

ARCHANA BHARDWAJ
(POST GENOMIC ERA)
GBIO0002

STRUCTURE OF THIS LECTURE

- **Recap of some concepts**
- **All about Post GWAS**
- **Data bases : Post GWAS**
- **Web servers : Post GWAS**

RECAP OF SOME CONCEPT



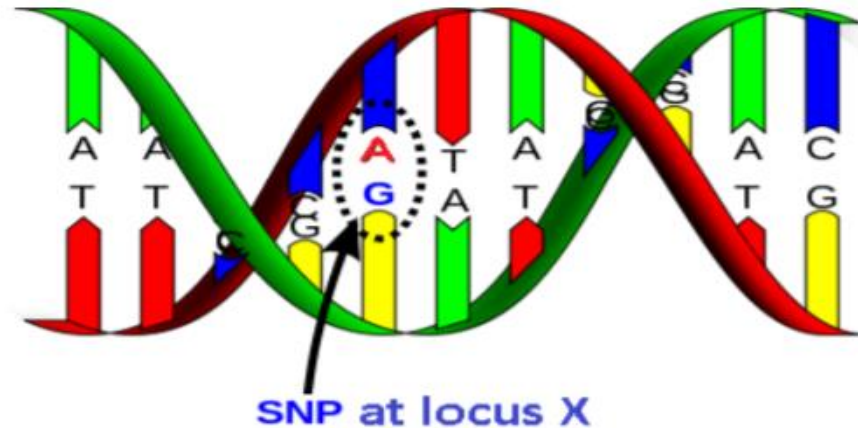
- What are SNPs ?
- What are minor and major alleles ?
- What is LD dis equilibrium ?
- What do you mean by association test ?
- What are manhattan plots ?
- What are q-q plot ?



All about GWAS (Recap)

Important genetic terms

- Given position in the genome (i.e. locus) has several associated alleles (**A** and **G**) which produce genotypes r_A/r_G



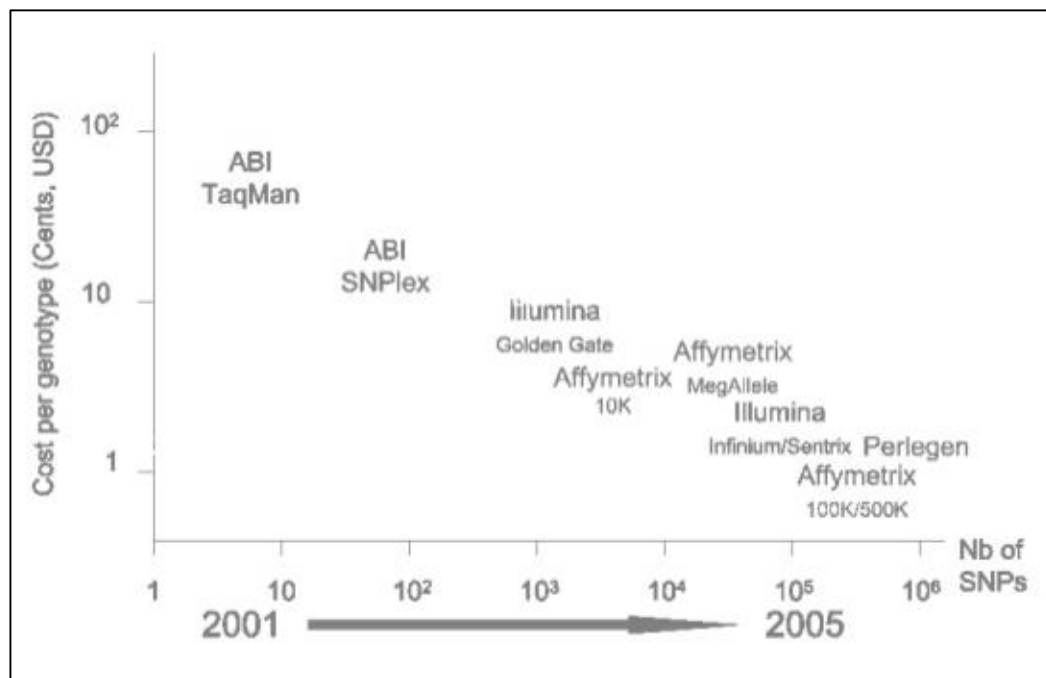
- Haplotypes
 - Combination of alleles at different loci

GWAS visually

- GWAS tries to uncover links between genetic basis of the disease
- Which set of SNPs explain the phenotype?

Genotype	Phenotype
ATGC A GTT	control
TTGC A GTT	control
CTGC A GTT	control
ATGC G GTT	case
TTGC G GTT	case
CTGC C GTT	case

SNP



Running a GWAS: Getting your genotype data

- Select your chip
- Complete your genotyping



The era of hypothesis generating research



EXTENDED PDF FORMAT
SPONSORED BY
More Stem Cell
Characterization
With Less Variation
RD
www.rdgateway.com

Complement Factor H Variant Increases the Risk of Age-Related Macular Degeneration
Jonathan L. Haines *et al.*
Science **308**, 419 (2005);
DOI: 10.1126/science.1110359

This copy is for your personal, non-commercial use only.

➤ Data Preparation

➤ Quality Control

➤ Clustering

➤ GWAS

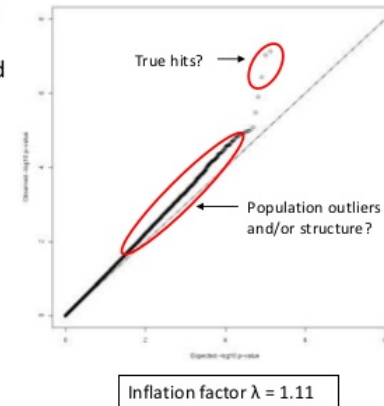
The Problems of population substructure

Devlin and Roeder (1999) used theoretical arguments to propose that with population structure, the distribution of Cochran-Armitage trend tests, genome-wide, is inflated by a constant multiplicative factor λ .

We can estimate the multiplicative inflation factor using the statistic $\lambda = \text{median}(X_i^2)/0.456$.

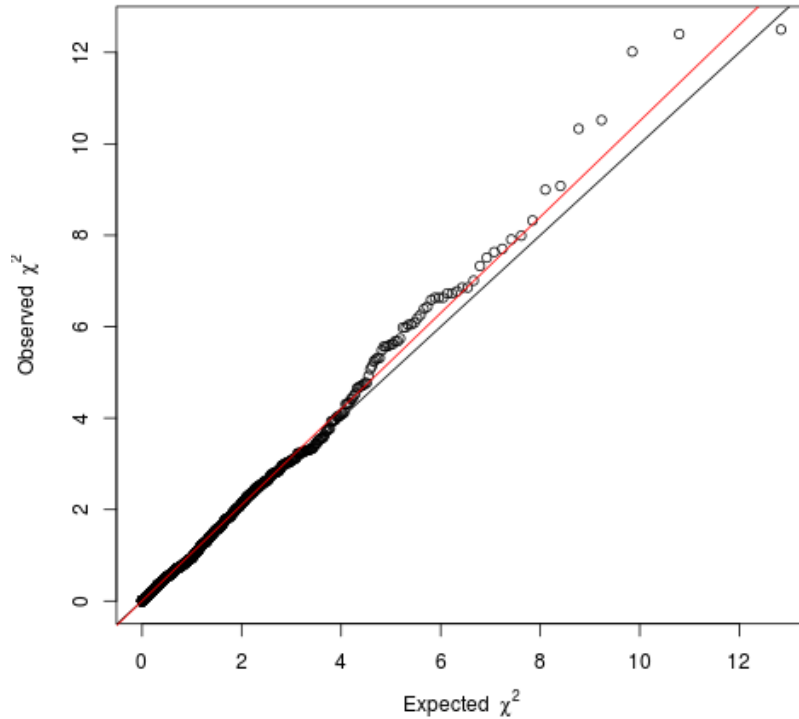
Inflation factor $\lambda > 1$ indicates population structure and/or genotyping error.

We can carry out an adjusted test of association that takes account of any mismatching of cases/controls at any SNP using the statistic X_i^2/λ .



GWAS based on Linkage Disequilibrium (LD)

- LD is the non-random correlation or association of alleles at two loci
- D , D' (normalized), and r^2 are commonly used summary statistics to estimate pairwise LD
- r^2 is preferred in association studies because it is more indicative of how markers might correlate with QTL



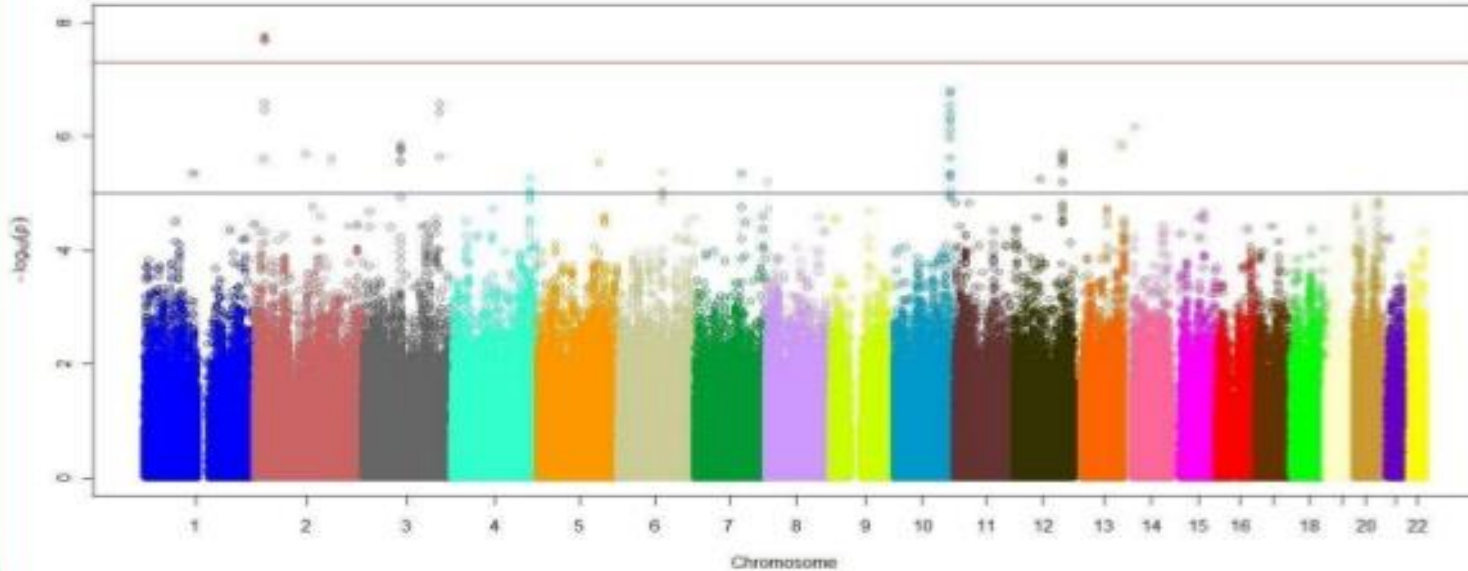
Lambda should range from 0 to 1 (ideal condition).

The resulting Q-Q plot clearly depicts a trend line ($\lambda = 1$, red), overlapping with $x = y$ (black) and a slight deviation in the right tail.

so we can be more confident about our results.

Manhattan (multi chromosome view)

Running a GWAS: Visualize your results



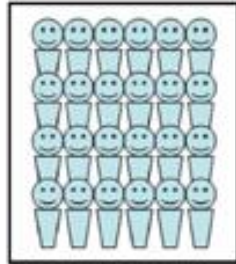
GWAS

Phenotyping

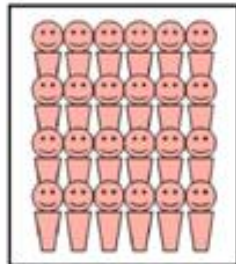
Genotyping

Mapping

Case



Control



DNA

Commercial
array of SNPs



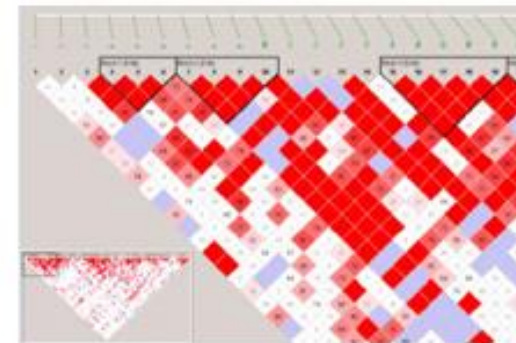
Information

Statistics

$$\begin{aligned}
 W &= |1 - \Phi(\mu_2, 0)| \int_{\Phi^{-1}(\alpha_1/2)}^{\infty} \psi(\mu_1, z_1) dz_1 \\
 &+ \int_{\Phi^{-1}(1-\alpha_1/2)}^{\Phi^{-1}(1-\gamma/2)} \psi(\mu_1, z_1) [1 - \Phi(\mu_2, \Phi^{-1}\{1 - \frac{\gamma}{4[1 - \Phi(z_1)]}\})] dz_1 \\
 &+ \Phi(\mu_2, 0) \int_{\Phi^{-1}(1-\gamma/2)}^{\infty} \psi(\mu_1, z_1) dz_1 + \int_{\Phi^{-1}(1-\alpha_1/2)}^{\Phi^{-1}(1-\gamma/2)} \psi(\mu_1, z_1) \Phi(\mu_2, \Phi^{-1}\{1 - \frac{\gamma}{4[1 - \Phi(z_1)]}\}) dz_1 \\
 &+ |1 - \Phi(\mu_2, 0)| \int_{-\infty}^{\Phi^{-1}(\gamma/2)} \psi(\mu_1, z_1) dz_1 + \int_{\Phi^{-1}(\gamma/2)}^{\Phi^{-1}(\alpha_1/2)} \psi(\mu_1, z_1) [1 - \Phi(\mu_2, \Phi^{-1}\{1 - \frac{\gamma}{4\Phi(z_1)}\})] dz_1 \\
 &= \Phi(\mu_2, 0) \int_{-\infty}^{\Phi^{-1}(\gamma/2)} \psi(\mu_1, z_1) dz_1 + \int_{\Phi^{-1}(\alpha_1/2)}^{\Phi^{-1}(\gamma/2)} \psi(\mu_1, z_1) \Phi(\mu_2, \Phi^{-1}\{1 - \frac{\gamma}{4\Phi(z_1)}\}) dz_1,
 \end{aligned}
 \tag{25}$$

Associated SNP

Chromosome



Linkage disequilibrium block



A guide to genome-wide association analysis and post-analytic interrogation

Eric Reed, Sara Nunez, David Kulp, Jing Qian, Muredach P. Reilly, Andrea S. Foulkes

First published: 06 September 2015 | <https://doi.org/10.1002/sim.6605> | Cited by: 21

Get it@ULiège

Support for this research is provided by NIH/NHLBI R01-HL107196.

SECTIONS



PDF



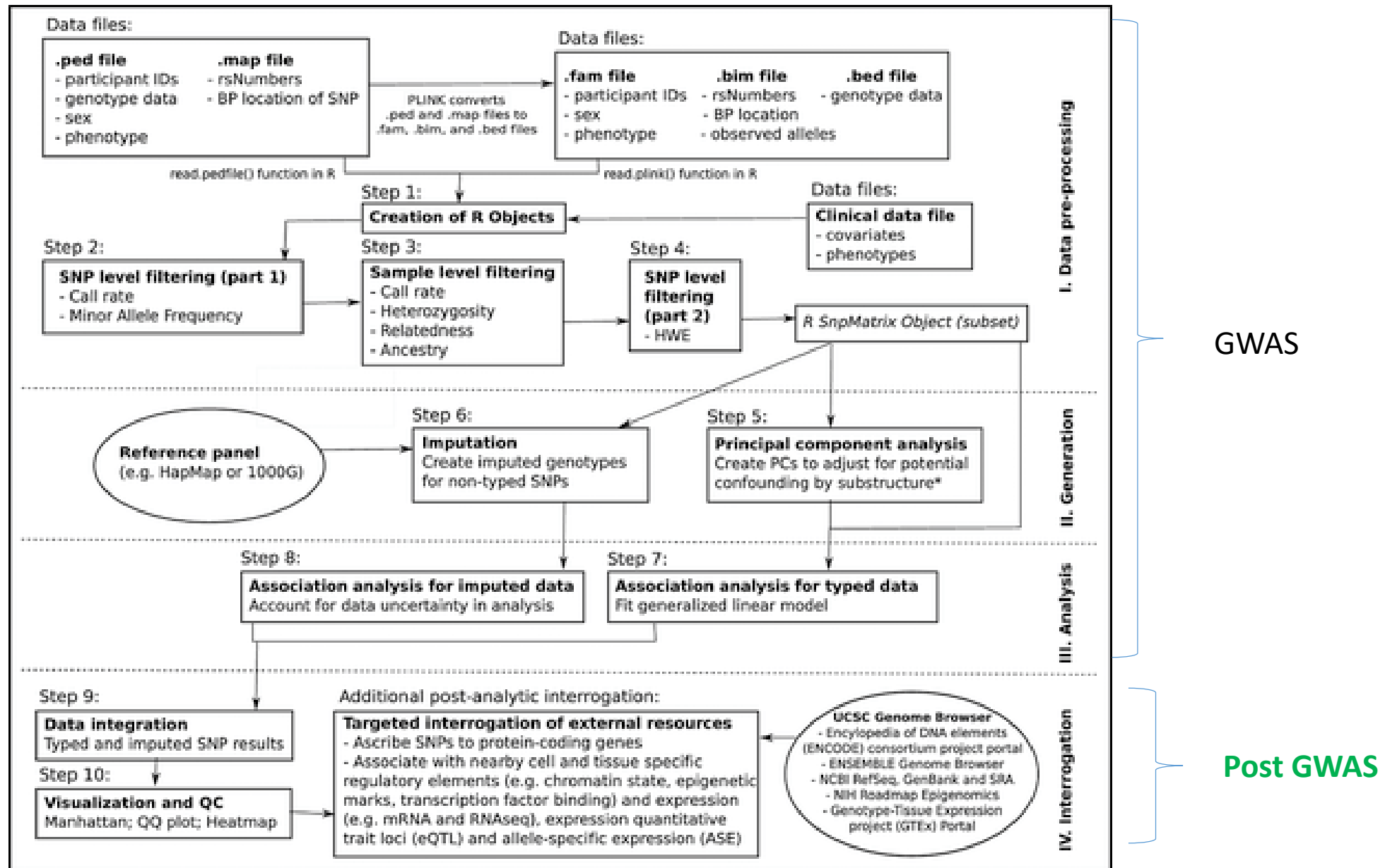
TOOLS

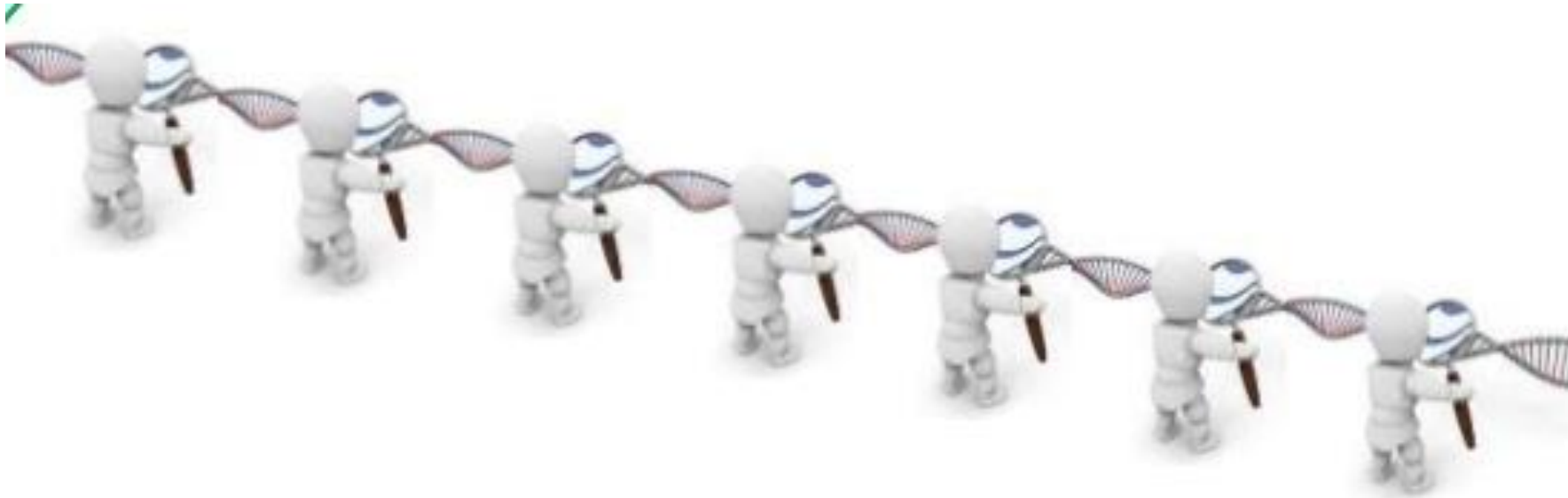


SHARE

Abstract

This tutorial is a learning resource that outlines the basic process and provides specific software tools for implementing a complete genome-wide association analysis. Approaches to post-analytic visualization and interrogation of potentially novel findings are also presented. Applications are illustrated using the free and open-source R statistical computing and graphics software environment, Bioconductor software for bioinformatics and the UCSC Genome Browser. Complete genome-wide association data on 1401 individuals across 861,473 typed single nucleotide polymorphisms from the PennCATH study of coronary artery disease are used for illustration. All data and code, as





All about Post GWAS

Post GWAS : Interpreting SNPs

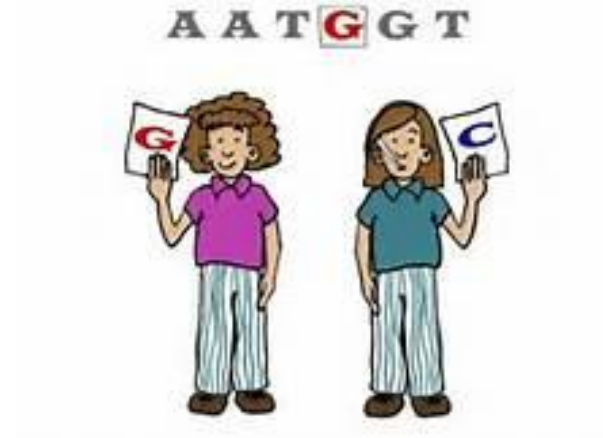
Look at the functionality of your SNP (SNPdoc)

Literature search – can you give biological plausibility?

