Finding Enrichments of Functional Annotations for Disease-Associated Single-Nucleotide Polymorphisms

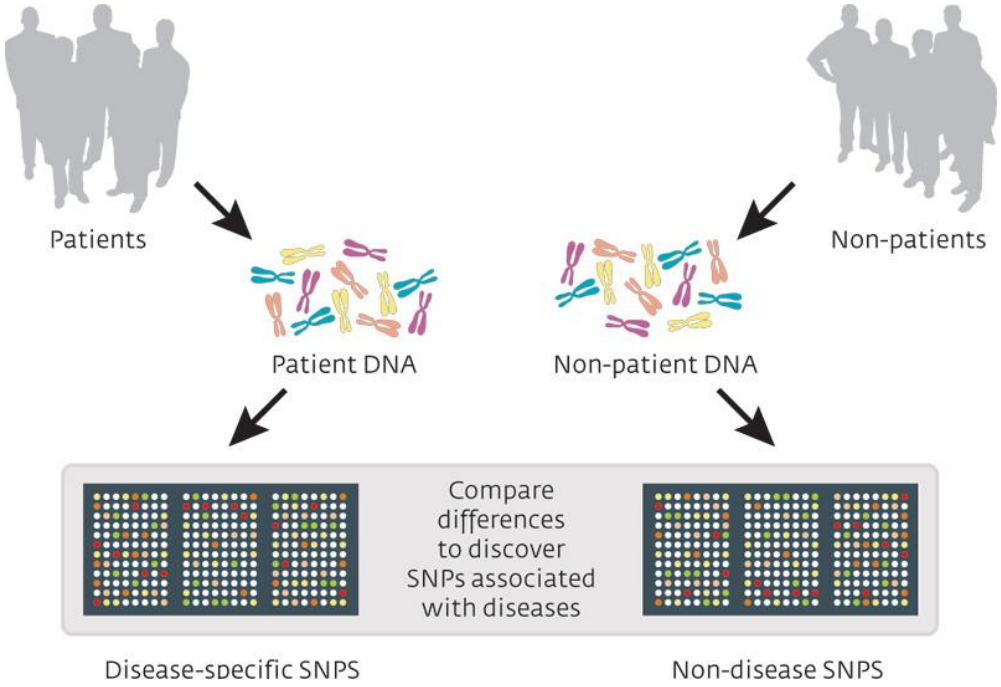
Steven Homberg
Mentor: Dr. Luke Ward

Third Annual MIT PRIMES Conference, May 18-19 2013

Motivation

- Relevance of genome in human traits, specifically disease
- Some of the genome's function is known, much is not
- Enrichments provide links between observable traits and candidates for biological explanation
- Combine GWAS, genomic annotation to extract more information from each

GWAS



- **G**enome-**W**ide **A**ssociation **S**tudy
  - Associations between mutations (SNPs) and traits/diseases
  - Does not provide information about the type of variant

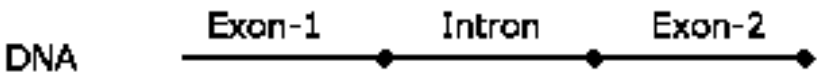
SNP

- SNP = Single Nucleotide Polymorphism
  - change of a single nucleotide (A, C, G, T)
  - can be insertion, deletion, substitution

Genomic Annotation

- Annotations differentiate non-coding regions of the genome
- Diverse information
  - Presence in coding exon, intron, intergenic region, untranslated region
  - Behavior as enhancer or promoter binding site
  - Member or Regulatory Motif

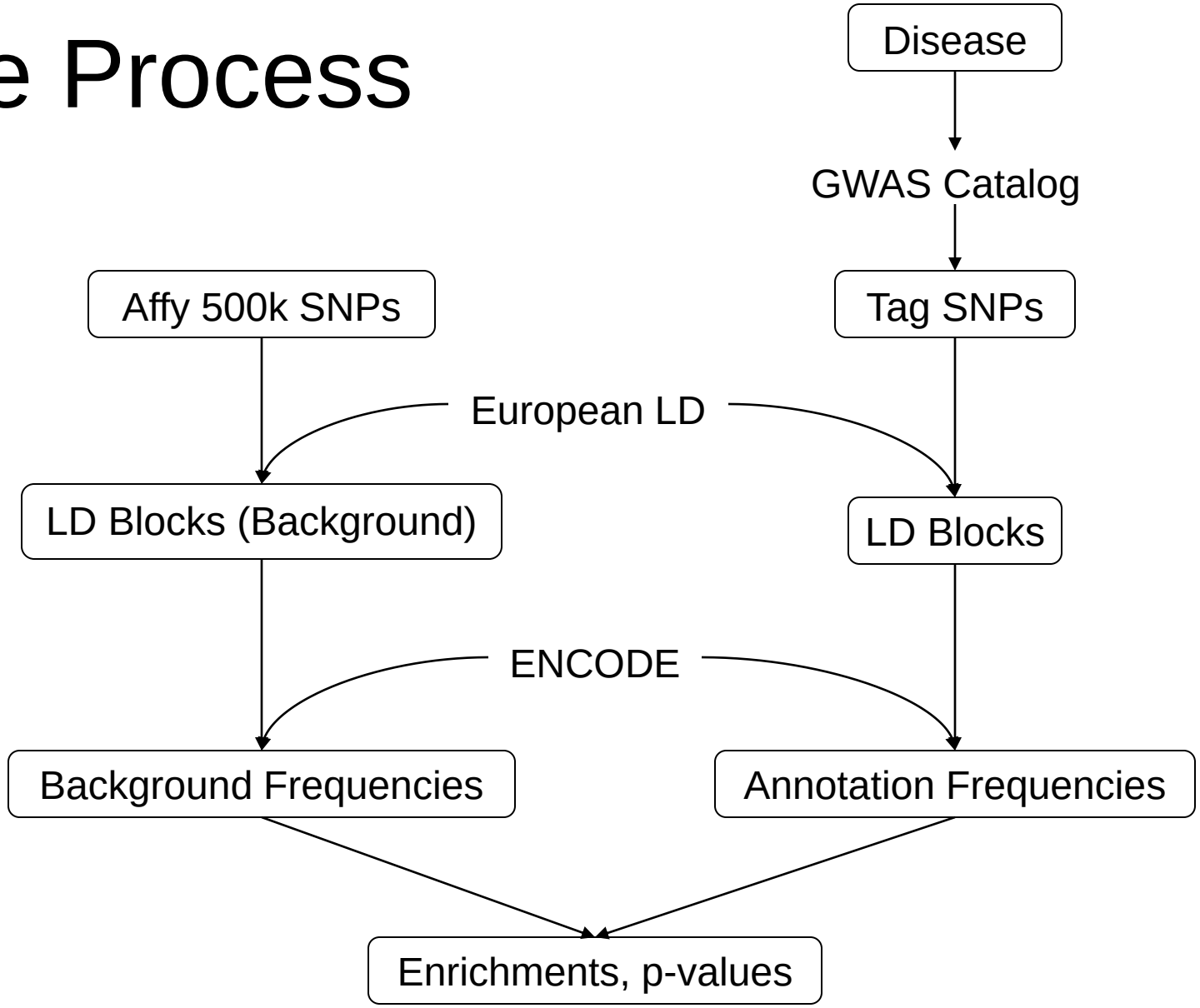
- Different segments of the genome serve different purposes



Enrichment

- GWAS links traits with genomic locations
- Annotations link genomic locations with genetic functions
- Enrichments bridge the gap, linking traits with genetic functions
  - Statistical process
  - Identifies increased frequency of an annotation at disease-associated locations compared to background frequency

The Process


Disease
GWAS Catalog
Affy 500k SNPs
Tag SNPs
European LD
LD Blocks (Background)
LD Blocks
ENCODE
Background Frequencies
Annotation Frequencies
Enrichments, p-values

Disease to Tag SNPs

- GWAS catalogs compile GWAS results associating diseases with SNPs
- Collect all SNPs associated with one disease

1	rs903263
16	rs3803662
14	rs1314913
6	rs9383938
19	rs8100241
20	rs2284378
6	rs17530068
10	rs3750817
3	rs6788895
...	...

Tag SNPs to LD Blocks

5	rs10069690									
9	rs1011970	rs77283072								
7	rs10263639	rs10270452								
2	rs10490113	rs2418490	rs78794871	rs139648605	rs17050034	rs76100924	rs79879739	rs3062468	rs10595980	▪▪▪

- Genetic locations grouped population-wide by genetic linkage
  - Association with a SNP is association with a block
- Any SNP in a block may be causal
- European LD

LD Blocks to Annotations

- Annotations from each SNP may influence trait regulation
- From ENCODE project, compiled in HaploReg

Head SNP	TSS DIST	EUR FREQ	LD SNP COUNT	HSMM	HSMMtube	...
rs10069690	13225	0.27	2	0	0	
rs1011970	-16459	0.16	3	0	0	
rs10263639	434967	0.16	3	0	0	
rs10490113	214918	0.1	12	0	0	
rs10466033	-18789	0.01	43	3	3	
...						

Annotations to Enrichment

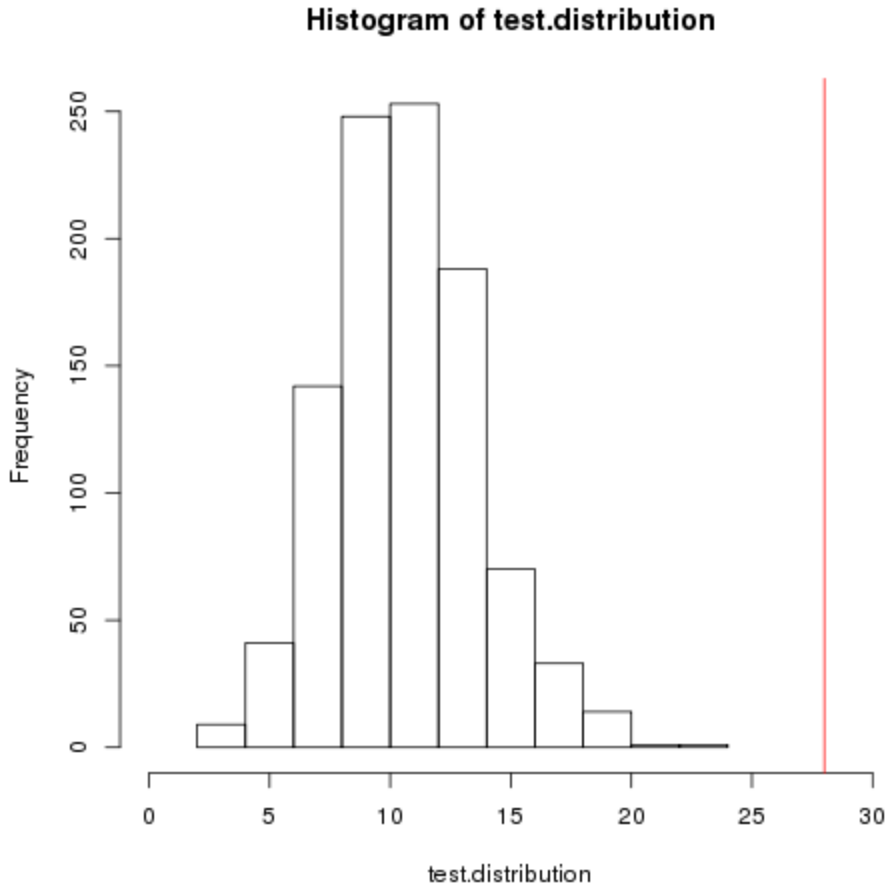
- Compare annotation frequencies in LD blocks associated with a disease to background frequencies
- Affy 500K
- LD blocks for background SNPs as well

Enrichment Statistics

- Empirical Null Distribution
- Annotation tally is a function of LD block length, TSS (transcription start site) distance, allelic frequency
- Null distribution generated from randomized simulation, controlling for confounding factors
- Normal model for null distribution, with the annotation tally, gives a p-value

Sample Test

MCF.7 Frequency	28
Mean of empirical null distribution	11.13
Standard deviation of empirical null distribution	3.02
z-score	5.58
p-value	1.20E-08



Results: Breast Cancer

ANNOT (DNase)	P.VALUE
MCF.7	1.75E-08
HUVEC	8.82E-07
HGF	1.02E-05
HRE	2.47E-05
ProgFib	3.55E-05
HMVEC.LBl	5.03E-05

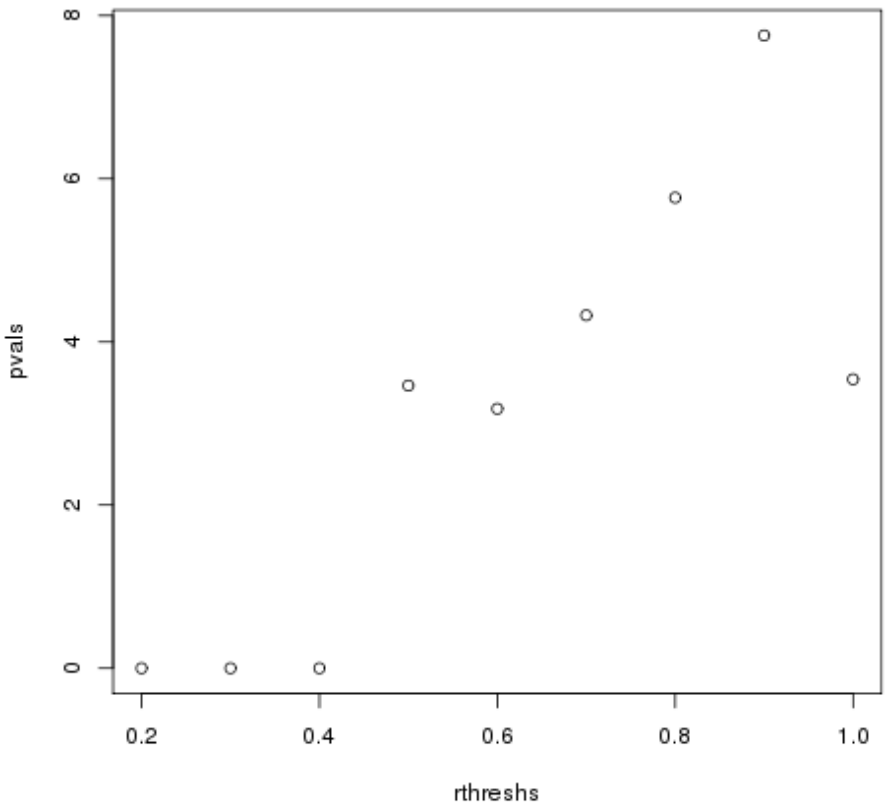
- Significant enrichments of functional annotations for disease
  - MCF.7 cell type enriched for a DNase hypersensitivity site, breast cancer

More Breast Cancer Results

ANNOT (Protein)	P.VALUE
ZNF274	1.63E-11
ZEB1	2.62E-08
GATA2	3.45E-06
CJUN	3.48E-06
NRF1	7.11E-05
TCF4	7.83E-05

ANNOT (Enhancer)	P.VALUE
PFK.3	2.78E-06
ESO	0.000269
ADI.NUC	0.000783
PFK.2	0.001384
R.SMUS	0.001603
GAS	0.001772

R^2 Threshold



- Linkage disequilibrium between SNPs varies in strength
- R^2 measures strength of phenomenon
- Different threshold, different LD blocks, different results

Summary

- Enrichment statistically links diseases to potential biological mechanisms
  - Bridge between GWAS and genomic annotation
  - Platform for further biological investigation
- The procedure is subject to improvement
  - R-squared threshold
  - Correction factors in null-distribution generation

Future Work

- Investigate other diseases
- Biological hypotheses to explain statistical enrichments
- Optimize parameters, correction factors
- Release computational tool for community use on new annotations, SNPs, diseases

Special Thanks to:

The MIT PRIMES Program

Mentor Dr. Luke Ward

Supervisor Prof. Manolis Kellis

Citations

Broad Institute Logo. N.d. Graphic. Broad Institute Website Web. 11 May 2013. <http://www.broadinstitute.org/ >.

MIT PRIMES Logo. N.d. Photograph. MIT PRIMES Website Web. 11 May 2013. <http://web.mit.edu/primes/>.

LD map. 2009. Photograph. MINGKHWAN'S BIOINFORMATICS Web. 11 May 2013. <http://woratanti.wordpress.com/>.

Single Nucleotide Polymorphisms. N.d. Photograph. Chinese Medical and Biological Information Web. 11 May 2013. <http://cmbi.bjmu.edu.cn/cmbidata/snp/index00.htm>.

GWAS. N.d. Photograph. Max-Planck-Gesellschaft Web. 11 May 2013. <http://www.mpg.de/10680/Modern_psychiatry?print=yes >.

RNA Splicing. N.d. Photograph. The Montegen's Pocket Science Web. 11 May 2013. <http://www.montegen.com/Montegen/Nature_of_Business/The_Library/The_Pocket-Science/The_Pocket-Science_Vol__7/the_pocket-science_vol__7.htm >.