

# **INTRODUCTION TO GENETIC EPIDEMIOLOGY**

## **(GBIO0015-1)**

Prof. Dr. Dr. K. Van Steen

## **CHAPTER 6: FAMILY-BASED GENETIC ASSOCIATION STUDIES**

### **1 Setting the scene**

#### **1.a Introduction**

#### **1.b Association analysis**

Linkage vs association

#### **1.c GWAs**

Scale issues

## 2 Families versus cases/controls

### 2.a Every design has statistical implicationse

How does design change the selection of analysis tool?

### 2.b Power considerations

Reasons for (not) selecting families?

### 2.c The transmission disequilibrium test

Pros and cons of TDT

### 2.d The FBAT test

Pros and cons of FBAT

## **3 From complex phenomena to models**

### **3.a Introduction**

### **3.b When the number of tests grows**

Multiple testing

### **3.c When the number of variables grows**

Prescreening and variable selection

## 4 Family-based screening strategies

### 4.a PBAT screening

Screen first and then test using *all* of the data

### 4.b GRAMMAR screening

Removing familial trend first and then test

## 5 Validation

### 5.a Replication

What is the relevance if results cannot be reproduced?

### 5.b Proof of concept

### 5.c Unexplained heritability

What are we missing?

Concepts: heterogeneity

## 6 Beyond main effects

### 6.a Dealing with multiplicity

Multiple testing explosion ...

### 6.b A bird's eye view on a road less travelled by

Analyzing multiple loci jointly

FBAT-LC

### 6.c Pure epistasis models

MDR and FAM-MDR

## 7 Future challenges

# 1 Setting the scene

## 1.a Introduction to genetic associations

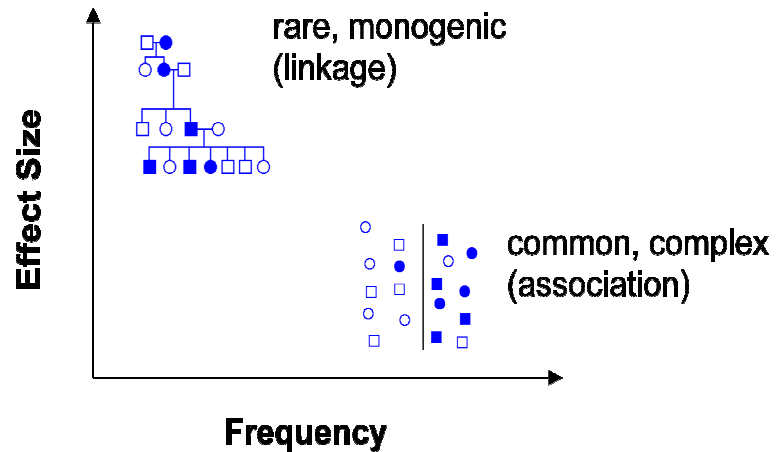
A genetic association refers to statistical relationships in a population between an individual's phenotype and their genotype at a genetic locus.

- Phenotypes:
  - Dichotomous
  - Measured
  - Time-to-onset
- Genotypes:
  - Known mutation in a gene (CKR5 deletion, APOE4)
  - Marker or SNP with/without known effects on coding



## 1.b Basic mapping strategies

Which gene hunting method is most likely to give success?

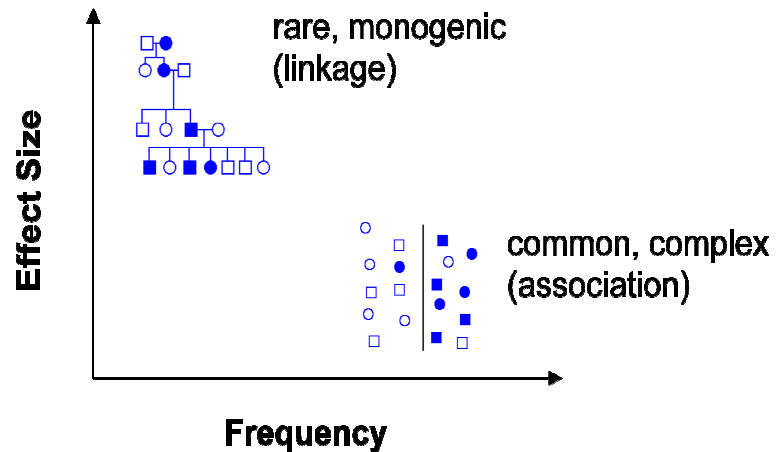


- Monogenic “Mendelian” diseases
  - Rare disease
  - Rare variants
    - Highly penetrant
- Complex diseases
  - Rare/common disease
  - Rare/common variants
    - Variable penetrance

(Slide: courtesy of Matt McQueen)

## Complex diseases

Which gene hunting method is most likely to give success?

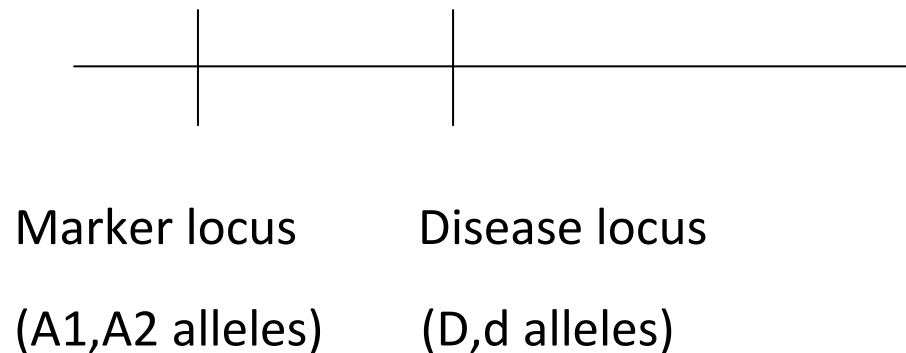


- Monogenic “Mendelian” diseases
  - Rare disease
  - Rare variants
    - Highly penetrant
- Complex diseases
  - Rare/common disease
  - Rare/common variants
    - Variable penetrance

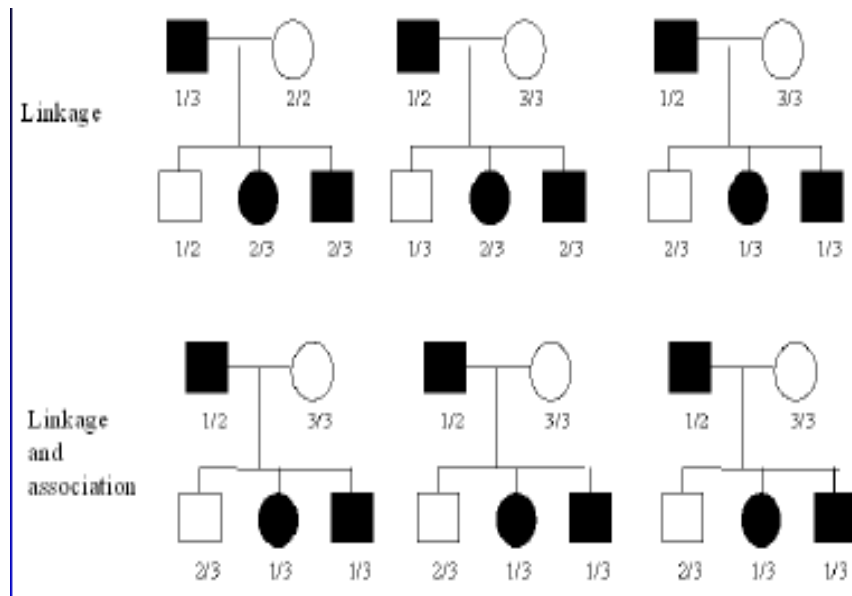
(Slide: courtesy of Matt McQueen)

## Using families: linkage versus association

- Linkage is a physical concept: The two loci are “close’ together on the same chromosome. There is hardly any recombination between disease locus and marker locus
- Association is a population concept: The allelic values at the two loci are associated. A particular marker allele tends to be present with disease allele.



## Features of linkage studies

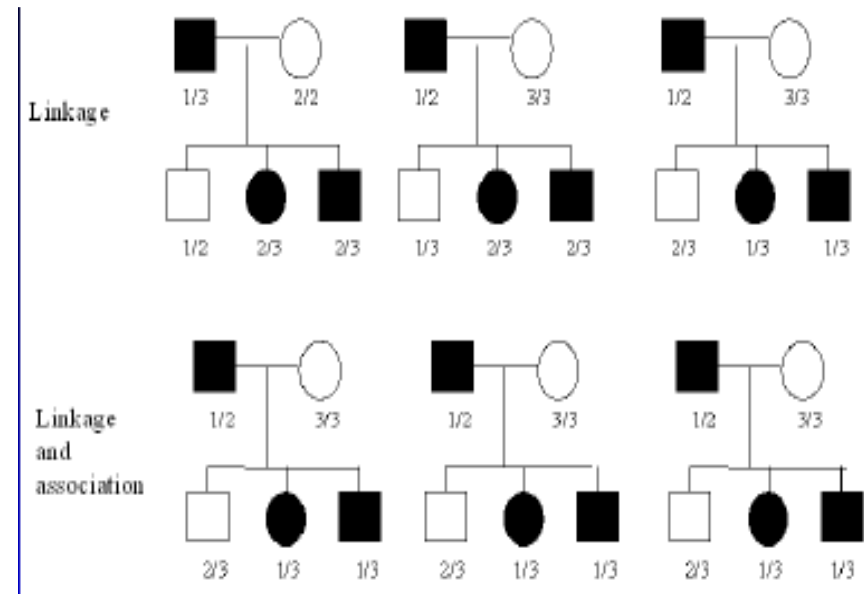


(Figure: courtesy of Ed Silverman)

- Linkage exists over a very broad region, entire chromosome can be done using data on only 400-800 DNA markers
- Broad linkage regions imply studies must be followed up with more DNA markers in the region
- Must have family data with more than one affected subject

## Features of association studies

- Association exists over a narrow region; markers must be close to disease gene
  - The basic concept is linkage disequilibrium (LD)
- Used for candidate genes or in linked regions
- Can use population-based (unrelated cases) or family-based design



## The Future of Genetic Studies of Complex Human Diseases

Neil Risch and Kathleen Merikangas


SCIENCE • VOL. 273 • 13 SEPTEMBER 1996

## 1.c Genome wide association analyses (GWAs)

- Genome-wide association analysis is an approach that involves rapidly scanning markers across the complete sets of DNA, or genomes, of many people to find genetic variations associated with a particular disease.
- Once new genetic associations are identified, researchers can use the information to develop better strategies to detect, treat and prevent the disease.
- Such studies are particularly useful in finding genetic variations that contribute to common, complex diseases, such as asthma, cancer, diabetes, heart disease and mental illnesses.

(<http://www.genome.gov/pfv.cfm?pageID=20019523>)

U.S. Department of Health & Human Services
www.hhs.gov



[Contact Us](#) | [Print Version](#)  
 Search:    
[Advanced Search](#) | [Site Map](#)

[Home](#) | [About Grants](#) | [Funding](#) | [Forms & Deadlines](#) | [Grants Policy](#) | [News & Events](#) | [About OER](#) | [NIH Home](#)

**Funding Opportunities**

[Funding Opportunities \(RFAs, PAs\) & Notices](#)

[Unsolicited Applications \(Parent Announcements\)](#)

[Research Training & Career Development](#)

[Small Business \(SBIR/STTR\)](#)

[Contract Opportunities](#)

**NIH-Wide Initiatives**

[Stem Cell Information](#)

[New and Early Stage Investigators](#)

[Genome-Wide Association Studies \(GWAS\)](#)

[NIH Roadmap for Medical Research](#)

**Global OER Resources**

[Glossary & Acronyms](#)

## Genome-Wide Association Studies (GWAS)

The NIH is interested in advancing genome-wide association studies (GWAS) to identify common genetic factors that influence health and disease. For the purposes of this policy, a genome-wide association study is defined as any study of genetic variation across the entire human genome that is designed to identify genetic associations with observable traits (such as blood pressure or weight), or the presence or absence of a disease or condition. Whole genome information, when combined with clinical and other phenotype data, offers the potential for increased understanding of basic biological processes affecting human health, improvement in the prediction of disease and patient care, and ultimately the realization of the promise of personalized medicine. In addition, rapid advances in understanding the patterns of human genetic variation and maturing high-throughput, cost-effective methods for genotyping are providing powerful research tools for identifying genetic variants that contribute to health and disease. The purpose of this Website is to support the implementation of the GWAS Policy.

The NIH will continue to release additional guidance information on this site. Please e-mail [GWAS@mail.nih.gov](mailto:GWAS@mail.nih.gov) with any questions.


