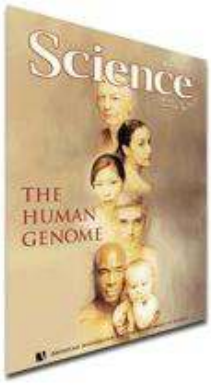
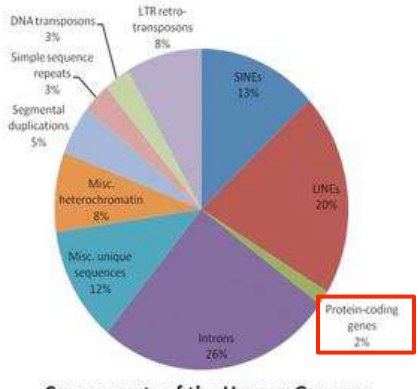
GenomeRunner web server: epigenomic similarity and differences define the functional impact of SNP sets

Mikhail Dozmorov, Ph.D.
Department of Biostatistics, Virginia Commonwealth University
mikhail.dozmorov@vcuhealth.org
03/05/2016

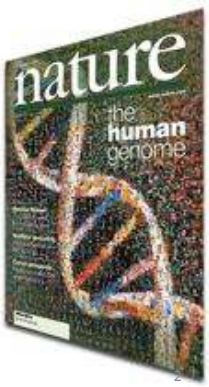
The human genome in a nutshell

- **Human genome is big - ~3.2 billion base pairs (~6ft long)**
- **20,000-25,000 genes (~2% of the genome)**



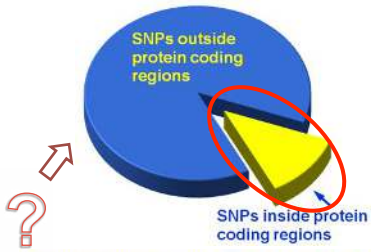


Components of the Human Genome



Genome research is gene-centric

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



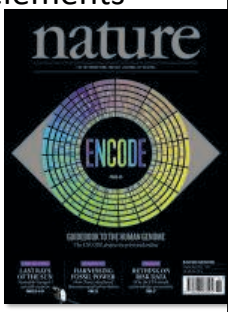
Only 12% of SNPs are located in, or occur in tight linkage disequilibrium with, protein-coding regions

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.
Proc Natl Acad Sci U S A. 2009 Jun 9;106(23):9362-7. Epub 2009 May 27.

3

Genome gets annotated with regulatory information

- **ENCODE project** – annotating regulatory elements
  - DNaseI hypersensitive sites
  - Histone modification marks
  - Transcription Factor Binding Sites
  - DNA methylation



- **Epigenomic (regulatory) data** - genomic regions annotated as carrying functional and/or regulatory potential

4

Even more epigenomic data

- **Roadmap Epigenomics** – uniformly processed annotations of regulatory elements across normal tissues.

ROADMAP epigenomics PROJECT — Integrative analysis of 111 reference human epigenomes

GTEx Portal

Human Epigenome Project

FANTOM — FUNCTIONAL ANNOTATION OF THE MAMMALIAN GENOME

IHEC — International Human Epigenome Consortium

5

Gap between data generation & data understanding

The Road to the $1,000 Genome

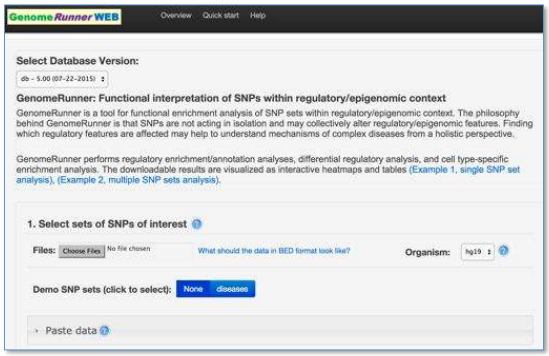
"**There is a growing gap between the generation of massively parallel sequencing output and the ability to process and analyze the resulting data**," says Canadian cancer research John

