GenomeRunnerWEB

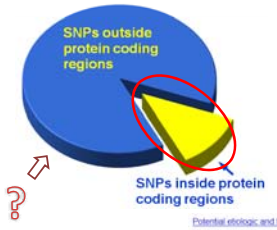
# Systematic classification of common disease-associated SNPs by their epigenomic relationship

Mikhail Dozmorov
mikhail.dozmorov@gmail.com
02-13-2014

http://www.genomerunner.org

## Genomic variants located everywhere

- SNPs – single nucleotide polymorphisms – and other genomic variants (CNVs, InDels, SVs) are located everywhere




Only 12% of SNPs are located in, or occur in tight linkage disequilibrium with, protein-coding regions.

Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.
Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA.

## Individual SNPs vs. multiple SNPs

- Hypothesis: SNPs may have additive effect – need to consider their collective impact on regulation

Example:

**Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits**

Olivia Corradin, Alina Saiakhova, Batool Akhtar-Zaidi, et al.

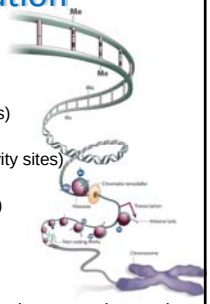
*Genome Res.* 2014 24: 1-13 originally published online November 6, 2013

**How to understand the collective regulatory impact of multiple SNPs?**

## Genome annotations as a means to understand regulation

Annotation

- Physiochemical properties (e.g., CpG islands)
- Sequence-based patterns (e.g., microsatellites)
- Experimental regions (e.g., DNA hypersensitivity sites)
- Predicted regions (e.g., TFBS, miRNA targets)
- 3D patterns (e.g., chromatin looping)



- **Epigenomic data** = genome annotation data = regions other than DNA sequence, annotated as carrying functional and/or regulatory potential or having a biological property

## Few methods for epigenomic data interpretation

- The use of epigenomic data is currently limited to a handful of features (e.g., DNA methylation, chromatin states, transcription factor binding sites), and on a local scale (e.g. visualization)

**Out of ALL epigenomic elements, which are enriched in my experimental data?**

## Two problems

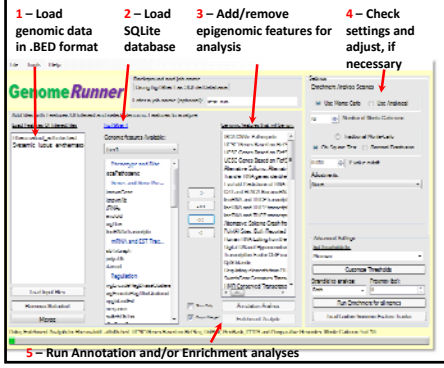
1. The collective regulatory impact of genomic variants is understudied
2. The use of epigenomic data for genomic variants' interpretation is inefficient

# Connecting genomics and epigenomics: GenomeRunner

**GenomeRunner: Automating genome exploration**

Mikhail G. Dozmorov[1*], Lukas R. Cara, Cory B. Giles, Jonathan D. Wren[1]

[1]Arthritis and Clinical Immunology Research Program, Oklahoma Medical
Oklahoma 73104-5005. *To whom correspondence should be addressed



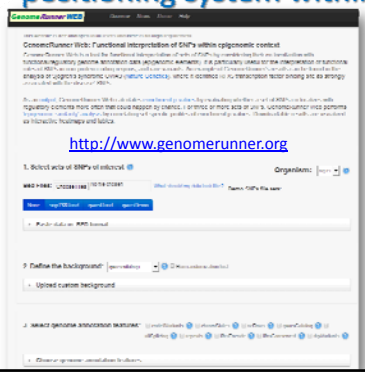

Finds statistically significant associations within the genome and epigenome

https://sourceforge.net/projects/genomerunner

## GenomeRunner highlights

- Annotation of genomic regions (ChIP-seq, DNA-methylation, gene promoters, SNPs etc.)
- Detection of the epigenomic elements enriched in a set of genomic regions
- Analysis of genomic regions of any length
- The ENCODE genome annotation data