Other tests: pathway analysis / Gene based tests

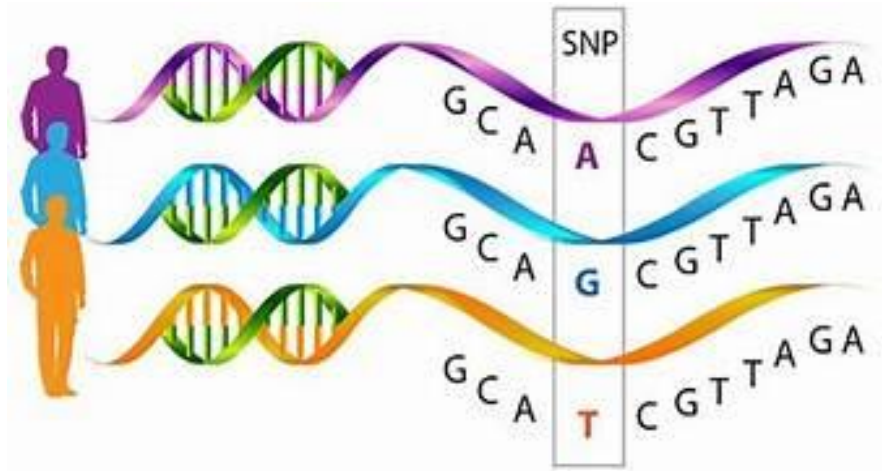
Manual Search = No

Multiple softwares are available

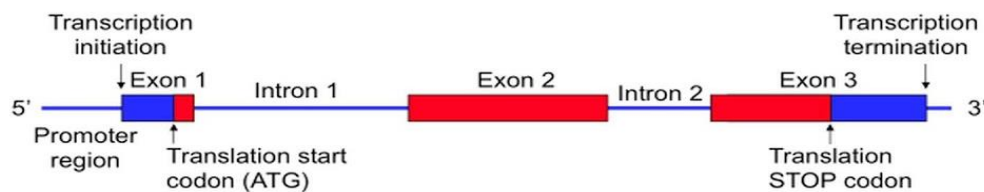


Genomic Positions of SNPs

IMPORTANT FINDING

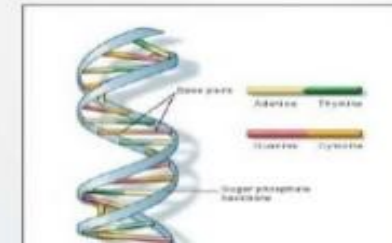


Gene Structure

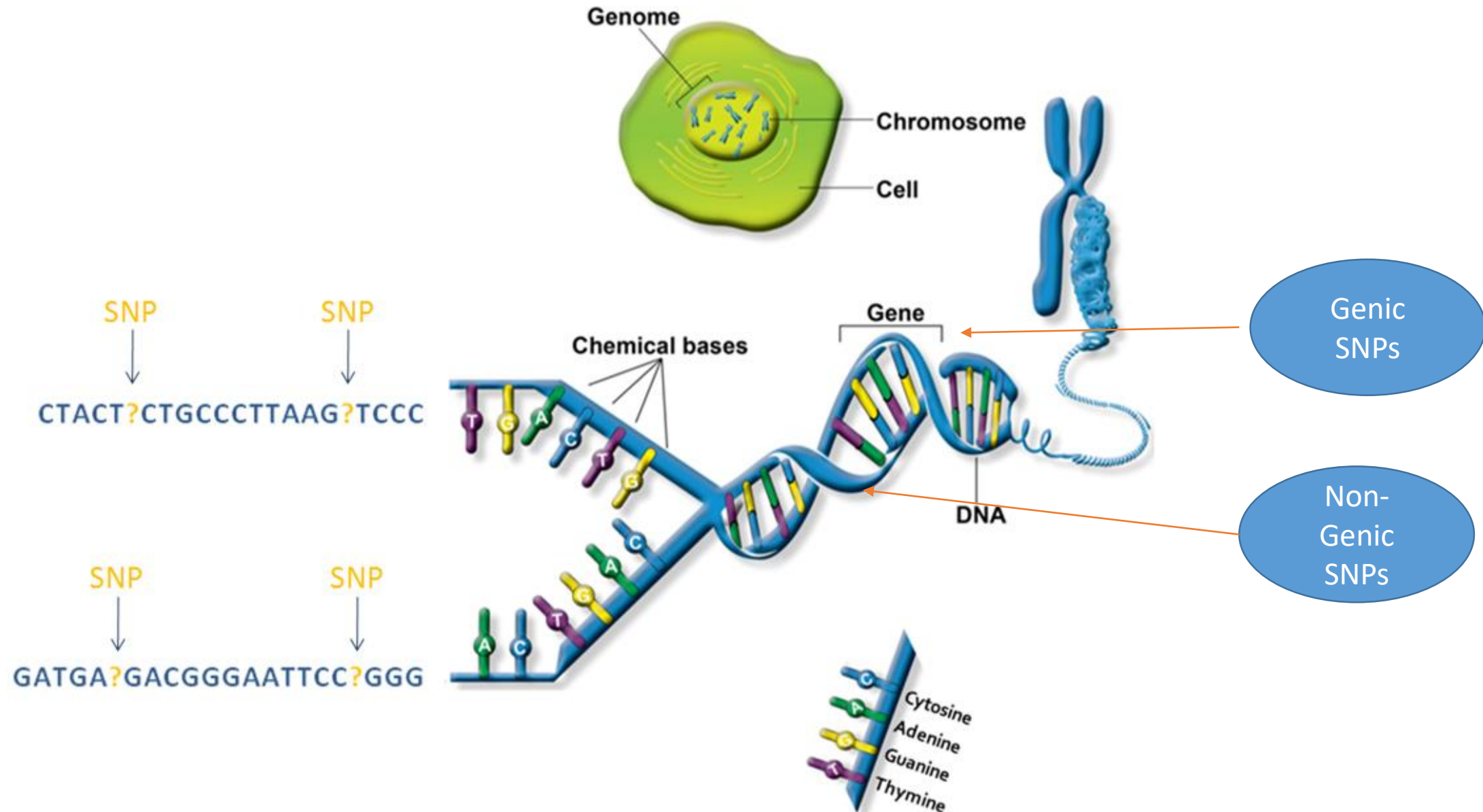


The Basics - Genes

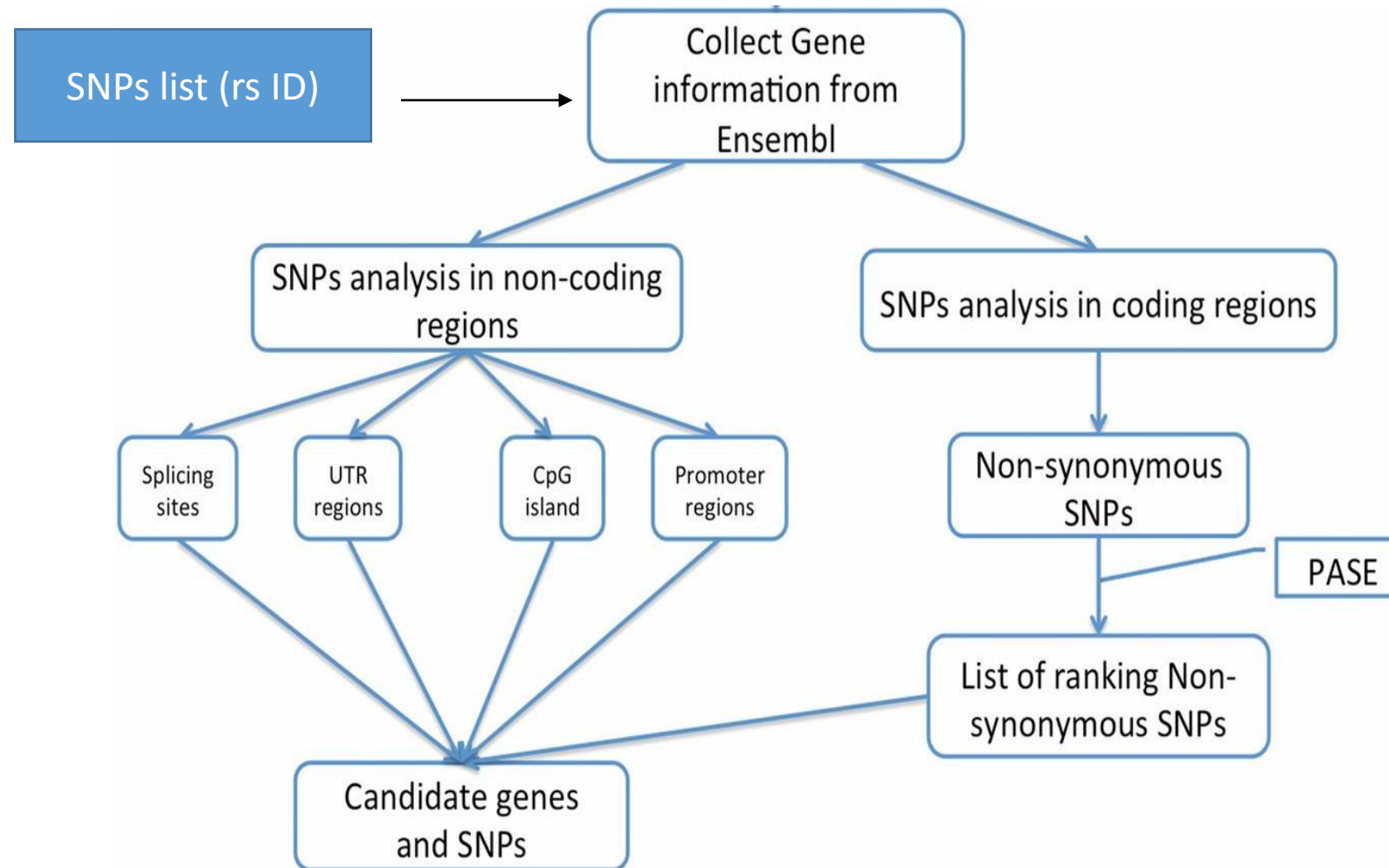
- Segments of DNA that encode instructions to our cells
- Nucleotides link the two strands of our DNA
- These bases are the alphabet of our genetic code



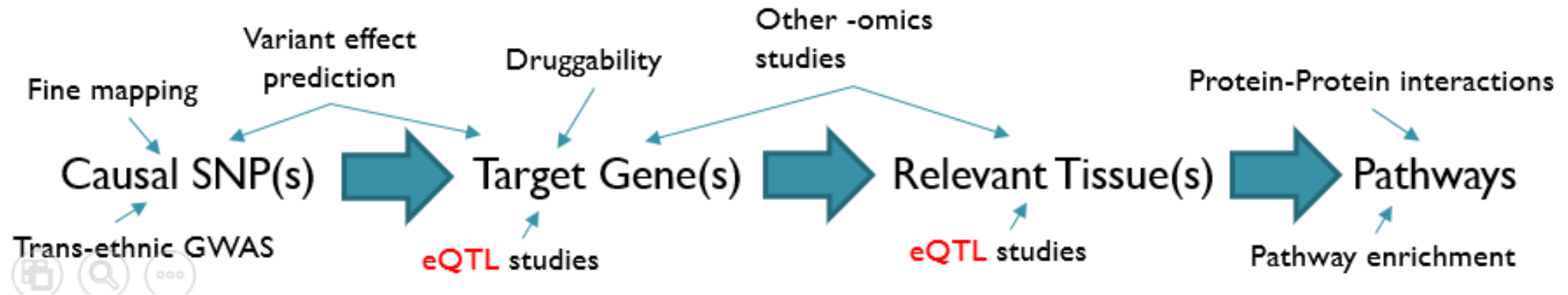
Genomic Positions of SNPs



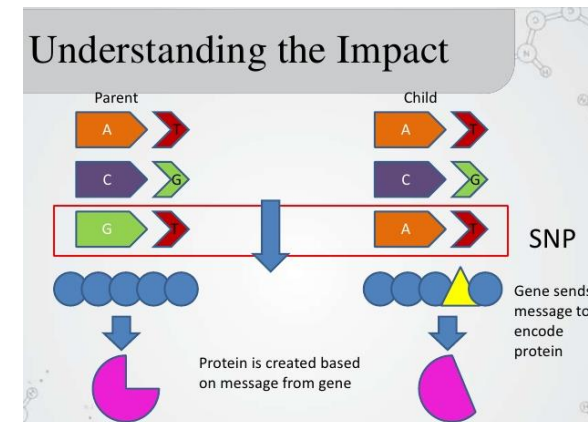
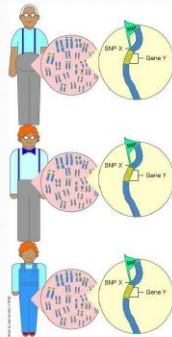
Classification of SNPs (Based on Genomic Position)



Why : From SNPs to Genes



SNPs act as gene markers



Examples: From SNPs to Genes

- **rs6311 and rs6313 are SNPs in the Serotonin 5-HT_{2A} receptor gene on human chromosome 13.**
- **rs3091244 is an example of a triallelic SNP in the CRP gene on human chromosome 1.**
- **rs148649884 and rs138055828 in the FCN1 gene encoding M-ficolin crippled the ligand-binding capability of the recombinant M-ficolin.**

List of Data sources for Post GWAS

Example data types	Select data sources*	UCSC genome browser navigation
<i>DNA level data (non-somatic; genEric to all cells):</i>		
I. Coordinates, e.g.		
(1) SNPs	NCBI dbSNP[a], ENSEMBL[b]	Variation: Common SNPs(141)
(2) Insertions and deletions (INDELs)		
(3) Copy number variants (CNVs)		
II. Gene elements, e.g.		
(1) Protein-coding genes	NCBI RefSeq[c], NCBI GenBank[d], ENSEMBL[b]	Gene and Gene Predictions: UCSC Genes
(2) Non-protein-coding genes	NCBI RefSeq[c], NCBI GenBank[d], ENSEMBL[b]	Gene and Gene Predictions: UCSC Genes
<i>Cell and tissue-specific regulation:</i>		
III. Chromatin state, e.g.		
(1) DNA hypersensitivity (DNase-Seq)	ENCODE[e], ENSEMBL[b]	Regulation: ENCODE Regulation
(2) FAIRE sequencing	ENCODE[e], ENSEMBL[b]	Regulation: ENC DNase/FAIRE
IV. Epigenetic marks, e.g.		
(1) Methylation promoter marks	ENCODE[e], NIH Roadmap Epigenomics[f]	Regulation: ENCODE Regulation
(2) Methylation enhancer marks	ENCODE[e], NIH Roadmap Epigenomics[f]	Regulation: ENCODE Regulation
(3) Acetylation marks (e.g. #H3K27Ac histone mark)	ENCODE[e], NIH Roadmap Epigenomics[f]	Regulation: ENCODE Regulation
V. Transcription factor binding, e.g.		
(1) ChIPSeq data	ENCODE[e], ENSEMBL[b], custom	Regulation: ENCODE Regulation
<i>Cell and tissue-specific expression:</i>		
VI. RNA expression, e.g.		
(1) historic mRNA	NCBI GenBank[d]	mRNA and EST: Human mRNAs
(2) genome-wide cell-specific RNA data (e.g. RNAseq)	ENCODE[e], GTex Portal[g], NCBI SRA[h]	Expression: ENC RNA-seq
VII. SNP-mRNA association, e.g.		
(1) Expression quantitative trait locus (eQTL)	GTex Portal[g], custom	N/A
(2) Allelic imbalance (AI); allele specific expression (ASE)	GTex Portal[g], custom	N/A
<i>Biomarkers endophenotype:</i>		
VIII. Other -omics data, e.g.		
(1) Proteomic (e.g. pQTLs)	UniProtKB[i]	N/A
(2) Metabolomic	HMDB[j]	N/A

Post GWAS : Terminology

- Indels
- Epigenetic markers
- eQTL

SNPs could be linked to epigenetic markers and regulate the expression of other genes

What are indels ?

- Indels can be contrasted with a point mutation.
- An indel inserts and deletes nucleotides from a sequence, while a point mutation is a form of substitution that replaces one of the nucleotides without changing the overall number in the DNA.

wild-type sequence

ATCTTCAGCCATAAAAGATGAAGTT

3 bp deletion

ATCTTCAGCCAAAGATGAAGTT

4 bp insertion (orange)

ATCTTCAGCCATATGTGAAAGATGAAGTT

Identify the insertion or deletion in following sequences

Wild CAT CAT CAT CAT CAT CAT

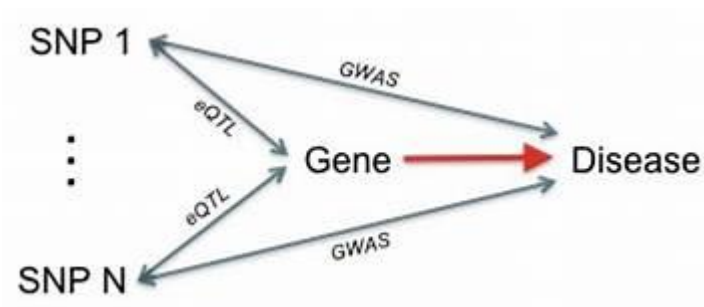
Mutation CAT CAT CAT ACA TCA TCA

Wild CAT CAT CAT CAT CAT CAT

Mutation CAC ATC ATA CAT CAT CA

eQTL

- SNPs can be located in gene regions or intergenic ones.
- eQTL= expression Quantitative Trait Locus.
- This is a genomic locus that influences the expression level of mRNA (how much a gene is transcribed).
- This locus can be physically located close to the gene that gets regulated, or far away (even on another chromosome).



Databases and Softwares

Data source/tool	Used for	Links	Last update	Reference
1000 Genome Project Phase 3	Reference panel used to compute r^2 and MAF.	Info: http://www.internationalgenome.org/ Data: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/	27 May 2019	1000 Genomes Project Consortium, et al. 2015. A global reference for human genetic variation. <i>Nature</i> 526, 86-74. PMID:26432245
PLINK v1.9	Used to compute r^2 and MAF.	Info and download: https://www.cog-genomics.org/plink2	27 May 2019	Purcell, S., et al. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. <i>Am. J. Hum. Genet.</i> 81, 559-575. PMID:17701901
MAGMA v1.07	Used for gene analysis and gene-set analysis.	Info and download: https://ctg.cncr.nl/software/magma	13 Feb 2019	de Leeuw, C., et al. 2015. MAGMA: Generalized gene-set analysis of GWAS data. <i>PLoS Comput. Biol.</i> 11, DOI:10.1371/journal.pcbi.1004219. PMID:26044016
ANNOVAR	A variant annotation tool used to obtain functional consequences of SNPs on gene functions.	Info and download: http://annovar.openbioinformatics.org/en/latest/	5 Dec 2016	Wang, K., Li, M. and Hakonarson, H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. <i>Nucleic Acids Res.</i> 38 e164 PMID:20801885
CADD v1.4	A deleterious score of variants computed by integrating 83 functional annotations. The higher the score, the more deleterious.	Info: http://cadd.gs.washington.edu/ Data: http://cadd.gs.washington.edu/download	27 May 2019	Kicler, M., et al. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. <i>Nat. Genet.</i> 46, 310-315. PMID:24487279
RegulomeDB v1.1	A categorical score to guide interpretation of regulatory variants.	Info: http://regulomedb.org/index Data: http://regulomedb.org/downloads/RegulomeDB.dbSNP141.txt.gz	5 Dec 2016	Boyle, AP., et al. 2012. Annotation of functional variation in personal genomes using RegulomeDB. <i>Genome Res.</i> 22, 1790-7. PMID:22855989
15-core chromatin state	Chromatin state for 127 epigenomes was learned by ChromHMM derived from 6 chromatin markers (H3K4me3, H3K4me1, H3K36me3, H3K27me3, H3K9me3).	Info: http://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html Data: http://egg2.wustl.edu/roadmap/data/byFileType/chromHMMSegmentations/ChmmModels/coreMarks/jointModel/final/all.mnemonics.bedFiles.gz	5 Dec 2016	Roadmap Epigenomics Consortium, et al. 2015. Integrative analysis of 111 reference human epigenomes. <i>Nature</i> 518, 317-330. PMID:25893583 Ernst, J. and Kellis, M. 2012. ChromHMM: automating chromatin-state discovery and characterization. <i>Nat. Methods</i> 28, 215-8. PMID:22373907
GTEx v6/v7/v8	eQTLs and gene expression used in the pipeline were obtained from GTEx.	Info and data: http://www.gtexportal.org/home/	14 Oct 2019	GTEx Consortium. 2015. Human genomics. The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. <i>Science</i> 348, 648-60. PMID:25954001 GTEx Consortium. 2017. Genetic effects on gene expression across human tissues. <i>Nature</i> 550, 204-213. PMID:29022597 Aguet, et al. 2019. The GTEx consortium atlas of genetic regulatory effects across human tissues. <i>bioRxiv</i> . doi: https://doi.org/10.1101/787903 . https://doi.org/10.1101/787903

Blood eQTL Browser	eQTLs of blood cells. Only cis-eQTLs with FDR ≤ 0.05 are available in FUMA.	Info and data: http://genenetwork.nl/bloodeqtlbrowser/	17 January 2017	Westra et al. 2013. Systematic identification of trans eQTLs as putative drivers of known disease associations. <i>Nat. Genet.</i> 45, 1238-1243. PMID:24013639
BIOS QTL browser	eQTLs of blood cells in Dutch population. Only cis-eQTLs (gene-level) with FDR ≤ 0.05 are available in FUMA.	Info and data: http://genenetwork.nl/biosqtlbrowser/	17 January 2017	Zhemakova et al. 2017. Identification of context-dependent expression quantitative trait loci in whole blood. <i>Nat. Genet.</i> 49, 138-145. PMID:27918533
BRAINEAC	eQTLs of 10 brain regions. Cis-eQTLs with nominal P-value < 0.05 are available in FUMA.	Info and data: http://www.braineac.org/	28 January 2017	Ramasamy et al. 2014. Genetic variability in the regulation of gene expression in ten regions of the human brain. <i>Nat. Neurosci.</i> 17, 1418-1428. PMID:27918533
MuTHER	eQTLs in Adipose, LCL and Skin samples (only cis eQTLs).	Info: http://www.mutther.ac.uk/ Data: http://www.mutther.ac.uk/Data.html	21 January 2018	Grundberg et al. 2012. Mapping cis and trans regulatory effects across multiple tissues in twins. <i>Nat. Genet.</i> 44, 1084-1089. PMID:22841192
xQTLServer	eQTLs in dorsolateral prefrontal cortex samples.	Info and data: http://mostafavilab.stat.ubc.ca/xqtl/	21 January 2018	Ng et al. 2017. An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. <i>Nat. Neurosci.</i> 20, 1418-1428. PMID:28895584
CommonMind Consortium	eQTLs in brain samples. Both cis and trans eQTLs are available	Info and data: https://www.synapse.org/#/Synapse:syn5585484	21 January 2018	Fromer et al. 2016. Gene expression elucidates functional impact of polygenic risk for schizophrenia. <i>Nat. Neurosci.</i> 19, 1442-1453. PMID:27883389
eQTLGen	Meta-analysis of cis and trans eQTLs based on 37 data sets (in total of 31,884 individuals).	Info: http://www.eqtngen.org/index.html Data: https://molgenis28.gsc.rug.nl/downloads/eqtngen/cis-eqt/cis-eqtls_full_20180905.txt.gz , https://molgenis28.gsc.rug.nl/downloads/eqtngen/trans-eqt/trans-eqtls_significant_20181017.txt.gz	20 Oct 2018	Vosa et al. 2018. Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. <i>bioRxiv</i> https://doi.org/10.1101/144737
DICE	eQTLs of 15 types of immune cells.	Info: https://dice-database.org/landing Data: https://dice-database.org/downloads	27 May 2019	Schmiedel et al. 2018. Impact of genetic polymorphisms on human immune cell gene expression. <i>Cell</i> 175, 1701-1715 e18. PMID:30449822
van der Wijst et al. scRNA eQTLs	eQTLs based on scRNA-seq of 9 cell types.	Info and data: https://molgenis28.target.rug.nl/downloads/scrna-seq/	27 May 2019	van der Wijst et al. 2018. Single-cell RNA sequencing identifies cell-type-specific eQTLs and co-expression QTLs. <i>Nat. Genet.</i> 50, 493-497. PMID:29810479
PsychENCODE	SNP annotations (enhancer, H3K27ac markers), eQTLs and HiC based enhancer-promoter interactions.	Info and data: http://resource.psychencode.org/	27 May 2019	Wang et al. 2018. Comprehensive functional genomic resource and integrative model for the human brain. <i>Science</i> 14, eaat8484. PMID:30545857
FANTOM5	SNP annotations (enhancer and promoter) and enhancer-promoter correlations.	Info: http://fantom.gsc.riken.jp/5/ Data: http://fantom.gsc.riken.jp/5/data/ , http://slidebase.binf.ku.dk/human_enhancers/presets	27 May 2019	Andersson et al. 2014. An atlas of active enhancers across human cell types and tissues. <i>Nature</i> 507, 455-461. PMID:24870763 FANTOM Consortium. A promoter-level mammalian expression atlas. <i>Nature</i> 507, 462-470. PMID:24870764

Databases and Softwares

BrainSpan	Gene expression data of developmental brain samples.	Info and data: http://www.brainspan.org/static/download	31 January 2018	Kang et al. 2011. Spatio-temporal transcriptome of the human brain. <i>Nature</i> 478, 483-489. PMID:22031440
GSE87112 (Hi-C)	Hi-C data (significant loops) of 21 tissue/cell types. Pre-processed data (output of Fit-Hi-C) is used in FUMA.	Info and data: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE87112	9 May 2017	Schmitt, A.D. et al. 2016. A compendium of chromatin contact maps reveals spatially active regions in the human genome. <i>Cell Rep.</i> 17, 2042-2059. PMID:27851967
Giusti-Rodriguez et al. 2019 (Hi-C)	Hi-C data (significant loops) of adult and fetal cortex. Only significant loops after Bonferroni correction ($P_{bon} < 0.001$) are available.	The data was kindly shared by Patric F. Sullivan.	13 Feb 2019	Giusti-Rodriguez, P. et al. 2019. Using three-dimensional regulatory chromatin interactions from adult and fetal cortex to interpret genetic results for psychiatric disorders and cognitive traits. <i>bioRxiv</i> . https://doi.org/10.1101/406330
Enhancer and promoter regions	Predicted enhancer and promoter regions (including dyadic) from Roadmap Epigenomics Projects. 111 epigenomes are available.	Info: http://egg2.wustl.edu/roadmap/web_portal/DNase_reg.html Data: http://egg2.wustl.edu/roadmap/data/byDataType/dnase/	9 May 2017	Roadmap Epigenomics Consortium, et al. 2015. Integrative analysis of 111 reference human epigenomes. <i>Nature</i> 518, 317-330. PMID:25803583 Ernst, J. and Kellis, M. 2012. ChromHMM: automating chromatin-state discovery and characterization. <i>Nat. Methods</i> 28, 215-8. PMID:22373907
MsigDB v7.0	Collection of publicly available gene sets. Data sets include e.g. KEGG, Reactome, BioCarta, GO terms and so on.	Info and data: http://software.broadinstitute.org/gsea/msigdb	14 Oct 2019	Liberzon, A. et al. 2011. Molecular signatures database (MSigDB) 3.0. <i>Bioinformatics</i> 27, 1739-40. PMID:21546383
WikiPathways v20191010	The curated biological pathways.	Info: http://wikipathways.org/index.php/WikiPathways Data: http://data.wikipathways.org/20181110/gmt/wikipathways-20181110-gmt-Homo_sapiens.gmt	14 Oct 2019	Kutmon, M., et al. 2016. WikiPathways: capturing the full diversity of pathway knowledge. <i>Nucleic Acids Res.</i> 44, 488-494. PMID:26481357
GWAS-catalog e98 2019-09-24	A database of reported SNP-trait associations.	Info: https://www.ebi.ac.uk/gwas/ Data: https://www.ebi.ac.uk/gwas/downloads	14 Oct 2019	MacArthur, J., et al. 2016. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). <i>Nucleic Acids Res.</i> pii:gkw1133. PMID:27899670
DrugBank v5.1.4	Targeted genes (protein) of drugs in DrugBank was obtained to assign drug ID for input genes.	Info: https://www.ncbi.nlm.nih.gov/pubmed/27899670 Data: https://www.drugbank.ca/releases/latest#protein-identifiers	14 Oct 2019	Wishart, D.S., et al. 2008. DrugBank: a knowledgebase for drugs, drug actions and drug targets. <i>Nucleic Acids Res.</i> 36, D901-8. PMID:18048412
pLI	A gene score annotated to prioritized genes. The score is the probability of being loss-of-function intolerance.	Info: http://exac.broadinstitute.org/ Data: ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3.1/functional_gene_constraint	27 April 2017	Lek, M. et al. 2016. Analyses of protein-coding genetic variation in 60,708 humans. <i>Nature</i> 536, 285-291. PMID:27535533
ncRVIS	A gene score annotated to prioritized genes. The score is the non-coding residual variation intolerance score.	Info: http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1005492 Data: http://journals.plos.org/plosgenetics/article/file?type=supplemental&id=info:doi/10.1371/journal.pgen.1005492.s011	27 April 2017	Petrovski, S. et al. 2015. The intolerance of regulatory sequence to genetic variation predict gene dosage sensitivity. <i>PLOS Genet.</i> 11, e1005492. PMID:26332131

