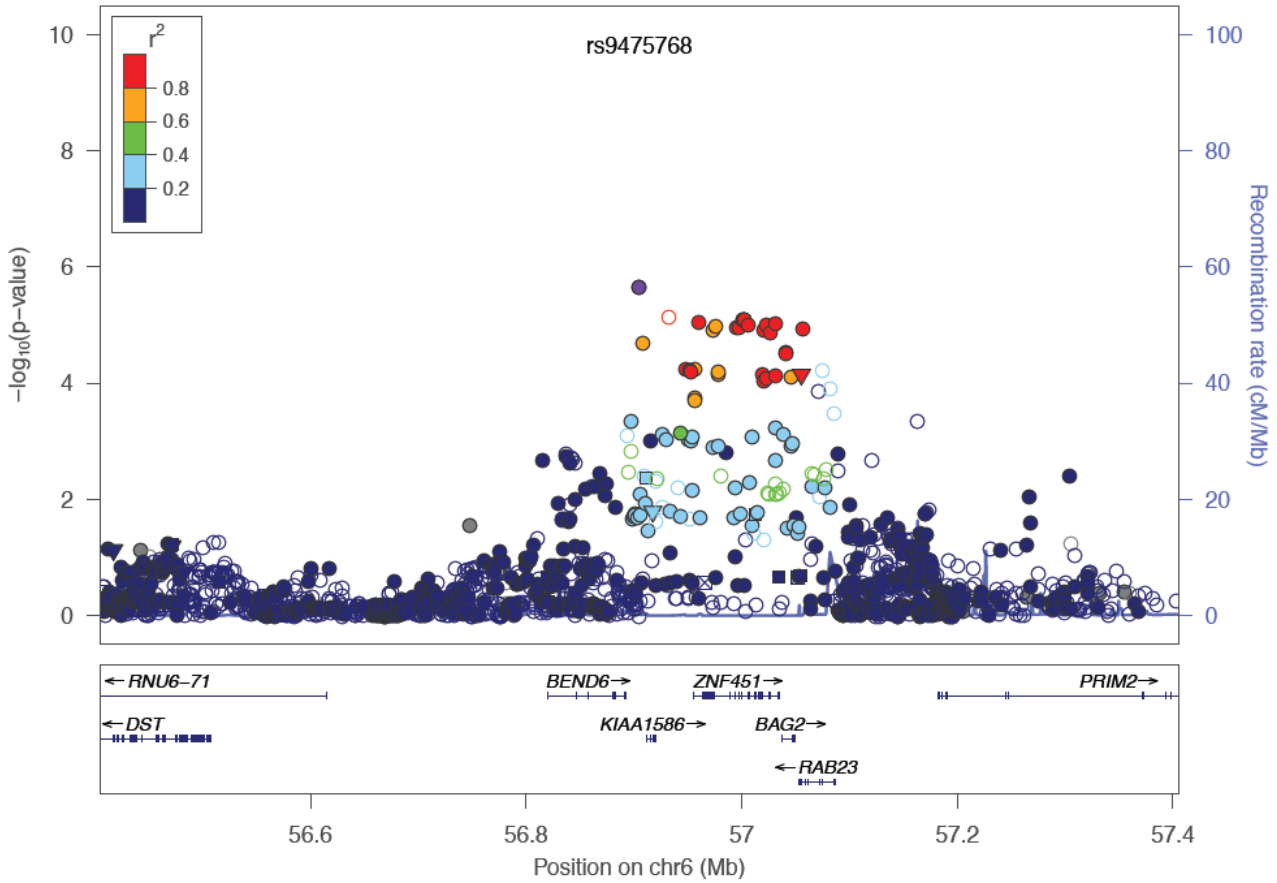
# Rare Variant Testing and Burden Testing as techniques in WGS analysis

# Cohort

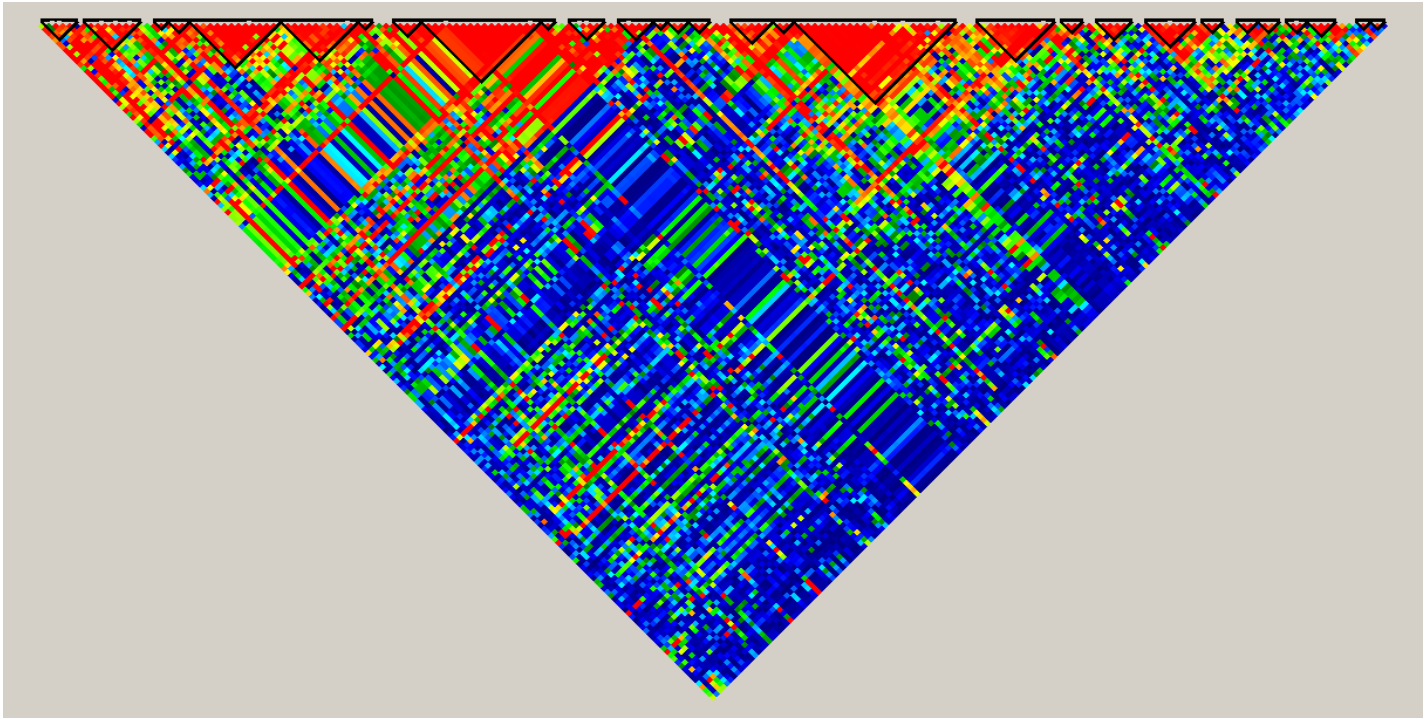
- 62 Whole Genomes of persons with SLE ESRD
- 62 Whole Genomes of persons with RA
- 176 healthy controls
- CGI-sequenced
- Extreme phenotype design
- Caucasian and African American

# GWAS – associated locus

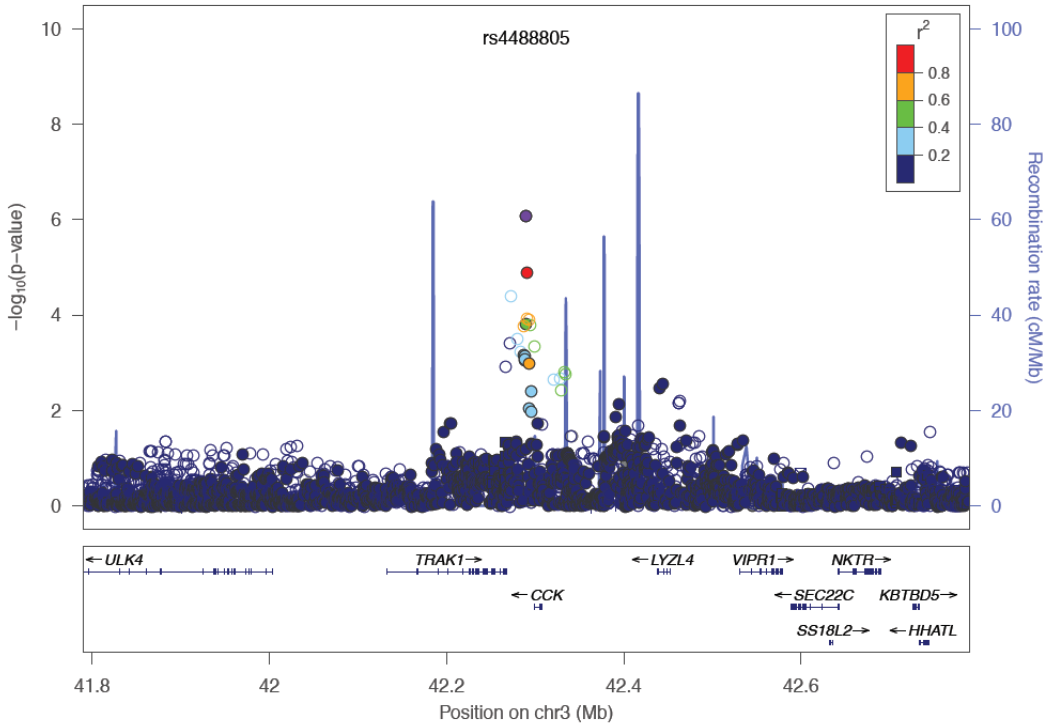



chr6:56,400,000-57,400,000
1,000,001 bp.
enter position, gene symbol or search terms
go
Scale
hg19
Chromosome Bands Localized by FISH Mapping Clones
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
RefSeq Genes
Pseudogene Annotation Set from GENCODE Version 19
Publications: Sequences in Scientific Articles
ClinVar Variants
NHGRI Catalog of Published Genome-Wide Association Studies
OMIM Allelic Variant SNPs
Human mRNAs from GenBank
Human ESTs That Have Been Spliced
Non-Human mRNAs from GenBank
Poly(A)-sequencing from Merck Research Laboratories
Alternative Splicing Graph from Swiss Institute of Bioinformatics
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Digital DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
100 Vertebrates Basewise Conservation by PhyloP
Multiz Alignments of 100 Vertebrates
Simple Nucleotide Polymorphisms (dbSNP 138) Found in >= 1% of Samples

# Heatmap



# Can we use next generation sequencing to support GWAS findings?



# Can look at…

- CNV / SV / MEI
  - Challenge – low sensitivity and specificity in short read sequencing
    - Why?
- Rare Variants (not tested by GWAS)
  - Challenge – very hard to study?
    - Why?
- How to overcome these challenges

# One way - rare variant burden testing

- 3 ontologies
  - GWAS associated genes (102)
  - Genes associated with joint erosion
  - Randomly generated loci
- 6 Tests
  - Simple Burden, C-alpha, Frequency Weighted, Variable Threshhold, Unique Alleles
- NOT Skat/Skat-O (limitations section)

# Rationale for Burden Testing

- Alpha level of 0.05, corrected by number of bp in the genome= $1.6*10^{-11}$.
- Not going to get that with 62 cases, even if you find "the variant."
- What do you do?
- Region tested can be anything
  - Domain
  - Gene
  - Locus

- Problems with burden testing
  - Linkage
  - Gets you no closer

# Alternatives to Burden Testing

- Rely on functional annotation
- Example

# Example of life after statistics

| Chr:Pos | Ref/Alt | Gene Region | Gene Symbol | Protein Varia | Translation Ir | SIFT Score | PhyloP | ENCODE TFBS | P | OR | dbSNP ID |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 20:43280349 | C/T | 5'UTR | ADA | | | | | TAF1; HEY1; POLR2A | 0.005 | 0.144 | 36216718 |
| 6:46620310 | -/C | Promoter; Ex | SLC25A27; CYP39A1 | p.I4fs*25 | frameshift | | | TRIM28; Pdx1; POLR2A; Nobox; HC | 0.006 | 3.192 | |
| 1:161518333 | A/C | Exonic | FCGR3A | p.L65R; p.L1( | missense | 0.25 | | | 0.007 | 18.236 | 10127939 |
| 1:161495563 | C/A | Exonic | HSPA6 | p.A372D | missense | 0 | 1.77E-03 | | 0.009 | 5.219 | |
| 1:161474960 | C/T | Promoter | FCGR2A | | | | | Mafb; Arnt::Ahr; TBP; FOXC1 | 0.013 | 0.070 | |
| 1:161601872 | T/C | Promoter | FCGR3B | | | | | YY1; BRCA1 | 0.013 | 0.070 | 377982 |
| 1:161474292 | A/G | Promoter | FCGR2A | | | | | FOXO3; Hltf; FOXD1; HOXA5 | 0.015 | 3.417 | |
| 12:105152193 | A/T | 3'UTR | CHST11 | | | | | | 0.017 | 2.077 | |
| 12:105155243 | ACACAC/- | 3'UTR | CHST11 | | | | | | 0.022 | 0.268 | 72056652 |
| 12:105136592 | C/G | Intronic | CHST11 | | | | 7.21E-03 | | 0.031 | 0.199 | |
| 6:46563812 | A/T | Exonic | CYP39A1 | L326H | missense | 0.04 | | | 0.033 | 5.375 | |
| 6:46563813 | G/C | Exonic | CYP39A1 | L326V | missense | 1 | | | 0.033 | 5.375 | |
| 6:46563815 | A/G | Exonic | CYP39A1 | L325P | missense | 0 | 1.98E-04 | | 0.081 | 3.542 | |

# Types of Burden Tests

- So-called collapsing tests

# Types of rare variant association tests

I. Burden Tests
II. Adaptive Burden Tests
III. Variance-Component Tests
IV. Combined tests
V. EC test

# Outline

I. Burden Tests
II. Adaptive Burden Tests
III. Variance-Component Tests
IV. Combined tests
V. EC test

# Burden Tests

- Collapse many variants into single risk score
- Several approaches
  - Combine minor allele counts into a single risk score (dominant genetic model)
  - Cohort Allelic Sums Test (CAST)
  - Combined Multivariate and Collapsing (CMC)
- Weighting
  - Variant type
  - Variant Rarity

# Burden Tests: Assumptions & Caveats

- Assume all rare variants in a set are causal and associated with a trait in the same direction.
- If this is untrue, power is lost.
- Some autoimmune GWAS loci already known to be associated with different diseases in different directions of effect

# Outline

I. Burden Tests
II. Adaptive Burden Tests
III. Variance-Component Tests
IV. Combined tests
V. EC test

# Adaptive Burden Tests

- Are still burden tests, but allow for +, -, and neutral variants.
  - W = -1 if unlikely to be associated
  - W = 1 if likely to be associated

# Types of Adaptive Burden Tests

- Data Adaptive Sum method (aSum)
  - Estimates direction of effect with w= -1, 0, or 1
  - Permutations to estimate p-values
  - Could use LD with Lead SNP in GWAS study to
- Estimated Regression Coefficient (EREC)
  - Uses Regression Coefficient ß of each variant as its weight (ß should be an unbiased estimator)
  - But MAF and our cohort size are both small..
- Variable threshold method (VT)
  - KBAC – kernel based adaptive weighting of risk / non-risk classification and association testing in an adaptive fashion.

# Adaptive Burden Tests, Summary

- Generally regarded as better-powered because they require fewer assumptions (e.g. all variants harmful).

# Outline

I. Burden Tests
II. Adaptive Burden Tests
III. Variance-Component Tests
IV. Combined tests
V. EC test

# Variance-Component Tests

- Test for association by evaluating the **distribution of test statistics.**

- **C-alpha**
  - 1st generation tests
- SKAT
  - better than C-alpha because it can accommodate covariates and SNP-SNP interactions.
  - SKAT = (weight)^2 * (ß)^2  (for each variant)

- These tests are robust to + or – effects due to squared term and kernel effects.

# Variance-Component Tests: notes and caveats

- Not stable for small cohorts having different numbers of cases / controls.
- SKAT can include covariates
- Tend to outperform Burden tests provided that many variants in the region are non-causal.

# Outline

I. Burden Tests
II. Adaptive Burden Tests
III. Variance-Component Tests
IV. Combined tests
V. EC test

# Combined Tests

- Burden > variance if many variants are causal
- Variance > burden if many variants non-causal
- Therefore, a test that combines both in different scenarios is useful.
  - Better powered than burden or variance if the truth is somewhere in between
  - Less powered if the assumptions of one or the other test are more or less accurate
- SKAT-O is such a test.
  - $Q = (1-p)Q_{SKAT} + pQ_{BURDEN}$

# Outline

I. Burden Tests
II. Adaptive Burden Tests
III. Variance-Component Tests
IV. Combined tests
V. EC test

# EC test

- Both burden and variance tests are based on linear & quadratic sums of $S_j$.
- EC test uses an exponential term sum of $S_j^2$, so the Q term rises rapidly in the presence of a causal variant.
- Better powered $n_{\text{causal variants}}$ is small.
- Worse powered if $n_{\text{causal variants}}$ is large.

# Comparison

- Best powered test completely depends on the kind of causal variant.
  - Loci with many rare causal variants much more likely to do well on a gene burden test.
  - Loci with few moderately rare causal variants better powered to be identified via single variant test or EC test.
- Pertinent question seems to be how to choose test on a per locus basis
- For a pathway-wide analysis, I would use SKAT-O combined test for aforementioned reason.

# Back to the data

# Procedure

- Generated genomic ranges based on the VCF files, +50kB on either side for each gene
- Generated 500 Random Loci matched to the loci of interest in terms of size and genic content
  - No overlap with loci of interest was permitted
  - Use is in understanding the tests, not for association testing

# Procedure, cont'd

- Then loaded each locus into Pseq
- Then excluded based on I=value

# Procedure, Cont'd

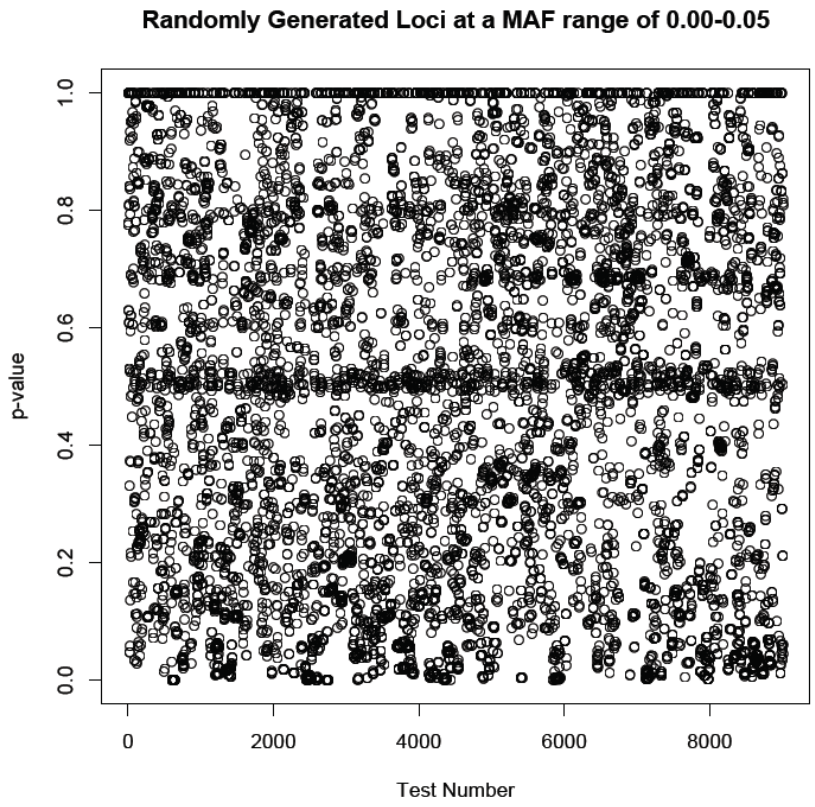
- Two sets of burden tests
  - All variants
  - Only non-synonymous coding variants
- Then conducted gene-level burden tests for the ontologies as well as the randomly generated loci

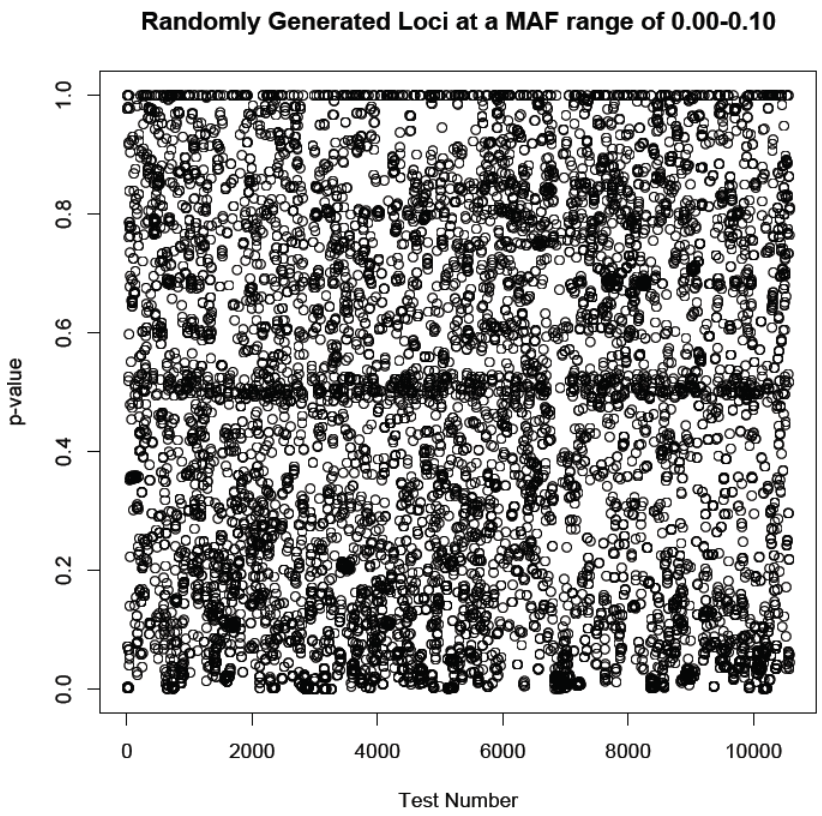
# Visualization

- Purpose of these visualizations is to understand what we should do
- Used the Random Data
- By MAF
- By Test

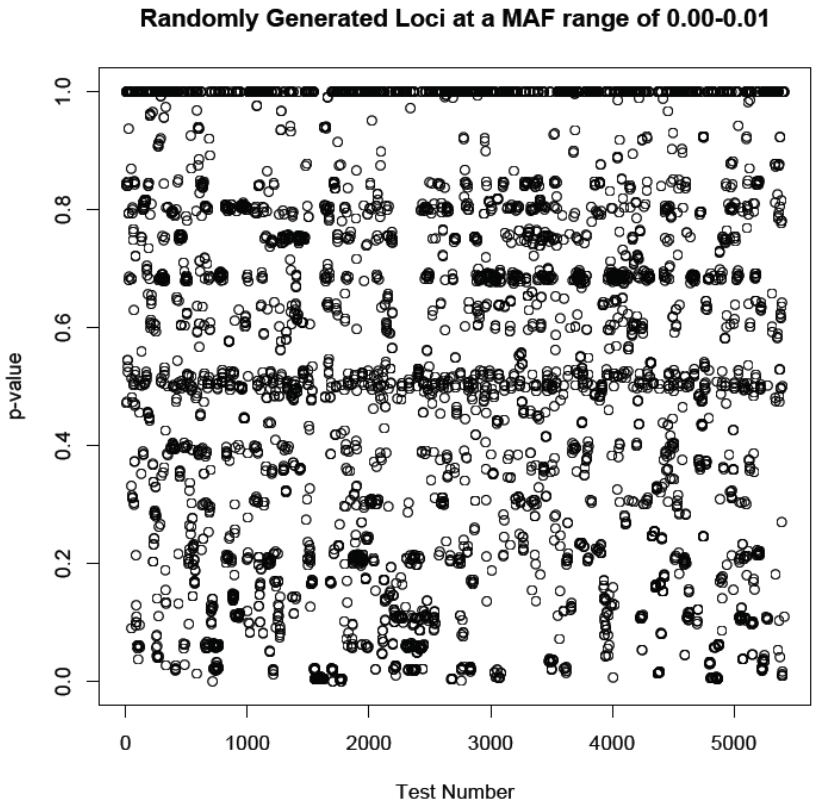
# By MAF

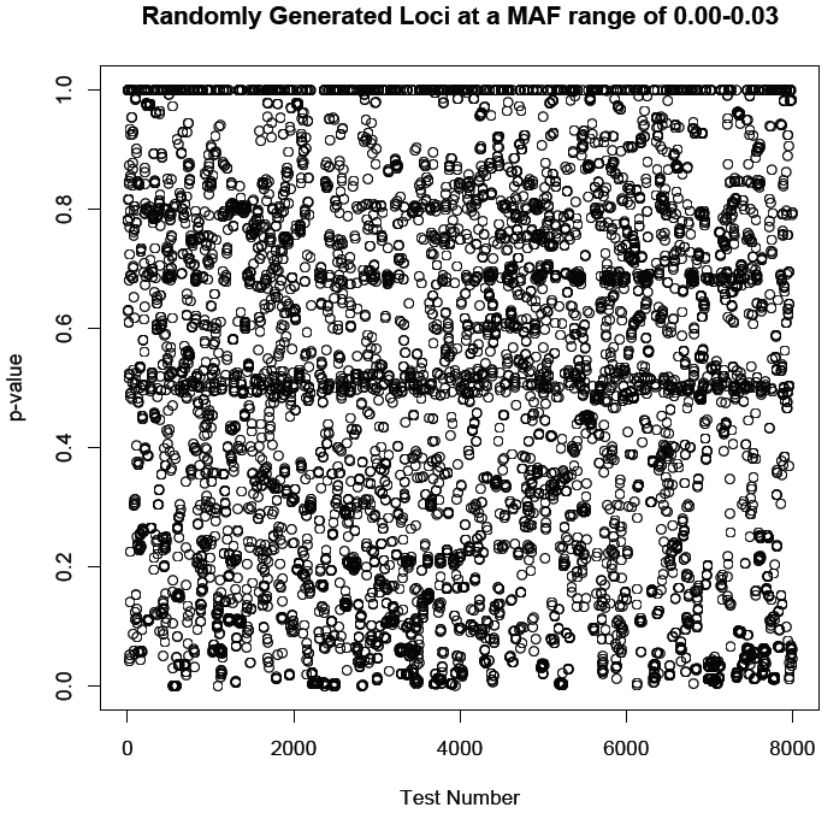
# Randomly Generated Loci, by MAF level



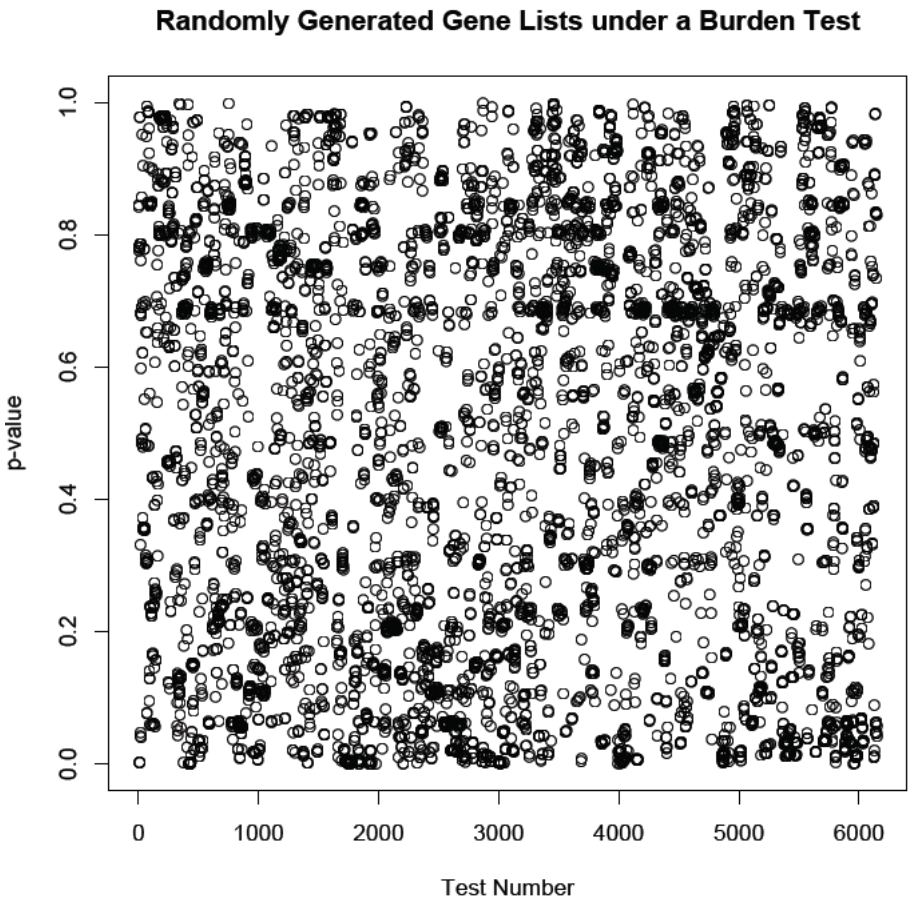
# Randomly Generated Loci, by MAF level



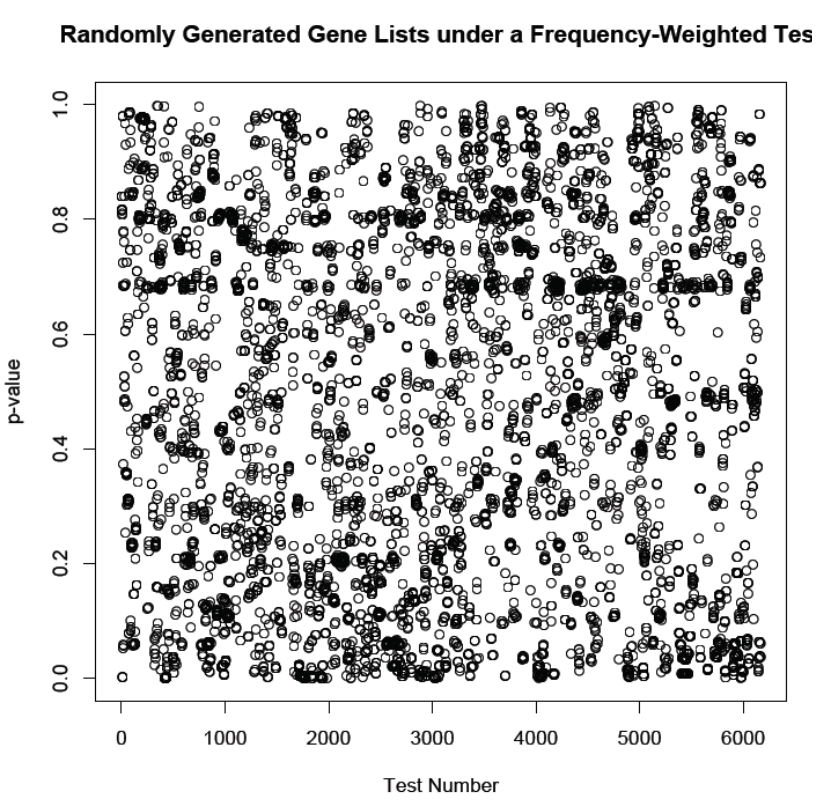



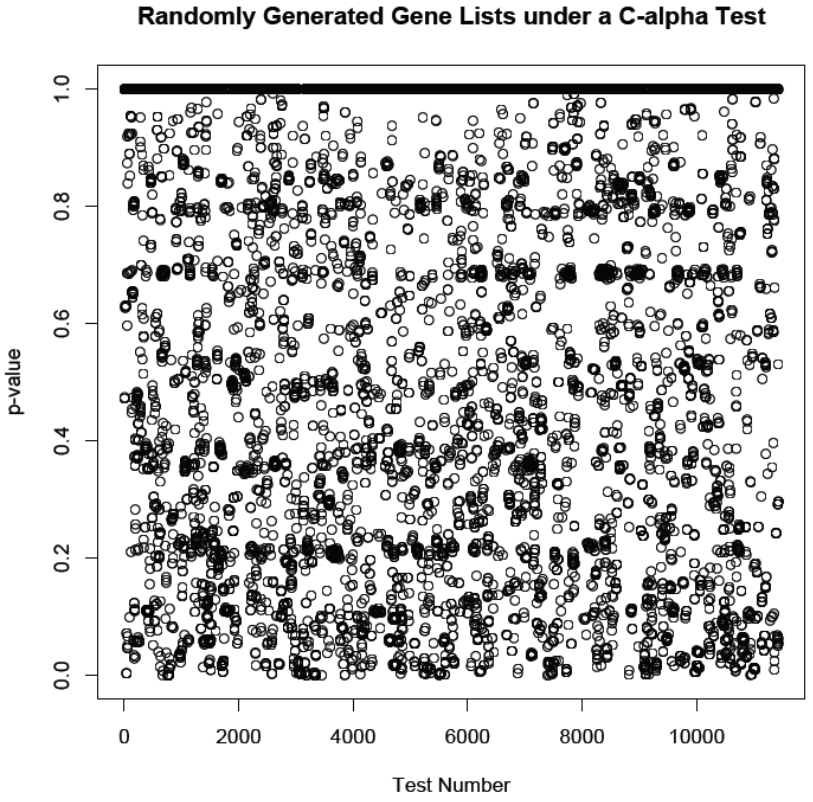
# By Test

# Randomly Generated Loci, by Test type



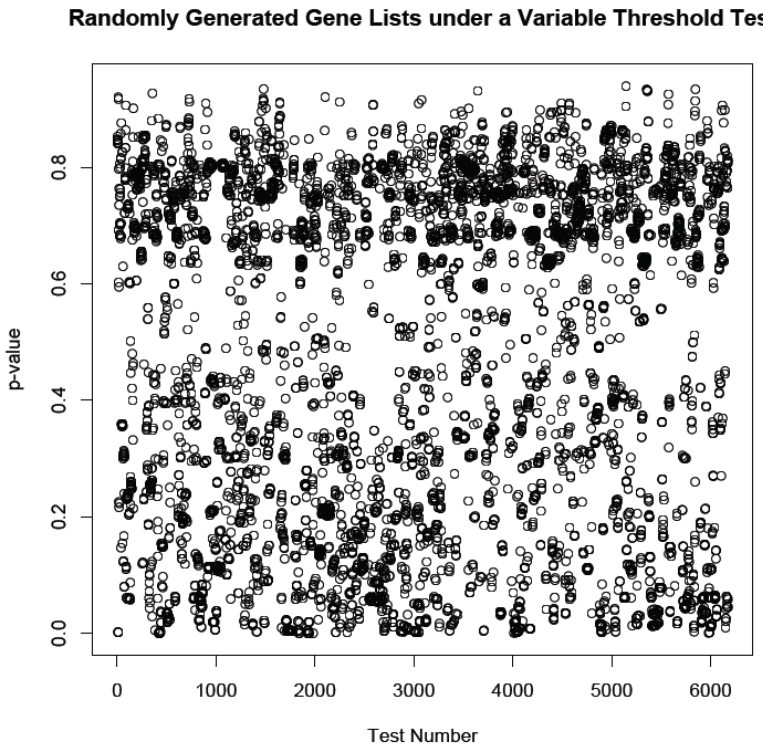
# By Test

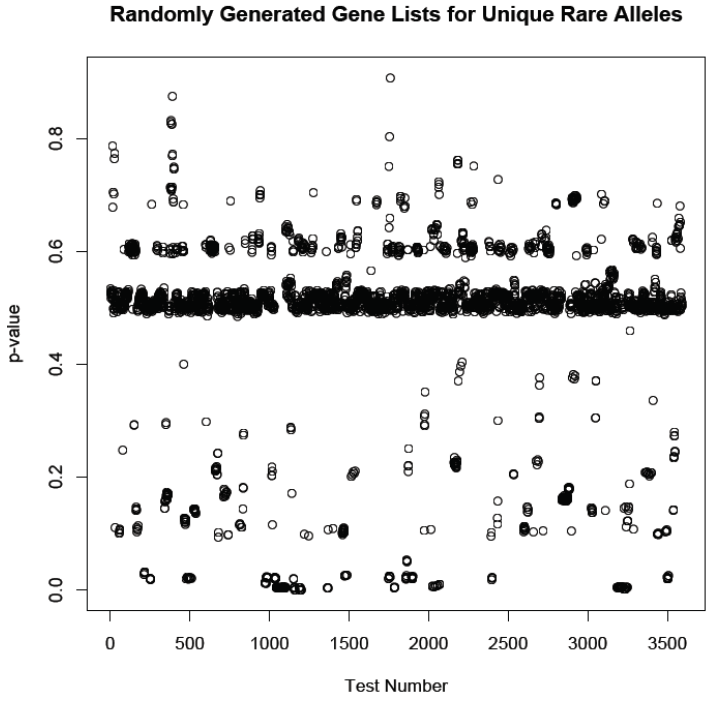

Randomly Generated Gene Lists under a Frequency-Weighted Tes
p-value
0.0
0.2
0.4
0.6
0.8
1.0
0
1000
2000
3000
4000
5000
6000
Test Number


Randomly Generated Gene Lists under a C-alpha Test
p-value
0.0
0.2
0.4
0.6
0.8
1.0
0
2000
4000
6000
8000
10000
Test Number

# Randomly Generated Loci, by Test type

?



# Selecting Tests for future

- Can continue to generate all tests all MAF levels. Actually, this is already done.
- However at some point might want to settle on a Test and an MAF.
- This brings me to next point

# Limitations

- In general, except for the Rare Allele test, most of these tests were very similarly powered for our data
- So, doing SKAT / SKAT-O might be superfluous
- IF NOT for the fact that they can accept covariates

# Covariates

- Have generated them for Sequencing Pipeline, Sex, top 8 PCs.
- Can employ once we get SKAT-O working
- If it were not for this, it would definitely not be worth using SKAT-O, but in this case I think it is probably worth it.

# Summary of Recommendations

- I would select:
- Test
  - Rare Allele Test
  - C-alpha Test
  - SKAT-O
- MAF
  - 0.03 (as per RefSeq dbase)

# Brief aside before data presentation

- I also ran a variant-level association test according to a logistic model with these 10 PCs.
- Not surprisingly, the SVD did not converge due to N=300 (62; 238)

# Results

- The following are genes that passed the burden test, which burden test it was, and the association p-value *by ontology*

# SLE genes RPK Curated

## All variants

- **FCGR2A-FCGR3A-FCGR3B**

maf05 **HSPA6** BURDEN 0.00379962 0.0003
maf10 **HSPA6** BURDEN 0.00019998 0.0001
maf05 **HSPA6** CALPHA 0.00429957 0.0005
maf10 **HSPA6** CALPHA 0.00139986 0.0001
maf05 **HSPA6** FRQWGT 0.00359964 0.0003
maf10 **HSPA6** FRQWGT 9.999e-05 0.0001
maf10 **HSPA6** SUMSTAT 0.00129987 0.0001
maf05 **HSPA6** VT 0.00409959 0.0004
maf10 **HSPA6** VT 0.00059994 0.0002

- **NEGR1**

maf01 NEGR1 BURDEN 0.00329967 0.0001
maf03 NEGR1 BURDEN 0.00489951 0.0001
maf02 NEGR1 BURDEN 0.00359964 0.0002
maf02 NEGR1 FRQWGT 0.00379962 0.0001
maf01 NEGR1 SUMSTAT 0.00169983 0.0001
maf03 NEGR1 SUMSTAT 0.00359964 0.0001
maf05 NEGR1 SUMSTAT 0.00359964 0.0001
maf10 NEGR1 SUMSTAT 0.00349965 0.0001
maf02 NEGR1 VT 0.00479952 0.0001
maf03 NEGR1 VT 0.00329967 0.0001
maf03 NEGR1 VT 0.00329967 0.0001

## Non-synonymous

**FCGR2A-FCGR3A-FCGR3B Locus, but**

- maf05 **HSPA6** BURDEN 0.00369963 0.0001
- maf10 **HSPA6** BURDEN 0.00419958 0.0002
- maf05 **HSPA6** CALPHA 0.0029997 0.0004
- maf10 **HSPA6** CALPHA 0.00319968 0.0003
- maf05 **HSPA6** FRQWGT 0.00389961 0.0002
- maf10 **HSPA6** FRQWGT 0.00369963 0.0003
- maf05 **HSPA6** VT 0.00439956 0.0003

# CKD GWAS

## All variants

- KCNQ1 maf01 KCNQ1 UNIQ 0.00119988 0.0002
- KCNQ1 maf01 KCNQ1 UNIQ 0.00119988 0.0002
- KCNQ1 maf02 KCNQ1 UNIQ 0.00079992 0.0001
- KCNQ1 maf02 KCNQ1 UNIQ 0.00089991 0.0001
- KCNQ1 maf03 KCNQ1 UNIQ 0.00119988 0.0001
- KCNQ1 maf03 KCNQ1 UNIQ 0.00139986 0.0001
- KCNQ1 maf05 KCNQ1 UNIQ 0.00129987 0.0001
- KCNQ1 maf05 KCNQ1 UNIQ 0.00129987 0.0001
- KCNQ1 maf10 KCNQ1 UNIQ 0.00059994 0.0001
- KCNQ1 maf10 KCNQ1 UNIQ 0.00049995 0.0001
- PVT1 maf01 PVT1 SUMSTAT 0.0009999 0.0012
- PVT1 maf02 PVT1 SUMSTAT 0.0009999 0.0024
- PVT1 maf03 PVT1 SUMSTAT 0.0009999 0.0012
- PVT1 maf05 PVT1 SUMSTAT 0.00119988 0.0019
- SLC7A9 maf10 CEP89 BURDEN 0.00409959 0.0003
- SLC7A9 maf10 CEP89 CALPHA 0.00479952 0.0001
- SLC7A9 maf10 CEP89 FRQWGT 0.00379962 0.0001
- SLC7A9 maf10 CEP89 SUMSTAT 0.00459954 0.0001
- SLC7A9 maf10 CEP89 VT 0.00389961 0.0001
- TSTD1 maf03 PVRL4 BURDEN 0.00389961 0.0001
- TSTD1 maf05 F11R BURDEN 0.00319968 0.0001
- TSTD1 maf10 ARHGAP30 BURDEN 0.0029997 0.0001
- TSTD1 maf10 ARHGAP30 BURDEN 0.00329967 0.0001
- TSTD1 maf03 PVRL4 CALPHA 0.00249975 0.0005
- TSTD1 maf05 F11R CALPHA 0.00289971 0.0001
- TSTD1 maf05 F11R FRQWGT 0.00479952 0.0001
- TSTD1 maf10 ARHGAP30 FRQWGT 0.00149985 0.0001
- TSTD1 maf10 ARHGAP30 FRQWGT 0.00149985 0.0001
- TSTD1 maf03 PVRL4 SUMSTAT 0.0049995 0.0002
- TSTD1 maf05 F11R SUMSTAT 0.00369963 0.0001
- TSTD1 maf10 ARHGAP30 SUMSTAT 0.00449955 0.0001
- TSTD1 maf10 ARHGAP30 SUMSTAT 0.00449955 0.0001
- TSTD1 maf03 PVRL4 VT 0.00389961 0.0001
- TSTD1 maf05 F11R VT 0.00229977 0.0001
- TSTD1 maf10 ARHGAP30 VT 0.00359964 0.0001
- TSTD1 maf10 ARHGAP30 VT 0.00369963 0.0001

## Non-synonymous

- SLC7A9 maf10 CEP89 BURDEN 0.00309969 0.0005
- SLC7A9 maf10 CEP89 CALPHA 0.00369963 0.0003
- SLC7A9 maf10 CEP89 FRQWGT 0.00279972 0.0002
- SLC7A9 maf10 CEP89 SUMSTAT 0.00309969 0.0001
- SLC7A9 maf10 CEP89 VT 0.00259974 0.0002

# Next?

- Implement SKAT-O?
- Conduct test by genic region?
- Do pathway, rather than gene-level
- Other?