## GenomeRunner Web – a global positioning system within the genome

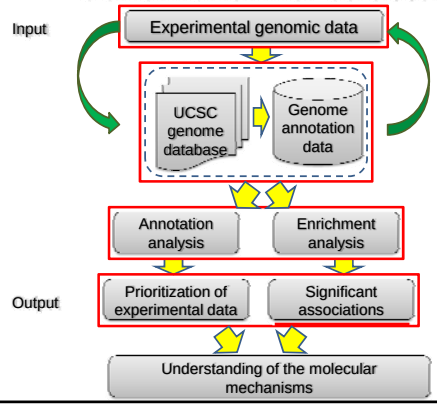



- SNP set analysis
- Visualization of the most significant enrichments
- Epigenomic similarity analysis and visualization

## GenomeRunner Web highlights

- **SNP-specific analysis**
  - in contrast to a "one-SNP-at-a-time" approach, GenomeRunner consider sets of SNPs as a whole, and determine their potential impact upon (cell type-specific, if available) epigenomic landscape
- **Enrichment- and epigenomic similarity analyses**
  - Enrichment analysis answers the question whether a set of SNPs of interest collectively enriched or depleted in regulatory regions, as compared with randomly selected set of SNPs.
  - Epigenomic similarity analysis visualizes similarity among enrichment profiles for three or more sets of SNPs of interest. It answers the question whether different sets of SNPs are enriched in similar epigenomic elements, hence, may affect similar regulatory networks.
- **Visualization and download of the results**

## GenomeRunner structure



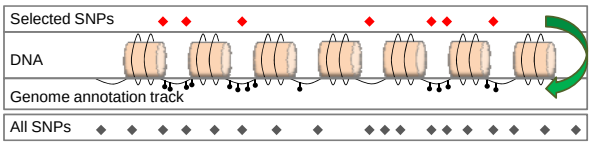
## Input data format (BED)

- Browser Extensible Data (BED) format
  - Tab-delimited text file
  - Three columns minimum
    - chrom - The name of the chromosome (e.g. chr3, chrY).
    - chromStart - The starting position of the feature in the chromosome . The first base is 0.
    - chromEnd - The ending position

| | | | | | |
|---|---|---|---|---|---|
| chr14 | 93644378 | 93644379 | rs1268843 | 0 | + |
| chr20 | 31277093 | 31277094 | rs210135 | 0 | + |
| chr17 | 41807330 | 41807331 | rs1513670 | 0 | - |
| chr7 | 12713069 | 12713070 | rs10488226 | 0 | + |

https://genome.ucsc.edu/FAQ/FAQformat.html#format1

## Enrichment analysis finds statistically significant associations




- 6 out of 7 selected SNPs overlap with an epigenomic mark
- How likely this to be observed by chance? (Fisher's exact test)

## Questions GenomeRunner can answer

- Which cell type-specific epigenomic elements are most statistically significantly associated with and potentially altered by, a set of SNPs of interest?
- What is the potential functional impact of a set of SNPs of interest from intergenic regions? As compared with the SNPs of interest from intronic/exonic regions?
- How do SNPs in one population differ from SNPs in another population in their associations with epigenomic elements, and which elements differ?
- How similar, or different, are sets of patient-specific rare SNPs, in their associations with all known genome annotation regions? As compared with common SNPs?

