeqFE feature.

M. **Other overlap degree** – degree of overlap (overlap length/other dataset feature length) for the comparative dataset feature.

- **Supplemental Table S5E.** Output for mouse AR 109 RefSeqFE intersections with other gene regulatory datasets, including feature and overlap lengths and degrees of overlap with respect to each dataset feature. Columns A-M are as described for Supplemental Table S5D, where genome locations pertain to the mouse GRCm39 reference assembly.
- **Supplemental Table S5F.** Human AR 109.20201120 and mouse AR 109 RefSeqFE features that lack overlap with the other gene regulatory datasets. Features are provided in BED 4+1 format with the fifth column indicating the organism. Genome locations pertain to the human GRCh38.p13 or mouse GRCm39 reference assemblies.

Supplemental Figures



OPSIN-LCR **opsin locus control region [*Homo sapiens* (human)]**

Gene ID: 107604627, updated on 24-Nov-2020

Summary

Gene symbol	OPSIN-LCR
Gene description	opsin locus control region
Primary source	MIM:300824
Gene type	biological region
Feature type(s)	misc_feature: conserved_region protein_bind regulatory: TATA_box, locus_control_region, promoter, transcriptional_cis_regulatory_region
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as	OPN1C; L/M-LCR; LCR-OPSIN
Summary	This genomic region represents a locus control region (LCR) that controls cone cell-specific and mutually exclusive expression of the red and green cone pigment genes. These genes, which are present in an array on chromosome X, include one red pigment gene (opsin 1, long-wave-sensitive; OPN1LW) and one or more green pigment genes (opsin 1, medium-wave-sensitive; OPN1MW, OPN1MW2 and OPN1MW3 in the reference genome assembly). The LCR core region, which is located approximately 3.1-3.7 kb upstream of the OPN1LW gene, binds to transcription factors associated with retinal development, and it forms looping interactions with the cone pigment gene promoters. It is thought that mutually exclusive expression in individual cone cells is achieved by LCR interaction with only one gene promoter, and that genes in the cone pigment array may be preferentially selected based on distance from the LCR and/or local chromatin domain conformation. This LCR has been used to drive expression of heterologous genes in cone cells, and it has a potential use for gene therapy of achromatopsia and other retinal diseases. Mutations in this LCR result in X-linked blue cone monochromacy (BCM), and loss of the LCR results in a disrupted cone mosaic organization. [provided by RefSeq, Apr 2016]
Orthologs	mouse all

Genomic context

Location: Xq28

See OPSIN-LCR in Genome Data Viewer

Annotation release	Status	Assembly	Chr	Location
109.20201120	current	GRCh38.p13 (GCF_000001405.39)	X	NC_000023.11 (154137727..154144286)