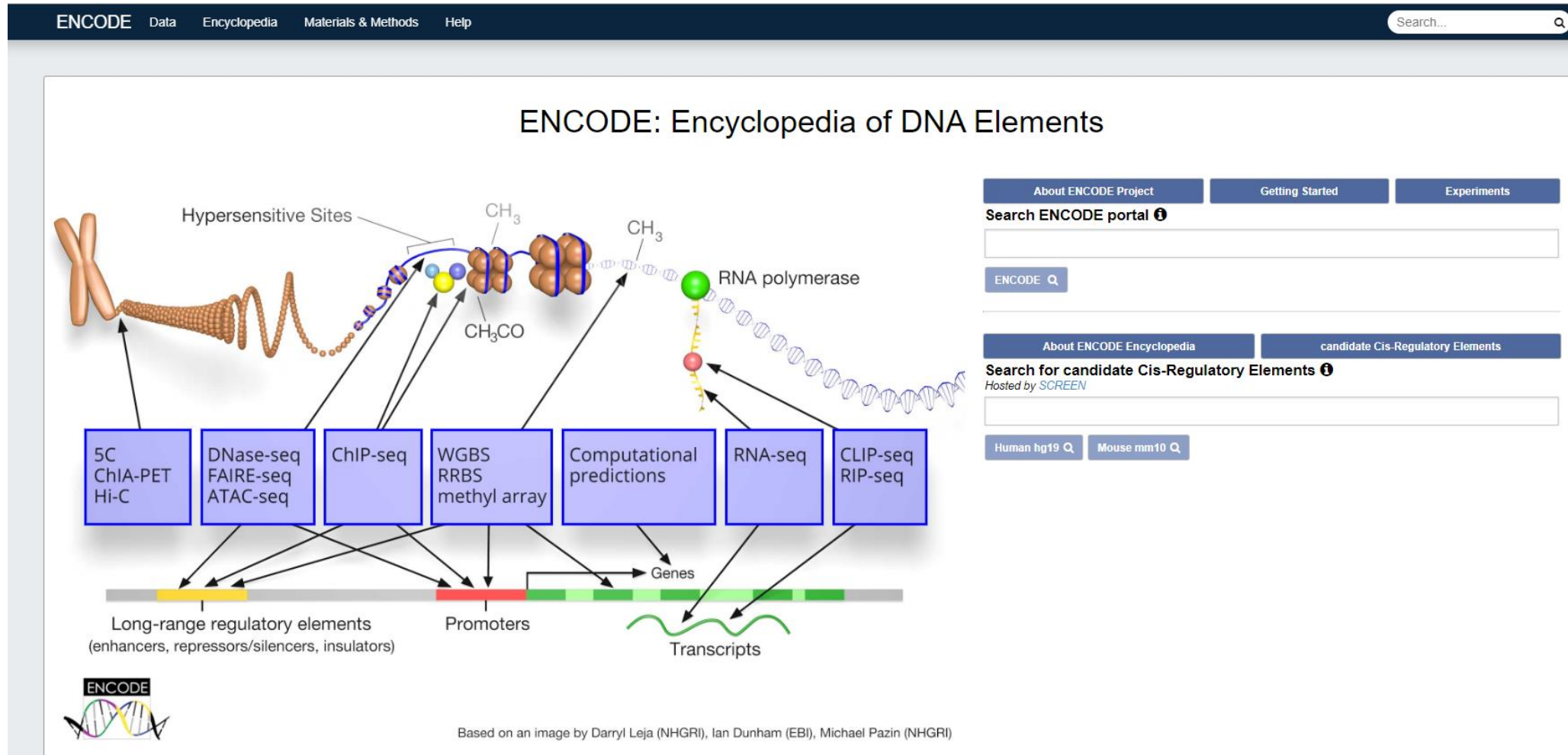
Data bases and web servers

Let us discuss :

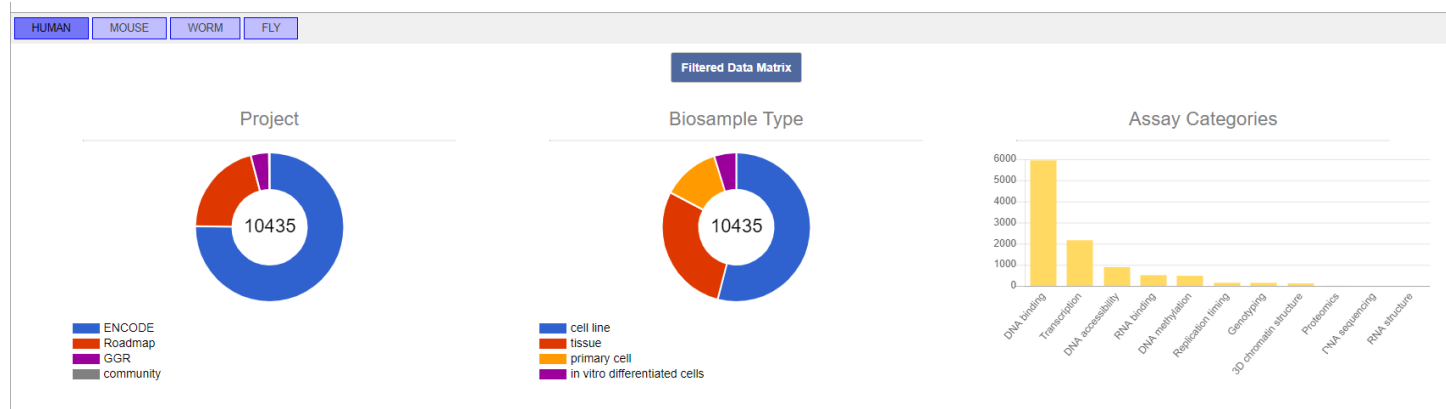
- **ENCODE**
- **HelgoDB**
- **RegulomeDB**
- **UniprotKB**
- **ENSEMBL**
- **FUMA**

ENCODE: Encyclopedia of DNA Elements

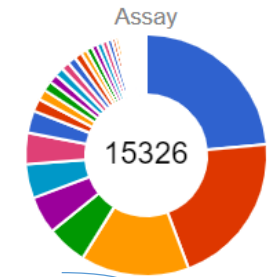
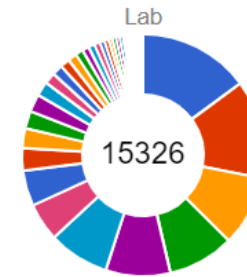
<https://www.encodeproject.org/>



Encode : Data structures



Most of data represents



Multiple resources

- Michael Snyder, Stanford
- Bradley Bernstein, Broad
- John Stamatoyannopoulos, UW
- Richard Myers, HAIB
- Bing Ren, UCSD
- Kevin White, UChicago
- Brenton Graveley, UConn
- Thomas Gingeras, CSHL
- Gene Yeo, UCSD
- Joseph Costello, UCSF
- Valerie Reinke, Yale
- Susan Celniker, LBNL
- Tim Reddy, Duke
- Ali Mortazavi, UCI
- Robert Waterston, UW
- Barbara Wold, Caltech
- Joe Ecker, Salk
- Chris Burge, MIT
- Gregory Crawford, Duke
- Ross Hardison, PennState
- Peggy Farnham, USC
- Xiang-Dong Fu, UCSD

GBIO002 AB 2019

Multiple platform

- TF ChIP-seq
- Histone ChIP-seq
- Control ChIP-seq
- DNase-seq
- polyA plus RNA-seq
- total RNA-seq
- shRNA RNA-seq
- eCLIP
- DNAme array
- small RNA-seq
- WGBS
- microRNA-seq
- ATAC-seq
- RNA microarray
- RAMPAGE
- RNA Bind-n-Seq
- genotyping array
- CAGE
- microRNA counts
- siRNA RNA-seq
- Repli-seq
- RRBS

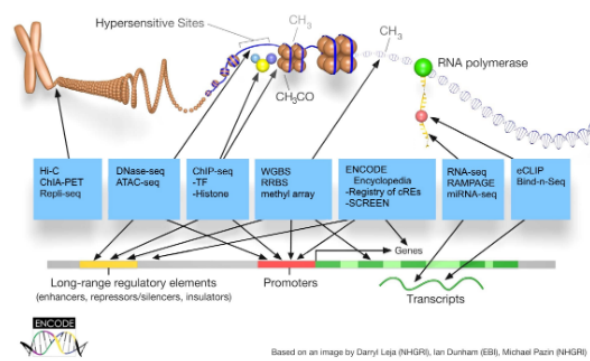
Let us use Encode

Go to link <http://screen.encodeproject.org/>

Enter snp id : *rs4846913*

SCREEN: Search Candidate cis-Regulatory Elements by ENCODE

[Overview](#) [About](#) [Tutorials](#) [Downloads](#) [Versions](#)



The diagram illustrates the genomic architecture of a gene. At the top, a DNA double helix is shown with various regulatory features: Hypersensitive Sites (orange), CH₃CO (yellow), CH₃ (green), and RNA polymerase (blue). Below the DNA, several data tracks are shown: Hi-C ChIA-PET Repli-seq, DNase-seq ATAC-seq, ChIP-seq -TF -Histone, WGBS rRBS methyl array, ENCODE Encyclopedia Registry of cREs -SCREEN, RNA-seq RAMPAGE mRNA-seq, and eCLIP Bind-n-Seq. These tracks are linked to genomic features: Long-range regulatory elements (enhancers, repressors/silencers, insulators), Promoters, Genes, and Transcripts. The ENCODE logo is also present.

SCREEN is a web interface for searching and visualizing the Registry of candidate cis-Regulatory Elements (ccREs) derived from [ENCODE data](#). The Registry contains 1.31M human ccREs in hg19 and 0.43M mouse ccREs in mm10, with orthologous ccREs cross-referenced. SCREEN presents the data that support biochemical activities of the ccREs and the expression of nearby genes in specific cell and tissue types.

You may launch SCREEN using the search box below or browse a curated list of SNPs from the NHGRI-EBI Genome Wide Association Study (GWAS) catalog to annotate genetic variants using ccREs. [Browse GWAS](#)

Enter a gene name or alias, a SNP rsID, a ccRE accession, or a genomic region in the form chr:start-end. You may also enter a cell type name to filter results.
Examples: "K562 chr11:5226493-5403124", "SOX4", "rs4846913", "EH37E0204974"

[Search Human \(hg19\)](#) [Search Mouse \(mm10\)](#)

© 2017 Weng Lab @ UMass Med, ENCODE Data Analysis Center

Click

new class x watanabe x Functional x Table 1 Fe x Functional x Functional x Functional x MAGMA: C x W A guide to x encode da x SCREEN h: x encode da x PPT - Dec x Settings - x + -

Not secure | screen.encodeproject.org/search/?q=rs4846913&assembly=hg19&uid=bac162f5-b32a-4f92-85ac-600b27d016bb

SCREEN hg19

chr1:230294714-230294715

Search

Biosamples

TSV

Search:

	cell type	tissue
<input type="radio"/>	A172	brain
<input type="radio"/>	A549	lung
<input type="radio"/>	A549 treated with dexamethasone	lung
<input type="radio"/>	A549 treated with ethanol	lung
<input type="radio"/>	A673	muscle
<input type="radio"/>	ACC112	salivary glands
<input type="radio"/>	adipocyte	adipose
<input type="radio"/>	adipose derived mesenchymal stem cell in vitro differentiated cells	stem cell
<input type="radio"/>	adrenal gland female adult (51 years)	adrenal
<input type="radio"/>	adrenal gland female fetal (108 days)	adrenal


Total: 622

« < 1 2 3 ... 63 > »

Chromosome

chr1

Coordinates: chr1:230294714-230294715



230294714 - 230294715

Maximum across cell types

ccRE

Search Results

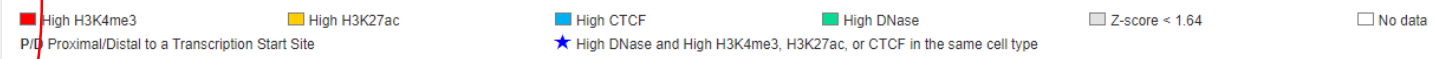
Bed

Upload

Candidate cis-Regulatory Elements (ccREs) that meet your search criteria are listed in the table below.

- Click a ccRE accession to view details about the ccRE, including top tissues, nearby genomic features, etc.
- Click a gene ID to view the expression profile of the gene.

Search:

accession	DNase	H3K4me3	H3K27ac	CTCF	chr	start	length	experimental evidence	nearest genes: protein-coding / all	cart	genome browsers
<input checked="" type="radio"/> EH37E0145522 ★ D 	3.48	2.13	4.09	1.20	chr1	230,294,315	813	--	pc: GALNT2, PGBD5, COG2 all: GALNT2, RP5-956O18.2, BX323860.1		UCSC

Add all to cart Clear cart Download bed Download JSON found 1 results

Select this row

download.png

TF_targets_FUMA_....png

gtex_v8_ts_genera....png

expHeat_FUMA_g_....png

lociPlot_FUMA_jo....png

snpAnnotPlot_FU....png

manhattan_FUMA_....pdf

Show all

12:40
04/11/2019

H3K4me3 Z-scores ⓘ

TSV

Tri methylation (me3): Chromatin markers

cell type	H3K4me3 and DNase	H3K4me3 only
OCI-LY1	--	2.13
HepG2	2.71	2.11
mid-neurogenesis radial glial cells derived from H9 stably expressing fusion protein	--	1.96
Caco-2	2.14	1.94
BE2C	2.31	1.87
radial glial cell derived from H9 stably expressing fusion protein	--	1.83
neuroepithelial stem cell derived from H9 stably expressing fusion protein	--	1.83
skeletal muscle male adult (54 years)	--	1.82
stomach smooth muscle female adult (84 years)	--	1.80
germinal matrix male fetal (20 weeks)	--	1.70

Total: 210

Chromatin markers

CTCF Z-scores ⓘ

TSV

cell type	CTCF and DNase	CTCF only
BE2C	1.97	1.20
H54	--	1.17
MCF-7 treated with estradiol	1.74	1.14
HGPS cell	--	1.13
skin fibroblast female	0.51	0.97
epithelial cell of proximal tubule	1.94	0.94
spleen adult	--	0.94
GM19240	--	0.92
GM12874	--	0.91
GM10266	--	0.89

Total: 101

Acetylation (AC) Chromatin markers

H3K27ac Z-scores ⓘ

TSV

cell type	H3K27ac and DNase	H3K27ac only
KMS-11	--	4.09
HepG2	3.52	3.73
neuroepithelial stem cell derived from H9 stably expressing fusion protein	--	3.68
right lobe of liver female adult (53 years)	3.47	3.58
HUES64-derived CD184+	--	3.54
small intestine male fetal (108 days)	3.23	3.40
hepatocyte derived from H9	2.98	3.39
KOPT-K1	--	3.30
liver male adult (31 years)	--	3.29
OCI-LY1	--	3.25

Total: 136

Chromatin markers

DNase Z-scores ⓘ

TSV

cell type	Z-score
large intestine female fetal (108 days)	3.48
large intestine female fetal (107 days)	3.42
small intestine male fetal (105 days)	3.37
small intestine female fetal (108 days)	3.35
right lobe of liver female adult (53 years)	3.35
large intestine female fetal (91 days)	3.35
HepG2	3.31
small intestine female fetal (105 days)	3.29
small intestine female fetal (98 days)	3.26
large intestine female fetal (110 days)	3.21

Total: 462

Z score data from multiple chromatin markers in different cell types

GTEx : Genotype-Tissue Expression (GTEx)

Go to link <https://gtexportal.org/home/>

Enter snp id : rs712 [Homo sapiens]

The screenshot displays the GTEx Portal homepage. The top navigation bar includes links for Home, Datasets, Expression, QTLs & Browser, Sample Data, and Documentation. A search bar is located on the right, and a 'Sign In' button is in the top right corner. Below the navigation bar is a banner image with a survey announcement. The main content area is divided into two columns. The left column, titled 'Resource Overview', contains links for 'Current Release (V8)', 'Tissue & Sample Statistics', 'Tissue Sampling Info (Anatomogram)', 'Access & Download Data', 'Release History', and 'How to cite GTEx?'. A red bracket highlights the 'Current Release (V8)' section, which includes a paragraph about the project and a list of recent releases: 2019-08-26 GTEx Portal V8 Release, 2019-07-24 GTEx V8 data release, 2019-03-07 New Histology Image Viewer, and 2017-10-18 ASHG GTEx Workshop Materials. Below this is a 'News and Events' section and a 'Documentation' section with links for 'Publication Policy', 'Consortium', and 'Analysis Methods'. The right column, titled 'Explore GTEx', is organized into four categories: 'Browse', 'Expression', 'QTL', and 'eQTL'. The 'Browse' category includes 'By gene ID', 'By variant or rs ID' (circled in red), 'By Tissue', and 'Histology Image Viewer'. The 'Expression' category includes 'Multi-Gene Query', 'Top 50 Expressed Genes', and 'Transcript Browser'. The 'QTL' category includes 'Locus Browser', 'IGV eQTL Browser' (circled in red), 'eQTL Dashboard', and 'eQTL Calculator'. The 'eQTL' category includes 'Data coming soon!'. The bottom of the page features a footer with 'GBIO002 AB 2019' and social media links.

Resource Overview

Current Release (V8)

Tissue & Sample Statistics
Tissue Sampling Info (Anatomogram)
Access & Download Data
Release History
How to cite GTEx?

The Genotype-Tissue Expression (GTEx) project is an ongoing effort to build a comprehensive public resource to study tissue-specific gene expression and regulation. Samples were collected from 54 non-diseased tissue sites across nearly 1000 individuals, primarily for molecular assays including WGS, WES, and RNA-Seq. Remaining samples are available from the GTEx Biobank. The GTEx Portal provides open access to data including gene expression, QTLs, and histology images.

News and Events

2019-08-26 GTEx Portal V8 Release
2019-07-24 GTEx V8 data release
2019-03-07 New Histology Image Viewer
2017-10-18 ASHG GTEx Workshop Materials

Documentation

Publication Policy
Consortium
Analysis Methods

Follow us
Contact us
External Links: dbGaP | NIH Common Fund | NHGRI

Explore GTEx

Browse

By gene ID
By variant or rs ID
By Tissue
Histology Image Viewer

Expression

Multi-Gene Query
Top 50 Expressed Genes
Transcript Browser

QTL

Locus Browser
IGV eQTL Browser
eQTL Dashboard
eQTL Calculator

eQTL

Data coming soon!

DNA, RNA methylation, ChIP-seq and more

- Top
- Single-Tissue eQTLs
- Single-Tissue sQTLs

Variant Page

Search:

Show

10

 entries

Variant ID	Shorthand	rs ID (v151)	Chromosome	Position	MAF >= 1%	Ref Allele	Alt Allele	b37 Variant ID
chr12_25209618_A_C_b38		rs712	chr12	25209618	true	A	C	12_25362552_A_C_b37

Showing 1 to 1 of 1 entries

Previous

1

Next

Single-Tissue eQTLs for chr12_25209618_A_C_b38

Data Source: GTEx Analysis Release V8 (dbGaP Accession phs000424.v8.p2)

eQTLs of chr12_25209618_A_C_b38

Copy

CSV

Search:

Show

10

 entries

GeneCode Id	Gene Symbol	Variant Id	SNP	P-Value	NES	Tissue	Actions
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	8.7e-17	0.23	Whole Blood	eQTL violin plot , IGV eQTL Browser , Multi-tissue eQTL Plot
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	1.7e-16	0.20	Skin - Sun Exposed (Lower leg)	eQTL violin plot , IGV eQTL Browser , Multi-tissue eQTL Plot
ENSG00000133703.11	KRAS	chr12_25209618_A_C_b38	rs712 dbSNP	5.5e-15	-0.18	Cells - Cultured fibroblasts	eQTL violin plot , IGV eQTL Browser , Multi-tissue eQTL Plot
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	6.2e-8	0.11	Testis	eQTL violin plot , IGV eQTL Browser , Multi-tissue eQTL Plot
ENSG00000118307.18	CASC1	chr12_25209618_A_C_b38	rs712 dbSNP	6.2e-8	-0.11	Testis	eQTL violin plot , IGV eQTL Browser , Multi-tissue eQTL Plot
ENSG00000118307.18	CASC1	chr12_25209618_A_C_b38	rs712 dbSNP	2.0e-7	0.20	Skin - Sun Exposed (Lower leg)	eQTL violin plot , IGV eQTL Browser , Multi-tissue eQTL Plot
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	2.0e-7	0.14	Skin - Not Sun Exposed (Suprapubic)	eQTL violin plot , IGV eQTL Browser , Multi-tissue eQTL Plot
ENSG00000118307.18	CASC1	chr12_25209618_A_C_b38	rs712 dbSNP	2.4e-7	0.25	Nerve - Tibial	eQTL violin plot , IGV eQTL Browser , Multi-tissue eQTL Plot
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	0.0000020	0.13	Thyroid	eQTL violin plot , IGV eQTL Browser , Multi-tissue eQTL Plot
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	0.000027	0.23	Brain - Cerebellum	eQTL violin plot , IGV eQTL Browser , Multi-tissue eQTL Plot

Showing 1 to 10 of 16 entries

First

Previous

1

2

Next

Last

eQTLs of chr12_25209618_A_C_b38

[Copy](#) [CSV](#)

Gencode Id	Gene Symbol	Variant Id	SNP	P-Value	NE
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	8.7e-17	0.23
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	1.7e-16	0.20
ENSG00000133703.11	KRAS	chr12_25209618_A_C_b38	rs712 dbSNP	5.5e-15	-0.14
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	6.2e-8	0.11
ENSG00000118307.18	CASC1	chr12_25209618_A_C_b38	rs712 dbSNP	6.2e-8	-0.17
ENSG00000118307.18	CASC1	chr12_25209618_A_C_b38	rs712 dbSNP	2.0e-7	0.20
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	2.0e-7	0.14
ENSG00000118307.18	CASC1	chr12_25209618_A_C_b38	rs712 dbSNP	2.4e-7	0.25
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	0.0000020	0.13
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	0.000027	0.23

Showing 1 to 10 of 16 entries

Single-Tissue sQTLs for chr12_25209618_A_C_b38

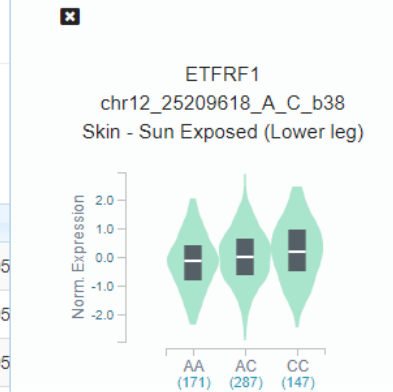
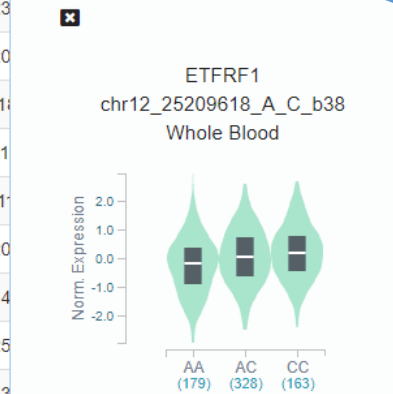
Data Source: GTEx Analysis Release V8 (dbGaP Accession phs000424.v8.p2)

[Copy](#) [CSV](#)

Gencode Id	Gene Symbol	Variant Id	SNP	Intron Id
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	25195337:25195708:cln_3401
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	25195337:25195708:cln_3401
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	25195337:25195708:cln_3401
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	25195337:25195708:cln_3401
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	25195337:25195708:cln_3401
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	25195337:25195708:cln_3401
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	25195337:25195708:cln_3401
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	25195337:25195708:cln_3401
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	25195337:25195708:cln_3401
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	25195337:25195708:cln_3401

eQTL Violin Plots

Clear All



Search: Show 10 entries

Actions
eQTL violin plot, IGW eQTL Browser, Multi-tissue eQTL Plot
eQTL violin plot, IGW eQTL Browser, Multi-tissue eQTL Plot
eQTL violin plot, IGW eQTL Browser, Multi-tissue eQTL Plot
eQTL violin plot, IGW eQTL Browser, Multi-tissue eQTL Plot
eQTL violin plot, IGW eQTL Browser, Multi-tissue eQTL Plot
eQTL violin plot, IGW eQTL Browser, Multi-tissue eQTL Plot
eQTL violin plot, IGW eQTL Browser, Multi-tissue eQTL Plot
eQTL violin plot, IGW eQTL Browser, Multi-tissue eQTL Plot
eQTL violin plot, IGW eQTL Browser, Multi-tissue eQTL Plot
eQTL violin plot, IGW eQTL Browser, Multi-tissue eQTL Plot

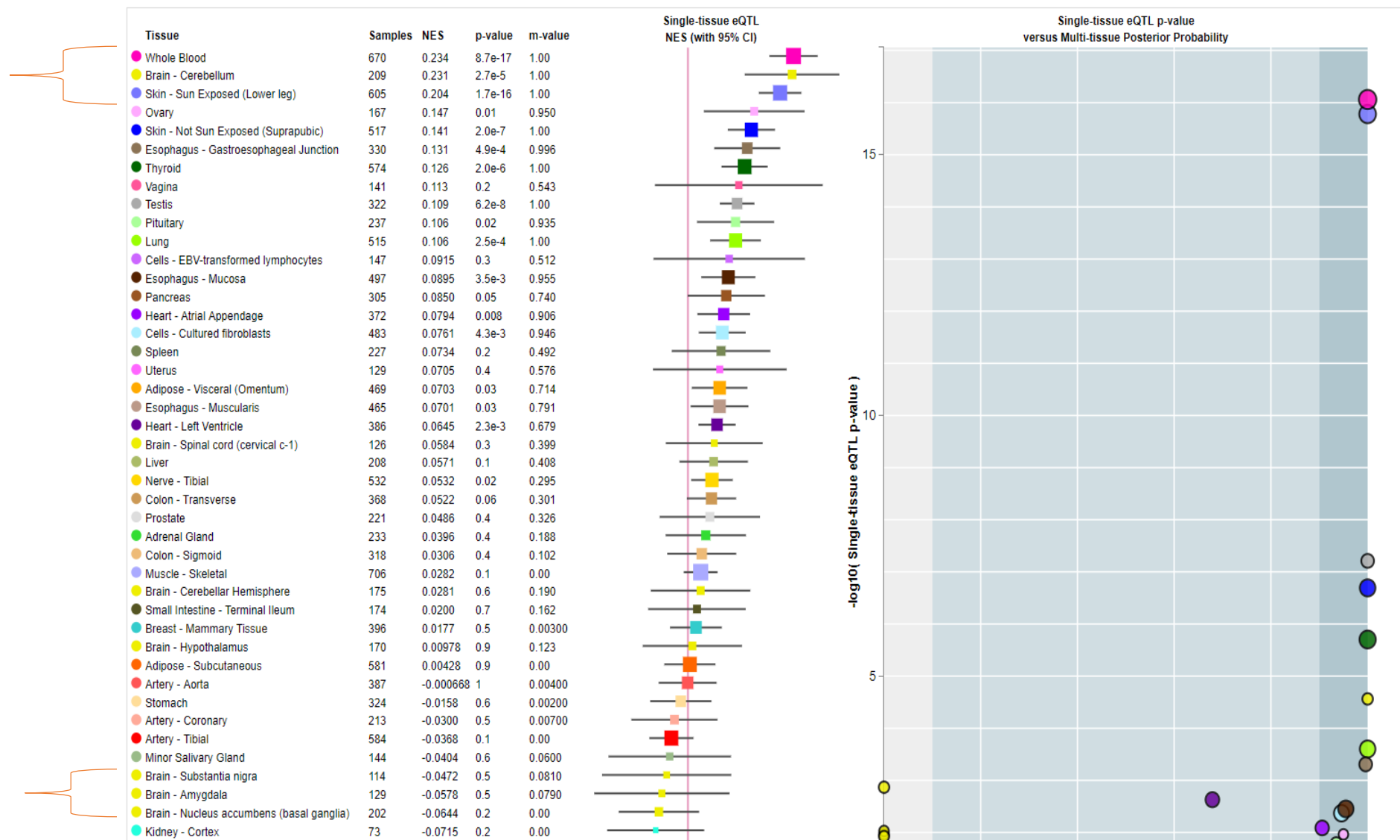
Search: Show 10 entries

Tissue	Actions
Thyroid	sQTL violin plot
Skin - Sun Exposed (Lower leg)	sQTL violin plot
Pituitary	sQTL violin plot
Testis	sQTL violin plot
Esophagus - Mucosa	sQTL violin plot
Artery - Tibial	sQTL violin plot
Artery - Aorta	sQTL violin plot

Indicates snps has high expression in human blood and skin tissues

Multi-tissue eQTL Comparison

ENSG00000205707.10 ETRF1 and chr12_25209618_A_C_b38 eQTL (Meta Analysis RE2 P-Value: 1.9385099999999995e-60)



Ensembl Database

<https://www.ensembl.org/index.html>

The screenshot shows the Ensembl Database homepage. A red oval highlights the top navigation bar, which includes the Ensembl logo, links to BLAST/BLAT, VEP, Tools, BioMart, Downloads, Help & Docs, and Blog, and a dropdown menu for the current species, Human (GRCh38.p13). A red rectangle highlights the 'Variation' section, which includes a search bar for variants, a link to 'More about variation in Ensembl', a link to 'Download all variants (GVF)', and a link to the 'Variant Effect Predictor (VEP)'. The 'Variation' section also features an 'Example variant' and an 'Example structural variant'.

ensembl.org/Homo_sapiens/info/index?db=core

Ensembl BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

Human (GRCh38.p13) ▾

Search Human (*Homo sapiens*)

Search all categories ▾ Search Human... Go

e.g. BRCA2 or 17:63992802-64038237 or rs699 or osteoarthritis

Genome assembly: GRCh38.p13 (GCA_000001405.28)

- More information and statistics
- Download DNA sequence (FASTA)
- Convert your data to GRCh38 coordinates
- Display your data in Ensembl

Other assemblies

GRCh37 Full Feb 2014 archive with BLAST, VEP and BioMart ▾ Go

Gene annotation

What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein sequences, non-coding RNAs.