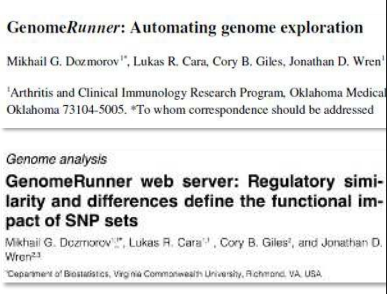
being solved. Thus, the idea of a '$1000 genome, $1 000 000 genome interpretation' was expressed by the president of the American College of Medical Genetics, Bruce Korf. In this review, we

6

GenomeRunner – a global positioning system within the genome

- **Finds significantly enriched co-localizations between 'omics' data and genome-wide epigenomic data**





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

Genome analysis

GenomeRunner web server: Regulatory similarity and differences define the functional impact of SNP sets

Mikhail G. Dozmorov[1,*], Lukas R. Cara[1], Cory B. Giles[2], and Jonathan D. Wren[2,3]

[1]Department of Biostatistics, Virginia Commonwealth University, Richmond, VA, USA

http://www.integrativegenomics.org

7

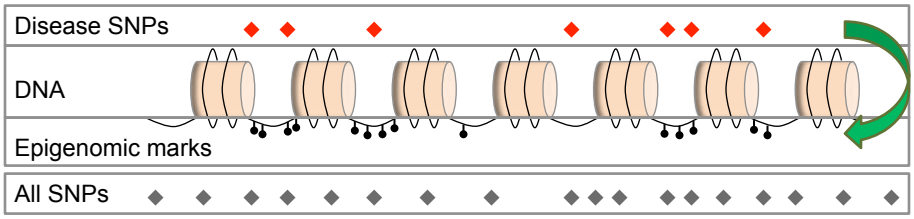
Epigenomic enrichment analysis

Gene enrichment and epigenome enrichment

- **Gene set enrichment analysis**
  - **Genes**
  - enriched in
  - **gene ontology (functional) annotations**
- **Epigenome enrichment analysis**
  - **SNPs**
  - enriched in
  - **regulatory (epigenomic) annotations**

9

Statistics of epigenomic enrichments



- 6 out of 7 disease-associated SNPs overlap with epigenomic marks
- How likely this to be observed by chance? (Chi-square test/Binomial test/Permutation test)

10

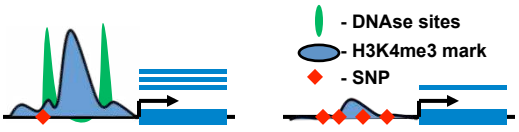
GenomeRunner vs. other enrichment analysis tools

Name	URL	Interface					Database		Analyses					Updated in 2015
		Web	Visualization	Command line	Scalable /HPC	Source code	Data filtering	Database provided	User-provided database support	Annotation analysis	Enrichment analysis	Cell type enrichment analysis	Regulatory similarity	Differential regulatory analysis	
EpiGraph	http://epigraph.mpi-inf.mpg.de	+	-	-	-	-	+	+	+	+	+	-	-	-	-
EpiExplorer	http://epiexplorer.mpi-inf.mpg.de	+	+	-	-	Through registration	+	+	+	+	Via other tools	-	-	-	-
Genomic HyperBrowser	https://hyperbrowser.uio.no/hb	+	Via other tools	Limited	-	-	+	+	+	+	+	-	+	-	+
Genomic Association Tester (GAT)	https://github.com/AndreasHeger/gat	-	-	+	-	+	-	-	+	+	+	-	-	-	+
PocBet	http://www.pocbat.org	+	+	+	-	+	-	Limited	+	+	+	-	-	-	-
EpiRegNet	http://wanglab.hku.hk/EpiRegNet	+	+	-	-	-	+	+	+	-	+	-	-	-	-
GREAT	http://bejerano.stanford.edu/great	+	+	+	-	+	-	Limited	-	-	+	-	-	-	+
BEDTools	https://github.com/arq5x/bedtools2	Via other tools	-	+	-	+	+	-	+	+	+	-	-	-	+
GenometriCorr	http://genometricorr.sourceforge.net	Via other tools	+	+	-	+	-	Limited	+	-	+	-	-	-	+
Genome Track Analyzer (AnCorr)	http://ancorr.eimb.ru/	+	-	-	-	-	+	Limited	+	-	-	-	+	-	+
ENCODE ChIP-Seq Significance Tool	http://encodenet.stanford.edu	+	-	-	-	-	+	+	-	-	+	Limited	-	-	+
The Genboree Epigenome Toolset	http://www.genboree.org/site/epigenome	+	Via other tools	-	-	-	+	+	-	-	+	-	+	-	+
INRICH	https://atgu.mgh.harvard.edu/inrich	-	-	+	-	+	-	-	+	-	+	-	-	-	-
Enrichr	http://amp.pharm.mssm.edu/Enrichr	+	+	Limited	-	Limited	-	+	-	-	+	+	-	-	+
BioMart Enrichment Tool	http://central.biomart.org/enrich	+	-	-	-	-	+	+	+	-	+	-	-	-	+
KGWAS	https://github.com/joepickrell/kgwas	-	-	+	-	+	+	-	+	-	+	-	-	-	+
GoShifter	https://www.broadinstitute.org/mpg/goshifter	-	-	+	-	+	-	-	+	-	+	-	-	-	+
FORGE	http://browser.1000genomes.org	+	+	+	-	+	+	Limited	-	+	+	+	-	-	+
ChIPSeeker	http://www.bioconductor.org/packages/release/bioc/html/ChIPseeker.html	-	+	+	-	+	+	Limited	+	-	+	-	-	-	+
[illegible]	[illegible]	-	-	-	-	-	-	-	-	-	-	-	-	-	-
GenomeRunner	http://www.integratomics.com	+	+	+	+	+	+	+	+	+	+	+	+	+	+

11

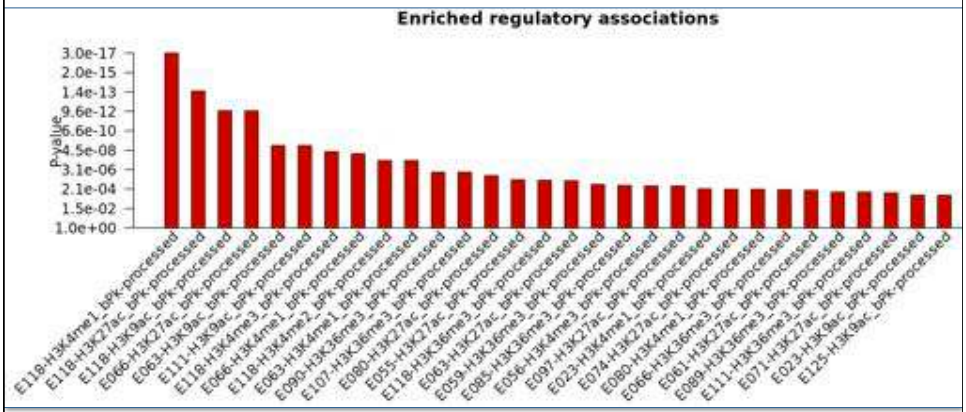
Functional impact of SNPs

- Hypothesis: SNPs in epigenomic regions may disrupt regulation
- Epigenomic regions enriched in SNPs (SNP burden) are disrupted more
- Epigenomic signature = types of epigenomic marks most enriched in SNPs



12

Functional impact of SNPs
HDL cholesterol



Epigenomic enrichments are sorted from most to least significant

E118-H3K4me1_bPk-processed = <cell/tissue ID>-<factor ID>-<source info>

13

Functional impact of SNPs
HDL cholesterol

Show 10 entries Search:

	epigenomic_name	p.value	direction	adj.p.val	cell	cell_desc	factor	factor_desc	source	source_desc
342	E118-H3K4me1_bPk-processed	3.86e-20	Overrepresented	3.78e-17	E118	HepG2 Hepatocellular Carcinoma Cell Line	H3K4me1	Histone H3 (mono methyl K4). Is associated with en ...	processed	Roadmap consolidated
345	E118-H3K27ac_bPk-processed	3.01e-16	Overrepresented	1.47e-13	E118	HepG2 Hepatocellular Carcinoma Cell Line	H3K27ac	Histone H3 (acetyl K27). As with H3K9ac, associate ...	processed	Roadmap consolidated
346	E118-H3K9ac_bPk-processed	3.32e-14	Overrepresented	1.08e-11	E118	HepG2 Hepatocellular Carcinoma Cell Line	H3K9ac	Histone H3 (acetyl K9). As with H3K27ac, associate ...	processed	Roadmap consolidated

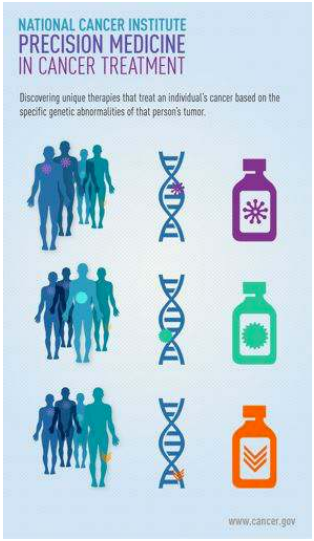
HDL cholesterol-associated SNPs are enriched in liver-specific activating H3K4me1 and H3K27ac histone modification marks

E118-H3K4me1_bPk-processed = <cell/tissue ID>-<factor ID>-<source info>

14

Epigenomic similarity analysis

Epigenomic similarity among SNP sets



- Similar SNP sets may have similar functional impact = be enriched in similar types of epigenomic regions
- Comparing functional impact of SNP sets may help understanding epigenomic mechanisms of complex diseases

16

Epigenomic similarity among disease-associated SNP sets

Name	SNP count	Category
Bone mineral density	430	other
C reactive protein	141	other
Creatinine levels	47	other
Renal function related traits	128	other
Urate levels	192	other
Platelet counts	368	other
Red blood cell traits	299	other
Alzheimers combined	175	neurologic
Progressive supranuclear palsy	74	neurologic
Restless legs syndrome	46	neurologic
Migraine	63	neurologic
Type 2 diabetes	522	metabolic
Chronic kidney disease	197	metabolic
Fasting glucose related traits	110	metabolic
HDL cholesterol	331	metabolic
LDL cholesterol	261	metabolic
Liver enzyme levels	173	metabolic
Triglycerides	234	metabolic
Allergy	160	immunologic
Behcets disease	66	immunologic
Autoimmune thyroiditis	48	immunologic
Type 1 diabetes	339	immunologic
Multiple sclerosis	696	immunologic
Kawasaki disease	56	immunologic
Asthma	121	immunologic
Crohns disease	787	immunologic
Ulcerative colitis	447	immunologic
Primary biliary cirrhosis	275	immunologic
Primary sclerosing cholangitis	80	immunologic
Celiac disease	370	immunologic
Atopic dermatitis	118	immunologic
Psoriasis	264	immunologic
Alopecia areata	66	immunologic
Vitiligo	155	immunologic
Systemic lupus erythematosus	221	immunologic
Systemic sclerosis	43	immunologic
Rheumatoid arthritis	195	immunologic
Juvenile idiopathic arthritis	236	immunologic
Ankylosing spondylitis	207	immunologic

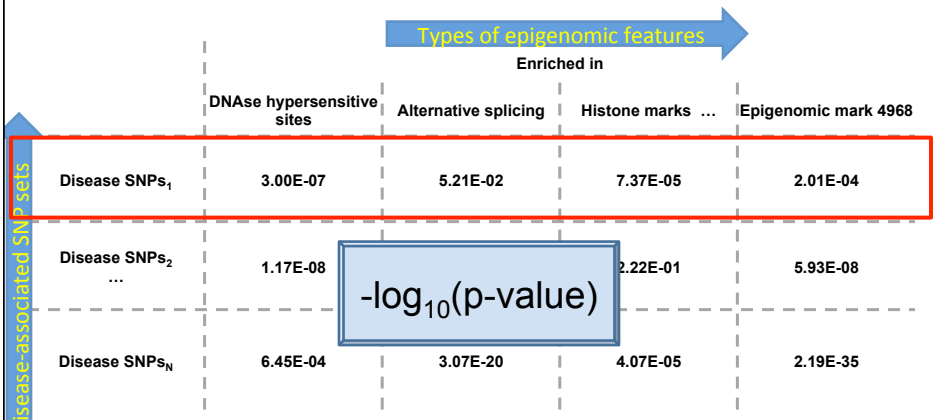
- We analyzed SNP sets associated with 39 complex diseases and traits
  - 21 immunologic diseases
  - 4 neurologic diseases/traits
  - 7 metabolic diseases/traits
  - 7 other traits

doi:10.1038/nature13835

Genetic and epigenetic fine mapping of causal autoimmune disease variants

Kyle Kai-How Farh[1,2]*, Alexander Marson[2]*, Jiang Zhu[1,4,5,6], Markus Kleinewietfeld[1,7,8], William J. Housley[7], Samantha Beik[2], Noam Shoresh[1], Holly Whitton[1], Russell J. H. Ryan[1,2], Alexander A. Shishkin[1,9], Meital Hatan[1], Marlene J. Carrasco-Alfonso[9], Dita Mayer[9], C. John Luckey[9], Nikolaos A. Patsopoulos[1,10,11], Philip L. De Jager[1,10,11], Vijay K. Kuchroo[12], Charles B. Epstein[1], Mark J. Daly[1,2], David A. Hafler[1,7]§ & Bradley E. Bernstein[1,4,5,6]§

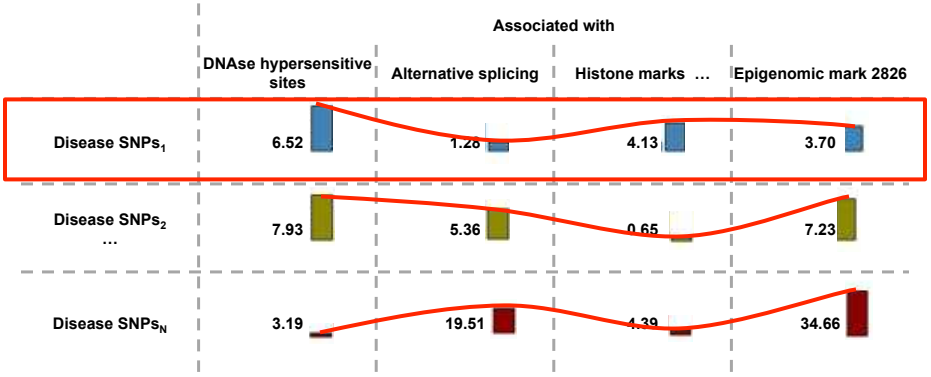
Defining epigenomic enrichment profiles of disease-associated SNP sets



- **Epigenomic enrichment profile** – SNP set-specific vector of epigenomic enrichment p-values

Comparing epigenomic enrichment profiles of disease-associated SNP sets

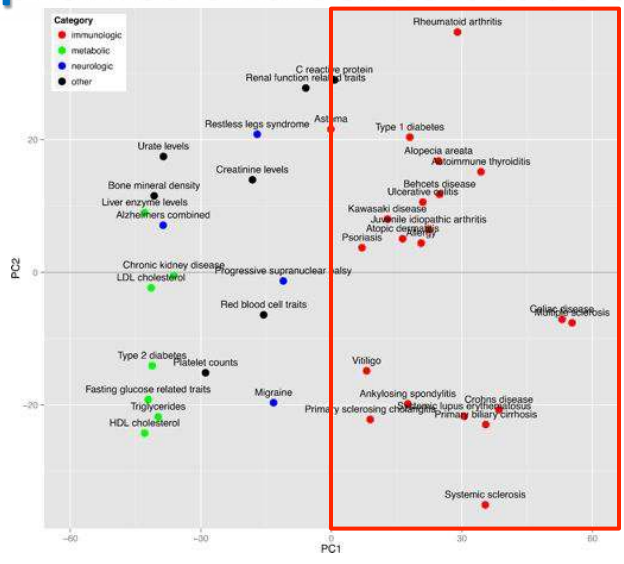
-log10-transform p-values – larger numbers = more significant p-values



- **Epigenomic enrichment profile** – SNP set-specific vector of –log10-transformed epigenomic enrichment p-values
- Compare them using PCA, Spearman correlation

19

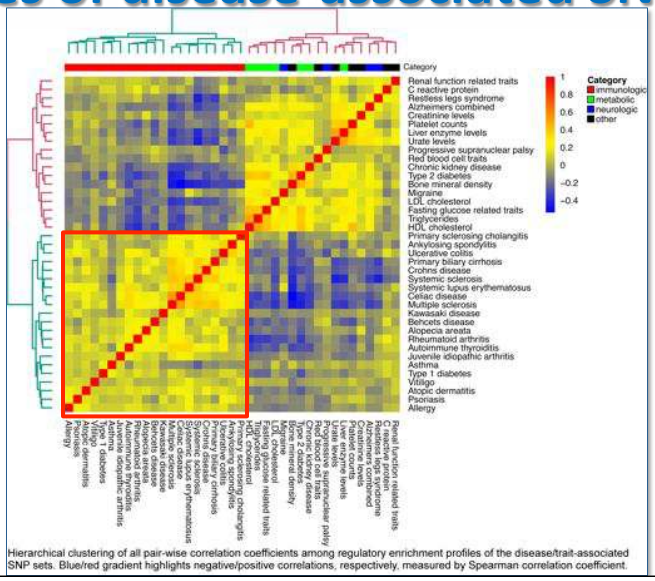
Comparing epigenomic enrichment profiles of disease-associated SNP sets



PCA separates immunologic disease-associated SNP sets as the most epigenomically distinct from others

20

Comparing epigenomic enrichment profiles of disease-associated SNP sets

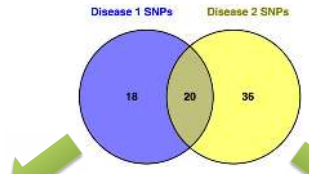


Hierarchical clustering of all pair-wise correlation coefficients among regulatory enrichment profiles of the disease/trait-associated SNP sets. Blue/red gradient highlights negative/positive correlations, respectively, measured by Spearman correlation coefficient.

21

Validating epigenomic similarity

Shared genomic loci – overlap between SNP sets



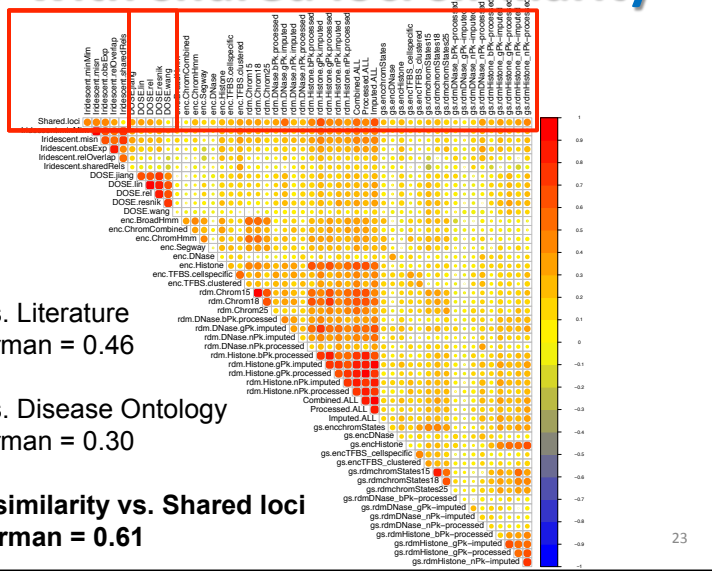
Semantic (literature) similarity

Disease Ontology similarity

Epigenomic similarity

22

Epigenomic similarity correlates best with shared loci similarity



Shared loci vs. Literature
Median Spearman = 0.46

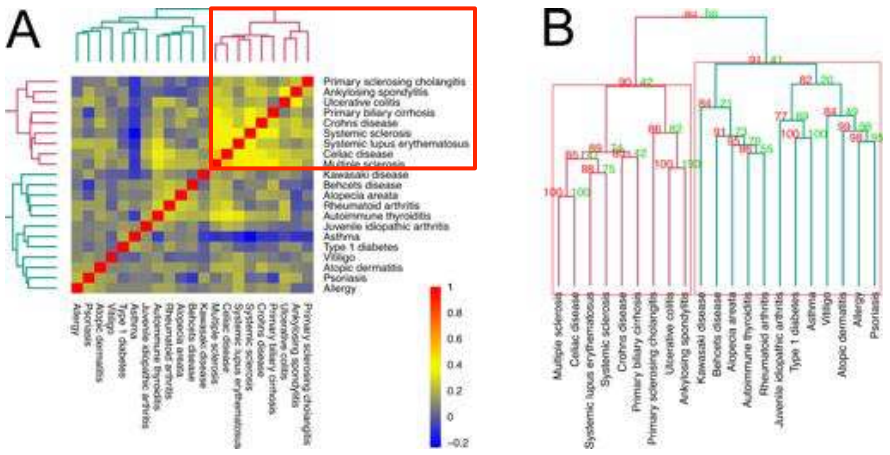
Shared loci vs. Disease Ontology
Median Spearman = 0.30

Epigenomic similarity vs. Shared loci
Median Spearman = 0.61

23

Differential epigenomic analysis

Two groups of immunologic diseases form epigenomically distinct clusters

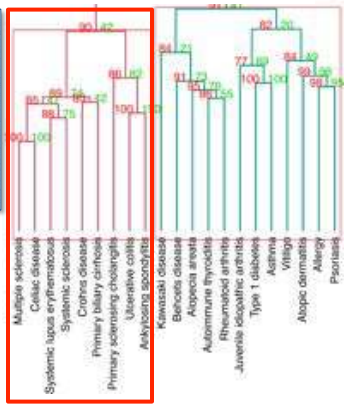


Two groups of immunologic diseases form epigenomically distinct clusters

'c1'/'c2' - average enrichment p-values in groups 'c1' and 'c2', respectively.
'adj.p.val' - adjusted for multiple testing p-value of the enrichment differences between group 1 'c1' and group 2 'c2' of immunologic diseases

epigenomic_name	cell	cell_desc	factor	factor_desc	c1	c2	adj.p.val
E043-7Enh-chromStates15	E043	Primary T helper cells from peripheral blood	7Enh	Enhancers	1.35E-01	2.72E-05	5.20E-03
E047-7Enh-chromStates15	E047	Primary T CD8+ naive cells from peripheral blood	7Enh	Enhancers	3.90E-01	2.71E-04	7.92E-03
E038-7Enh-chromStates15	E038	Primary T helper naive cells from peripheral blood	7Enh	Enhancers	3.45E-01	4.77E-04	5.59E-02
E039-7Enh-chromStates15	E039	Primary T helper naive cells from peripheral blood	7Enh	Enhancers	4.02E-01	1.64E-03	5.81E-02
E040-5TxWk-chromStates15	E040	Primary T helper memory cells from peripheral blood 1	5TxWk	Weak transcription	8.12E-01	2.68E-03	6.50E-02
E029-7Enh-chromStates15	E029	Primary monocytes from peripheral blood	7Enh	Enhancers	9.21E-01	5.59E-03	4.96E-03
E035-7Enh-chromStates15	E035	Primary hematopoietic stem cells	7Enh	Enhancers	5.94E-01	1.12E-02	3.41E-02
E124-6EnhG-chromStates15	E124	Monocytes-CD14+ RO01746 Primary Cells	6EnhG	Genic enhancers	8.91E-01	1.18E-02	5.81E-02
E030-7Enh-chromStates15	E030	Primary neutrophils from peripheral blood	7Enh	Enhancers	8.31E-01	1.25E-02	5.97E-02
E033-2TssAFlnk-chromStates15	E033	Primary T cells from cord blood	2TssAFlnk	Flanking Active TSS	7.69E-01	1.63E-02	8.43E-02
E124-2TssAFlnk-chromStates15	E124	Monocytes-CD14+ RO01746 Primary Cells	2TssAFlnk	Flanking Active TSS	5.00E-01	2.21E-02	8.52E-02
E039-2TssAFlnk-chromStates15	E039	Primary T helper naive cells from peripheral blood	2TssAFlnk	Flanking Active TSS	9.84E-01	3.11E-02	4.76E-02
E062-6EnhG-chromStates15	E062	Primary mononuclear cells from peripheral blood	6EnhG	Genic enhancers	8.66E-01	3.76E-02	9.08E-02
E039-15Quies-chromStates15	E039	Primary T helper naive cells from peripheral blood	15Quies	Quiescent-Low	6.68E-01	-4.53E-03	5.44E-02
E030-15Quies-chromStates15	E030	Primary neutrophils from peripheral blood	15Quies	Quiescent-Low	1.46E-01	-4.49E-03	2.52E-02
E047-15Quies-chromStates15	E047	Primary T CD8+ naive cells from peripheral blood	15Quies	Quiescent-Low	6.51E-01	-2.09E-03	7.92E-03
E038-15Quies-chromStates15	E038	Primary T helper naive cells from peripheral blood	15Quies	Quiescent-Low	8.41E-01	-1.64E-04	6.04E-02
E043-14ReprPCWk-chromStates15	E043	Primary T helper cells from peripheral blood	14ReprPCWk	Weak Repressed PolyCom	5.99E-01	-1.20E-04	1.83E-02
E037-15Quies-chromStates15	E037	Primary T helper memory cells from peripheral blood 2	15Quies	Quiescent-Low	9.83E-01	-1.36E-05	7.87E-02

group 2 c2 **group 1 c1**



- Disease-associated SNPs of nine immunologic diseases (group 2) are significantly **enriched in T helper cell-specific enhancers**
- .. and **depleted in T helper cell-specific quiescent chromatin states**

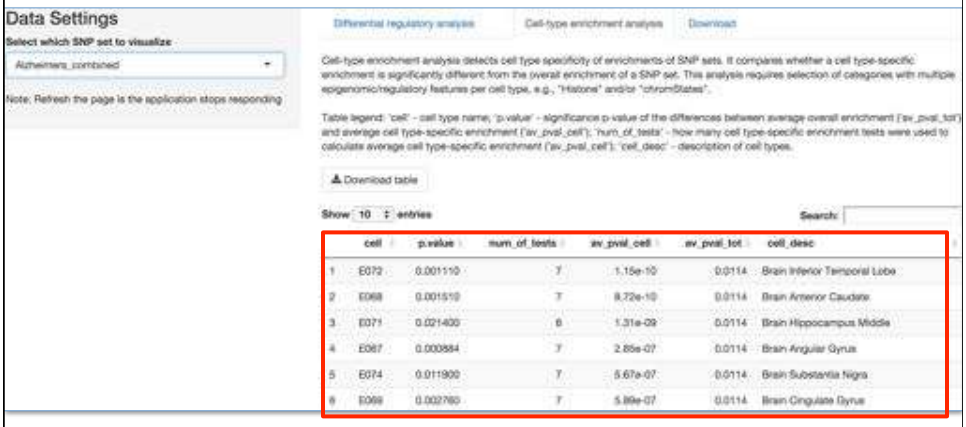
Cell type-specific epigenomic enrichment analysis

Cell type-specific epigenomic enrichment analysis defines the origin of the functional impact

- Epigenomic elements are cell- and tissue type specific
- Cell type-specific epigenomic enrichment analysis identifies cell types where epigenomic enrichments are the most significant

28

Cell type specificity of the functional impact of disease SNPs is relevant to disease pathology



Alzheimer SNPs are enriched in brain-specific epigenomic marks

29

Conclusions

- GenomeRunner defines potential functional impact of SNP sets via **epigenomic enrichment analysis**
- **Epigenomic similarity analysis** identifies regulatory **similarity and differences** among SNP sets
- **Cell type-specific enrichment** analysis prioritizes cell/tissue type specificity of the epigenomic enrichments
- Epigenomic enrichment analyses can be applied to any genomic signature, from disease-associated SNP sets to patient-specific genotypes

30

Thank you

Genome analysis

GenomeRunner web server: Regulatory similarity and differences define the functional impact of SNP sets

Mikhail G. Dozmorov[1,†,*], Lukas R. Cara[1,†], Cory B. Giles[2], and Jonathan D. Wren[2,3]

[1]Department of Biostatistics, Virginia Commonwealth University, Richmond, VA, USA

Mikhail Dozmorov, Ph.D.
Department of Biostatistics, Virginia Commonwealth University
mikhail.dozmorov@vcuhealth.org
03/05/2016
https://github.com/mdozmorov/presentations