Chromosome X - NC_000023.11

Table of contents

Genome Browsers

Genome Data Viewer

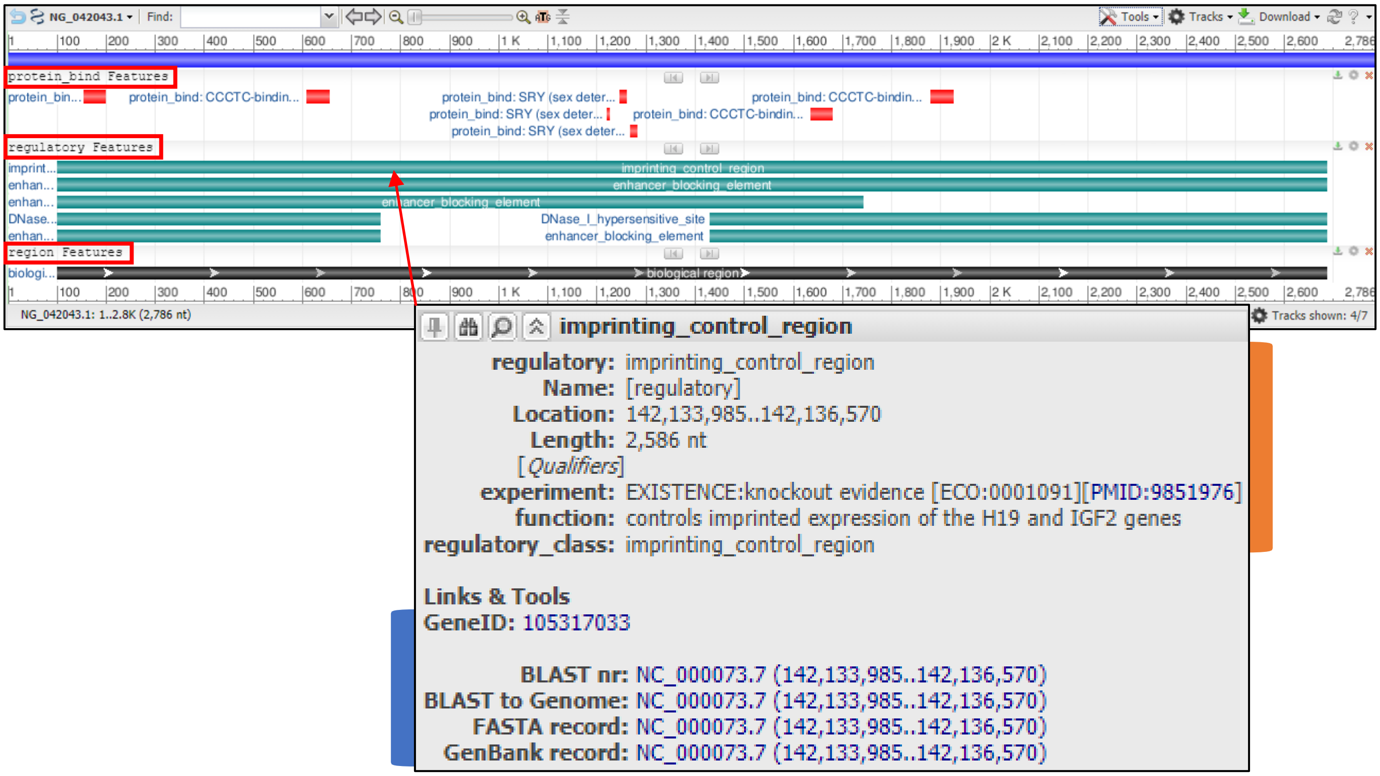
Related information

BioProjects
ClinVar
Full text in PMC
Full text in PMC_nucleotide
Gene neighbors
Nucleotide
OMIM
PubMed
PubMed (OMIM)
PubMed(nucleotide/PMC)
Taxonomy

Links to other resources

Supplemental Figure S1. Example of a biological region record in the Gene database. The image shows the top portion of the opsin locus control region record (*OPSIN-LCR*, GeneID:107604627). The 'biological region' locus type is indicated in the 'Gene type' field (red tab), and annotated feature types are listed in the 'Feature type(s)' field (green tab). The record includes biological region-specific nomenclature and a summary derived from published functional information (orange tab), genomic context information (grey tab), and several other sections as indicated in the 'Table of contents' (blue tab), including associated publications, sequence, variation and other related information when applicable.

RefSeq accession graphical view


NG_042043.1
Find:
Tools
Tracks
Download
protein_bind Features
protein_bin...
protein_bind: CCCTC-bindin...
protein_bind: SRY (sex deter...
protein_bind: CCCTC-bindin...
protein_bind: SRY (sex deter...
protein_bind: CCCTC-bindin...
protein_bind: SRY (sex deter...
regulatory Features
imprint...
enhan...
enhan...
DNase...
enhan...
imprinting_control_region
enhancer_blocking_element
enhancer_blocking_element
DNase_I_hypersensitive_site
enhancer_blocking_element
region Features
biologi...
biological region
NG_042043.1: 1..2.8K (2,786 nt)
Tracks shown: 4/7