- More about this genebuild
- Download FASTA files for genes, cDNAs, ncRNA, proteins
- Download GTF or GFF3 files for genes, cDNAs, ncRNA, proteins
- Update your old Ensembl IDs

Comparative genomics

What can I find? Homologues, gene trees, and whole genome alignments across multiple species.

- More about comparative analysis
- Download alignments (EMF)

Regulation

What can I find? DNA methylation, transcription factor binding sites, histone modifications, and regulatory features such as enhancers and repressors, and microarray annotations.

- More about the Ensembl regulatory build and microarray annotation
- Experimental data sources
- Download all regulatory features (GFF)

Variation

What can I find? Short sequence variants and longer structural variants; disease and other phenotypes

- More about variation in Ensembl
- Download all variants (GVF)
- Variant Effect Predictor **VEP**

Variant annotation

Example gene tree

Example region

Example transcript

Example variant

Example phenotype

Example structural variant

UNIPROT KB

Available at <https://www.uniprot.org/>

- **The UniProt Knowledgebase (UniProtKB) is the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation.**
- **In addition to capturing the core data mandatory for each UniProtKB entry (mainly, the amino acid sequence, protein name or description, taxonomic data and citation information), as much annotation information as possible is added.**



Cross-referenced databases ▼

Advanced ▼

Search

BLAST Align Retrieve/ID mapping Peptide search

Help Contact

Database - dbSNP

Map to

Format

UniProtKB (12,533)

Name	Database of single nucleotide polymorphism
Servers	https://www.ncbi.nlm.nih.gov/SNP/
URL template	https://www.ncbi.nlm.nih.gov/SNP/snp_ref.cgi?type=rs&rs=%s
Citation	[PubMed:17170002][DOI:10.1093/nar/gkl1031]
Link type	Explicit
Category	Polymorphism and mutation databases

Tools

BLAST
Align
Retrieve/ID mapping
Peptide search

Core data

Protein knowledgebase (UniProtKB)
Sequence clusters (UniRef)
Sequence archive (UniParc)
Proteomes

Supporting data

Literature citations
Taxonomy
Keywords
Subcellular locations
Cross-referenced databases
Diseases

Information

About UniProt
Help
FAQ
UniProtKB manual
Technical corner
Expert biocuration



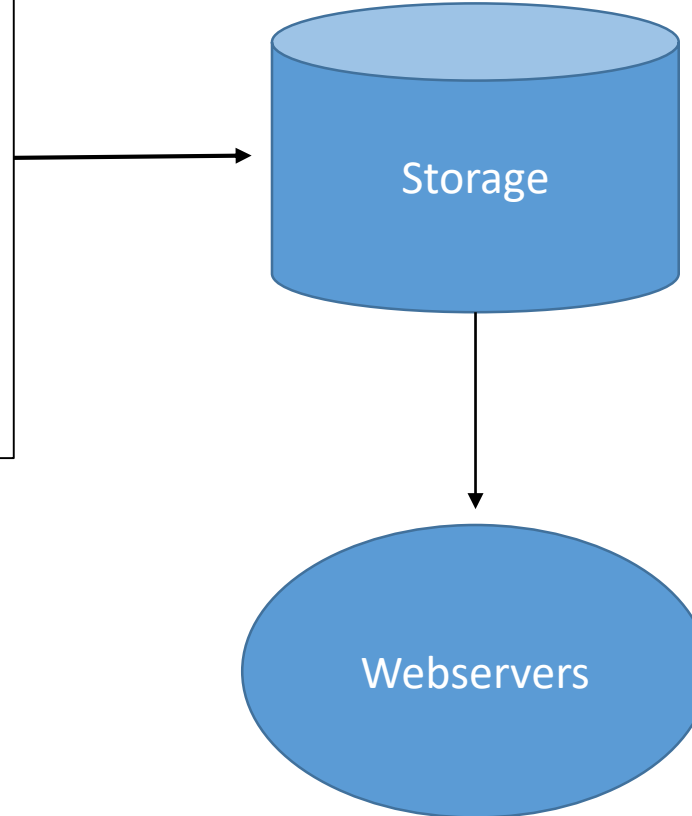
© 2002 – 2019 UniProt Consortium | License & Disclaimer | Privacy Notice



Multiple web servers (for Post GWAS)

- Identifying causal variants remains a key challenge in post-GWAS (genome-wide association study) era, as many GWAS single-nucleotide polymorphisms (SNPs) (including imputed ones) fall into non-coding regions.
- Its making it difficult to associate statistical significance with predicted functionality.
- Therefore, researches developed web-based multiple tools which overlays functional annotation information, such as histone modification states, methylation patterns, transcription factor binding sites, eQTL and higher-order chromosomal structure, to GWAS results.

- **functional annotation information, such as histone modification states**
- **methylation patterns,**
- **transcription factor binding sites**
- **eQTL and**
- **higher-order chromosomal structure**



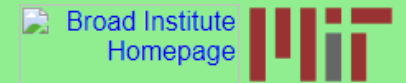
HaploReg web server

<https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php>

stitute.org/mammals/haploreg/haploreg.php



HaploReg v4.1



HaploReg is a tool for exploring annotations of the noncoding genome at variants on haplotype blocks, such as candidate regulatory SNPs at disease-associated loci. Using LD information from the 1000 Genomes Project, linked SNPs and small indels can be visualized along with chromatin state and protein binding annotation from the Roadmap Epigenomics and ENCODE projects, sequence conservation across mammals, the effect of SNPs on regulatory motifs, and the effect of SNPs on expression from eQTL studies. HaploReg is designed for researchers developing mechanistic hypotheses of the impact of non-coding variants on clinical phenotypes and normal variation.

Update 2015.11.05: Version 4.1 GWAS and eQTL have been updated; a simpler pruning strategy is applied when combining GWAS; and links out to other NHGRI/EBI GWAS hits and GRASP QTL hits are provided.

Update 2015.09.15: Version 4.0 now includes many recent eQTL results including the GTEx pilot, four different options for defining enhancers using Roadmap Epigenomics data, and a complete set of source files for download and local analysis. Older versions available: [v3](#), [v2](#), [v1](#).

[Build Query](#) [Set Options](#) [Documentation](#)

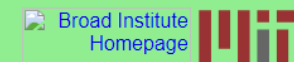
Use one of the three methods below to enter a [Documentation](#) If an r^2 threshold is specified (see the Set Options tab), results for each variant will be shown in a separate table along with other variants in LD. If r^2 is set to NA, only queried variants will be shown, together in one table.

Query (comma-delimited list of rsIDs OR a single region as chrN:start-end):

or, upload a text file (one refSNP ID per line): No file chosen

or, select a GWAS:

HaploReg v4.1



HaploReg is a tool for exploring annotations of the noncoding genome at variants on haplotype blocks, such as candidate regulatory SNPs at disease-associated loci. Using LD information from the 1000 Genomes Project, linked SNPs and small indels can be visualized along with chromatin state and protein binding annotation from the Roadmap Epigenomics and ENCODE projects, sequence conservation across mammals, the effect of SNPs on regulatory motifs, and the effect of SNPs on expression from eQTL studies. HaploReg is designed for researchers developing mechanistic hypotheses of the impact of non-coding variants on clinical phenotypes and normal variation.

Update 2015.11.05: Version 4.1 GWAS and eQTL have been updated; a simpler pruning strategy is applied when combining GWAS; and links out to other NHGRI/EBI GWAS hits and GRASP QTL hits are provided.

Update 2015.09.15: Version 4.0 now includes many recent eQTL results including the GTEx pilot, four different options for defining enhancers using Roadmap Epigenomics data, and a complete set of source files for download and local analysis. Older versions available: [v3](#), [v2](#), [v1](#).

[Build Query](#) [Set Options](#) [Documentation](#)

Use one of the three methods below to enter a set of variants. If an r^2 threshold is specified (see the Set Options tab), results for each variant will be shown in a separate table along with other variants in LD. If r^2 is set to NA, only queried variants will be shown, together in one table.

Query (comma-delimited list of rsIDs OR a single region as chrN:start-end):

or, upload a text file (one refSNP ID per line): [Choose File](#) No file chosen

or, select a GWAS:

[Submit](#)

Query SNP: **rs9271055** and variants with $r^2 \geq 0.8$

chr	pos (hg38)	LD (r ²)	LD (D')	variant	Ref	Alt	AFR freq	AMR freq	ASN freq	EUR freq	SiPhy cons	Promoter histone marks	Enhancer histone marks	DNAse	Proteins bound	Motifs changed	NHGRI/EBI GWAS hits	GRASP QTL hits	Selected eQTL hits	GENCODE genes	dbSNP func annot
6	32602082	0.88	0.94	rs9270815	A	G	0.83	0.88	0.81	0.85			BLD			HNF4A,PPAR			265 hits	12kb 5' of HLA-DRB1	intronic
6	32604152	0.81	0.96	rs4367411	C	T	0.79	0.86	0.78	0.84		BLD, FAT	BLD	10 tissues	POL2	Maf,Spz1			263 hits	14kb 5' of HLA-DRB1	intronic
6	32604684	0.91	0.97	rs9270928	G	T	0.82	0.88	0.81	0.85		BLD, FAT	BLD, BRN, GI	16 tissues	5 bound proteins				265 hits	15kb 5' of HLA-DRB1	intronic
6	32606132	0.88	0.98	rs9270980	C	A	0.82	0.88	0.81	0.84			BLD			Evi-1			264 hits	16kb 5' of HLA-DRB1	intronic
6	32606283	0.95	0.98	rs9270986	A	C	0.83	0.89	0.81	0.85			BLD	BLD		Ascl2		34 hits	273 hits	16kb 5' of HLA-DRB1	intronic
6	32606473	0.95	0.98	rs9270994	T	C	0.83	0.89	0.81	0.85			BLD	BLD, BLD					265 hits	17kb 5' of HLA-DRB1	
6	32606597	0.94	0.97	rs9270997	G	A	0.83	0.89	0.81	0.85			BLD	BLD		FAC1,Pou1f1,STAT			265 hits	17kb 5' of HLA-DRB1	
6	32607592	1	1	rs9271055	G	T	0.83	0.88	0.81	0.85		BLD	BLD	5 tissues	BATF,EGR1,NFKB	4 altered motifs		4 hits	299 hits	18kb 5' of HLA-DRB1	
6	32607601	1	1	rs9271056	T	C	0.83	0.88	0.81	0.85		BLD	BLD	5 tissues	BATF,EGR1,NFKB	BDP1,MIF-1,Myf			265 hits	18kb 5' of HLA-DRB1	
6	32607767	0.97	0.99	rs9271061	A	T	0.83	0.89	0.81	0.85		BLD	ESC, BLD, FAT	BLD, BLD, BLD	5 bound proteins	Hoxa13,Hoxb13			265 hits	18kb 5' of HLA-DRB1	
6	32607798	0.94	0.99	rs9271062	T	A	0.83	0.89	0.81	0.85		BLD	ESC, BLD, FAT	4 tissues	5 bound proteins	STAT			267 hits	18kb 5' of HLA-DRB1	
6	32607842	0.82	0.96	rs9271065	C	G	0.83	0.94	0.88	0.87		BLD	BLD, FAT	4 tissues	4 bound proteins				228 hits	18kb 5' of HLA-DRB1	
6	32608299	0.8	0.97	rs9271080	C	T	0.79	0.86	0.78	0.83		BLD	BLD	BLD, BLD	NFKB, TBP	HNF1,Ncx			264 hits	18kb 5' of HLA-DRB1	
6	32608309	0.81	0.98	rs9271082	T	C	0.79	0.86	0.77	0.83		BLD	BLD	BLD, BLD	NFKB, TBP	Pax-6			229 hits	18kb 5' of HLA-DRB1	
6	32608375	0.86	0.98	rs9271085	T	C	0.82	0.88	0.80	0.84		BLD	BLD	BLD, BLD, BLD	NFKB, TBP	4 altered motifs			264 hits	19kb 5' of HLA-DRB1	
6	32608564	0.9	0.95	rs9271093	G	A	0.82	0.88	0.81	0.85		BLD	BLD	5 tissues	CTCF,NFKB,TBP	6 altered motifs			263 hits	19kb 5' of HLA-DRB1	
6	32609754	0.8	0.9	rs9271152	T	G	0.83	0.88	0.81	0.86		5 tissues	11 tissues	16 tissues	6 bound proteins				265 hits	18kb 5' of HLA-DQA1	

Advantage

- It was developed to systematically mine chromatin state data, along with conservation data and regulatory motif alterations.
- It uses Gtex , Encode databases in backend.
- Most importantly, it gives motif based regulatory impact of SNPs

SNP causes 4 altered motifs due to change in nucleotide from G to T

chr	pos (hg38)	LD (r ²)	LD (D')	variant	Ref	Alt	AFR freq	AMR freq	ASN freq	EUR freq	SiPhy cons	Promoter histone marks	Enhancer histone marks	DNAse	Proteins bound	Motifs changed	NHGRI/EBI GWAS hits	GRASP QTL hits	Selected eQTL hits	GENCODE genes	dbSNP func annot
6	32602082	0.88	0.94	rs9270815	A	G	0.83	0.88	0.81	0.85			BLD			HNF4,PPAR			265 hits	12kb 5' of HLA-DRB1	intronic
6	32604152	0.81	0.96	rs4367411	C	T	0.79	0.86	0.78	0.84		BLD, FAT	BLD	10 tissues	POL2	Maf,Spz1			263 hits	14kb 5' of HLA-DRB1	intronic
6	32604684	0.91	0.97	rs9270928	G	T	0.82	0.88	0.81	0.85		BLD, FAT	BLD, BRN, GI	16 tissues	5 bound proteins				265 hits	15kb 5' of HLA-DRB1	intronic
6	32606132	0.88	0.98	rs9270980	C	A	0.82	0.88	0.81	0.84			BLD			Evi-1			264 hits	16kb 5' of HLA-DRB1	intronic
6	32606283	0.95	0.98	rs9270986	A	C	0.83	0.89	0.81	0.85			BLD	BLD		Ascl2		34 hits	273 hits	16kb 5' of HLA-DRB1	intronic
6	32606473	0.95	0.98	rs9270994	T	C	0.83	0.89	0.81	0.85			BLD	BLD, BLD					265 hits	17kb 5' of HLA-DRB1	
6	32606597	0.94	0.97	rs9270997	G	A	0.83	0.89	0.81	0.85			BLD	BLD		FAC1,Pou1f1,STAT			265 hits	17kb 5' of HLA-DRB1	
6	32607592	1	1	rs9271055	G	T	0.83	0.88	0.81	0.85		BLD	BLD	5 tissues	BATF,EGR1,NFKB	4 altered motifs		4 hits	299 hits	18kb 5' of HLA-DRB1	
6	32607601	1	1	rs9271056	T	C	0.83	0.88	0.81	0.85		BLD	BLD	5 tissues	BATF,EGR1,NFKB	BDP1,MIF-1,Myf			265 hits	18kb 5' of HLA-DRB1	
6	32607767	0.97	0.99	rs9271061	A	T	0.83	0.89	0.81	0.85		BLD	ESC, BLD, FAT	BLD, BLD, BLD	5 bound proteins	Hoxa13,Hoxb13			265 hits	18kb 5' of HLA-DRB1	
6	32607798	0.94	0.99	rs9271062	T	A	0.83	0.89	0.81	0.85		BLD	ESC, BLD, FAT	4 tissues	5 bound proteins	STAT			267 hits	18kb 5' of HLA-DRB1	
6	32607842	0.82	0.96	rs9271065	C	G	0.83	0.94	0.88	0.87		BLD	BLD, FAT	4 tissues	4 bound proteins				228 hits	18kb 5' of HLA-DRB1	
6	32608299	0.8	0.97	rs9271080	C	T	0.79	0.86	0.78	0.83		BLD	BLD	BLD, BLD	NFKB,TBP	HNF1,Ncx			264 hits	18kb 5' of HLA-DRB1	
6	32608309	0.81	0.98	rs9271082	T	C	0.79	0.86	0.77	0.83		BLD	BLD	BLD, BLD	NFKB,TBP	Pax-6			229 hits	18kb 5' of HLA-DRB1	
6	32608375	0.86	0.98	rs9271085	T	C	0.82	0.88	0.80	0.84		BLD	BLD	BLD, BLD, BLD	NFKB,TBP	4 altered motifs			264 hits	19kb 5' of HLA-DRB1	
6	32608564	0.9	0.95	rs9271093	G	A	0.82	0.88	0.81	0.85		BLD	BLD	5 tissues	CTCF,NFKB,TBP	6 altered motifs			263 hits	19kb 5' of HLA-DRB1	
6	32609754	0.8	0.9	rs9271152	T	G	0.83	0.88	0.81	0.86		5 tissues	11 tissues	16 tissues	6 bound proteins				265 hits	18kb 5' of HLA-DQA1	

RegulomeDB

Access to the database at <http://RegulomeDB.org/>

[Download](#) [About](#) [Help](#)



v 1.1 TRY NEW BETA SITE

Enter dbSNP IDs, 0-based coordinates, BED files, VCF files, GFF3 files (hg19).

Submit

Use RegulomeDB to identify DNA features and regulatory elements in non-coding regions of the human genome by entering ...

dbSNP IDs

Single nucleotides

A chromosomal region

Enter dbSNP ID(s) (example) or upload a list of dbSNP IDs to identify DNA features and regulatory elements that contain the coordinate of the SNP(s).

 A project of the Center for Genomics and Personalized Medicine at Stanford University. 

RegulomeDB (TM) Copyright ©2011 The Board of Trustees of Leland Stanford Junior University. Permission to use the information contained in this database was given by the researchers/institutes who contributed or published the information. Users of the database are solely responsible for compliance with any copyright restrictions, including those applying to the author abstracts. Documents from this server are provided "AS-IS" without any warranty, expressed or implied. The RegulomeDB project at Stanford University is supported by a Genome Research Resource Grant from the US National Human Genome Research Institute, part of the US National Institutes of Health.

Input Files Format

- The integrated database is fully searchable using common variant formats (VCF, BED, GFF3, rsIDs) and through file upload of the same formats.

rsID FORMAT

rs33914668
rs35004220
rs78077282
rs7881236

VCF FORMAT

#CHROM	POS	REF	ALT	INFO
chr1	100	G	A	AC=10;AF=0.05
chr1	200	C	T	AC=40;AF=0.20
chr1	300	G	T	AC=20;AF=0.10
...				

BED FORMAT

1	#Chromosome	Start	End	SNP Id	Allele	
2	chr1	174	175	1	T/C	
3	chr1	5073	5074	2	T/G	
4	chr1	5635	5636	3	T/C	
5	chr1	6240	6241	4	T/C	
6	chr1	39160	39161	5	T/C	
7	chr1	50111	50112	6	C/T	
8	chr1	126968	126969	7	C/A	
9	chr1	223601	223602	8	C/T	
10	chr1	226507	226508	9	T/A	
11	chr1	251874	251875	10	C/T	
12	chr1	523060	523061	11	C/T	

Output Files

- **The initial results table provides a list of the coordinates of the variants, a dbSNP rsID (if it exists), a score assigned by method, and links to external resources for each variant**
- **The list is sorted by our classification scheme, with the SNVs most likely to be functional listed first. This list of SNVs is also downloadable by the user for their own analysis.**

The search has evaluated 5 input line(s) and found 4 SNP(s).

Summary of SNP analysis

Show 10 entries

Coordinate (0-based)	dbSNP ID	? Regulome DB Score	Other Resources
chr11:5246957	rs33914668	2a	UCSC ENSEMBL dbSNP
chrX:53101683	rs7881236	2c	UCSC ENSEMBL dbSNP
chr11:5248049	rs35004220	4	UCSC ENSEMBL dbSNP
chr14:100741725	rs78077282	4	UCSC ENSEMBL dbSNP

Showing 1 to 4 of 4 entries

Download

BED

GFF

Full Output

Click on each score one by one



A project of the Center for Genomics and Personalized Medicine at Stanford University.



RegulomeDB (TM) Copyright ©2011 The Board of Trustees of Leland Stanford Junior University. Permission to use the information contained in this database was given by the researchers/institutes who contributed or published the information. Users of the database are solely responsible for compliance with any copyright restrictions, including those applying to the author abstracts. Documents from this server are provided "AS-IS" without any warranty, expressed or implied. The RegulomeDB project at Stanford University is supported by a Genome Research Resource Grant from the US National Human Genome Research Institute, part of the US National Institutes of Health.

- This display includes six major categories: Protein Binding, Motifs, Chromatin Structure, eQTLs, Histone Modifications, and Related Data (which includes gene information and other manual annotations).

Table 1. Database content

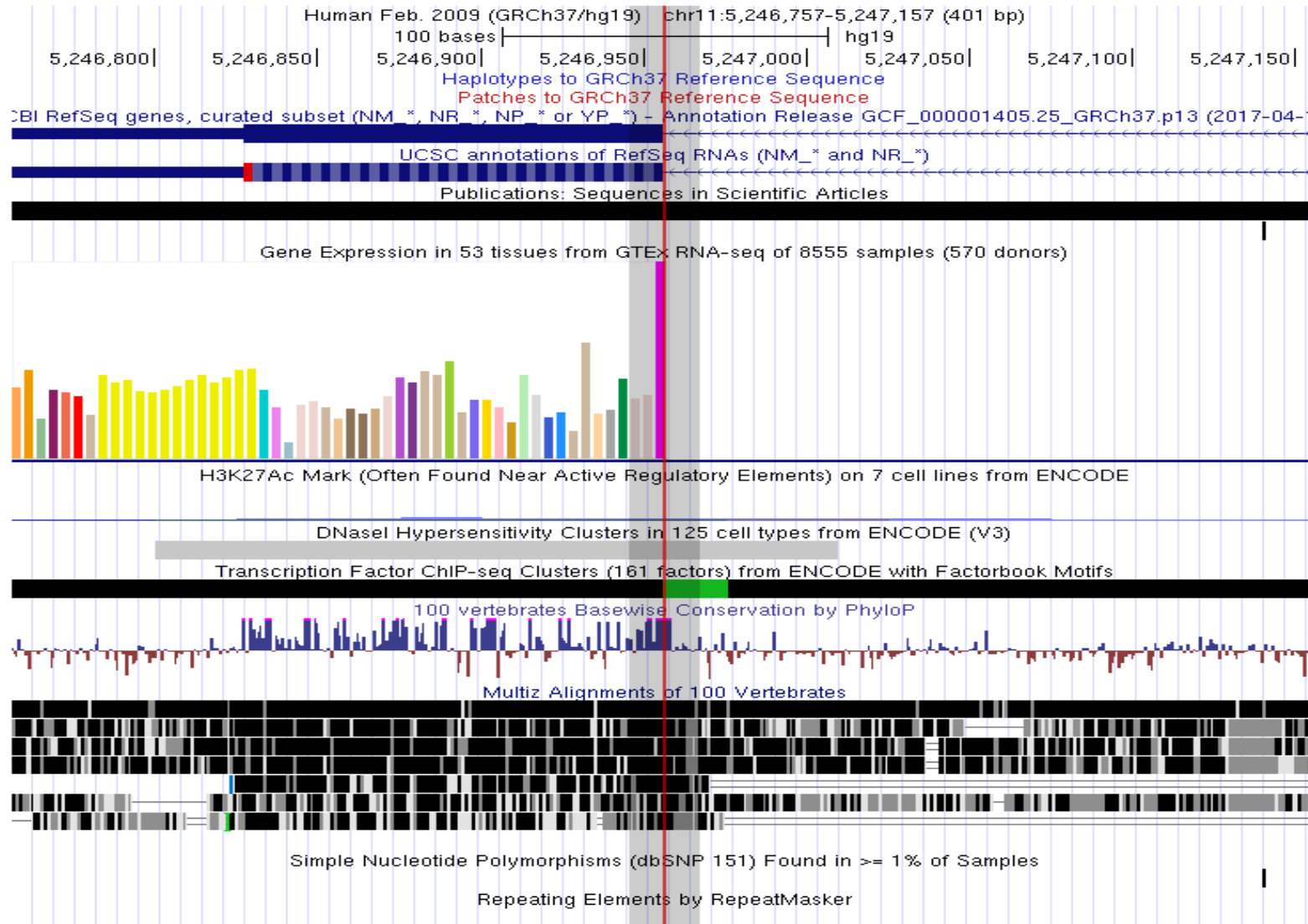
Data type	Types	Features	Genomic coverage (bp)
Transcription factor ChIP-seq (ENCODE)	495 conditions/cell lines	7,721,822	230,795,743
Transcription factor ChIP-seq (non-ENCODE)	32 conditions/cell lines	397,534	140,534,725
Transcription factor ChIP-exo	1 condition	35,161	2,604,066
Histone modifications	284 conditions/cell lines/marks	23, 055, 241	2,805,205,184
DNase I hypersensitive sites	114 conditions/cell lines	20,710,098	614,973,579
FAIRE sites	25 conditions/cell lines	4,816,196	476,386,909
DNase I footprints	50 cell lines	128,266,803	178,722,370
Predicted binding (PWMs)	1158 motifs	239,713,973	1,151,732,122
eQTLs	142,945 SNPs	142,945	142,945
dsQTLs	6069 SNPs	6069	6069
Manual annotations	6 genomic regions	282	11,607
VISTA enhancers	1448 enhancers	1325	1,658,146
Validated SNPs affecting binding	855 SNPs	855	855

Sources of data currently included in RegulomeDB. (Features) Specific entries in the database. (Genomic coverage) Total unique base pairs covered by each data type.

