Correspondence

https://doi.org/10.1038/s41588-023-01365-3

Annotating and prioritizing human noncoding variants with RegulomeDB v.2

In the format provided by the authors and unedited

Supplementary Information

1. Data sources

Genomic variants

The information of genomic variants was retrieved from dbSNP153¹, including the positions and allele frequencies from different projects, such as the 1000 genome project ², TOPMED ³ and GnomAD ⁴.

ChIP-seq and chromatin accessibility experiments

We collected the peaks of ChIP-seq targeting transcription factors (TF), DNase-seq and ATAC-seq experiments called by uniform pipeline from the latest release of the ENCODE portal, which includes the experiments from the Roadmap project ⁵.

PWM matching

We downloaded the PWMs (position weight matrices) of 746 non-redundant TF motifs from JASPAR 2020 database ⁶.The kmers matching to TF motifs were called by TFM P-value with a threshold at 4⁻⁸ for each PWM ⁷. Bowtie was used to map the kmers on the genome to determine the final PWM matching positions for the TF motifs ⁸. The information content from each PWM was also integrated into the database and used as a feature to calculate the probabilistic score from the random forest model.

Footprints

Footprints were predicted with signals from 642 DNase-seq experiments and 591 TF motifs by the *TRACE* pipeline: <u>https://www.encodeproject.org/search/?type=Annotation&internal_tags=RegulomeDB_2_2&ann_otation_type=footprints&software_used.software.name=trace</u>⁹. TRACE is a computational method that incorporates DNase-seq signals and PWMs within a multivariate hidden Markov model to detect footprint regions with matching motifs.

Chromatin states

Chromatin states in 833 biosamples were called from chromHMM in EpiMap ¹⁰, and were directly retrieved from the ENCODE portal.

eQTLs

The eQTLs from the GTEx project across 49 human tissues were downloaded from the GTEx portal

(https://storage.googleapis.com/gtex_analysis_v8/single_tissue_qtl_data/GTEx_Analysis_v8_e QTL.tar) ¹¹. The variant-gene pairs with the corresponding tissue were added as annotations in the database.

caQTLs

The chromatin accessibility QTLs (caQTLs) were collected from 9 publications ^{12–20} <u>https://www.encodeproject.org/search/?type=Annotation&internal tags=RegulomeDB 2 2&ann</u>

<u>otation_type=caQTLs</u>. Only SNVs were included and lifted over from hg19 to GRCh38 if necessary ²¹.

Prediction scores

We provide a heuristic ranking and a probabilistic score for each query variant representing its potential of being a functional variant in regulatory elements. The heuristic ranking is defined in the same way as in the previous version of RegulomeDB²². The probabilistic score is calculated from a random forest model, TURF, trained with allele-specific TF binding SNVs²³. We used a simplified version here only including binary features from functional genomic evidence as used in the heuristic ranking, as well as numeric features from information content in matched PWMs. We will include the whole feature set in a future release.

2. Database and web server design

RegulomeDB annotates a variant by intersecting its position with genomic intervals identified from a massive number of experiments and computational approaches. The database directly integrates the datasets from the ENCODE portal creating a genomic data service (https://github.org/ENCODE-DCC/genomic-data-service). The genomic intervals are parsed from BED formatted files and associated with metadata of the source experiments and computational pipelines from the ENCODE portal. These BED files are then indexed in Elasticsearch (https://www.elastic.co/) as in integer range type to enable efficient search against a query position. In total, over two billion genomic intervals representing ChIP-seq and DNase-seq peaks, matches to PWMs and DNase footprints, eQTLS, caQTLs and chromatin states are indexed in Elasticsearch. After each search, the JSON objects associated with the intersected intervals are returned and passed on to generate ranking scores from RegulomeDB 1.1 and new probabilistic scores from TURF ^{23,24}. The query results are displayed with a web interface (https://github.org/ENCODE-DCC/regulome-encoded) that contains charts and interaction figures, which can be customized by users.