### Recent News

- [NIH Background Fact Sheet on GWAS Policy Update](#) - (08/28/2008) (PDF - 40 KB)
- [NIH Modifications to Genome-Wide Association Studies \(GWAS\) Data Access](#) - (08/28/2008) (PDF - 43 KB)

### Data Access Information

- [Senior Oversight Committee \(SOC\) Charge and Roster](#) - (07/10/2008) (PDF - 103 KB)
- [Data Access Committees \(DACs\) Charge and Roster](#) - (07/10/2008) (PDF - 50 KB)

This website wants to run the following add-on: 'Adobe Flash Player' from 'Adobe Systems Incorporated'. If you trust the website and the add-on and want to allow it to run, click here...



genome.gov  
National Human Genome Research Institute  
National Institutes of Health

Google™ Search  Search

Home | About NHGRI | Newsroom | Staff

Research Grants Health Policy & Ethics Educational Resources Careers & Training

Home > Educational Resources > Fact Sheets > **Genome-Wide Association Studies**

Share this page Print

### See Also:

Genome-Wide Association Studies for the Rest of Us: Adding Genome-Wide Association to Population Studies  
Boston, Mass.  
June 22, 2007

## Genome-Wide Association Studies

- [What is a genome-wide association study?](#)
- [Why are such studies possible now?](#)
- [How will genome-wide association studies benefit human health?](#)
- [What have genome-wide association studies found?](#)
- [How are genome-wide association studies conducted?](#)
- [How can researchers access data from genome-wide association studies?](#)
- [What is NIH doing to support genome-wide association studies?](#)

### What is a genome-wide association study?

A genome-wide association study is an approach that involves rapidly scanning markers across the complete sets of DNA, or genomes, of many people to find genetic variations associated with a particular disease. Once new genetic associations are identified, researchers can use the information to develop better strategies to detect, treat and prevent the disease. Such studies are particularly useful in finding genetic variations that contribute to common, complex diseases, such as asthma, cancer, diabetes, heart disease and mental illnesses.

Top of page

### Why are such studies possible now?

With the completion of the Human Genome Project in 2003 and the International HapMap Project in 2005, researchers now have a set of research tools that make it possible to find the genetic contributions to common diseases. The tools include computerized databases that contain the reference human genome sequence, a map of human genetic variation and a set of new technologies that can quickly and



genome.gov  
National Human Genome Research Institute  
National Institutes of Health

SEARCH

Research Funding
Research at NHGRI
Health
Education
Issues in Genetics
Newsroom
Careers & Training
About
For You

[Home](#) > [About](#) > [Organizational Structure](#) > [About the Office of the Director](#) > [Office of Population Genomics](#) > [A Catalog of Published Genome-Wide Association Studies](#)

## Office of Population Genomics

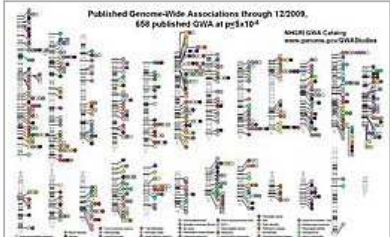
[Overview](#) | [A Catalog of Genome-Wide Association Studies](#) | [Research Programs](#) | [Publications](#) | [Meetings & Workshops](#) | [Notices & Funding Opportunities](#) | [Staff Biographies](#) | [Contact](#)

### A Catalog of Published Genome-Wide Association Studies

[Potential etiologic and functional implications of genome-wide association loci for human diseases and traits](#)

Click here to read our recent *Proceedings of the Academy of Sciences (PNAS)* article on catalog methods and analysis.

[Go to the Catalog](#)



**The genome-wide association study (GWAS) publications listed here** include only those attempting to assay at least 100,000 single nucleotide polymorphisms (SNPs) in the initial stage. Publications are organized from most to least recent date of publication, indexing from online publication if available. Studies focusing only on candidate genes are excluded from this catalog. Studies are identified through weekly PubMed literature searches, daily NIH-distributed compilations of news and media reports, and occasional comparisons with an existing database of GWAS literature ([HuGE Navigator](#)).

SNP-trait associations listed here are limited to those with p-values <  $1.0 \times 10^{-5}$  (see full methods for additional details). Multipliers of powers of 10 in p-values are rounded to the nearest single digit; odds ratios and allele frequencies are rounded to two decimals. Standard errors are converted to 95 percent confidence intervals where applicable. Allele frequencies, p-values, and odds ratios derived from the largest sample size, typically a combined analysis (initial plus replication studies), are recorded below if reported; otherwise statistics from the initial study sample are recorded. For quantitative traits, information on % variance explained, SD increment, or unit difference is reported where available. Odds ratios < 1 in the original paper are converted to OR > 1 for the alternate allele. Where results from multiple genetic

Share this page

Keywords: [GWAS](#)

- > [GWAS](#)
- > [GWAS studies](#)
- > [OPG](#)
- > [Teri Manolio](#)
- > [RSS Web Feeds](#)

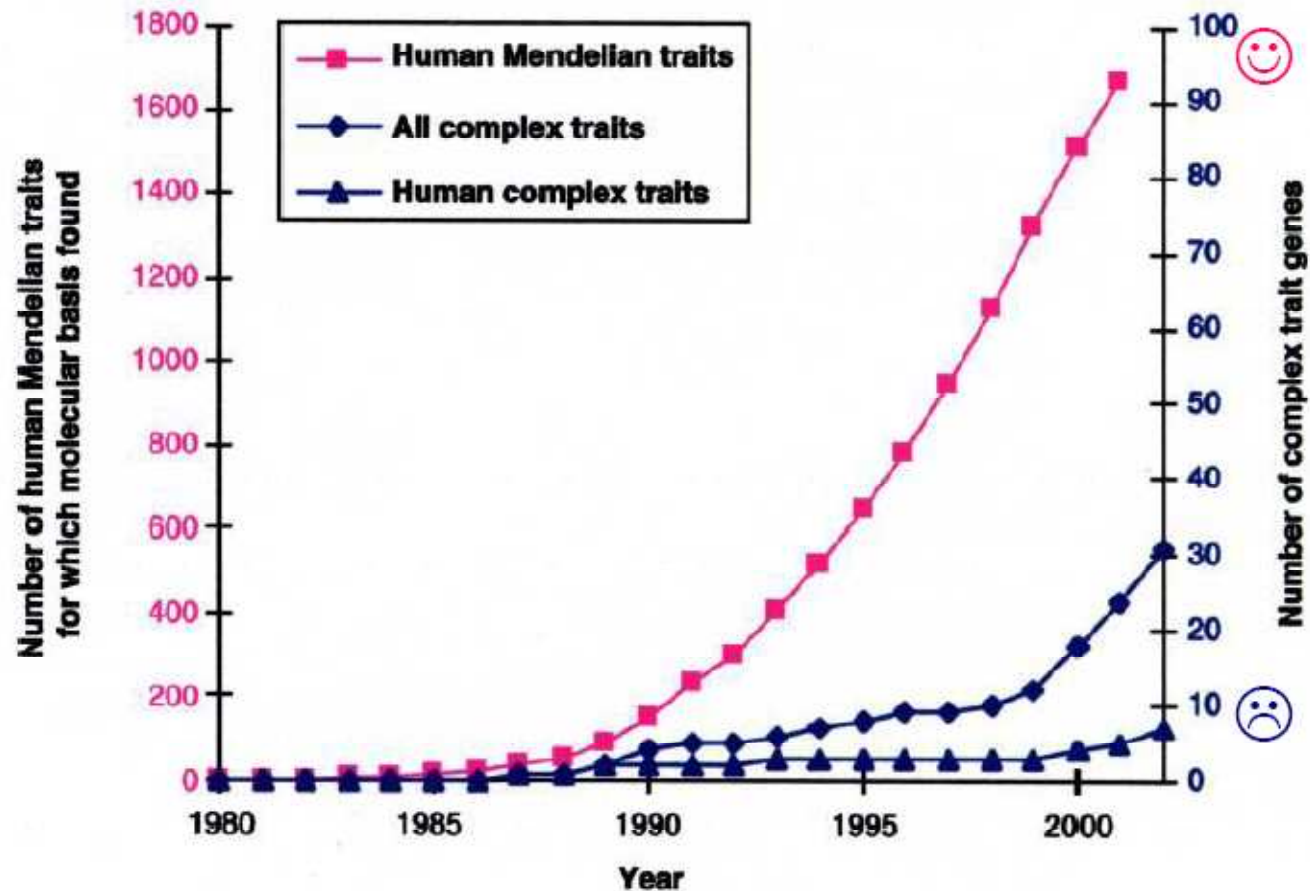
(<http://www.genome.gov/26525384>)

## Genome wide association analyses

- GWAs have become possible with the completion of the Human Genome Project in 2003 and the International HapMap Project in 2005. Hence researchers have a set of research tools that make it possible to find the genetic contributions to common diseases.
- The tools include
  - computerized databases that contain the reference human genome sequence,
  - a map of human genetic variation and
  - a set of new technologies that can quickly and accurately analyze whole-genome samples for genetic variations that contribute to the onset of a disease.

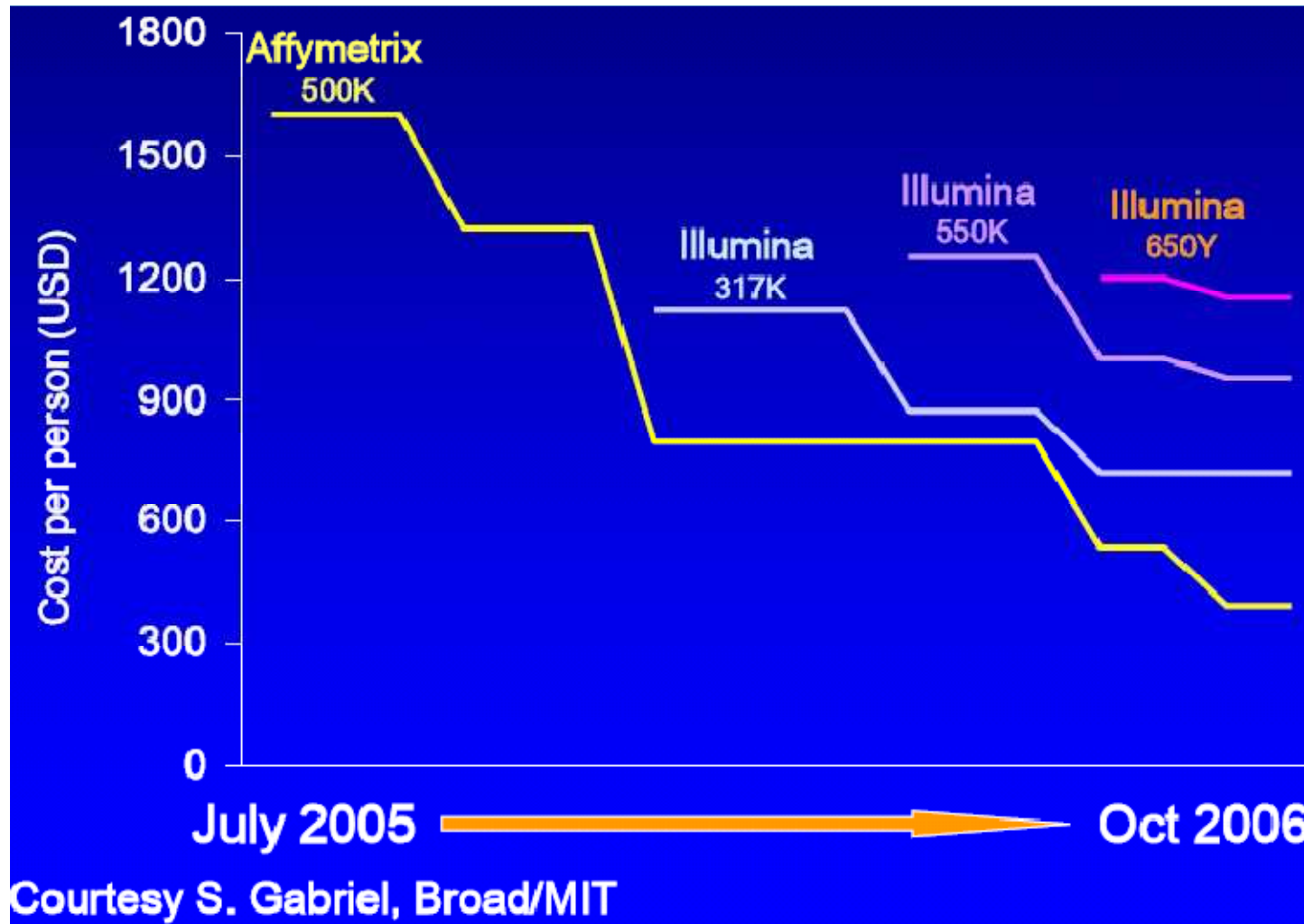
(<http://www.genome.gov/pfv.cfm?pageID=20019523>)

## GWAs: historical evolution of their struggle and success



(Glazier et al 2002)