Data supporting chr11:5246957 (rs33914668)

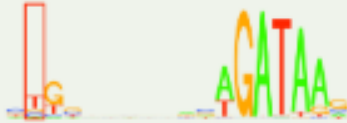
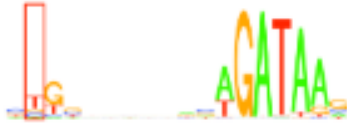
Score: 2a

Likely to affect binding



GBIO002 AB 2019

- Each of these categories provides detailed information about the transcription factor, cell line, and a literature source of the information to provide the user with direct access for addressing their hypothesis.

Motifs						Filter: <input type="text"/>
Method	Location	Motif	? Cell Type	PWM	Reference	
Footprinting	chr11:5246956..5246974	Tal1::Gata1	K562		21106904	
PWM	chr11:5246956..5246974	Tal1::Gata1			18006571	

Result indicate SNP is present in Gata Motif which could have regulatory impact on the gene expresion

Histone modifications					Filter: <input type="text"/>
Method	Location	Chromatin State	Tissue Group	Tissue	Reference
ChromHMM	chr11:4648200..5617400	Quiescent/Low	Digestive	Colonic Mucosa	REMC
ChromHMM	chr11:4648400..5255400	Quiescent/Low	Thymus	Thymus	REMC
ChromHMM	chr11:4658600..5617400	Quiescent/Low	Digestive	Rectal Mucosa Donor 29	REMC
ChromHMM	chr11:4687400..5545600	Quiescent/Low	Digestive	Rectal Mucosa Donor 31	REMC
ChromHMM	chr11:4704000..5530600	Quiescent/Low	ES-deriv	H9 Derived Neuronal Progenitor Cultured Cells	REMC
ChromHMM	chr11:4742400..5617400	Quiescent/Low	Sm. Muscle	Colon Smooth Muscle	REMC
ChromHMM	chr11:4772600..5273800	Quiescent/Low	Blood & T-cell	Primary T helper memory cells from peripheral blood 2	REMC
ChromHMM	chr11:4815200..5351800	Quiescent/Low	Blood & T-cell	Primary T helper memory cells from peripheral blood 1	REMC
ChromHMM	chr11:4820400..5617400	Quiescent/Low	Digestive	Stomach Mucosa	REMC
ChromHMM	chr11:4859800..5371600	Quiescent/Low	Blood & T-cell	Primary T CD8+ naive cells from peripheral blood	REMC
ChromHMM	chr11:4885000..5272600	Quiescent/Low	Other	Placenta Amnion	REMC
ChromHMM	chr11:5086000..5617800	Quiescent/Low	Blood & T-cell	Primary T cells effector/memory enriched from peripheral blood	REMC
ChromHMM	chr11:5080800..5605600	Quiescent/Low	Blood & T-cell	Primary T CD8+ memory cells from peripheral blood	REMC

Result indicates SNP has chromatin regulatory impact

Related data					Filter: <input type="text"/>
Method	Location	Cell Type	Annotation	Reference	
Transcript_expression_evidence	chr11:5246957..5246958	Cho	Canonical Three Prime Splice Site	2987809	

Result indicates SNP has expression in cho cell type and affect Splice site

Advantage of RegulomeDB

- **An integrated database to quickly generate prioritized hypotheses for the function of variants affecting both coding and noncoding regions in a genome by combining a large array of data sources into a single, integrated database.**
- **In particular, it include extensive information on annotated and computed regulatory elements in the human genome.**
- **Access to this novel approach via a simple and straightforward interface allows for easy query submission, and the scoring system provides for instant classification of significant variants.**
- **In addition, the SNV summary page will allow a user to quickly form a hypothesis as to the true functional consequence of a variant.**
- **While our examples deal with single nucleotide variants only, the database can also be used to annotate insertions and deletions.**

Comparision of HaploReg and RegulomeDB

- [Ward and Kellis \(2012\)](#) published the HaploReg database which aims to provide a similar annotation by providing an intersect of SNVs with chromatin state ([Ernst and Kellis 2010](#)).
- RegulomeDB database provides additional information well beyond this by prioritizing SNVs within general regulatory regions based on specific TF, chromatin, eQTL, and PWM information.
- Furthermore, RegulomeDB allow for a query of personal SNPs which account for a large proportion of variation in the population.

How many of these SNPs alter motifs sequence ?

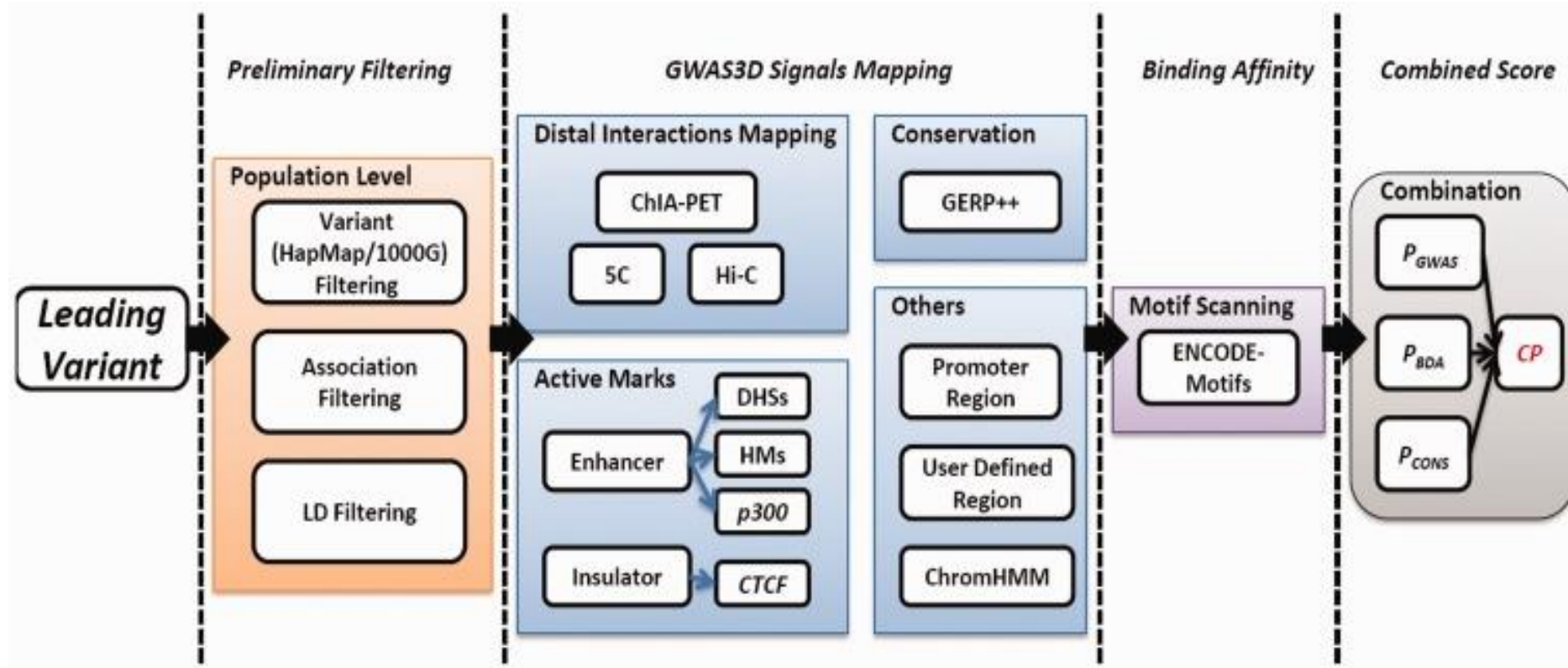
rs4468290

rs11201609

GWAS3D/GWAS4D

- GWAS3D: detecting human regulatory variants by integrative analysis of genome-wide associations, chromosome interactions and histone modifications

<http://mulinlab.tmu.edu.cn/gwas4d/gwas4d/gwas4d>



From GWAS to Regulatory Function

- Majority of GWAS risk loci localize to the noncoding genomic region with gene regulatory signal, suggesting that most trait/disease casual SNPs exert their phenotypic effects by altering gene expression. GWAS4D systematically analyzes GWAS summary data and identify context-specific regulatory variants by integrating latest multidimensional functional genomics resources and our recently published algorithms.

Context-dependent Prediction

- By incorporating roadmap 127 tissue/cell type-specific epigenomes data, GWAS4D uses joint likelihood framework to measure the regulatory probability of genetic variants in a context-dependent manner. It also estimates possible altered TFBSs using large-scale motif collections and annotates non-coding variant with comprehensive functional predictions.

Link Variant to Target

Connecting non-coding variant to their gene targets under particular chromatin organization is crucial to understand variant regulatory mechanism. GWAS4D uniformly processes Hi-C data and reports significant interactions at 5kb resolution across tissues/cell types of multiple human organs and different development stages. It also equips a highly interactive visualization function for variant-target interaction.

Comparision with RegulomeDB and HaploReg

- Compared with recent software and databases such as HaploReg and RegulomeDB, GWAS3D integrates more features and can be used in many scenarios.
- User can identify the most probable functional variant associated with interesting trait in one risk locus or prioritize the leading variants when given a full list of GWAS result or evaluate the deleteriousness of genetic variants affecting the gene regulation without any prior effect.
- GWAS3D also provides flexible configurations, such as human population, cell type specificity and TF family classification, for users to deal with different aspects of complex disease/trait. For example, user may select a matched cell type/tissue satisfying with a specific phenotype or manually define motifs of interested TFs used in following scanning when considering the tissue specificity of TFs.
- Recently, researchers found that the disease/trait-associated variants are highly related to active chromatin marks in relevant cell types. Therefore, these distinct features will greatly facilitate the discovery of regulatory variants under particular condition.

Comparision with RegulomeDB and HaploReg

- The computational process of our system is real-time, which is different from databases such as HaploReg and RegulomeDB, where the function annotations are pre-computed and stored in the database in advance.
- Therefore, it can dynamically deal with the genetic variants input by users with maximum flexibility.
- Despite large computational burden in the background when LD is considered, our system can finish the job of a meta GWAS data set (thousands of variants with moderate GWAS significance, $P < 1.0 \times 10^{-5}$) within a few hours even with LD from the 1000 Genomes Project. It will be much quicker when using HapMap LD.
- To exploit the regulatory properties of personal genomics data, GWAS3D accepts VCF-like format and can evaluate the deleteriousness of rare/novel variation altering gene regulation associated with personalized trait.

List of Tools

As
discussed
before

Analyses and
visualization

Tools	Format	GWAS summary statistics	LD	Functional consequences on genes	Regulatory elements	eQTLs	3D chromatin interactions	Prioritize SNPs	Map SNPs to genes	Gene expression	Pathways and gene sets	Prioriti genes
<i>LD calculation</i>												
PLINK	St	x	x									
<i>Variant annotations</i>												
ANNOVAR	St			x	x			x	x			
VEP	St			x	x			x	x			
SCAN	Web		x			x		x		x		
ReglomeDB	Web				x	x		x				
HaploReg	Web		x		x	x		x				
<i>Gene-based test/Gene-set analyses</i>												
VEGAS	St	x							x			x
MAGMA	St	x							x		x	x
Pascal	St	x							x		x	x
MAGENTA	St	x							x		x	x
INRICH	St	x							x		x	
DEPICT	St	x							x		x	x
<i>Visualization tools</i>												
LocusZoom	St/Web	x										
LocusTrack	St/Web	x			x							
3D genome browser	Web						x					
<i>FUMA</i>												
	Web	x	x	x	x	x	x	x	x	x	x	x



PLoS Comput Biol. 2015 Apr; 11(4): e1004219.

PMCID: PMC4401657

Published online 2015 Apr 17. doi: [10.1371/journal.pcbi.1004219](https://doi.org/10.1371/journal.pcbi.1004219)

PMID: [25885710](https://pubmed.ncbi.nlm.nih.gov/25885710/)

MAGMA: Generalized Gene-Set Analysis of GWAS Data

[Christiaan A. de Leeuw](#), ^{1,2,*} [Joris M. Mooij](#), ³ [Tom Heskes](#), ² and [Danielle Posthuma](#) ^{1,4}

Hua Tang, Editor

► [Author information](#) ► [Article notes](#) ► [Copyright and License information](#) [Disclaimer](#)

This article has been [cited by](#) other articles in PMC.

Associated Data

- [Supplementary Materials](#)
- [Data Availability Statement](#)

Abstract

[Go to:](#)

By aggregating data for complex traits in a biologically meaningful way, gene and gene-set analysis constitute a valuable addition to single-marker analysis. However, although various methods for gene and gene-set analysis currently exist, they generally suffer from a number of issues. Statistical power for most methods is strongly affected by linkage disequilibrium between markers, multi-marker associations are often hard to detect, and the reliance on permutation to compute p-values tends to make the analysis computationally very expensive. To address these issues we have developed MAGMA, a novel tool for gene and gene-set analysis. The gene analysis is based on a multiple regression model, to provide better statistical performance. The gene-set analysis is built as a separate layer around the gene analysis for additional flexibility. This gene-set analysis also uses a regression structure to allow generalization to analysis of continuous properties of genes and simultaneous analysis of multiple gene sets and other gene

Gene analysis

- The gene analysis in MAGMA is based on a multiple linear principal components regression model, using an F-test to compute the gene p-value.
- This model first projects the SNP matrix for a gene onto its principal components (PC), pruning away PCs with very small eigenvalues, and then uses those PCs as predictors for the phenotype in the linear regression model.
- This improves power by removing redundant parameters, and guarantees that the model is identifiable in the presence of highly collinear SNPs.

Gene-set analysis

- To perform the gene-set analysis, for each gene g the gene p-value p_g computed with the gene analysis is converted to a Z-value $z_g = \Phi^{-1}(1 - p_g)$, where Φ^{-1} is the probit function. This yields a roughly normally distributed variable Z with elements z_g that reflects the strength of the association each gene has with the phenotype, with higher values corresponding to stronger associations.
- Gene based and Gene set based analysis are included as feature of FUMA webserver

FUMA : interrogation of GWAS

Article | [Open Access](#) | Published: 28 November 2017

Functional mapping and annotation of genetic associations with FUMA

Kyoko Watanabe, Erdogan Taskesen, Arjen van Bochoven & Danielle Posthuma 

Nature Communications **8**, Article number: 1826 (2017) | [Cite this article](#)

9563 Accesses | **170** Citations | **23** Altmetric | [Metrics](#)

Abstract

A main challenge in genome-wide association studies (GWAS) is to pinpoint possible causal variants. Results from GWAS typically do not directly translate into causal variants because the majority of hits are in non-coding or intergenic regions, and the presence of linkage disequilibrium leads to effects being statistically spread out across multiple variants. Post-GWAS annotation facilitates the selection of most likely causal variant(s). Multiple resources are available for post-GWAS annotation, yet these can be time consuming and do not provide integrated visual aids for data interpretation. We, therefore, develop FUMA: an integrative web-based platform using information from multiple biological resources to facilitate functional annotation of GWAS results, gene prioritization and interactive visualization. FUMA accommodates positional, expression quantitative trait loci (eQTL) and chromatin interaction mappings, and provides gene-based, pathway and tissue enrichment results. FUMA results directly aid in generating hypotheses that are testable in functional experiments aimed at proving causal relations.

<http://fuma.ctglab.nl/>

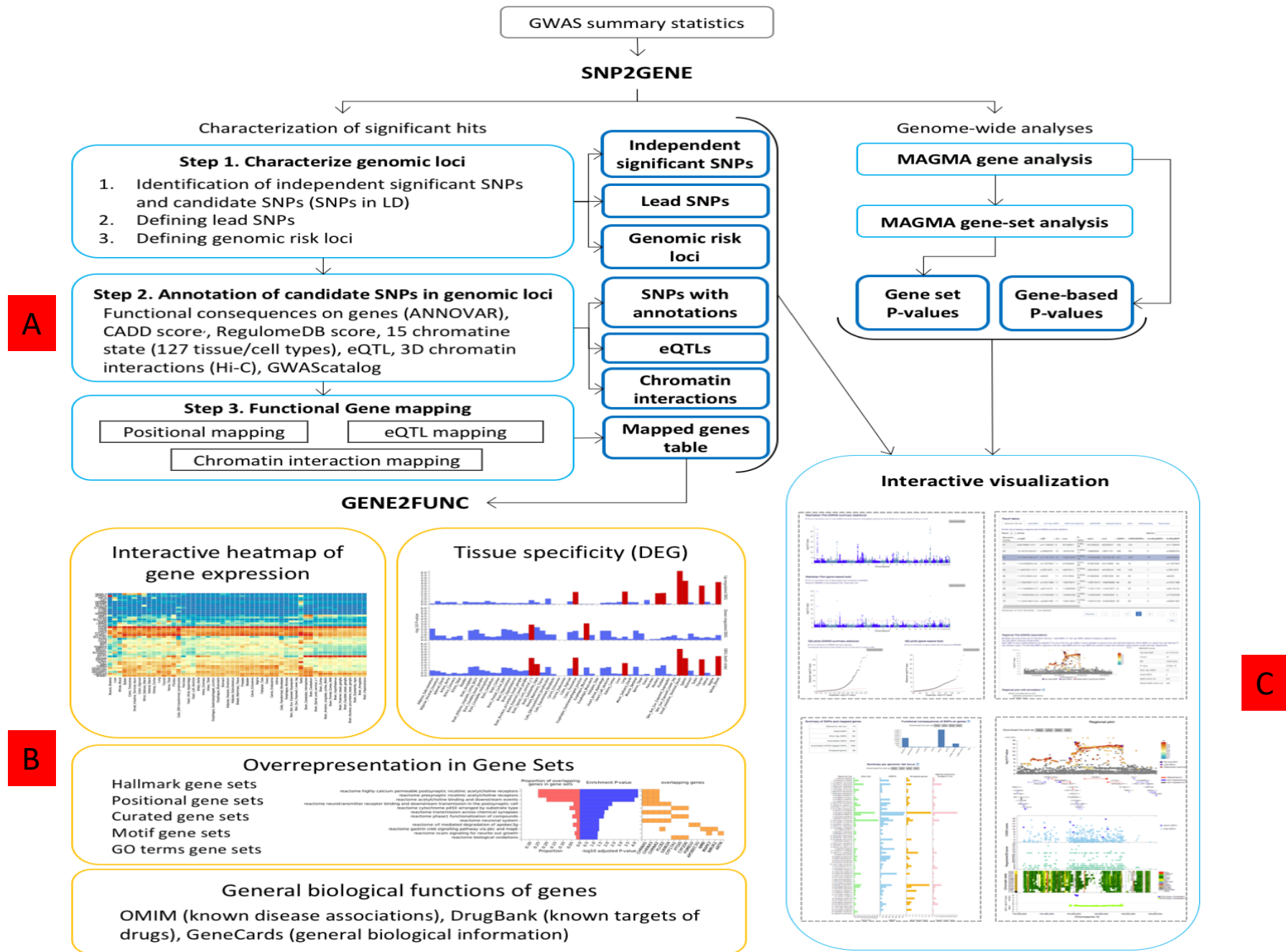
FUMA : Muti Steps

- The main purpose of FUMA is to use functional, biological information to prioritize genes based on GWAS outcomes.
- FUMA consists of two separate process; SNP2GENE and GENE2FUNC.
- To annotate and prioritize SNPs and genes from your GWAS summary statistics, go to SNP2GENE which compute LD structure, annotates functions to SNPs, and prioritize candidate genes.
- You can then use the prioritized genes as input to GENE2FUNC to check expression patterns and shared molecular functions between genes. GENE2FUNC can also be used for any list of pre-selected genes (i.e. created outside of SNP2GENE).

FUMA : Discuss

<https://www.nature.com/articles/s41467-017-01261-5>

Ready to use FUMA Webserver !!!



FUMA GWAS

Functional Mapping and Annotation of genome-wide association results

2) Login

1) Register

FUMA is a platform that can be used to annotate, prioritize and visualize and interpret GWAS results.

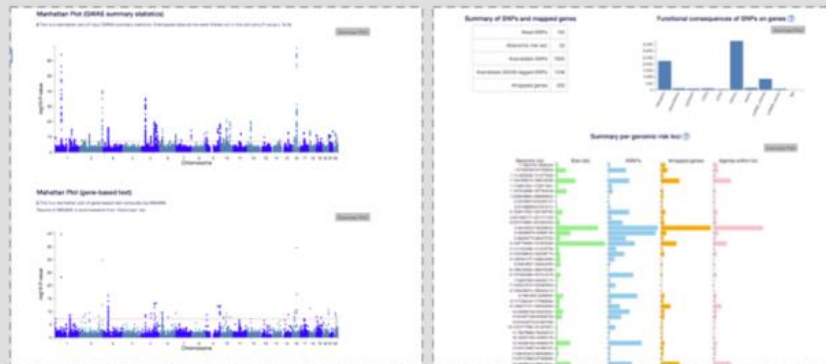
The [SNP2GENE](#) function takes GWAS summary statistics or a list of rsid's as input, and provides extensive functional annotation for all SNPs in genomic areas identified by lead SNPs.

The [GENE2FUNC](#) function takes a list of geneids (as identified by SNP2GENE or as provided manually) and annotates genes in biological context

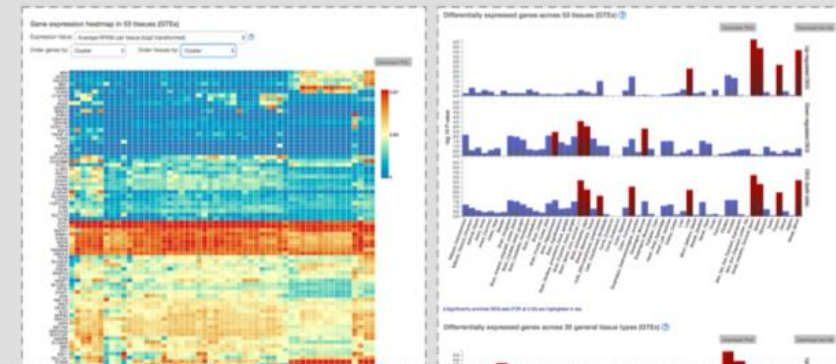
Please log in to use FUMA. If you have't registered yet, you can do from [here](#).

When using FUMA, please acknowledge Watanabe et al. xxx

SNP2GENE



GENE2FUNC



2. Submit new job at SNP2GENE

A new job starts with a GWAS summary statistics file. A variety of file formats are supported. Please refer the section of [Input files](#) for details. If your input file is an output from PLINK, SNPTEST or METAL, you can directly submit the file without specifying column names.

The input GWAS summary statistics file could be a subset of SNPs (e.g. only SNPs which are interesting in your study), but in this case, MAGMA results are not relevant anymore.

Optionally, if you would like to pre-specify lead SNPs, you can upload a file with 3 columns; rsID, chromosome and position. FUMA will then use these SNPs to select LD-related SNPs for annotation and mapping, instead of using lead SNPs identified by FUMA (it requires to disable an option for "identify additional lead SNPs").

In addition, if you are interested in specific genomic regions, you can also provide them by uploading a file with 3 columns; chromosome, start and end position. FUMA will then use these genomic regions to select LD-related SNPs for annotation and mapping, instead of determining the regions itself.

The screenshot shows the FUMA GWAS web interface. The top navigation bar includes links for Home, Tutorial, **SNP2GENE** (highlighted with a red box), GENE2FUNC, Links, and example. On the left sidebar, the 'New Job' button is highlighted with a red box and an arrow pointing to it from a grey box labeled 'Submit new job'. The main content area is titled 'Upload your GWAS summary statistics and set parameters to obtain functional annotations of the genomic loci associated with your trait'. Below this, the '1. Upload input files' section is highlighted with a pink background. It contains a 'Choose file' button (labeled 'No file chosen') and a text box for 'GWAS summary statistics' with a help icon. A red box labeled 'Mandatory input' is next to the file upload area. Below the file upload, there are input fields for 'Chromosome:', 'Position:', 'rsID:', 'P-value:', 'Risk allele:', 'Other allele:', and 'OR:'. A blue box with an information icon contains the text: 'Optional. Please fill as much as you can. It is not necessary to fill all column names.'

3. Set parameters

- On the same page as where you specify the input files, there are a variety of optional parameters that control the prioritization of genes.
- Please check your parameters carefully. The default settings are to perform identification of independent genome-wide significant SNPs at r^2 0.6 and lead SNPs at r^2 0.1, to map SNPs to genes up to 10kb apart.
- To filter SNPs by specific functional annotations and to use eQTL mapping, please change parameters
- If all inputs are valid, 'Submit Job' button will be activated. Once you submit a job, this will be listed in My Jobs.

The image shows two screenshots of the FUMA GWAS web application. The top screenshot is the 'New Job' page where users upload GWAS summary statistics and set parameters. The bottom screenshot is the 'My Jobs' page showing a list of submitted jobs.

