PERSPECTIVES ON FAMILY-BASED GWAs

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 implications

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 tests 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

Main references:

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)
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.

<table>
<thead>
<tr>
<th>Marker locus</th>
<th>Disease locus</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A1,A2 alleles)</td>
<td>(D,d alleles)</td>
</tr>
</tbody>
</table>
Features of linkage studies

- 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

(Figure: courtesy of Ed Silverman)
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
1.c Genome wide association analyses (GWAs)

Reasons for continuing popularity of GWAs using SNPs

• They 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

*statistical issues cannot be ruled out*
Scale of the study

candidate gene approach

vs

genome-wide screening approach

Can’t see the forest for the trees

Can’t see the trees for the forest
GWA screening is a complicated process

- 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...


Study designs

Multi-stage

- Less expensive
- More complicated
- Less powerful

Single-stage

- More expensive
- Less complicated
- More powerful

(slide: courtesy of McQueen)
2 FAMILIES VERSUS CASES/CONTROLS

Main references:


2.a Every design has statistical implications

There are many possible designs for a genetic association study

<table>
<thead>
<tr>
<th>Details</th>
<th>Advantages</th>
<th>Disadvantages</th>
<th>Statistical analysis method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross-sectional</td>
<td>Genotype and phenotype (i.e. note disease status or quantitative trait value) in a random sample from the population</td>
<td>Inexpensive, Provides estimate of disease prevalence</td>
<td>Few affected individuals if disease rare</td>
</tr>
<tr>
<td>Cohort</td>
<td>Genotype subsection of population and follow disease incidence for a specified time period</td>
<td>Provides estimate of disease incidence</td>
<td>Expensive to follow up, issues with drop-out</td>
</tr>
<tr>
<td>Case-control</td>
<td>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 consensus sample.</td>
<td>No need for follow up, Provides estimates of exposure effects</td>
<td>Requires careful selection of controls, Potential for confounding (e.g., population stratification)</td>
</tr>
<tr>
<td>Extreme values</td>
<td>Genotype individuals with extreme high or low values of a quantitative trait, as established from initial cross-sectional or cohort sample</td>
<td>Genotype only most informative individuals hence save on genotyping costs</td>
<td>No estimate of true genetic effect sizes</td>
</tr>
<tr>
<td>Case-parent triads</td>
<td>Genotype affected individuals plus their parents (afflicted individuals determined from initial cross-sectional, cohort, or disease outcome based sample).</td>
<td>Robust to population stratification, Can estimate maternal and imprinting effects</td>
<td>Less powerful than case-control design</td>
</tr>
<tr>
<td>Case-parent-grandparent septets</td>
<td>Genotype affected individuals plus their parents and grandparents</td>
<td>Robust to population stratification, Can estimate maternal and imprinting effects</td>
<td>Granparents rarely available</td>
</tr>
<tr>
<td>General pedigrees</td>
<td>Genotype random sample or disease outcome based sample of families from general population. Phenotype for disease trait or quantitative trait</td>
<td>Higher power with large families, Sample may already exist from linkage studies</td>
<td>Sensitive to genotype, Many missing individuals</td>
</tr>
<tr>
<td>Case-only</td>
<td>Genotype only affected individuals, obtained from initial cross-sectional, cohort, or disease outcome based sample</td>
<td>Most powerful design for detection of interaction effects</td>
<td>Can only estimate interaction effects, Very sensitive to population stratification</td>
</tr>
<tr>
<td>DNA-pooling</td>
<td>Applies to variety of disease designs, but genotyping is of pools of anywhere between two and 100 individuals, rather than on an individual basis</td>
<td>Potentially inexpensive compared within-bulk genotyping (but technology still under development)</td>
<td>Hard to estimate different experimental sources of variance</td>
</tr>
</tbody>
</table>

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

Transmission
Disequilibrium
Test (TDT)
2.b Power considerations

Rare versus common diseases (Lange and Laird 2006)
Power

- 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)
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

<table>
<thead>
<tr>
<th></th>
<th>Control</th>
<th>Case exposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>exposed</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>No</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>Yes</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

- 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.

(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

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, \( Y=0 \) if child is unaffected in a trio design
  - \( T=Y \)
  - \( X \) 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)
3 FROM COMPLEX PHENOMENA TO MODELS

Main references:

3.a Introduction

- 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.

