Design and Analysis Issues in Genomewide Association Scans

Duncan Thomas, Jonathan Buckley,
David Conti, Jim Gauderman,
Juan Pablo Lewinger, Dan Stram

USC
UNIVERSITY
OF SOUTHERN
CALIFORNIA

Multistage Designs for Genetic Associations

- Satagopan et al. (2002- 4): two-stage design, testing all markers in stage I followed by testing a subset on additional subjects in stage II
- We propose adding additional tagging SNPs in all regions initially flagged before proceeding to stage II
- and take differences in genotyping costs into account

Satagopan et al., *Genet Epidemiol* 2003;25:149-57

Multistage Design

- **Stage I:** full scan of 500,000 SNPs on sample of size N_1
- **Stage II:** genotype only SNPs "significant" at level α_1 from stage I on a new sample of size N_2
- Final analysis combines both samples at significance level α_2, chosen to ensure an overall Type I error rate α
  - Significance assessed conditionally on hit in stage I
- Optimize choice of N_1 and α_1 to minimize cost subject to constraint on α and power

Satagopan et al., *Genet Epidemiol* 2003;25:149-57

Optimal Designs

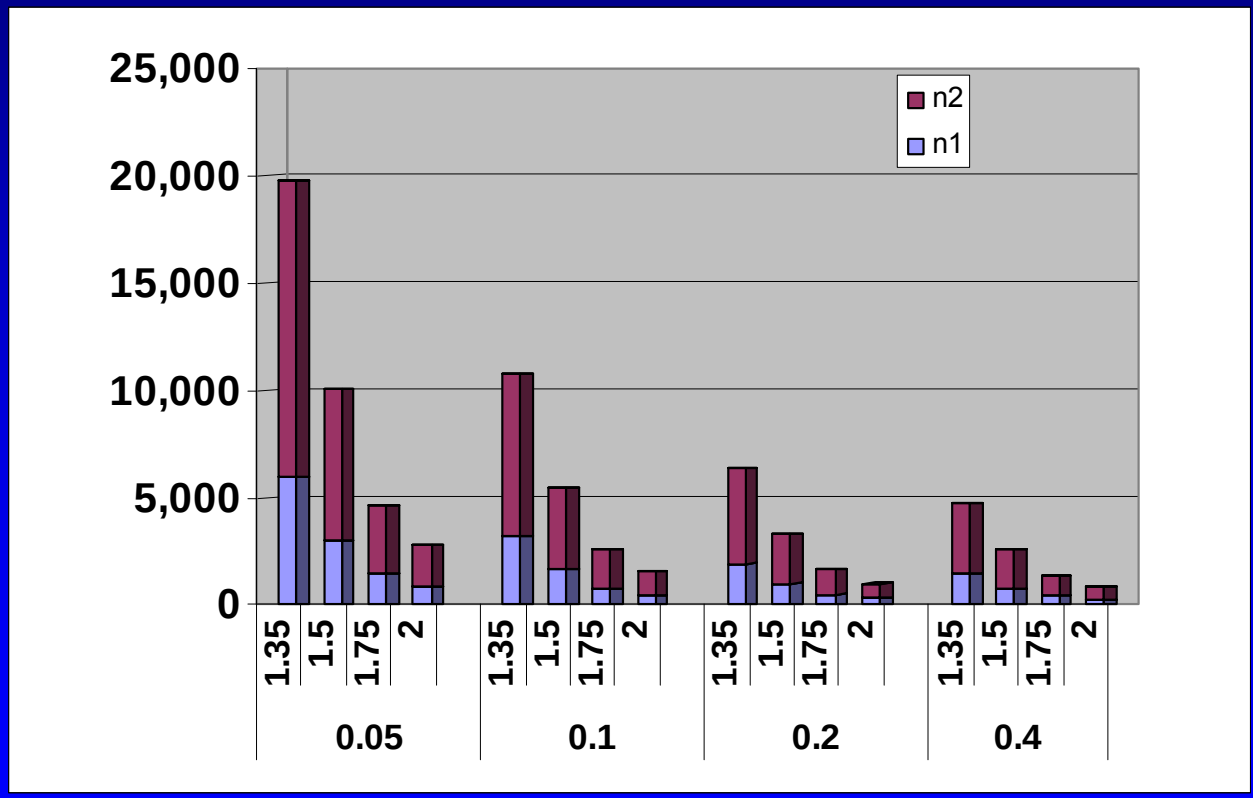
Per-Genotype Cost Ratio = 17.5 for Stages II / I:

Genomewide $\alpha = .05$, $1 - \beta = 0.9$

Minimizing Total Cost

$\alpha_1 = .0038$
$1 - \beta_1 = 0.907$
$\alpha_2 = 1.7 \times 10^{-7}$
$1 - \beta_2 = 0.987$

