FORGEdb: systematic analysis of candidate causal variants to uncover target genes and mechanisms in complex traits.

Charles E. Breeze[1,2,3,*], Eric Haugen[2], María Gutierrez-Arcelus[4,5], Xiaozheng Yao[1], Andrew Teschendorff[6], Stephan Beck[3], Ian Dunham[7], John Stamatoyannopoulos[2], Nora Franceschini[8], Mitchell J. Machiela[1], and Sonja I. Berndt[1]

[1] Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892

[2] Altius Institute for Biomedical Sciences, 2211 Elliott Avenue 98121 Seattle

[3] UCL Cancer Institute, University College London, 72 Huntley Street, London WC1E 6BT, United Kingdom.

[4] Division of Immunology, Department of Pediatrics, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA

[5] Broad Institute of MIT and Harvard, Cambridge, MA, USA

[6] CAS Key Lab of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute for Biological Sciences, Chinese Academy of Sciences, 320 Yue Yang Road, Shanghai 200031, China.

[7] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

[8] Department of Epidemiology, University of North Carolina, Chapel Hill, NC, USA

* Email address: c.breeze@ucl.ac.uk

Abstract

The majority of disease-associated variants identified through genome-wide association studies (GWAS) are located outside of protein-coding regions, and are collectively overrepresented in sequences that regulate gene expression. Prioritizing candidate regulatory variants and potential biological mechanisms for further functional experiments, such as genome editing, can be challenging, especially in regions with a high number of variants in strong linkage disequilibrium or multiple proximal gene targets. Improved annotation of the regulatory genome can help identify promising variants and target genes for further experiments and accelerate translation of identified GWAS loci into important biological insights. To advance this area, we developed FORGEdb (https://forge2.altiusinstitute.org/files/forgedb.html), a web-based tool that can rapidly integrate data for individual genetic variants, providing information on associated regulatory elements, transcription factor (TF) binding sites and target genes. FORGEdb uses annotations derived from data across a wide range of biological samples to delineate the regulatory context for each variant at the cell type level. Different datatypes, including CADD scores, expression quantitative trait loci (eQTLs), activity-by-contact (ABC) interactions, Contextual Analysis of TF Occupancy (CATO) scores, TF motifs, DNase I hotspots, histone mark ChIP-seq and chromatin states in FORGEdb are made available for >37 million variants, and these annotations are integrated into a FORGEdb score to guide assignment of functional importance. The inclusion of a wide range of genomic annotations, such as ABC interactions and CADD scores, provides a comprehensive resource for researchers seeking to prioritize variants for functional validation. In summary, FORGEdb provides an expansive and unique resource for the analysis of genomic variants associated with complex traits and diseases.

Introduction

Genome-wide association studies (GWAS) have been remarkably successful in identifying genetic loci associated with many different diseases and traits[1]. The latest version of the GWAS catalog (as of 2022) comprises 228,157 distinct variants associated with >3,000 diseases and traits[2]. Many loci identified from GWAS are intergenic and locate to non-protein-coding regions of the genome[3]. Although the functional mechanisms of some variants have been reported[4,5], most genomic loci have not been studied and little is known about target genes, pathways or mechanisms of action. There are multiple reports that GWAS variants are overrepresented in sequences that regulate gene expression[3]. Therefore, to aid interpretation of GWAS variants in the context of gene regulation, researchers have used large-scale mapping data for enhancers and other regulatory elements from ENCODE[6], Roadmap Epigenomics[7], and other consortia[8]. Several webtools, such as Haploreg[9], RegulomeDB[10] and others[11], have been developed to help researchers link these data to individual variants. However, these methods do not include more recent high-dimensional ENCODE data from contemporary technologies, such as Hi-C[12], or expanded expression quantitative trait locus (eQTL) data from large consortia, such as the Genotype-Tissue Expression Project (GTEx)[13] or the eQTLGen project[14]. Gathering information from many different data sources and linking the data to individual genetic variants is challenging in terms of computational resources and in terms of quality control and data processing.