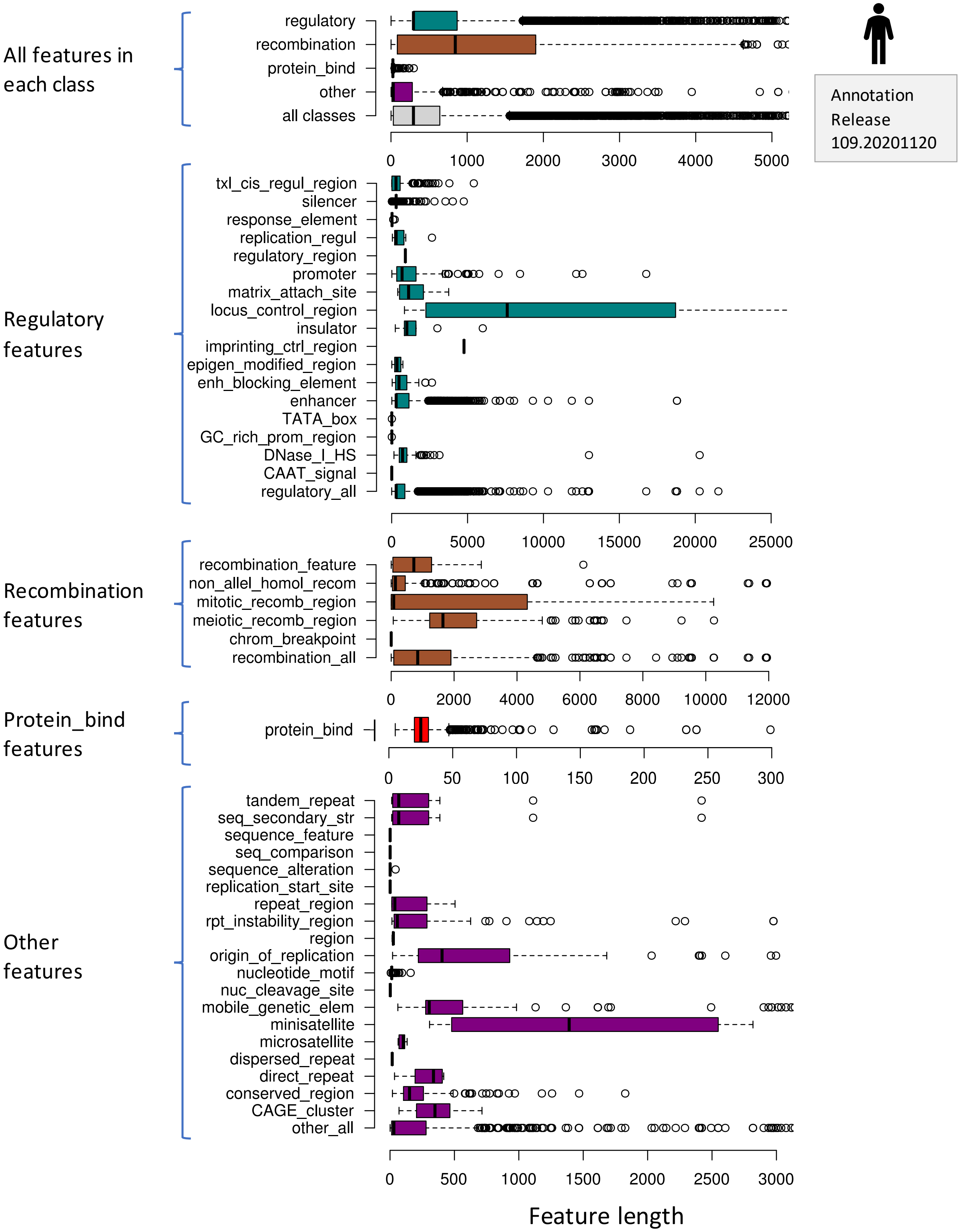
imprinting_control_region

regulatory: imprinting_control_region
Name: [regulatory]
Location: 142,133,985..142,136,570
Length: 2,586 nt
[*Qualifiers*]
experiment: EXISTENCE:knockout evidence [ECO:0001091][PMID:9851976]
function: controls imprinted expression of the H19 and IGF2 genes
regulatory_class: imprinting_control_region

