Gene Set Analysis for Sequence Data: An Exploration of Tools and Methods

Shaurya Jauhari
Mora Lab, A2-212.
Email: shauryajauhari@gzhmu.edu.cn
September 26th 2018

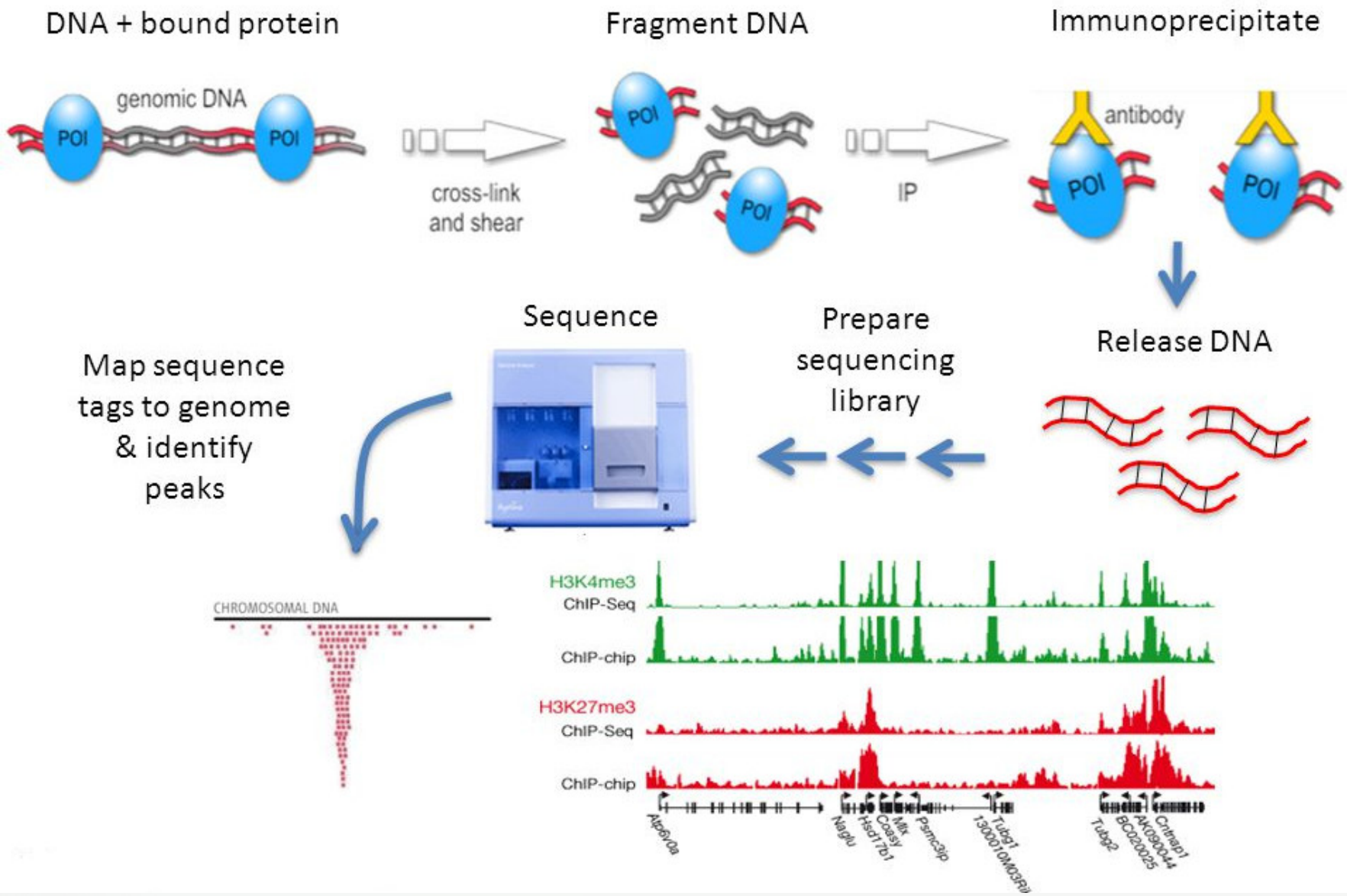
Outline

- Technology
  - ChIP Seq
- Tools
  - GREAT
  - ChIP- Enrich
  - Broad- Enrich
  - Seq2Pathway
  - Enrichr

ChIP Seq

- Chromatin Immunoprecipitation Sequencing
- Technology to discern protein binding sites in the DNA
- These proteins *may* contribute to gene regulation
- Gene Regulation
  - Transcription Factor binding (narrow peaks)
  - Histone modification (broad domains)

ChIP Seq Overview

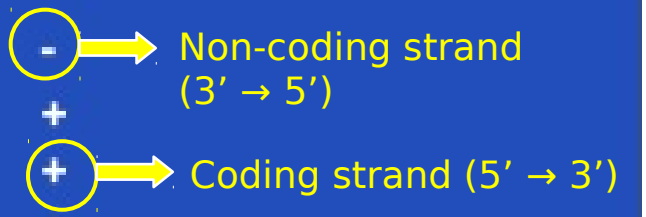


Credit: Stuart M. Brown, NYU.

BED File

- Defines coordinates for genomic region
- Tab Delimited format

```
$ head ../data/Input_tags.bed
chr1  233604    233639    0    2    -
chr1  559767    559802    0    3    +
chr1  742600    742635    0    2    +
chr1  742600    742635    0    0    +
chr1  744231    744266    0    0    +
chr1  744307    744342    0    2    -
chr1  746885    746920    0    2    +
chr1  746958    746993    0    1    +
chr1  748226    748261    0    2    +
chr1  748357    748392    0    0    -
```



CENTRAL DOGMA

- Gene based enrichment considers coding regions only.
- ~ 99 % of the nucleotides in the human genome **do not** code for proteins.
  - Nearly half of the disease associated SNPs are located in the “junk DNA”.
- Pathway affiliation restricted to genes’ definitions in databases.
- Henceforth, Sequence based enrichment is preferred.

GREAT

- Genomic Regions Enrichment of Annotations Tool
- Bejerano Lab, Stanford
- Online interface
- Highest cited tool
- Input: BED file from the Chip-Seq experiment
- Employs Binomial Distribution

