

Issues in Case-Control Association Studies to Detect Disease-Associated SNPs

Ruth Pfeiffer, Ph.D.

Biostatistics Branch
Division of Cancer Epidemiology & Genetics
National Cancer Institute

Strategies to identify disease producing mutations

- **Linkage analysis:** sharing of same allele within diseased family members
- **Association studies:** correlation between a genotype and disease status
 - **Typical design:** unrelated cases and population controls
 - **Candidate gene/pathway association**
 - **Genome wide association studies (GWAS)**

Issues in Association Studies

- Sources of bias:
 - Selection bias
 - Confounding
 - Differential miss-classification (genotyping error)
- Variance distortion: cryptic relatedness
- Marker versus true functional SNP
- Multiple Comparisons
 - Discovery versus confirmation study

3

Selection Bias

- Non-random selection of cases and controls can create bias if
 - The selection mechanism depends on exposure of interest and is differential by disease status
- Less of a concern for genetic association studies, but influence of family history and behavioral genes on participation of cases and controls can create problem
 - E.g. healthy subjects are more likely to participate if they have family history

4

Confounding

- **Confounding bias:** disease outcome seems associated with exposure merely due to its correlation with true risk-factor for disease
- **Population-stratification (PS):** particular type of confounding: **unobserved sub-populations with different allele frequencies and different disease frequencies;** causes
 - spurious association, i.e. false positive results, or
 - failure to detect true associations

to **avoid** PS cases and controls should have comparable genetic ancestry background

5

Population-stratification, cont.

No methods for candidate gene studies

In **GWAS**, availability of many **null** SNPs allows to

- Monitor the extent of PS
 - Over-dispersion factor
 - Q-Q plot
- Estimate the population ancestry and correct for PS (at the cost of power)
 - **STRUCTURE** (Pritchard et al. 2000)
 - **EIGENSTRAT** (Price et al. 2006)
 - Other methods: Satten et al. 2001, Epstein et al, 2007

6

Differential Genotype Miss-Classification

Failure to call genotype not independent of case-control status

- Case-control studies could be prone to differential genotyping error due to differences in
 - DNA quality
 - Sample storage
 - shipment
- Differential measurement error can cause bias away from null

Clayton et al, Nature Genetics, 2005

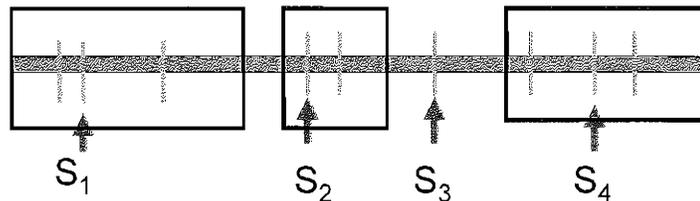
7

Variance distortion: cryptic relatedness

- Existence of sub-populations with same allele frequencies
- **Alleles** in same subpopulation are **correlated (IBD sharing)**
- Marker alleles across different sub-populations are uncorrelated
- Causes **variance distortion**
- Corrected with **genomic control** approach:
compute **variance inflation factor** (Devlin and Roeder, 1999, Setakis et al, 2006, Zheng et al, 2006, Reich and Goldstein, 2001)
- **Genomic control methods do NOT correct for population stratification!** Marchini et al, 2004; Campell et al, 2005

Which SNPs to Genotype?

- 10 million common **single nucleotide polymorphisms** (SNPs)
- Utilization of correlation (**Linkage Disequilibrium, LD**) of genetic variants near each other to obtain a reduced set of 300K-1m SNPs that captures most of genetic variation



9

Linkage Disequilibrium

Bi-allelic disease locus: disease allele G (p_G), wild type allele g (p_g); Bi-allelic marker: a, A (p_a , p_A)

Linkage disequilibrium (LD) defined as

$$D = P(A, G) - p_A p_G$$

$$D' = D/D_{\max} = D / \min(p_G p_a, p_g p_A) \text{ for } D > 0$$

$$r^2 = (D')^2 p_a p_G / p_A p_g$$

D' is upper bound of r^2

Genotype based test for independent cases & controls

	Marker Genotype			
	aa	aA	AA	total
Cases	r_0	r_1	r_2	R
Controls	s_0	s_1	s_2	S
Total counts	n_0	n_1	n_2	N

