RefSeq Functional Elements as experimentally assayed nongenic reference standards and functional interactions in human and mouse

Catherine M Farrell, Tamara Goldfarb, Sanjida H Rangwala, et al.

Genome Res. published online December 7, 2021
Access the most recent version at doi:10.1101/gr.275819.121

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:

Catherine M. Farrell, Ph.D.
National Center for Biotechnology Information (NCBI)
National Library of Medicine
National Institutes of Health
Bethesda, MD 20894
USA
E-mail: farrelca@ncbi.nlm.nih.gov

Running title: NCBI RefSeq Functional Elements

Keywords: functional element; genome annotation; non-coding genome; non-genic; gene regulation; epigenomic; database; curation

ABSTRACT:

Eukaryotic genomes contain many non-genic elements that function in gene regulation, chromosome organization, recombination, repair or replication, and mutation of those elements can affect genome function and cause disease. While numerous epigenomic studies provide high coverage of gene regulatory regions, those data are not usually exposed in traditional genome annotation, and can be difficult to access and interpret without field-specific expertise. The National Center for Biotechnology Information (NCBI) therefore provides RefSeq Functional Elements (RefSeqFEs), which represent experimentally validated human and mouse non-genic elements derived from the literature. The curated dataset is comprised of richly annotated sequence records, descriptive records in the NCBI Gene database, reference genome feature annotation, and activity-based interactions between non-genic regions, target genes and each other. The dataset provides succinct functional details and transparent experimental evidence, leverages data from multiple experimental sources, is readily accessible and adaptable, and utilizes a flexible data model. The data have multiple uses for basic functional discovery, bioinformatics studies, genetic variant interpretation, as known positive controls for epigenomic data evaluation, and as reference standards for functional interactions. Comparisons to other gene regulatory datasets show that the RefSeqFE dataset includes a wider range of feature types representing more areas of biology, but it is comparatively smaller and subject to data selection biases. RefSeqFEs thus provide an alternative and complementary resource for experimentally assayed functional elements, with future dataset growth expected.

Eukaryotic genomes contain many types of functional elements, including conventional protein-coding and non-coding genes, gene regulatory elements, architectural elements, and elements associated with DNA replication, recombination and repair. Among those, conventional genes have received the most attention for representation in major genome annotation resources, e.g., RefSeq (O'Leary et al. 2016), GENCODE (Frankish et al. 2021) and others. Gene products, which include alternatively spliced transcripts and proteins, are abundantly represented in genome annotation databases with a heavy focus on protein-coding regions, which occupy less than 1.5% of the mammalian genome. Moreover, genes are major focal points for the curation of disease-associated genetic variation, where there is an emphasis on anchoring variation and linking human disease to specific genes (Wang et al. 2010; Vihinen et al. 2016; Xin et al. 2016; Rivera-Munoz et al. 2018; Amberger et al. 2019; Landrum et al. 2020). Aside from the obvious need to identify gene products due to their importance in biology, such a gene-centric focus is not surprising given that genes and gene-associated variation are generally more amenable for discovery and experimentation than non-genic functional elements, and they tend to offer more tangible avenues for therapeutic treatment of human disease.

The genome includes many non-genic elements that function in diverse biological processes, including gene regulation, chromosome organization, recombination or replication. Genome function can be adversely affected by mutation of those elements and result in disease (Lupianez et al. 2016; Chatterjee and Ahituv 2017; Perenthaler et al. 2019; Nesta et al. 2020), supported by genome-wide association studies (GWASs) showing that >90% of disease-associated variation occurs outside of coding regions (Ward and Kellis 2012; Gusev et al. 2014; Albert and Kruglyak 2015; Visscher et al. 2017; Gallagher and Chen-Plotkin 2018; Boix et al. 2021). While a lot of progress has been made in characterizing non-genic functional elements in specialist research fields, that information is not always adequately disseminated to other research fields, most notably to bioclinical research that relies on genome annotation for personal genomic or disease-associated variant interpretation (Perenthaler et al. 2019). Gene regulatory elements are the most abundantly studied among the non-genic element types and their epigenetic signatures are

indicated in several large-scale resources, including the Encyclopedia of DNA Elements (The ENCODE Project Consortium 2012), NIH Roadmap Epigenomics (Roadmap Epigenomics Consortium et al. 2015), International Human Epigenome Consortium (Stunnenberg et al. 2016), Ensembl Regulation (Zerbino et al. 2016) and EpiMap (Boix et al. 2021) projects, among others (Garda et al. 2021). However, those data are not usually exposed in traditional genome annotation, can be difficult to interpret, and have not been reconciled with region-specific experimental data in the literature. Thus, complexities in epigenomic data and its consumption disadvantages for non-field-specific experts indicate a need for more highly visible and easily accessible annotation of the non-coding genome. Furthermore, because genome function can only be truly elucidated by taking the entire genome into account, the current gene-centric focus of traditional genome annotation and variant curation points to a general need for better definitions of non-genic functional regions and an ethos to move beyond the genes.

To address this, NCBI created RefSeq Functional Elements (RefSeqFEs), a literature-derived dataset that provides reference genome annotation of experimentally validated and well-characterized non-genic regions in human and mouse. The dataset also links functional regions to target genes and to each other when there is activity-based support for functional interactions. Here we describe the creation of this freely available and readily accessible dataset, its multiple components, access options and uses. We also compare RefSeqFEs to other gene regulatory resources, and we report current dataset statistics with feature and genomic distribution analyses, which provide insights into current dataset content and offer suggestions for future needs.

Results

Dataset scope and design

To distinguish the non-genic dataset from RefSeq conventional genes, which include protein-coding genes, non-coding genes, pseudogenes and gene segments, we defined RefSeqFEs (https://www.ncbi.nlm.nih.gov/refseq/functionalelements; Supplemental Table S1) as any genomic

element with experimentally validated function, and which is not otherwise considered a conventional gene. For element types we included gene regulatory elements (e.g., enhancers, protein binding sites), known structural elements (e.g., boundary elements, chromatin conformation-associated regions), and other elements of functional importance (e.g., well-defined recombination hotspots or replication origins). While any experimentally validated non-genic element would fall in scope, including elements from high-throughput experimental studies, we prioritized genomic regions that are implicated in human disease or are otherwise of significant interest to the research community. Since we did not aim to replicate the numerous gene regulatory resources that already exist based on well-processed epigenomic or other multi-omics data (Garda et al. 2021), and because re-processing of available omics-derived data was not feasible for us at this time, we decided that RefSeqFEs, at least in the earlier stages of the project, would be focused on smaller-scale experimental data from the literature. Thus, it would be an alternative but complementary literature-derived resource with an emphasis on functional activity. That approach provides flexibility for representing a wide range of feature types in different areas of biology, fills a void to help reconcile other data resources with traditional experimental data in the literature, and allows for robust functional metadata provision such as direct links to publications. Consequently, the current dataset, which is focused on human and mouse, excludes elements from large-scale epigenomic studies and elements that exist solely based on disease-associated variation. It also excludes elements that have indefinite extents or are very large (tens of kilobase or greater lengths), such as telomeres, centromeres, topologically associating domains (TADs) and their broad boundaries, where those are less tractable for genome annotation.

For producing the data, we used the existing platforms and workflows already in place for the RefSeq transcript project to take full advantage of NCBI services, such as NCBI search engines, graphical displays and tools, full indexing and versioning of sequence records, and the ability to update records and genome annotation, including on new genome assemblies. Since all RefSeqs are incorporated in NCBI's Nucleotide database (Benson et al. 2018), RefSeqFE sequences and feature annotations adhere to data

standards defined by the International Nucleotide Sequence Database Collaboration (INSDC) (Karsch-Mizrachi et al. 2018) with robust use of those standards and ontologies, including recently introduced controlled vocabularies for 'regulatory_class' and 'recombination_class' features (Supplemental Table S2A,B), and feature qualifiers for metadata displays. Terms from the Sequence Ontology (SO) (Eilbeck et al. 2005) were additionally used to provide further specificity for features lacking a specific INSDC feature or class, and SO terms were also used to define genome-anchored features in GFF3- and bigBed-formatted download files.

The dataset was structured to include the following components: 1) Sequence records with curated underlying feature annotation, represented by genomic RefSeq accessions in NCBI's Nucleotide database; 2) Locus-level curated records to integrate metadata, graphical displays and sequences for the underlying region, represented as biological regions in NCBI's Gene database (Brown et al. 2015); 3) NCBI genome annotation on the human and mouse reference genome assemblies, represented as annotated features with concise and formatted metadata for download and display; and 4) Interaction data to link biological regions to target genes and each other, represented as pairwise interactions.

An overview of the RefSeqFE workflow is shown in Figure 1. Briefly, curation was based on experimental data from the literature, with bulk extraction from external databases for large-scale validated datasets, and with supplemental data from researchers if necessary. RefSeq and Gene database records were curated simultaneously, and those records were used as input for genome annotation by NCBI's Eukaryotic Annotation Pipeline (Pruitt et al. 2014; McGarvey et al. 2015; Supplemental Table S1, annotation pipeline links). Resulting FTP download files and graphical displays were produced for individual RefSeq sequences, Gene database records and genome-annotated features. Those data were further integrated and linked to NCBI-annotated genes and each other to produce additional FTP download files and displays, including a track hub. The following sections expand upon this workflow and provide more details on the components that make up the dataset, followed by analyses of the dataset and its contents.

Sequence records

Genomic RefSeq sequence records with 'NG_' accession prefixes were created to represent the range of one or more experimentally validated non-genic features. We grouped features that were closely located and functionally related in single RefSeqs, such as multiple adjacent or overlapping regulatory elements. The range of those grouped features was used to define a parental biological region (Supplemental Table S2B), as represented in Gene database records described below. To distinguish these non-genic RefSeqs from other genomic 'NG_' accessions represented by RefSeq, all RefSeqFE accessions are associated with NCBI BioProject accession PRJNA343958 (Barrett et al. 2012) and include the keyword RefSeqFE, as indicated in GenBank flat files (Fig. 2A).

Both manual curation and automated or semi-automated methods were used to create the RefSeqs. 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, then identifying specific well-characterized elements described therein and following links to citations. Additional in-scope elements were identified from 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), from searches for publications that employ bulk screening techniques, from specific experimental validation term searches in PubMed or PubMed Central, and through outreach efforts and user requests. The current dataset has some inevitable biases for experimentally validated elements that are easily findable (e.g., have been discussed in reviews or are associated with a biomedically important gene), for readily apparent evidence that is well-described and presented in main text of open access publications with a PubMed ID, and for data that are straightforward to curate directly from the publication. Additional details of the data selection and curation process are provided in Supplemental Material. All data were derived from evidence in the literature, either based on individual locus studies or on experimentally validated subsets from larger-scale studies. Examples of high-throughput evidence types used include but are not limited to clustered

regularly interspaced short palindromic repeats interference (CRISPRi) assays (e.g., Fulco et al. 2019; Gasperini et al. 2019), massively parallel reporter assays (MPRAs) (e.g., Kheradpour et al. 2013; Ernst et al. 2016), and reporter or transgenic assays from the VISTA project (Visel et al. 2007), FANTOM5 project (Andersson et al. 2014) and other bulk-screened datasets (e.g., Wang et al. 2006; Roh et al. 2007; Petrykowska et al. 2008; Narlikar et al. 2010). Those represent a sampling of the available evidence in scope for curation, where new datasets and many more focused region data are continually being identified and will be added to the RefSeqFE dataset over time.