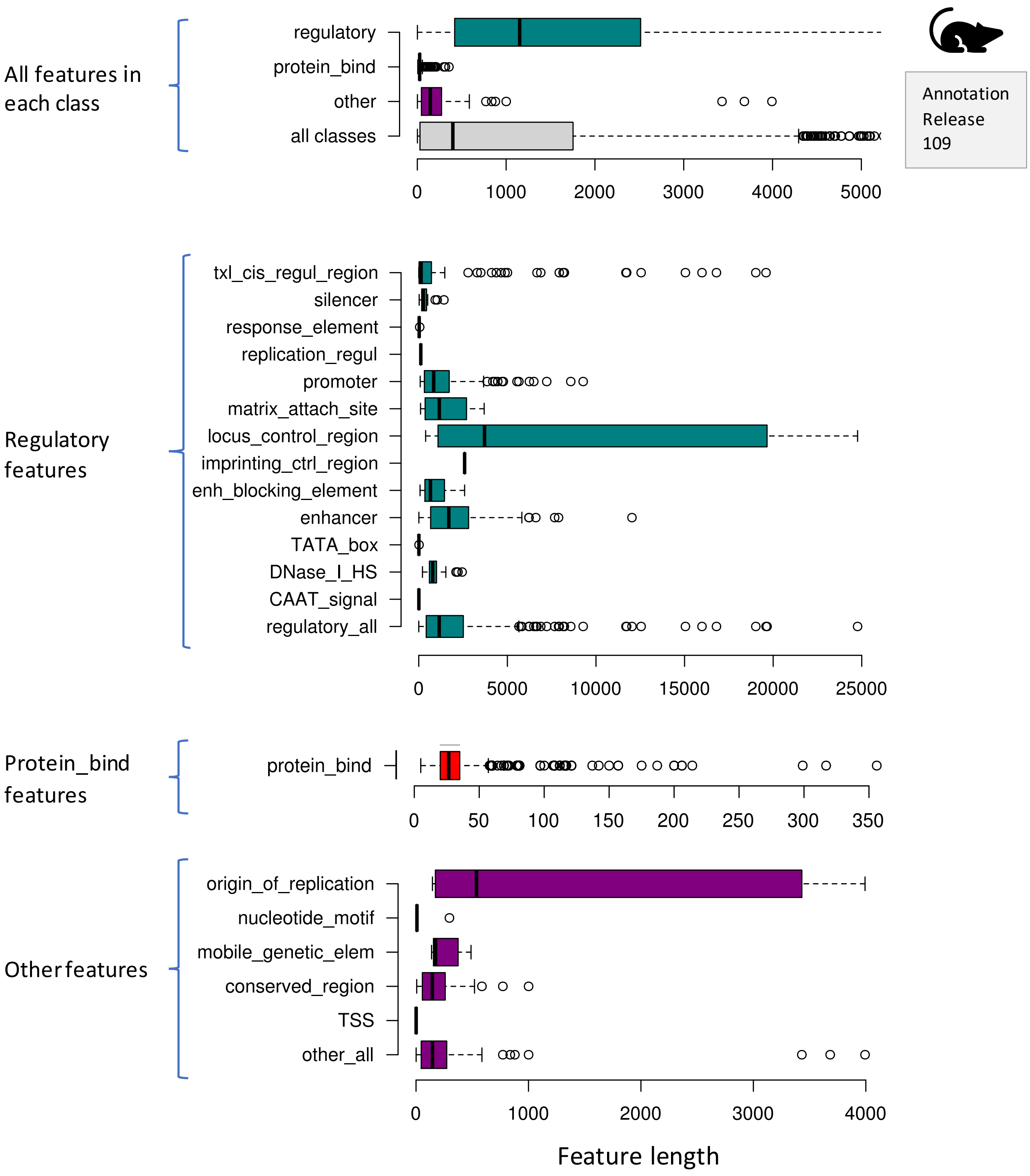
Links & Tools
GeneID: 105317033

BLAST nr: NC_000073.7 (142,133,985..142,136,570)
BLAST to Genome: NC_000073.7 (142,133,985..142,136,570)
FASTA record: NC_000073.7 (142,133,985..142,136,570)
GenBank record: NC_000073.7 (142,133,985..142,136,570)