Remark on the null hypothesis

H_0 : "no association between SNP and disease"

is true if either one of the following

1. Disease has no genetic component
2. true disease locus not in LD with observed SNP

**Sample Sizes N for 100% attainable LD
between marker and causal locus,
OR=2.7, $\alpha=0.05$, power = .80, AR=0.1**

Prob. Dis.	Genetic model	Disease SNP P_G	Disease SNP N	Tag SNP $P_A=0.2$ N	Tag SNP $P_A=0.5$ N
0.1	Additive	0.04	312	1316 $r^2=.17$	4938 $r^2=.05$
0.1	Dom.	0.04	330	1402	5300
0.1	Recess.	0.29	594	938	1322
0.01	Additive	0.03	338	1574	5858

**Genome wide association
approach**

- Try to get closer to disease locus by high density SNP coverage: **tag SNPs**
- Obtain good coverage of untyped SNPs by $r^2 > 0.8$ within bin
- **Pay price of multiplicity**
- Still estimated that 25% of SNPs not captured adequately by LD and tag SNPs

Designs for association studies

- **Single stage design:** all markers measured on all samples
- **Two stage design:**
 - Stage 1:** Proportion of available samples genotyped on large number of markers
 - Stage 2:** Proportion of these markers are followed up by genotyping them on remaining samples

15

Two stage designs

Skol et al, Nature Genetics, 2006

Stage 1: proportion of samples, π , genotyped on all SNPs

Stage 2: SNPs that have test statistic $T_1 > C_1$ for some significance threshold C_1 followed up by genotyping on remaining samples.

Replication analysis: view stage 2 as replication study; final significance of SNPs based on analogous test statistic $T_2 > C_2$

Joint analysis: for each of the SNPs selected in stage 1, compute

$$T_{\text{joint}} = T_1\sqrt{\pi} + T_2\sqrt{1-\pi} > C_3$$

16

Example 1: Power computations

- **8000 cases/8000 controls**
- $\alpha=10^{-7}$ (500,000 tests)
- Disease prevalence=10%, disease allele frequency=20%, additive model, **OR=1.2**
- 25% samples genotyped in stage 1

Power for

- one stage analysis: **97%**
- replication analysis: **86%**
- joint analysis: **93%**

17

Example 2: 3000 cases/3000 controls

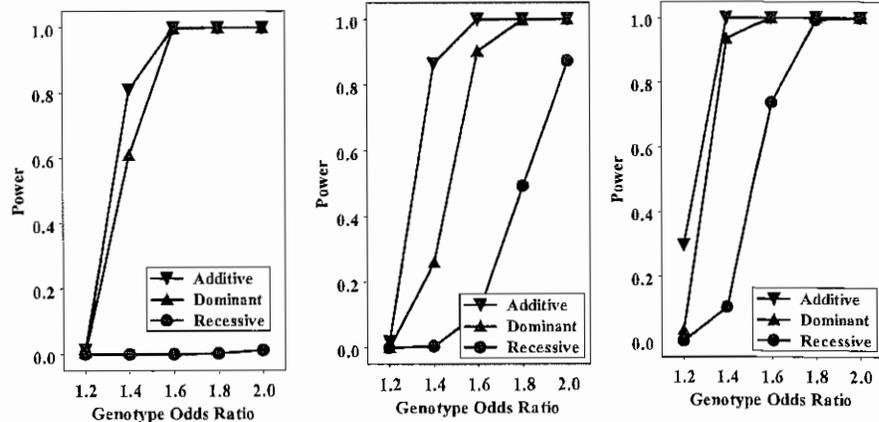
- $\alpha=10^{-7}$ (500,000 tests)
- Disease prevalence=10%, disease allele frequency=20%, additive model, **OR=1.3**
- 50% samples genotyped in stage 1

Power for

- one stage analysis: **84%**
- replication analysis: **27%**
- joint analysis: **83%**

18

Power for two stage genome wide scan (4000/4500 cases/controls)



risk allele frequency equal to (a) 0.1; (b) 0.3; and (c) 0.5.