$n_1/n_* = 30\%$



Wang, Thomas & Stram, *Genet Epidemiol* 2006:30:3

Designs Using Additional Markers

- **Plan A:** type additional markers on stage I sample around each “hit”; then type subset of most significant original or extra markers on stage II sample
- **Plan B:** type additional markers on stage II sample only for each hit from stage I; combined analysis uses indirect haplotype-based associations for stage I samples
- **Plan C:** no additional markers until stage III

Indirect SNP Associations

- Suppose in stage I we observe markers M_i on $i = 1,\ldots,N_1$ subjects, and in stage II markers M_j on $j = 1,\ldots,N_2$ subjects
- We wish to draw inference about a particular SNP A in M_j that was not included in M_i

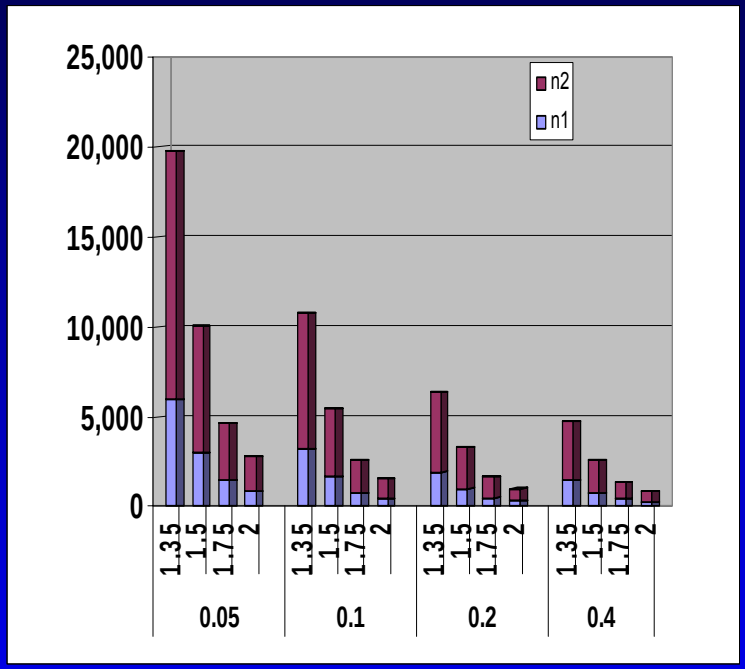
$$L_A(\beta,\alpha) = \prod_{i=1}^{N_1} \sum_a P_\beta(Y_i \mid A{=}a)\, P_\alpha(A{=}a \mid M_i)$$

$$\times \prod_{j=1}^{N_2} P_\beta(Y_j \mid A_j)\, P_\alpha(A_j \mid M_j)$$

– Thomas et al, Genet Epidemiol 2004;27:401-14

Optimal Designs

Per-Genotype Cost Ratio = 17.5 for Stages II / I: Genomewide $\alpha = .05$, $1 - \beta = 0.9$