Top Screenshot: New Job Page

Navigation: Home, Tutorial, **SNP2GENE**, GENE2FUNC, Links, example ▾

Left sidebar: New Job (active), My Jobs

Header: Upload your GWAS summary statistics and set parameters to obtain functional annotations of the genomic loci associated with your trait

Parameters:

1. Upload input files
2. Parameters for lead SNPs and candidate SNPs identification
- 3-1. Gene Mapping (positional mapping)
- 3-2. Gene Mapping (eQTL mapping)
4. Gene types
5. MHC region
6. Title of job submission

Submit Job button (highlighted with a red box and an arrow pointing to it with the text 'Click to Submit Job'). A text overlay says: 'Make sure all parameters here have non-red message!!'

Bottom Screenshot: My Jobs Page

Navigation: Home, Tutorial, **SNP2GENE**, GENE2FUNC, Links, example ▾

Left sidebar: New Job, My Jobs (active)

Header: My Jobs

List of Jobs (refresh icon)

Buttons: Delete selected jobs

Job ID	Job name	Submit date	Status ?	Select
89	example	2017-01-19 14:31:01	NEW	<input type="checkbox"/>
22	example2	2016-12-23 13:31:37	Go to results	<input type="checkbox"/>
20	example3	2016-12-23 13:31:37	Go to results	<input type="checkbox"/>

A red box highlights the first row of the table. A text overlay says: 'Submitted job will appear here'.

4. Check your results

After you submit files and parameter settings, a JOB has the status NEW which will be updated to QUEUES to RUNNING. Depending on the number of significant genomic regions, this may take between a couple of minutes and an hour. Once a JOB has finished running, you will receive an email. Unless an error occurred during the process, the email includes the link to the result page (this again requires login). You can also access to the results page from My Jobs page.

The result page displays 4 additional side bars.

Genome-wide plots: Manhattan plots and Q-Q plots for GWAS summary statistics and gene-based test by MAGMA, results of MAGMA gene-set analysis and tissue expression analysis.

Summary of results: Summary of results such as the number of lead and LD-related SNPs, and mapped genes for overall and per identified genomic risk locus.

Results: Tables of lead SNPs, genomic risk loci, candidate SNPs with annotations, eQTLs (only when eQTL mapping is performed), mapped genes and GWAS-catalog reported SNPs matched with candidate SNPs. You can also create interactive regional plots with functional annotations from this tab.

Downloads: Download all results as text files.



1. Input files

Parameter	Mandatory	Description	Type	Default
GWAS summary statistics	Mandatory	Input file of GWAS summary statistics. Plain text file or zipped or gzipped files are acceptable. The maximum file size which can be uploaded is 600Mb. As well as full results of GWAS summary statistics, subset of results can also be used. e.g. If you would like to look up specific SNPs, you can filter out other SNPs. Please refer to the Input files section for specific file format.	File upload	none
Pre-defined lead SNPs	Optional	Optional pre-defined lead SNPs. The file should have 3 columns, rsID, chromosome and position.	File upload	none
Identify additional lead SNPs	Optional only when predefined lead SNPs are provided	If this option is CHECKED, FUMA will identify additional independent lead SNPs after defining the LD block for pre-defined lead SNPs. Otherwise, only given lead SNPs and SNPs in LD of them will be used for further annotations.	Check	Checked
Pre-defined genetic region	Optional	Optional pre-defined genomic regions. FUMA only looks at provided regions to identify lead SNPs and SNPs in LD of them. If you are only interested in specific regions, this option will increase the speed of process.	File upload	none

FUMA : Parameter detail

Parameter	Mandatory	Description	Type	Default	Direction
Sample size (N)	Mandatory	The total number of individuals in the GWAS or the number of individuals per SNP. This is only used for MAGMA to compute the gene-based P-values. For total sample size, input should be an integer. When the input file of GWAS summary statistics contains a column of sample size per SNP, the column name can be provided in the second text box. i When column name is provided, please make sure that the column only contains integers (no float or scientific notation). If there are any float values, they will be rounded up by FUMA.	Integer or text	none	Does not affect any candidates
Maximum lead SNP P-value (\leq)	Mandatory	FUMA identifies lead SNPs with P-value less than or equal to this threshold and independent from each other.	numeric	5e-8	lower: decrease #lead SNPs. higher: increase #lead SNPs.
Maximum GWAS P-value (\leq)	Mandatory	This is the P-value threshold for candidate SNPs in LD of independent significant SNPs. This will be applied only for GWAS-tagged SNPs as SNPs which do not exist in the GWAS input but are extracted from 1000 genomes reference do not have P-value.	numeric	0.05	higher: decrease #candidate SNPs. lower: increase #candidate SNPs.
r^2 threshold for independent significant SNPs (\geq)	Mandatory	The minimum r^2 for defining independent significant SNPs, which is used to determine the borders of the genomic risk loci. SNPs with $r^2 \geq$ user defined threshold with any of the detected independent significant SNPs will be included for further annotations and are used for gene prioritisation.	numeric	0.6	higher: decrease #candidate SNPs and increase #independent significant SNPs. lower: increase #candidate SNPs and decrease #independent significant SNPs.
2nd r^2 threshold for lead SNPs (\geq)	Mandatory	The minimum r^2 for defining lead SNPs, which is used for the second clumping (clumping of the independent significant SNPs). Note that when this threshold is same as the first r^2 threshold, lead SNPs are identical to independent significant SNPs.	numeric	0.1	higher: increase #lead SNPs. lower: decrease #lead SNPs.
Reference panel	Mandatory	The reference panel to compute r^2 and MAF. Five populations from 1000 genomes Phase 3 and 3 versions of UK Biobank are available. See here for details.	Select	1000G Phase EUR	-
Include variants from reference panel	Mandatory	If Yes, all SNPs in strong LD with any of independent significant SNPs including non-GWAS-tagged SNPs will be included and used for gene mapping.	Yes/No	Yes	-
Minimum MAF (\geq)	Mandatory	The minimum Minor Allele Frequency to be included in annotation and prioritisation. MAF is based the user selected reference panel. This filter also applies to lead SNPs. If there is any pre-defined lead SNPs with MAF less than this threshold, those SNPs will be skipped. When this value is 0 (by default), SNPs with MAF>0 are considered.	numeric	0	higher: decrease #candidate SNPs. lower: increase #candidate SNPs.
Maximum distance of LD blocks to merge (\leq)	Mandatory	This is the maximum distance between LD blocks of independent significant SNPs to merge into a single genomic locus. When this is set at 0, only physically overlapping LD blocks are merged. Defining genomic loci does not affect identifying which SNPs fulfil selection criteria to be used for annotation and prioritization. It will only result in a different number of reported risk loci, which can be desired when certain loci are partly overlapping or physically very close.	numeric	250kb	higher: decrease #genomic loci. lower: increase #genomic loci.

3.1 Positional mapping

Parameter	Mandatory	Description	Type	Default	Direction
Positional mapping	Optional	Check this option to perform positional mapping. Positional mapping is based on ANNOVAR annotations by specifying the maximum distance between SNPs and genes or based on functional consequences of SNPs on genes. These parameters can be specified in the option below.	Check	Checked	-
Distance to genes or functional consequences of SNPs on genes to map	Mandatory if positional mapping is activated.	<p>Positional mapping criterion either map SNPs to genes based on physical distances or functional consequences of SNPs on genes.</p> <p>When maximum distance is provided SNPs are mapped to genes based on the distance given the user defined maximum distance. Alternatively, specific functional consequences of SNPs on genes can be selected which filtered SNPs to map to genes. Note that when functional consequences are selected, all SNPs are locating on the gene body (distance 0) except upstream and downstream SNPs which are up to 1kb apart from TSS or TSE.</p> <p>i When the maximum distance is set at > 0kb and < 1kb all upstream and downstream SNPs are included since the actual distance is not provided by ANNOVAR. Therefore, the maximum distance > 0kb and < 1kb is same as the maximum distance 1 kb.</p> <p>i For SNPs which are locating on a genomic region where multiple genes are overlapped, ANNOVAR has its own prioritization criteria to report the most deleterious function. For those SNPs, only prioritized annotations are used.</p>	Integer / Multiple selection	Maximum distance 10 kb	-

3.2 eQTL mapping

Parameter	Mandatory	Description	Type	Default	Direction
eQTL mapping	Optional	Check this option to perform eQTL mapping. eQTL mapping will map SNPs to genes which likely affect expression of those genes up to 1 Mb (cis-eQTL). eQTLs are highly tissue specific and tissue types can be selected in the following option. eQTL mapping can be used together with positional mapping.	Check	Unchecked	-
Tissue types	Mandatory if eQTL mapping is CHECKED	All available tissue types with data sources are shown in the select boxes. From FUMA v1.3.0, GTEx v7 became available but GTEx v6 are kept available. Therefore, when "all" is selected, both GTEx v6 and v7 are used for mapping. For detail of eQTL data resources, please refer to the eQTL section in this tutorial.	Multiple selection	none	-
eQTL maximum P-value (\leq)	Optional	The P-value threshold of eQTLs. Two options are available, Use only significant snp-gene pairs or nominal P-value threshold. When Use only significant snp-gene pairs is checked, only eQTLs with $FDR \leq 0.05$ will be used. Otherwise, defined nominal P-value is used to filter eQTLs. i Some of eQTL data source only contained eQTLs with a certain FDR threshold. Please refer to the eQTLs section for details of each data sources.	Check / Numeric	Checked / 1e-3	lower: increase #eQTLs and #mapped genes. higher: decrease #eQTLs and #mapped genes.

3.3 Chromatin interaction mapping

Parameter	Mandatory	Description	Type	Default	Direction
chromatin interaction mapping	Optional	Check this option to perform chromatin interaction mapping.	Check	Unchecked	-
Builtin chromatin interaction data	Optional	Build in chromatin interaction data can be selected in this option. Details of available build in data are available in the Chromatin interactions section in this tutorial.	Multiple selection	none	-
Custom chromatin interaction matrices	Optional	In addition to build in chromatin interaction data, user can upload custom data. The data should be pre-computed chromatin loops with significance (ideally FDR but another score can be used, see the Chromatin interactions section for details). The file should be gzipped and named as "(name-of-data).txt.gz". Multiple files can be uploaded. For each data, user can also provide data type, such as Hi-C, ChIA-PET or C5 which is not mandatory but will be used in the result table and regional plot. The file format is described in the Chromatin interactions section in this tutorial. ⚠ Please avoid uploading more than one file with identical file names. In that case, the files are over-written by the last uploaded one.	File upload (multiple)	none	-
FDR threshold (α)	Mandatory if chromatin interaction mapping is CHECKED	FDR threshold for significant loops. The default value is set at 1e-6 which is suggested by Schmitt et al. (2016) ⚠ This threshold will be applied both build in and user uploaded chromatin loops.	Numeric	1e-6	lower: increase #chromatin interactions and #mapped genes. higher: decrease #chromatin interactions and #mapped genes.
Promoter region window	Mandatory if chromatin interaction mapping is CHECKED	Promoter regions of genes to map in significantly interacting regions. The input format should be "(upstream bp)-(downstream bp)" from transcription start site (TSS). For example, the default "250-500" means that promoter regions are defined as 250bp upstream and 500bp downstream of the TSS. By the chromatin interaction mapping, genes whose user defined promoter regions are overlapped with the significantly interacting regions will be mapped. Please refer the Chromatin interactions section in this tutorial for details.	text	250-500	lower: increase #mapped genes. smaller: decrease #mapped genes.
Annotate enhancer/promoter regions (Roadmap 111 epigenomes)	Optional	Predicted enhancer and promoter regions from Roadmap epigenomics project for 111 epigenomes can be annotated to significantly interaction regions. If any epigenome is not selected, enhancer and promoter regions are not annotated. Annotated enhancer/promoter regions can be used to filter SNPs and mapped genes in the next two options.	Multiple selection	none	-
Filter SNPs by enhancers	Optional	This option is only available when at least one epigenome is selected in the previous option to annotate enhancer/promoter regions. When this option is checked, SNPs are filtered on such that overlap with one of the annotated enhancer regions for chromatin interaction mapping. Please refer the Chromatin interactions section in this tutorial for details.	Check	Unchecked	-
Filter genes by promoters	Optional	This option is only available when at least one epigenome is selected in the previous option to annotate enhancer/promoter regions. When this option is checked, chromatin interaction mapping is only performed for genes whose promoter regions are overlap with one of the annotated promoter regions. Please refer the Chromatin interactions section in this tutorial for details.	Check	Unchecked	-



3.4 Functional annotation filtering

Positional, eQTL and chromatin interaction mappings have the following options separately, for the filtering of SNPs based on functional annotation. All filters below apply to selected SNPs in LD with independent significant SNPs that are used to prioritize genes and influence the number of SNPs that are mapped to genes, and consequently influence the number of prioritized genes.

Parameter	Mandatory	Description	Type	Default	Direction
CADD score	Optional	Check this if you want to perform filtering of SNPs by CADD score. This applies to selected SNPs in LD with independent significant SNPs that are used to prioritize genes. CADD score is the score of deleteriousness of SNPs predicted by 63 functional annotations. 12.37 is the threshold to be deleterious suggested by Kicher et al (2014). Please refer to the original publication for details from links .	Check	Unchecked	-
Minimum CADD score (z)	Mandatory if CADD score is checked	The higher the CADD score, the more deleterious.	numeric	12.37	higher: less SNPs will be mapped to genes. lower: more SNPs will be mapped to genes.
RegulomeDB score	Optional	Check if you want to perform filtering of SNPs by RegulomeDB score. This applies to selected SNPs in LD with independent significant SNPs that are used to prioritize genes. RegulomeDB score is a categorical score representing regulatory functionality of SNPs based on eQTLs and chromatin marks. Please refer to the original publication for details from links .	Check	Unchecked	-
Minimum RegulomeDB score (z)	Mandatory if RegulomeDB score is checked	RegulomeDB score is a categorical score from 1a to 7) Score 1a means that those SNPs are most likely affecting regulatory elements and 7 means that those SNPs do not have any annotations. SNPs are recorded as NA if they are not present in the database. SNPs with NA will not be included for filtering on RegulomeDB score.	string	7	higher: more SNPs will be mapped to genes. lower: less SNPs will be mapped to genes.
15-core chromatin state	Optional	Check if you want to perform filtering of SNPs by chromatin state. This applies to selected SNPs in LD with independent significant SNPs that are used to prioritize genes. The chromatin state represents accessibility of genomic regions (every 200bp) with 15 categorical states predicted by ChromHMM based on 5 chromatin marks for 127 epigenomes.	Check	Unchecked	-
15-core chromatin state tissue/cell types	Mandatory if 15-core chromatin state is checked	Multiple tissue/cell types can be selected from the list.	Multiple selection	none	-
Maximum state of chromatin(s)	Mandatory if 15-core chromatin state is checked	The maximum state to filter SNPs. Between 1 and 15. Generally, between 1 and 7 is open state.	numeric	7	higher: more SNPs will be mapped to genes. lower: less SNPs will be mapped to genes.
Method for 15-core chromatin state filtering	Mandatory if 15-core chromatin state is checked	When multiple tissue/cell types are selected, either any (filtered on SNPs which have state above than threshold in any of selected tissue/cell types), majority (filtered on SNPs which have state above than threshold in majority (≥50%) of selected tissue/cell type), or all (filtered on SNPs which have state above than threshold in all of selected tissue/cell type).	Selection	any	-
Annotation datasets	Optional	Additional functional annotations can be annotated to candidate SNPs. All available data are regional based annotation (bed file format).	Multiple selection	none	-
Annotation filtering method	Mandatory if any of Annotation datasets is selected.	By default, SNPs are not filtered by the annotations selected in Annotation datasets . To filter SNPs based on the selected annotation, select this options from any (filtered on SNPs which are overlapping with any selected annotations), majority (filtered on SNPs which are overlapping with majority (≥50%) of selected annotations), or all (filtered on SNPs which are overlapping with all of selected annotations).	Selection	No filtering	-

4. Gene types

Biotype of genes to map can be selected. Please refer to Ensembl for details of biotypes.

Parameter	Mandatory	Description	Type	Default
Gene type	Mandatory	Gene type to map. This is based on gene_biotype obtained from BioMart of Ensembl build 85. Please see here for details	Multiple selection.	Protein coding genes.

5. MHC region

The MHC region is often excluded due to its complicated LD structure. Therefore, this option is checked by default. Please uncheck to include MHC region. Note that it doesn't change any results if there is no significant hit in the MHC region.

Parameter	Mandatory	Description	Type	Default
Exclude MHC region	Optional	Check if you want to exclude the MHC region. The default region is defined as between "MOG" and "COL11A2" genes.	Check	Checked
Options for excluding MHC region	Optional	MHC region can be excluded only from either annotations or MAGMA gene analysis, or from both by selecting this option.	Select	Only from annotations
Extended MHC region	Optional	User specified MHC region to exclude (for extended or shorter region). The input format should be like "25000000-34000000" on hg19.	Text	Null

6. MAGMA analysis

MAGMA gene and gene-set analyses are performed for the input summary statistics by default, but user can also select to omit MAGMA process that reduce the run time of SNP2GENE process. Gene expression data sets for MAGMA gene expression analysis can be also selected from here.

Parameter	Mandatory	Description	Type	Default
Perform MAGMA	Optional	UNCHECK to SKIP MAGMA analyses.	Check	Checked
MAGMA gene annotation window	Mandatory when MAGMA is active.	The window of the genes to assign SNPs (symmetric). e.g. when 5kb is selected, SNPs within 5kb window of a gene (both side) will be assigned to that gene. The option is available from 0, 5, 10, 15, 20kb window.	Select	0kb from both side of the genes
MAGMA gene expression analysis	Mandatory when MAGMA is active.	Gene expression data sets used for MAGMA gene-property analysis to test positive association between genetic associations and gene expression in a given label.	Select	GTEEx v6

Gene expression database used by Fuma

Gene expression data sets

1. GTEx v6

Data source

RNAseq data set was downloaded from <http://www.gtexportal.org/home/datasets>. Gene level RPKM was used (GTEx_Analysis_v6_RNA-seq_RNA-SeQCv1.1.8_gene_rpkm.gct.gz).

Pre-process

Primary gene ID was Ensemble ID. In total, 8,555 samples were available. From 56,318 annotated genes, genes were filtered on such that average RPKM per tissue is >1 in at least on of the 53 tissues. This resulted in 28,577 genes. RPKM was winsorized at 50 (replaced $\text{RPKM} > 50$ with 50). Then average of log transformed RPKM with pseudocount 1 ($\log_2(\text{RPKM}+1)$) per tissue (for either 53 detail or 30 general tissues) was used as the covariates conditioning on the average across all the tissues.

2. GTEx v7

Data source

RNAseq data set was downloaded from <http://www.gtexportal.org/home/datasets>. Gene level TPM was used (GTEx_Analysis_2016-01-15_v7_RNASeQCv1.1.8_gene_rpm.gct.gz).

Pre-process

Primary gene ID was Ensemble ID. In total, 11,688 samples were available. From 56,203 annotated genes, genes were filtered on such that average TPM per tissue is >1 in at least on of the 53 tissues. This resulted in 32,335 genes. TPM was winsorized at 50 (replaced $\text{TPM} > 50$ with 50). Then average of log transformed TPM with pseudocount 1 ($\log_2(\text{TPM}+1)$) per tissue (for either 53 detail or 30 general tissues) was used as the covariates conditioning on the average across all the tissues.

3. BrainSpan

Data source

RNAseq data set was downloaded from <http://www.brainspan.org/static/download>. Gene level RPKM was used (genes_matrix_csv.zip).

Pre-process

Primary gene ID was Ensemble ID. In total, 524 samples were available. General developmental stages were annotated for each sample based on the age. We used 11 developmental stages and 29 ages as the label. For the label of age, we excluded age groups with <3 samples (25 pcw and 35 pcw). From 52,376 annotated genes, genes were filtered on such that average RPKM per label is >1 in at least one of the either developmental stage or age. This resulted in 19,601 and 21,001 genes for developmental stages and age groups, respectively. RPKM was winsorized at 50 (replaced $\text{RPKM} > 50$ with 50). Then average of log transformed RPKM with pseudocount 1 ($\log_2(\text{RPKM}+1)$) per label (for either 11 developmental stages or 29 age groups) was used as the covariates conditioning on the average across all the labels.

Fuma : Genomic risk loci Identification

Characterization of genomic risk loci based on GWAS

To define genomic loci of interest to the trait based on provided GWAS summary statistics, pre-calculated LD structure based on 1000G of the relevant reference population (EUR for BMI, CD and SCZ) is used. First of all, independent significant SNPs with a genome-wide significant P-value ($< 5e-8$) and independent from each other at $r^2 < 0.6$ are identified. For each independent significant SNP, all known (i.e., regardless of being available in the GWAS input) SNPs that have $r^2 \geq 0.6$ with one of the independent significant SNPs are included for further annotation (candidate SNPs). These SNPs may thus include SNPs that were not available in the GWAS input, but are available in the 1000G reference panel and are in LD with an independent significant SNP. Candidate SNPs can be filtered based on a user-defined minor allele frequency (MAF, ≥ 0.01 by default).

Based on the identified independent significant SNPs, independent lead SNPs are defined if they are independent from each other at $r^2 < 0.1$. Additionally, if LD blocks of independent significant SNPs are closely located to each other (< 250 kb based on the most right and left SNPs from each LD block), they are merged into one genomic locus. Each genomic locus can thus contain multiple independent significant SNPs and lead SNPs.

Besides using FUMA to determine lead SNPs based on GWAS summary statistics, users can provide a list of pre-defined lead SNPs. In addition, users can provide a list of pre-defined genomic regions to limit all annotations carried out by FUMA to those regions.

Fuma : Gene and Gene set analysis

MAGMA for gene analysis and gene set analysis

FUMA uses input GWAS summary statistics to compute gene-based P-values (gene analysis) and gene set P-value (gene set analysis) using the MAGMA³⁵ tool. For gene analysis, the gene-based P-value is computed for protein-coding genes by mapping SNPs to genes if SNPs are located within the genes. For gene set analysis, the gene set P-value is computed using the gene-based P-value for 4728 curated gene sets (including canonical pathways) and 6166 GO terms obtained from MsigDB v5.2. For both analyses, the default MAGMA setting (SNP-wise model for gene analysis and competitive model for gene set analysis) are used, and the Bonferroni correction (gene) or FDR (gene-set) was used to correct for multiple testing. 1000G phase 3²⁷ is used as a reference panel to calculate LD across SNPs and genes.

Lets run SNP2GENE

1. Upload input files

GWAS summary statistics ?	<div>Choose File No file chosen</div> <div><input checked="" type="checkbox"/> : Use example input (Crohn's disease, Franke et al. 2010).</div>	✓ OK. An example file will be used.
GWAS summary statistics file columns ?	<div>i case insensitive</div> <div>Chromosome: <input type="text"/></div> <div>Position: <input type="text"/></div> <div>rsID: <input type="text"/></div> <div>P-value: <input type="text"/></div> <div>Effect allele*: <input type="text"/></div> <div>**A1* is effect allele by default</div> <div>Non effect allele: <input type="text"/></div> <div>OR: <input type="text"/></div> <div>Beta: <input type="text"/></div> <div>SE: <input type="text"/></div>	<div>Optional. Please fill as much as you can. It is not necessary to fill all column names.</div>
Pre-defined lead SNPs ?	<div>Choose File No file chosen</div>	Optional.
Identify additional independent lead SNPs ?	<input checked="" type="checkbox"/>	Optional. This is only valid when predefined lead SNPs are provided.
Predefined genomic region ?	<div>Choose File No file chosen</div>	Optional.

2. Parameters for lead SNPs and candidate SNPs identification

Sample size (N) ?	<div>Total sample size (integer): 21389</div> <div>OR</div> <div>Column name for N per SNP (text): <input type="text"/></div>	✓ OK. The total sample size will be applied to all SNPs.
Minimum P-value of lead SNPs (<)	5e-8	✓ OK
Maximum P-value cutoff (< ?)	0,05	✓ OK
r ² threshold to define independent significant SNPs (≥)	0,6	✓ OK
2nd r ² threshold to define lead SNPs (≥ ?)	0,1	✓ OK
Reference panel population	1000G Phase3 EUR	✓ OK
Include variants in reference panel (non-GWAS tagged SNPs in LD) ?	Yes	✓ OK
Minimum Minor Allele Frequency (≥) ?	0	✓ OK
Maximum distance between LD blocks to merge into a locus (< kb) ?	250 kb	

3-1. Gene Mapping (positional mapping)

Positional mapping		
Perform positional mapping ?	<input checked="" type="checkbox"/>	✓ OK.
Distance to genes or functional consequences of SNPs on genes to map ?	<div>Maximum distance: <input type="text" value="10"/> kb</div> <div>OR</div> <div>Functional consequences of SNPs on genes: <div>clear</div><div>exonic splicing intronic 3UTR 5UTR</div></div>	✓ OK. SNPs are mapped to genes up to 10 kb
Optional SNP filtering by functional annotations for positional mapping i This filtering only applies to SNPs mapped by positional mapping criterion. When eQTL mapping is also performed, this filtering can be specified separately. All these annotations will be available for all SNPs within LD of identified lead SNPs in the result tables, but this filtering affect gene prioritization.		
CADD	Perform SNPs filtering based on CADD score. ?	<input type="checkbox"/> Optional.
	Minimum CADD score (≥) ?	<input type="text" value="12,37"/> Optional.
RegulomeDB	Perform SNPs filtering based on RegulomeDB score ?	<input type="checkbox"/> Optional.
	Maximum RegulomeDB score (categorical) ?	<input type="text" value="7"/> Optional.
15-core chromatin state	Perform SNPs filtering based on chromatin state ?	<input type="checkbox"/> Optional.
	<div>Select all Clear</div> <div>Adrenal (1) E080 (Other) Fetal Adrenal Gland Blood (27) E029 (HSC & B-cell) Primary monocytes from peripheral blood E030 (HSC & B-cell) Primary neutrophils from peripheral blood E031 (HSC & B-cell) Primary B cells from cord blood E032 (HSC & B-cell) Primary B cells from peripheral blood E033 (Blood & T-cell) Primary T cells from cord blood E034 (Blood & T-cell) Primary T cells from peripheral blood E035 (HSC & B-cell) Primary hematopoietic stem cells</div>	Optional.
	Tissue/cell types for 15-core chromatin state i Multiple tissue/cell types can be selected.	
	15-core chromatin state maximum state ?	<input type="text" value="7"/> Optional.
	15-core chromatin state filtering method ?	<input type="text" value="any"/> Optional.

3-2. Gene Mapping (eQTL mapping)

eQTL mapping
Perform eQTL mapping ? <input type="checkbox"/> Optional.