(Weiss and Terwilliger 2000)

(Moore 2008)
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)

- Chapter 5: with too many SNPs
  
  - 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

Main references:


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

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

<table>
<thead>
<tr>
<th></th>
<th>SNP1</th>
<th>SNP2</th>
<th>SNP3</th>
<th>SNP4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method I (top 1)</td>
<td>0.05</td>
<td>0.587 (0.268)</td>
<td>0.690 (0.264)</td>
<td>0.455 (0.091)</td>
</tr>
<tr>
<td></td>
<td>0.07</td>
<td>0.771 (0.400)</td>
<td>0.841 (0.333)</td>
<td>0.783 (0.116)</td>
</tr>
<tr>
<td></td>
<td>0.10</td>
<td>0.950 (0.511)</td>
<td>0.964 (0.379)</td>
<td>0.958 (0.125)</td>
</tr>
<tr>
<td>Method II (top 1)</td>
<td>0.05</td>
<td>0.406 (0.152)</td>
<td>0.460 (0.092)</td>
<td>0.318 (0.046)</td>
</tr>
<tr>
<td></td>
<td>0.07</td>
<td>0.506 (0.293)</td>
<td>0.729 (0.116)</td>
<td>0.600 (0.130)</td>
</tr>
<tr>
<td></td>
<td>0.10</td>
<td>0.957 (0.345)</td>
<td>0.960 (0.179)</td>
<td>0.958 (0.167)</td>
</tr>
<tr>
<td>Method III</td>
<td>0.05</td>
<td>0.000 (0.000)</td>
<td>0.000 (0.000)</td>
<td>0.000 (0.000)</td>
</tr>
<tr>
<td></td>
<td>0.07</td>
<td>0.000 (0.000)</td>
<td>0.000 (0.000)</td>
<td>0.000 (0.000)</td>
</tr>
<tr>
<td></td>
<td>0.10</td>
<td>0.032 (0.032)</td>
<td>0.040 (0.032)</td>
<td>0.057 (0.057)</td>
</tr>
<tr>
<td>Method IV</td>
<td>0.05</td>
<td>0.024 (0.008)</td>
<td>0.008 (0.008)</td>
<td>0.008 (0.008)</td>
</tr>
<tr>
<td></td>
<td>0.07</td>
<td>0.041 (0.041)</td>
<td>0.006 (0.006)</td>
<td>0.033 (0.033)</td>
</tr>
<tr>
<td></td>
<td>0.10</td>
<td>0.103 (0.103)</td>
<td>0.113 (0.103)</td>
<td>0.098 (0.098)</td>
</tr>
</tbody>
</table>

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)

- SNP: Screen Meth = Cond Power
- Gene (SNP in): Screen Meth = “ “
- Gene (SNP out): Screen Meth = “ “
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.
- 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)
**Motivation**

- 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.

- This 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)
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 \( \text{MVN}(0, \phi \sigma^2_G) \)
- \( \phi \) is relationship matrix
- \( \sigma^2_G \) 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

Main references:

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 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,¹* Norman P. Gerry,¹ Matthew B. McQueen,² Iris M. Heid,³,⁴ Arne Pfeufer,⁵,⁶ Thomas Illig,³,⁴ H.-Erich Wichmann,³,⁴,⁷ Thomas Meitinger,⁵,⁶ David Hunter,²,⁸,⁹ Frank B. Hu,²,⁸,⁹ Graham Colditz,⁸,⁹ Anke Hinney,¹⁰ Johannes Hebebrand,¹⁰ Kerstin Koberwitz,⁶,¹⁰ Xiaofeng Zhu,¹¹ Richard Cooper,¹¹ Kristin Ardlie,¹² Helen Lyon,¹³,¹⁴,¹⁵ Joel N. Hirschhorn,¹³,¹⁴,¹⁵ Nan M. Laird,¹⁶ Marc E. Lenburg,¹ Christoph Lange,⁹,¹³ Michael F. Christman¹*
Genome wide association study of BMI

- A surrogate measure for obesity
- BMI = weight / (height)$^2$ in kg / 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