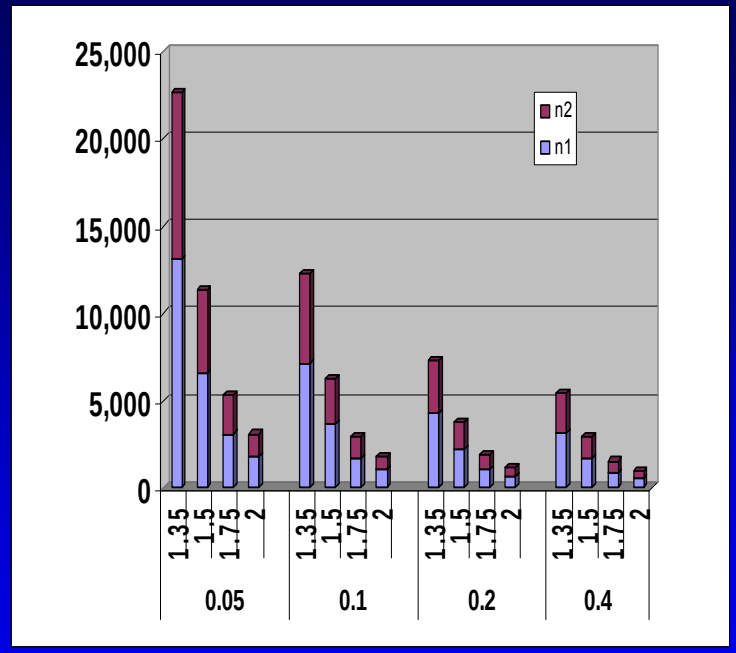


Minimizing Total Cost

$\alpha_1 = .0038$ $1 - \beta_1 = 0.907$

$\alpha_2 = 1.7 \times 10^{-7}$ $1 - \beta_2 = 0.987$

$n_1/n_* = 30\%$



5 Additional Markers Typed at Stage II

$R_s^2 = 0.6$ at stage I and 0.9 stage II

$\alpha_1 = .0005$ $1 - \beta_1 = 0.906$

$\alpha_2 = 0.5 \times 10^{-7}$ $1 - \beta_2 = 0.975$

$n_1/n_* = 49\%$

Wang, Thomas, Pe'er & Stram, *Genet Epidemiol* 2006:30:356-68

Other Possible Options

- More that two stages
- Other constraints:
  - Total sample size fixed
  - Stage 1 sample size fixed,
    optimize significance levels at stages I and II
- Different designs at stages I and II
  - E.g., population-based vs. family-based
  - SNP vs. haplotype tests
  - When to test for interactions?

Hierarchical Approach to Prioritizing SNPs

- Standard multistage designs assume the α_1 most significant SNPs from the first stage will be tested in later stage(s)
- Can we do better?
- False discovery rate using a weights by prior knowledge (Roeder et al, *AJHG* 2006:78:243-42)
- Bayesian FDR (Whittemore, *CEBP* 2005;14:1359)
- Empirical Bayes ranking, using an exchangeable mixture prior with a large mass at *RR* = 1
- Adding prior knowledge to hierarchical Bayes

Empirical Bayes Ranking

- Assume an "exchangeable" distribution of noncentrality parameter λ_m for the observed unsigned chi statistics χ_m for markers *m=1…M*
  - $\Pr(\lambda_m \neq 0) = \pi$
  - $\Pr(\lambda_m \mid \lambda_m \neq 0) = fN(\mu,\sigma^2)$
- Estimate parameters $\Theta = (\pi,\mu,\sigma^2)$ given set of observed chi statistics $\mathcal{D} = \{\chi_m\}$, $\chi_m \sim fN(\lambda_m,1)$
- Then estimate $\hat{p}_m = \Pr(\lambda_m \neq 0 \mid \chi_m, \Theta)$
  and $\hat{e}_m = E(\lambda_m \mid \lambda_m \neq 0, \chi_m, \Theta)$
- Rank unconditional expectations $\hat{E}_m = \hat{p}_m \hat{e}_m$

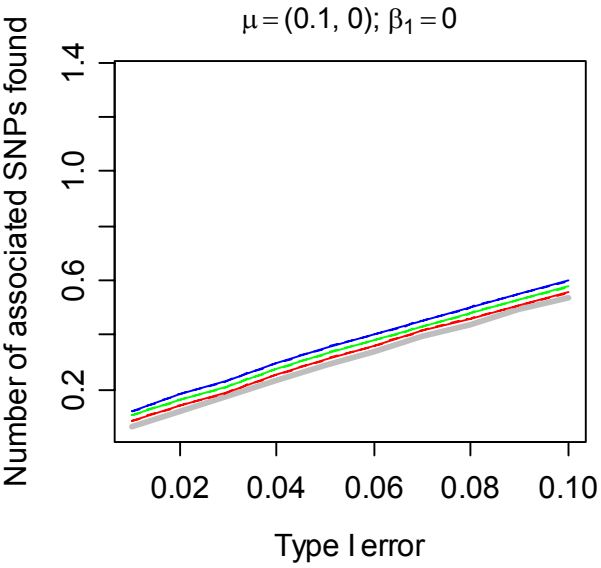
Incorporating Genomic Annotation

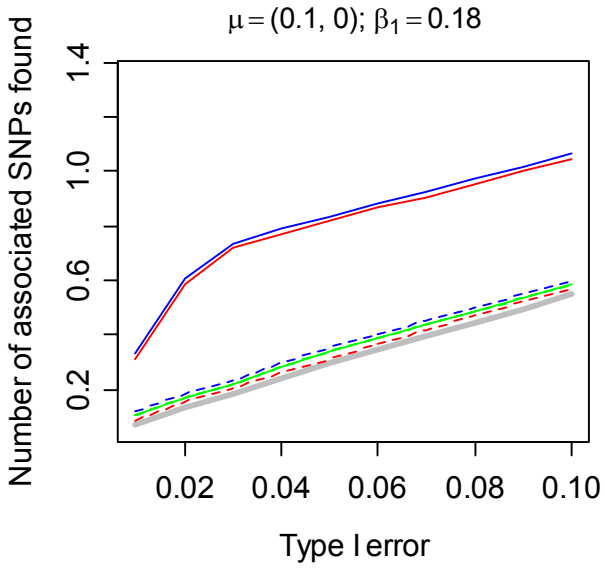
- Extend the mixture prior to incorporate a vector of *prior covariates* Z