3-3. Gene Mapping (3D Chromatin Interaction mapping)

chromatin interaction mapping
Perform chromatin interaction mapping ? <input type="checkbox"/> Optional.

4. Gene types

Ensembl version	v92	✓ OK.
Gene type ?	<div>All</div> <div>Protein coding</div> <div>lncRNA</div> <div>ncRNA</div> <div>Processed transcript</div>	✓ OK.
i Multiple gene type can be selected.		

5. MHC region

Exclude MHC region ?



from only annotations

✓ OK. Normal MHC region will be excluded from only annotations.

Extended MHC region ?

i.e.g. 25000000-33000000

Optional.

6. MAGMA analysis

Perform MAGMA ?



✓ OK. MAGMA will be performed.

Gene windows ?

0

kb

One value will set same window size both sides, two values separated by comma will set different window sizes for up- and downstream. e.g. 2,1 will set window sizes 2kb upstream and 1kb downstream of the genes.
Maximum window size is limited to 50.

✓ OK.

MAGMA gene expression analysis ?

GTEx v8: 54 tissue types
GTEx v8: 30 general tissue types
GTEx v7: 53 tissue types
GTEx v7: 30 general tissue types
GTEx v6: 52 tissue types

✓ OK.

Title of job submission:

trail

This is not mandatory, but job title might help you to track your jobs.

Submit Job

⚠ After submitting, please wait until the file is uploaded, and do not move away from the submission page.

My Jobs

List of Jobs



Delete selected jobs

Job ID

Job name

Submit date

Status 

Jump to GENE2FUNC

Publish

Select

60609

trail

2019-11-04 10:51:43

[Go to results](#)

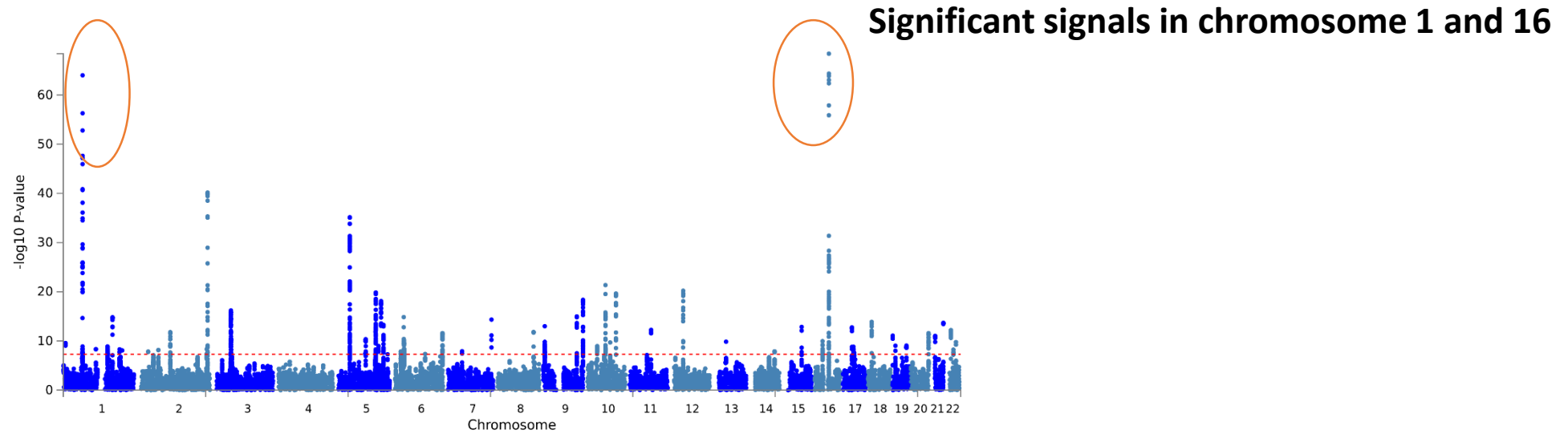
GENE2FUNC

Publish



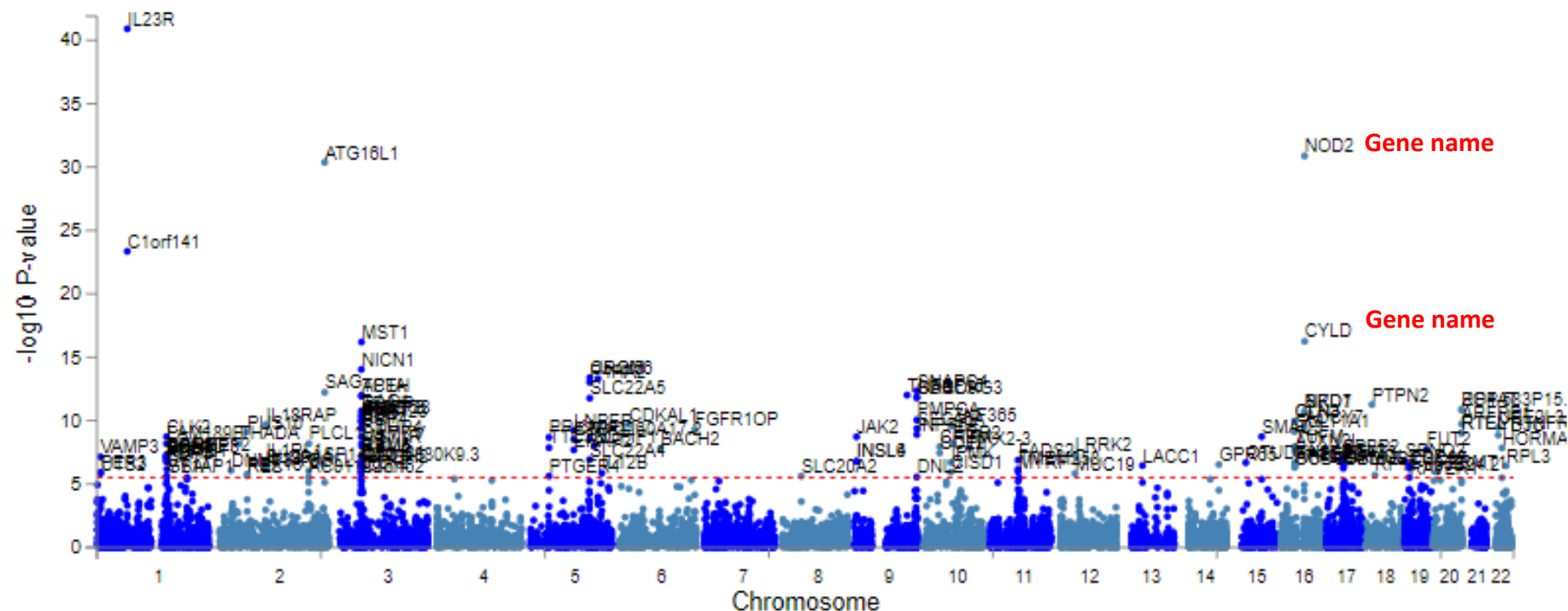
Result

GWAS PLOTS



Manhattan Plot (GWAS summary statistics)

GWAS PLOTS (gene based test)



i This is a manhattan plot of the gene-based test as computed by MAGMA based on your input GWAS summary statistics. The gene-based P-value is downloadable from 'Download' tab from the left side bar.

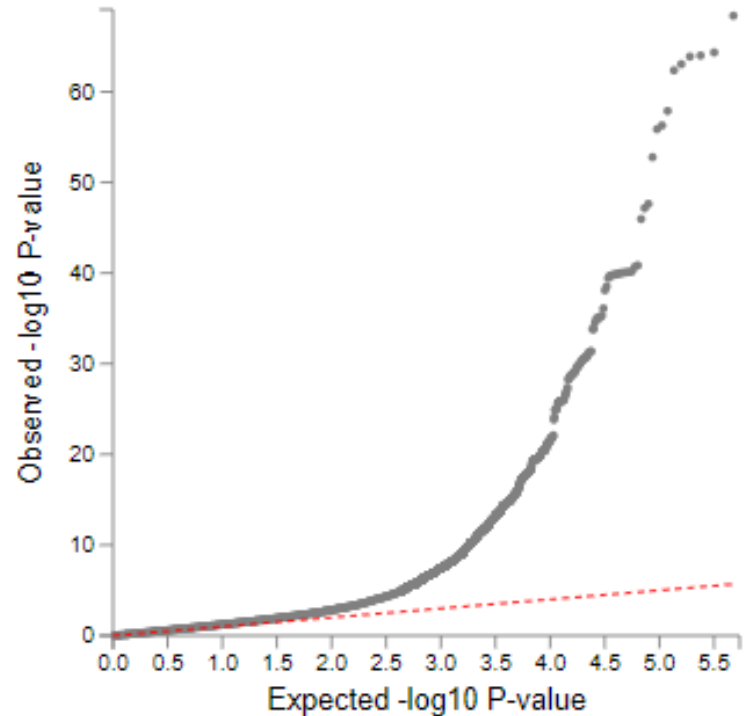
Input SNPs were mapped to 16510 protein coding genes. Genome wide significance (red dashed line in the plot) was defined at $P = 0.05/16510 = 3.028e-6$.

Q-Q PLOTS (GWAS/gene based test)

QQ plot (GWAS summary statistics)

i This is a Q-Q plot of GWAS summary statistics.
For plotting purposes, overlapping data points are not drawn (filtering was performed only for SNPs with P-value $\geq 1e-5$, see tutorial for details).

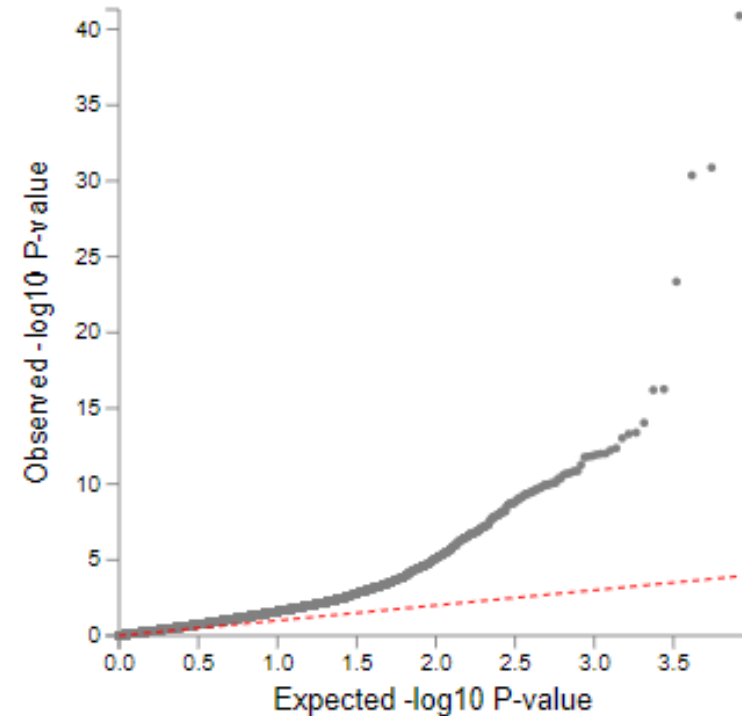
Download the plot as [PNG](#) [JPG](#) [SVG](#) [PDF](#)



QQ plot (gene-based test)

i This is a Q-Q plot of the gene-based test computed by MAGMA.

Download the plot as [PNG](#) [JPG](#) [SVG](#) [PDF](#)



Slight variation in Plots (From SNPs to Gene based QQ plot)

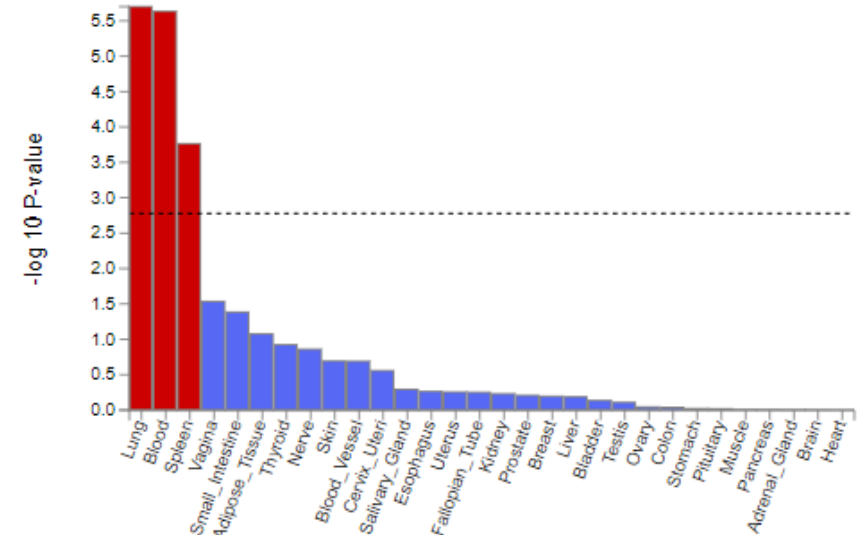
MAGMA gene set analysis

Over represented Gene ontology :

Gene Set	N genes	Beta	Beta STD	SE	P	P _{bon}
GO_bp:go_defense_response	1286	0.17241	0.046207	0.028022	3.9114e-10	6.05171808e-06
GO_bp:go_cytokine_production	627	0.22151	0.04234	0.039019	6.9857e-09	0.0001080757647
GO_bp:go_inflammatory_response	589	0.22743	0.042186	0.040501	9.9697e-09	0.000154231259
GO_bp:go_cytokine_mediated_signaling_pathway	614	0.21695	0.041054	0.038923	1.2674e-08	0.000196054106
GO_bp:go_positive_regulation_of_signaling	1541	0.13826	0.04022	0.025461	2.8612e-08	0.000442570416
GO_bp:go_response_to_cytokine	958	0.17057	0.039878	0.031566	3.3194e-08	0.000513411598
GO_bp:go_positive_regulation_of_intracellular_signal_transduction	845	0.17471	0.038502	0.033458	8.9777e-08	0.001388491082
Curated_gene_sets:reactome_signaling_by_interleukins	538	0.21607	0.038364	0.041524	9.9138e-08	0.00153316917
GO_bp:go_positive_regulation_of_rna_biosynthetic_process	1351	0.13521	0.037064	0.026186	1.2264e-07	0.00189650496
Curated_gene_sets:qi_plasmacytoma_up	208	0.3429	0.038246	0.067423	1.8522e-07	0.00286405686

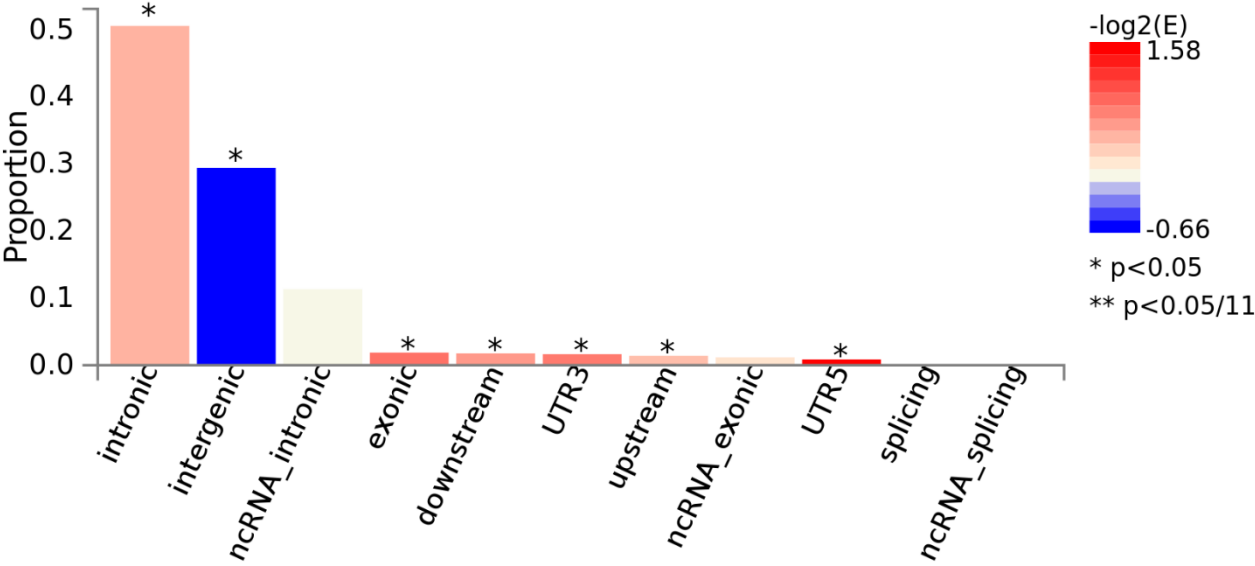
Defense response specific regulatory genes are highly significantly OR in this data.

Signifiant expression observed in Lung, Blood and spleen tissue.

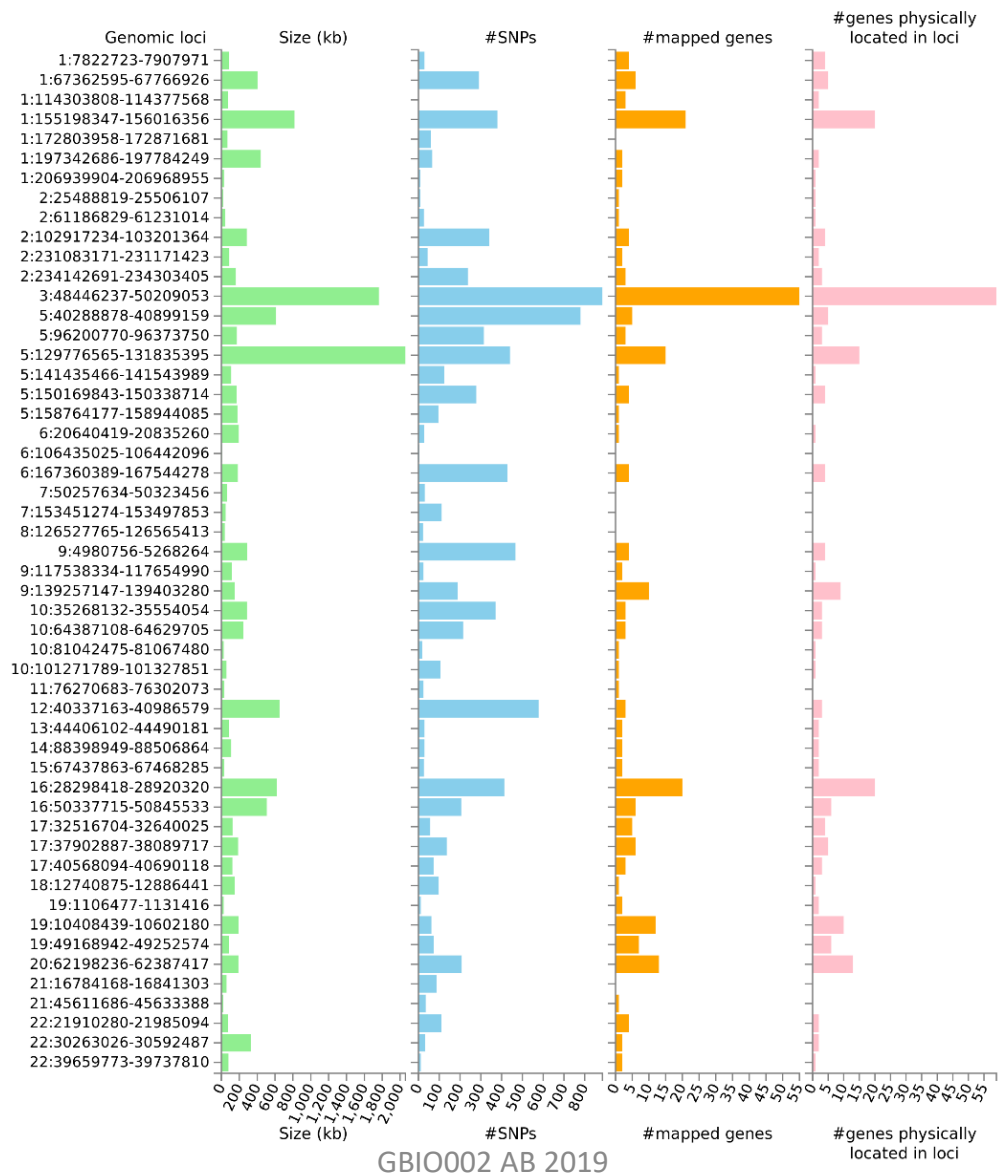


Summary of SNPs and mapped genes

#Genomic risk loci	52
#lead SNPs	75
#Ind. Sig. SNPs	164
#candidate SNPs	8717
#candidate GWAS tagged SNPs	1247
#mapped genes	256



Distribution of SNPs



Fuma : Regional Plots

Result tables

Genomic risk loci

lead SNPs

Ind. Sig. SNPs

SNPs (annotations)

ANNOVAR

Mapped Genes

GWAScatalog

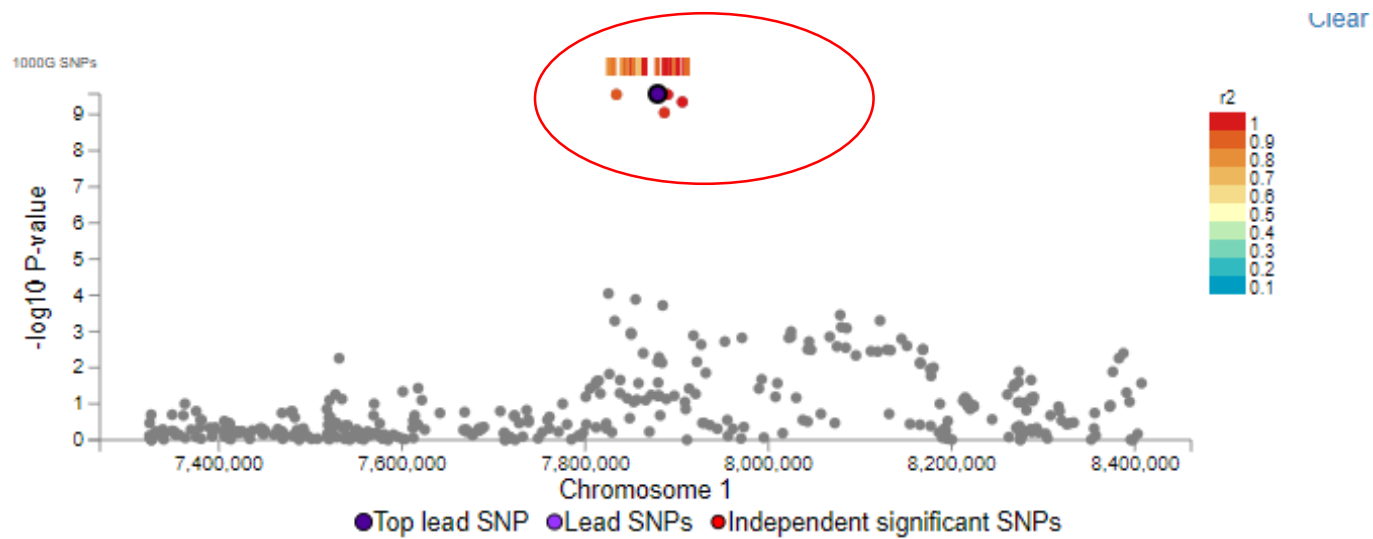
Parameters

Click row to display a regional plot of GWAS summary statistics.

Show 10 entries

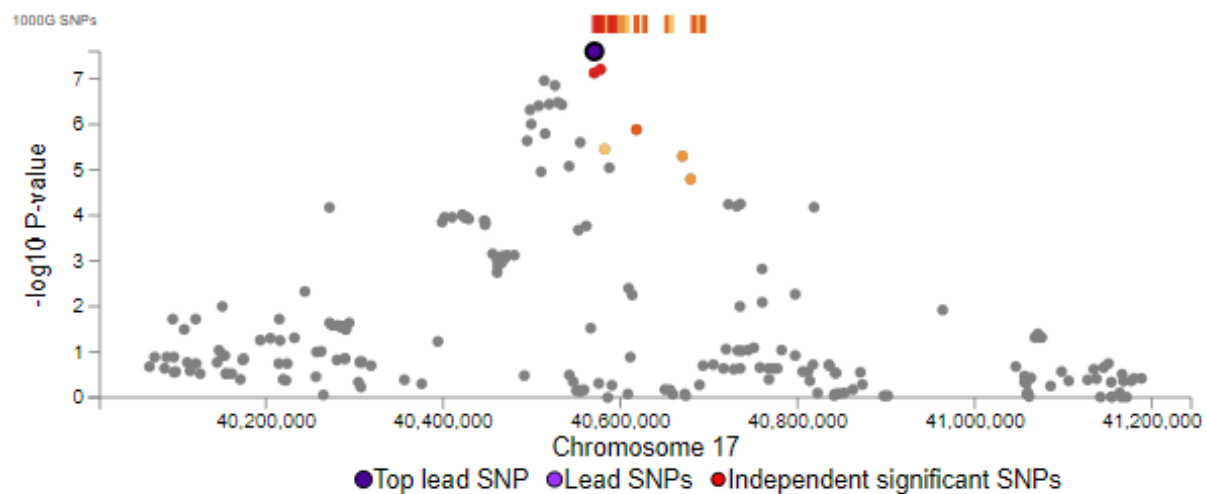
Search:

Genomic Locus	uniqID	rsID	chr	pos	P-value	start	end	nSNPs	nGWASSNPs	nIndSigSNPs	IndSigSNPs
21	6:106435025:A:G	rs6568421	6	106435025	4.4e-08	106435025	106442096	4	2	1	rs6568421
42	17:40570772:A:C	rs11871801	17	40570772	2.5e-08	40568094	40690118	72	7	1	rs11871801
8	2:25492467:A:G	rs13428812	2	25492467	1.4e-08	25488819	25506107	9	2	1	rs13428812
20	6:20728731:C:T	rs6908425	6	20728731	1.4e-08	20640419	20835260	27	7	2	rs6908425;rs
36	14:88472595:C:T	rs8005161	14	88472595	1.3e-08	88398949	88506864	29	4	1	rs8005161
23	7:50304461:C:T	rs1456896	7	50304461	1.2e-08	50257634	50323456	30	5	1	rs1456896
7	1:206939904:A:G	rs3024505	1	206939904	8.3e-09	206939904	206968955	8	1	1	rs3024505
9	2:61224259:C:T	rs10181042	2	61224259	6.6e-09	61186829	61231014	26	6	1	rs10181042
51	22:30592487:C:G	rs713875	22	30592487	5.7e-09	30263026	30592487	32	8	1	rs713875
6	1:197727642:A:G	rs1998598	1	197727642	4.9e-09	197342686	197784249	66	11	1	rs1998598



Selected Locus

top lead SNP	rs6568421
Chrom	6
BP	106435025
P-value	4.4e-08
#Ind. Sig. SNPs	1
#lead SNPs	1
SNPs within LD	4
GWAS SNPs within LD	2



Moving from SNP2Gene to Gene2FUNC

<

New Job

Redo gene mapping

My Jobs

My Jobs

Q

My Jobs

List of Jobs

Job ID

Job name

Submit date

Status ?

