Breast and Colon CFR Genomewide Association Studies

Duncan Thomas

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**
- ***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**

CFR and Related Initiatives

- Breast CFR GWA study (Ahsan)
- Colon CFR GWA study (Casey)
- Others underway:
  - UK breast (Ponder, Easton, …)
- Others pending:
  - Multiethnic Cohort (Breast, Colon, Prostate)
  - Design and Analysis U01

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 Dugga

[1]Department of Preventive Medicine, University of Southern California, Los Angeles; and 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 Habib Ahsan (Columbia University), Fernando Arena (National Cancer Institute, National Institutes of Health), Paul de Bakker (Massachusetts General Hospital), Timothy Bishop (Leeds University), Jonathan Buckley (University of Southern California), Graham Casey (Cleveland Clinic Foundation), David Clayton (Cambridge Institute of Medical Research), Mariza de Andrade (Mayo Clinic), David Duggan (TGen), Eleazar Eskin (University of California at San Diego), Nelson Freimer (University of California at Los Angeles), Ellen Goode (Mayo Clinic), Derek Gordon (Rockefeller University), Robert Haile (University of Southern California), Brian Henderson (University of Southern 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).

Some Issues in GWAS

- Multistage study designs
- Multiple comparisons
- Prioritizing SNPs for second stage
- Haplotype analyses using tag SNPs:
  unifying association and sharing
- Control of population stratification
- GxE and GxG interactions