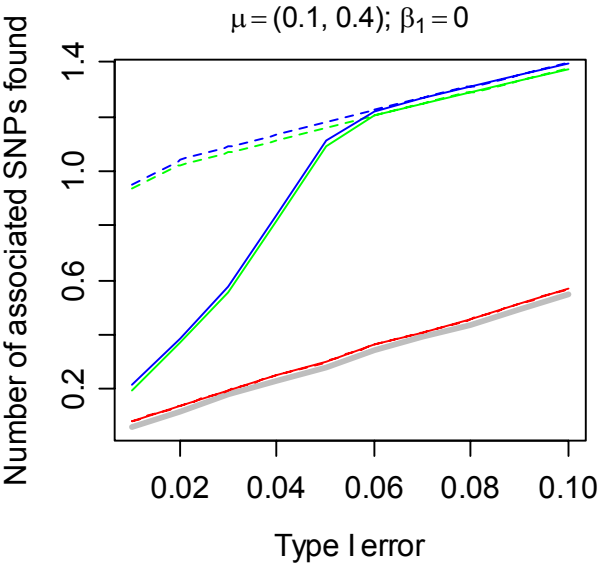
$$\text{logit Pr}(\lambda_m \neq 0) = \pi_0 + \pi_1' Z_m$$

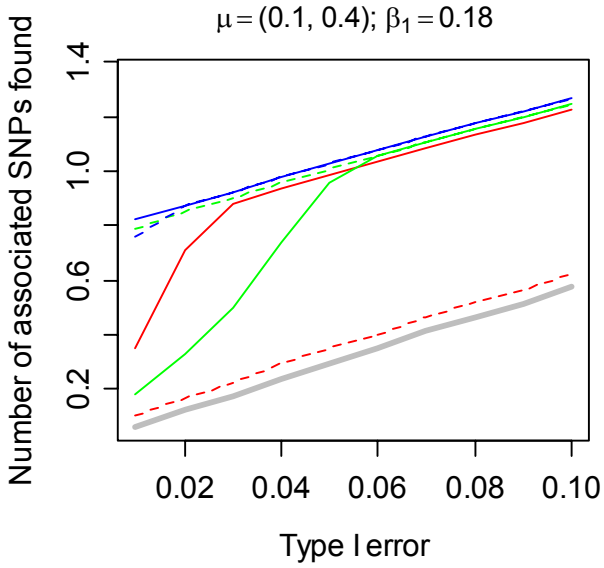
$$E(\lambda_m \mid \lambda_m \neq 0) = \mu_0 + \mu_1' Z_m$$

- Examples of prior covariates:
  - Location relative to known or predicted genes
  - Predicted function or evolutionary conservation
  - Prior linkage or association results









Covars in Prob Model?	Covariates in Means Model?	
	No	Yes
No	All rankings equal	$E[\lambda > 0 \mid \chi]$ covariates in both
Yes	$Pr[\lambda > 0 \mid \chi]$ covariates in both or probability only	$E[\lambda > 0 \mid \chi]$ covariates in both

- $E[\lambda \mid \chi]$, means model only
- $Pr(\lambda > 0 \mid \chi)$, means model only
- $E[\lambda \mid \chi]$, probability model only
- $Pr(\lambda > 0 \mid \chi)$, probability model only
- $E[\lambda \mid \chi]$, both
- $Pr(\lambda > 0 \mid \chi)$, both

Methodological Issues

- TagSNP selection and haplotype analysis
  - "Bake-off" of alternative methods
  - Unifying haplotype association & sharing
- Multistage sampling and multiple comparisons
  - Study designs using additional markers
  - Resampling methods for 2-stage designs
  - Hierarchical models for selecting SNPs for stage 2
- Family- vs. population-based studies
  - Hybrid design/analysis using both
  - Adjustments for population stratification
- GxE & GxG interactions

Practicalities: What We Decided

- Balancing main effects and interactions
- Ethnic heterogeneity
- Prioritization to SNPs to carry forward
- Multiple endpoints
- Single SNP vs haplotype tests
- Additional SNPs
- Family-based vs population-based designs
- Replication
- Etc.

- Short list of criteria for stage I
- More detailed analyses of joint stage I/II data
- To select SNPs for stage II,
  - form separate rankings for each criterion
  - Build consolidated list from weighted ranks, eliminating redundant SNPs, until target number obtained

Practicalities: What We Decided

- Balancing main effects and interactions
- Ethnic heterogeneity; genomic control
  - Treated like interactions in building consolidated list
  - Criterion 1: pan-ethnic effects (race adjusted)
  - Criterion 2: test of between-group heterogeneity
    (Other projects adopted ethnic-specific tests)
- Prioritization to SNPs to carry forward
- Multiple endpoints
- Single SNP vs haplotype tests
- Additional SNPs
  - Selection of top-ranked rather than fixed significance level is implicitly a form of genomic control
- Family-based vs population-based designs
- Replication
- Etc.
  - Joint stage I/II analysis will use more powerful structured association methods

Practicalities: What We Decided

- Balancing main effects and interactions
- Ethnic heterogeneity: genomic control
- **Prioritization to SNPs to carry forward**
- Multiple endpoints
  - **Use hierarchical modeling strategy for main effects only**
- Single SNP vs haplotype tests
- Additional SNPs
- Family-based vs population-based designs
- Replication
- Etc.