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

(genomewide sign: 0.005; rec model)
# Replication

## Family-based design

<table>
<thead>
<tr>
<th>STUDY</th>
<th>FAMILIES</th>
<th>TEST</th>
<th>P-VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>FHS (Original)</td>
<td>288</td>
<td>PBAT</td>
<td>0.003</td>
</tr>
<tr>
<td>Maywood (Dichotimous)</td>
<td>342</td>
<td>PBAT</td>
<td>0.009</td>
</tr>
<tr>
<td>Maywood (Quantitative)</td>
<td>342</td>
<td>PBAT</td>
<td>0.070</td>
</tr>
<tr>
<td>Essen (Children)</td>
<td>368</td>
<td>TDT</td>
<td>0.002</td>
</tr>
</tbody>
</table>

## Cohort design

<table>
<thead>
<tr>
<th>STUDY</th>
<th>SUBJECTS</th>
<th>TEST</th>
<th>P-VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>KORA (QT)</td>
<td>3996</td>
<td>Regression</td>
<td>0.008</td>
</tr>
<tr>
<td>NHS (QT)</td>
<td>2726</td>
<td>Regression</td>
<td>&gt; 0.10</td>
</tr>
</tbody>
</table>

(Example on Framingham 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
Success stories of GWAs (nearly 100 loci, 40 common diseases/traits)

(Manolio et al 2008)
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 and
Analytically, it can be difficult to distinguish between *interactions* and *heterogeneity*.

(Weiss and Terwilliger 2000)
Definitions for Heterogeneity

<table>
<thead>
<tr>
<th>Allelic heterogeneity</th>
<th>Locus heterogeneity</th>
<th>Phenocopy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Definition</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>When two or more alleles of a single locus are independently associated with the same trait</td>
<td>When two or more DNA variations in distinct genetic loci are independently associated with the same trait</td>
<td>The presence of a disease phenotype that has a non-genetic (random or environmental) basis</td>
</tr>
<tr>
<td><strong>Diagram</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allelic variant i of locus A</td>
<td>Allelic variant ii of locus A</td>
<td>Allelic variant i of locus A</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Trait heterogeneity</th>
<th>Phenotypic variability</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Definition</strong></td>
<td></td>
</tr>
<tr>
<td>When a trait, or disease, has been defined with insufficient specificity such that it is actually two or more distinct underlying traits</td>
<td>Variation in the degree, severity or age of onset of symptoms exhibited by persons who actually have the same trait or disease process</td>
</tr>
<tr>
<td><strong>Diagram</strong></td>
<td></td>
</tr>
<tr>
<td>Trait I</td>
<td>Trait II</td>
</tr>
</tbody>
</table>

Moderate or severe | Early onset or late onset
Presentation 1 of trait I | Presentation 2 of trait I
Disease X

(Thornton-Wells et al 2004)
Two main types of Interactions

<table>
<thead>
<tr>
<th>Gene-gene interaction</th>
<th>Gene-environment interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Definition</strong></td>
<td>When two or more DNA variations interact either directly (DNA-DNA or DNA-mRNA interactions), to change transcription or translation levels, or indirectly by way of their protein products, to alter disease risk separate from their independent effects.</td>
</tr>
<tr>
<td></td>
<td>When a DNA variation interacts with an environmental factor, such that their combined effect is distinct from their independent effects</td>
</tr>
<tr>
<td><strong>Diagram</strong></td>
<td>Allelic variant i of locus A</td>
</tr>
<tr>
<td></td>
<td>Allelic variant ii of locus B</td>
</tr>
<tr>
<td></td>
<td>No disease</td>
</tr>
<tr>
<td></td>
<td>Allelic variant i of locus A</td>
</tr>
<tr>
<td></td>
<td>Environmental factor K</td>
</tr>
<tr>
<td></td>
<td>No disease</td>
</tr>
</tbody>
</table>

(Thornton-Wells et al 2004)
6 BEYOND MAIN EFFECTS

Main references:


6.a Dealing with multiplicity

Multiple testing explosion

~500,000 SNPs span 80% of common variation in genome (HapMap)
Ways to handle multiplicity

Recall that several strategies can be adopted, including:

- clever multiple corrective procedures
- pre-screening strategies,
- multi-stage designs,
- adopting haplotype tests or
- multi-locus tests

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

• The multiple testing problem becomes “unmanageable” when looking at multiple loci jointly?
6.b A bird’s eye view on roads less travelled by

Multiple disease susceptibility loci (mDSL)

• Dichotomy between
  - Improving single markers strategies to pick up multiple signals at once (PBAT)
  - Testing groups of markers (FBAT multi-locus tests)
PBAT screening for mDSL