## GWAs: historical evolution of their struggle and success



## 2007: a turning point (Pennisi 2007)

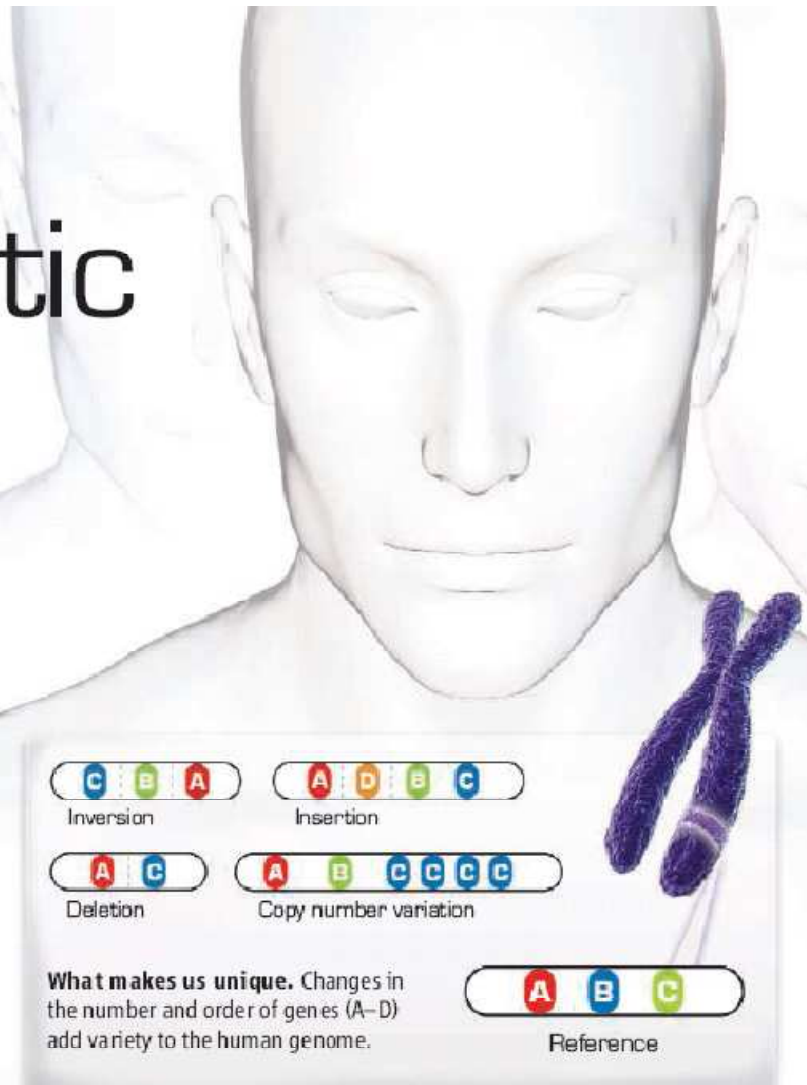
### BREAKTHROUGH OF THE YEAR

# Human Genetic Variation

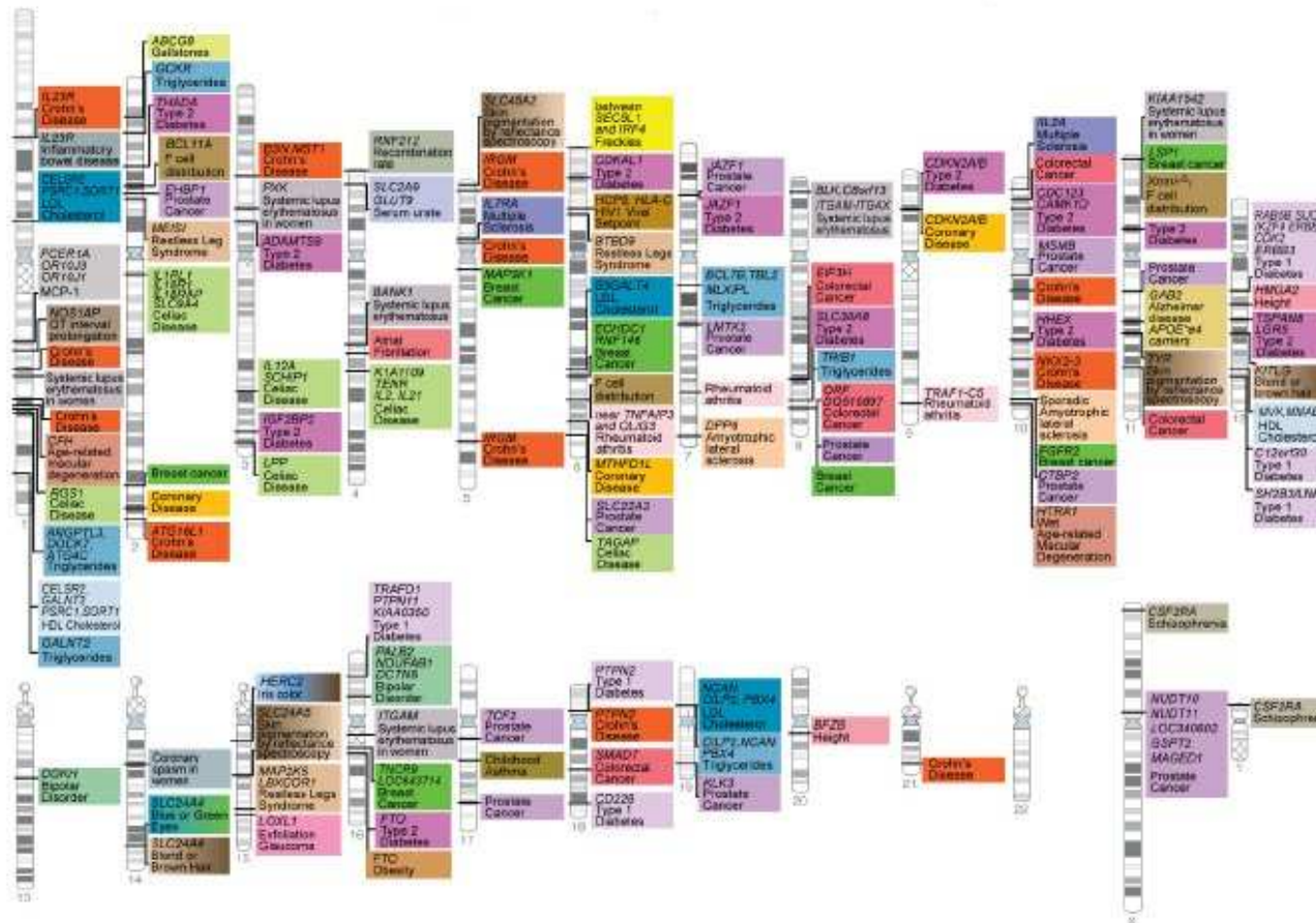
Equipped with faster, cheaper technologies for sequencing DNA and assessing variation in genomes on scales ranging from one to millions of bases, researchers are finding out how truly different we are from one another

THE UNVEILING OF THE HUMAN GENOME ALMOST 7 YEARS AGO cast the first faint light on our complete genetic makeup. Since then, each new genome sequenced and each new individual studied has illuminated our genomic landscape in ever more detail. In 2007, researchers came to appreciate the extent to which our genomes differ from person to person and the implications of this variation for deciphering the genetics of complex diseases and personal traits.

Less than a year ago, the big news was triangulating variation between us and our primate cousins to get a better handle on genetic changes along the evolutionary tree that led to humans. Now, we have moved from asking what in our DNA makes us human to striving to



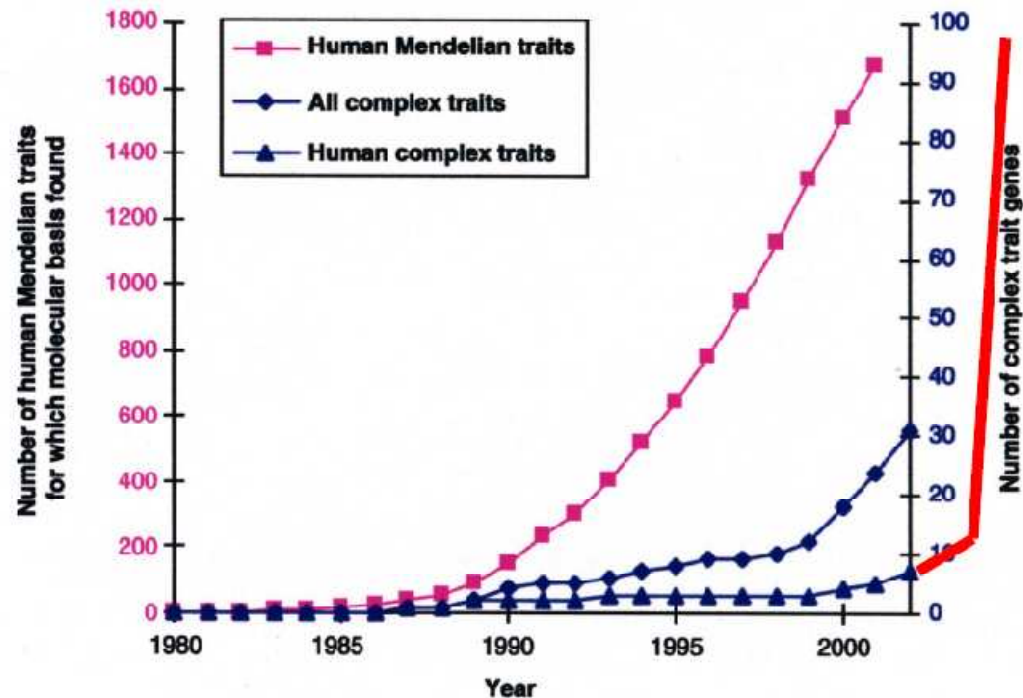
# 2007: a turning point (nearly 100 loci, 40 common diseases/traits)



(Manolio et al 2008 – first quarter 2008)

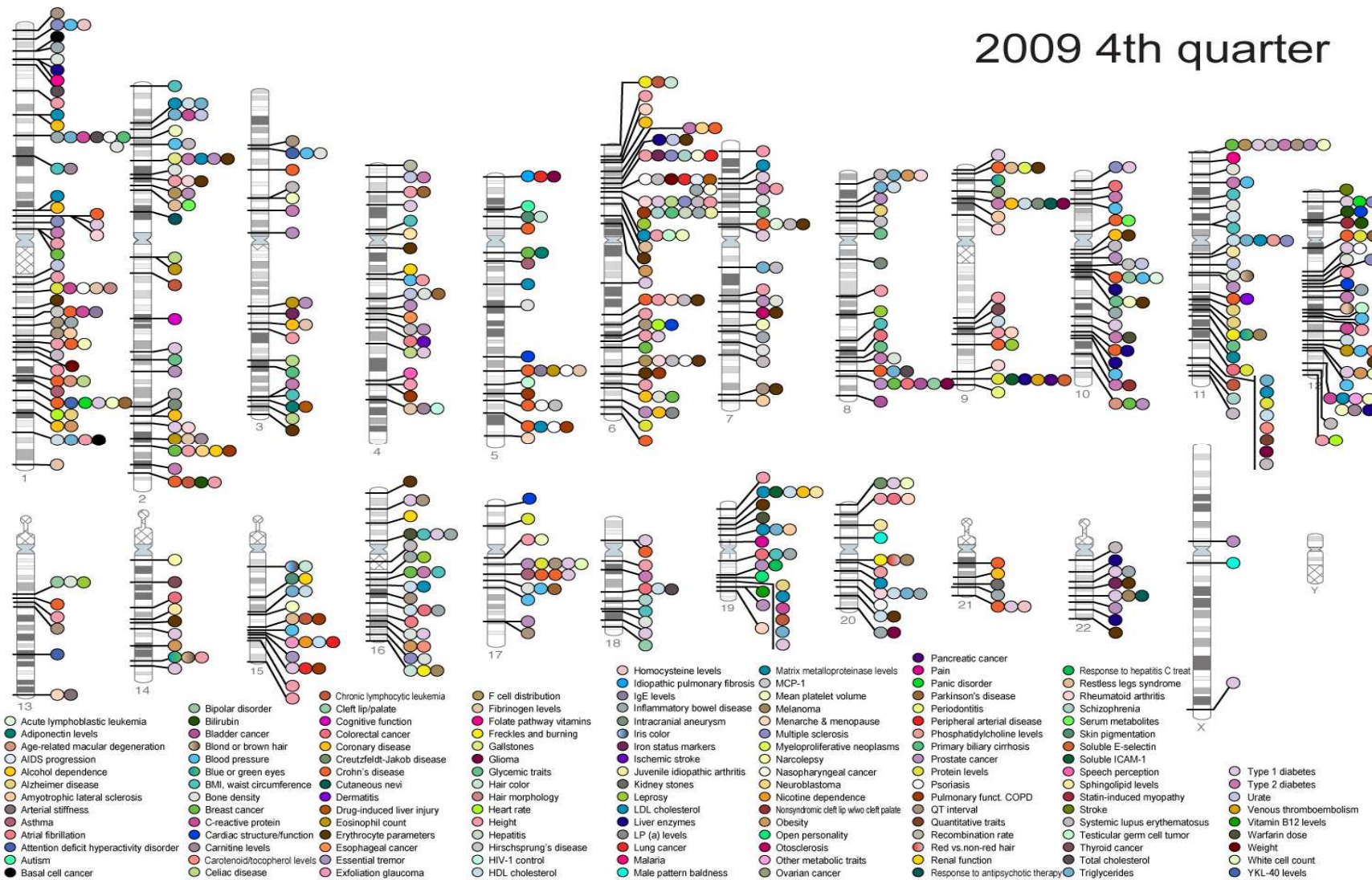
## 2007: a turning point

- By the end of March 2009, more than 90 diseases and traits have been identified with published GWA results ... (Feero 2009)



