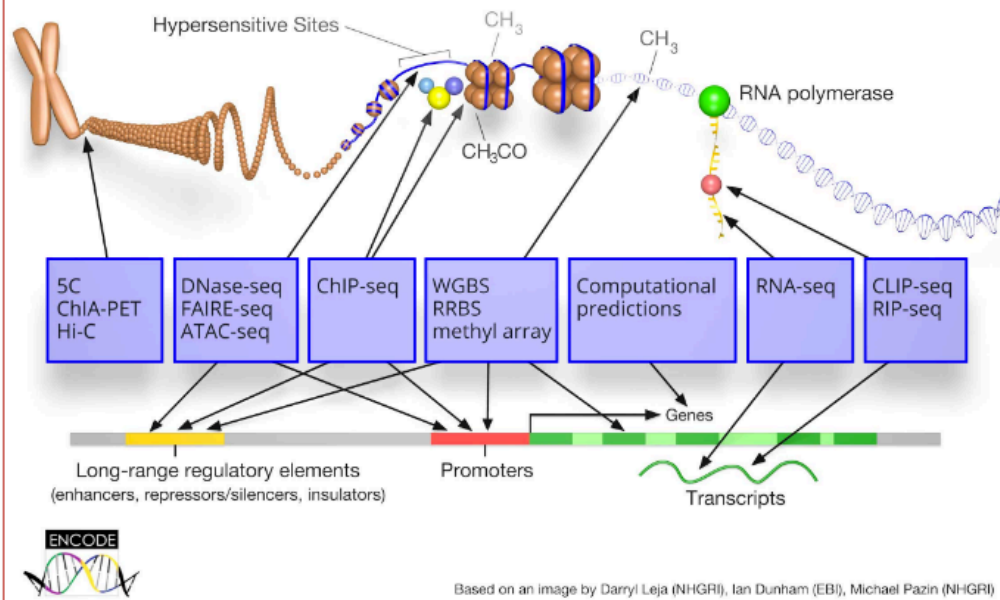


3D Genome Workshop - Overview of ENCODE

ENCODE: Encyclopedia of DNA Elements



The ENCODE (Encyclopedia of DNA Elements) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

[Get Started](#)

Based on an image by Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

Dan Gilchrist

National Human Genome Research Institute

October 18, 2016

3D Genome Workshop - Overview of ENCODE

- The ENCODE Project
- 3D Data Access Through the ENCODE Portal
- The ENCODE Encyclopedia
- Tools for Investigating Genetic Variants



ENCODE Tutorials

- ENCODE Tutorials

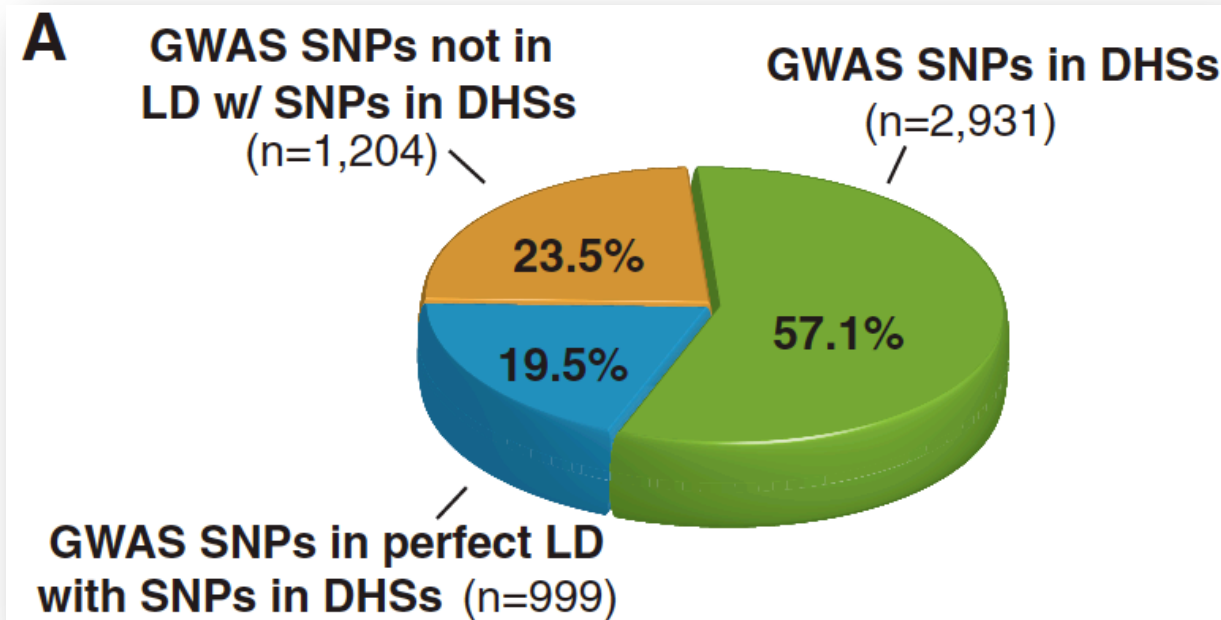
<https://www.encodeproject.org/tutorials/>