We used a wide range of functional features to represent various element types, as indicated in the feature table on our webpage (www.ncbi.nlm.nih.gov/refseq/functionalelements/#Feature_table) and in Supplemental Table S2A,B. A parental biological region feature was annotated on all RefSeqs in addition to one or more underlying functional features. To standardize the curation process, we used the feature definitions provided by INSDC or SO and established policies for the annotation of each feature type (Supplemental Table S2B, columns E,F). In the vast majority of cases, the annotated feature range was defined by the exact fragment tested in an experimental assay, with a minority of features being based on ranges asserted by authors or by sequence analysis tools (see policies per feature type in Supplemental Table S2B, column F). Overlapping feature annotation was allowed, where each feature with a distinct range or type was treated as a unique entity and annotated separately, thus enabling the end user to see each feature as it was assayed in the linked publication(s). Feature type annotation was strictly based on the activity or characteristics primarily shown by the experimental evidence; for example, a protein binding assay and separate evidence that the bound protein functions in a regulatory activity would be represented by separate but overlapping protein binding and regulatory features, such as the *HBB-LCR* 5'HS5 CTCF binding site and overlapping enhancer-blocking element features shown in Figure 2B. Adhering to such guidelines enabled straightforward and consistent annotation decisions among curators. Experimental evidence was displayed in INSDC '/experiment' qualifiers on flat files (Fig. 2B, blue tabs), including an evidence type derived from the Evidence & Conclusion Ontology (ECO) (Giglio et al.

2019), followed by relevant publication evidence indicated by PubMed IDs. Other feature qualifiers included '/note' and '/function' for additional descriptive and functional information (e.g., cell type activity details), '/db_xref' to link to the associated Gene database record, and feature type-specific INSDC qualifiers. An example of GenBank flat file feature annotation with qualifier and ontology formatting is shown in Figure 2B.

Gene database records

Whereas the RefSeq records provide standalone annotated sequences with feature-specific metadata stored in various INSDC qualifiers, the Gene database serves as the central location for storing various types of metadata at the locus level, while also integrating sequence, genome annotation and graphical display data. Types of locus-level metadata include nomenclature, the locus type designation, a summary based on a synopsis of information from the literature, related publications, orthology information and other standard Gene database fields, as described previously for conventional genes (Brown et al. 2015).

To support the RefSeqFE project we created Gene database records identified by a new 'biological region' Gene type (Supplemental Fig. S1, red tab). We also added a new 'Feature type(s)' field to indicate the types of underlying features annotated on the associated RefSeq (Supplemental Fig. S1, green tab). Since the provision of official nomenclature by the HUGO Gene Nomenclature Committee (HGNC) (Bruford et al. 2020) or Mouse Genome Informatics (MGI) (Zhu et al. 2015) is generally out of scope for non-genic regions, official nomenclature was included for only a few biological regions that had pre-existing official nomenclature. Otherwise, all names, symbols and descriptions were based on curator derivation from the literature, with default symbols containing a '*LOC*' prefix appended with the integer GeneID assigned to the locus.

Genome annotation

All RefSeqFE features (Supplemental Table S3A,B) were annotated by NCBI's Eukaryotic Genome Annotation Pipeline together with NCBI's conventional gene-related features, initially in interim human

and mouse annotation releases (ARs) starting in 2017 up to the current ARs on the human GRCh38.p13 and mouse GRCm39 reference genome assemblies. Following genome annotation, genomic coordinates for annotated biological regions were propagated to relevant Gene records, both in text and graphical formats. Graphical displays of our genome annotation are described below (Fig. 3). Genome-anchored feature annotation was provided in both GFF3 (Moore et al. 2010) and bigBed (Kent et al. 2010) formats for FTP download (Supplemental Table S1, genome annotation data paths), with further details in the 'Accessing RefSeq Functional Elements data' section below.

Interaction data

An important aspect of our non-genic annotations is how these regions interplay with each other and with target genes. Therefore, during data curation we internally tracked regulatory element-to-target gene interactions and recombination partner pairings when there was sufficient experimental support in the literature. For regulatory interactions, our linkages were based on either direct experimental evidence for modulation of target gene promoter activity by methods such as reporter gene assays or transgenesis, or by genetic perturbation assays showing regulatory effects on target gene expression. We excluded linkages based on reporter assays that used heterologous promoters, and based on gene proximity predictions, which may only be accurate less than half of the time (Fulco et al. 2019). Thus, only a fifth of the biological regions are linked to target loci, but these linkages have been experimentally assayed and can be used as reference standards for activity-validated interactions. We also tracked regulatory interactions for biological regions that regulate each other, such as distal enhancer activation of a curated promoter region (e.g., the *CFTR* -44 kb enhancer and the *CFTR* promoter, *LOC111674478* and *LOC111674463*, respectively), or when regulatory elements from distinct biological regions have known cooperative activity (e.g., the *CFTR* -44 kb and +36.6 kb enhancers, *LOC111674478* and *LOC111674479*, respectively). Our recombination partner pairings were based on either experimental evidence for non-allelic homologous recombination (e.g., *LOC106804612* and *LOC106804613* representing alpha-globin recombination regions; Fig. 3B), or on direct assays showing translocations or other reproducible

recombination events on both sides of a breakpoint (e.g., the *LOC107980440* and *LOC107963955* major breakpoint regions involved in *BCR-ABL* translocations). Both the gene regulatory and recombination interactions were tracked at the parental biological region level, where they are relevant to at least one but not necessarily all underlying features within the biological region.

Following reference genome annotation, we determined genomic coordinates for relevant biological regions and target genes and then assembled the pairwise interactions in bigInteract format (Haeussler et al. 2019), also including a custom column listing supporting publications. These data are available for download on our FTP site (https://ftp.ncbi.nlm.nih.gov/refseq/FunctionalElements/trackhub/data/; Supplemental Table S1) and can also be visualized in regulatory interaction and recombination tracks in our track hub (Fig. 3B) described below.

Accessing RefSeq Functional Elements data

We provided a variety of data access options for different levels of our data, including for individual RefSeq and Gene database records, and for further processed genome annotation and interaction data. We employed FAIR (Findable, Accessible, Interoperable, Reusable) data principles (Wilkinson et al. 2016) to incorporate compatibility across multiple NCBI and non-NCBI tools and platforms. Our access options are summarized in Supplemental Table S1, where various links are provided for data downloads, sample queries and relevant help documentation. Options are available to access RefSeqFEs via NCBI's Gene database, Nucleotide database, BLAST searching, the BioProject database, NCBI graphical displays, the RefSeqFE Hub (see below), and the NCBI RefSeq, Gene and Genomes FTP sites. In addition, we periodically announce news about the dataset in the NCBI Insights blog (https://ncbiinsights.ncbi.nlm.nih.gov/tag/refseq-functional-elements/) and other NCBI social media.

To visualize the non-genic biological regions and features, multiple graphical displays were provided for standalone RefSeqs and their genome-annotated contexts (Fig. 3). Each standalone RefSeqFE record can be viewed in graphical format (Supplemental Fig. S2) via a 'Graphics' link at the top of each flat file

(Rangwala et al. 2021). Genome-annotated features are color coded according to feature class, and displayed in a 'Biological regions, aggregate' track for the indicated NCBI AR (Fig. 3A). The track can be viewed in NCBI graphical view embeds (e.g., in Gene records) and in NCBI's Genome Data Viewer (GDV) (www.ncbi.nlm.nih.gov/genome/gdv/; Rangwala et al. 2021), enabling the features to be viewed in the context of other data tracks such as variation data, user-uploaded data, remotely connected files or track hubs.

To expand the range of genome browsers RefSeqFE annotations can be viewed in and to graphically display the interaction data, we also created a RefSeqFE track hub (Fig. 3B; Supplemental Material). It is 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. It provides parental biological region and underlying feature tracks with custom metadata in bigBed format, and separate tracks for regulatory and recombination interactions in bigInteract format. Additional details on the RefSeqFE Hub and NCBI graphical displays are described in Supplemental Material and on our webpage.

While some of our access options are applicable for data querying and use at the biological region level, the ability to query and extract genome-annotated features is likely to be of higher interest. We therefore provided features for the entire set of NCBI-annotated features (including conventional genes) in GFF3 format with RefSeqFE features indicated in 'source' column 2, and for standalone RefSeqFE features in bigBed format, as described for the RefSeqFE Hub. Links to all our downloadable data can be found in Supplemental Table S1 and on our webpage, where we also provide feature and metadata extraction examples (www.ncbi.nlm.nih.gov/refseq/functionalelements/#Feat_extraction).

Current dataset statistics and content

To qualitatively and quantitatively assess the RefSeqFE dataset, we performed multiple analyses to assess the depth of curation used to produce the dataset, to determine the distribution of features and their genomic locations relative to conventional genes, to assess the relevance of the dataset to clinically relevant genes, and to compare the dataset to other available gene regulatory datasets.

To determine the depth of curation used to produce the dataset, we quantified the number of publications used for feature evidence and assessed the number of features derived from each publication (Supplemental Table S2C). In total, we used 2,219 distinct publications as evidence for human AR 109.20201120 and mouse AR 109 features combined. A broad set of publications were used as evidence for just a few features (e.g., 85% of publications were used for 1-4 features alone), while a small set of publications for large-scale studies contributed greater than 50 features each and were used as evidence for almost half of the features in the dataset, indicating that the dataset contains a good balance between large-scale and focused study evidence. We additionally assessed the biological regions with respect to single or multiple feature presence and according to study type derivation (Supplemental Table S2D). Approximately 21-23% of biological regions contained multiple features, while 70-73% of biological regions contained a single feature derived from a large-scale study. In summary, these analyses indicate that the dataset is deeply curated from diverse publications with a mix of large-scale and focused studies, and they attest to the high volume of laborious literature review used to create the dataset.