## Interpreting disease- and trait-specific SNPs within epigenomic context

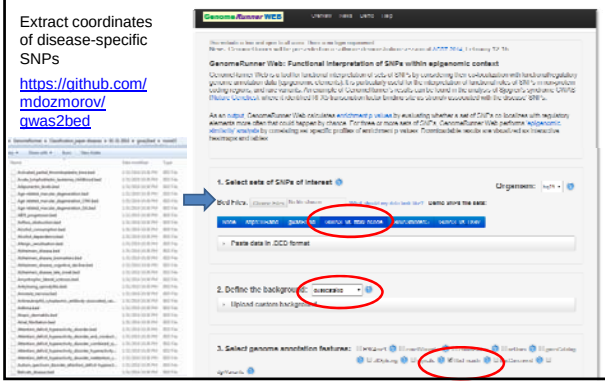
## GWAScatalog: extract disease- and trait-associated sets of SNPs


https://github.com/mdozmorov/gwas2bed - scripts to extract disease-specific genomic coordinates

## Load the data and prepare the run

Extract coordinates of disease-specific SNPs

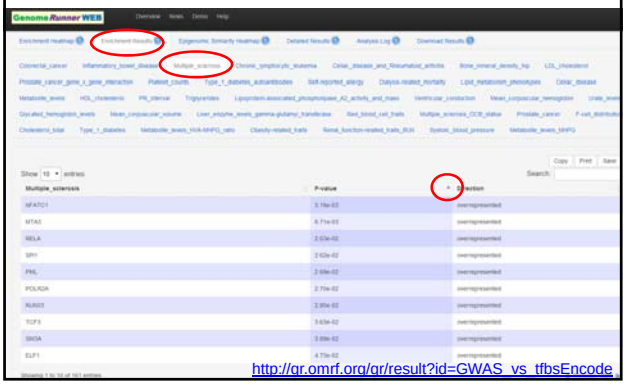
https://github.com/mdozmorov/gwas2bed



## Enrichment analysis heatmap


http://gr.omrf.org/gr/result?id=GWAS_vs_tfbsEncode

## Enrichment analysis tables

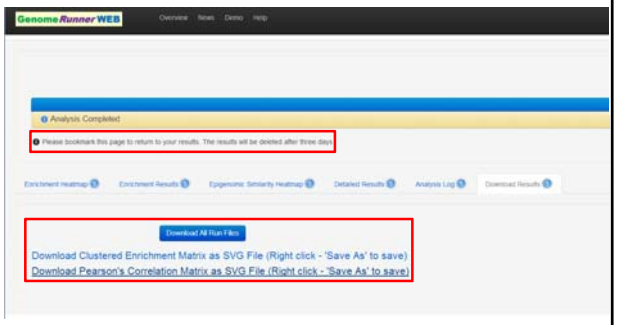


http://gr.omrf.org/gr/result?id=GWAS_vs_tfbsEncode

## Epigenomic similarity heatmap


http://gr.omrf.org/gr/result?id=GWAS_vs_tfbsEncode

## Download the results

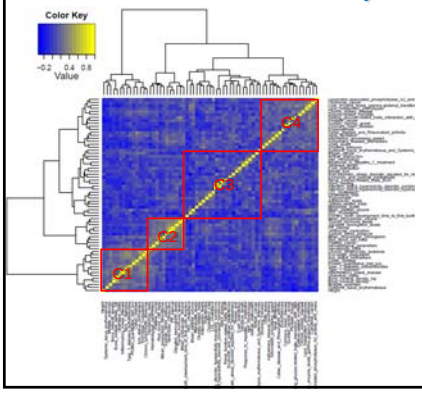



## GenomeRunner: Deeper exploration of the results

## Deeper exploration of the results

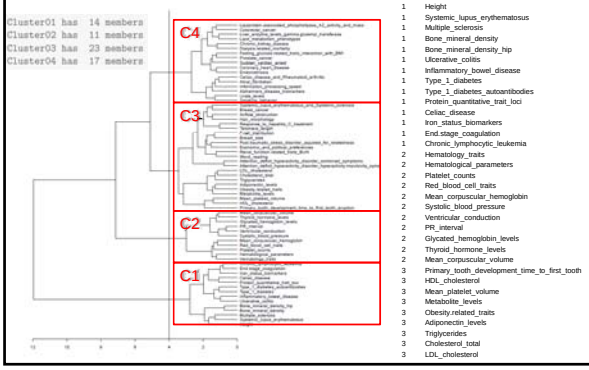
- https://github.com/mdozmorov/R.genomerunner
- Tutorial
- R scripts for data filtering
- Visualization of the enrichment- and epigenomic similarity results
- Max/Min epigenomic similarity
- Works best with RStudio