(Glazier et al 2002)

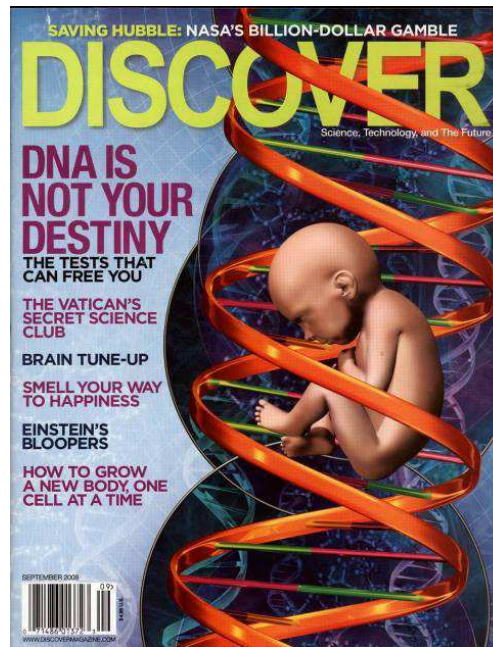
2009 4th quarter





## Reasons for continuing popularity of GWAs

- The impact on medical care from genome-wide association studies could potentially be substantial. Such research is laying the groundwork for the era of personalized medicine, in which the current one size-fits-all approach to medical care will give way to more customized strategies.



## ... It will take more than SNPs alone

---

### Genetic Risk Prediction — Are We There Yet?

Peter Kraft, Ph.D., and David J. Hunter, M.B., B.S., Sc.D., M.P.H.

A major goal of the Human Genome Project was to facilitate the identification of inherited genetic variants that increase or decrease the risk of complex diseases. The completion of the International HapMap Project and the development of new methods for genotyping individual DNA samples at 500,000 or more loci have led to a wave of discoveries through genomewide association studies. These analyses have identified common genetic variants that are associated with the risk of more than 40 diseases and human phenotypes. Several companies have begun offering direct-to-consumer testing that uses the

tests of genetic predisposition to important diseases would have major clinical, social, and economic ramifications. But the great majority of the newly identified risk-marker alleles confer very small relative risks, ranging from 1.1 to 1.5,<sup>2</sup> even though such analyses meet stringent statistical criteria (i.e., the identification of associations with disease that have very small P values and hence are unlikely to be false positives). However, even when alleles that are associated with a modest increase in risk are combined, they generally have low discriminatory and predictive ability.<sup>3</sup>

One argument in favor of us-

est relative risks are almost certainly overrepresented in the first wave of findings from genomewide association studies, since considerations of statistical power predict that they will be identified first. However, a striking fact about these first findings is that they collectively explain only a very small proportion of the underlying genetic contribution to most studied diseases. (Some exceptions exist — notably, age-related macular degeneration, for which a few alleles explain a substantial fraction of the genetic contribution.) Several lines of evidence support this overall conclusion.

(Kraft and Hunter 2009)

## ... It will take more than SNPs alone

### PERSPECTIVES

#### GENETICS

## Getting Closer to the Whole Picture

Uwe Sauer, Matthias Heinemann, Nicola Zamboni

A major challenge of biology is to unravel the organization and interactions of cellular networks that enable complex processes such as the biochemistry of growth or cell division. The underlying complexity arises from intertwined nonlinear and dynamic interactions among large numbers of cellular constituents, such as genes, proteins, and metabolites. As well, interactions among these components vary in nature (regulatory, structural, and catalytic), effect, and strength. The reductionist approach has successfully identified most of the components and many interactions but, unfortunately, offers no convincing concepts and methods to comprehend how system properties emerge. To understand how and why cells function the way they do, comprehensive and quantitative data on component concentrations are required to quantify component interactions. On page 593 of this issue, Ishii *et al.* (1) provide unsurpassed complete and quantitative data of components at the various constituent levels in a bacterial cell.

better addressed by observing, through quantitative measures, multiple components simultaneously, and by rigorous data integration with mathematical models (2). Such a systemwide perspective (so-called systems biology) on component interactions is required so that network properties, such as a particular functional state or robustness (3), can be quantitatively understood and rationally manipulated.

The technical challenges of the systems biological approach (4) are mainly along four lines (see the figure): (i) systemwide component identification and quantification ("omics" data) at the level of mRNA, proteins, and small molecular weight metabolites; (ii) experimental identification of physical component interactions, primarily for information processing networks; (iii) computational inference of structure, type, and quantity of component interactions from data; and (iv) rigorous integration of heterogeneous data. The last step is required to

A quantitative data set of RNA, proteins, and metabolites provides an unprecedented starting point to understand, at a systems level, the effects of perturbations on a cell.

ods relating to the first challenge has made tremendous advances in the past decade, but the level of sophistication and the associated costs have led to a situation where primarily single-component data—that is, data solely on genes, proteins, or metabolites—are available. Until the study by Ishii *et al.*, at best two different types of component data were reported for a given experiment, which severely limited progression along the iterative cycle between experiments and theory.

By joining forces among specialized labs, Ishii *et al.* report systemwide data on three main component layers of cells—transcriptome (mRNA), proteome (protein), and metabolome (metabolites)—with a particular focus on central carbon metabolism of the model bacterium *Escherichia coli*. Beyond component concentrations, the functional endpoint of gene, protein, and metabolite interactions—the intracellular metabolic fluxes—were quantified from <sup>13</sup>C-labeling experiments (5). In a laborious procedure,

(Sauer et al 2007)

## Reasons for continuing popularity of GWAs using SNPs

- There is a large compendium of validated SNP data
- SNP GWAs are able to potentially use *all* of the data
- They are more powerful for genes of small to moderate effect (see before)
- They allow for covariate assessment, detection of interactions, estimation of effect size, ...

BUT

**ALL statistical issues cannot be ruled out**

## PERSPECTIVE

DRINKING FROM THE FIRE HOSE — STATISTICAL ISSUES IN GENOMEWIDE ASSOCIATION STUDIES

## STATISTICS AND MEDICINE

## Drinking from the Fire Hose — Statistical Issues in Genomewide Association Studies

David J. Hunter, M.B., B.S., and Peter Kraft, Ph.D.

Related article, page 443

The past 3 months have seen the publication of a series of studies examining the inherited genetic underpinnings of common diseases such as prostate cancer, breast cancer, diabetes, and in this issue of the *Journal*, coronary artery disease (reported by Samani et al., pages 443–453). These genomewide association studies have been able to examine interpatient differences in inherited genetic variability at an unprecedented level of resolution, thanks to the development of microarrays, or chips, capable of assessing more than 500,000 single-

ating the need for guessing which genes are likely to harbor variants affecting risk. Most of the robust associations seen in this type of study have not been with genes previously suspected of being related to the disease. Some of these associations have been found in regions not even known to harbor genes, such as the 8q24 region, in which multiple variants have been found to be associated with prostate cancer.<sup>2</sup> Such findings promise to open up new avenues of research, through both the discovery of new genes relevant to specific diseases and the

The main problem with this strategy is that, because of the high cost of SNP chips, most studies are somewhat constrained in terms of the number of samples and thus have limited power to generate P values as small as  $10^{-7}$ . In addition, most variants identified recently have been associated with modest relative risks (e.g., 1.3 for heterozygotes and 1.6 for homozygotes), and many true associations are not likely to exceed P values as extreme as  $10^{-7}$  in an initial study. On the other hand, a “statistically significant” finding in an underpowered study is more

(Hunter and Kraft 2007)

## Using all of the data for case/control designs?

candidate gene approach

vs

genome-wide screening approach



## Using all of the data for case/control designs ?

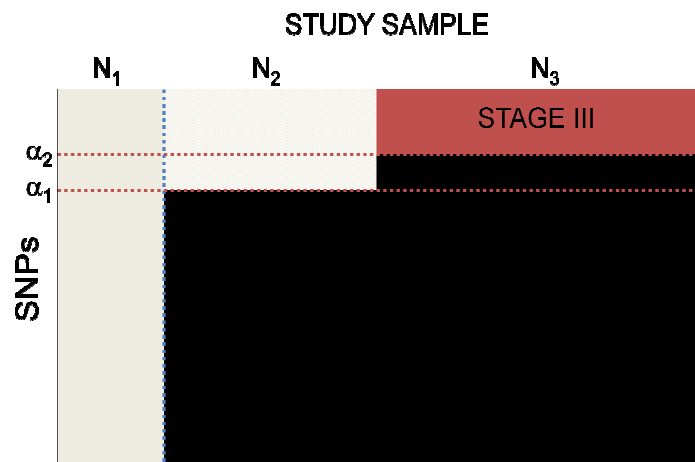
- There are many (single locus) tests to perform
- The multiplicity can be dealt with in several ways
  - clever multiple corrective procedures (see later)
  - adopting multi-locus tests (see later) or
  - haplotype tests,
  - pre-screening strategies (see later), or
  - multi-stage designs.

Which of these approaches are more powerful is  
still under heavy debate...

## Using all of the data ?

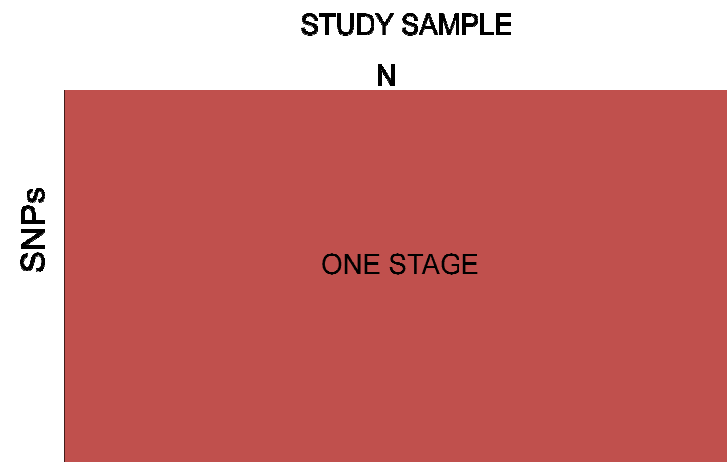
### Multi-stage

- Less expensive
- More complicated
- Less powerful



### Single-stage

- More expensive
- Less complicated
- More powerful



(slide: courtesy of McQueen)



## 2 Families versus unrelated cases and controls

### 2.a Every design has statistical implications

There are many possible designs for a genetic association study

	Details	Advantages	Disadvantages	Statistical analysis method
Cross-sectional	Genotype and phenotype (ie, note disease status or quantitative trait value) a random sample from population	Inexpensive. Provides estimate of disease prevalence	Few affected individuals if disease rare	Logistic regression, $\chi^2$ tests of association or linear regression
Cohort	Genotype subsection of population and follow disease incidence for specified time period	Provides estimate of disease incidence	Expensive to follow-up. Issues with drop-out	Survival analysis methods
Case-control	Genotype specified number of affected (case) and unaffected (control) individuals. Cases usually obtained from family practitioners or disease registries, controls obtained from random population sample or convenience sample	No need for follow-up. Provides estimates of exposure effects	Requires careful selection of controls. Potential for confounding (eg, population stratification)	Logistic regression, $\chi^2$ tests of association
Extreme values	Genotype individuals with extreme (high or low) values of a quantitative trait, as established from initial cross-sectional or cohort sample	Genotype only most informative individuals hence save on genotyping costs	No estimate of true genetic effect sizes	Linear regression, non-parametric, or permutation approaches
Case-parent triads	Genotype affected individuals plus their parents (affected individuals determined from initial cross-sectional, cohort, or disease-outcome based sample)	Robust to population stratification. Can estimate maternal and imprinting effects	Less powerful than case-control design	Transmission/disequilibrium test, conditional logistic regression or log-linear models
Case-parent-grandparent septets	Genotype affected individuals plus their parents and grandparents	Robust to population stratification. Can estimate maternal and imprinting effects	Grandparents rarely available	Log-linear models
General pedigrees	Genotype random sample or disease-outcome based sample of families from general population. Phenotype for disease trait or quantitative trait	Higher power with large families. Sample may already exist from linkage studies	Expensive to genotype. Many missing individuals	Pedigree disequilibrium test, family-based association test, quantitative transmission/disequilibrium test
Case-only	Genotype only affected individuals, obtained from initial cross-sectional, cohort, or disease-outcome based sample	Most powerful design for detection of interaction effects	Can only estimate interaction effects. Very sensitive to population stratification	Logistic regression, $\chi^2$ tests of association
DNA-pooling	Applies to variety of above designs, but genotyping is of pools of anywhere between two and 100 individuals, rather than on an individual basis	Potentially inexpensive compared with individual genotyping (but technology still under development)	Hard to estimate different experimental sources of variance	Estimation of components of variance