3. New interface for variant functionality exploration

The RegulomeDB v2 web server accepts any query variant on the whole genome in either GRCh38 or hg19 genome assembly. A toggle above the search box allows users to switch between the two assemblies. The search box allows any user to input multiple queries (up to 500 at a time) (Supplementary Figure 3). The input query can be in three formats: 1) rsID (from dbSNP database v153); 2) chromosome position for a single nucleotide variant; 3) chromosome position for a chromosome region. In the third case, all variants on the chromosome region at >1% allele frequency from dbSNP153 will be queried. The backend then intersects the variant(s) position with the genomic intervals of annotations obtained predicted from functional genomics experiments and returns a sortable summary table of variant scores (Supplementary Figure 3), including a ranking score and a probabilistic score showing its potential of being a regulatory variant. In addition, a dbSNP rsID will link to the query variant if it exists.

After clicking on any field of a row in the score table, a more detailed information page on genomic evidence is shown for the variant of interest (Supplementary Figure 4, Supplementary Figure 5). The top of the page shows some basic information on the variant position, scores, and allele frequencies from the dbSNP database. While on the bottom is the initial summary section on genomic annotations' hits. Since a single query can hit up to 2,000 results, the initial summary section is divided into five data types; TF binding sites from ChIP-seq, chromatin states from chromHMM, chromatin accessibility, PWM matching or footprint predictions, and eQTLs or caQTLs. In addition, a genome browser section is also available to view the specific DNase-seq and ChIP-seq data, which can aid in variant interpretation.

Each of the six sections can be clicked to display more details on the genomic hits from specific assays, such as the biosample of DNase peaks and the transcription factors of ChIP-seq peaks. The chromatin state tab shows the chromHMM state for each of the 833 biosamples, which also includes an intuitive body map colored by the most active chromatin state in each organ. Furthermore, the genome browser tab provides an interaction view for exploring the gene transcripts along with DNase-seq and ChIP-seq peaks near the variant of interest (shown as a yellow highlight). The tracks on the genome browser can be further filtered using a modal that allows one to sub-select by specific organ/cell types, biosample types, file types, assay methods, or by TF targets.

References

1. Sherry, S. T. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research* vol.

29 308–311 Preprint at https://doi.org/10.1093/nar/29.1.308 (2001).

1000 Genomes Project Consortium *et al.* A global reference for human genetic variation.
 Nature 526, 68–74 (2015).

Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program.
 Nature 590, 290–299 (2021).

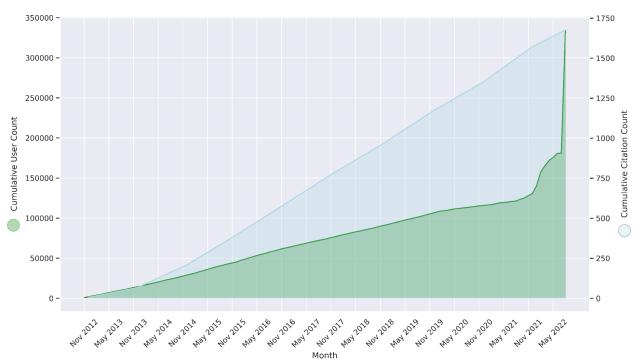
- Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 434–443 (2020).
- Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
- Fornes, O. *et al.* JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 48, D87–D92 (2020).
- 7. Touzet, H. & Varré, J.-S. Efficient and accurate P-value computation for Position Weight

Matrices. Algorithms Mol. Biol. 2, 15 (2007).