## Enrichment similarity visualization




- Data filtering
- Kendall correlation
- Euclidean/ ward clustering
- Visual assessment of cluster groups

## Closer look at the cluster groups




## Differentially enriched elements

`degs.matrix` – total counts of differentially enriched elements between cluster pairs

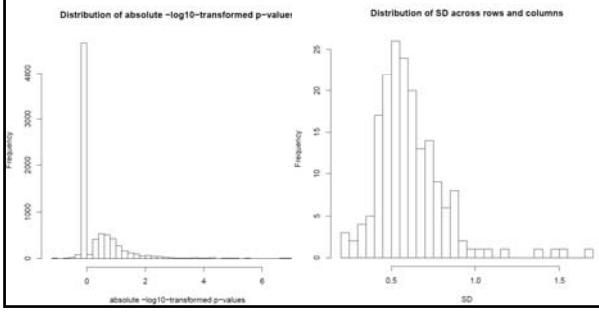
| | c1 | c2 | c3 | c4 |
|---|---|---|---|---|
| c1 | 0 | 11 | 7 | 6 |
| c2 | | 0 | 8 | 9 |
| c3 | | | 0 | 0 |
| c4 | | | | 0 |

*limma* analysis results, and average enrichments in the cluster groups

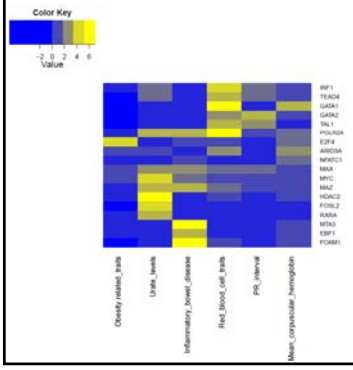
| | logFC | AveExpr | t | P.Value | adj.P.Val | B | l.vs.j | l.av | j.av |
|---|---|---|---|---|---|---|---|---|---|
| GATA2 | -1.636 | 0.489 | -5.343 | 1.81E-08 | 2.14E-06 | 9.19 | c1 vs. c2 | 0.06 | 1.69 |
| NR2F2 | -1.212 | 0.378 | -5.770 | 1.89E-07 | 1.12E-05 | 6.92 | c1 vs. c2 | 0.15 | 1.35 |
| SMARCA4 | -1.203 | 0.290 | -5.105 | 2.63E-06 | 9.56E-05 | 4.39 | c1 vs. c2 | 0.02 | 1.22 |
| GATA1 | -1.932 | 0.571 | -5.051 | 3.25E-06 | 9.58E-05 | 4.19 | c1 vs. c2 | 0.07 | 2.00 |
| TAL1 | -1.336 | 0.400 | -4.923 | 7.77E-06 | 1.53E-04 | 3.36 | c1 vs. c2 | 0.09 | 1.41 |
| RELA | 1.174 | 0.455 | 4.703 | 1.22E-05 | 2.05E-04 | 2.93 | c1 vs. c2 | 1.26 | 0.09 |
| RUNX3 | 1.003 | 0.484 | 4.620 | 1.66E-05 | 2.45E-04 | 2.63 | c1 vs. c2 | 1.38 | 0.38 |
| CCNT2 | -1.133 | 0.422 | -4.471 | 2.87E-06 | 3.77E-04 | 2.11 | c1 vs. c2 | 0.10 | 1.23 |
| TEAD4 | -1.233 | 0.654 | -4.258 | 6.20E-05 | 6.65E-04 | 1.38 | c1 vs. c2 | 0.19 | 1.42 |
| ARID3A | -1.193 | 0.560 | -3.790 | 3.12E-04 | 3.06E-03 | -0.14 | c1 vs. c2 | 0.02 | 1.20 |
| MTA3 | 1.070 | 0.492 | 3.592 | 6.00E-04 | 5.05E-03 | -0.75 | c1 vs. c2 | 1.48 | 0.41 |