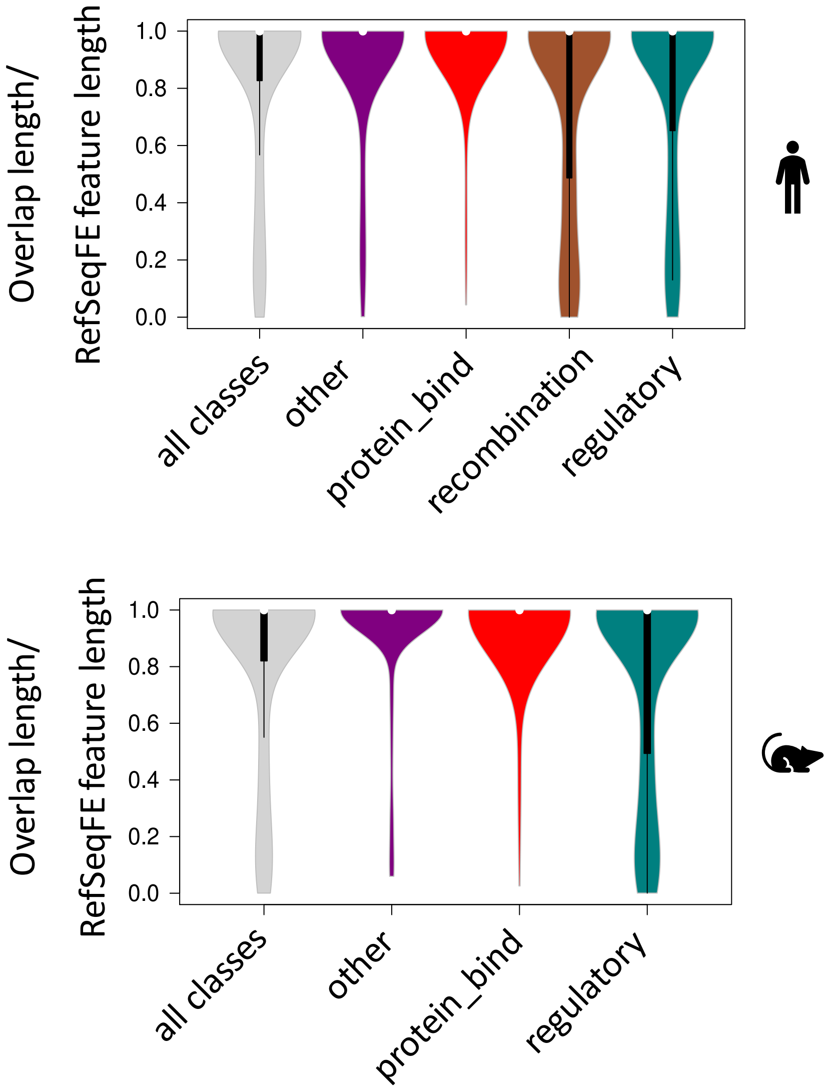
Supplemental Figure S2. RefSeq accession graphical display of NG_042043.1 representing the mouse *H19*/*Igf2* imprinting control region (*H19-icr*, GeneID:105317033). Annotated features are displayed in different tracks (outlined in red) according to feature class, here showing an imprinting control region, enhancer-blocking elements, DNase I hypersensitive sites and transcription factor binding sites. The parental ‘biological region’ feature that spans all underlying features is displayed in black in the ‘region Features’ track. Coordinates are based on positions within the RefSeq accession. An example of a mouseover-activated pop-up box is shown (overlaid grey box). These boxes contain descriptive and functional information (orange tab) including experimental evidence and links to publications, as well as a ‘Links & Tools’ area (blue tab) linking to the related Gene database record and to sequences and BLAST analyses.