To assess the wide range of functional features represented in the dataset (Supplemental Table S2A,B), we determined genome coverage and feature distributions following human AR 109.20201120 (GRCh38.p13 assembly) and mouse AR 109 (GRCm39 assembly). In total, and not including parental biological region features, we annotated 9,862 features representing 4,450 distinct biological regions across 6.1 Mb in human, and 2,271 features representing 889 distinct biological regions across 2.2 Mb in mouse (Fig. 4E). The number of annotations per feature type and other feature statistics are shown in Supplemental Table S2A. To further summarize feature distributions, we grouped features into four main types: By INSDC regulatory class, recombination class (represented for human only), protein binding

sites and miscellaneous others, as charted in Figure 4A,C and indicated in Supplemental Table S2A. In both human and mouse 63-65% of features were regulatory class. Enhancers were by far the most common among regulatory class features, which may reflect a preference for performing enhancer assays in the literature, likely because their epigenetic signatures make them easier to identify. Protein binding sites were the second-most common feature type in the dataset, accounting for 14% of human and 30% of mouse features, indicating that protein binding assays are also popular in the literature, likely due to the molecular-level functional insights they provide. We noted an underrepresentation of silencer features (only 1-5% of regulatory class features) in the dataset, but we expect to increase silencer representation in the near future based on evidence from recent bulk screens (e.g., Huang et al. 2019; Doni Jayavelu et al. 2020; Pang and Snyder 2020).

We determined feature length distributions and other length-related statistics for all features combined, per feature class and for individual feature types (Fig. 4B,D; Supplemental Figs. S3 and S4; Supplemental Table S2A). The average length for all features was 781 bps for human and 1,125 bps for mouse, with recombination class features being generally the longest at 2,590 bps, and protein binding sites being generally the shortest at 29-35 bps (Fig. 4B,D). Feature length variability was also apparent between individual feature types within each feature class (Supplemental Figs. S3 and S4; Supplemental Table S2A), e.g., locus control regions were longer than other regulatory class features.

To assess the genomic distribution of RefSeqFE features relative to conventional genes and gene subregions, RefSeqFE features were first overlapped with annotated gene ranges, which include introns. We found that more than half (53-54%) of the features were gene range-overlapping in both human and mouse (Fig. 5A,B; Supplemental Table S3A). We further assessed the gene-overlapping features relative to gene subparts (exons, introns, CDS and UTR), and the intergenic features relative to gene 5'-proximal (2 kb upstream of transcript starts) or gene 5'-distal regions (Fig. 5A,B; Supplemental Table S3). For gene overlaps, 37% of all features overlapped introns and 16% overlapped exons in both human and mouse. The majority of exon-overlapping features were UTR-overlapping (12% and 15% of the human

and mouse datasets, respectively), while 4% of human and 1% of mouse features were CDS-overlapping, indicating that protein-coding regions may have non-coding biological functions too, a point that may impact genetic variant interpretation as described previously (Hirsch and Birnbaum 2015; Ahitiv 2016). Of the intergenic features, approximately two-thirds were gene-distal, corresponding to 33% of all human and 27% of all mouse features. For all genomic locations, feature overlap completeness was generally high (>75% of features showed >80% overlap with relevant genomic subregions overall; Fig. 5C,D; Supplemental Table S3A), especially with larger genomic segments (whole gene ranges, introns, intergenic regions) or for shorter feature classes (Supplemental Fig. S5), while shorter genomic segments (exons, CDS, UTR) tended to show more partial RefSeqFE feature overlaps.

Among the genes that overlapped RefSeqFE features, 64-69% were protein-coding, 28-34% were long non-coding RNA (lncRNA) genes, and 2-4% were other biotypes (Fig. 5E; Supplemental Table S3B,C). In total RefSeqFE features overlapped 2,455 and 565 distinct human and mouse genes, respectively. We also determined that 45% of the human overlapping genes were in at least one clinically relevant gene dataset (Supplemental Table S3B, column 6 square bracket indications), where 833 genes were represented in the RefSeqGene (RSG) dataset (Pruitt et al. 2014), 197 in the Locus Reference Genomic (LRG) dataset (Dalgleish et al. 2010), and 835 were genes used for pathogenic (or likely pathogenic) variant submissions to the ClinVar database (Landrum et al. 2020). Cumulatively, RefSeqFE features overlapped 13% of clinically relevant genes from those gene datasets combined, and further gene overlaps are expected upon future dataset growth. This indicates that alternative biological roles may be relevant when interpreting genetic variation in genes of clinical interest. We additionally quantified clinically relevant genes represented as target genes in RefSeqFE human regulatory interactions (Supplemental Table S4). Of 667 distinct target genes, 388 (58%) were represented in at least one of the RSG, LRG and ClinVar gene lists. This likely reflects a high focus on clinically relevant genes in the literature and/or our prioritization of clinical genes for regulatory annotation provision.

To further assess RefSeqFEs, we compared the dataset to other gene regulatory resources. Notwithstanding the availability of numerous gene regulatory resources (Garda et al. 2021), we selected just a sampling of those for comparison, namely ENCODE candidate *cis*-regulatory elements (cCREs; The ENCODE Project Consortium et al. 2020), Ensembl Regulation (Zerbino et al. 2016), FANTOM5 enhancers (Andersson et al. 2014), VISTA enhancers (Visel et al. 2007) and dbSUPER super-enhancer (Khan and Zhang, 2016). Compared to literature-derived RefSeqFEs, the other resources had different data derivation (Fig. 6A; Supplemental Table S5A), including from epigenomic signatures (ENCODE cCREs, Ensembl and dbSUPER), CAGE data (FANTOM5 enhancers) and transgenic assays (VISTA enhancers). Those resources represented only one or a few feature types (Fig. 6A; Supplemental Table S5A) compared to the over 40 feature types covering more areas of biology in the RefSeqFE resource (Supplemental Table S2A,B). RefSeqFE feature lengths were generally on a par with those from the other datasets (Fig. 6C,E; Supplemental Fig. S6A,C,D; Supplemental Table S5A) except for dbSUPER features, which were longer overall (Supplemental Fig. S6A). However, dataset size and genome coverage comparisons (Fig. 6A; Supplemental Fig. S6B; Supplemental Table S5A) show that the current RefSeqFE dataset is considerably smaller than the comparative datasets with the exception of VISTA enhancers, thereby indicating the major limitation of the dataset, as expected based on its literature-derived nature.

To determine feature-level similarity, we intersected RefSeqFE features with features in the other datasets, either individually with each dataset or with features from the comparative datasets combined (Fig. 6B,D; Supplemental Table S5B-F). When all RefSeqFE features were compared to all features in the other resources, approximately 80% of RefSeqFE features overlapped a feature(s) in at least one of the other datasets, with higher overlap percentages being apparent with the larger resources. As expected based on the considerably smaller RefSeqFE dataset size, with the exception of the more similarly sized VISTA dataset, these overlaps represented very low percentages of features in the comparative datasets (Supplemental Table S5B,C, columns E-G), indicating that a lot more content can be gleaned from those

large-scale datasets. Nevertheless, a fifth of RefSeqFE features did not show any overlap with the comparative datasets (Fig. 6B,D; Supplemental Table S5B,C,F) and a further 8-25% of pairwise overlaps were poor (<=10% of the RefSeqFE feature was overlapped, Supplemental Table S5D,E, column L), indicating that the dataset contains novel content not represented in the other resources. The non-overlapping features were distributed across all feature classes (30% regulatory, 19% recombination, 22% protein-binding and 29% other types in human and mouse combined; Supplemental Table S5F). A higher proportion of RefSeqFE regulatory or enhancer features overlapped features in the other datasets. (Fig. 6B,D). We noted better overlap with ENCODE cCRE enhancers than Ensembl enhancers, likely because ENCODE data were used to identify screening candidates in most of the large-scale studies used as evidence for RefSeqFE enhancers. Many RefSeqFE enhancers correlated with promoter flanking regions, CTCF binding sites and promoters that are abundantly represented in the Ensembl dataset (pairwise feature overlaps in Supplemental Table S5D,E), and indeed non-equivalent feature type overlaps existed with all the comparative datasets, likely due to differences in cell type activity, the versatility of gene regulatory elements, or dataset derivation and completeness differences. The enhancer-only comparisons also indicated high similarities with VISTA positive enhancers (Fig. 6B,D; Jaccard statistics in Supplemental Table S5B,C), as expected given that VISTA positive enhancers are incorporated in the RefSeqFE dataset and are a major source of RefSeqFE enhancers in mouse.

In summary, comparisons to other gene regulatory resources indicate that RefSeqFEs represent an alternative but smaller resource based on more traditional experimental evidence from the literature. The dataset offers a greater variety of nuanced feature types covering additional areas of biology, the features generally overlap well with features in comparative resources, and the dataset includes content not found in the other resources. Importantly, the currently smaller and more selective RefSeqFE resource should be considered complementary to other gene regulatory resources.

Discussion

We described here a new literature-derived dataset that provides annotation of experimentally assayed non-genic functional elements in human and mouse, and which uses a robust data model with rich but succinct metadata, and with accessibility options for a wide range of researchers. The dataset includes non-genic elements with diverse biological functions, ranging from gene regulatory elements, replication origins, genomic instability and recombination regions, to gene regulatory and recombination partner interactions. To our knowledge, this combination of functional element annotation is not available in other comparative non-genic data resources. The dataset is unique from a biocuration perspective, where we maximized use of INSDC feature types and qualifiers to format descriptive and functional information from hundreds of publications, with all formatting being accessible and extractable from both standalone RefSeqs and genome annotation. Our provision of RefSeq accessions for standalone use enables sequence findability through various NCBI avenues, including from the Nucleotide and Gene databases and by BLAST analysis. These are more consumable for focused genomic region studies without needing genome-scale extraction, e.g., for sequence determination for subsequent experimental assays, or for using with small-scale sequence analysis tools in the absence of high-performance computation.

Our integrative approach with respect to literature-derived data combines diverse experimental data types with a unified metadata structure, and it eliminates user need for exhaustive searching of the literature, or the need to remap data types between different genome assembly versions. Integrating different evidence types can also result in stronger evidence and better inform on function than individual evidence types alone. Although the literature-derived dataset is not inclusive of all available data sources and additional support can be gained from evidence in larger complementary resources, we have already observed strengthened functional support in some biological regions based on multiple evidence types. Examples include *LOC110121455*, *LOC112997545* and *LOC111501765*, where we were able to determine the element type based on reporter assay evidence and link to target genes based on CRISPRi evidence. Further such evidence type combinations are likely to yield more functional insights as the dataset grows.