Practicalities: What We Decided

- Balancing main effects and interactions
- Ethnic heterogeneity; genomic control
- Prioritization to SNPs to carry forward
- **Multiple endpoints**
  - Adopt a single genome-wide significance level for each endpoint (and type of analysis)
- Single SNP vs haplotype tests
- Additional SNPs
  - Form consolidated list of SNPs across endpoints
- Family-based vs population-based designs
- Replication
- Etc.

Practicalities: What We Decided

- Balancing main effects and interactions
- Ethnic heterogeneity; genomic control
- Prioritization to SNPs to carry forward
- Multiple endpoints
- **Single SNP vs haplotype tests**
  - Test all typed SNPs directly
  - And all common untyped SNPs indirectly using haplotypes that predict them in stage I
- Additional SNPs
- Family-based vs population-based designs
- Replication
- Etc.
  - Prioritize SNPs separately and take top-ranked SNPs forward to stage II

Practicalities: What We Decided

- Balancing main effects and interactions
- Ethnic heterogeneity; genomic control
- Prioritization to SNPs to carry forward
- Multiple endpoints
- Single SNP vs haplotype tests
- **Additional SNPs**
  - In stage II, genotype single best untyped SNP near any selected SNP that is more strongly associated
  - Joint analysis will combine tested SNPs from stage II and expected SNP dosage from stage I using available typed SNPs
- Family-based vs. population-based designs
- Replication
- Etc.

Practicalities: What We Decided

- Balancing main effects and interactions
- Ethnic heterogeneity; genomic control
- Prioritization to SNPs to carry forward
- Multiple endpoints
- Single SNP vs haplotype tests
- Additional SNPs
- **Family-based vs population-based designs**
- Replication
- Etc.

  - **Stage 1 uses population-based unrelated cases & controls; stage II is family-based (some overlap)**
  - **We will combine two samples in joint analysis**