## Properties of the transformed p-values

Visually assess p-value and SD cutoffs for filtering non-significant results



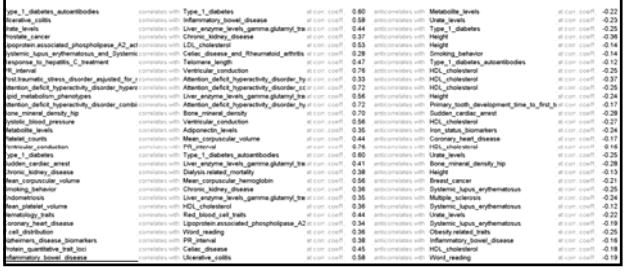

## Enrichment analysis visualization




- Filtering out rows/columns with non-significant p-values
- Clustering highlights groups of differentially enriched epigenomic elements

## Extreme epigenomic (dis-)similarities

- Maximum and minimum correlations among disease-specific SNP sets

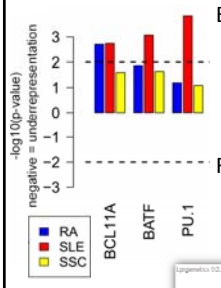


## Conclusions

- Disease-specific SNP sets show distinct epigenomic associations and can be grouped by their epigenomic similarity
- Epigenomic similarity identifies known and novel relationships among the diseases
- Enrichment analysis identifies disease-specific epigenomic elements, such as MTA3, EBF1 and FOXM1 transcription factor binding sites enriched in "Inflammatory bowel disease"-associated SNPs

## Other examples of GenomeRunner analysis

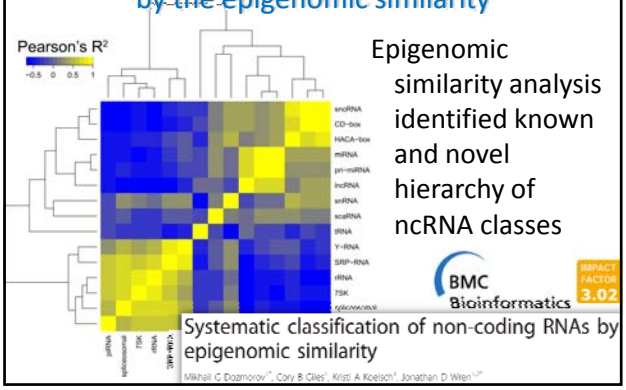
## Epigenomic elements enriched in the promoters of autoimmunity susceptibility genes




Enrichment analysis of the promoters of the disease-specific gene sets revealed PU.1 and H3K4me1/2 histone marks strongly enriched in SLE genes

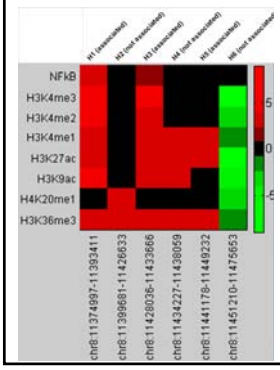
Reverse epigenomic engineering analysis identified known and novel immune-related genes sharing similar set of epigenomic marks in their promoters


## Systematic classification of non-coding RNAs by the epigenomic similarity




Epigenomic similarity analysis identified known and novel hierarchy of ncRNA classes

## Epigenomic differences in the disease-associated SNP haploblocks




Haploblock-specific epigenomic differences suggest disease mechanisms


## RFX5 transcription factor binding site found to be affected in Sjögren's syndrome