The dataset has multiple uses, ranging from basic functional discovery, to genetic variant interpretation, to use as experimentally validated reference standards in multiple bioinformatic and epigenomic studies. Furthermore, the activity-supported interactions can be used as reference standards for gene regulatory or recombination interactions (also see discussion on their intended use below). Our multiple data accessibility options allow data usage through visual inspection on genome browsers (e.g., for region-specific comparisons to other datasets of interest), or through computational methods based on feature, sequence or metadata extraction, where we have incorporated compatibility with multiple tools and platforms. For basic research, the detailed experimental metadata can inform a researcher on experimental approaches for further in-depth characterization of features of interest. Both the feature annotations and target gene linking may be particularly useful for assigning function to clinically relevant genetic variants, and the experimentally validated features can be used as positive controls for assessing calls in various bioinformatic and epigenomic studies. We have already noted some applications of the dataset in diverse studies, including use of the RefSeqs as a source of locus control regions in a bioinformatics study (Sharma et al. 2019), use of the feature annotation for determining a DNase I hypersensitive site location and sequence in a focused research study (Uchida et al. 2019), and use of the mouse enhancer and promoter features to validate ChIP-seq calls in an epigenomic study (Roller et al. 2021). RefSeqFEs have recently become one of the gene regulatory data sources for the GeneHancer resource (Fishilevich et al. 2017). The biological region records have also been used in other resources such as the GeneCards database (Stelzer et al. 2016), and some variation resources link to the biological regions when there is variant overlap, including the Medical Genomics Japan Variant Database (MGeND) (Kamada et al. 2019) and ClinVar Miner (Henrie et al. 2018). NCBI's dbSNP database (Sherry et al. 2001) includes placements relative to RefSeqFE 'NG_' accessions for some SNP entries (e.g., rs11036238), while NCBI's ClinVar resource includes biological regions in the 'Gene(s)' tab for some variant records (e.g., Variation ID:96742). Some biological regions are also reported loci for ClinVar submissions (e.g., *LOC111365204*). Reciprocally, a link to overlapping ClinVar variants can be found in the 'Variation'

section for most human biological regions in the Gene database (Supplemental Table S1, biological regions with ClinVar variants link).

The RefSeqFE interactions may appear akin to interactions observed in 3D genomics studies, such as 3D-FISH or chromosome conformation capture-based assays (3C, Hi-C and similar derivatives; Kempfer and Pombo 2020). However, RefSeqFE interactions primarily provide basic element-to-target information as opposed to informing on higher order genome structure, and they are not intended to be comprehensive. They are derived either from genetic manipulation evidence for element-to-target activity or from genomic rearrangement characterization, as opposed to 3D genomics studies which assess physical contacts that are usually mapped at high density. Nevertheless, high-density 3D data can be difficult to interpret and visualize, usually requiring different visualization displays, data formats and specialized analysis tools, thus some users requiring basic element-to-target information may find the relatively simple RefSeqFE interactions easier to use, with the main limitation being the low numbers of interactions in the current dataset, notwithstanding future expected growth. As is the case for the RefSeqFE feature data, the interactions provide complementary data based on an alternative data model.

While the RefSeqFE dataset has accessibility, visibility and other advantages, it should be noted that this is a growing dataset where many regulatory elements from the literature still need to be curated, or many functional elements still need to be experimentally validated. This results in obvious disadvantages with respect to genome coverage, and consequently, the current dataset is less useful for researchers seeking comprehensive genome-wide data, for which we encourage the use of larger-scale complementary datasets. The current RefSeqFE dataset is more useful for seeking region-specific functional information (when present) or as an experimentally assayed subset for comparative evaluation of larger-scale data. Other limitations include the data selection biases indicated earlier, including selectivity for data that are easier to curate or automate, and our focus on regions that have been assayed in the literature. The literature itself may have limitations that affect data representation, such as absent, incomplete or inaccurate details in published methods. We expect all of the above limitations to decrease over time as

the dataset grows. Other limitations include caveats for some experimental evidence types used in the dataset (Catarino and Stark 2018; Perenthaler et al. 2019). For instance, *in vitro* experimental approaches may not always mimic *in vivo* conditions, including lack of an endogenous chromatin environment, use of heterologous promoters and absence of adjacent accessory sequences in reporter gene assays, lack of a chromatin context in direct protein binding assays, or ectopic genomic integrations resulting in altered chromatin landscapes in transgenic assays. Nevertheless, our representation of features based on those evidence types catalogs them on the genome, alerts researchers about their existence and could potentially prompt further in-depth characterization by other approaches. We also note that the majority of RefSeqFE features based on those evidence types were originally identified as screening candidates from epigenomic or other indicative data in supporting publications, or overlapping features based on alternative evidence types may be present, which boosts confidence in them. While we aim to convey as much functional information about each non-genic element as possible, we recommend that users critically assess experimental evidence and its context.

Future plans for RefSeqFEs include dataset growth and qualitative improvements based on research community needs. We aim for significant growth over the next several years, and are particularly interested in engaging with researchers who have data suitable for inclusion in the RefSeqFE dataset. Incorporation of improved and evolving high-throughput functional assays will contribute to dataset growth, including multiplex assays given their high-confidence nature, e.g., epigenetic or 3D genomics information combined with activity assays such as the ChIP-STARR-seq method (Barakat et al. 2018). We plan to increase representation of currently underrepresented feature types and to diversify the sources of high-throughput evidence in the dataset. We will also explore ways to incorporate high-value subsets of large-scale multi-omics data, for which we welcome research community input. We will continue to review our access options and make improvements where necessary, e.g., backfilling and improving cell/tissue type activity data, which is currently only accessible as free-text in feature qualifiers, by converting it to an extractable format. We will provide additional details on our webpage and periodically

announce any dataset improvements in the NCBI Insights blog. All community feedback is welcome either directly by e-mail, by using the 'Feedback' button on the RefSeqFE webpage or through the RefSeq user mail interface (https://www.ncbi.nlm.nih.gov/projects/RefSeq/update.cgi).

As the dataset continues to grow, we hope that our literature-derived annotations will provide further insights into how genes are regulated and how the genome functions, with a goal to inform on mechanisms of human disease. Now that we are in the exciting era of genomics in the 2020s, our dataset fulfills a timely need in moving traditional genome annotation beyond the genes, and in disseminating non-genic functional annotation to mainstream research in a more accessible format.

Methods

RefSeq Functional Elements dataset creation

An overview of the dataset and criteria used for data representation are described in the Results, on our webpage (www.ncbi.nlm.nih.gov/refseq/functionalelements/) and in Figure 1. Procedures to provide sequence records, Gene database records, genome annotation, interaction data and graphical displays are described in relevant sections of the Results and on our webpage. Further specific details on each are available in Supplemental Material.

Data analyses

Data analyses were based on 'RefSeqFE' source features extracted from GFF3 files for human AR 109.20201120 and mouse AR 109 (FTP download paths in Supplemental Table S1). Full-length gene, gene subpart ('exon', 'CDS') and 2 kb 5'-proximal features were also extracted from the same GFF3 files. Publication metrics were based on extraction of supporting PubMed IDs from bigBed feature files for the same ARs. Clinically relevant gene list sources are provided in Supplemental Table S1 and in Supplemental Material. Comparative datasets were obtained and processed as described in Supplemental Material. Standard UNIX command line methods were used together with the BEDTools software

package (Quinlan 2014) to extract and count features, to determine genome coverage and feature length statistics, to deduce intron, UTR and intergenic feature subsets, to determine publication-to-feature and biological region-to-feature metrics, to convert to BED format, to perform feature intersections, and to obtain statistics for overlapping genes, clinically relevant genes, comparative datasets and regulatory target genes. Further specific details on each are available in Supplemental Material.

Competing interest statement

The authors declare no competing interests.

Acknowledgements

This work was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine. We thank Drs. Donna Maglott and James Ostell for insights and support during the initiation of this project, Dr. Axel Visel and colleagues for VISTA enhancer discussions, Drs. Daniel Camerini-Otero, Florencia Pratto and Kevin Brick for meiotic recombination region discussions, the Sequence Ontology and INSDC teams for assisting in feature term provision and definitions, numerous NCBI teams and colleagues for contributing to NCBI tools and supporting workflows, and UCSC Genome Browser staff including Drs. Maximilian Haeussler and Brian Lee for track hub support. We are indebted to the countless research scientists who published the experimental evidence used to create the RefSeqFE dataset, and to those who contributed suggestions, supplementary details and data clarifications.

Author Contributions: The project was conceived, designed and managed by C.M.F. with assistance from T.D.M. and K.D.P. Data curation was performed by C.M.F., T.G and S.H.R. Database development and support were led by T.D.M. and carried out by A.A., O.D.E., V.H., K.S.K., V.K.K., F.L., T.D.M., K.D.P. and C.L.W. with input from C.M.F. Webpage documentation was provided by C.M.F. with assistance

from S.H.R. and T.D.M. Track hub preparation, associated download file provision, user outreach, all data analyses and manuscript writing were carried out by C.M.F. with assistance from T.D.M.

References

Ahituv N. 2016. Exonic enhancers: proceed with caution in exome and genome sequencing studies. *Genome Med* **8**: 14.

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.

Albert FW, Kruglyak L. 2015. The role of regulatory variation in complex traits and disease. *Nat Rev Genet* **16**: 197-212.

Amberger JS, Bocchini CA, Scott AF, Hamosh A. 2019. OMIM.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res* **47**: D1038-D1043.

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.

Barakat TS, Halbritter F, Zhang M, Rendeiro AF, Perenthaler E, Bock C, Chambers I. 2018. Functional Dissection of the Enhancer Repertoire in Human Embryonic Stem Cells. *Cell Stem Cell* **23:** 276-288.e8

Barrett T, Clark K, Gevorgyan R, Gorelenkov V, Gribov E, Karsch-Mizrachi I, Kimelman M, Pruitt KD, Resenchuk S, Tatusova T et al. 2012. BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Res* **40**: D57-63.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Ostell J, Pruitt KD, Sayers EW. 2018. GenBank. *Nucleic Acids Res* **46**: D41-D47.

Boix CA, James BT, Park YP, Meuleman W, Kellis M. 2021. Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature* **590**: 300-307.

Brown GR, Hem V, Katz KS, Ovetsky M, Wallin C, Ermolaeva O, Tolstoy I, Tatusova T, Pruitt KD, Maglott DR et al. 2015. Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res* **43**: D36-42.

Bruford EA, Braschi B, Denny P, Jones TEM, Seal RL, Tweedie S. 2020. Guidelines for human gene nomenclature. *Nat Genet* **52**: 754-758.

Bulger M, Schubeler D, Bender MA, Hamilton J, Farrell CM, Hardison RC, Groudine M. 2003. A complex chromatin landscape revealed by patterns of nuclease sensitivity and histone modification within the mouse beta-globin locus. *Mol Cell Biol* **23**: 5234-5244.

Catarino RR, Stark A. 2018. Assessing sufficiency and necessity of enhancer activities for gene expression and the mechanisms of transcription activation. *Genes Dev* **32**: 202-223.

Chan PK, Wai A, Philipsen S, Tan-Un KC. 2008. 5'HS5 of the human beta-globin locus control region is dispensable for the formation of the beta-globin active chromatin hub. *PLoS One* **3**: e2134.

Chatterjee S, Ahituv N. 2017. Gene Regulatory Elements, Major Drivers of Human Disease. *Annu Rev Genomics Hum Genet* **18**: 45-63.