• Little has been done in the context of family-based screening for epistasis
• First assess how a method is capable of detecting multiple DSL
• Simulation strategy (10,000 replicates):
  - Genetic data from Affymetrix SNPChip 10K array on 467 subjects from 167 families
  - Select 5 regions; 1 DSL in each region
  - Generate traits according to normal distribution, including up to 5 genetic contributions
  - For each replicate: generate heritability according to uniform distribution with mean $h = 0.03$ for all loci considered (or $h = 0.05$ for all loci)

(Van Steen et al 2005)
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 H₀ 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)
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)

(Lange and Laird 2006)
**Power to detect genes with multiple DSL**

<table>
<thead>
<tr>
<th>Identified genes ($h = 0.03$)</th>
<th>Identified genes ($h = 0.05$)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DSLs</strong></td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0.646</td>
</tr>
<tr>
<td>3</td>
<td>0.855</td>
</tr>
<tr>
<td>3</td>
<td>0.776</td>
</tr>
<tr>
<td>4</td>
<td>0.823</td>
</tr>
<tr>
<td>4</td>
<td>0.846</td>
</tr>
<tr>
<td>5</td>
<td>0.914</td>
</tr>
<tr>
<td>5</td>
<td>0.730</td>
</tr>
<tr>
<td>5</td>
<td>0.822</td>
</tr>
</tbody>
</table>

*top: top 5 SNPs in the ranking*
*bottom: top 10 SNPs in the ranking*

(Van Steen et al 2005)
### Power to detect genes with multiple DSL

<table>
<thead>
<tr>
<th>DSLs</th>
<th>Identified genes ($h = 0.03$)</th>
<th>Identified genes ($h = 0.05$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>0.059</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>0.201</td>
<td>0.000</td>
</tr>
<tr>
<td>3</td>
<td>0.138</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>0.303</td>
<td>0.000</td>
</tr>
<tr>
<td>4</td>
<td>0.258</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>0.485</td>
<td>0.010</td>
</tr>
<tr>
<td>5</td>
<td>0.368</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>0.563</td>
<td>0.008</td>
</tr>
</tbody>
</table>

**top:** Benjamini-Yekutieli FDR control at 5% (general dependencies)

**bottom:** Benjamini-Hochberg FDR control at 5%

(Van Steen et al 2005)
FBAT multi-locus tests

FBAT-SNP-PC: An Approach for Multiple Markers and Single Trait in Family-Based Association Tests

Gmil S. Rakovski\textsuperscript{a} Scott T. Weiss\textsuperscript{b} Nan M. Laird\textsuperscript{b} Christophe Lanape\textsuperscript{a,b}

\textsuperscript{a}Department of Biostatistics, Harvard School of Public Health, and \textsuperscript{b}Channing Laboratory, Harvard Medical School.
Bostok, Mass., USA.

\begin{itemize}
\item FBAT-SNP-PC attains higher power in candidate genes with lower average pair-wise correlations and moderate to high allele frequencies with large gains (up to 80%).
\end{itemize}

(Rakovski et al 2008)

An Efficient Family-Based Association Test Using Multiple Markers

Xin Xu\textsuperscript{a,\ast}, Cyril Rakovski\textsuperscript{a}, Xiping Xu\textsuperscript{a} and Nan Laird\textsuperscript{b}

\textsuperscript{a}Program for Population Genetics, Harvard School of Public Health, Boston, Massachusetts
\textsuperscript{b}Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts

In genetic association studies, multiple markers are usually employed to cover a genomic region of interest for localizing a trait locus. To this report, we present a novel multi-marker family-based association test (\textit{T}_\text{MMD}) that linearly combines the single-marker test statistics using data-driven weights. We examined the type-I error rate in a simulation study and compared its power to identify a common trait locus using both single nucleotide polymorphisms (SNP) within the same haplotype block that the trait locus resides with three competing tests including a global haplotype test \textit{T}_\text{H}, a multi-marker test similar to the Horsting\textsuperscript{\textsuperscript{1}}'s test for the population-based data (\textit{T}_\text{HPOP}), and a single-marker test with Bonferroni's correction for multiple testing (\textit{T}_\text{BB}). The type-I error rate of \textit{T}_\text{MMD} is well maintained in our simulation study. In all the scenarios we examined, \textit{T}_\text{MMD} is the most powerful, followed by \textit{T}_\text{BB} and \textit{T}_\text{H}. The \textit{T}_\text{BB} is the most powerful test when both parents are missing. However, when both parents are missing, \textit{T}_\text{MMD} is substantially more powerful than \textit{T}_\text{BB}. We also apply this new test on a data set from a previous association study on nicotine dependence. Curr. Epidemiol. 30:682–686, 2006. © 2006 Wiley-Liss, Inc.