Snapshot

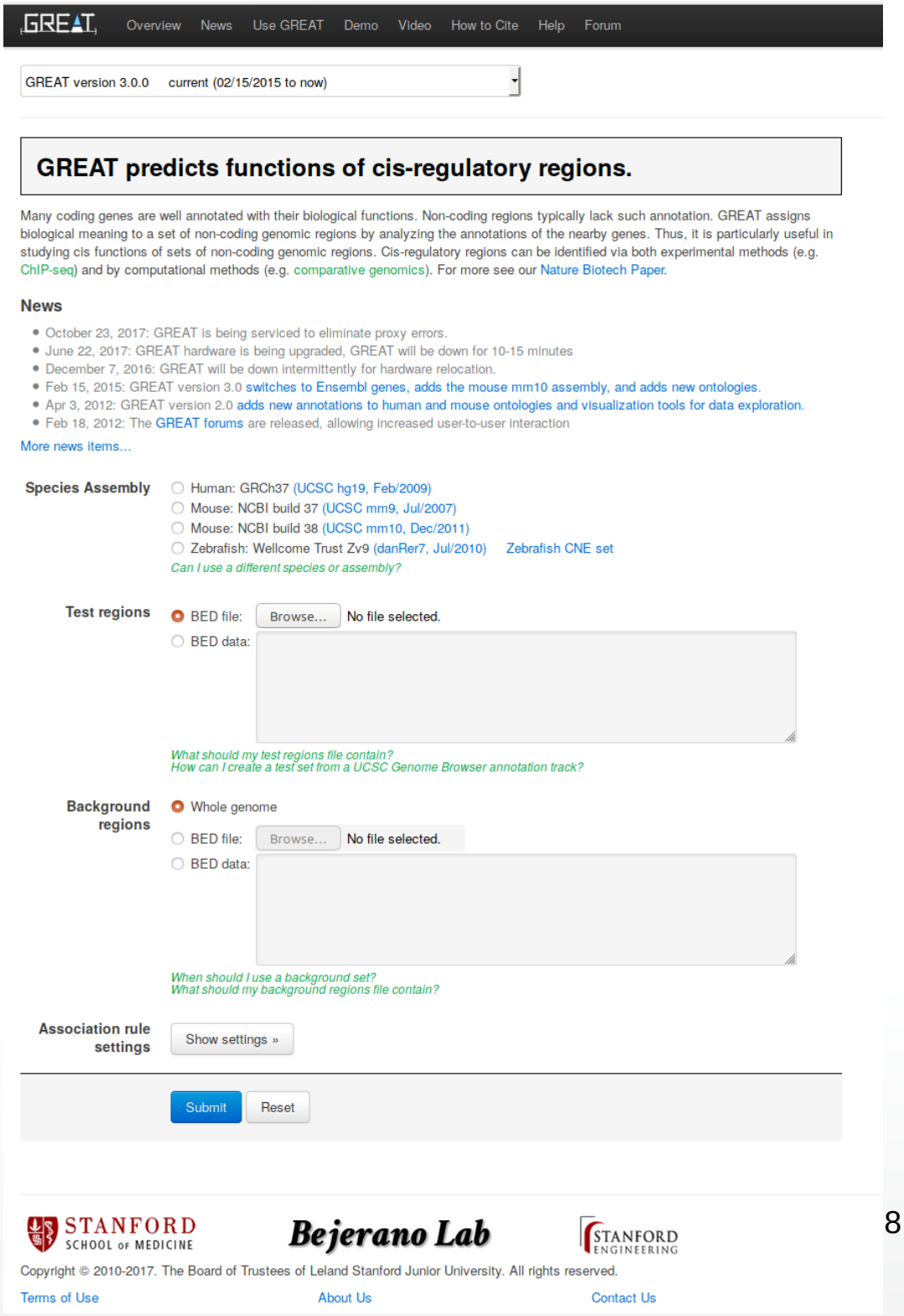
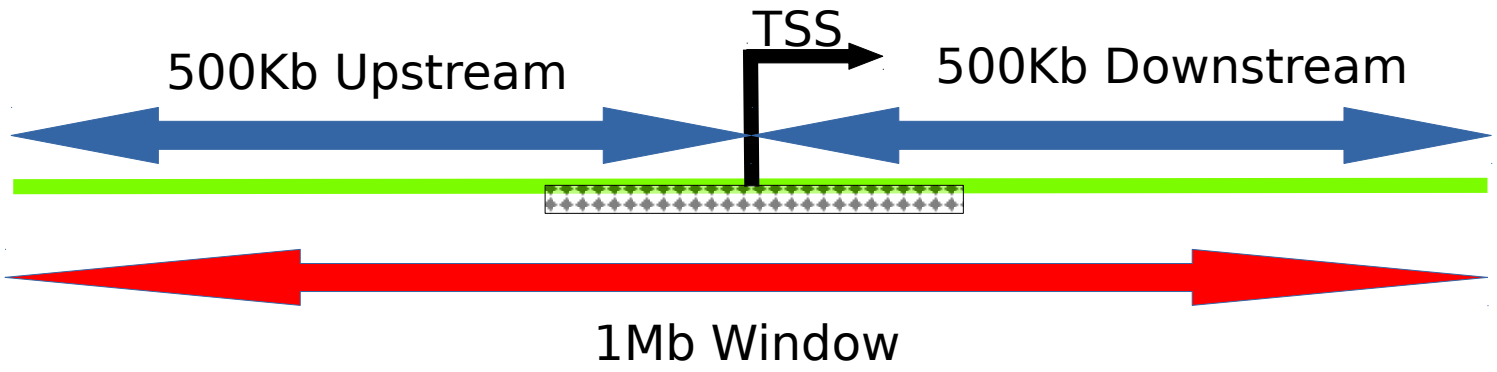

GREAT
Overview
News
Use GREAT
Demo
Video
How to Cite
Help
Forum
GREAT version 3.0.0 current (02/15/2015 to now)
GREAT predicts functions of cis-regulatory regions.
Many coding genes are well annotated with their biological functions. Non-coding regions typically lack such annotation. GREAT assigns biological meaning to a set of non-coding genomic regions by analyzing the annotations of the nearby genes. Thus, it is particularly useful in studying cis functions of sets of non-coding genomic regions. Cis-regulatory regions can be identified via both experimental methods (e.g. ChIP-seq) and by computational methods (e.g. comparative genomics). For more see our Nature Biotech Paper.
News
October 23, 2017: GREAT is being serviced to eliminate proxy errors.
June 22, 2017: GREAT hardware is being upgraded, GREAT will be down for 10-15 minutes
December 7, 2016: GREAT will be down intermittently for hardware relocation.
Feb 15, 2015: GREAT version 3.0 switches to Ensembl genes, adds the mouse mm10 assembly, and adds new ontologies.
Apr 3, 2012: GREAT version 2.0 adds new annotations to human and mouse ontologies and visualization tools for data exploration.
Feb 18, 2012: The GREAT forums are released, allowing increased user-to-user interaction
More news items...
Species Assembly
Human: GRCh37 (UCSC hg19, Feb/2009)
Mouse: NCBI build 37 (UCSC mm9, Jul/2007)
Mouse: NCBI build 38 (UCSC mm10, Dec/2011)
Zebrafish: Wellcome Trust Zv9 (danRer7, Jul/2010) Zebrafish CNE set
Can I use a different species or assembly?
Test regions
BED file: Browse... No file selected.
BED data:
What should my test regions file contain?
How can I create a test set from a UCSC Genome Browser annotation track?
Background regions
Whole genome
BED file: Browse... No file selected.
BED data:
When should I use a background set?
What should my background regions file contain?
Association rule settings
Show settings »
Submit
Reset
STANFORD SCHOOL OF MEDICINE
Bejerano Lab
STANFORD ENGINEERING
Copyright © 2010-2017. The Board of Trustees of Leland Stanford Junior University. All rights reserved.
Terms of Use
About Us
Contact Us