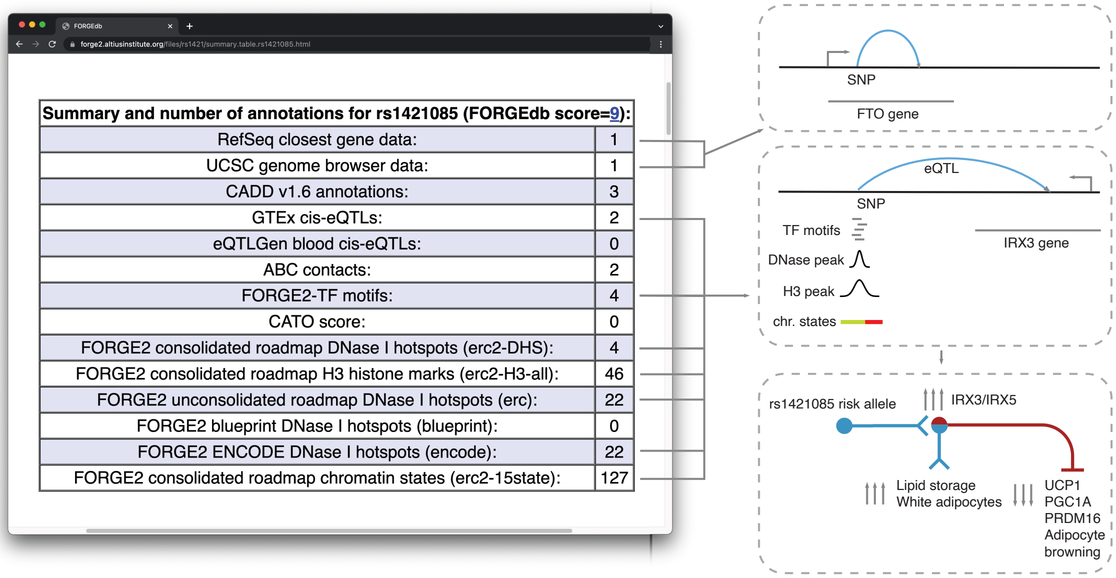
Description

To address this issue and provide researchers with a state-of-the-art web tool for variant annotation that includes these updated resources, we have developed FORGEdb (https://forge2.altiusinstitute.org/files/forgedb.html, **Table 1**). FORGEdb incorporates a range of datasets covering three broad areas relating to gene regulation: regulatory regions, transcription factor (TF) binding, and target genes. First, using genome-wide epigenomic track data from ENCODE, Roadmap Epigenomics, and BLUEPRINT consortia, FORGEdb links SNPs with data for candidate regulatory regions (e.g., enhancers or promoters). Specifically, FORGEdb annotates variants for overlap with DNase I hotspots, histone mark broadPeaks, and chromatin states as implemented in FORGE2[15]. Second, within these candidate regulatory regions, FORGEdb integrates SNPs with transcription factor (TF) binding data via a) the overlap with TF motifs as implemented in FORGE2-TF (https://forge2-tf.altiusinstitute.org/) and b) SNP-specific Contextual Analysis of TF Occupancy (CATO) scores, which provide a complementary line of evidence for TF binding computed from allele-specific TF occupancy data measured by DNase I footprinting[16]. Third, FORGEdb links SNPs to target genes via a) the overlap between SNPs and enhancer-to-promoter looping regions (or other looping regions) using Activity-By-Contact (ABC) data[17] and b) expression quantitative trait locus (eQTL) annotations using large-scale data from GTEx[13] and eQTLGen[14].

To integrate and summarize these annotations, we developed a new scoring system combining all datasets relating to gene regulation: FORGEdb scores. FORGEdb scores are designed to prioritize genetic variants for functional validation. Given the need to ensure that no single dataset would dominate or skew this scoring system, we chose a points-based approach that scored distinct experimental or technological approaches separately. FORGEdb scores are thus computed based on the presence or absence of 5 different lines of evidence for regulatory function:

1. DNase I hotspot, marking accessible chromatin (2 points)
2. Histone mark ChIP-seq broadPeak, marking different regulatory states (2 points)
3. TF motif (1 point) and CATO score, marking potential TF binding (1 point)
4. Activity-by-contact (ABC) interaction, marking gene looping (2 points)
5. Expression quantitative trait locus (eQTL), marking an association with gene expression (2 points)

All of these different features are independently associated with gene regulation and are thus scored separately. FORGEdb scores are calculated by summing the number of points across all lines of evidence present for each SNP and range between 0 and 10. A score of 9 or 10 suggests a large amount of evidence for functional relevance, whereas 0 or 1 indicate a low amount of evidence. For example, SNP rs1421085 shows data regarding eQTLs (including for IRX3, a key target gene[4]), chromatin looping, TF motifs, DNase I hotspots, and histone mark broadPeaks, but does not have a CATO score (**Figure 1**). Together, these data sources provide strong evidence for a regulatory role for this SNP, and the SNP has a FORGEdb score of 9 (**Figure 1**). This high score for rs1421085 is consistent with independent experimental analyses, which have demonstrated a regulatory role for this SNP.[4]