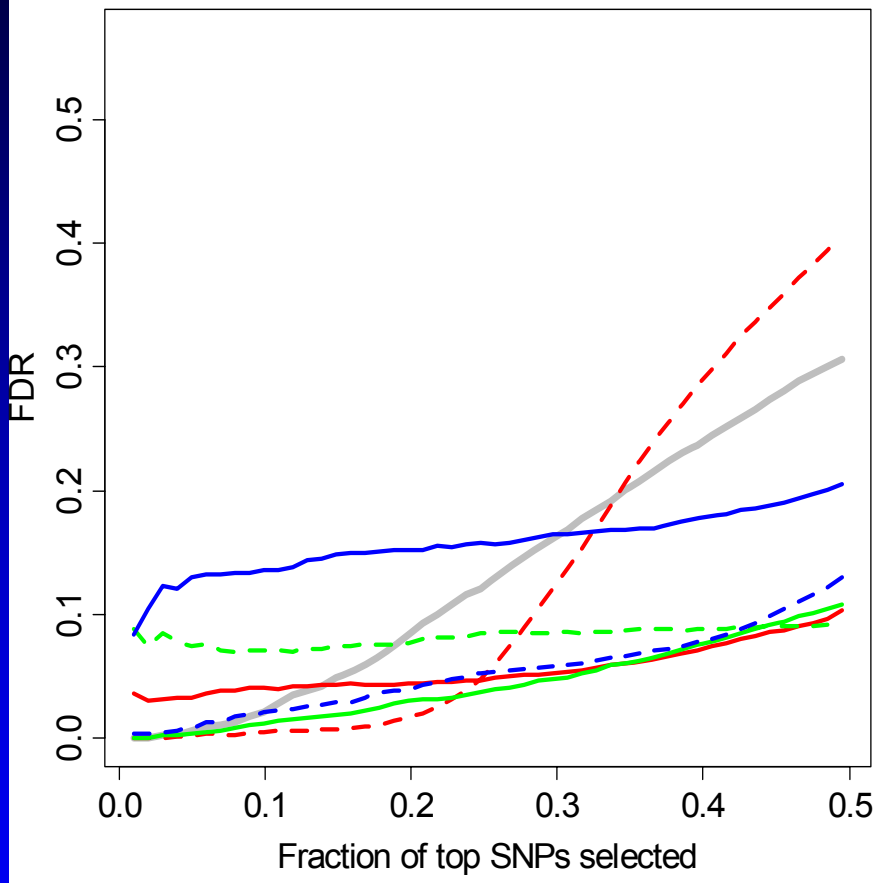
Practicalities: What We Decided

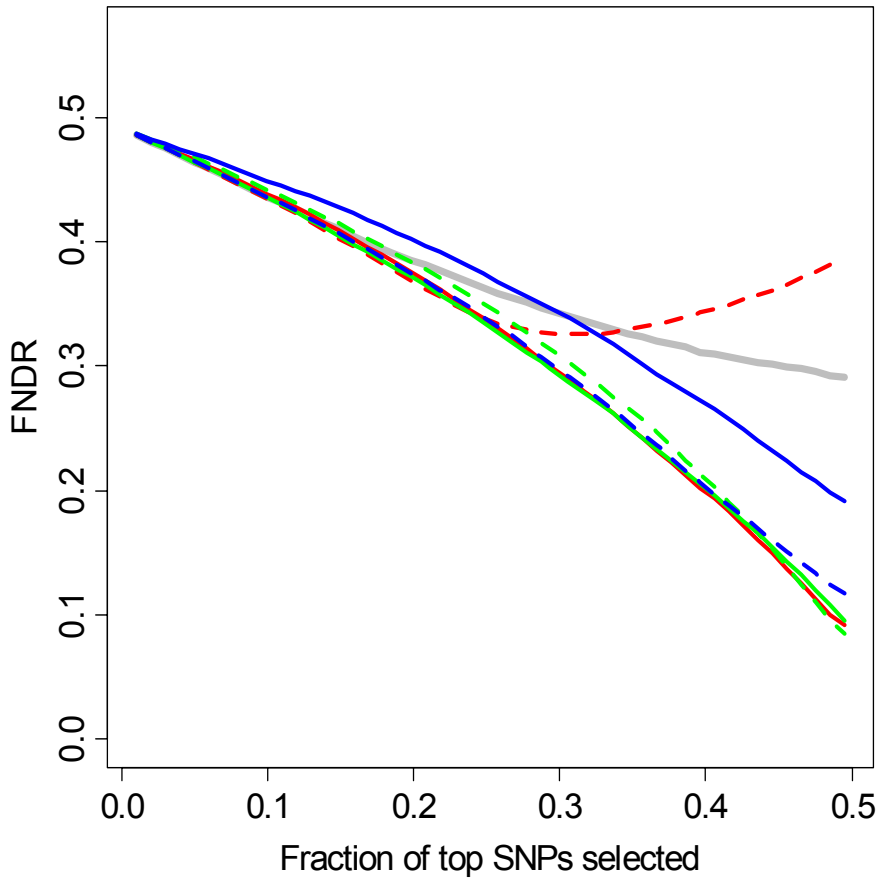
- Balancing main effects and interactions
- Ethnic heterogeneity; genomic control
- Prioritization to SNPs to carry forward
- Multiple endpoints
- Single SNP vs haplotype tests
- Additional SNPs
- Family-based vs population-based designs
- **Replication**
  - **Independent samples, depending on specific study**
- Etc.

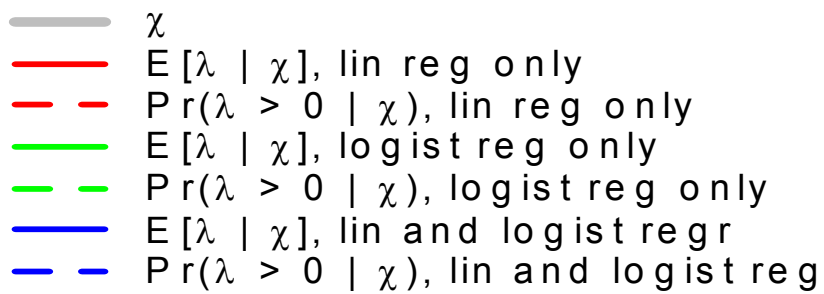
Conclusions

- Costs have now become feasible: many such studies now being undertaken
- Efficient design and analysis strategies essential
- Rich area for statistical research