Illustration of Regulatory Domain

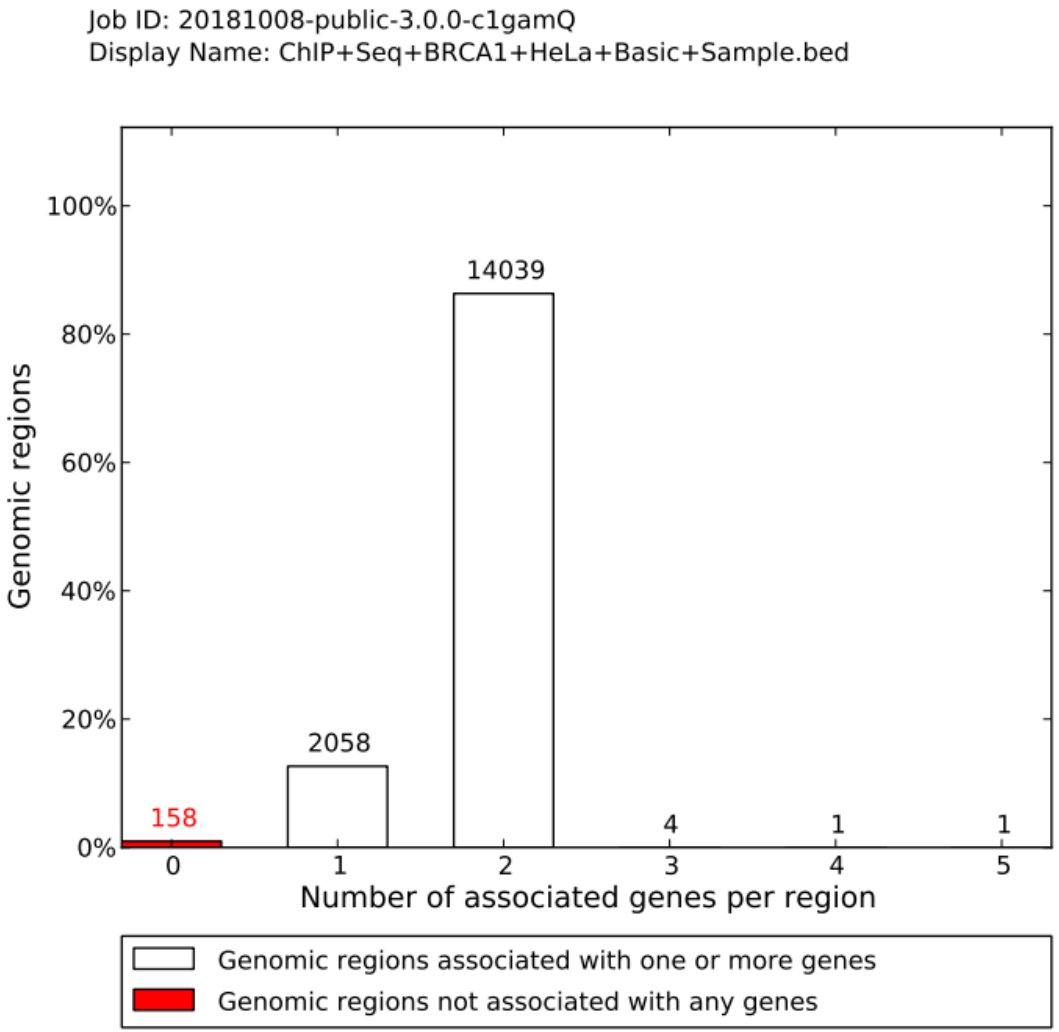


Basal/ Core

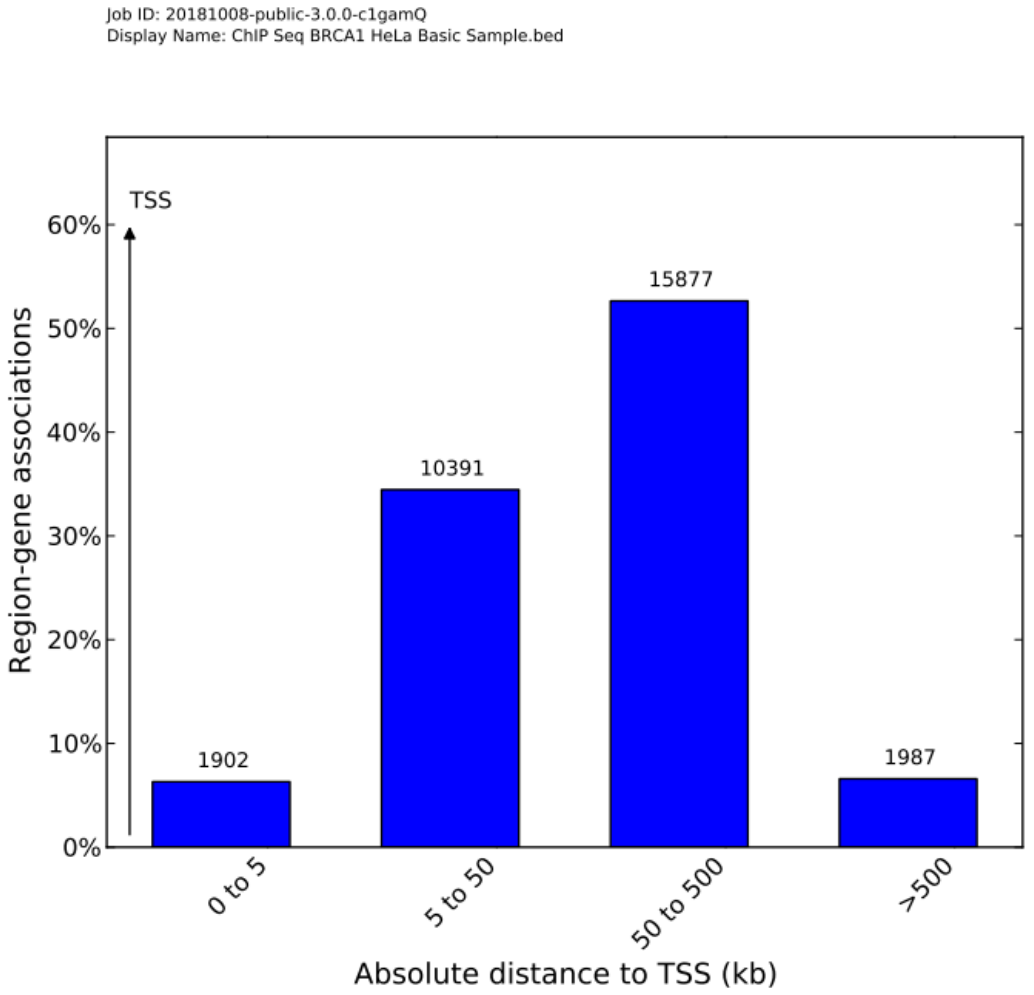
Assumptions

- Invariable to the gene locus length, the 1Mb window (upstream and downstream TSS) outside 5Kb(upstream) and 1Kb(downstream) holds all vital non-coding elements.
- The above landscape, in unision, constitutes an ideal gene regulatory region.
- The basal gene length has been accounted for normalization.