- 8. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
- 9. Ouyang, N. & Boyle, A. P. TRACE: transcription factor footprinting using chromatin accessibility data and DNA sequence. *Genome Res.* **30**, 1040–1046 (2020).
- 10. Boix, C. A., James, B. T., Park, Y. P., Meuleman, W. & Kellis, M. Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature* **590**, 300–307 (2021).
- GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 369, 1318–1330 (2020).
- 12. Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–394 (2012).
- Schwartzentruber, J. *et al.* Molecular and functional variation in iPSC-derived sensory neurons. *Nat. Genet.* **50**, 54–61 (2018).
- Khetan, S. *et al.* Type 2 Diabetes–Associated Genetic Variants Regulate Chromatin Accessibility in Human Islets. *Diabetes* 67, 2466–2477 (2018).
- Gate, R. E. *et al.* Genetic determinants of co-accessible chromatin regions in activated T cells across humans. *Nat. Genet.* 50, 1140–1150 (2018).
- Tehranchi, A. *et al.* Fine-mapping cis-regulatory variants in diverse human populations. *Elife* 8, (2019).
- 17. Kumasaka, N., Knights, A. J. & Gaffney, D. J. High-resolution genetic mapping of putative causal interactions between regions of open chromatin. *Nat. Genet.* **51**, 128–137 (2019).
- Zhao, Q. *et al.* Molecular mechanisms of coronary disease revealed using quantitative trait loci for TCF21 binding, chromatin accessibility, and chromosomal looping. *Genome Biol.* 21, 135 (2020).
- Liang, D. *et al.* Cell-type-specific effects of genetic variation on chromatin accessibility during human neuronal differentiation. *Nat. Neurosci.* 24, 941–953 (2021).

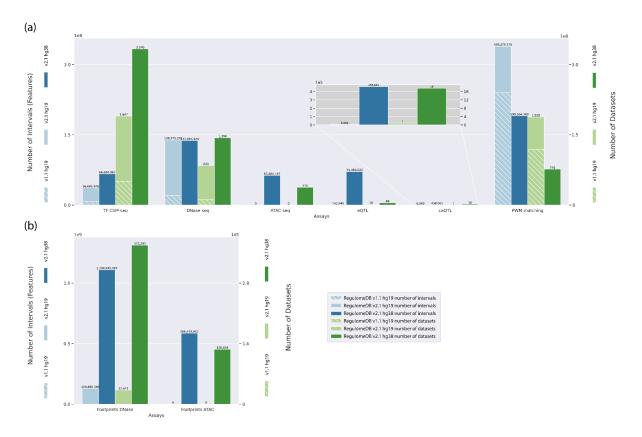
- 20. Currin, K. W. *et al.* Genetic effects on liver chromatin accessibility identify disease regulatory variants. *Am. J. Hum. Genet.* **108**, 1169–1189 (2021).
- Kuhn, R. M., Haussler, D. & Kent, W. J. The UCSC genome browser and associated tools.
 Brief. Bioinform. 14, 144–161 (2013).
- Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* 22, 1790–1797 (2012).
- Dong, S. & Boyle, A. P. Prioritization of regulatory variants with tissue-specific function in the non-coding regions of human genome. *Nucleic Acids Res.* (2021) doi:10.1093/nar/gkab924.
- Dong, S. & Boyle, A. P. Predicting functional variants in enhancer and promoter elements using RegulomeDB. *Hum. Mutat.* 40, 1292–1298 (2019).

Supplementary Figures



Supplementary Figure 1. Popularity of RegulomeDB.

The x-axis is month and year since RegulomeDB first published in 2012. The left y-axis is cumulative user count (green). The right y-axis is cumulative citation count (light blue). The citation count data are derived from Clarivate Web of Science. © Copyright Clarivate 2022. All rights reserved.



Supplementary Figure 2. Overview of RegulomeDB Version 2 Data Growth and Refinement. Statistics on database content. Numbers under each data type include all experiments across different treatment conditions and biosamples. All numbers are RegulomeDB v2 stats, in hg19 or hg38.



Q Search by dbSNP ID or coordinate range: GRCh38 _____ hg19

rs75982468	
rs10117931	
rs11749731	
rs11160830	
rs2808110	
rs2839467	
rs147375898	
rs111686660	