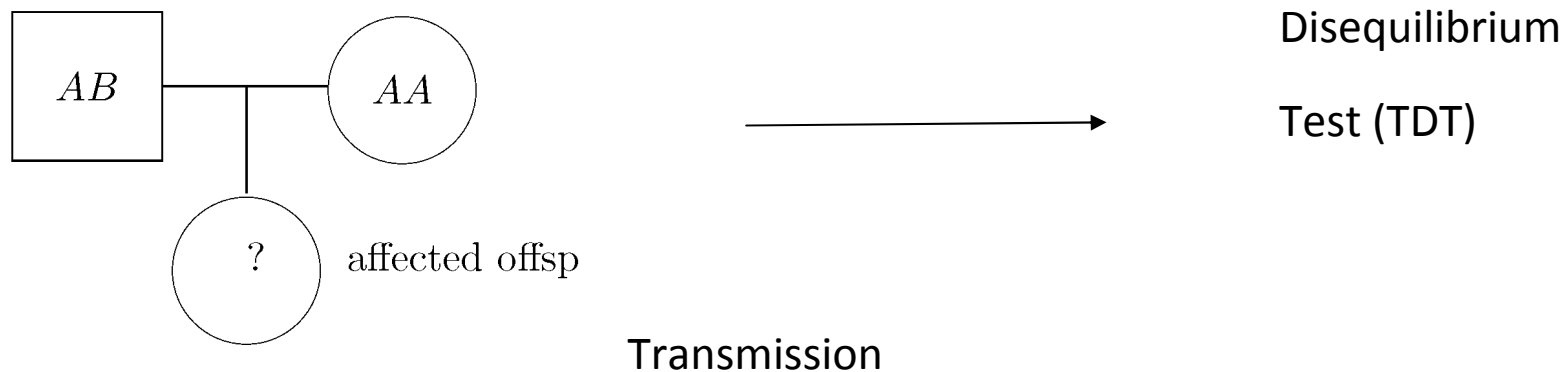
Table 2: Study designs for genetic association studies

(Cordell and Clayton, 2005)

## Family-based designs

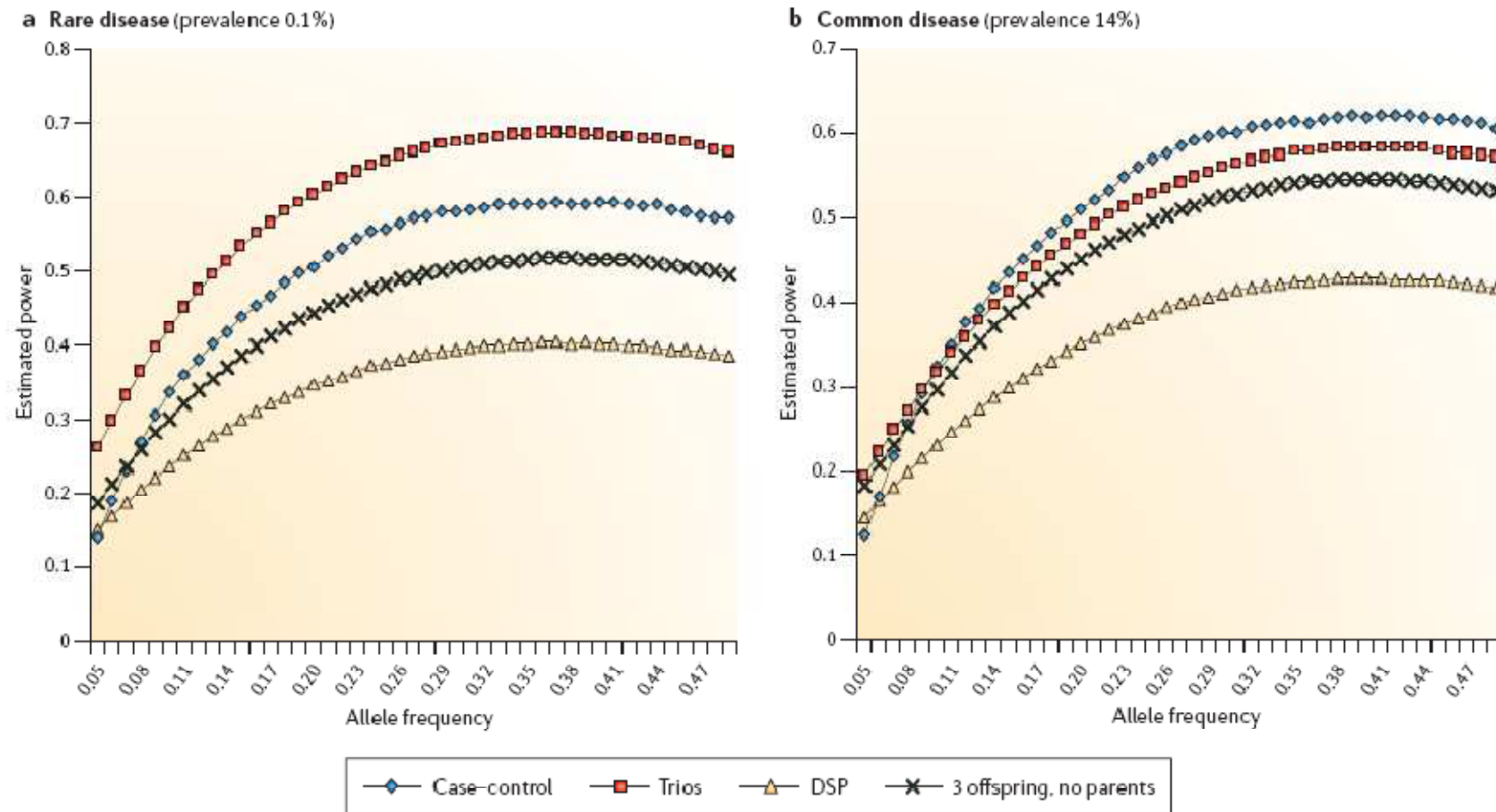
- Cases and their parents
- Test for both linkage and association
- Robust to population substructure: admixture, stratification, failure of HWE
- Offer a unique approach to handle multiple comparisons

## Using trios

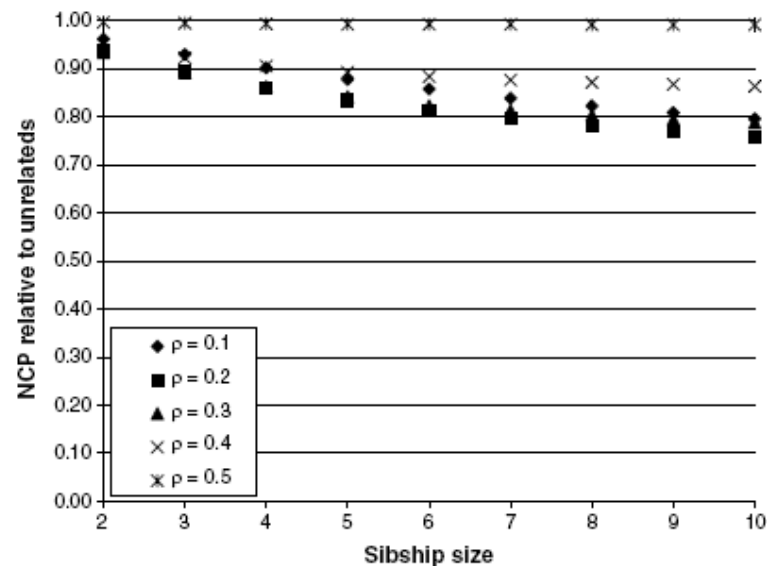


## 2.b Power considerations

### Rare versus common diseases (Lange and Laird 2006)



## Power



**Figure 1** Relative power of GWAS for sibships versus unrelated individuals, for the same cost of genotyping.  $\rho$  is the phenotypic correlation between siblings.

- Little power lost by analysing families relative to singletons
- It may be efficient to genotype only some individuals in larger pedigrees
- Pedigrees allow error checking, within family tests, parent-of-origin analyses, joint linkage and association, ...

(Visscher et al 2008)

## Power of GWAs (whether or not using related individuals)

- Critical to success is the development of robust study designs to ensure high power to detect genes of modest risk while minimizing the potential of false association signals due to testing large numbers of markers.
- Key components include
  - sufficient sample sizes,
  - rigorous phenotypes,
  - comprehensive maps,
  - accurate high-throughput genotyping technologies,
  - sophisticated IT infrastructure,
  - rapid algorithms for data analysis, and
  - rigorous assessment of genome-wide signatures.

## The role of population resources

- Critical to success is the collection of sufficient numbers of rigorously phenotyped cases and matched control groups or family trios to have sufficient power to detect disease genes conferring modest risk.
- Power studies have shown that at least 2,000 to 5,000 samples for both cases and controls groups are required when using general populations.
- This large number of samples makes the collection of rigorously consistent clinical phenotypes across all cases quite challenging.
- In addition, matching of cases and controls with respect to geographic origin and ethnicity is critical for minimizing false positive signals due to population substructure (especially when non-family specific tests are used).

## The role of SNP Maps and Genotyping

- A second key success factor is having a comprehensive map of hundreds of thousands of carefully selected SNPs.
- Currently there are several groups offering SNP arrays for genotyping, with **Affymetrix** ([www.affymetrix.com](http://www.affymetrix.com)) and **Illumina** ([www.illumina.com](http://www.illumina.com)) both providing products containing more than 500,000 SNPs.
- Achieving high call rates and genotyping accuracy are also critically important, because small decreases in accuracy or increases in missing data can result in relatively large decreases in the power to detect disease genes.

([http://www.genengnews.com/articles/chitem\\_print.aspx?aid=1970&chid=0](http://www.genengnews.com/articles/chitem_print.aspx?aid=1970&chid=0))

## The role of IT and Analytic Tools

- Genotyping instruments now have sufficient capacity to enable genotyping of thousands of subjects in only a few weeks.
- A study of 1,000 cases and 1,000 control subjects using a 550,000 SNP array produces over 1 billion genotypes.
- To properly store, manage, and process the enormous data sets arising from GWAS, a highly sophisticated IT infrastructure is needed, including computing clusters with sufficient CPUs and automated, robust pipelines for rapid data analysis.
- Given this wealth of genotypic data, the availability of efficient analytical tools for performing association analyses is critical to the successful identification of disease-associated signals.

([http://www.genengnews.com/articles/chitem\\_print.aspx?aid=1970&chid=0](http://www.genengnews.com/articles/chitem_print.aspx?aid=1970&chid=0))



## The role of IT and Analytic Tools

- Primary genome-wide analyses include a comparison of allele and genotype frequencies between case and control cohorts or for child-affected trios, a comparison of the frequencies of transmitted (case) and non-transmitted (control) alleles.
- An alternative test of association when using child-affected trios is the transmission disequilibrium test for the overtransmission of alleles to affected offspring (see next section).
- Since these analyses require considerable computing power to handle terabytes of data, genome-wide analyses are often limited to single SNPs with haplotype analyses performed once candidate regions are identified.
- But the field is changing ... STAY TUNED !!!

([http://www.genengnews.com/articles/chitem\\_print.aspx?aid=1970&chid=0](http://www.genengnews.com/articles/chitem_print.aspx?aid=1970&chid=0))

## Software

- With recent technical advances in high-throughput genotyping technologies the possibility of performing GWAs becomes increasingly feasible for a growing number of researchers.
- A number of packages are available in the R Environment to facilitate the analysis of these large data sets.
  - **GenABEL** is designed for the efficient storage and handling of GWAS data with fast analysis tools for quality control, association with binary and quantitative traits, as well as tools for visualizing results.
  - **pbatR** provides a GUI to the powerful PBAT software which performs family and population based family and population based studies. The software has been implemented to take advantage of parallel processing, which vastly reduces the computational time required for GWAS.