Dalgleish R, Flicek P, Cunningham F, Astashyn A, Tully RE, Proctor G, Chen Y, McLaren WM, Larsson P, Vaughan BW et al. 2010. Locus Reference Genomic sequences: an improved basis for describing human DNA variants. *Genome Med* **2**: 24.

Dhar V, Nandi A, Schildkraut CL, Skoultchi AI. 1990. Erythroid-specific nuclease-hypersensitive sites flanking the human beta-globin domain. *Mol Cell Biol* **10**: 4324-4333.

Doni Jayavelu N, Jajodia A, Mishra A, Hawkins RD. 2020. Candidate silencer elements for the human and mouse genomes. *Nat Commun* **11**: 1061.

Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, Ashburner M. 2005. The Sequence

602 Ontology: a tool for the unification of genome annotations. *Genome Biol* **6**: R44.
603 The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human
604 genome. *Nature* **489**: 57-74.
605 The ENCODE Project Consortium, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, Adrian J,
606 Kawli T, Davis CA, Dobin A et al. 2020. Expanded encyclopaedias of DNA elements in the
607 human and mouse genomes. *Nature* **583:** 699-710.
608 Ernst J, Melnikov A, Zhang X, Wang L, Rogov P, Mikkelsen TS, Kellis M. 2016. Genome-scale high-
609 resolution mapping of activating and repressive nucleotides in regulatory regions. *Nat Biotechnol*
610 **34**: 1180-1190.
611 Farrell CM, West AG, Felsenfeld G. 2002. Conserved CTCF insulator elements flank the mouse and
612 human beta-globin loci. *Mol Cell Biol* **22**: 3820-3831.
613 Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, Rosen N, Kohn A, Twik M et al.
614 2017. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards.
615 *Database (Oxford)* **2017**: bax028.
616 Frankish A, Diekhans M, Jungreis I, Lagarde J, Loveland JE, Mudge JM, Sisu C, Wright JC, Armstrong
617 J, Barnes I et al. 2021. Gencode 2021. *Nucleic Acids Res* **49**: D916-D923.
618 Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, Subramanian V, Grossman SR, Anyoha R,
619 Doughty BR, Patwardhan TA et al. 2019. Activity-by-contact model of enhancer-promoter
620 regulation from thousands of CRISPR perturbations. *Nat Genet* **51**: 1664-1669.
621 Gallagher MD, Chen-Plotkin AS. 2018. The Post-GWAS Era: From Association to Function. *Am J Hum*
622 *Genet* **102**: 717-730.
623 Garda S, Schwarz JM, Schuelke M, Leser U, Seelow D. 2021. Public data sources for regulatory genomic
624 features. *Med Genet* **33:** 167-177.
625 Gasperini M, Hill AJ, McFaline-Figueroa JL, Martin B, Kim S, Zhang MD, Jackson D, Leith A,
626 Schreiber J, Noble WS et al. 2019. A Genome-wide Framework for Mapping Gene Regulation
627 via Cellular Genetic Screens. *Cell* **176**: 377-390 e319.
628 Giglio M, Tauber R, Nadendla S, Munro J, Olley D, Ball S, Mitraka E, Schriml LM, Gaudet P, Hobbs ET
629 et al. 2019. ECO, the Evidence & Conclusion Ontology: community standard for evidence
630 information. *Nucleic Acids Res* **47**: D1186-D1194.
631 Gusev A, Lee SH, Trynka G, Finucane H, Vilhjalmsson BJ, Xu H, Zang C, Ripke S, Bulik-Sullivan B,
632 Stahl E et al. 2014. Partitioning heritability of regulatory and cell-type-specific variants across 11
633 common diseases. *Am J Hum Genet* **95**: 535-552.
634 Haeussler M, Zweig AS, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, Lee CM, Lee BT, Hinrichs AS,
635 Gonzalez JN et al. 2019. The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res*
636 **47:** D853-D858.
637 Henrie A, Hemphill SE, Ruiz-Schultz N, Cushman B, DiStefano MT, Azzariti D, Harrison SM, Rehm
638 HL, Eilbeck K. 2018. ClinVar Miner: Demonstrating utility of a Web-based tool for viewing and
639 filtering ClinVar data. *Hum Mutat* **39**: 1051-1060.
640 Hirsch N, Birnbaum RY. 2015. Dual Function of DNA Sequences: Protein-Coding Sequences Function
641 as Transcriptional Enhancers. *Perspect Biol Med* **58**: 182-195.
642 Huang D, Petrykowska HM, Miller BF, Elnitski L, Ovcharenko I. 2019. Identification of human silencers
643 by correlating cross-tissue epigenetic profiles and gene expression. *Genome Res* **29**: 657-667.
644 Kamada M, Nakatsui M, Kojima R, Nohara S, Uchino E, Tanishima S, Sugiyama M, Kosaki K,
645 Tokunaga K, Mizokami M et al. 2019. MGeND: an integrated database for Japanese clinical and
646 genomic information. *Hum Genome Var* **6**: 53.
647 Karsch-Mizrachi I, Takagi T, Cochrane G, International Nucleotide Sequence Database C. 2018. The
648 international nucleotide sequence database collaboration. *Nucleic Acids Res* **46**: D48-D51.
649 Kempfer R, Pombo A. 2020. Methods for mapping 3D chromosome architecture. *Nat Rev Genet* **21**: 207-
650 226.
651 Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. 2010. BigWig and BigBed: enabling browsing
652 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.

Landrum MJ, Chitipiralla S, Brown GR, Chen C, Gu B, Hart J, Hoffman D, Jang W, Kaur K, Liu C et al. 2020. ClinVar: improvements to accessing data. *Nucleic Acids Res* **48**: D835-D844.

Long Q, Bengra C, Li C, Kutlar F, Tuan D. 1998. A long terminal repeat of the human endogenous retrovirus ERV-9 is located in the 5' boundary area of the human beta-globin locus control region. *Genomics* **54**: 542-555.

Lupianez DG, Spielmann M, Mundlos S. 2016. Breaking TADs: How Alterations of Chromatin Domains Result in Disease. *Trends Genet* **32**: 225-237.

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.

Moore B, Fan G, Eilbeck K. 2010. SOBA: sequence ontology bioinformatics analysis. *Nucleic Acids Res* **38**: W161-164.

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.

Nesta AV, Tafur D, Beck CR. 2020. Hotspots of Human Mutation. *Trends Genet* doi: 10.1016/j.tig.2020.10.003.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**: D733-745.

Pang B, Snyder MP. 2020. Systematic identification of silencers in human cells. *Nat Genet* **52**: 254-263.

Perenthaler E, Yousefi S, Niggl E, Barakat TS. 2019. Beyond the Exome: The Non-coding Genome and Enhancers in Neurodevelopmental Disorders and Malformations of Cortical Development. *Front Cell Neurosci* **13**: 352.

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.

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.

Rivera-Munoz EA, Milko LV, Harrison SM, Azzariti DR, Kurtz CL, Lee K, Mester JL, Weaver MA, Currey E, Craigen W et al. 2018. ClinGen Variant Curation Expert Panel experiences and standardized processes for disease and gene-level specification of the ACMG/AMP guidelines for sequence variant interpretation. *Hum Mutat* **39**: 1614-1622.

Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518**: 317-330.

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.

Roller M, Stamper E, Villar D, Izuogu O, Martin F, Redmond AM, Ramachanderan R, Harewood L,

Odom DT, Flicek P. 2021. LINE retrotransposons characterize mammalian tissue-specific and evolutionarily dynamic regulatory regions. *Genome Biol* **22**: 62.

Sharma BS, Swain PK, Verma RJ. 2019. A Systematic Bioinformatics Approach to Motif-Based Analysis of Human Locus Control Regions. *J Comput Biol* **26**: 1427-1437.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**: 308-311.

Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, Stein TI, Nudel R, Lieder I, Mazor Y et al. 2016. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Curr Protoc Bioinformatics* **54**: 1.30.1-1.30.33.

Stunnenberg HG, International Human Epigenome Consortium, Hirst M. 2016. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell* **167**: 1145-1149.

Tuan D, Solomon W, Li Q, London IM. 1985. The "beta-like-globin" gene domain in human erythroid cells. *Proc Natl Acad Sci U S A* **82**: 6384-6388.

Uchida N, Hsieh MM, Raines L, Haro-Mora JJ, Demirci S, Bonifacino AC, Krouse AE, Metzger ME, Donahue RE, Tisdale JF. 2019. Development of a forward-oriented therapeutic lentiviral vector for hemoglobin disorders. *Nat Commun* **10**: 4479.

Vihinen M, Hancock JM, Maglott DR, Landrum MJ, Schaafsma GC, Taschner P. 2016. Human Variome Project Quality Assessment Criteria for Variation Databases. *Hum Mutat* **37**: 549-558.

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.

Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 2017. 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* **101**: 5-22.

Wai AW, Gillemans N, Raguz-Bolognesi S, Pruzina S, Zafarana G, Meijer D, Philipsen S, Grosveld F. 2003. HS5 of the human beta-globin locus control region: a developmental stage-specific border in erythroid cells. *EMBO J* **22**: 4489-4500.

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.

Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**: e164.

Ward LD, Kellis M. 2012. Interpreting noncoding genetic variation in complex traits and human disease. *Nat Biotechnol* **30**: 1095-1106.

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**: 160018.

Xin J, Mark A, Afrasiabi C, Tsueng G, Juchler M, Gopal N, Stupp GS, Putman TE, Ainscough BJ, Griffith OL et al. 2016. High-performance web services for querying gene and variant annotation. *Genome Biol* **17**: 91.

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:** bav119.

Zhu Y, Richardson JE, Hale P, Baldarelli RM, Reed DJ, Recla JM, Sinclair R, Reddy TB, Bult CJ. 2015. A unified gene catalog for the laboratory mouse reference genome. *Mamm Genome* **26**: 295-304.

Figure legends

Figure 1. Workflow for RefSeqFE dataset production. Full cylinders represent databases, the half-cylinder represents the indicated data source, and rectangles represent actions. Relevant links to additional information and data access are provided in Supplemental Table S1.

Figure 2. Example of a biological region RefSeqFE flat file. Segments of RefSeq accession NG_052895.1 representing the beta-globin locus control region (*HBB-LCR*) are shown. (*A*) Top section of the flat file with a link to BioProject accession PRJNA343958 and the 'RefSeqFE' keyword outlined in red. (*B*) Segment of the feature annotation section. Features are displayed for the 5'HS5 DNase I hypersensitive site (Tuan et al. 1985; Dhar et al. 1990; Wai et al. 2003), a transcriptional *cis*-regulatory region (Long et al. 1998), a CTCF binding site (Farrell et al. 2002; Bulger et al. 2003; Chan et al. 2008) and an enhancer-blocking element (Farrell et al. 2002). Features include '/experiment' qualifiers with experimental evidence from the literature as indicated by ECO strings and IDs and links to publications (blue tabs), '/note' qualifiers with descriptive information (grey tabs), '/function' qualifiers describing the function of each feature where applicable (green tabs), and a '/bound_moiety' qualifier for the protein binding site (red tab). All features include a '/db_xref' qualifier (black tabs) linking to the biological region record in the Gene database (GeneID:109580095), and an INSDC class qualifier when relevant (orange tabs).