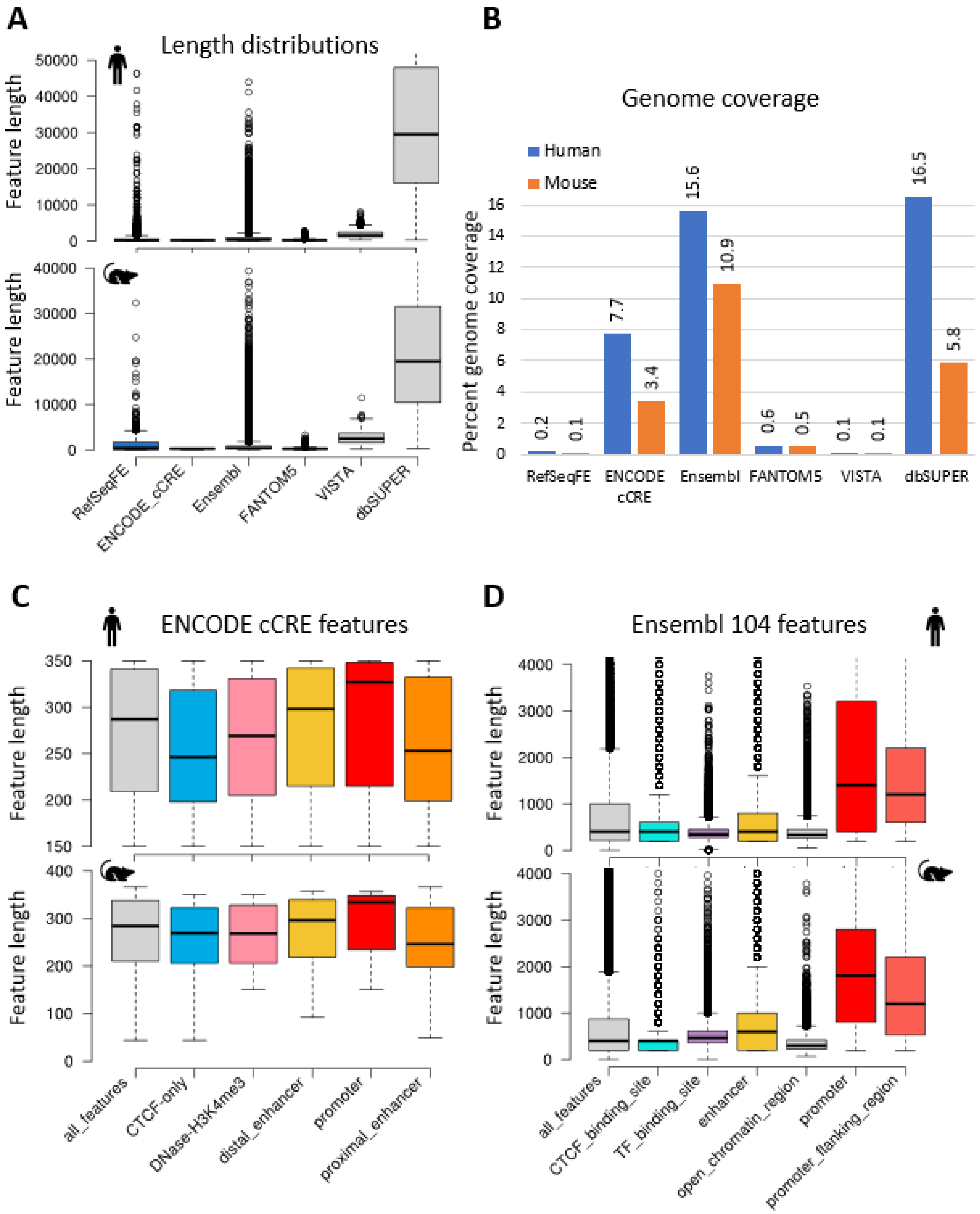
Supplemental Figure S3. RefSeqFE feature distributions for each feature type annotated in human AR 109.20201120. Horizontal boxplots showing feature length distributions for all human features, with coloring per feature class as used in Figure 4 and similar to that used in NCBI graphical displays. Center lines within boxes show the medians, box limits indicate the 25th and 75th percentiles, whiskers extend 1.5 times from the interquartile ranges, and outliers are represented by dots. Some outliers are not displayed because X-axis scales are variably customized to better visualize distributions of shorter features within each class. Some Y-axis feature labels are abbreviated to fit within the image. Full labels and supporting statistics (counts, minimums, maximums, averages and standard deviations from the mean) are provided in Supplemental Table S2A.



Supplemental Figure S4. RefSeqFE feature distributions for each feature type annotated in mouse AR 109. Horizontal boxplots showing feature length distributions for all mouse features, with coloring per feature class as used in Figure 4 and similar to that used in NCBI graphical displays. Center lines within boxes show the medians, box limits indicate the 25th and 75th percentiles, whiskers extend 1.5 times from the interquartile ranges, and outliers are represented by dots. Some outliers are not displayed because X-axis scales are variably customized to better visualize distributions of shorter features within each class. Some Y-axis feature labels are abbreviated to fit within the image. Full labels and supporting statistics (counts, minimums, maximums, averages and standard deviations from the mean) are provided in Supplemental Table S2A.



Supplemental Figure S5. Completeness of RefSeqFE feature overlaps at all gene-relative locations per feature class. (*A*) Violin plot showing degrees of human AR 109.20201120 feature overlaps (overlap length/RefSeqFE feature length) based on cumulative intersections with all gene-relative subregions shown in Figure 5 and Supplemental Table S3A (intersections with each subregion carried out separately, then results accumulated). Degree of overlap distributions are displayed for either all features (light grey distribution) or the indicated feature classes (purple, red, golden brown and teal distributions). White circles show the medians, box limits (where applicable) indicate the 25th and 75th percentiles, whiskers extend 1.5 times from the interquartile ranges, and curved shapes represent density estimates with extensions to extreme values. n = 25029, 3183, 2996, 2608 and 16242 sample points. Supporting statistics (Fisher p-values, Jaccard statistics, degree of overlap minimums, maximums, averages and standard deviations) are provided in Supplemental Table S3A. (*B*) Violin plot showing degrees of mouse AR 109 feature overlaps at all gene-relative locations per feature class as described for human in *A*. n = 5810, 255, 1503 and 4052 sample points. Additional statistics are provided in Supplemental Table S3A.