Key words: multi-marker family-based association; FBAT

(FBAT-LC : Xin et al 2008)
Multi-locus tests for unrelateds

• Parametric methods:
  - Regression
  - Logistic or (Bagged) logic regression

• Non-parametric methods:
  - Combinatorial Partitioning Method (CPM)
    ▪ quantitative phenotypes; interactions
  - Multifactor-Dimensionality Reduction (MDR)
    ▪ qualitative phenotypes; interactions
  - Machine learning and data mining

• The multiple testing problem becomes “unmanageable” when looking at (genetic) interaction effects?
6.c Pure epistasis models

What’s in a name?

• Distortions of Mendelian segregation ratios due to one gene masking the effects of another (William Bateson 1861-1926).

• Deviations from linearity in a statistical model (Ronald Fisher 1890-1962).

“Epistasis:
what it means,
what it doesn't mean,
and statistical methods to detect it in humans”

(Cordell 2002)
Interpretation of epistasis

- The study of epistasis poses problems of interpretability. Statistically, epistasis is usually defined in terms of deviation from a model of additive effects, but this might be on either a linear or logarithmic scale, which implies different definitions.

(Moore 2004)

- Despite the aforementioned concerns, there is evidence that a direct search for epistatic effects can pay dividends.
- It is expected to have an increasing role in future analyses...
Slow shift from main towards epistatic effects

(Motsinger et al 2007)
Some “philosophical” considerations

• A variant with small marginal effect is not necessarily clinically insignificant:
  - It might turn out to have a strong effect in certain genetic or environmental backgrounds,
  - and in any case might give clues to mechanisms of disease causation.
• Most analyses of population association data focus on the marginal effect of individual variants, mostly because looking out for multiple interacting variants simultaneously is a daunting business:

  Is the indirect approach of first seeking marginal effects a better strategy than tackling epistatic effects directly?
Some “philosophical” considerations (continued)

• Gene-gene interactions are readily incorporated into SNP-based or haplotype-based regression models and related tests.

What about the “hierarchy rule” in statistical parametric models under the assumption of “pure epistasis”?

• It is commonly known that in “interaction analyses”, the case-only study design that looks for association between two genes can give greater power than the heavily used case-control design.

What about family-based designs?
Non-parametric multifactor dimensionality reduction methods

(adapted from Lou et al 2008)  
(Ritchie et al 2003)
FAM-MDR as a semi-parametric approach for families

• Family-adapted Model-Based Multifactor Dimensionality Reduction (MB-MDR) technique
  (MB-MDR: Calle et al 2008)

• Uses GRAMMAR principles (Aulchenko et al 2007, Amin et al 2007), but now for genome-wide epistasis screening:
  - Step 1: Perform a polygenic analysis using the complete pedigree structure
    ▪ Does not use measured genotypes in the mean model statement
  - Step 2: Derive residuals from the model in step 1
    ▪ Gives rise to familial correlation-free “new” traits
  - Step 3: Submit to MB-MDR
    (Cattaert et al 2009 - in preparation)
MB-MDR as a semi-parametric approach for unrelated

- Step 1: New risk cell identification via association test on each genotype cell $c_j$
  - Parametric or non-parametric test of association

- Step 2: Test one-dimensional “genetic” construct $X$ on $Y$

- Step 3: assess significance
  - $W = [b/se(b)]^2$, $b=\ln(OR)$
  - Adjust for number of combined cells in high and low risk category

(Calle et al 2007, Calle et al 2008)
7 FUTURE CHALLENGES

Integration of –omics data in GWAs

• Post-analysis
  - As validation tool in main effects GWAs

• During the analysis:
  - Epistasis screening (FAM-MDR)
    ▪ Use expression values to prioritize multi-locus combinations
  - Main effects screening (PBAT)
    ▪ Construct an overall phenotype for each marker based on the linear combination of expression values (e.g., within 1Mb from the marker) that maximizes heritability and perform FBAT-PC screening to prioritize SNPs
Extensive boundary crossing collaborations

Statistical Genetics Research Club (www.statgen.be)