## Software

- A number of packages are available in the R Environment to facilitate the analysis of these large data sets.
  - **SNPassoc** is another package for carrying out GWAS analysis. It offers descriptive statistics of the data (including patterns of missing data) and tests for Hardy-Weinberg equilibrium. Single-point analyses with binary or quantitative traits are implemented via generalized linear models, and multiple SNPs can be analysed for haplotypic associations or epistasis. It is also possible to perform interaction analyses.
- Check out Zhang 2008: R Packages for Genome-Wide association Studies

## 2.c The Transmission Disequilibrium Test

- Assumptions:
  - Parents' and offspring genotypes known
  - dichotomous phenotype, only affected offspring
- Count transmissions from heterozygote parents, compare to expected transmissions
- Expected computed using parents' genotypes and Mendel's laws of segregation (differ from case-control)
  - Conditional test on offspring affection status and parents' genotypes
- Special case of McNemar's test (columns: alleles not transmitted; rows: alleles transmitted)

(Spielman et al 1993)

## Recall for binary outcomes

	Case exposed	
	No	Yes
Control exposed	<i>a</i>	<i>b</i>
Yes	<i>c</i>	<i>d</i>

- For a single binary exposure, the relevant data may be presented in the table above, which counts sets not subjects.
- Estimation of odds ratio:

$$\hat{\theta} = \frac{b}{c}, \quad SE(\log \hat{\theta}) = \sqrt{\frac{1}{b} + \frac{1}{c}}$$

## McNemar's test

- Score test of the null hypothesis,  $\theta = 1$

$$U = b - \frac{b + c}{2} = \frac{b - c}{2},$$

$$V = \frac{b + c}{4}$$

- $\frac{U^2}{V} = \frac{(b-c)^2}{b+c}$  is distributed as chi-square (1 df) in large samples
- This test discards concordant pairs and tests whether discordant sets split equally between those with case exposed and those with control exposed
- McNemar's test is a special case of the Mantel-Haenszel test

## Attraction of TDT

- $H_0$  relies on Mendel's laws, not on control group
- $H_A$  linkage disequilibrium is present: DSL and marker loci are linked, and their alleles are associated

- Intuition:

If no linkage but association at population level, no systematic transmission of a particular allele. If linkage, but no association, different alleles will be transmitted in different families.

- Consequence:

TDT is robust to population stratification, admixture, other forms of confounding

(model free). The same properties hold for FBAT statistics of which the TDT is a special case.

Human  
Heredit

---

### Advances in Family-Based Association Analysis

15 Years of Practical Experience with the  
Original Transmission Disequilibrium Test

---

Guest Editor  
Derek Gordon, Piscataway, N.J.

13 figures and 19 tables, 2008

(Spielman et al 1993)

## Disadvantages of TDT

- Only affected offspring
- Only dichotomous phenotypes
- Biallelic markers
- Single genetic model (additive)
- No allowance for missing parents/pedigrees
- Method for incorporating siblings is limited
- Does not address multiple markers or multiple phenotypes



## Generalization of the TDT

Need for a unified framework that flexible enough to encompass:

- standard genetic models
- other phenotypes, multiple phenotypes
- multiple alleles
- additional siblings; extended pedigrees
- missing parents
- multiple markers
- haplotypes

(Horvath et al 1998, 2001; Laird et al 2000, Lange et al 2004)

## 2.d FBAT test statistic

$T$ : code trait, based on phenotype  $Y$  and offset  $\mu$

$X$ : code genotype (harbors genetic inheritance model)

$P$ : parental genotypes

$$U = \sum T(X - E(X|P))$$
$$U = \sum (Y - \mu)(X - E(X|P))$$

- $\sum$  is *sum* over all offspring ,
- $E(X|P)$  is the expected marker score computed under  $H_0$ , conditional on  $P$
- $Var(U) = \sum T^2 Var(X|P)$
- $Var(X|P)$  computed from offspring distribution, conditional on  $P$  and  $T$ .

## FBAT test statistic

$$Z = U / \sqrt{\text{Var}(U)}$$

- Asymptotic distributions
  - $Z \sim N(0,1)$  under  $H_0$
  - $Z^2 \sim \chi^2$  on 1 df under  $H_0$
- $Z^2_{FBAT} = \chi^2_{TDT}$  when
  - $Y=1$  if child is affected
  - $T=Y$
  - $X$  follows an additive coding
  - no missing data

(Horvath et al 1998, 2001; Laird et al 2000)

## General theory on FBAT testing

- Test statistic:

- works for any phenotype, genetic model
- use covariance between offspring trait and genotype

$$U = \sum (Y - \mu)(X - E(X|P))$$

- Test Distribution:

- computed assuming  $H_0$  true; random variable is offspring genotype
- condition on parental genotypes when available, extend to family configurations (avoid specification of allele distribution)
- condition on offspring phenotypes (avoid specification of trait distribution)

(Horvath et al 1998, 2001; Laird et al 2000)

## Key features of TDT are maintained

- Random variable in the analysis is the offspring genotype
- Parental genotypes are fixed (condition on the parental genotypes)
- Trait is fixed (condition on all offspring being affected)

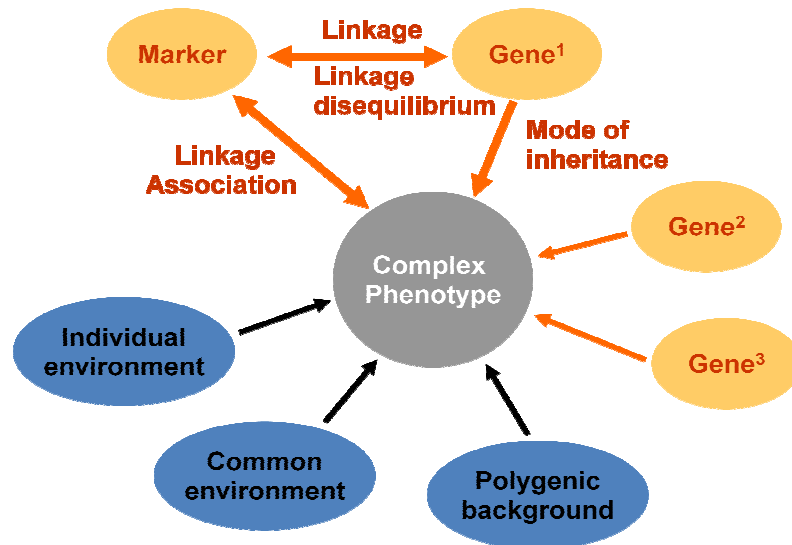
## Missing genotypes revisited

- You may think otherwise, but there ARE some additional advantages to impute missing marker data ... Which ones?
- IMPUTE is a software package that is able to rapidly impute missing data in unrelated individuals.
- Note that an imputation process generally becomes more complicated when genotypes need to be imputed in studies of related individuals.
- Two important packages that allow for proper genotype imputation in family-based designs include MERLIN and MENDEL
- The latest developments can be retrieved from Gonçalo Abecasis or Jonathan Marchini
  - <http://www.sph.umich.edu/csg/abecasis/>
  - <http://www.stats.ox.ac.uk/~marchini/>

(Li et al 2009)

## 3 From complex phenomena to models

### 3.a Introduction



(Weiss and Terwilliger 2000)

(Moore 2008)

- There are likely to be *many* susceptibility genes each with combinations of *rare and common* alleles and genotypes that impact disease susceptibility primarily through *nonlinear interactions* with *genetic and environmental* factors
- Analytically, it can be difficult to distinguish between *interactions* and *heterogeneity*.

## 3.b When the number of tests grows

### Multiple testing revisited

- Multiple testing is a thorny issue, the bane of statistical genetics.
  - The problem is not really the number of tests that are carried out: even if a researcher only tests one SNP for one phenotype, if many other researchers do the same and the nominally significant associations are reported, there will be a problem of false positives.

(Balding 2006)



## Multiple testing (continued)

- With too many SNPs, the classical corrective measures still perform adequately?
  - Family-wise error rate (FWER)
    - Bonferroni Threshold:  $< 10^{-7}$
  - Permutation data sets
    - Enough compute capacity?
  - False discovery rate (FDR) and variations thereof
    - it starts to break down
    - the power over Bonferroni is minimal
  - Bayesian methods such as false-positive report probability (FPRP)
    - Could work but for now not yet well documented
    - What are the priors?

### 3.c When the number of SNPs grows

#### **Variable selection** (reduces multiple testing burden)

- Pre-screening for subsequent testing:
  - Independent screening and testing step (PBAT screening)
  - Dependent screening and testing step
- Identify linkage disequilibrium blocks according to some criterion and infer and analyze haplotypes within each block, while retaining for individual analysis those SNPs that do not lie within a block
- Multi-stage designs ...

## 4 Family-based screening strategies

### 4.a PBAT screening

#### Addressing GWA's multiple testing problems

- Adapted from Fulker model with "between" and "within" component (1999):

$$E[Y] = \mu + a_w(X - E[X|P]) + a_b(E[X|P])$$

↑
↙

Family-based      Population-based  
 association

$X$ : coded genotype

$P$ : parental genotypes

## Screen

- Use ‘between-family’ information  
[ $f(S, Y)$ ]
- Calculate conditional power  
( $a_b, Y, S$ )
- Select *top N* SNPs on the basis of power

$$E[Y] = \mu + a_w(X - E[X|P]) + a_b(E[X|P])$$

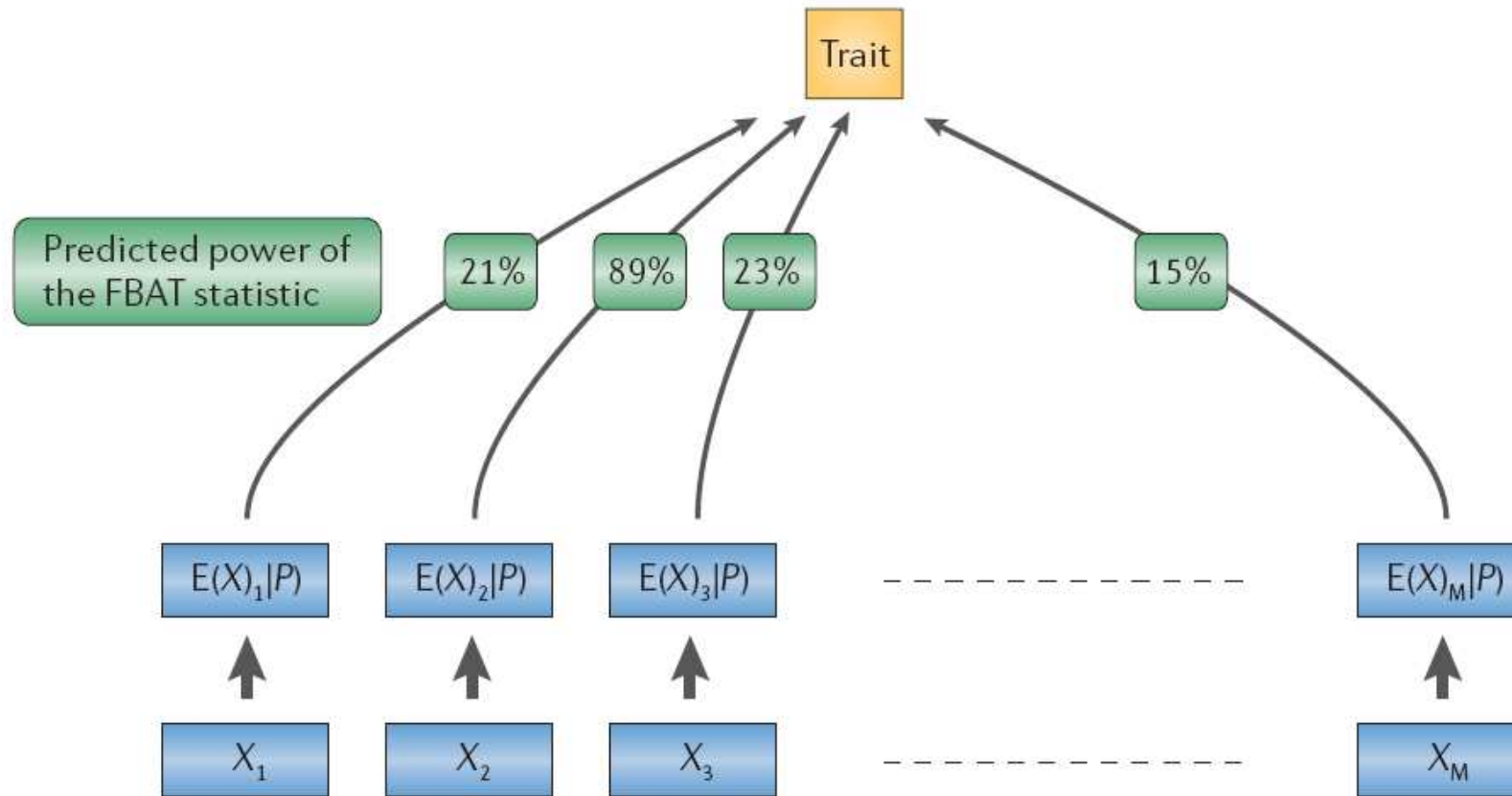
## Test

- Use ‘within-family’ information  
[ $f(X/S)$ ] while computing the FBAT statistic
- This step is independent from the screening step
- Adjust for *N* tests (not 500K!)

$$E[Y] = \mu + a_w(X - E[X|P]) + a_b(E[X|P])$$

(Van Steen et al 2005)

## PBAT screening



(Lange and Laird 2006)

## Detection of 1 DSL

(Van Steen et al 2005)

- SNPChip 10K array on prostate cancer (467 subjects from 167 families) taken as genotype platform in simulation study (10,000 replicates)