Supplemental Figure S6. Feature length distributions and genome coverage for RefSeqFEs and other gene regulatory datasets. (*A*) Boxplot showing feature length distributions for the indicated human and mouse datasets as shown in Figure 6C,E, here with the Y-axes scaled to view longer features in the dbSUPER dataset. Some outliers (maximum 498572 for human and 202016 for mouse, dbSUPER dataset) are not displayed. n = 9862, 926535, 622457, 63285, 1989 and 69340 sample points for the human panel and 2271, 343747, 364670, 49802, 1291 and 12103 sample points for the mouse panel. Additional statistics including minimums, maximums, averages and standard deviations from the mean are provided in Supplemental Table S5A. (*B*) Bar graph showing percent genome coverage of all collapsed features from each dataset relevant to the total lengths of the human GRCh38.p13 (blue bars) and mouse GRCm39 (orange bars) reference assemblies. RefSeqFE features are based on human AR 109.20201120 and mouse AR 109, while Ensembl Regulation features are based on Ensembl Release 104. The versions used and/or remapping of the other datasets on the relevant genome assembly are described in Supplemental Methods. The values for genome coverage (in bps) and percent genome coverage are provided in Supplemental Table S5A. (*C*) Boxplot showing feature length distributions for different feature types in the ENCODE cCRE dataset in human and mouse. Feature type coloring is based on that used by the ENCODE cCRE resource. n = 926535, 56766, 25537, 667599, 34803 and 141830 sample points for the human panel and 343747, 24079, 10537, 211189, 24115 and 73827 sample points for the mouse panel. (*D*) Boxplot showing feature length distributions for different feature types in the Ensembl Release 104 dataset in human and mouse. Feature type coloring is based on that used by the Ensembl Regulation resource. n = 622457, 175885, 30870, 127935, 110622, 36597 and 140548 sample points for the human panel and 364670, 110887, 17486, 70004, 61964, 25131 and 79198 sample points for the mouse panel. Some outliers (maximum 69402 for human and 78202 for mouse) are not displayed to better visualize the distributions of shorter features. The FANTOM5, VISTA and dbSUPER datasets have only one feature type (enhancer or super-enhancer), thus further feature breakdowns are not shown for those. Length distributions for individual feature types in the RefSeqFE dataset are shown in Supplemental Figures S3 and S4.

Supplemental References

Aken BL, Achuthan P, Akanni W, Amode MR, Bernsdorff F, Bhai J, Billis K, Carvalho-Silva D, Cummins C, Clapham P et al. 2017. Ensembl 2017. *Nucleic Acids Res* **45**: D635-D642.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**: 403-410.

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T et al. 2014. An atlas of active enhancers across human cell types and tissues. *Nature* **507**: 455-461.

The ENCODE Project Consortium, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, Adrian J, Kawli T, Davis CA, Dobin A et al. 2020. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583:** 699-710.

Ernst J, Melnikov A, Zhang X, Wang L, Rogov P, Mikkelsen TS, Kellis M. 2016. Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. *Nat Biotechnol* **34**: 1180-1190.

Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, Subramanian V, Grossman SR, Anyoha R, Doughty BR, Patwardhan TA et al. 2019. Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat Genet* **51**: 1664-1669.

Gasperini M, Hill AJ, McFaline-Figueroa JL, Martin B, Kim S, Zhang MD, Jackson D, Leith A, Schreiber J, Noble WS et al. 2019. A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell* **176**: 377-390 e319.

Haeussler M, Zweig AS, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, Lee CM, Lee BT, Hinrichs AS, Gonzalez JN et al. 2019. The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res* **47**: D853-D858.

Hinrichs AS, Raney BJ, Speir ML, Rhead B, Casper J, Karolchik D, Kuhn RM, Rosenbloom KR, Zweig AS, Haussler D et al. 2016. UCSC Data Integrator and Variant Annotation Integrator. *Bioinformatics* **32**: 1430-1432.

Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, Bhai J et al. 2021. Ensembl 2021. *Nucleic Acids Res* **49**: D884-D891.

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32**: D493-496.

Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. 2010. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**: 2204-2207.

Khan A, Zhang X. 2016. dbSUPER: a database of super-enhancers in mouse and human genome. *Nucleic Acids Res* **44:** D164-171.

Kheradpour P, Ernst J, Melnikov A, Rogov P, Wang L, Zhang X, Alston J, Mikkelsen TS, Kellis M. 2013. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res* **23**: 800-811.

Kitts PA, Church DM, Thibaud-Nissen F, Choi J, Hem V, Sapojnikov V, Smith RG, Tatusova T, Xiang C, Zherikov A et al. 2016. Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Res* **44**: D73-80.

Lee CM, Barber GP, Casper J, Clawson H, Diekhans M, Gonzalez JN, Hinrichs AS, Lee BT, Nassar LR, Powell CC et al. 2020. UCSC Genome Browser enters 20th year. *Nucleic Acids Res* **48**: D756-D761.

Mangan ME, Williams JM, Kuhn RM, Lathe WC, 3rd. 2014. The UCSC Genome Browser: What Every Molecular Biologist Should Know. *Curr Protoc Mol Biol* **107**: 19.9.1-19.9.36.

McGarvey KM, Goldfarb T, Cox E, Farrell CM, Gupta T, Joardar VS, Kodali VK, Murphy MR, O'Leary

NA, Pujar S et al. 2015. Mouse genome annotation by the RefSeq project. *Mamm Genome* **26**: 379-390.

Narlikar L, Sakabe NJ, Blanski AA, Arimura FE, Westlund JM, Nobrega MA, Ovcharenko I. 2010. Genome-wide discovery of human heart enhancers. *Genome Res* **20**: 381-392.

NCBI Resource Coordinators. 2015. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **43**: D6-17.

Petrykowska HM, Vockley CM, Elnitski L. 2008. Detection and characterization of silencers and enhancer-blockers in the greater CFTR locus. *Genome Res* **18**: 1238-1246.

Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, Farrell CM, Hart J, Landrum MJ, McGarvey KM et al. 2014. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res* **42**: D756-763.

Quinlan AR. 2014. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr Protoc Bioinformatics* **47**: 11.12.1-11.12.34.

Rajput B, Murphy TD, Pruitt KD. 2015. RefSeq curation and annotation of antizyme and antizyme inhibitor genes in vertebrates. *Nucleic Acids Res* **43**: 7270-7279.

Raney BJ, Dreszer TR, Barber GP, Clawson H, Fujita PA, Wang T, Nguyen N, Paten B, Zweig AS, Karolchik D et al. 2014. Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics* **30**: 1003-1005.

Rangwala SH, Kuznetsov A, Ananiev V, Asztalos A, Borodin E, Evgeniev V, Joukov V, Lotov V, Pannu R, Rudnev D et al. 2021. Accessing NCBI data using the NCBI Sequence Viewer and Genome Data Viewer (GDV). *Genome Res* **31**: 159-169.

Roh TY, Wei G, Farrell CM, Zhao K. 2007. Genome-wide prediction of conserved and nonconserved enhancers by histone acetylation patterns. *Genome Res* **17**: 74-81.

Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M et al. 2015. The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res* **43**: D670-681.

Sayers EW, Beck J, Brister JR, Bolton EE, Canese K, Comeau DC, Funk K, Ketter A, Kim S, Kimchi A et al. 2020. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **48**: D9-D16.

Spitzer M, Wildenhain J, Rappsilber J, Tyers M. 2014. BoxPlotR: a web tool for generation of box plots. *Nat Methods* **11:**121-122.

Visel A, Minovitsky S, Dubchak I, Pennacchio LA. 2007. VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic Acids Res* **35**: D88-92.

Wang H, Zhang Y, Cheng Y, Zhou Y, King DC, Taylor J, Chiaromonte F, Kasturi J, Petrykowska H, Gibb B et al. 2006. Experimental validation of predicted mammalian erythroid *cis*-regulatory modules. *Genome Res* **16**: 1480-1492.

Zerbino DR, Johnson N, Juetteman T, Sheppard D, Wilder SP, Lavidas I, Nuhn M, Perry E, Raffaillac-Desfosses Q, Sobral D et al. 2016. Ensembl regulation resources. Database (Oxford) 2016: bav11.