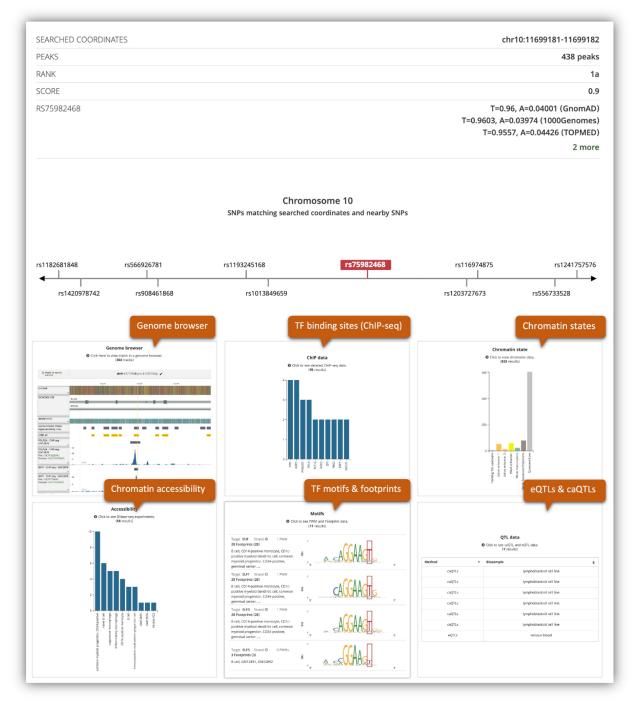
Click for example entry: multiple dbSNPs or coordinates ranges

Search

This s	earch has	found 15 varia	nt(s).				
Do	wnload BEI	D Download T	sv				
Chromosome location	\$	dbSNP IDs	\$	Rank	•	Score	4
chr10:1169918111699182		rs75982468		1a		0.97	
chr9:45751194575120		rs10117931		1b		0.77931	
chr14:105015650105015651		rs11160830		1f		0.36978	
chr17:4019387440193875		rs74792881		1f		0.33586	
chr21:4209308942093090		rs2839467		1f		0.22271	
chr5:142120870142120871		rs11749731		1f		0.55324	
chr5:149667681149667682		rs147375898		2a		0.9135	
chr7:7373189273731893		rs190318542		2a		0.9943	
chr19:5066393950663940		rs3087079		3a		0.52739	
chr1:8837392288373923		rs2808110		3b		0.72329	
chr9:7704057877040579		rs11145227		4		0.60906	
chr12:128972886128972887		rs111686660		5		0.86083	
chr11:1863949018639491		rs2166521		6		0.23675	
chr5:117154078117154079		rs148232663		6		0.27391	
chr4:8226387282263873		rs62319725		7		0.51392	

Supplementary Figure 3. RegulomeDB Query Interface.

An example query with the rsIDs of variants from dbSNP database. Upon clicking the search buttons, a summary table representing prediction scores for all query variants will be displayed. (See Supplementary Note 3 for more details.)



Supplementary Figure 4. RegulomeDB Result Overview Page of rs75982468.

For any variant of interest, a results page with more information on the hits of genomic annotations is available. Each of the six sections at the bottom can be clicked to expand more details on each data type.