Figure 3. Graphical displays of RefSeqFE data. (*A*) NCBI Genome Data Viewer display of genome-annotated features at the human opsin locus control region (*OPSIN-LCR*, GeneID:107604627 also shown in Supplemental Fig. S1). Underlying features are aggregated and displayed in the 'Biological regions, aggregate' track (outlined in red). Depending on user track set options or the entry point to GDV, the track may need to be turned on via the configuration interface, as detailed on our webpage (Supplemental Table S1, graphical displays link). Features are color-coded according to class or type. Coordinates are based on positions on the genome sequence. 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. (*B*) RefSeqFE Hub view of parental biological regions, underlying features, and gene regulatory and recombination partner interactions in the UCSC Genome Browser. Regulatory interactions are shown between the alpha-globin locus control region (*HBA-LCR*, GeneID:106144573) and the downstream *HBZ*, *HBA2*, *HBA2* and *HBQ1* genes (blue curved lines), while the recombination partners track visualizes recombination (green curved line)

between two alpha-globin recombination regions (*LOC106804612* and *LOC106804613*). Parental biological regions are denoted by black rectangles in the biological regions track, while the features track uses color coding as described for *A*. Further item-specific metadata, display options and links to related data and tools can be found within item- and track-specific details pages. Depending on the density of interactions in a region, appropriate zoom levels or configuration modes may need to be adjusted, or specific hub settings such as multi-region view can be used for viewing interactions between distally located regions.

Figure 4. RefSeqFE feature distributions. (*A*) Categorized feature counts from human AR 109.20201120 on the GRCh38.p13 genome assembly with grouping by feature class. The pale blue labels indicate the feature counts per category, where categories and a full breakdown of feature types and counts are available in Supplemental Table S2A. (*B*) Boxplot showing feature length distributions for all human features (light grey) and individual feature classes, with coloring as in *A*. Some outliers (maximum length 141940) are not displayed because the Y-axis was scaled to better visualize the distributions of shorter features. Length distributions per feature class are provided in Supplemental Figure S3 with customized scaling for each class. n = 9862, 1357, 1379, 926 and 6200 sample points. Additional statistics including minimums, maximums, averages and standard deviations from the mean are provided in Supplemental Table S2A. (*C*) Categorized feature counts from mouse AR 109 on the GRCm39 genome assembly as shown for human in *A*. (*D*) Boxplot showing feature length distributions for all mouse features (light grey) and individual feature classes, as described for human in *B*. n = 2271, 109, 690 and 1472 sample points. Additional details are provided in Supplemental Figure S4 and Supplemental Table S2A. (*E*) Summary table with overall counts of annotated features, biological region loci and genome coverage for the indicated AR.

Figure 5. Locations of RefSeqFE features relative to genes. (*A*) Locations of features from human AR 109.20201120 compared to NCBI annotated genes and subparts from the same AR. The horizontal bar graph shows the overall locations (gene-overlapping or intergenic), while the bar-of-pie chart shows more detailed locations. Blue tones denote genes and subparts while grey tones denote intergenic regions. The pale blue labels indicate overlapping feature counts for each location, as shown for called overlaps in Supplemental Table S3A. (*B*) Locations of features from mouse AR 109 as shown for human in *A*. (*C*) Violin plot showing completeness of human RefSeqFE feature overlaps (overlap length/RefSeqFE feature length) at each gene-relative location (blue- and grey-tone coloring as in *A*) and cumulative results for all locations (blue-grey distribution at left). n = 25029, 5468, 2084, 4373, 743, 1735, 5235, 1906 and 3485

sample points. Supporting statistics (Fisher p-values, Jaccard statistics, degree of overlap minimums, maximums, averages and standard deviations) are provided in Supplemental Table S3A. (*D*) Violin plot showing completeness of mouse feature overlaps at each gene-relative location as described for human in *C*. n = 5810, 1249, 502, 981, 97, 459, 1237, 578 and 707 sample points. Supporting statistics are provided in Supplemental Table S3A. (*E*) Biotype statistics for genes that are overlapped by RefSeqFE features. The count columns indicate the number of distinct genes overlapped by one or more features, while the % total columns indicate percentages of the total number of genes (2455 human, 565 mouse) overlapped by RefSeqFE features for each biotype.

Figure 6. Comparison of RefSeqFEs to other gene regulatory datasets. (*A*) Overview showing data derivation, feature type representation and current sizes of each dataset on the human GRCh38.p13 and mouse GRCm39 reference assemblies. Additional information for each dataset is provided in Supplemental Table S5A. (*B*) Bar graph showing human AR 109.20201120 RefSeqFE feature intersections with the indicated datasets, where the Y-axis represents the percent of input RefSeqFE features showing overlap. All features in comparative datasets were intersected with either all RefSeqFE features (medium blue bars), RefSeqFE regulatory features (grey bars) or RefSeqFE enhancer features (light blue bars). Enhancer features from each dataset were additionally intersected with RefSeqFE enhancer features (dark blue bars). Full statistics including input and overlapping feature counts, overlap percentages with respect to each dataset, Fisher p-values and Jaccard statistics are provided in Supplemental Table S5B, with raw intersection output, feature lengths and degrees of overlap with respect to each dataset in Supplemental Table S5D. Datasets showing overlap with each RefSeqFE feature are also indicated in Supplemental Table S3B, column G. (*C*) Boxplot showing feature length distributions for the indicated human datasets. Some outliers and dbSUPER feature lengths (maximum 498572) are not displayed because the Y-axis was scaled to better visualize shorter feature distributions; see Supplemental Figure S6A for a 50 kb Y-axis scale with dbSUPER data included. n = 9862, 926535, 622457, 63285 and 1989 sample points. Additional statistics including minimums, maximums, averages and standard deviations from the mean are provided in Supplemental Table S5A. (*D*) Bar graph showing mouse AR 109 RefSeqFE feature intersections with the indicated datasets, as described for human in *B*. Supporting details are provided in Supplemental Tables S3C and S5C,E. (*E*) Boxplot showing feature length distributions for the indicated mouse datasets, as described for human in *C*. n = 2271, 343747, 364670, 49802 and 1291 sample points. Supporting details are provided in Supplemental Table S5A and Supplemental Figure S6.

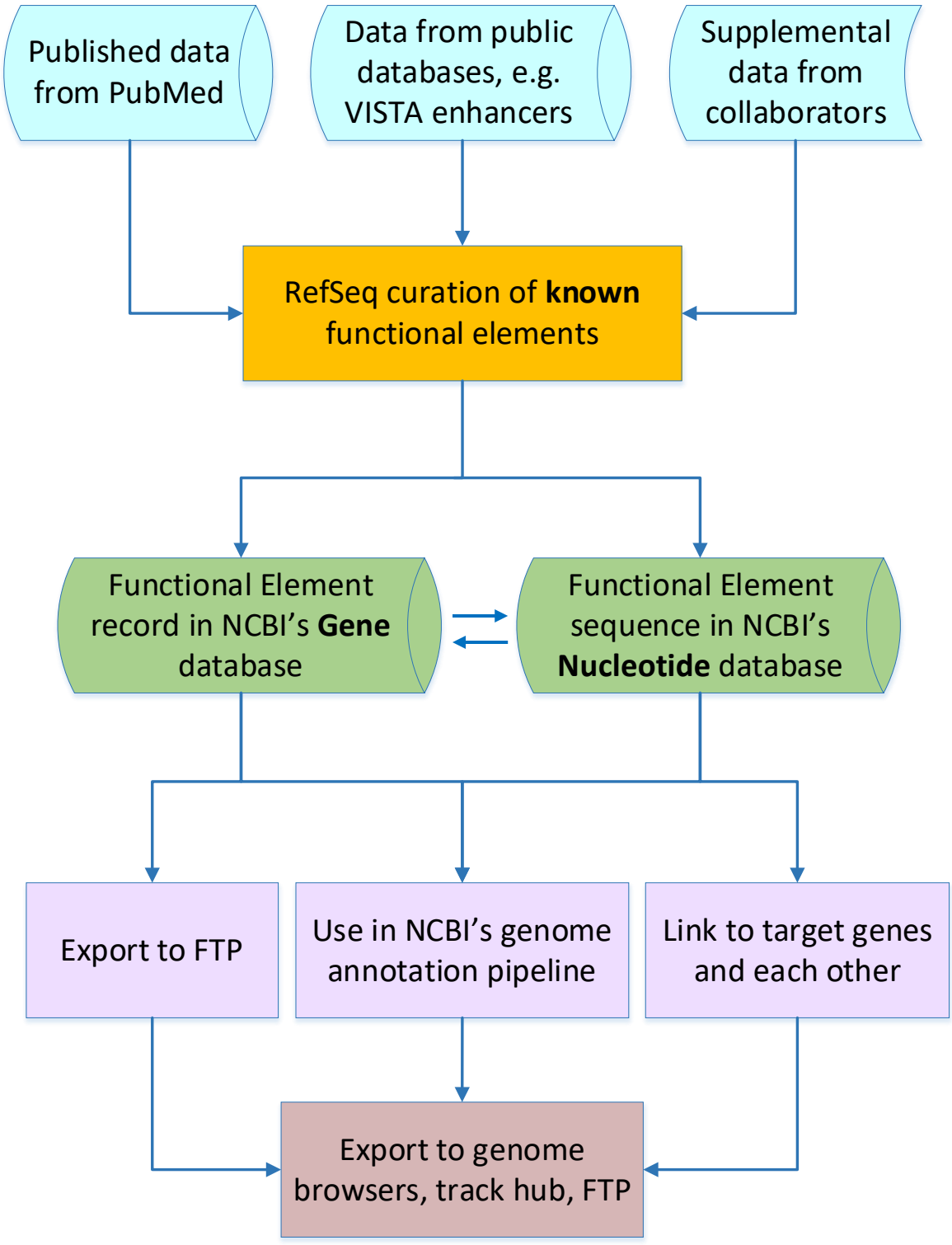

Published data from PubMed
Data from public databases, e.g. VISTA enhancers
Supplemental data from collaborators
RefSeq curation of **known** functional elements
Functional Element record in NCBI's **Gene** database
Functional Element sequence in NCBI's **Nucleotide** database
Export to FTP
Use in NCBI's genome annotation pipeline
Link to target genes and each other
Export to genome browsers, track hub, FTP