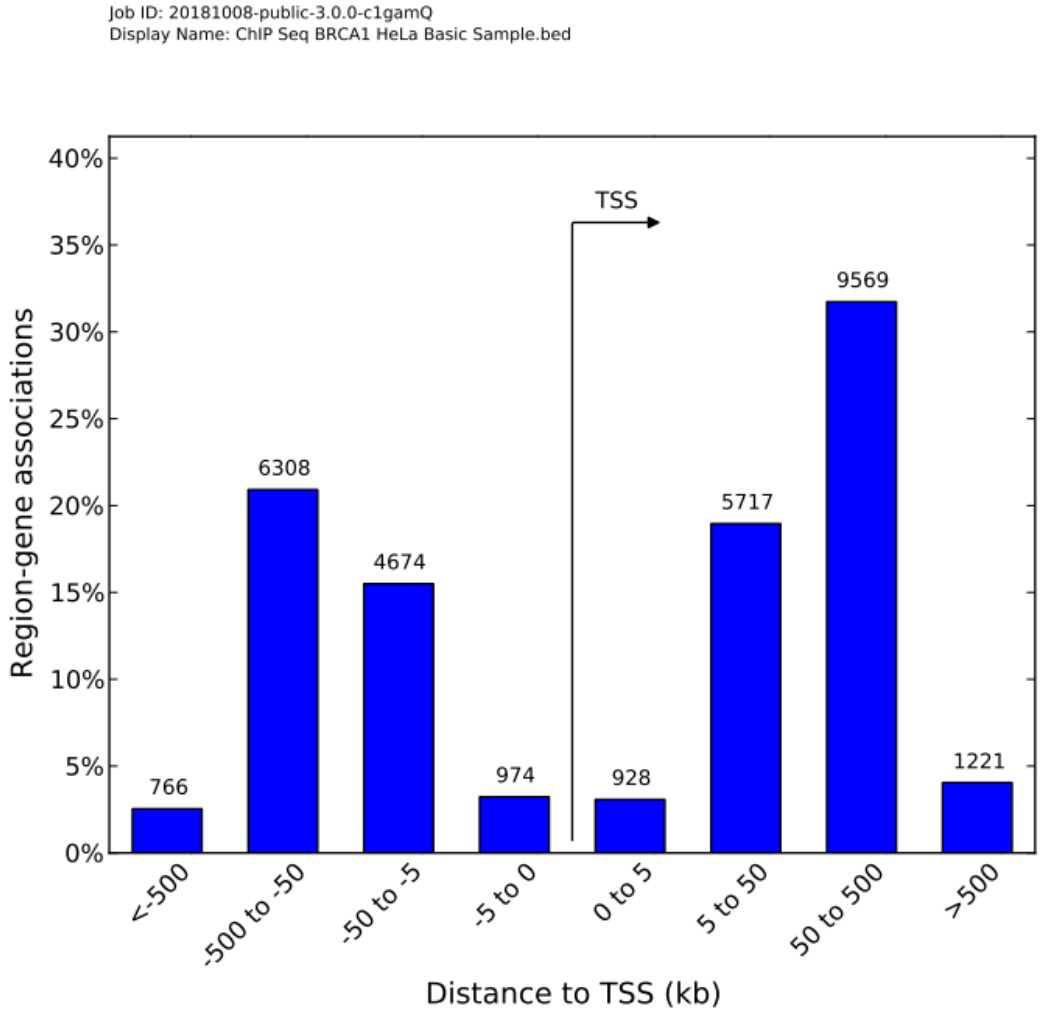
Results 1


Job ID: 20181008-public-3.0.0-c1gamQ
Display Name: ChIP+Seq+BRCA1+HeLa+Basic+Sample.bed
Genomic regions
100%
80%
60%
40%
20%
0%
158
2058
14039
4
1
1
0
1
2
3
4
5
Number of associated genes per region
Genomic regions associated with one or more genes
Genomic regions not associated with any genes

Results 2


Job ID: 20181008-public-3.0.0-c1gamQ
Display Name: ChIP Seq BRCA1 HeLa Basic Sample.bed
TSS
60%
50%
40%
30%
20%
10%
0%
Region-gene associations
1902
10391
15877
1987
0 to 5
5 to 50
50 to 500
>500
Absolute distance to TSS (kb)

Results 3

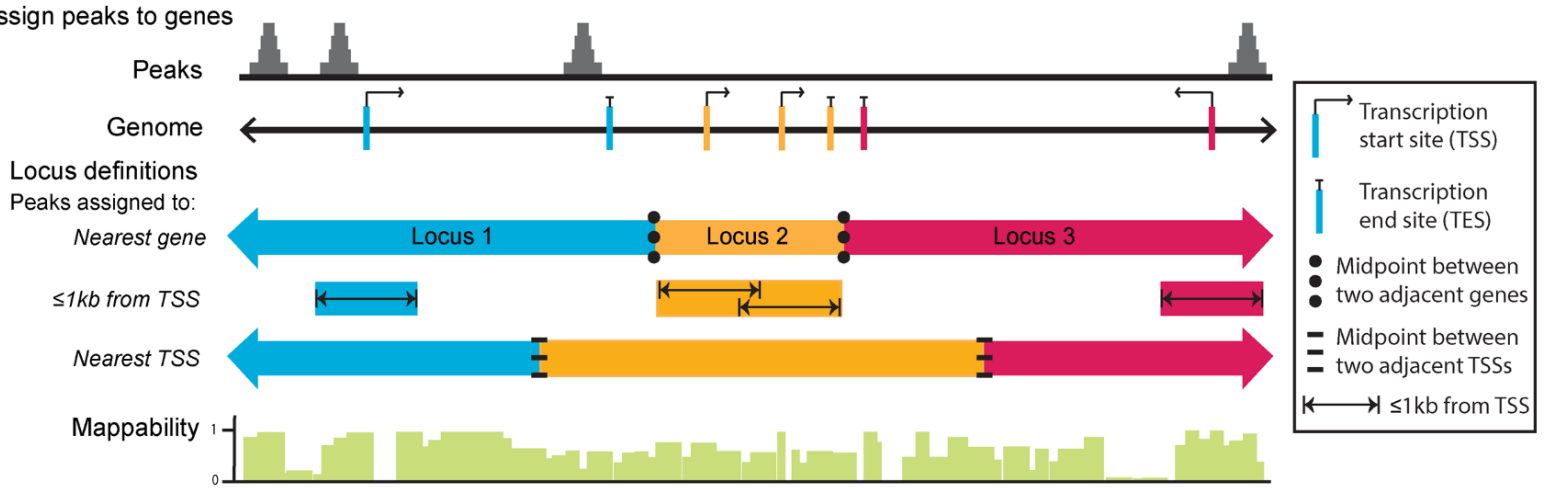

Job ID: 20181008-public-3.0.0-c1gamQ
Display Name: ChIP Seq BRCA1 HeLa Basic Sample.bed
Region-gene associations
40%
35%
30%
25%
20%
15%
10%
5%
0%
TSS
766
6308
4674
974
928
5717
9569
1221
<-500
-500 to -50
-50 to -5
-5 to 0
0 to 5
5 to 50
50 to 500
>500
Distance to TSS (kb)

ChIP- Enrich

- ChIP -Seq data enrichment tool
- University of Michigan
- Available as “chipenrich” R package and also as an online tool: http://chip-enrich.med.umich.edu
- Better handles Type I error rate as compared to GREAT
- Employs logistic regression approach
- ChIP-Enrich is designed for use with 1,000s or 10,000s of narrow peaks which results in fewer gene loci containing a peak overall. For example, ChIP-seq experiments for transcription factors

Core Model

1. Assign peaks to genes



2. Determine presence of peaks in genes

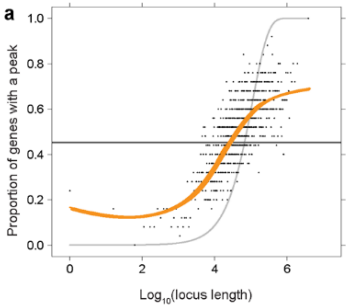
Gene	Locus length	Presence of peak
ACP1	11,541	0
CHL1	447,985	1
HES4	23,485	0
ITPR1	24,602	1
MYT1L	500,221	1
SAMD11	266,255	0
...	...	...

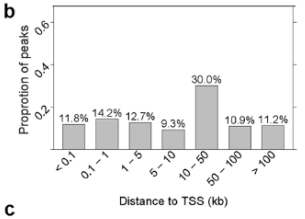
3. Test for gene set enrichment

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 g + f\left(log_{10}(mL+1)\right)$$

Logistic regression model
- Adjust for (mappable, m) locus length (L)
- Estimate gene set (g) effect size (β_1)

