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

Annotating and prioritizing human non-coding variants with RegulomeDB v.2

Nearly 90% of the disease risk-associated variants identified by genome-wide association studies are in non-coding regions of the genome. The annotations obtained by analyzing functional genomics assays can provide additional information to pinpoint causal variants, which are often not the lead variants identified from association studies. However, the lack of available annotation tools limits the use of such data. To address the challenge, we previously built the 'RegulomeDB database' to prioritize and annotate variants in non-coding regions[1], which has been a highly utilized resource for the research community (Supplementary Fig. 1).

Here we present an update of the RegulomeDB web server, RegulomeDB v.2 (http://regulomedb.org). RegulomeDB annotates a variant by intersecting its position with genomic intervals identified from functional genomic assays and computational approaches. It also incorporates variant hits into a heuristic ranking score, representing its potential to be functional in regulatory elements. We improve and boost annotation power by incorporating thousands of newly processed data from functional genomic assays in GRCh38 assembly and include probabilistic scores from the SURF algorithm that was the top performing non-coding variant predictor in the Fifth Critical Assessment of Genome Interpretation (CAGI-5)[2].

The update of RegulomeDB now includes more than 650 million and 1.5 billion genomic intervals in hg19 and GRCh38, respectively – a fivefold increase compared with the previous version (Supplementary Fig. 2). We included approximately 5,000 chromatin immunoprecipitation followed by sequencing experiments targeting transcription factors (TF ChIP–seq), and chromatin accessibility experiments from the ENCODE project[3], the Roadmap Epigenomics program[4], and the Genomics of Gene Regulation project. We also produced a comprehensive set of footprint predictions using over 800 chromatin accessibility experiments and 591 transcription factor motifs in GRCh38 using the TRACE pipeline[5]. In addition, we refined the included transcription factor motifs by using the non-redundant vertebrates set from the JASPAR database[6]. We also integrated approximately 71 million variant–gene pairs in expression quantitative trait loci (eQTL) studies from the GTEx project[7], and 450,000 chromatin-accessibility QTLs (caQTLs) from 9 recent publications (Supplementary Information). Finally, we included chromatin state annotations known as from chromHMM in EpiMap for 833 biosamples[8].

RegulomeDB accepts any query variants genome-wide in either GRCh38 or hg19 genome assembly by rsID or genome coordinates. The query variants can then be prioritized by functional prediction scores shown in a sortable table. For any variant of interest, an information page on five types of supported genomic evidence, as well as a genome browser view is displayed. Each of the six sections can be clicked to show more detail for functionality exploration (Supplementary Figs. 3–5).

RegulomeDB enables researchers to quickly separate functional variants from a large pool of variants and assign tissue or organ specificity for each variant. Here we showcase this using four verified variants from recent literature[9–13], and demonstrate the applicability of RegulomeDB to annotate those variants based on various sources of data (Fig. 1).

Transcription factor motifs and ChIP–seq data together provide evidence about how a variant is likely to affect phenotype in a cell-specific context. For example, rs213641 is known to affect behavioral responses to fear and anxiety stimuli[9]. The POLR2A binding and the active transcriptional start site (TSS) state in the brain indicate that rs213641 is likely to function in the brain by disrupting the TSS of *STMN1*. We also examined rs7789585, in which RegulomeDB transcription factor motif evidence suggests that mutation to the reference allele G would disrupt the binding of GCM1, which may interrupt the active enhancer state at the locus in the heart. Hocker et al.[10] recently confirmed this hypothesis using reporter assays, and discovered that rs7789585 disrupts a *KCNH2* enhancer and affects cardiomyocyte electrophysiologic function.

DNase-seq assays and underlying footprint predictions identify open chromatin regions with mapped transcription factor binding sites in hundreds of biosamples and can also be used to assign putative function to variants. rs190509934 has been associated with the risk of COVID-19 infection by affecting *ACE2* expression[11]. RegulomeDB shows hits to several DNase-seq peaks in lung-related biosamples. Furthermore, RegulomeDB extends this tissue effect with the hypothesis that *ACE2* expression may be regulated by CEBP by its overlap with DNase footprints in the lung found in the upstream promoter region of *ACE2*[12]. In addition, eQTL studies provide correlation evidence between the variants and their target genes. For example, rs72635708 is predicted as a regulatory variant by RegulomeDB with a high probability of 0.91 due to its locus overlapping with DNase and ChIP–seq peaks, footprints, and it is an eQTL that associates with *LINC01714* gene expression in the right lobe liver. Because rs72635708 lies in the FOS motif, it is likely to be a functional variant in the liver by modulating the binding of the AP-1 complex[13].