<http://www.genome.gov/27553900>



ENCODE: Annotating the non-coding genome

- Vast majority of common disease associations and heritability lie outside of coding regions
- Non-coding variants known to cause human diseases and alter human traits (FXS, ALS)



Goals of ENCODE

- Identify all candidate functional elements in the human and mouse genomes

Promoters, enhancers, transcribed regions

- Make resource freely available to the community for use in studies of:

Genetic basis of disease

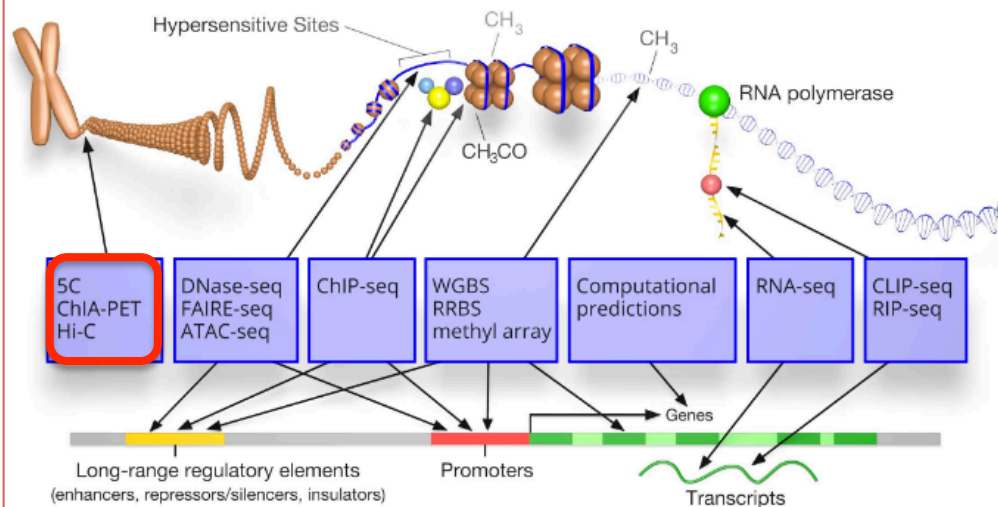
Gene regulation

Computational tool development



ENCODE Data Types

ENCODE: Encyclopedia of DNA Elements



The ENCODE (Encyclopedia of DNA Elements) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

[Get Started](#)



Based on an image by Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

- 1000s of data sets
- 100s of biosamples



3D Genome Workshop - Overview of ENCODE

- The ENCODE Project
- 3D Data Access Through the ENCODE Portal
- The ENCODE Encyclopedia
- Tools for Investigating Genetic Variants



ENCODE Portal – Unrestricted data access

The screenshot displays the ENCODE Portal's Experiment Matrix interface. At the top, there is a navigation bar with 'ENCODE', 'Data', 'Encyclopedia', 'Materials & Methods', and 'Help', along with a search bar and the user name 'Dan Gilchrist'. The main content area is divided into several sections:

- Experiment Matrix:** A search box with the text 'Enter search term(s)'. Below it is a diagram of a chromosome with various assays (5C, ChIA-PET, Hi-C, DNase-seq, FAIRE-seq, ATAC-seq) and a 'Long-range regulatory element'.
- Assay Summary Table:**

Assay	Assay category	Target of assay	Date released	Available data
ChIA-PET	DNA binding	transcription factor	May, 2016	fastq
Hi-C	Transcription	histone	May, 2012	bigWig
5C	RNA binding	histone modification	October, 2011	bed bed3+
	DNA accessibility	narrow histone mark	July, 2014	hd5
	DNA methylation	chromatin remodeler	June, 2016	bed bed12
- Filters:** Organism (Homo sapiens), Biosample type (Immortalized cell line), Project (ENCODE), Genome assembly (hg19), Audit category, and Status (released).
- ASSAY Results:** A table showing 54 results for 'Immortalized cell line' across various cell lines and assays. A red box highlights the 'ChIA-PET', 'Hi-C', and '5C' columns. Below the table are 'Download' and 'Visualize' buttons.
- Project Summary:** A donut chart showing 56 projects, with ENCODE being the largest.
- Social Media:** A Twitter feed and a News section with a headline '104 ChIP-seq from Snyder Lab datasets released'.

Select data sets

Click to download or visualize

Data also available through programatic web services



3D Genome Workshop - Overview of ENCODE

- The ENCODE Project
- 3D Data Access Through the ENCODE Portal
- The ENCODE Encyclopedia
- Tools for Investigating Genetic Variants



ENCODE Annotations – the Encyclopedia

ENCODE Data Encyclo

Top Level Annotations

Ground

Gene expression

The expression levels of genes across different cell types.

[[Long RNA-seq Data](#)]

Transcriptomics

Peaks (enriched genomic regions) generated by the WashU genome browser.

[[Raw Data](#) | [Peak](#)]

Histone marks

Peaks of a variety of histone marks.

[[Raw Data](#) | [Peak](#)]

Open chromatin

DNase I hypersensitive regions identified in various cell types.

[[Raw Data](#) | [Peak](#)]

Topological

Topologically Associating Domains (TADs) and Hi-C data.

[[Raw Data](#) | [Visualize](#)]

Promoter-enrichment

Links between promoter regions and enhancers.

[[Raw Data](#) | [Links](#)]

RNA binding protein occupancy (eCLIP-seq)

Peaks computed from eCLIP-seq data in human cell lines K562 and HepG2 for a large number of RNA Binding Proteins (RBPs).

[[Raw Data](#) | [Peaks](#)]

Middle

Promoter-like

DNase hypersensitive regions, promoter function, and H3K4me3 signal are used to predict regions with higher accuracy than DNase-seq alone. The procedure is "semi-automated" because states are then manually compared with known biological information in order to designate each state as an enhancer-like, promoter-like, gene body, etc.

Chromatin states

Semi-automated genomic annotation methods such as ChromHMM and Segway take as input a panel of epigenomic data (including histone mark ChIP-seq and DNase-seq) in a particular cell type and use machine learning methods to simultaneously partition the genome into segments and assign chromatin states to these segments; the states are assigned such that two segments with the same state exhibit similar epigenomic patterns. The procedure is "semi-automated" because states are then manually compared with known biological information in order to designate each state as an enhancer-like, promoter-like, gene body, etc.

[[Search](#)]

Variant Annotation

Over the past decade, Genome Wide Association Studies (GWAS) have provided insights into how genetic variations contribute to human diseases. However, over 80% of the variants reported by GWAS are in noncoding regions of the genome and the mechanism of how they contribute to disease onset is unknown. By integrating data from the ENCODE project and other public sources, RegulomeDB and HaploReg are two resources developed by ENCODE labs to aid the research community in annotating GWAS variants. FunSeq2 is another ENCODE resource for annotating both germline and somatic variants, particularly in the noncoding regions of cancer genomes.