4. Summarize data and enrichment results



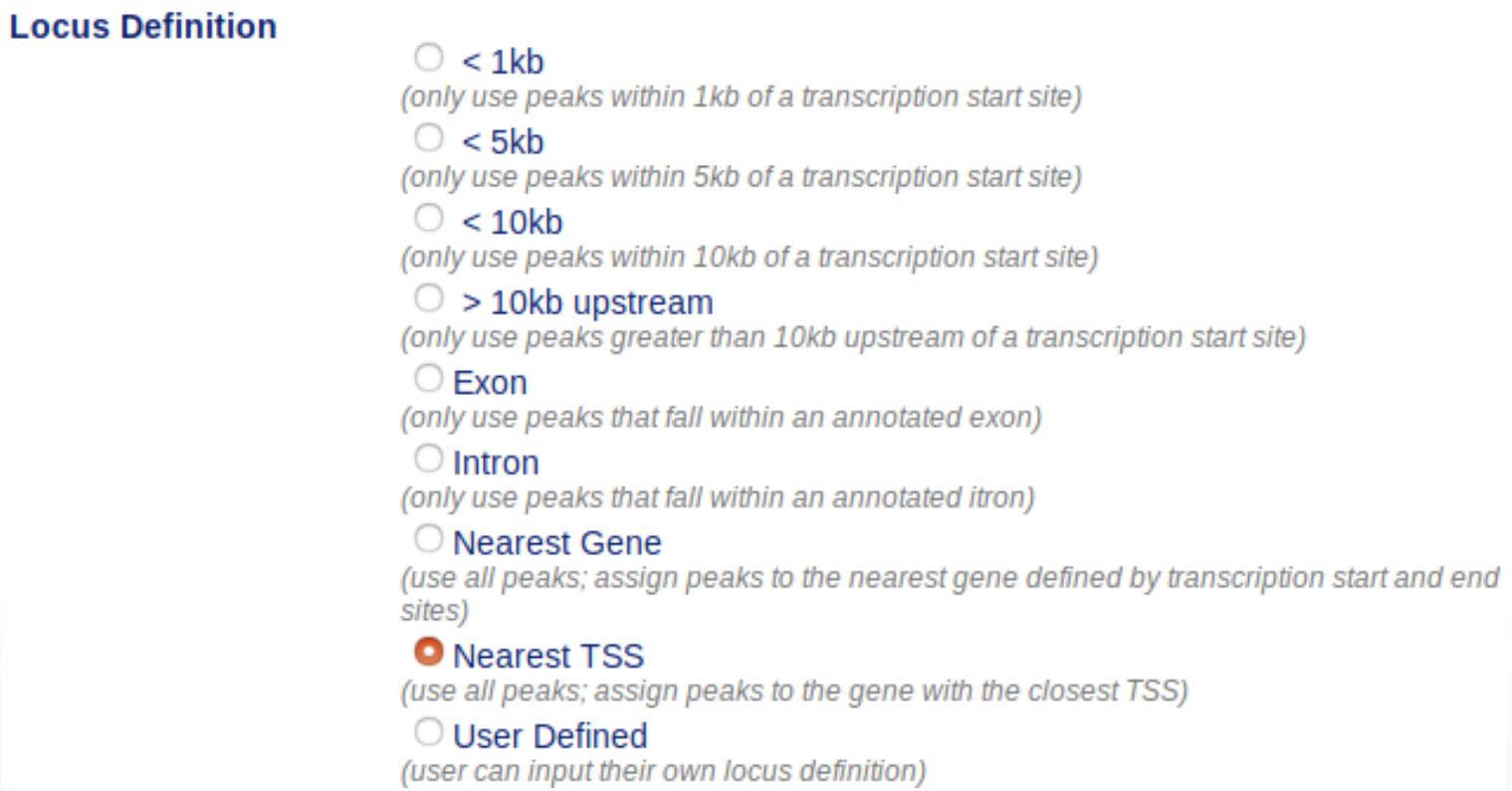


c

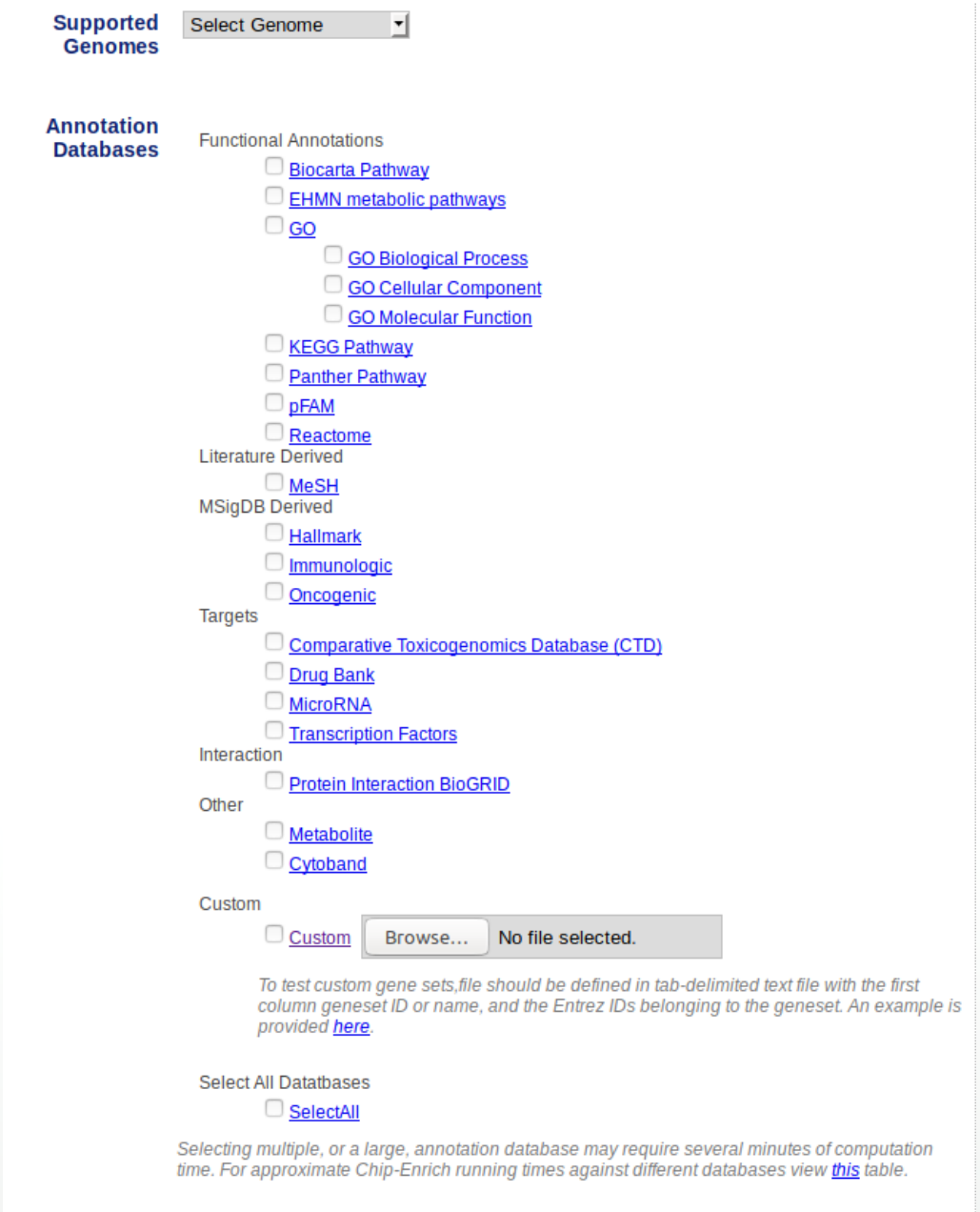
Gene set ID	Odds ratio	p-value	q-value
GO:0044424	2.7	2.8×10^{-38}	1.5×10^{-35}
GO:0005622	2.5	6.5×10^{-38}	1.7×10^{-35}
GO:0043227	2.2	4.5×10^{-36}	7.8×10^{-34}
...	...	...	...

Web Interface

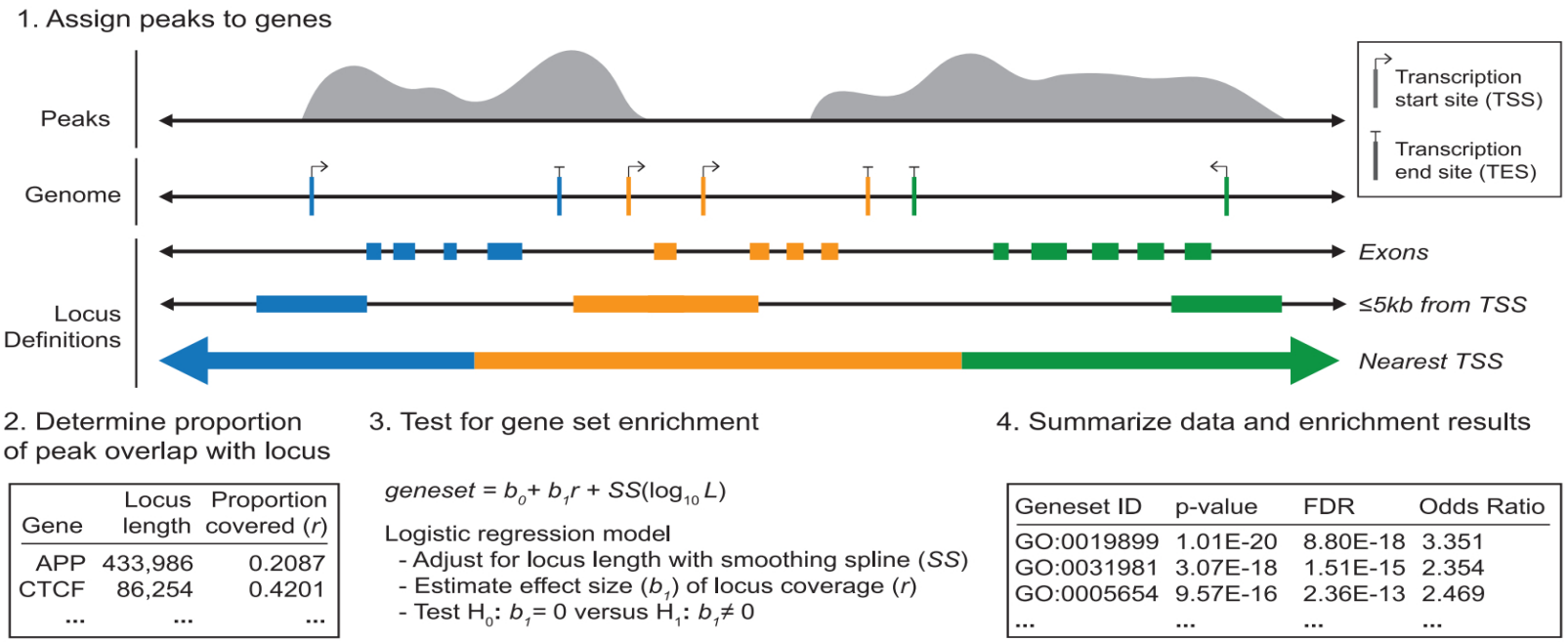
- Gene locus definitions/ available choices