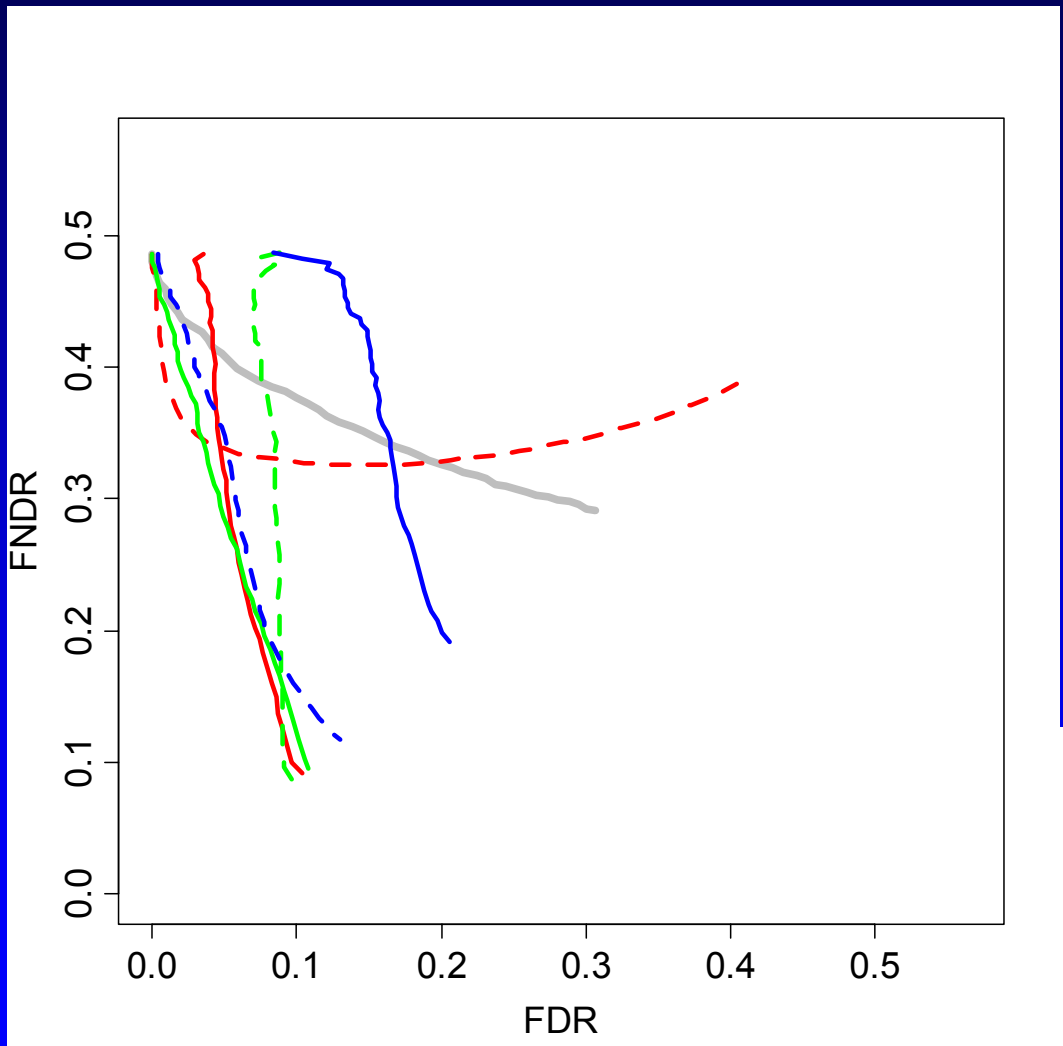
FDR vs FNDR

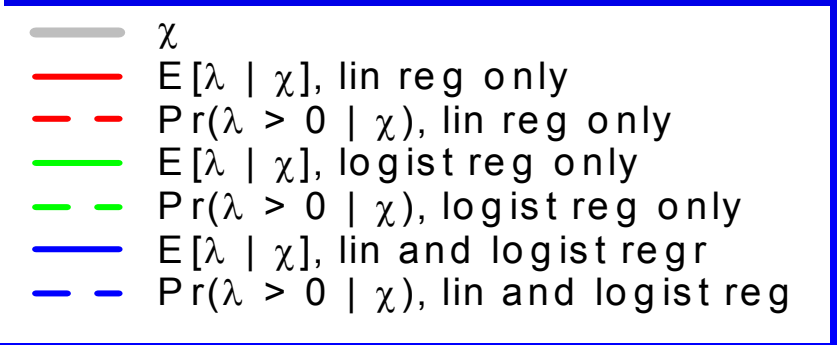

FDR
0.0
0.1
0.2
0.3
0.4
0.5
Fraction of top SNPs selected


FNDR
0.0
0.1
0.2
0.3
0.4
0.5
Fraction of top SNPs selected


χ
E[λ | χ], lin reg only
Pr(λ > 0 | χ), lin reg only
E[λ | χ], logist reg only
Pr(λ > 0 | χ), logist reg only
E[λ | χ], lin and logist regr
Pr(λ > 0 | χ), lin and logist reg

ROC Curve


FNDR
0.0
0.1
0.2
0.3
0.4
0.5
FDR
0.0
0.1
0.2
0.3
0.4
0.5


χ
E[λ | χ], lin reg only
Pr(λ > 0 | χ), lin reg only
E[λ | χ], logist reg only
Pr(λ > 0 | χ), logist reg only
E[λ | χ], lin and logist regr
Pr(λ > 0 | χ), lin and logist reg

What and Why GWAS?