Summary and number of annotations for rs1421085 (FORGEdb score=9):	
RefSeq closest gene data:	1
UCSC genome browser data:	1
CADD v1.6 annotations:	3
GTEx cis-eQTLs:	2
eQTLGen blood cis-eQTLs:	0
ABC contacts:	2
FORGE2-TF motifs:	4
CATO score:	0
FORGE2 consolidated roadmap DNase I hotspots (erc2-DHS):	4
FORGE2 consolidated roadmap H3 histone marks (erc2-H3-all):	46
FORGE2 unconsolidated roadmap DNase I hotspots (erc):	22
FORGE2 blueprint DNase I hotspots (blueprint):	0
FORGE2 ENCODE DNase I hotspots (encode):	22
FORGE2 consolidated roadmap chromatin states (erc2-15state):	127

Figure 1: Example FORGEdb results for rs1421085. For this SNP, there is evidence for eQTL associations (with IRX3, shown to be a key target gene[4]), chromatin looping (ABC interactions), overlap with significant TF motifs, DNase I hotspot overlap, as well as overlap with histone mark broadPeaks. The only regulatory dataset that this SNP does not have evidence for is for CATO score (1 point). The resulting FORGEdb score for rs1421085 is therefore 9 = 2 (eQTL) + 2 (ABC) + 1 (TF motif) + 2 (DNase I hotspot) + 2 (histone mark ChIP-seq). Independent experimental analyses by Claussnitzer *et al.* have demonstrated a regulatory role for this SNP in the regulation of white vs. beige adipocyte proliferation via IRX3/IRX5.

In addition to regulatory datasets, FORGEdb also has datasets that aid interpretation of protein-coding changes (CADD scores, which measure the deleteriousness of SNPs using experimental data and simulated mutations[18]).

Conclusion

In summary, FORGEdb is a new web-based tool to aid genetic variant interpretation and prioritization for experimental analysis. FORGEdb includes a number of features from novel technologies not available in commonly used webtools (**Table 1**), providing a more comprehensive analysis of potential regulatory function. All of these features are accessible via a simple, easy-to-use search engine that can be found at https://forge2.altiusinstitute.org/files/forgedb.html. Annotations from FORGEdb can be accessed from https://ldlink.nci.nih.gov/?tab=ldproxy, https://ldlink.nci.nih.gov/?tab=ldassoc, https://ldlink.nci.nih.gov/?tab=ldmatrix, and https://forge2.altiusinstitute.org/.

Table 1: A comparison of features across FORGEdb, Haploreg and RegulomeDB:

	FORGEdb	Haploreg	RegulomeDB
Roadmap Chromatin states	yes	yes	yes
TF motifs	yes	yes	yes
SNP Scoring system	yes	no	yes
Roadmap DNase-seq	yes	yes	no
Roadmap H3 histone mark data	yes	yes	no
3D genomic data (ABC Hi-C-based data)	yes	no	no
CADD v1.6 data	yes	no	no
GTEx v8 data	yes	no	no
QTLGen data	yes	no	no
BLUEPRINT DNase-seq	yes	no	no
Allele-specific TF binding data (CATO)	yes	no	no

Acknowledgements

We would like to acknowledge the IHEC Integrative analysis project for supporting this research. This study was supported in part by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH. SB acknowledges funding from the Wellcome Trust (218274/Z/19/Z).

References

1. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).

2. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucl. Acids Res.* **42**, D1001–D1006 (2014).

3. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).

4. Claussnitzer, M. *et al.* FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *N. Engl. J. Med.* **373**, 895–907 (2015).

5. Sekar, A. *et al.* Schizophrenia risk from complex variation of complement component 4. *Nature* **530**, 177–183 (2016).
6. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
7. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
8. Stunnenberg, H. G. *et al.* The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell* **167**, 1145–1149 (2016).
9. Ward, L. D. & Kellis, M. HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res.* **44**, D877-881 (2016).
10. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* **22**, 1790–1797 (2012).
11. Tak, Y. G. & Farnham, P. J. Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. *Epigenetics & Chromatin* **8**, 57 (2015).
12. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
13. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
14. Võsa, U. *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat Genet* **53**, 1300–1310 (2021).

15. Breeze, C. E. *et al.* Integrative analysis of 3604 GWAS reveals multiple novel cell type-specific regulatory associations. *Genome Biology* **23**, 13 (2022).

16. Maurano, M. T. *et al.* Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. *Nat. Genet.* **47**, 1393–1401 (2015).

17. Fulco, C. P. *et al.* Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat Genet* **51**, 1664–1669 (2019).

18. Rentzsch, P., Schubach, M., Shendure, J. & Kircher, M. CADD-Splice—improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Medicine* **13**, 31 (2021).