Figure 1

A

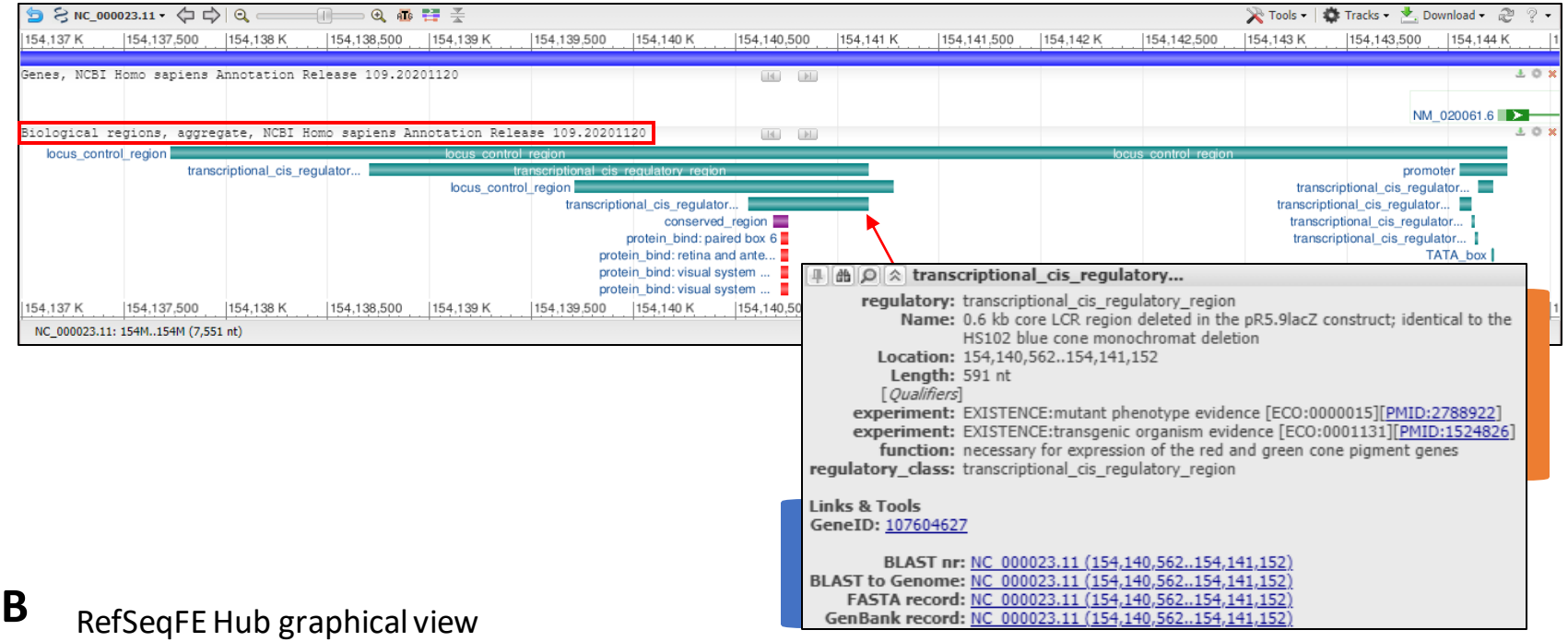
```
LOCUS       NG_052895              34462 bp    DNA     linear   CON 22-APR-2021
DEFINITION  Homo sapiens beta-globin locus control region (HBB-LCR) on
            chromosome 11.
ACCESSION   NG_052895
VERSION     NG_052895.1
DBLINK      BioProject: PRJNA343958
KEYWORDS    RefSeq; RefSeqFE.
SOURCE      Homo sapiens (human)
```

B

```
regulatory      20980..21979
                /regulatory_class="DNase_I_hypersensitive_site"
                /experiment="EXISTENCE:in vivo cleavage assay evidence
                [ECO:0001075][PMID:2370867, PMID:3879975, PMID:12941700]"
                /note="5'HS5, also known as HS5, HSS5, HSV or -21.4
                hypersensitive site; not exclusively erythroid; the
                nucleotide coordinates are approximate for this feature"
                /function="enhancer-blocking activity"
                /db_xref="GeneID:109580095"
regulatory      21054..22300
                /regulatory_class="transcriptional_cis_regulatory_region"
                /experiment="EXISTENCE:reporter gene assay evidence
                [ECO:0000049][PMID:9878258]"
                /note="1.2 kb HS5 fragment in the HS5-epsilon-p-CAT
                construct"
                /function="synergizes with the ERV-9 LTR enhancer in
                stably transfected K562 cells"
                /db_xref="GeneID:109580095"
protein_bind    21580..21651
                /experiment="EXISTENCE:protein binding evidence
                [ECO:0000024][PMID:11997516, PMID:12861010,
                PMID:18461170]"
                /note="5'HS5 CTCF-binding oligonucleotide"
                /bound_moiety="CCCTC-binding factor"
                /function="enhancer-blocking activity"
                /db_xref="GeneID:109580095"
regulatory      21580..21651
                /regulatory_class="enhancer_blocking_element"
                /experiment="EXISTENCE:reporter gene assay evidence
                [ECO:0000049][PMID:11997516]"
                /note="h5'HS5 enhancer-blocking fragment"
                /function="blocks activation of the Agamma-globin promoter
                by the mouse 5'HS2 enhancer"
                /db_xref="GeneID:109580095"
```

Figure 2

A Genome Data Viewer graphical view



B RefSeqFE Hub graphical view

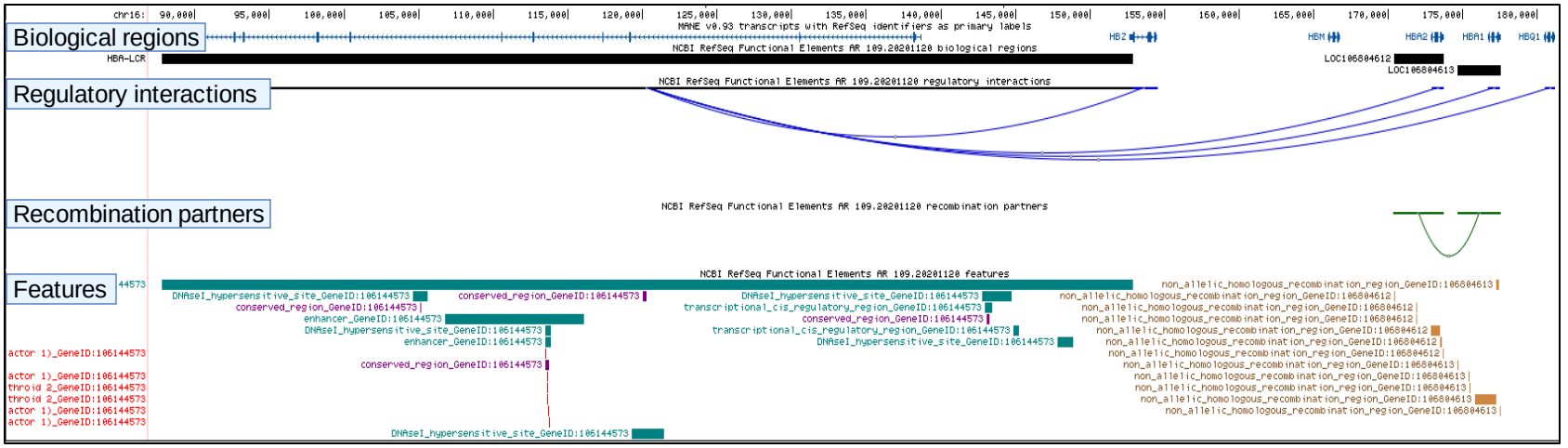
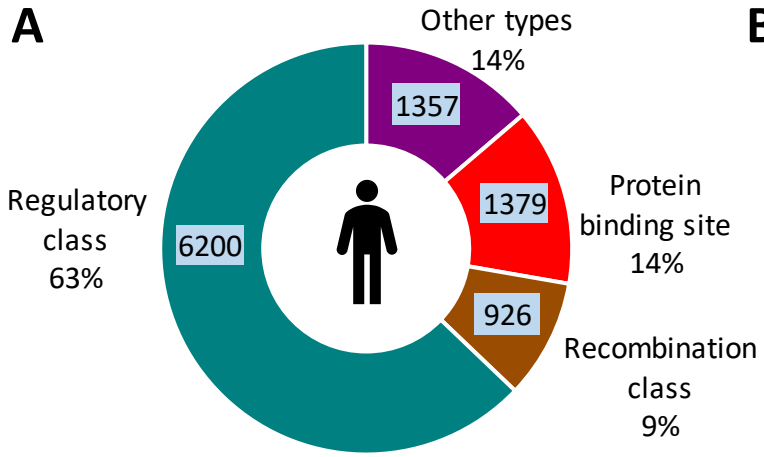
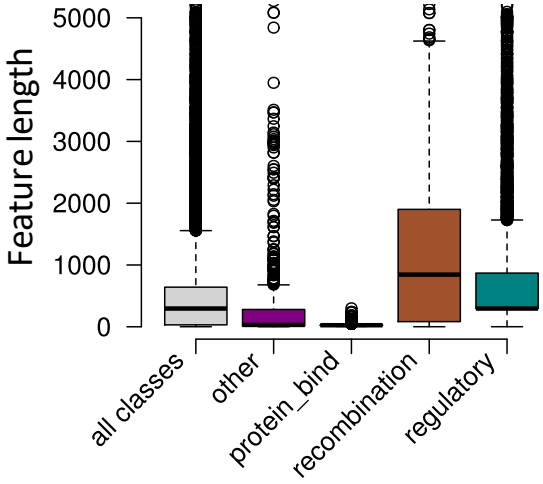
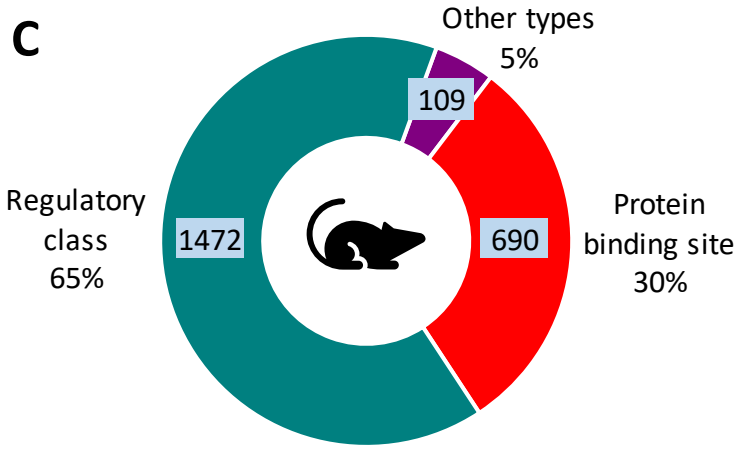
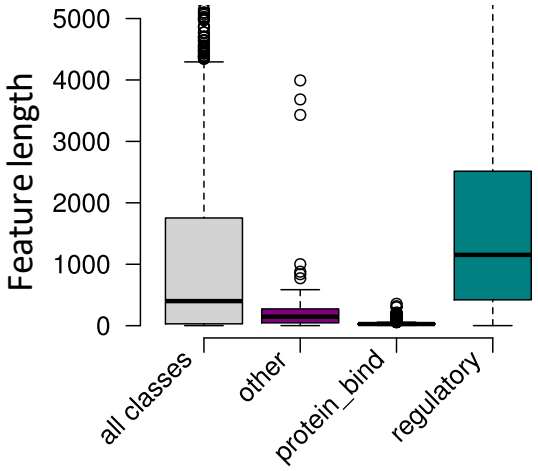