Web Interface Contd...



Broad- Enrich



Gene	Locus length	Proportion covered (*r*)
APP	433,986	0.2087
CTCF	86,254	0.4201
...	...	...

$geneset = b_0 + b_1 r + SS(\log_{10} L)$

Logistic regression model
- Adjust for locus length with smoothing spline (*SS*)
- Estimate effect size (b_1) of locus coverage (*r*)
- Test H_0: $b_1 = 0$ versus H_1: $b_1 \neq 0$

Geneset ID	p-value	FDR	Odds Ratio
GO:0019899	1.01E-20	8.80E-18	3.351
GO:0031981	3.07E-18	1.51E-15	2.354
GO:0005654	9.57E-16	2.36E-13	2.469
...	...	...	...

- Broad-Enrich is designed for use with broad peaks that may intersect multiple gene loci, and cumulatively cover greater than 5% of the genome. For example, ChIP-seq experiments for histone modifications
- HISTONE MODIFICATIONS SPAN POLYGENIC DOMAINS
- Greater loci/ coverage than Chip Enrich (TF sites)

Broad-Enrich: Model

- Agenda is to examine whether a gene with locus length *L* and coverage proportion *r* claims a gene set membership
- Variable *geneset* is a binary vector
- Logistic regression model applied

$$\log \frac{\pi}{1-\pi} = b_0 + b_1 r + SS(\log_{10} L)$$

Dependent variable/ odds

Web Interface

Broad-Enrich: gene set enrichment testing for sets of broad genomic regions

Overview

Broad-Enrich tests sets of broad genomic regions (e.g., from ChIP-seq data for histone modifications or copy number variations) for enriched biological pathways, Gene Ontology terms, or other gene sets. The pre-defined gene sets are the same as used in LRpath, and can be browsed here. Using an input .bed, .narrowPeak or.broadPeak file, Broad-Enrich determines the proportion of each gene locus covered by a peak, using a chosen "gene locus definition". The "locus" of a gene is the region from which the gene is predicted to be regulated. Broad-Enrich uses a logistic regression model to test for association between the proportion of each gene locus covered by a peak and gene set membership. It empirically adjusts for the bias due to locus length using a binomial cubic smoothing spline within the logistic model. Detailed methods are provided here. Output includes summary plots, peak to gene assignments, and enrichment (and depletion) results including odds ratio, p-value, and FDR for each gene set.
more

Broad-Enrich is also available as part of the Chip-Enrich R package: Broad-Enrich.zip
Vignette: pdf

Select input file Browse... No file selected.

Input file should be a standard .bed,.narrowPeak or.broadPeak file containing at least three columns: (1) chromosome (of the form "chr3") (2) start position, and (3) end position. Additional columns will be ignored.

Analysis Name

Please provide a meaningful name for this analysis (used to name output files).

Email

Please provide your email address to be notified when the analysis is complete.

Supported Genomes Select Genome

- Select Genome
- Human (hg19)
- Mouse (mm9)
- Rat (rn4)
- Mouse(mm10)

Annotation Databases

thway
bolic Pathways
GO

Web Interface

Seq2Pathway

- Only available as a R package
- *Seq2gene*
  - Many-many mapping
  - Links coding and non-coding regions to coding genes
- *Gene2pathway*
  - Takes into account the quantity of significance for gene members within a pathway compared to those outside pathway (*Competitive hypothesis*)