In summary, RegulomeDB provides a user-friendly tool to annotate and prioritize variants in non-coding regions of the human genome, which can aid variant function interpretation and guide follow-up experiments. We welcome user feedback through regulomedb@mailman.stanford.edu.

Reporting Summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

RegulomeDB v.2 can be accessed through the web server at https://regulomedb.org. All datasets collected in RegulomeDB are accessible through the ENCODE portal https://www.encodeproject.org/search/?internal_tags=RegulomeDB_2_2.

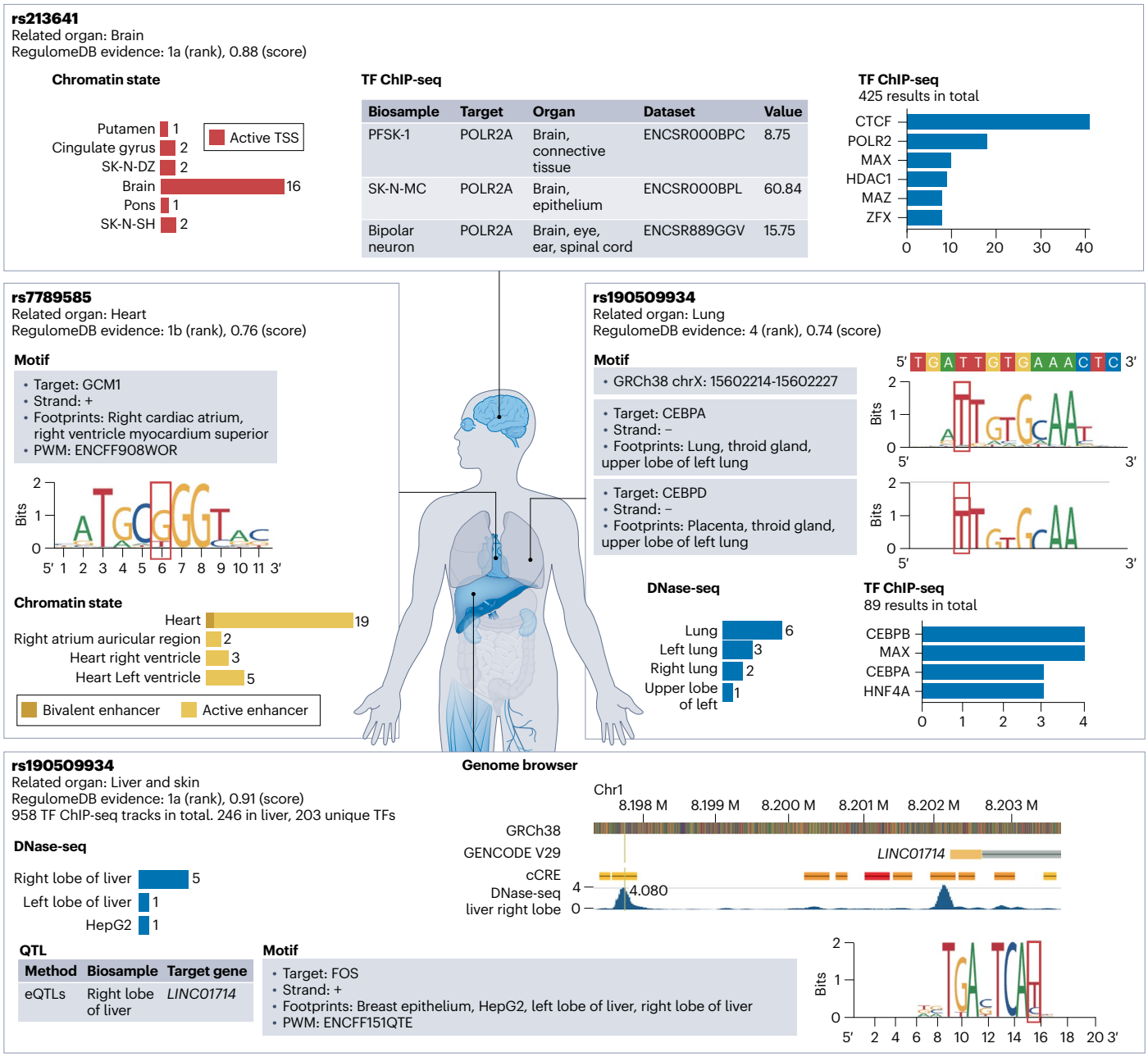

Fig. 1 | Prioritization of functional variants with RegulomeDB version 2. Four example variants with verified functions in related organs from recent literature. Various sources of evidence in RegulomeDB are indicated by gray boxes. RegulomeDB heuristic ranking score and probability score summarized all evidence.