Thomas et al, *AJHG* 2005:77:337-45

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

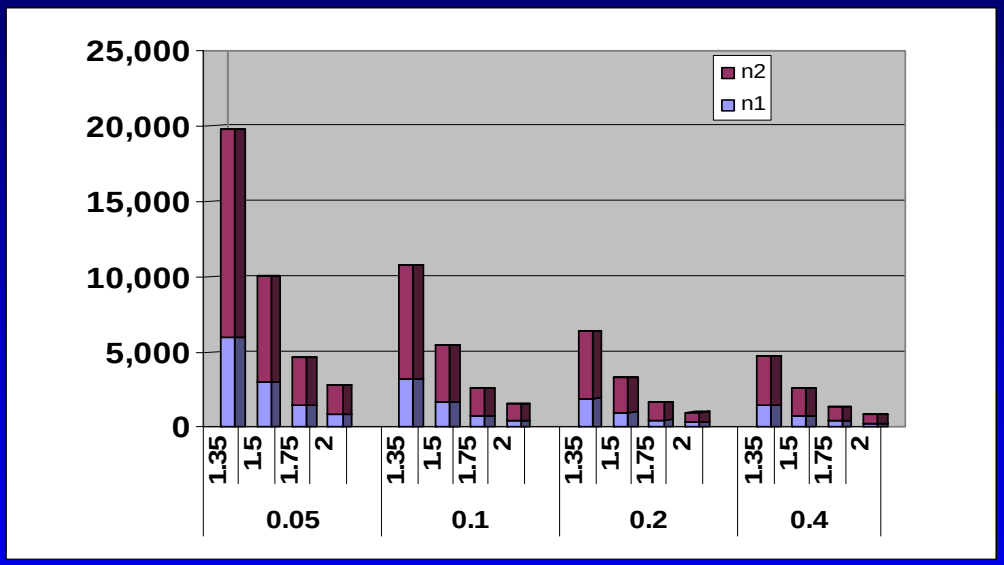
Multistage Design

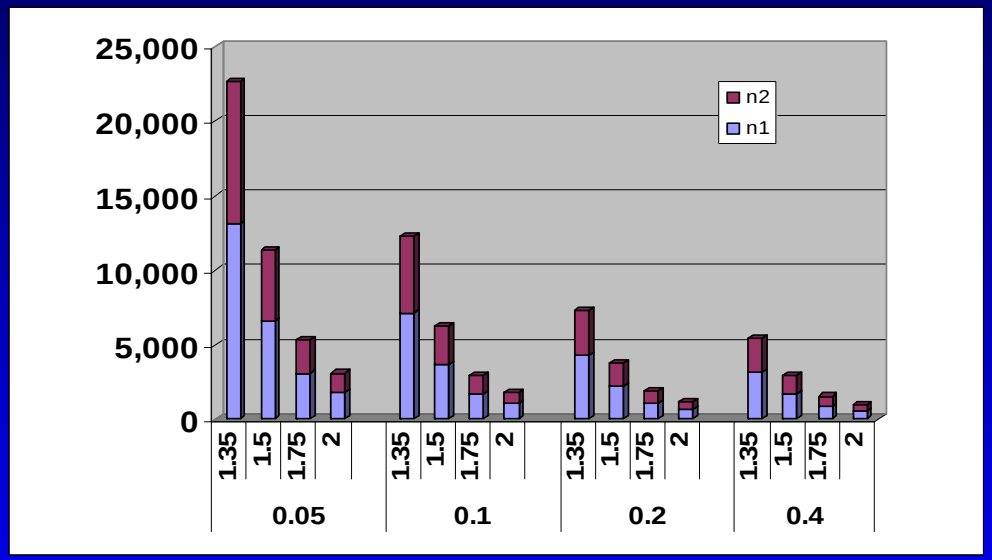
- **Stage I:** full scan of 500,000 SNPs on sample of size N_1
- **Stage II:** genotype only SNPs "significant" at level α_1 from stage I on a new sample of size N_2
- Final analysis combines both samples at significance level α_2, chosen to ensure an overall Type I error rate α
  - Significance assessed 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 α = .05, 1 – β = 0.9





Minimizing Total Cost

α_1 = .0038 $\quad$ 1 – β_1 = 0.907

$\alpha_2 = 1.7\text{x}10^{-7}$ $\quad$ 1 – β_1 = 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

α_1 = .0005 $\quad$ 1 – β_1 = 0.906

$\alpha_2 = 0.5\text{x}10^{-7}$ $\quad$ 1 – β_2 = 0.975

$n_1/n_* = 49\%$

Wang, Thomas & Stram, under review

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?

Some Issues in GWAS

- Multistage study designs
- Multiple comparisons
- Prioritizing SNPs for second stage
- Haplotype analyses using tag SNPs:
  unifying association and sharing
- Control of population stratification
- GxE and GxG 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, ASHG 2005)
- Empirical Bayes ranking, using an exchangeable mixture prior with a large mass at *RR* = 1
- Adding prior knowledge to hierarchical Bayes

Incorporating Genomic Annotation

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

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

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

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

Some Other Issues

- Haplotype analyses using tag SNPs: unifying association and sharing
- GxG interactions: all at once or staged?
- Ditto for GxE interactions?
- Control of population stratification
- Family-based or hybrid designs?
- DNA pooling?
- Large-scale copy number changes?

Haplotype-Based Tests

- **Association**
  - **Is a particular haplotype more common in cases than controls?**
  - **Multi-*df* omnibus tests for all "common" haplotypes used in absence of *a priori* hypotheses**
  - **Multiple (correlated) tests used for different haplotype windows**
- **Sharing**
  - **Are pairs of cases more likely to share haplotype segments around a particular locus**

Some Other Issues

- Haplotype analyses using tag SNPs: unifying association and sharing
- GxG interactions: all at once or staged?
- Ditto for GxE interactions?
- Control of population stratification
- Family-based or hybrid designs?
- DNA pooling?
- Large-scale copy number changes?

Hybrid Family- and Population-based Designs

- Breast and Colon CFRs plan GWASs to exploit unique advantages of family-based resource
  - Focus first stage on genetically enriched cases
  - Use FBATs to overcome population stratification in second stage
- Final analysis needs to combine both designs

Some Issues in GWAS

- Multistage study designs
- Multiple comparisons
- Prioritizing SNPs for second stage
- Haplotype analyses using tag SNPs:
  unifying association and sharing
- Control of population stratification
- GxE and GxG interactions

Genomewide GxG Interactions

- Marchini et al. (*Nat Genet* 2005;37:413-7) showed that it was computationally feasible and statistically efficient to scan the genome for all possible pairwise gene-gene interactions!
- Lin et al. (*Nat Genet* 2004;36:1181-8) considered all all possible haplotype associations genomewide
- At what stage of a genomewide scan are these best done?
- Should they be limited to those showing marginal effects?

Conclusions

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