Mapping in Seq2Pathway

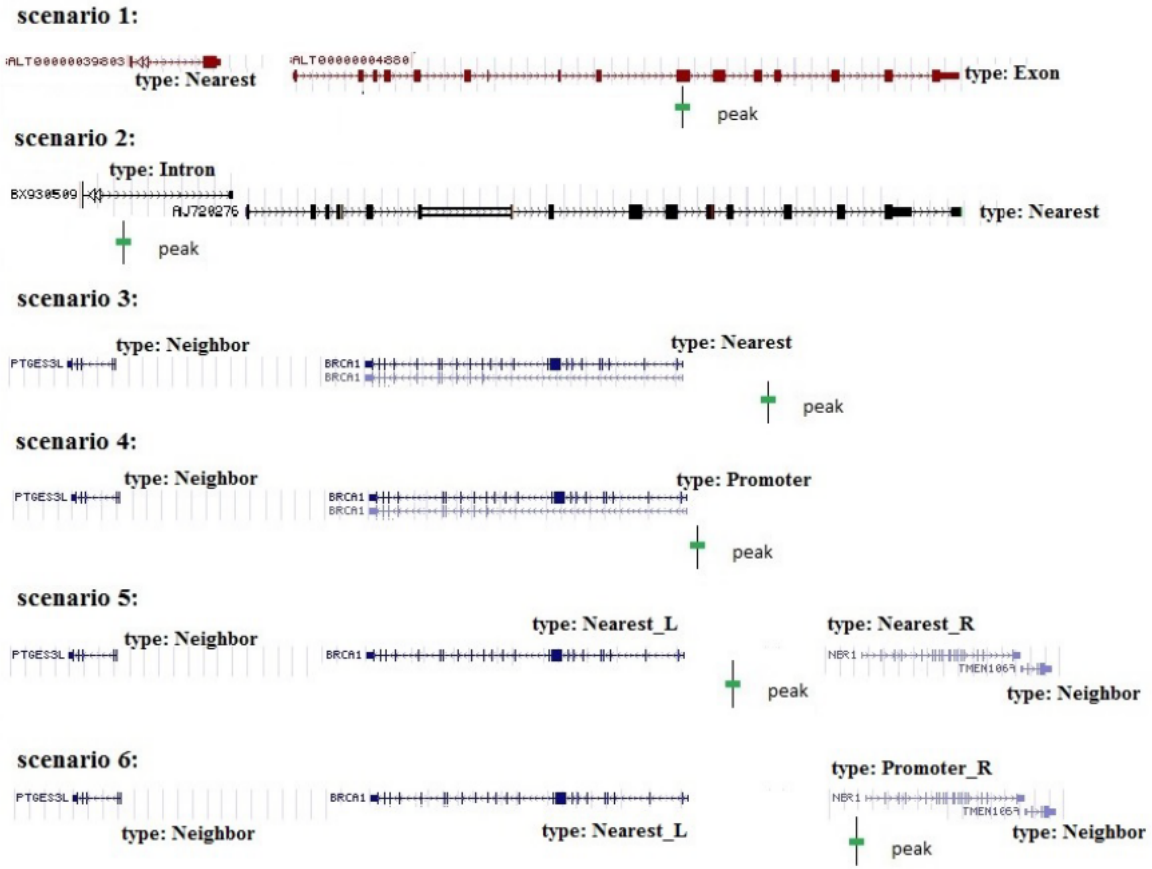


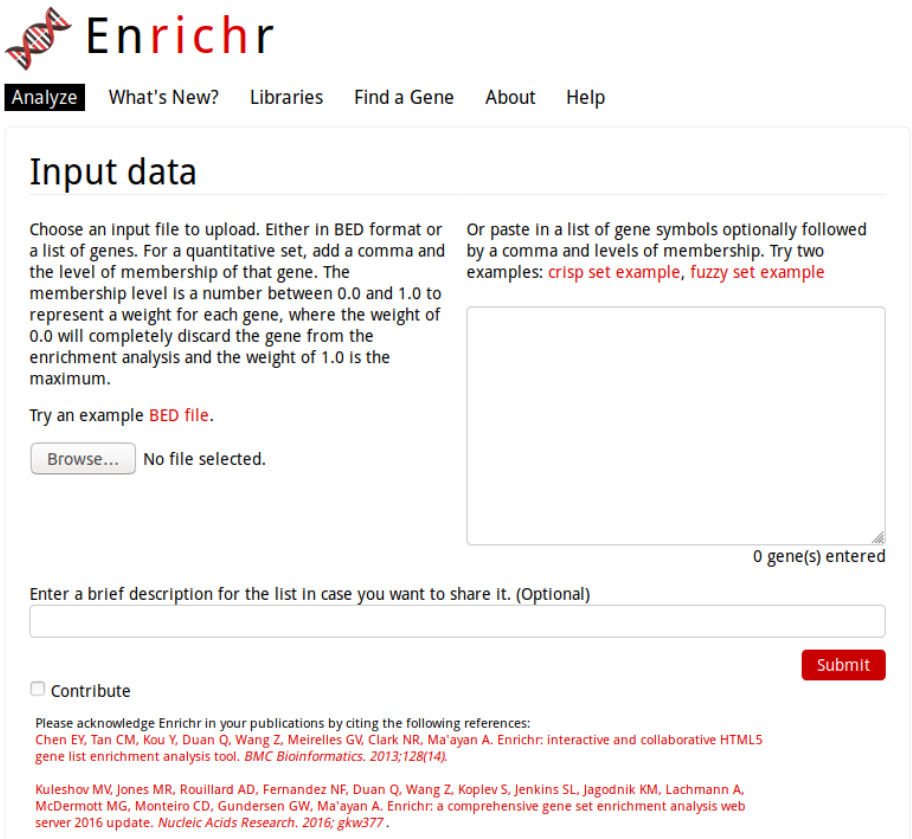
Figure 6: **Six output type values in several scenarios**. In each scenario, we map the genomic region of interest in green to the following types of a coding gene: exon (1), intron (2), the nearest (3), promoter (4), Nearest_L and Nearest_R (5), or Promoter_R (6).

Application

```
biocLite("seq2pathway.data")
biocLite("seq2pathway")
biocLite("chipenrich")

### Calling functional libraries ###

library(seq2pathway.data)
library(seq2pathway)
library(chipenrich.data)
library(chipenrich)


### Testing Seq2Pathway with a sample file ###

inputfile <- read.table("/home/postdoc1/Postdoc@GMU/Data/ChIP Seq BRCA1 HeLa Seq2Pathway.bed", sep = "\t", header = FALSE)
result_seq2gene <- runseq2gene(inputfile, search_radius = 150000, promoter_radius = 200, promoter_radius2 = 100, genome = "hg38",
            adjacent = FALSE, SNP = FALSE, PromoterStop = FALSE, NearestTwoDirection = TRUE, UTR3 = FALSE)

### Extracting gene list from the result of seq2gene algorithm ###

seq2gene_gene_list <- result_seq2gene$seq2gene_CodingGeneOnlyResult[ ,"gene_name"]
seq2gene_gene_list <- as.data.frame(seq2gene_gene_list)

### Executing gene list for enrichment ###
data(MsigDB_C5,package="seq2pathway.data")
result_gene2pathway <- gene2pathway_test(dat= seq2gene_gene_list, FisherTest=TRUE, EmpiricalTest=TRUE,
                        method="FAIME", alpha=5, logCheck= FALSE, na.rm=FALSE)
```

Enrichr

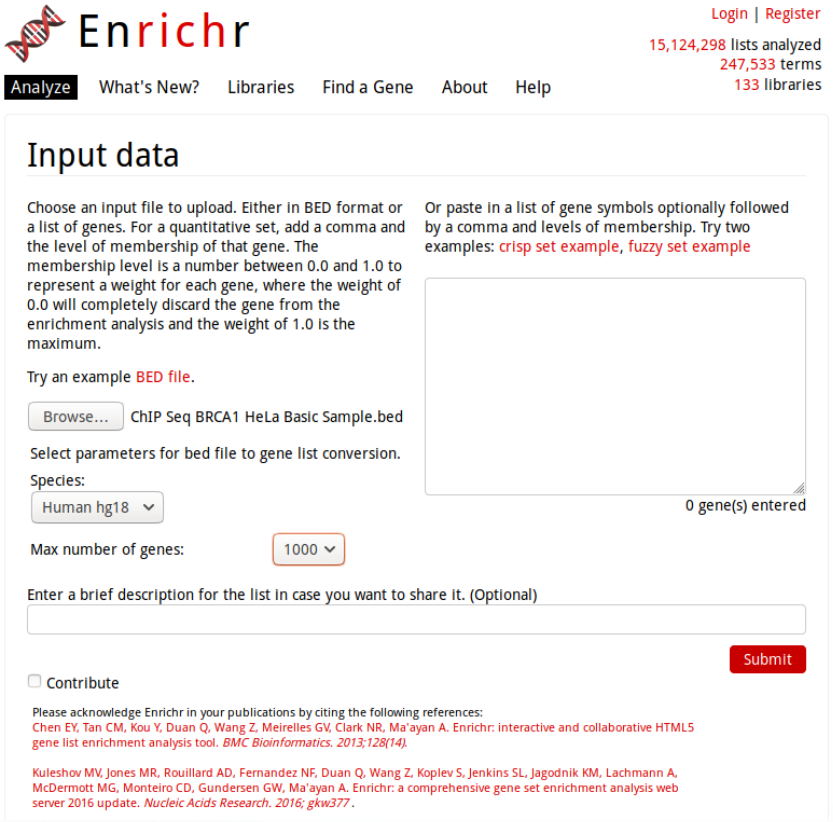
- http://amp.pharm.mssm.edu/Enrichr
- Freely available online tool
- Released in 2013 with updates in 2014 and 2016.
- HTML5, web-based application with mobile support for iPhone, Android, and Blackberry too.