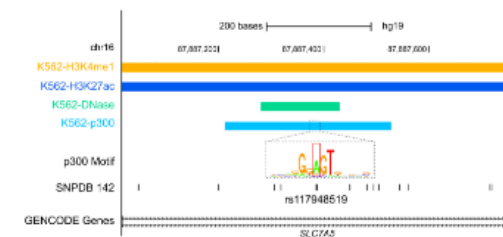
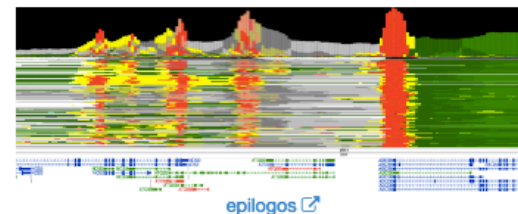
[[RegulomeDB](#) | [HaploReg](#) | [FunSeq2](#)]

Enhancer-like

DNase hypersensitive regions, enhancer function, and H3K27ac signal are tested on mouse and H3K27ac in mouse cell types. Roadmap Epigenomic consortia. For cell and tissues types with only H3K27ac or DNase data, we rank the peaks using the available data and make predictions of enhancer-like regions. You can query these enhancers by genomic locations, nearby genes, or SNPs, and visualize them in the UCSC and WashU genome browsers.

[[Visualize](#) | [Method](#)]

Enhancer-like genomic regions were tested on VISTA experimentally-validated enhancer elements: [[VISTA](#)]



VISTA Enhancer

PhastCons Conservation

RBFOX2 eCLIP

Size-matched input

RBFOX2 read density




3D Genome Workshop - Overview of ENCODE

- The ENCODE Project
- 3D Data Access Through the ENCODE Portal
- The ENCODE Encyclopedia
- Tools for Investigating Genetic Variants



Tools for Variant Annotation – FunSeq2



FunSeq2 - A flexible framework to prioritize regulatory mutations from cancer genome sequencing

Analysis **Results** **Downloads** **Documentation** **FAQ** **Whole Genoms Query**

Overview


This tool is specialized to prioritize somatic variants from cancer whole genome sequencing. It contains two components : 1) building data context from various resources; 2) variants prioritization. We provide downloadable scripts for users to customize the data context (found under '[Downloads](#)'). The variants prioritization step is downloadable, and also implemented as a web server (Right Panel), along with the pre-processed data context.

Instructions

- ✦ Input File - BED or VCF formatted. Click the "green" button to add multiple files. With multiple files, the tool will do recurrent analysis. (Note: for BED format, user can put variants from multiple genomes into one file, see [Sample input file](#) .)
- ✦ Recurrence DB - User can select particular cancer types from the database. The DB will continue to be updated with newly-available WGS data.
- ✦ Gene List - Option to analyze variants associated with a particular set of genes. Note: Please use Gene Symbols, with one row per gene.
- ✦ Differential Gene Expression Analysis - Option to detect differentially expressed genes in RNA-Seq data. Two files are needed: an expression file and a class label file. Please refer to [Expression input files](#) for instructions on how to prepare these files.

✦ Note: In addition to on-site calculation, we also provide scores for all possible noncoding SNVs of GRCh37/hg19 under '[Downloads](#)' (without annotation and recurrence analysis).

Input File: (only for hg19 SNVs)

no file selected 

BED or VCF files as input. [Sample input file](#)

Output Format:

▾

MAF:

Minor allele frequency threshold to filter polymorphisms from 1KG (value 0~1)

Cancer Type from Recurrence DB: [Summary table](#)

▾

[Add a gene list](#) (Optional)

[Add differential gene expression analysis](#) (Optional)


<http://funseq2.gersteinlab.org/>

Fu, Khurana, Gerstein, Genome Biology (15)10 2014



Tools for Variant Annotation - RegulomeDB

Download About Help



Chromatin structure Filter:

Method	Location	? Cell Type	Additional Info	Reference
DNase-seq	chr14:94207796..94208834	Cerebellumoc		ENCODE
DNase-seq	chr14:94208390..94208598	Panissets		ENCODE
DNase-seq	chr14:94208414..94209060	T47d		ENCODE
DNase-seq	chr14:94208516..94209130	Hepatocytes		ENCODE

RegulomeDB with current Chromatin to DNase f

Enter d

Histone modifications Filter:

Method	Location	Chromatin State	Tissue Group	Tissue	Reference
ChromHMM	chr14:94201400..94210600	Weak Repressed PolyComb	ESC	ES-UCSF4 Cell Line	REMC
ChromHMM	chr14:94201400..94211200	Quiescent/Low	ESC	H1 Cell Line	REMC
ChromHMM	chr14:94201400..94211400	Quiescent/Low	ESC	HUES6 Cell Line	REMC
ChromHMM	chr14:94201400..94211600	Quiescent/Low	ESC	ES-13 Cell Line	REMC

Summary of 5

Show 10 entries

Coordinate (0-based)	dbSNP ID	? Regulome DB Score	Other Resources
chr14:94208529	rs4900194	5	UCSC ENSEMBL dbSNP

Showing 1 to 1 of 1 entries

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.

<http://regulomedb.org/>

Boyle, Cherry, Snyder, Genome Research 22-1790,2012



Tools for Variant Annotation - HaploReg

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.

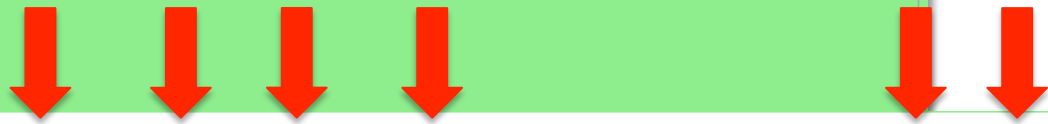
Update 2015.11.05: Version 4.1 GWAS and eQTL have been updated; a simpler pruning strategy is applied when combining GWAS; and links out to other NHGRI/EBI GWAS hits and GRASP QTL hits are provided.

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):



Query SNP: **rs6575353** 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
14	93736606	0.83	-0.96	rs77582682	G	A	0.03	0.06	0.00	0.08				9 tissues	CTCF,RAD21	4 altered motifs				PRIMA1	intronic
14	93738999	1	1	rs6575353	G	A	0.71	0.85	0.51	0.92						Sox,TCF4	1 hit	1 hit		PRIMA1	intronic
14	93739616	0.95	1	rs12896080	G	A	0.61	0.84	0.51	0.91						4 altered motifs				PRIMA1	intronic
14	93739927	0.83	1	rs4905084	A	G	0.60	0.83	0.51	0.90						4 altered motifs				PRIMA1	intronic
14	93742184	0.95	1	rs4900194	A	G	0.61	0.84	0.51	0.91			4 tissues	BRN	Gfi1,KAP1,Smad3		1 hit		PRIMA1	intronic	
14	93743570	0.97	1	rs11160137	A	G	0.69	0.85	0.52	0.91			5 tissues	ESC,HRT	4 altered motifs				PRIMA1	intronic	
14	93743836	0.85	0.93	rs12587586	T	G	0.59	0.84	0.52	0.91			6 tissues		Roaz				PRIMA1	intronic	



www.broadinstitute.org/mammals/haploreg/



Coming soon – Next Phase of ENCODE

- Starting in 2017
- Element mapping in normal and disease samples
- Characterizing function of candidate elements
- Likely to include expanded 3D data sets



ENCODE at ASHG2016

- Chat with ENCODE Data Coordination Center and NHGRI ENCODE Staff
 - NHGRI Booth #517
 - 2 - 3 PM Thursday 10/20

- ENCODE DCC Workshop - Open Science: Quality Assurance and Analysis of ChIP-seq Data Using the ENCODE Uniform Processing Pipeline
 - Ballroom B
 - 7:15 AM Friday 10/21

ENCODE Project Participants



Elise Feingold



Mike Pazin