		Causal mutation in Affymetrix block			
		SNP1	SNP2	SNP3	SNP4
Method I (top 1)	0.05	0.587 (0.268)	0.690 (0.264)	0.455 (0.091)	0.527 (0.054)
	0.07	0.771 (0.400)	0.841 (0.333)	0.783 (0.116)	0.794 (0.066)
	0.10	0.950 (0.511)	0.964 (0.379)	0.958 (0.125)	0.967 (0.069)
Method II (top 1)	0.05	0.406 (0.152)	0.460 (0.092)	0.318 (0.046)	0.365 (0.122)
	0.07	0.686 (0.293)	0.739 (0.116)	0.688 (0.130)	0.720 (0.241)
	0.10	0.957 (0.345)	0.950 (0.179)	0.958 (0.167)	0.937 (0.373)
Method III	0.05	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
	0.07	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.008 (0.000)
	0.10	0.032 (0.032)	0.040 (0.032)	0.057 (0.057)	0.049 (0.041)
Method IV	0.05	0.024 (0.008)	0.008 (0.008)	0.008 (0.008)	0.000 (0.000)
	0.07	0.041 (0.041)	0.008 (0.008)	0.033 (0.033)	0.024 (0.016)
	0.10	0.153 (0.153)	0.113 (0.105)	0.098 (0.098)	0.146 (0.138)

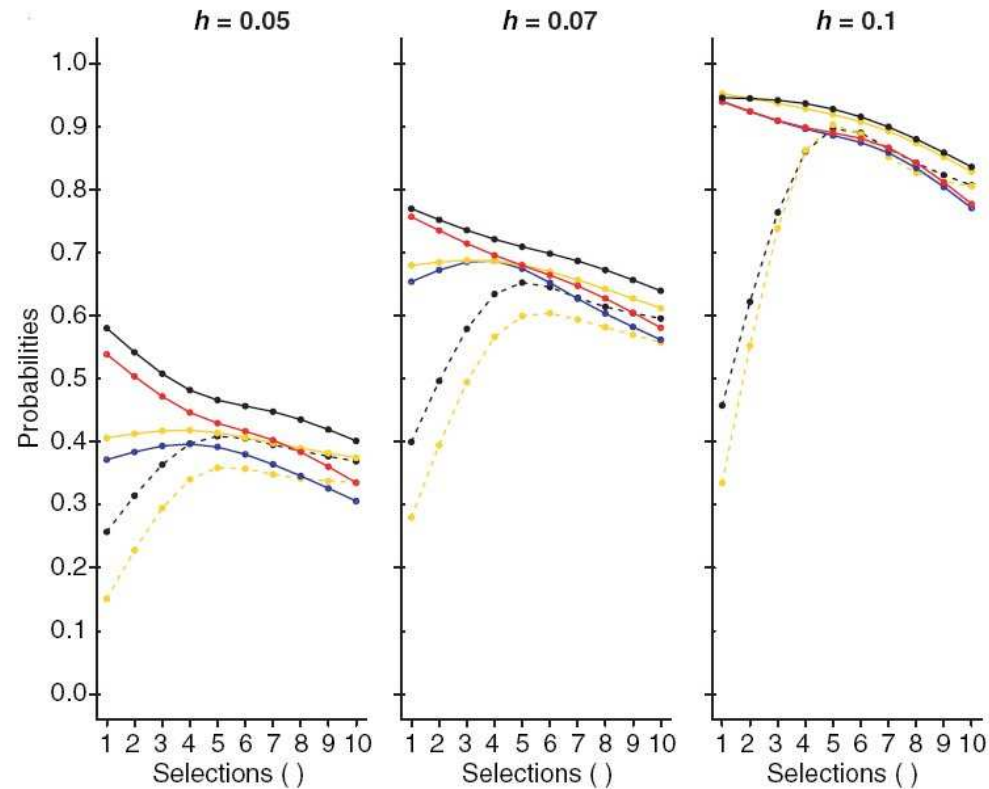
Method I: explained PBAT screening method

Method III: Benjamini-Yekutieli FDR control to 5% (general dependencies)

Method IV: Benjamini-Hochberg FDR control to 5%

# Power to detect 1 DSL

(Van Steen et al 2005)



—• Gene (SNP in): Screen Meth = Cond Power    
 —• Gene (SNP in): Screen Meth = Wald    
 -•- SNP: Screen Meth = Cond Power  
-•- Gene (SNP out): Screen Meth = Cond Power    
 -•- Gene (SNP out): Screen Meth = Wald    
 -•- SNP: Screen Meth = Wald

## One stage is better than multiple stages?

- Macgregor (2008) claims that a total test for family-based designs should be more powerful than a two-stage design
- However, these and similar conclusions are restricted by the methods they include in the comparative study:
  - Ranking based conditional power versus ranking based on  $p$ -values (which is much less informative)
  - Summing the conditional mean model statistic (from PBAT pre-screening stage) and FBAT statistic (from PBAT testing stage) to obtain a single-stage procedure
  - The top  $K$  approach of Van Steen et al (2005) versus the even more powerful weighted Bonferroni approach of Ionita-Laza (2007)



## Weighted Bonferroni Testing

### Screen

- Compute, for all genotyped SNPs, the conditional power of the family-based association test (FBAT) statistic on the basis of the estimates obtained from the conditional mean model
- Since these power estimates are statistically independent of the FBAT statistics that will be computed subsequently, the overall significance level of the algorithm does not need to be adjusted for the screening step.

$$E[Y] = \mu + a_w(X - E[X|P]) + a_b(E[X|P])$$

### Test

- The “new method” tests all markers, not just the 10 or 20 SNPs with the highest power ranking tested in the top K approach, at  $w_i\alpha$  ( $\sum_i w_i = 1$ )
- Unlike a Bonferroni or FDR approach, the new method incorporates the extra information obtained in the screening step (conditional power estimate of the FBAT statistic)

$$E[Y] = \mu + a_w(X - E[X|P]) + a_b(E[X|P])$$

(Ionita-Laza et al. 2007)

## The weighted Bonferroni method and its motivation

- Flexible adjustment:
  - Markers that have a high power ranking are tested at a significance level that is far less stringent than that used in a standard Bonferroni adjustment.
  - For SNPs with low power estimates, the evidence against the null hypothesis has to be extremely strong to overthrow the prior evidence against association from the screening step.
- Hence, the adjustment is made at the expense of the lower-ranked markers, which are tested using more-stringent thresholds.
- The adjustment follows the intuition that low conditional power estimates imply small genetic effect sizes and/or low allele frequencies, which makes such SNPs less desirable choices for the investment of relatively large parts of the significance level.

(Ionita-Laza et al. 2007)

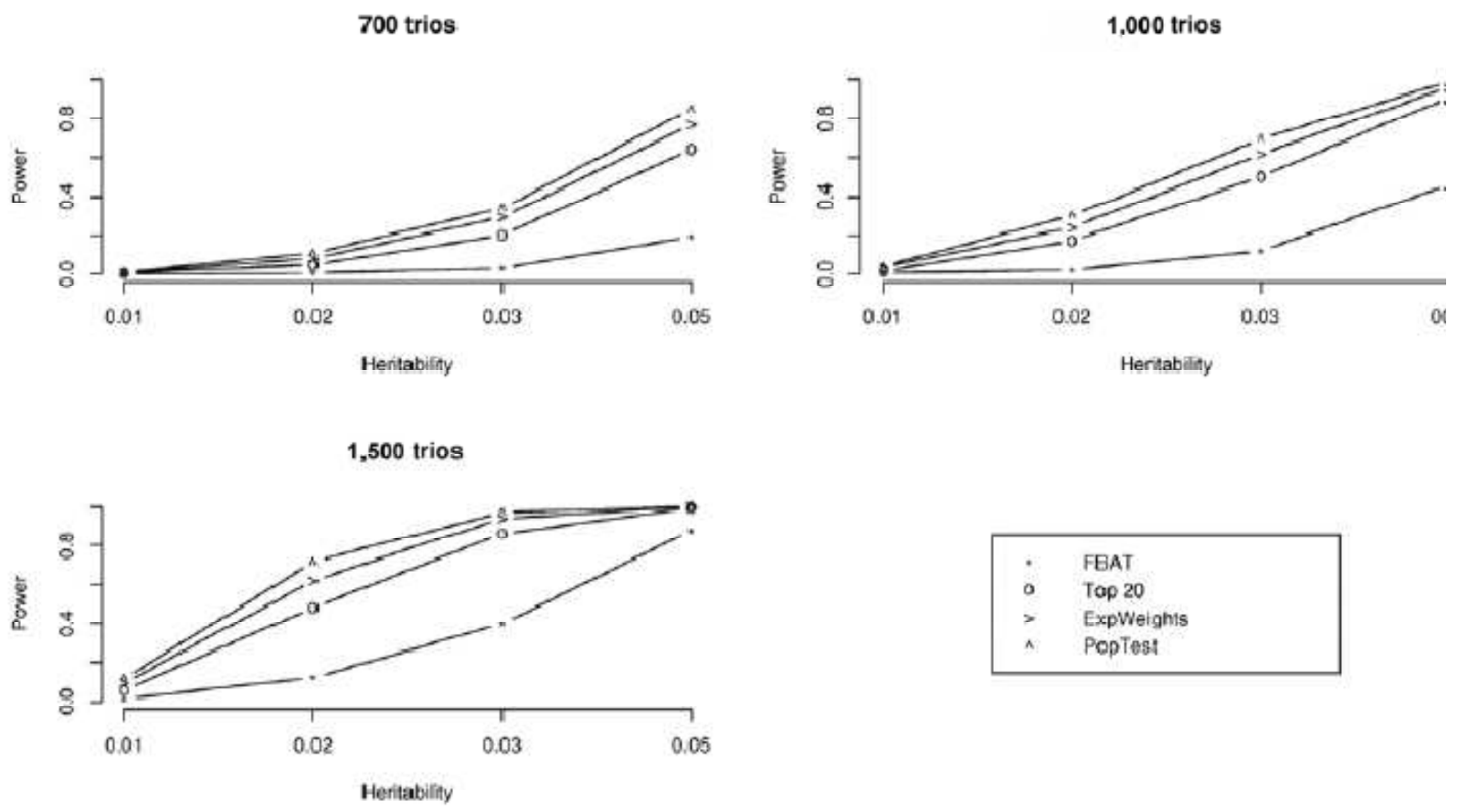
## Ionita-Laza Power to detect 1 DSL

(Ionita-Laza et al 2007)

- Power versus heritability for three sample sizes when there is no population stratification: 700, 1,000, and 1,500 trios.
- Minor-allele frequency at the DSL is 0.3. Other allele frequencies 0.1–0.5 resulted in similar trends and are therefore not shown.
- Comparative study:
  - FBAT is the simple Bonferroni method that uses the FBAT statistic,
  - Top 20 is the top K method with  $K=20$
  - ExpWeights is the Bonferroni weighted method (“new method”) with first partition size  $k_1=5$
  - PopTest is the standard population-based test, with identical numbers of probands, that uses Bonferroni adjustment.

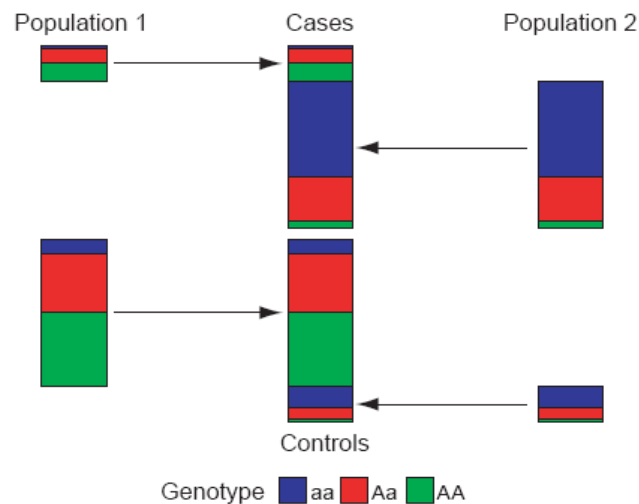
# Ionita-Laza Power to detect 1 DSL

(Ionita-Laza et al 2007)



## 4.b GRAMMAR screening

- Even though family-based design is adopted, when not conditioning on parental genotypes, a distinction should be made between:
  - Analysis of samples of relatives from genetically homogeneous population
  - Analysis of samples of relatives from genetically heterogeneous population



If we mix two populations that have both different disease prevalence and different marker distribution in each population, and there is no association between the disease and marker allele in each population, then there will be an association between the disease and the marker allele in the mixed population. (Marchini 2004)

## Mixed model for families

- A conventional polygenic model of inheritance, which is a statistical genetics' "gold standard", is a mixed model

$$Y = \mu + G + e$$

with an overall mean  $\mu$ , the vector of random polygenic effects  $G$ , and the vector of random residuals  $e$

- For association testing, we need an additional term  $kg$

$$Y = \mu + k g + G + e$$

where

$G$  is random polygenic effect distributed as  $MVN(0, \phi\sigma_G^2)$

$\phi$  is relationship matrix

$\sigma_G^2$  is polygenic variance

- This model is also known as the **measured genotype model (MG)**

## GRAMMAR

- The MG approach, implemented using (restricted) maximum likelihood, is a powerful tool for the analysis of quantitative traits
  - when ethnic stratification can be ignored and
  - pedigrees are small or
  - when there are few dozens or hundreds of candidate polymorphisms to be tested.
- This approach, however, is not efficient in terms of computation time, which hampers its application in genome-wide association analysis.