- ***What:*** **a scan of the entire genome for SNP polymorphisms associated with disease**
  - **typically ~ 100K – 1M markers used**
  - **most associations expected to due to LD with an unobserved causal locus, not directly causal**

What and Why GWAS?

- ***What:*** **a scan of the entire genome for SNP polymorphisms associated with disease**
- ***Why:*** **"common disease common variant" hypothesis – complex diseases involve multiple genes with common, low penetrance polymorphisms, interacting with each other and/or environmental factors**
  - **such associations are difficult to detect by linkage**
  - **contrary view: "multiple rare variants" hypothesis**

Terwilliger, *Eur J Hum Genet* 2006;14:426-37
Pritchard & Cox, Hum Mol Genet 2002;11:2417-23
Pritchard, AJHG 2001;69:124-37

The “Unit” of Analysis

- We take the view that our ultimate aim is to test association with all ~5M common variants
- 500K SNPs on chip effectively tag most of these, but additional markers will be needed to fully explore regions flagged (multistage design required)
  - But cf. proviso in

    Jorgenson & Witte, *AJHG* 2006:78:884-8
- These 5M tests are dependent, an “effective” number of ~1M independent tests

Methodological Issues

- TagSNP selection and haplotype analysis
  - "Bake-off" of alternative methods
  - Unifying haplotype association & sharing
- Multistage sampling and multiple comparisons
  - Study designs using additional markers
  - Resampling methods for 2-stage designs
  - Hierarchical models for selecting SNPs for stage 2
- Family- vs. population-based studies
  - Hybrid design/analysis using both
  - Adjustments for population stratification
- GxE & GxG interactions

Practicalities

- Balancing main effects and interactions
- Ethnic heterogeneity; genomic control
- Prioritization to SNPs to carry forward
- Multiple endpoints
- Single SNP vs haplotype tests
- Additional SNPs
- Family-based vs population-based designs
- Replication
- Etc.

Am. J. Hum. Genet. 77:337–345, 2005

Recent Developments in Genomewide Association Scans: A Workshop Summary and Review

Duncan C. Thomas,[1] Robert W. Haile,[1] and David Duggan[2]

[1]Department of Preventive Medicine, University of Southern California, Los Angeles; and [2]T Phoenix

Acknowledgments

This workshop was supported by the University of Southern California (USC) Center of Excellence in Genomic Sciences (grant 1P50 HG002790), the Southern California Environmental Health Sciences Center (grant 5P30 ES07048), and the USC Keck School of Medicine. Invited speakers included Ha- (Columbia University), Fernando Arena (National tute, National Institutes of Health), Paul de Bakker etts General Hospital), Timothy Bishop (Leeds Jonathan Buckley (University of Southern Cali- ham Casey (Cleveland Clinic Foundation), David ambridge Institute of Medical Research), Mariza (Mayo Clinic), David Duggan (TGen), Eleazar ersity of California at San Diego), Nelson Freimer of California at Los Angeles), Ellen Goode (Mayo ek Gordon (Rockefeller University), Robert Haile of Southern California), Brian Henderson (Uni- outhern California), John Hopper (University of Melbourne), Eric Jorgenson (University of California at San Francisco), Magnus Nordborg (University of Southern California), Lyle Palmer (University of Western Australia), Itsik Pe'er (Broad Institute), Chiara Sabatti (University of California at Los Angeles), Jaya Satagopan (Memorial Sloan Kettering Cancer Center), Nik Schork (University of California at San Diego), Daniela Seminara (National Cancer Institute, National Institutes of Health), Susan Service (University of California at Los Angeles), Dan Stram (University of Southern California), Simon Tavaré (University of Southern California), Nicole Tedeschi (University of Southern California), David Van Den Berg (University of Southern California), Alice Whittemore (Stanford University), and John Witte (University of California at San Francisco).

Editorial

Are We Ready for Genome-wide Association Studies?

Duncan C. Thomas

University of Southern California, Los Angeles, Califo

Cancer Epidemiol Biomarkers Prev 2006;15(4). April 2006

Genetic Epidemiology 30: 356–368 (2006)

Optimal Two-Stage Genotyping Designs for Genome-Wide Association Scans

Hansong Wang,[1] Duncan C. Thomas,[1] Itsik Pe'er,[2] and Daniel O. Stram[1]*

[1]*Division of Biostatistics and Genetic Epidemiology, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California*

[2]*Broad Institute of MIT and Harvard University, Cambridge, Massachusetts*

Methodological Issues

- TagSNP selection and haplotype analysis
  - "Bake-off" of alternative methods
  - Unifying haplotype association & sharing
- **Multistage sampling and multiple comparisons**
  - Study designs using additional markers
  - Resampling methods for 2-stage designs
  - **Hierarchical models for selecting SNPs for stage 2**
- Family- vs. population-based studies
  - Hybrid design/analysis using both
  - Adjustments for population stratification
- GxE & GxG interactions