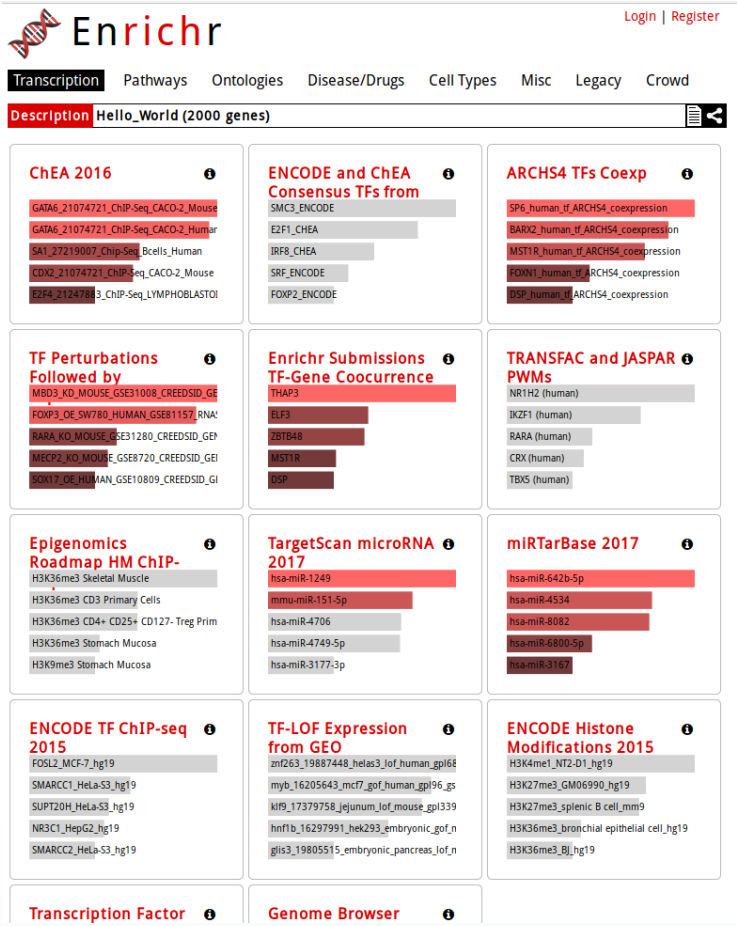


Gene- Set Libraries

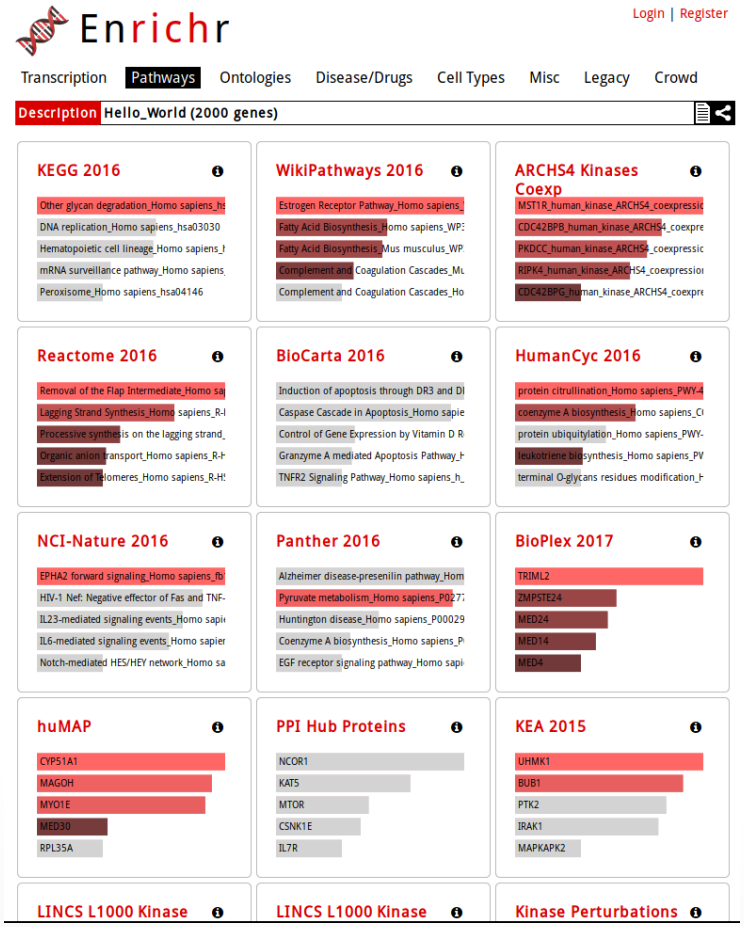
- *Transcription*
  - ChEA, PWMs(TRANSFAC, JASPAR, UCSC genome browser), TF targets from ENCODE, histone modifications from Roadmap Epigenomics Project, microRNA targets from TargetScan.
- *Pathways*
  - KEGG, BioCarta, Reactome, etc.
- *Ontologies*
  - Gene Ontology trees, Knockout Mouse Phenotypes Ontology from MGI-MP browser, List2Networks.
- *Diseases/ Drugs*
  - Connectivity Map database, GeneSigDB, MSigDB, OMIM, VirusMINT.
- *Cell Types*
  - Mouse and Human Gene Atlases, Cancer Cell Line Encyclopedia (CCLE), NCI-60.
- *Miscellaneous*
  - Chromosomal locations from MsigDB, Metabolite library from HMDB, structural domains library from PFAM and Interpro.

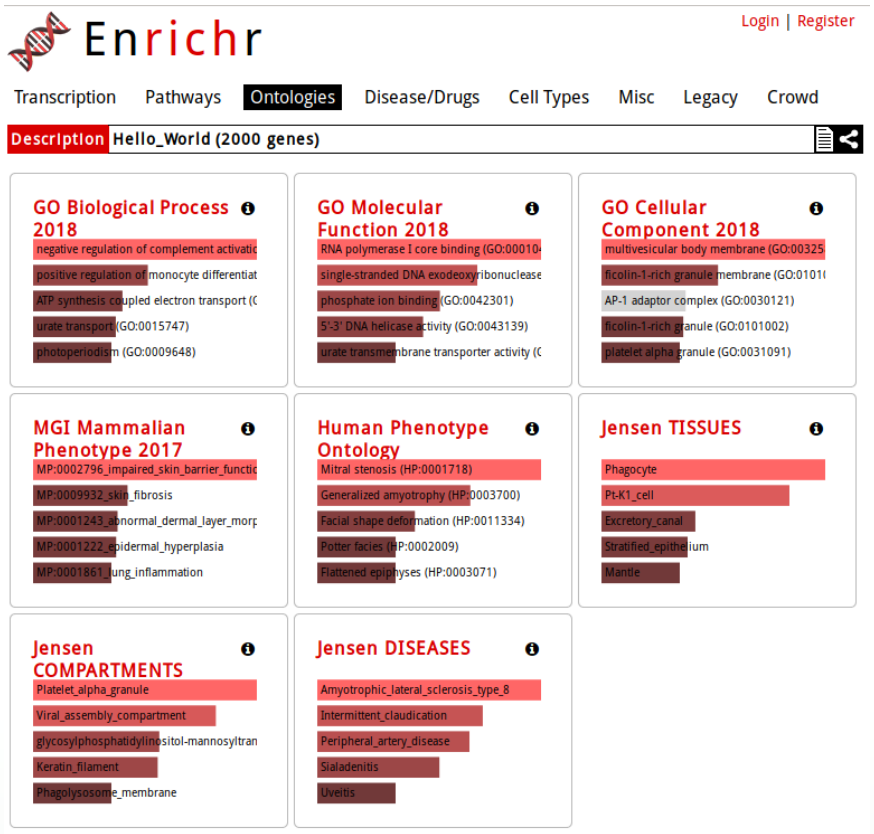
Enrichr: Input Screen, Transcription





Enrichr: Pathways, Ontologies





Deployment

Methods	R	Web Interface
GREAT		✔
SEQ2PATHWAY	✔	
CHIP ENRICH	✔	✔
BROAD ENRICH	✔	✔
ENRICHR		✔

References

- Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., … Ma'ayan, A. (2013). Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics, 14. https://doi.org/10.1186/1471-2105-14-128
- Welch, R. P., Lee, C., Imbriano, P. M., Patil, S., Weymouth, T. E., Smith, R. A., … Sartor, M. A. (2014). ChIP-Enrich: Gene set enrichment testing for ChIP-seq data. Nucleic Acids Research, 42(13), 1–13. https://doi.org/10.1093/nar/gku463
- Cavalcante, R. G., Lee, C., Welch, R. P., Patil, S., Weymouth, T., Scott, L. J., & Sartor, M. A. (2014). Broad-Enrich: Functional interpretation of large sets of broad genomic regions. Bioinformatics, 30(17), 393–400. https://doi.org/10.1093/bioinformatics/btu444
- McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., … Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nature Biotechnology, 28(5), 495–501. https://doi.org/10.1038/nbt.1630
- Wang, B., Cunningham, J. M., & Yang, X. (2015). Seq2pathway: An R/Bioconductor package for pathway analysis of next-generation sequencing data. Bioinformatics, 31(18), 3043–3045. https://doi.org/10.1093/bioinformatics/btv289