**Genomewide Rapid Association using Mixed Model And Regression**

(Aulchenko et al 2007; Amin et al 2007)

## GRAMMAR

- Step 1: Compute individual environmental residuals ( $r^*$ ) from the additive polygenic model
- Step 2: Test the markers for association with these residuals using simple linear regression

$$r^* = \mu + k g + e$$

Note that family-effects have been removed!

- Step 3: Due to multiple testing, one could think of type I levels being elevated. However, GRAMMAR actually leads to a conservative test
- Step 4: A genomic-control like procedure, computing the deflation factor as a corrective factor, solves this problem

(Aulchenko et al 2007, Amin et al 2007)



## GRAMMAR versus FBAT

- The GRAMMAR test becomes increasingly conservative and less powerful with the increase in number of large full-sib families and increased heritability of the trait.
- Interestingly, empirical power of GRAMMAR is very close to that of MG
- When no genealogical info on all generations, or when it is inaccurate, the most likely outcome for GRAMMAR (and GM) will be an inflated type I error.
- FBAT has increased power when heritability increases and uses “within” family information only from “informative” families
- FBAT does not explicitly rely on kinship matrices;
- FBAT is robust to population stratification

## 5 Validation

### 5.a Replication

- Replicating the genotype-phenotype association is the “gold standard” for “proving” an association is genuine
- Most loci underlying complex diseases will not be of large effect. It is unlikely that a single study will unequivocally establish an association without the need for replication
- SNPs most likely to replicate:
  - Showing modest to strong statistical significance
  - Having common minor allele frequency
  - Exhibiting modest to strong genetic effect size
- Note: Multi-stage design analysis results should not be seen as “evidence for replication” ...

## Guidelines for replication studies

- Replication studies should be of sufficient size to demonstrate the effect
- Replication studies should be conducted in independent datasets
- Replication should involve the same phenotype
- Replication should be conducted in a similar population
- The same SNP should be tested
- The replicated signal should be in the same direction
- Joint analysis should lead to a lower  $p$ -value than the original report
- Well-designed negative studies are valuable

## 5.b Proof of concept

# A Common Genetic Variant Is Associated with Adult and Childhood Obesity

Alan Herbert,<sup>1\*</sup> Norman P. Gerry,<sup>1</sup> Matthew B. McQueen,<sup>2</sup> Iris M. Heid,<sup>3,4</sup> Arne Pfeufer,<sup>5,6</sup>  
Thomas Illig,<sup>3,4</sup> H.-Erich Wichmann,<sup>3,4,7</sup> Thomas Meitinger,<sup>5,6</sup> David Hunter,<sup>2,8,9</sup> Frank B. Hu,<sup>2,8,9</sup>  
Graham Colditz,<sup>8,9</sup> Anke Hinney,<sup>10</sup> Johannes Hebebrand,<sup>10</sup> Kerstin Koberwitz,<sup>6,10</sup>  
Xiaofeng Zhu,<sup>11</sup> Richard Cooper,<sup>11</sup> Kristin Ardlie,<sup>12</sup> Helen Lyon,<sup>13,14,15</sup> Joel N. Hirschhorn,<sup>13,14,15</sup>  
Nan M. Laird,<sup>16</sup> Marc E. Lenburg,<sup>1</sup> Christoph Lange,<sup>9,13</sup> Michael F. Christman<sup>1\*</sup>

www.sciencemag.org **SCIENCE** VOL 312 14 APRIL 2006

## Genome wide association study of BMI

- A surrogate measure for obesity
- $BMI = \text{weight} / (\text{height})^2$  in  $\text{kg} / \text{m}^2$
- Classification
  - $\geq 25$  = overweight
  - $\geq 30$  = obese

## Epidemiology of BMI

- Prevalence (US)
  - 65% overweight
  - 30% obese
- Seen as risk factor for
  - Diabetes, Stroke, ...
- Non-genetic risk factors
  - Sedentary lifestyle, dietary habits, etc
- Genetic risk factors
  - Heritability = 30-70%

## Design

- Framingham Heart Study (FHS)
  - Public Release Dataset (NHLBI)
  - 694 offspring from 288 families
  - Longitudinal BMI measurements
- Genotypes
  - Affymetrix GeneChip 100K

## Analysis technique

- FBAT screening methodology (Van Steen et al. 2005)
- Exploit longitudinal character of the measurements:
  - Principal Components (PC) Approach
    - Maximize *heritability*
    - Univariate test (one combined trait per obs)
  - PBAT algorithm
    - Find maximum *heritability* of trait without biasing the testing step

Ranking from screen	SNP	Chromosome	Frequency	Informative families	<i>P</i> value FBAT
1	rs3897510	20p12.3	0.36	30	0.2934
2	rs722385	2q32.1	0.16	15	0.1520
3	rs3852352	8p12	0.33	34	0.7970
<b>4</b>	<b>rs7566605</b>	<b>2q14.1</b>	<b>0.37</b>	<b>39</b>	<b>0.0026</b>
5	rs4141822	13q33.3	0.29	27	0.0526
6	rs7149994	14q21.1	0.35	31	0.0695
7	rs1909459	14q21.1	0.39	38	0.2231
8	rs10520154	15q15.1	0.36	38	0.9256
9	rs440383	15q15.1	0.36	38	0.8860
10	rs9296117	6p24.1	0.40	44	0.3652

(genomewide sign: 0.005; rec model)



## Replication

### Family-based design

STUDY	FAMILIES	TEST	P-VALUE
<b>FHS</b> (Original)	288	PBAT	0.003
<b>Maywood</b> (Dichotimous)	342	PBAT	0.009
<b>Maywood</b> (Quantitative)	342	PBAT	0.070
<b>Essen</b> (Children)	368	TDT	0.002

### Cohort design

STUDY	SUBJECTS	TEST	P-VALUE
<b>KORA</b> (QT)	3996	Regression	0.008
<b>NHS</b> (QT)	2726	Regression	> 0.10

(Example on Framinham Study: courtesy of  
Matt McQueen)



**“If you consider the wind-chill factor, adjust for inflation and score on a curve, I only weigh 98 pounds!”**

## Why did this work so well?

- The Study Population
  - Unascertained sample
  - Family-based
  - Longitudinal measurements
- The Method
  - PBAT
- Good Fortune

## 5.c Unexplained heritability

### What are we missing?

- Despite these successes, it has become clear that usually only a small percentage of total genetic heritability can be explained by the identified loci.
- For instance:  
for inflammatory bowel disease (IBD), 32 loci significantly impact disease but they explain only 10% of disease risk and 20% of genetic risk (Barrett et al 2008).

## Possible reasons for poor “heritability” explanation

- This may be attributed to the fact that reality shows
  - multiple small associations (in contrast to statistical techniques that can only detect moderate to large associations),
  - dominance or over-dominance, and involves
  - non-SNP polymorphisms, as well as
  - epigenetic effects,
  - gene-environment interactions and
  - gene-gene interactions (Dixon et al 2000).

## GWA Gene-environment interactions



American Journal of Epidemiology

© The Author 2008. Published by the Johns Hopkins Bloomberg School of Public Health.

All rights reserved. For permissions, please e-mail: journals.permissions@oxfordjournals.org.

Vol. 169, No. 2

DOI: 10.1093/aje/kwn351

Advance Access publication November 20, 2008

---

### Invited Commentary

---

#### Invited Commentary: From Genome-Wide Association Studies to Gene-Environment-Wide Interaction Studies—Challenges and Opportunities

Muin J. Khoury and Sholom Wacholder

*Initially submitted May 30, 2008; accepted for publication July 25, 2008.*

---

The recent success of genome-wide association studies in finding susceptibility genes for many common diseases presents tremendous opportunities for epidemiologic studies of environmental risk factors. Analysis of gene-environment interactions, included in only a small fraction of epidemiologic studies until now, will begin to accelerate as investigators integrate analyses of genome-wide variation and environmental factors. Nevertheless, considerable methodological challenges are involved in the design and analysis of gene-environment interaction studies. The authors review these issues in the context of evolving methods for assessing interactions and discuss how the current agnostic approach to interrogating the human genome for genetic risk factors could be extended into a similar approach to gene-environment-wide interaction studies of disease occurrence in human populations.

environment; epidemiologic methods; genetics; genomics

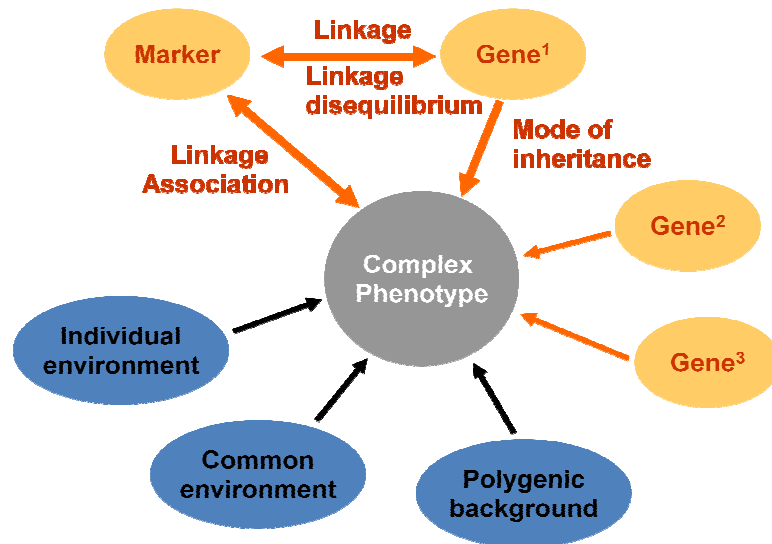
---

Abbreviations: GEWIS, gene-environment-wide interaction studies; GWAS, genome-wide association studies; HuGE, human genome epidemiology.

---

(Khoury et al 2009)

## GWA Gene-gene interactions



(Weiss and Terwilliger 2000)

## Heterogeneity

Analytically, it can be difficult to distinguish between *interactions* and *heterogeneity*.

(Moore 2008)

## References:

- Ziegler A and König I. *A Statistical approach to genetic epidemiology*, 2006, Wiley.
- Lawrence RW, Evans DM, and Cardon LR (2005). Prospects and pitfalls in whole genome association studie. *Philos Trans R Soc Lond B Biol Sci*. August 29; 360(1460): 1589–1595.
- Laird, N., Horvath, S. & Xu, X (2000). Implementing a unified approach to family based tests of association. *Genet. Epidemiol.* 19 Suppl 1, S36–S42.
- Lange, C. & Laird, N.M (2002). On a general class of conditional tests for family-based association studies in genetics: the asymptotic distribution, the conditional power, and optimality considerations. *Genet. Epidemiol.* 23, 165–180.
- Rabinowitz, D. & Laird, N (2000). A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Hum. Hered.* 50, 211–223.
- Aulchenko, Y. S.; de Koning, D. & Haley, C. (2007), 'Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis.', *Genetics* 177(1), 577--585.
- Fulker, D. W. et al (1999). Combined linkage and association sib-pair analysis for quantitative traits. *Am. J. Hum. Genet.* 64, 259–267.



## References (continued):

- Van Steen, K; McQueen, M. B.; Herbert, A.; Raby, B.; Lyon, H.; Demeo, D. L.; Murphy, A.; Su, J.; Datta, S.; Rosenow, C.; Christman, M.; Silverman, E. K.; Laird, N. M.; Weiss, S. T. & Lange, C. (2005), 'Genomic screening and replication using the same data set in family-based association testing.', *Nat Genet* 37(7), 683--691.
- Iles 2008. What can genome-wide association studies tell us about the genetics of common diseases? *PLoS Genetics* 4 (2): e33-.

## Background reading:

- Chen and Abecasis 2007. Family-based association tests for genomewide association scans. *AJHG* 81: 913-
- Johnson 2009. Single-nucleotide polymorphism bioinformatics – A comprehensive review of resources
- Kraft and Hunter 2009. Genetic risk prediction – are we there yet? *N Engl J Med* 360;17.
- Zhang 2008: PTT presentation on R Packages for Genome-Wide Association Studies
- Aulchenko et al 2009. Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nat Genet.* 2009 January ; 41(1): 47–55. doi:10.1038/ng.269.

## In-class discussion document

- Laird and Lange 2006. Family-based designs in the age of large-scale gene-association studies. *Nature Reviews Genetics* 7, 385-394

## Preparatory reading

- Hopper et al 2005. Genetic Epidemiology 6: Population-based family studies in genetic epidemiology. *The Lancet*; 366: 1397–406