Two stage designs, ranking

Stage 1: proportion of samples, π , genotyped on all SNPs

Stage 2: M_1 SNPs with smallest p values followed up by genotyping on remaining samples.

Replication analysis: view stage 2 as a replication study; final selection of SNPs depended on **ranking of p-values from stage 2 alone**

Joint analysis: for each of the M_1 SNPs selected in stage 1, compute

$$\lambda T_1 + (1-\lambda)T_2, \lambda = 0.0, 0.05, 0.1, \dots, 1.0$$

T_i test statistic for stage i

Ranking based approaches are for discovery!

Probability of detecting a disease SNP and optimal stage1 weight
 OR=1.2 per allele for 8000 cases/8000 controls for 500,000 SNPs

% sample in stage 1	Analysis	Number of true disease SNPs =1			
		$M_1 = 1000$		$M_1 = 25,000$	
		$M_2 = 1$	$M_2 = 100$	$M_2 = 1$	$M_2 = 100$
0.125	Replicate	.266	.269	.630	.664
	Joint	.267	.269	.635	.664
	λ_{opt}	.25	.27	.35	.25
0.25	Replicate	.612	.626	.803	.881
	Joint	.613	.626	.825	.885
	λ_{opt}	.32	.15	.40	.25
0.50	Replicate	.769	.897	.821	.912
	Joint	.843	.900	.858	.953
	λ_{opt}	.40	.20	.45	.40
1.00	One-stage^a	.882	.966	.882	.966

21

Combining case-control studies: Meta-analytic approach

First step: Estimate separate logistic models for each study

Second step: Estimate overall effect

$$\hat{\beta} = \sum_{s=1}^S w_s \hat{\beta}_s$$

where $w_s = Var(\hat{\beta}_s)^{-1}$

Q test statistic for between-studies heterogeneity among β_s

$$Q = \sum_{s=1}^S w_s (\hat{\beta}_s - \hat{\beta})^2 \sim \chi_{s-1}^2$$

Assumption: effect of SNP has same direction in all studies (not true if LD patterns are different in different study populations)

Alternative meta-analytic approach

First step: Estimate separate logistic models for each study

Second step: combine Wald test statistics to assess significance of SNP:

$$W=W_1 + \dots W_s \sim \text{chi square, } S \text{ d.f.}$$

where $W_i = \beta_i^2 / \text{Var}(\beta_i)$

Discovery only!

Aggregate Analysis

- Estimate single association parameter, adjusted for study.
- Assumptions: same SNP effect in all studies, same adjustments for confounding.
- Advantage: 1 d.f. chi square test for association: more powerful

References

- Skol et al. 2006. Nature Genetics, 38:209-13.
 Dudridge 2006, Am J of Hum Genetics. 78:1094-95.
 Marchini et al. 2006. Nature Genetics, 2005: 413-7.
 Freedman et al. 2004. Nature Genetics, 36: 388-395
 Marchini et al. 2004. Nature Genetics, 36: 512-517
 Campbell et al. 2005. Nature Genetics, 37: 868-872
 Devlin and Roeder, 1999. Biometrics, 55: 997-1004
 Pfeiffer and Gail, 2003. Genetic Epidemiology, 25:136-148
 Price et al. 2006. Nature Genetics, 38: 904-909.
 Pritchard JK et al. 2000. Genetics, 155: 945-959
 Reich and Goldstein, 2001. Genetic Epidemiology, 20: 4-16
 Satten GA et al. 2001. Am J. Hum. Genet. 68: 466-477
 Cochran WG. The combination of estimates from different experiments.
 Biometrics, 10, 101-129, 1954
 DerSimonian R, Laird N. Control Clin Trials 1986
 Gail, Pfeiffer, Wheeler, Pee. Biostatistics. 9(2):201-15. 2008.

Meta-analytic approach, cont.

Shortcoming of Q statistic:

- assumption: weights are known
- low power when number of studies included is small

Estimate between study variance

$$\hat{\sigma}_B^2 = \{Q - (k - 1)\} / \left\{ \sum_{s=1}^S w_s - \frac{\sum_{s=1}^S w_s^2}{\sum_{s=1}^S w_s} \right\}$$

Hardy RJ. Detecting and describing heterogeneity in meta analysis. Stat in Med,

17 : 841 1998