Genome browser ChiP data Chromatin state Accessibility wear Que saw	Genome browser ChiP data Chromatin state Accessibility Mours QIX eats
	NA accessibility experiments ✓ ✓ q Search for a biosample name x
2-	common myslold progenitor, CD34-positive 18
	File Dataset Organ Method ENCFF734TMQ ENCSR453EVC bone marrow, bone element Dhase-seq
	ENCFF620JAI ENCSR191EII bone marrow, bone element DNase-seq ENCFF667AVX ENCSR468EXN bone marrow, bone element DNase-seq
 9.9 (9) 9.9 (9)	ENCFF063IWN ENC5R281HBB bone marrow, bone element DNase-seq ENCFF843YAB ENC5R20TUJ bone marrow, bone element DNase-seq
Electronic and a second se	ENCFF239LXI ENCSR564TUY bone marrow bone element DNare-sig ENCFF752XGR ENCSR758BIM bone marrow bone element DNare-sig
hiP data	ENCFF116WZJ ENCSR271QSV bone marrow, bone element DNase-seq
Iteltend + Peak location Biosample or Biosample or Signal Targets or Biolog, bodly fluid Organ Dataset File Value or Signal hIP-seq chrl0:116990-0	ENCFF42QEL ENCSR122VUW bone marrow, bone element DNase-seq ENCFF785BMM ENCSR6392JI bone marrow, bone element DNase-seq
NIP-seq chr10:11698974.11699563 GM12878 PKNOX1 blood, bodily fluid ENCSR711XNY ENCFF946JFA 3359.30479	naive 8 cell
hIP-sea chr10:11699114_11699370 GM12878 TCF12 blood. bodily fluid ENC\$R00086Z ENCFF068YYR 22.82952	
Genome browser ChiP data Chromatin state Accessibility Voters ort data	eQTLS & CaQTL Genome browser ChiP data Chromatin state Accessibility Motifs L data ~
ttifs	caQTL data
RCh38 Reference CACAATGAGGAAGTGACATTTTC	Method ~ QTL location © Biosample ~ Population © Dataset © File caQTLs chritCr16991811699182 lymphoblastoid cell line ESN ENCSRS33BHN ENCFR212WCH
TABLE BHE Strand © CMM ENCEPTRESSD 2	caQTLs chr10:1169918111699182 lymphoblastoid cell line LWK ENCSR533BHN ENCFF694HXE
otorines 2016 a coll CO14-positive monocyte CO12-positive monocyte	caQTLs chr10.1169918111699182 lymphoblastoid cell line N/A ENCERBORASP ENCEF919URR
rget BJP1 Strand Ø PVM ENCEPHERCEN	caQTLs chr10:11699181.11699182 lymphoblastoid cell line YRI ENCSR533BHN ENCFF277UAK
bogannis (28) B call CD14 positive myeloid dendratic call	cxQTLs chr10:11697182 lymphoblassad cell line CWD ENCERSUBIN ENCERSUBIN cxQTLs chr10:11697182 lymphoblassad cell line ASW ENCERSUBIN ENCERSUBIN
vrgit: EU-3 Strand © PVIM ENC/FØS3AGK 2 VC 11	eQTL data
objirinis (20) III celli enotopite CDI c-positive monocyte CDI c-positive myreloid denotric cell	Method • QTL location • Biosample • Target genes • Dataset • File eQTLs dr10.11699181_11699182 venous blood ECHDC3 ENCR7360MV ENCR7360MV
Genome browser Chill data Coronastin state Accessibility Motifs Of Jata 2	Genome browser ChiP data Chromatin state Accessibility Molth 91 data
Filter by organ T Filter by biosample T Filter by chromatin state	Visualize files for chr10:11699181-11699182
lared by more lasters stare. Crossood by argues Contended by transcription actuary Contended by transcription actuary Contended by Tables Contende	There are 302 tracks. Selected fitem: To select a fiter, make a selection from "Refne your search" below.
naive B cell blood, bodly fuid 1 Active enhancer 2	REFINE YOUR SEARCH
T-cell blood, bodly fuid Weak transcription	Statest to query virset der10 116941812p to 117041822p ✓
GM1202 slock, fould 1 GM1202 slock, fould 1 GM1202 slock, fould 1 GM1202 slock, fould 1 GM1204 slock, fould /slock	I DAY DAY DAY DAY DAY DAY DAY DAY DAY
Two watching in the cost of th	RCDB
Control pred took body and progenitor cell blook body fuld control pred took body fuld	UNLUKE V/F
Child and the second se	
lected filters: O adreval gland blood	SNP (153) presentation (Name partnership) years
Gearalifiters	38. si
Chromatin state - Biosample © Classification © Organ © Dataset © File ©	20222 - CMP-Reg - M12278
ctive enhancer 1 nake 8 cell primary cell blood, bodily fuid ENCSR194USR ENCFF229FAN	DU572 - CMP4ee

Supplementary Figure 5. RegulomeDB Expanded Pages of rs75982468.

The expanded pages of each section shows details on the genomic experiments and annotations, such as the biosample, organ, TF target and the peak file called from the ENCODE project. The body map under the chromatin states view is colored by the most active state among all biosamples in each organ, which gives an intuitive way to explore the candidate organs where the query variant might be functional. Users can also explore the nearby genes of the query variant under the genome browser view.