Code availability

The code RegulomeDB uses is available on GitHub repository at https://github.com/ENCODE-DCC/regulome-encoded/releases/tag/v2.2 and https://github.com/ENCODE-DCC/genomic-data-service/releases/tag/v2.2.

Shengcheng Dong[1,4], Nanxiang Zhao[2,4], Emma Spragins[1], Meenakshi S. Kagda[1], Mingjie Li[1], Pedro Assis[1], Otto Jolanki[1], Yunhai Luo[1], J. Michael Cherry[1], Alan P. Boyle[2,3] ✉ & Benjamin C. Hitz[1] ✉

[1]Department of Genetics, Stanford University, Stanford, CA, USA. [2]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA. [3]Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA. [4]These authors contributed equally: Shengcheng Dong, Nanxiang Zhao. ✉e-mail: apboyle@umich.edu; hitz@stanford.edu

Published online: 25 April 2023

References

1. Boyle, A. P. et al. *Genome Res.* **22**, 1790–1797 (2012).
2. Dong, S. & Boyle, A. P. *Hum. Mutat.* **40**, 1292–1298 (2019).
3. ENCODE Project Consortium. *Nature* **583**, 699–710 (2020).

4. Roadmap Epigenomics Consortium. *Nature* **518**, 317–330 (2015).
5. Ouyang, N. & Boyle, A. P. *Genome Res.* **30**, 1040–1046 (2020).
6. Fornes, O. et al. *Nucleic Acids Res.* **48**, D87–D92 (2020).
7. GTEx Consortium. *Science* **369**, 1318–1330 (2020).
8. Boix, C. A., James, B. T., Park, Y. P., Meuleman, W. & Kellis, M. *Nature* **590**, 300–307 (2021).
9. Brocke, B. et al. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **153B**, 243–251 (2010).
10. Hocker, J. D. et al. *Sci. Adv.* **7**, eabf1444 (2021).
11. Horowitz, J. E. et al. *Nat. Genet.* **54**, 382–392 (2022).
12. Beacon, T. H., Delcuve, G. P. & Davie, J. R. *Genome* **64**, 386–399 (2021).
13. Kubota, N. & Suyama, M. *BMC Med. Genomics* **13**, 8 (2020).

Acknowledgements

We thank the RegulomeDB users and the scientific community for producing and sharing functional genomic experiments. We also thank all members in the Cherry and Boyle laboratories for constructive feedbacks. This research was supported by US National Institutes of Health (NIH) grants U24 HG009293 (A.P.B. and J.M.C).

Competing interests

The authors have no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-023-01365-3.

nature portfolio

Corresponding author(s): Boyle and Hitz

Last updated by author(s): Feb 16, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	All datasets collected in RegulomeDB are accessible through the ENCODE portal https://www.encodeproject.org/search/?internal_tags=RegulomeDB_2_2.
Data analysis	The code RegulomeDB uses is available on GitHub repository at https://github.com/ENCODE-DCC/regulome-encoded/releases/tag/v2.2

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

RegulomeDB v2 can be fully open accessed through the web server at https://regulomedb.org.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	*Use the terms sex (biological attribute) and gender (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms. Indicate if findings apply to only one sex or gender; describe whether sex and gender were considered in study design whether sex and/or gender was determined based on self-reporting or assigned and methods used. Provide in the source data disaggregated sex and gender data where this information has been collected, and consent has been obtained for sharing of individual-level data; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex- and gender-based analyses where performed, justify reasons for lack of sex- and gender-based analysis.*
Population characteristics	*Describe the covariate-relevant population characteristics of the human research participants (e.g. age, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."*
Recruitment	*Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*
Ethics oversight	*Identify the organization(s) that approved the study protocol.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We showcased four variants with RegulomeDB evidence. Selection was based on the recent publications and the intention to demonstrate a variety of RegulomeDB interfaces.
Data exclusions	No data were excluded.
Replication	No replication was performed in our study.
Randomization	Randomization was not application. We did not perform statistical tests.
Blinding	Blinding was not application. We did not perform statistical tests.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging