Association Analysis

Dr. Chris Carlson
FHCRC

NIEHS
January 31, 2006

Analyzing SNP Data

- Study Design
- SNPs vs Haplotypes
- Regression Analysis
- Population Structure
- Multiple Testing
- Whole Genome Analysis

Analyzing SNP Data

- Study Design
- SNPs vs Haplotypes
- Regression Analysis
- Population Structure
- Multiple Testing
- Whole Genome Analysis

Study Design

- Heritability
- Prior hypotheses
- Target phenotype(s)
- Power
- Ethnicity
- Replication

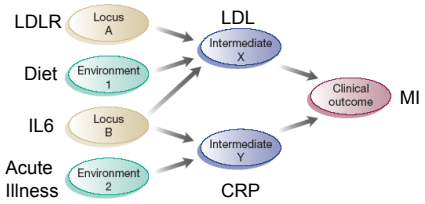
Heritability

- Is your favorite phenotype genetic?
- Heritability (h^2) is the proportion of variance attributed to genetic factors
  - h^2 ~ 100%: ABO Blood type, CF
  - h^2 > 80%: Height, BMI, Autism
  - h^2 50-80%: Smoking, Hypertension, Lipids
  - h^2 20- 50%: Marriage, Suicide, Religiousness
  - h^2 ~ 0: ??

Prior Hypotheses

- There will always be too much data
- There will (almost) always be priors
  - Favored SNPs
  - Favored Genes
- Make sure you've stated your priors (if any) explicitly BEFORE you look at the data

Target Phenotypes



Carlson et al., Nature v. 429 p. 446

Statistical Power

- Null hypothesis: all alleles are equal risk
- Given that a risk allele exists, how likely is a study to reject the null?
- Are you ready to genotype?

Genetic Relative Risk

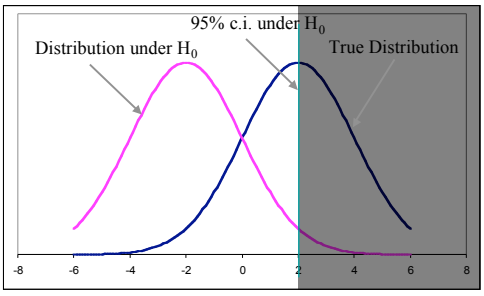
		Disease	Unaffected
SNP	Allele 1	p_{1D}	p_{1U}
	Allele 2	p_{2D}	p_{2U}

$$RR = \frac{p(Disease \mid Allele1)}{p(Disease \mid Allele2)} = \frac{\frac{p_{1D}}{p_{1D} + p_{1U}}}{\frac{p_{2D}}{p_{2D} + p_{2U}}}$$

Power Analysis

- Statistical significance
  - Significance = p(false positive)
  - Traditional threshold 5%
- Statistical power
  - Power = 1- p(false negative)
  - Traditional threshold 80%
- Traditional thresholds balance confidence in results against reasonable sample size

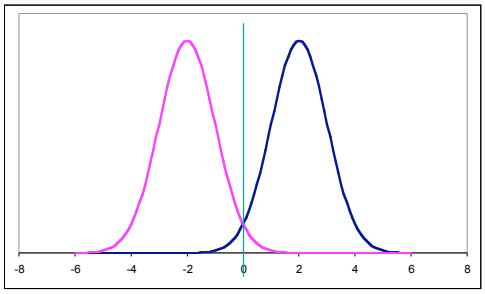
Small sample: 50% Power



Maximizing Power

- Effect size
  - Larger relative risk = greater difference between means
- Sample size
  - Larger sample = smaller SEM
- Measurement error
  - Less error = smaller SEM

Large sample: 97.5% Power

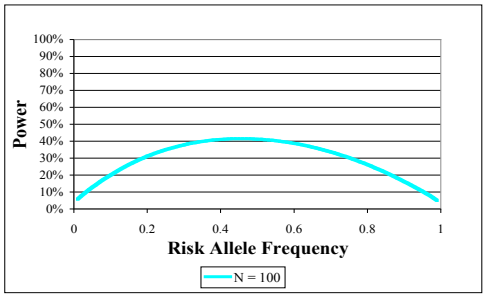


Risk Allele Example
10% Population Frequency

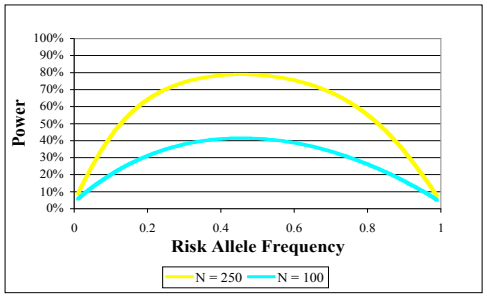
- Homozygous Relative Risk = 4
- Multiplicative Risk Model
  - Het RR = 2
- Case Freq
  - 18.2%
- Control Freq
  - 9.9%

- Homozygous Relative Risk = 2
- Multiplicative Risk Model
  - Het RR = 1.4
- Case Freq
  - 13.6%
- Control Freq
  - 9.96%

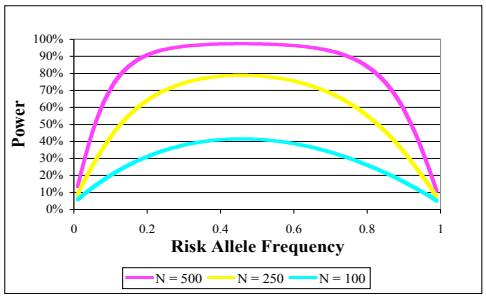
Power to Detect RR=2
N Cases, N Controls



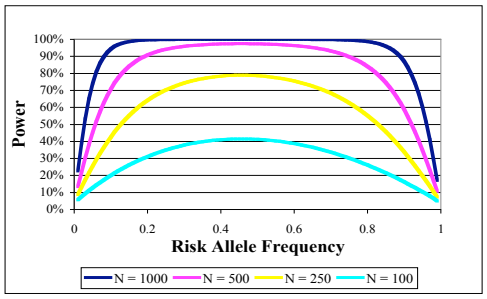
Power to Detect RR=2
N Cases, N Controls



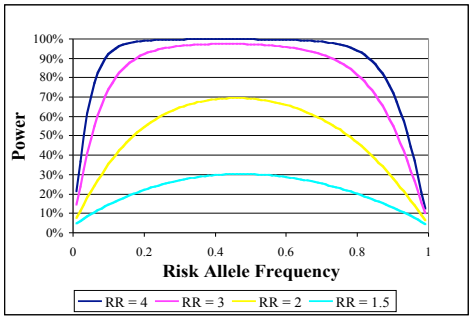
Power to Detect RR=2
N Cases, N Controls



Power to Detect RR=2
N Cases, N Controls



Power to Detect SNP Risk 200 Cases, 200 Controls



Power Analysis Summary

- For common disease, relative risk of common alleles is probably less than 4
- Maximize number of samples for maximal power
- For RR < 4, measurement error of more than 1% can significantly decrease power, even in large samples

SNP Selection for Association Studies

Direct:
Catalog and test all functional variants for association

Indirect:
Use dense SNP map and select based on LD

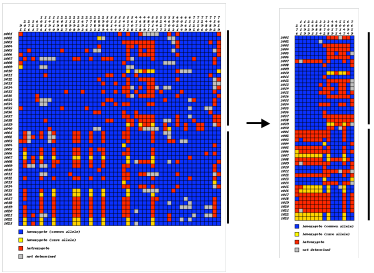
Collins, Guyer, Chakravarti (1997). Science 278:1580-81

Parameters for SNP Selection

- Allele Frequency
- Putative Function (cSNPs)
- Genomic Context (Unique vs. Repeat)
- Patterns of Linkage Disequilibrium

Focus on Common Variants - Haplotype Patterns

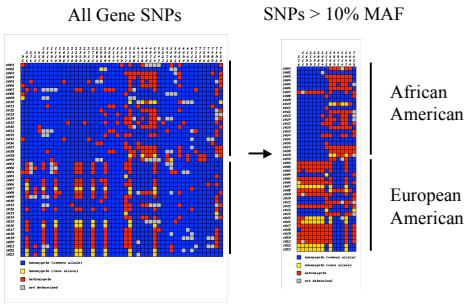
All Gene SNPs SNPs > 10% MAF



Why Common Variants?

- Rare alleles with large effect (RR > 4) should already be identified from linkage studies
- Association studies have low power to detect rare alleles with small effect (RR < 4)
- Rare alleles with small effect are not important, unless there are a lot of them
- Theory suggests that it is unlikely that many rare alleles with small effect exist (Reich and Lander 2001).

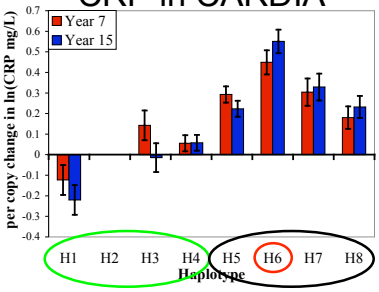
Ethnicity



Replication

- You WILL be asked to replicate
- Statistical replication
  - Split your sample
  - Arrange for replication in another study
  - Multiple measurements in same study
- Functional replication

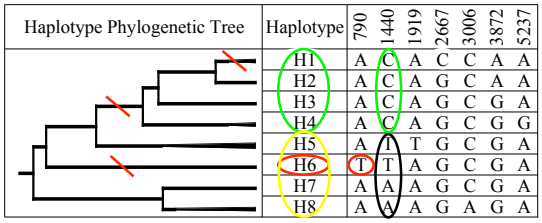
Multiple Measurements: CRP in CARDIA



Carlson et al, AJHG v77 p64

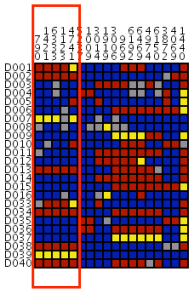
Haplo.glm: Lake et al, Hum Hered v. 55 p. 56

Haplotypes vs tagSNPs



Haplotype Phylogenetic Tree	Haplotype	790	1440	1919	2667	3006	3872	5237
	H1	A	C	A	C	C	A	A
	H2	A	C	A	G	C	A	A
	H3	A	C	A	G	C	G	A
	H4	A	C	A	G	C	G	G
	H5	A	T	T	G	C	G	A
	H6	T	T	A	G	C	G	A
	H7	A	A	A	G	C	G	A
	H8	A	A	A	G	A	G	A

High CRP Haplotype

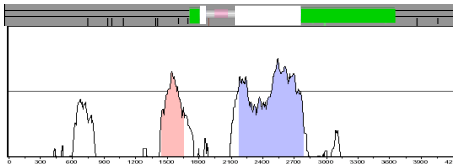


- 5 SNPs specific to high CRP haplotype

Functional Replication

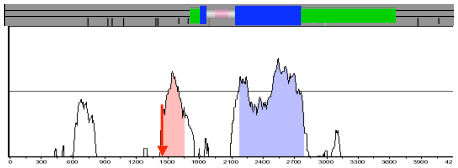
- Statistical replication is not always possible
- Association may imply mechanism
- Test for mechanism at the bench
  - Is predicted effect in the right direction?
  - Dissect haplotype effects to define functional SNPs

CRP Evolutionary Conservation



- TATA box: 1697
- Transcript start: 1741
- CRP Promoter region (bp 1444-1650) >75% conserved in mouse

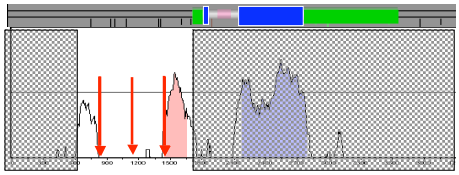
Low CRP Associated with H1-4



- USF1 (Upstream Stimulating Factor)
  - Polymorphism at 1440 alters USF1 binding site

```
          1420       1430       1440
H1-4 gcagctacCACGTGcacccagatggcCACTCGtt
H7-8 gcagctacCACGTGcacccagatggcCACTAGtt
H5-6 gcagctacCACGTGcacccagatggcCACTTGtt
```

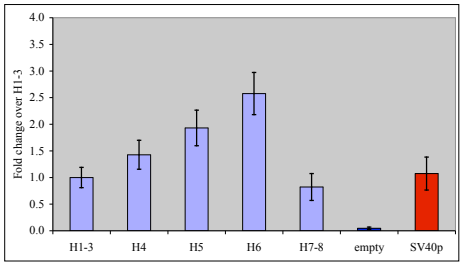
High CRP Associated with H6



- USF1 (Upstream Stimulating Factor)
  - Polymorphism at 1421 alters another USF1 binding site

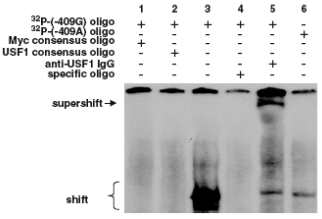
```
          1420       1430       1440
H1-4 gcagctacCACGTGcacccagatggcCACTCGtt
H7-8 gcagctacCACGTGcacccagatggcCACTAGtt
H5   gcagctacCACGTGcacccagatggcCACTTGtt
H6   gcagctacCACATGcacccagatggcCACTTGtt
```

CRP Promoter Luciferase Assay



Carlson et al, AJHG v77 p64

CRP Gel Shift Assay



Szalai et al, J Mol Med v83 p440

Study Design Summary

- State your priors
- Know your phenotypes
- Estimate your power
- Pay attention to ethnicity
- Set up replication ASAP
- Replication can be functional

Data Analysis

- Study Design
- SNPs vs Haplotypes
- Regression Analysis
- Population Structure
- Multiple Testing
- Whole Genome Analysis

SNPs or Haplotypes

- There is no right answer: explore both
- The only thing that matters is the correlation between the assayed variable and the causal variable
- Sometimes the best assayed variable is a SNP, sometimes a haplotype

Example: APOE

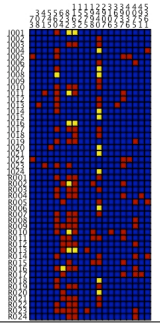
Table 1
ApoE genotype frequency in US population and AD risk

Genotype	Population[a] (%)	AD[b] (%)	#Population	#AD	Risk[c] (%)	If all US
ε2/2	1	0.1	0.5M	0.004M	0.08	0.4M
ε2/3	12	4	5.5M	0.18M	3.2	1.5M
ε3/3	60	35	27.6M	1.4M	5.1	2.3M
ε3/4	21	42	9.6M	1.7M	18	8.2M
ε4/4	2	16	0.9M	0.6M	67	30.7M

Please note that ε2/ε4 subjects are not included in table.
[a] Using estimate of 46 million in US over 60 y/o in 2000.
[b] Assuming 4 million individuals have AD.
[c] Data from [13,46,49].

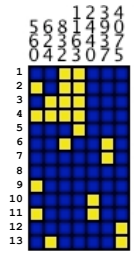
Raber et al, Neurobiology of Aging, v25 p641

Example: APOE



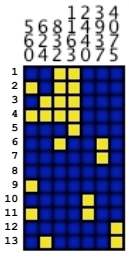
- Small gene (<6kb)
- 7 SNPs with MAF > 5%
- APOE ε2/ε3/ε4
  - Alzheimer's associated
  - ε2 = 4075
  - ε4 = 3937

Example: APOE



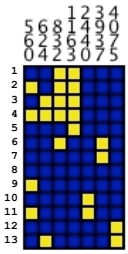
- Haplotype inferred with PHASE2
- 7 SNPs with MAF >5%
- APOE 2/3/4
  - E2 = 4075
  - E4 = 3937
  - E3 = ?

Example: APOE



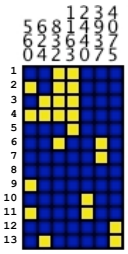
- 13 inferred haplotypes
- Only three meaningful categories of haplotype
- No single SNP is adequate

Example: APOE



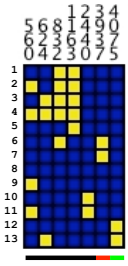
- SNP analysis:
  - 7 SNPs
  - 7 tests with 1 d.f.
- Haplotype analysis
  - 13 haplotypes
  - 1 test with 12 d.f.

Example: APOE



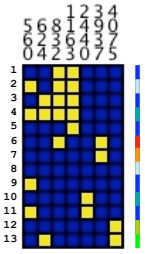
- Best marker is a haplotype of only the right two SNPs: 3937 and 4075

Building Up



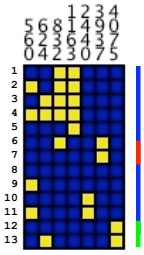
- Test each SNP for main effect
- Test SNPs with main effects for interactions

Paring Down



- Test all haplotypes for effects

Paring Down



- Test all haplotypes for effects
- Merge related haplotypes with similar effect

Data Analysis

- Study Design
- SNPs vs Haplotypes
- Regression Analysis
- Population Structure
- Multiple Testing
- Whole Genome Analysis

Exploring Candidate Genes: Regression Analysis

- Given
  - Height as "target" or "dependent" variable
  - Sex as "explanatory" or "independent" variable
- Fit regression model

  height = β*sex + ε

Regression Analysis

- Given
  - Quantitative "target" or "dependent" variable y
  - Quantitative or binary "explanatory" or "independent" variables x_i
- Fit regression model

  $y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i + \varepsilon$

Regression Analysis

- Works best for normal y and x
- Fit regression model

  $y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i + \varepsilon$
- Estimate errors on β's
- Use t-statistic to evaluate significance of β's
- Use F-statistic to evaluate model overall

Regression Analysis

```
Call:
lm(formula = data$TARGET ~ (data$CURR_AGE + data$CIGNOW +
    data$PACKYRS + data$SNP1 + data$SNP2 +
    data$SNP3 + data$SNP4))

Residuals:
     Min       1Q   Median       3Q      Max
-123.425  -25.794   -3.125   23.629  120.046

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     139.52703   13.80820  10.105  < 2e-16 ***
data$CURR_AGE    -0.04844    0.18492  -0.262  0.79345
data$CIGNOW     -10.11001    4.06797  -2.485  0.01327 *
data$PACKYRS      0.01573    0.05456   0.288  0.77320
data$SNP1         8.61749    3.31204   2.602  0.00955 **
data$SNP2       -19.71980    2.84816  -6.924 1.35e-11 ***
data$SNP3        -9.32590    2.96600  -3.144  0.00176 **
data$SNP4        -9.58801    3.05650  -3.137  0.00181 **
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 36.11 on 503 degrees of freedom
Multiple R-Squared: 0.2551, Adjusted R-squared: 0.2448
F-statistic: 24.61 on 7 and 503 DF,  p-value: < 2.2e-16
```

Coding Genotypes

Genotype	Dominant	Additive	Recessive
AA	1	2	1
AG	1	1	0
GG	0	0	0

- Genotype can be re-coded in any number of ways for regression analysis
- Additive ~ codominant

Fitting Models

- Given two models

  $y = \beta_1 x_1 + \varepsilon$

  $y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$
- Which model is better?
- More parameters will always yield a better fit
- Information Criteria
  - Measure of model fit penalized for the number of parameters in model
- AIC (most common)
  - Akaike's Info Criterion
- BIC (more stringent)
  - Bayesian Info Criterion

Tool References

- Haplo.stats (haplotype regression)
  - Lake et al, Hum Hered. 2003;55(1):56-65.
- PHASE (case/control haplotype)
  - Stephens et al, Am J Hum Genet. 2005 Mar;76(3):449-62
- Haplo.view (case/control SNP analysis)
  - Barrett et al, Bioinformatics. 2005 Jan 15;21(2):263-5.
- SNPHAP (haplotype regression?)
  - Sham et al Behav Genet. 2004 Mar;34(2):207-14.

Analyzing SNP Data

- Study Design
- SNPs vs Haplotypes
- Regression Analysis
- Population Structure
- Multiple Testing
- Whole Genome Analysis

Population Stratification

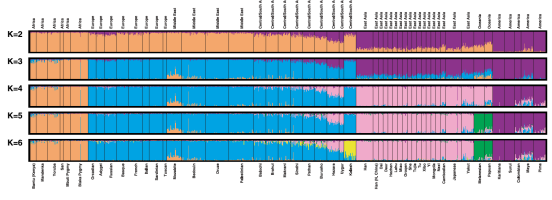
- Many diseases have different frequencies in ancestral groups
  - E.g. MS is more frequent in Europeans
- In admixed or stratified populations, markers correlated with ancestry may show spurious associations
  - E.g. Duffy and MS in African Americans

Population Stratification

- Admixture
  - Individuals with ancestry from multiple populations
  - E.g. Hispanic or African American
- Stratification
  - Subpopulations with distinct allele frequencies
  - E.g. Brazil, California
- STRUCTURE software
  - Pritchard et al, Genetics v155 p945

Genomic Controls

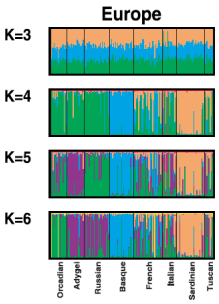
- Unlinked anonymous markers not chosen for known allele frequencies
- Allow unbiased estimation of population structure



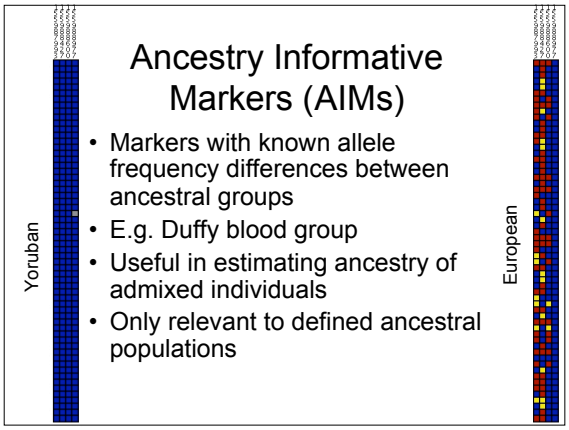
Rosenberg et al Science v298 p2381

Genomic Controls

- Warning: 377 microsatellites barely detects European structure
- Within continent resolution probably requires thousands of SNPs



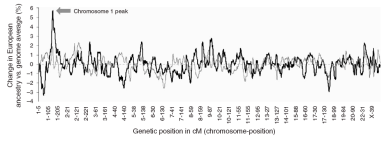
Ancestry Informative Markers (AIMs)



- Markers with known allele frequency differences between ancestral groups
- E.g. Duffy blood group
- Useful in estimating ancestry of admixed individuals
- Only relevant to defined ancestral populations

Admixture mapping

- Type several thousand AIMs
- Search for regions with excess allelic ancestry from a single population
- E.g. MS in AA: Reich et al, Nat Genet v37 p1113



Pop Structure Summary

- For known admixture, use AIMs to estimate ancestry
- For diseases with substantial differences in risk by ethnicity, use admixture mapping
- Detecting cryptic population structure requires hundreds to thousands of genomic controls

Analyzing SNP Data

- Study Design
- SNPs vs Haplotypes
- Regression Analysis
- Population Structure
- Multiple Testing
- Whole Genome Analysis

Multiple Testing

Study target	Technology	Samples	Studies
Gene 10 SNPs	TaqMan	100's	2
Pathway 1500 SNPs	Illumina SNPlex	1000's	2
Genome 500k SNPs	Affy Illumina	??	??

Multiple Testing

- Practical guidelines
  - Write down your priors
  - Bonferroni
  - FDR
  - Staged Study Design
  - Other approaches - Neural Nets

Bonferroni

- P-values of stats assume a single test
- For multiple tests, adjust significance by multiplying P-value by number of tests
  - Given 10 tests and unadjusted p = 0.02
  - p = 10 * 0.02 = 0.2
- Over conservative

Step-Down Bonferroni

- Given N SNPs to analyze
- Order SNPs using prior info
  - Evaluate the most interesting hypotheses first
- For first SNP, do not correct p-value
- For second SNP, adjust for 2 tests
- Etc.

Staged Study Design

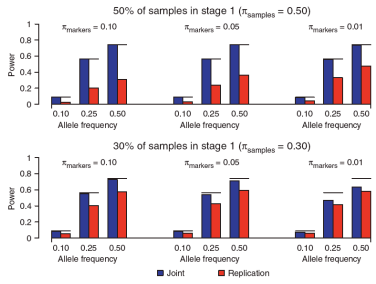
- Given 500,000 SNPs
- Bonferroni corrected significance threshold

  $p = 0.05 / 500000 = 10^{-7}$
- Significance in a single study is difficult to achieve

Staged Study Design

- Study I: Genotype 500k SNPs in 1000 cases/controls
  - Expect 5,000 false positives at $p < 0.01$
- Study II: Genotype best 5000 hits from stage I in additional 1000 cases/controls
  - Expect 50 false positives at $p < 0.01$
- Study 3: Genotype best 50 hits in a third set of 1000 cases/controls
  - Expect 0.5 false positives at $p < 0.01$

Joint Analysis



Skol et al, Nat Genet in press

Post-Hoc Analysis

- Significance
  - Probability of a single observation under H_0
- False Discovery Rate
  - Proportion of observed results inconsistent with H_0

FDR Example

- Assume 10 tests
- 5 with uncorrected p = 0.05
- No single significant result
- More than 5% below 5%
- At least one of the five is probably real, but we can't say which

Multiple Testing Summary

- Bonferroni can be useful, but overly conservative
- FDR can be more helpful
- Staged study designs don't improve power, but can be economically advantageous

Analyzing SNP Data

- Study Design
- SNPs vs Haplotypes
- Regression Analysis
- Population Structure
- Multiple Testing
- Whole Genome Analysis

SNP Selection

- cSNPs (~20-25k common genome wide)
- tagSNPs
  - 500k random ≈ 300k selected
  - Probably adequate in European
  - Possibly adequate in Asian
  - More needed for African (~750k)
  - Possibly adequate in South Asian, Hispanic

Case/Control WGAA

- Allele Counting
  - Assumes codominant risk model

	A1	A2
Case	p_{1+}	p_{2+}
Control	p_{1-}	p_{2-}

$$\chi^2 = N(p_{1+}p_{2-} - p_{1-}p_{2+})$$

Case/Control WGAA

- Allele Counting
  - Assumes codominant risk model

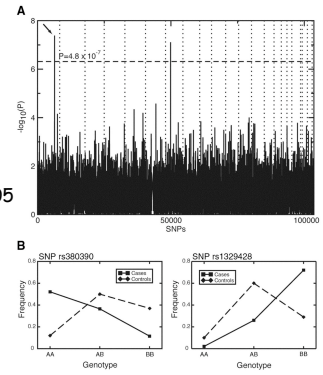
	A1	A2
Case	p_{1+}	p_{2+}
Control	p_{1-}	p_{2-}

$$\chi^2 = N(p_{1+}p_{2-} - p_{1-}p_{2+})$$

- Genotype Counting
  - Allows for dominance
  - Not important for rare SNPs

	11	12	22
Case	p_{11+}	p_{12+}	p_{22+}
Control	p_{11-}	p_{12-}	p_{22-}

Affymetrix's 100K Chip Analysis: Macular Degeneration Klein et al. Science 308: 385-389, 2005



P <0.05/ 103,611

4.8 X10^{-7}

Interaction Analysis

- SNP X SNP
- Within gene: haplotype
  - Modest interaction space
  - Most haplotype splits do not matter (APOE)
- Between genes: epistasis
  - Interaction space is vast (500k X 500k)
- SNP X Environment
  - Smaller interaction space (500k X a few environmental measures)

Limiting the Interaction Space

- Not all epistatic interactions make sense
  - Physical interactions (lock and key)
  - Physical interactions (subunit stoichiometry)
  - Pathway interactions
  - Regulatory interactions

Whole Genome Summary

- Low Hanging Fruit exist (e.g. AMD)
- Tier studies for economic purposes
  - Make sure N is large enough to be powered if all samples were 500k genotyped
- Interactions may be interesting
  - Explore sparingly for hypothesis testing
  - Explore comprehensively for hypothesis generation

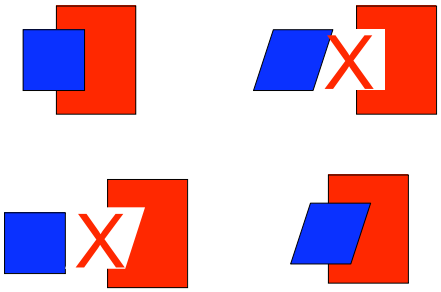
Conclusions

- Pay attention to study design
  - Sample size
  - Estimated power
  - Multiple Testing
- Analyze SNPs (and haplotypes)
- Keep population structure in mind
- Explore epistasis and environmental interactions after main effects

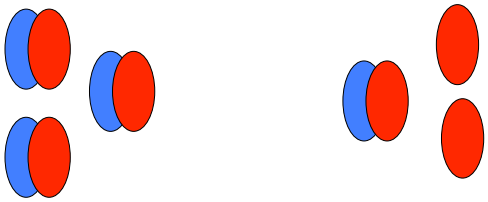
Limiting the Interaction Space

- Not all epistatic interactions make sense
  - Physical interactions (lock and key)
  - Physical interactions (subunit stoichiometry)
  - Pathway interactions
  - Regulatory interactions

Lock and Key



Stoichiometry



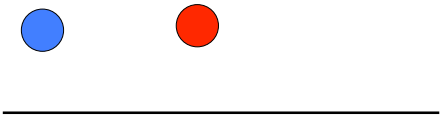
E.g. α and β globin in Thalassemia

Pathway

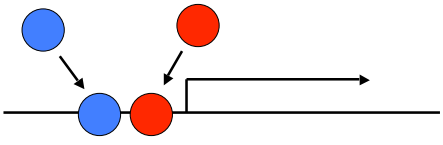
Pathway output can integrate across all steps within the pathway

BUT, many pathways have rate limiting step which can erase upstream variation

Regulatory



Regulatory



Tx factor X Tx factor (500 X 500)
Tx factor X gene (10 X 500k)

Epistasis: SNP X SNP Interactions

		AA	AC/CC
	OR		2
GG		1	2
GT/TT	2	2	4

Simple model: two dominant loci, two-fold relative risk (RR) to single carriers at either locus, four-fold risk to double carriers. Risk allele frequency 0.05 at both loci.

Epistasis I: Synergistic

		AA	AC/CC
	OR		2.533
GG		1	2
GT/TT	2.533	2	**10**

Simple model: two dominant loci, two-fold relative risk (RR) to single carriers at either locus, more than four-fold risk to double carriers. Risk allele frequency 0.05 at both loci.

Epistasis II: Permissive

		AA	AC/CC
	OR		1.878
GG		1	1
GT/TT	1.878	1	10

Simple model: two dominant loci, no risk (RR) to single carriers at either locus, more than four-fold risk to double carriers. Risk allele frequency 0.05 at both loci.

Epistasis III: Sufficient

		AA	AC/CC
	OR		1.822
GG		1	2
GT/TT	1.822	2	2

Simple model: two dominant loci, two-fold relative risk (RR) to single carriers at either locus, two-fold risk to double carriers. Risk allele frequency 0.05 at both loci.

Epistasis IV: Exclusive

		AA	AC/CC
	OR		1.733
GG		1	2
GT/TT	1.733	2	1

Simple model: two dominant loci, two-fold relative risk (RR) to single carriers at either locus, no risk to double carriers. Risk allele frequency 0.05 at both loci.

Rare Allele Epistasis

- Main effects are the observed effects analyzing one SNP at a time
- Main effects of rare alleles are not substantially affected by epistatic models
- Are common alleles more substantially affected by epistasis?

Common Allele, No Epistasis

		AA	AC/CC
	OR		2
GG		1	2
GT/TT	2	2	4

Simple model: two dominant loci, two-fold relative risk (RR) to single carriers at either locus, four-fold risk to double carriers. Risk allele frequency 0.3 at both loci (= risk genotype frequency 0.51 at either locus).

Epistasis I: Synergistic

		AA	AC/CC
	OR		4.026
GG		1	2
GT/TT	4.026	2	10

Simple model: two dominant loci, two-fold relative risk (RR) to single carriers at either locus, more than four-fold risk to double carriers. Risk allele frequency 0.3 at both loci.

Epistasis II: Permissive

		AA	AC/CC
	OR		5.59
GG		1	1
GT/TT	5.59	1	10

Simple model: two dominant loci, no risk (RR) to single carriers at either locus, more than four-fold risk to double carriers. Risk allele frequency 0.3 at both loci.

Epistasis III: Sufficient

		AA	AC/CC
	OR		1.325
GG		1	2
GT/TT	1.325	2	2

Simple model: two dominant loci, two-fold relative risk (RR) to single carriers at either locus, two-fold risk to double carriers. Risk allele frequency 0.3 at both loci.

Epistasis IV: Exclusive

		AA	AC/CC
	OR		0.987
GG		1	2
GT/TT	0.987	2	1

Simple model: two dominant loci, two-fold relative risk (RR) to single carriers at either locus, no risk to double carriers. Risk allele frequency 0.3 at both loci.

Main Effects Analysis

- In the vast majority of epistatic models, main effects exist, and point in the right direction
- Epistatic interaction is potentially more important for common alleles
- Limit epistatic exploration to common SNPs with main effects?