Jump to GENE2FUNC

Publish

Select

60609

trail

2019-11-04 10:51:43

[Go to results](#)

GENE2FUNC

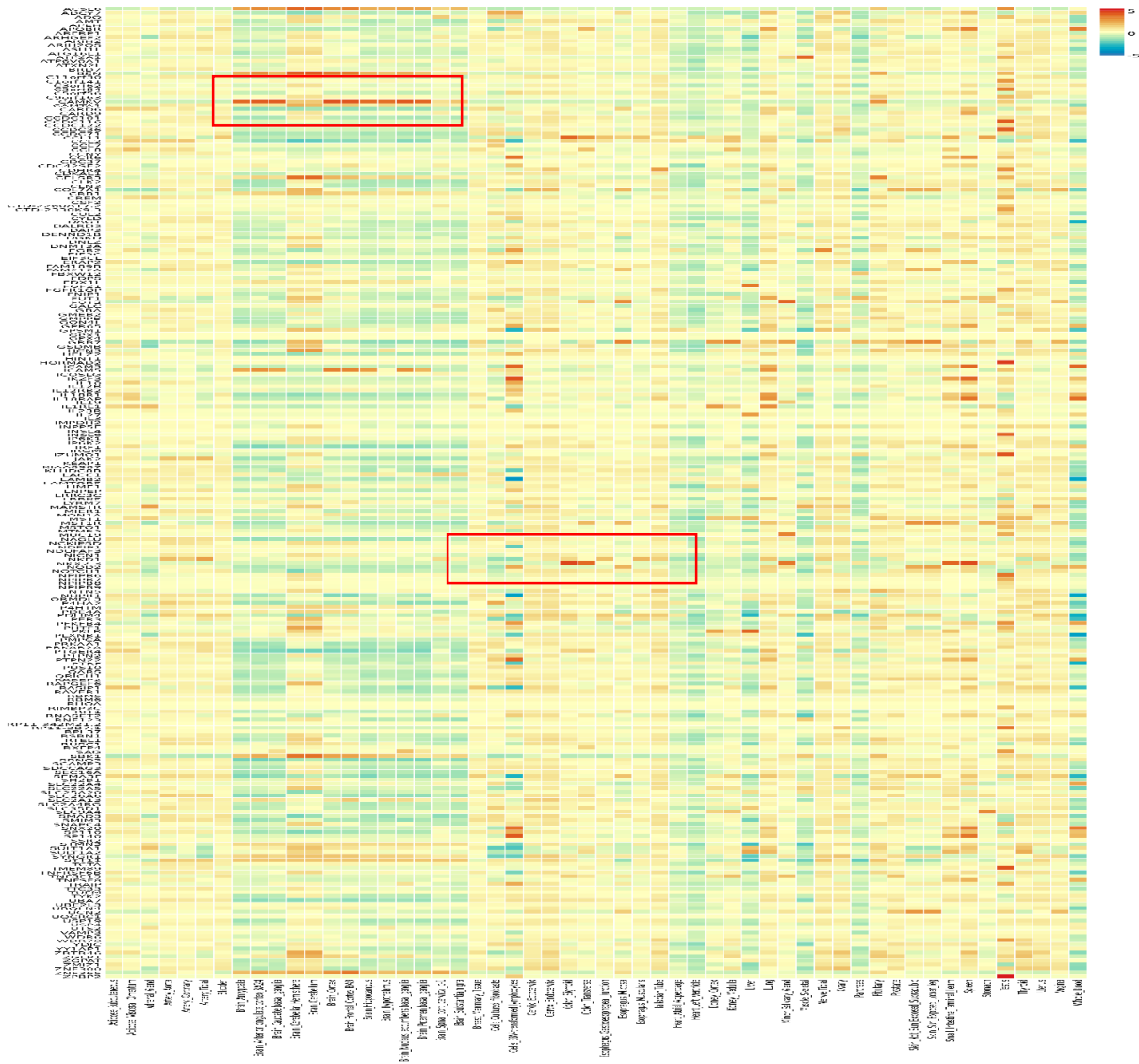
Publish

☐

Delete selected jobs

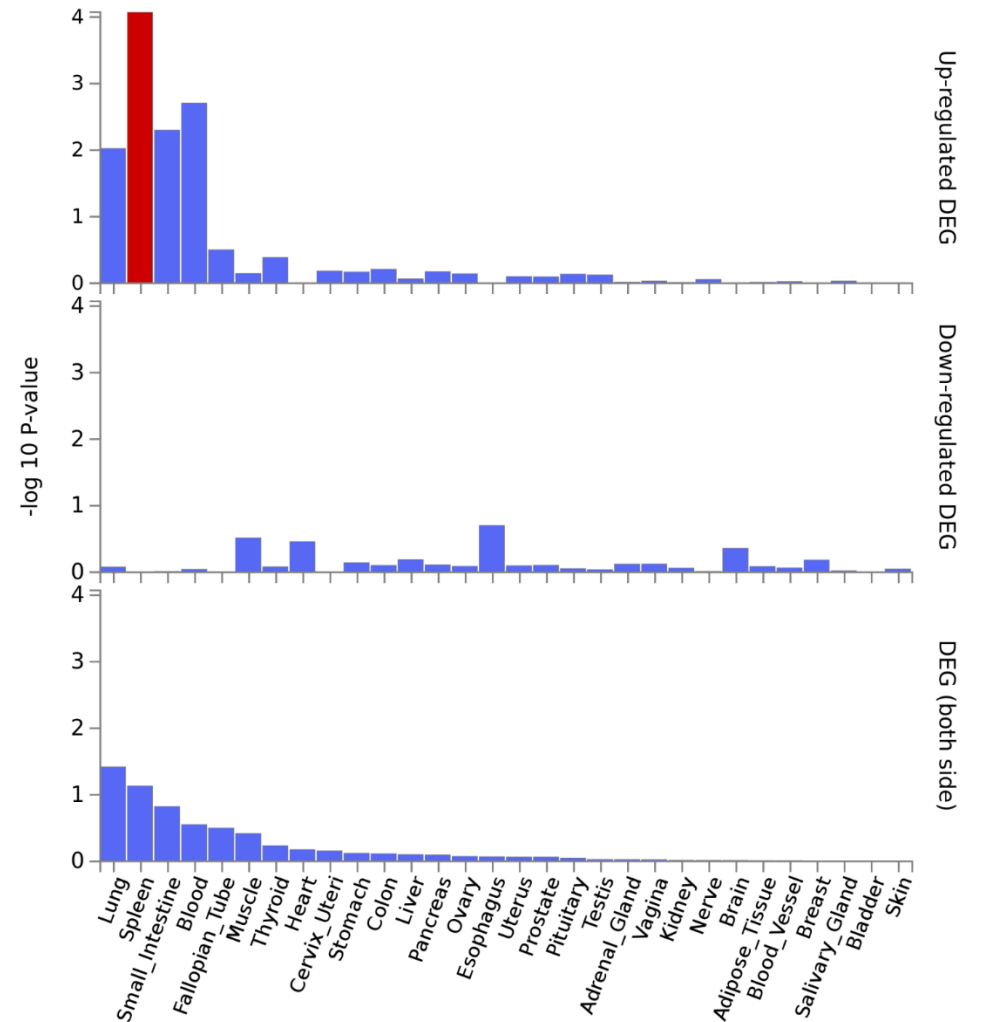
Click

Expression Heatmap plot



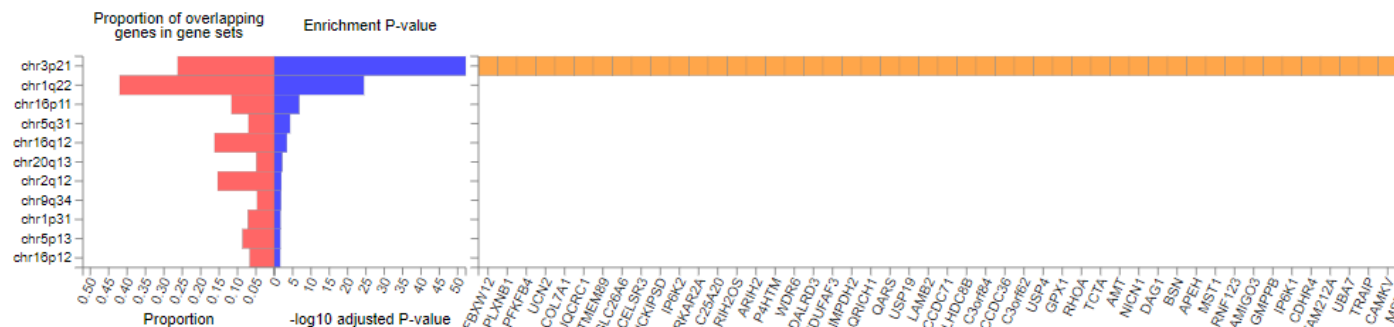
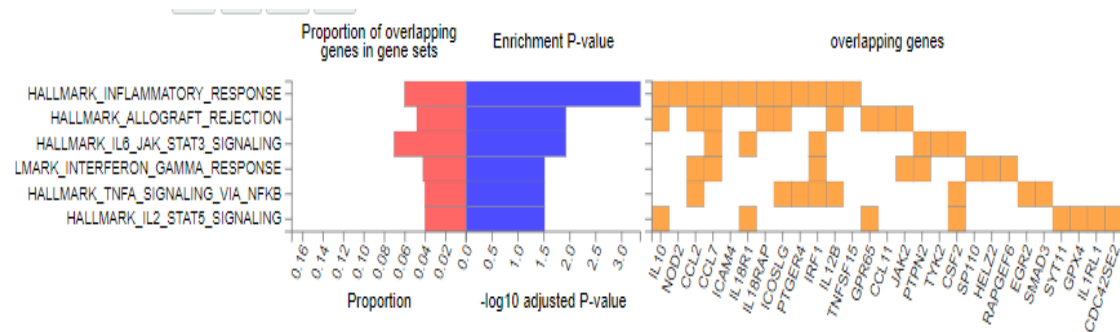
Dark red color : high expression

Tissue specific Expression

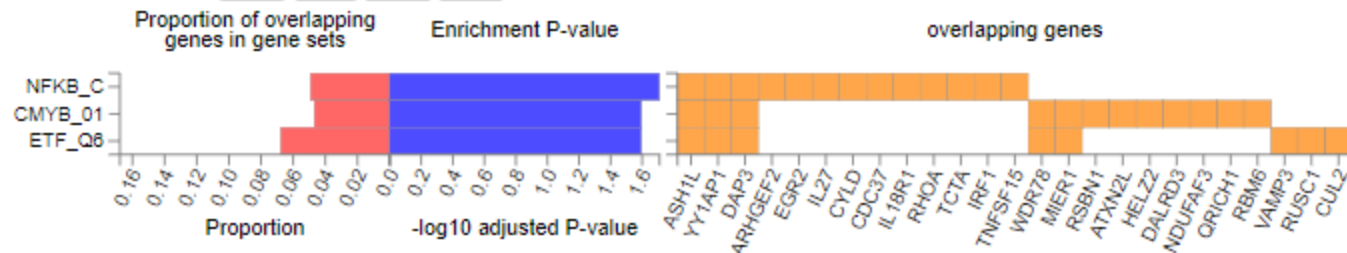


SNPs encoding genes have significant expression in spleen tissue

Functional Enrichment plots



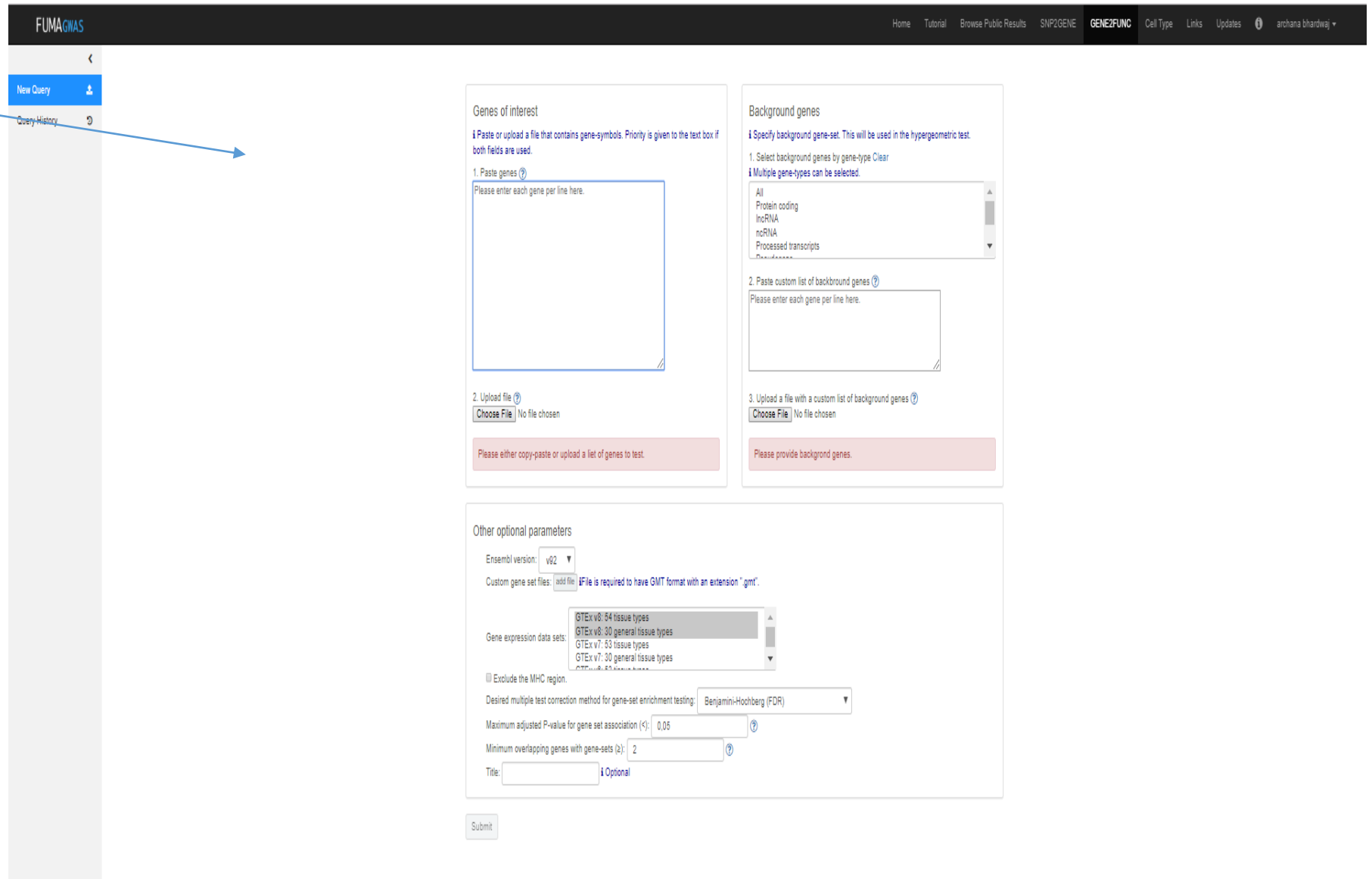
Download the plot as [PNG](#) [JPG](#) [SVG](#) [PDF](#)



Gene ID

Let us run Gene2FUNC

ANKRD44
FOSL2
RAP1GAP
CARMIL1
CACNA1S
CYLD
ATG16L1
DOCK3
TTC33
INSL6
ADCY7
NKD1
KSR1
OSMR
BABAM2
IFNGR2
IL23R
NOD2
SPNS1
FOSL1
TEX41
AL138720.1
AC067751.1
ZNF512
LINC00824
AP005482.1
AC007493.1
LINC02178
LINC02178
AF246928.1
ATG16L1
AC008703.1



FUMA-GS

Home Tutorial Browse Public Results SNP2GENE **GENE2FUNC** Cell Type Links Updates ⓘ archana bhardwaj ▾

New Query ⓘ

Query History ⓘ

Genes of interest

1. Paste or upload a file that contains gene-symbols. Priority is given to the text box if both fields are used.

1. Paste genes ⓘ
Please enter each gene per line here.

2. Upload file ⓘ
Choose File No file chosen

Please either copy-paste or upload a list of genes to test.

Background genes

1. Specify background gene-set. This will be used in the hypergeometric test.

1. Select background genes by gene-type **Clear**
Multiple gene-types can be selected.

All
Protein coding
lncRNA
miRNA
Processed transcripts
Pseudogenes

2. Paste custom list of background genes ⓘ
Please enter each gene per line here.

3. Upload a file with a custom list of background genes ⓘ
Choose File No file chosen

Please provide background genes.

Other optional parameters

Ensembl version: v92 ▾

Custom gene set files: add file ⓘ If file is required to have GMT format with an extension ".gmt".

Gene expression data sets:
GTEx v8: 54 tissue types
GTEx v8: 30 general tissue types
GTEx v7: 53 tissue types
GTEx v7: 30 general tissue types

☐ Exclude the MHC region.

Desired multiple test correction method for gene-set enrichment testing: Benjamini-Hochberg (FDR) ▾

Maximum adjusted P-value for gene set association (<): 0.05 ⓘ

Minimum overlapping genes with gene-sets (>): 2 ⓘ

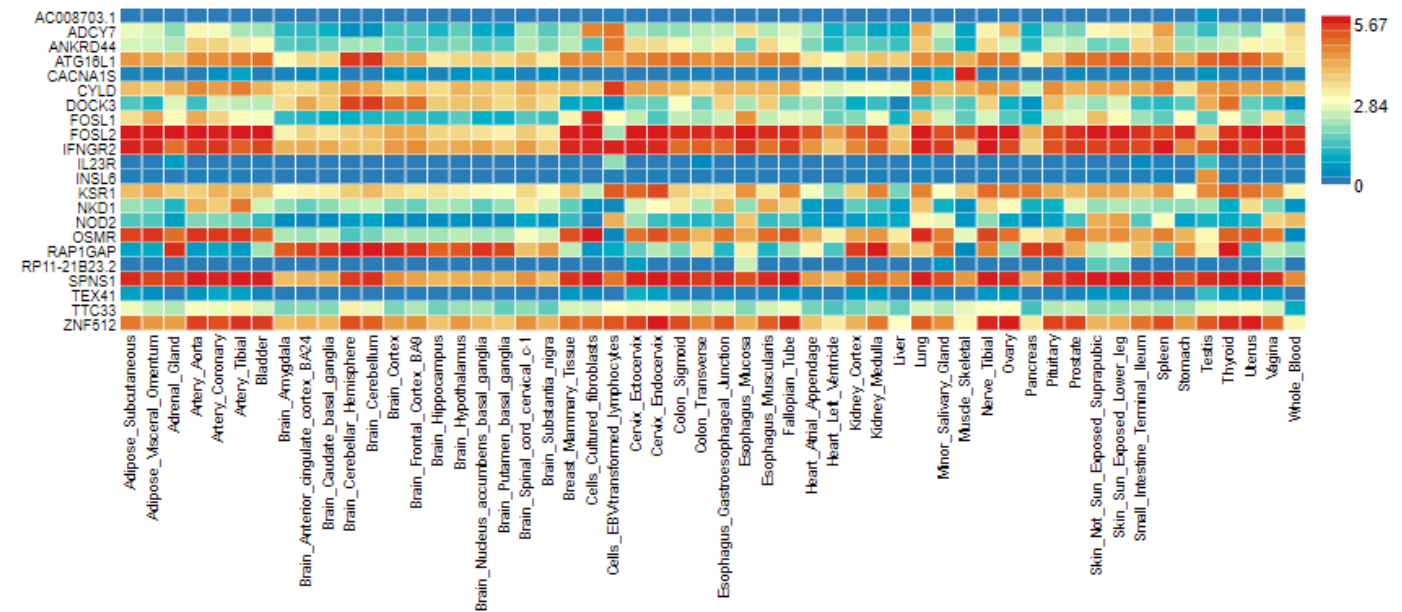
Title: ⓘ ⓘ Optional

Submit

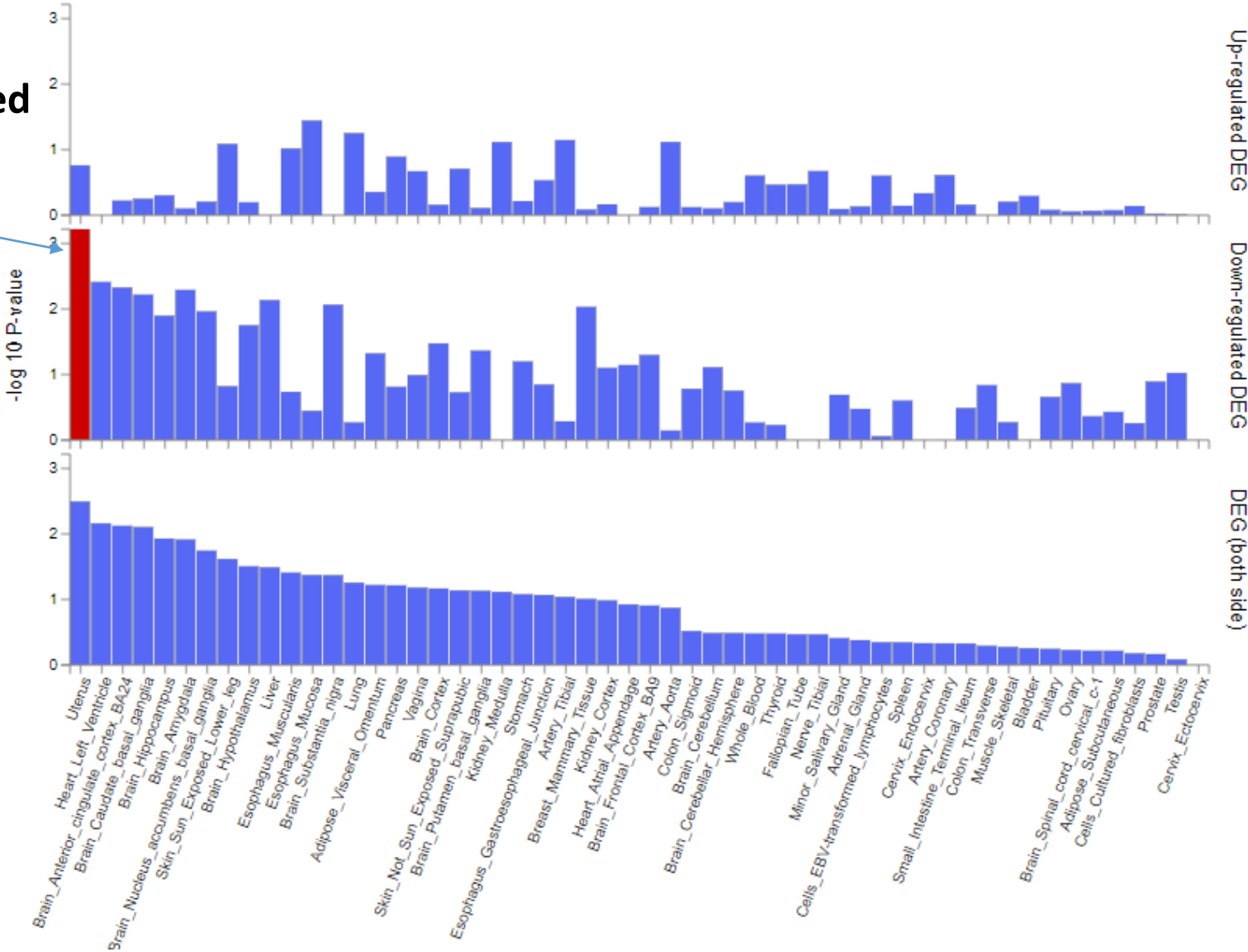
Summary of input genes

Number of input genes	32
Number of background genes	57241
Number of input genes with recognised Ensembl ID	26
Input genes without recognised Ensembl ID	CARMIL1, BABAM2, AL138720.1, LINC00824, AC007493.1, AF246928.1
Number of background genes with recognised Ensembl ID	57241
Background genes without recognised Ensembl ID	NA
Number of input genes with unique entrez ID	23
Number of background genes with unique entrez ID	35142

Download the plot as [PNG](#) [JPG](#) [SVG](#) [PDF](#)



Significantly differentially expressed



Enrichment : plots

Hallmark gene sets (MsigDB h)	(3)
Positional gene sets (MsigDB c1)	(1)
Curated_gene_sets	(0)
Chemical and Genetic perturbation gene sets (MsigDB c2)	(0)
All Canonical Pathways (MsigDB c2)	(0)
BioCarta (MsigDB c2)	(1)
KEGG (MsigDB c2)	(2)
Reactome (MsigDB c2)	(0)
microRNA targets (MsigDB c3)	(1)
TF targets (MsigDB c3)	(0)
All computational gene sets (MsigDB c4)	(0)
Cancer gene neighborhoods (MsigDB c4)	(0)
Cancer gene modules (MsigDB c4)	(0)
GO biological processes (MsigDB c5)	(2)
GO cellular components (MsigDB c5)	(0)
GO molecular functions (MsigDB c5)	(1)
Oncogenetic signatures (MsigDB c6)	(0)
Immunologic signatures (MsigDB c7)	(0)
WikiPathways	(0)
GWAS catalog reported genes	(8)

there are two signifiant pathways

there are two signifiant gene ontology

Informations found in GWAS catalog

Exercise

1. Classify SNPs list based on genomic location (genic and non genic)
2. Identify chromatin markers affected by given SNPs list .
3. Identify over represented KEGG pathways and GO enrichment based on SNPs encoding genes
4. Identify which tissue is differentially expressed due to given SNP list (via genes)?

rs4468290
rs11201609
rs4933212
rs701546
rs1241901
rs8087497
rs2409457
rs1666559
rs12943387
rs2036660

FROM TODAY SESSION

- Performed Post GWAS of identified SNPs (different ways)
- Multiple databases
- Multiple servers
- FUMA

NEXT SESSION

Multi Omics Integration :

- Working on multiple omics profiles (gene expression/methylation/SNPs)
- Integrate data using multiple algorithms for gene prioritization