Figure 3





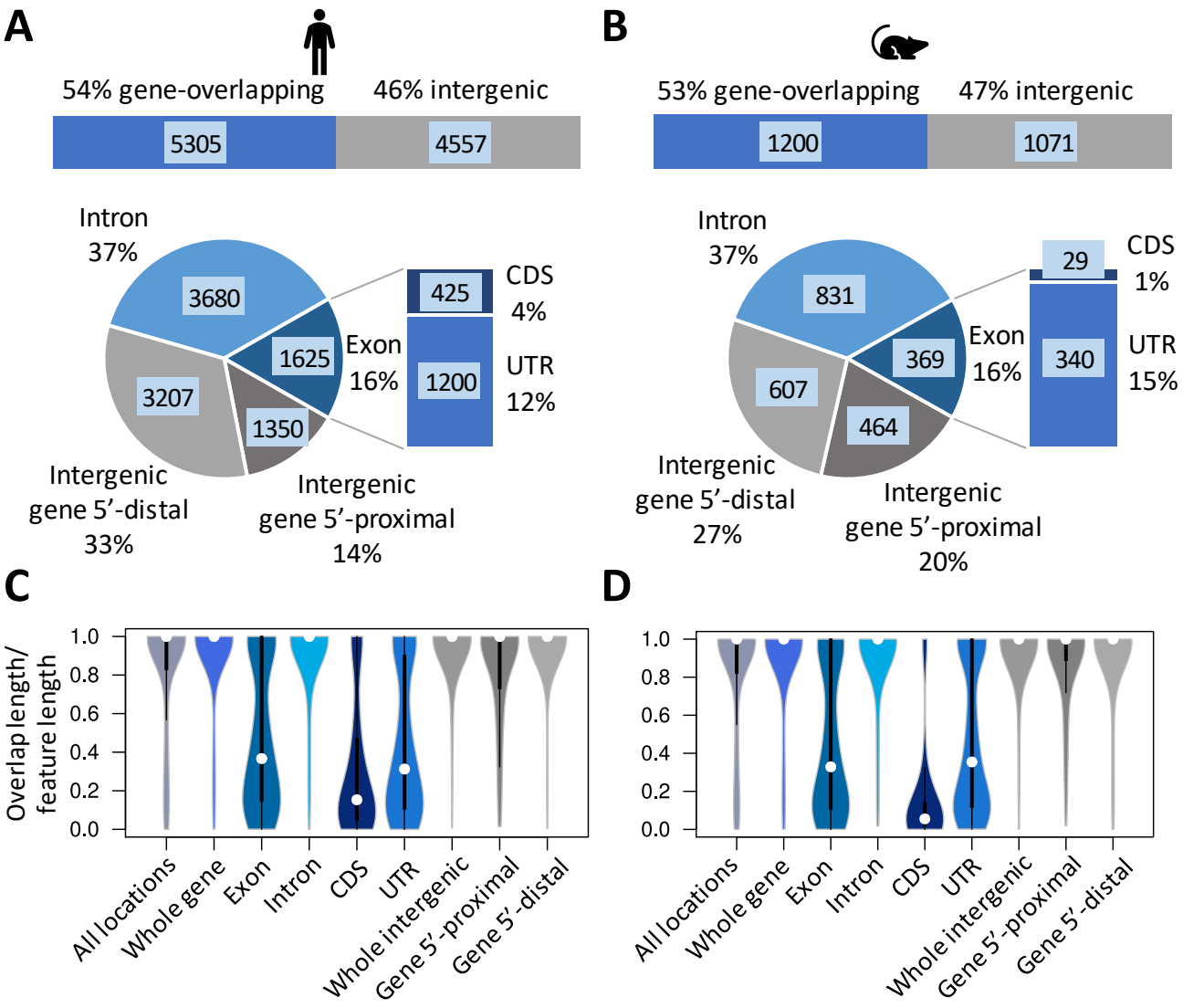




E

	Human	Mouse
Annotation Release/Assembly	109.20201120/GRCh38.p13	109/GRCm39
Feature count	9862	2271
Locus count	4450	889
Genome coverage	6.1 Mb	2.2 Mb

Figure 4



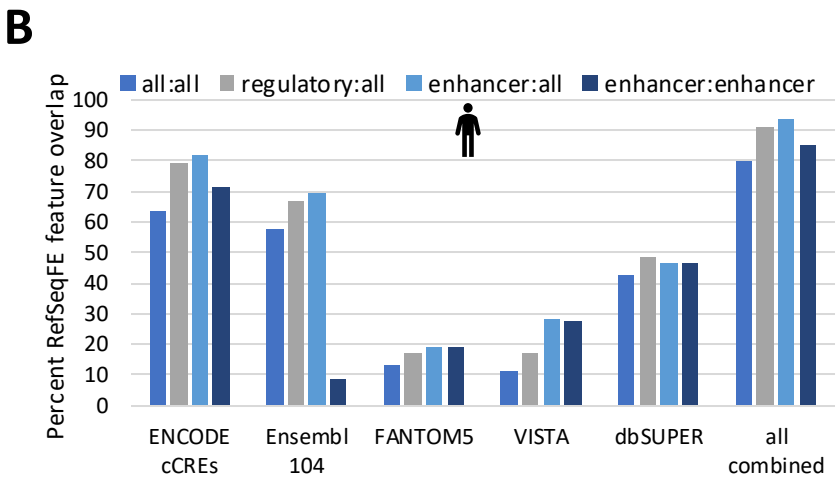
E

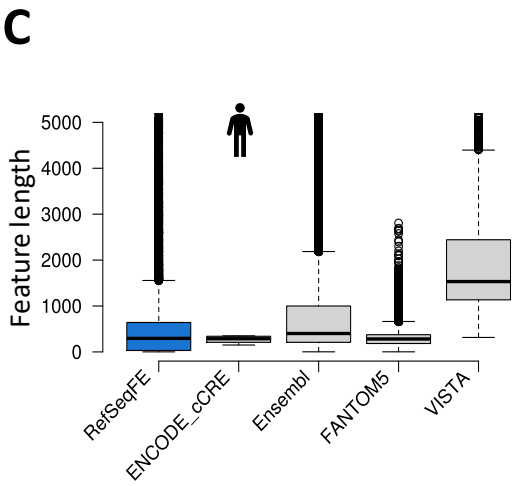
	Human		Mouse	
Overlapping gene biotype	**% total**	**count**	**% total**	**count**
protein coding	68.6	1683	63.9	361
lncRNA	27.6	677	34.0	192
pseudogene type	2.9	71	0.9	5
other RNA class	1.0	24	1.2	7

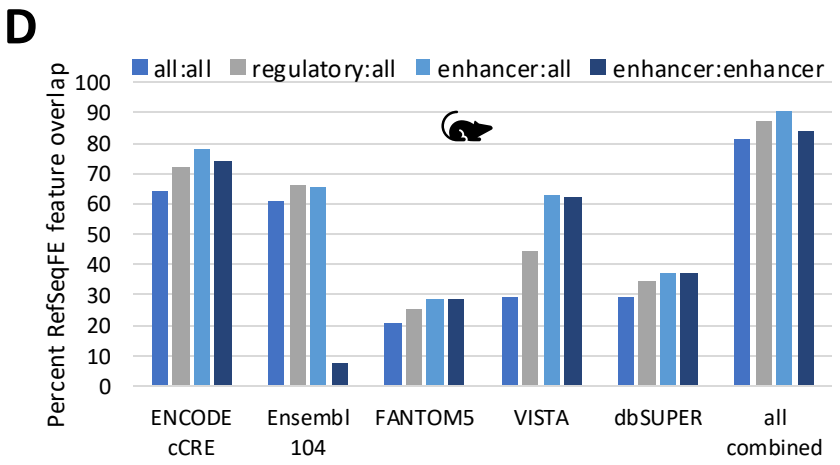
Figure 5

A

Dataset	Data derivation	Feature types	Size
RefSeq Functional Elements	Various experimental assays from the literature	>40 regulatory, protein binding, recombination and other types as in Supplemental Table S2A,B	Human: 9862 features, 6.1 Mb Mouse: 2271 features, 2.2 Mb
ENCODE cCREs	Epigenomic signatures, ChIP-seq, accessibility assays	CTCF-only, DNase-H3K4me3, distal enhancer, promoter, proximal enhancer	Human: 926535 features, 253 Mb Mouse: 343747 features, 298 Mb
Ensembl Regulation	Epigenomic signatures, ChIP-seq, accessibility assays	CTCF binding site, TF binding site, enhancer, open chromatin region, promoter, promoter flanking region	Human: 622457 features, 510 Mb Mouse: 364670 features, 298 Mb
FANTOM5 enhancers	Bidirectional balanced CAGE data, reporter assays for a subset	enhancer	Human: 63285 features, 18.6 Mb Mouse: 49802 features, 14.8 Mb
VISTA enhancers	Comparative genomics, transgenic assays	enhancer	Human: 1989 features, 3.8 Mb Mouse: 1291 features, 3.5 Mb
dbSUPER super-enhancers	Epigenomic signatures, ChIP-seq	super-enhancer	Human: 69340 features, 540 Mb Mouse: 12103 features, 159 Mb







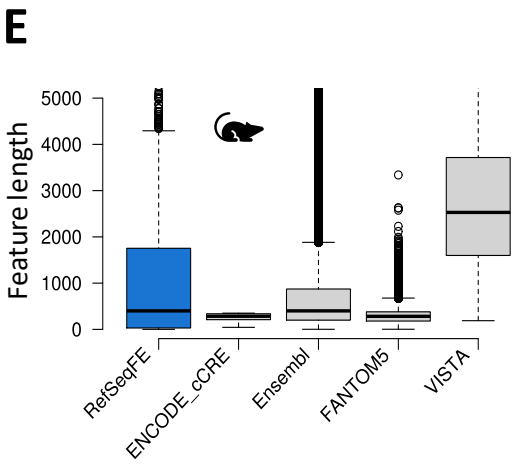


Figure 6