Use of DNA Pooling in Large-Scale Association Studies

Prof. Pak Sham

Social, Genetic & Developmental Psychiatry Centre
Institute of Psychiatry
King's College, London

Scale of Association Studies

Genome-wide

<table>
<thead>
<tr>
<th>Strong LD</th>
<th>Moderate LD</th>
<th>Weak LD</th>
</tr>
</thead>
<tbody>
<tr>
<td>$50,000</td>
<td>$100,000</td>
<td>$500,000</td>
</tr>
</tbody>
</table>

Gene-based

30,000 x 10

@ $0.1 / genotype, Cost / Sample $30,000

Sample Size for Complex Diseases

Small Effect Size

Interactions

Large Sample Size 10,000

Incomplete LD

Struggled Significance
Use of DNA Pooling in Large-Scale Association Studies
Prof. Pak Sham

Feasibility of genome-wide screen

2,000 samples @ $30,000 / sample

$60,000,000

Not currently feasible

Efficient screening using pooled DNA

Pool Construction

Allele frequency Measurements

Pool-based association tests

Individual-based association tests

Cost and Throughput of Pooling

Pool Construction:
$1 / sample
Our technician-week / 100 samples

Genotyping:
$2 per marker per pool
2 technician-days / 284 plate

The screen versions of these slides have full details of copyright and acknowledgements
Use of DNA Pooling in Large-Scale Association Studies
Prof. Pak Sham

Feasibility of screen using pooling

2 pools of 1,000 individuals
Duplicate measurements
300,000 markers

@ $2 / measurement → $2,400,000

(less if pool-based genotyping becomes cheaper)

Pool Construction

Individual DNA samples
Equal amounts of DNA
Pooled DNA

Dilute DNA sample to 40-80 ng/μl using spectrophotometer

Dilute DNA to 10 ng/μl

Measure concentration by fluorometer

Dilute DNA to 5 ± 0.5 ng/μl

Measure out required volume to give target amount of DNA

The screen versions of these slides have full details of copyright and acknowledgements.
Use of DNA Pooling in Large-Scale Association Studies
Prof. Pak Sham

Quantitative Genotyping Assays

- Restriction fragment length polymorphism
- Real-time pyrophosphate DNA sequencing
- Single-base extension with fluorescently-labelled ddNTPs
- Homogeneous S' -nuclease assay
- MALDI-TOF mass spectrometry

“can all serve for quantification of allele frequencies in DNA pools with reasonable accuracy”


SNaPshot Principles

![SNaPshot diagram]

Experimental Errors in Pooling

- Experimental Errors
  - Systematic
    - Differential Amplification
  - Random
    - Pool Construction
    - Genotyping

The screen versions of these slides have full details of copyright and acknowledgements
Use of DNA Pooling in Large-Scale Association Studies
Prof. Pak Sham

Differential Amplification

Peak Heights from GENESCAN
Heterozygous Genotype
Differential Amplification Ratio
\[ k = \frac{a}{b} \]

Differential Amplification: Correction

Peak Heights from a DNA pool
Allele frequencies estimates:
\[ p = \frac{c}{(c+d)} \]
\[ q = \frac{kd}{(c+kd)} \]

Estimated K-values in 5 markers

The screen versions of these slides have full details of copyright and acknowledgements
Differential Amplification: Consequences

Uncorrected Differential Amplification

\[ \text{Var}(\hat{\beta}) = \left( \frac{k}{1 + (k-1)p} \right)^2 \frac{p(1-p)}{2n} \]

→ Liberal / Conservative P-values

---

Differential Amplification: Consequences

Correction using an unbiased estimate of \( k \), with coefficient of variation CV

Additional variance in allele frequency estimate

\[ \text{Var}(\hat{e}_k) = (p(1-p)CV)^2 \]

Additional variance in difference between allele frequency estimates

\[ \text{Var}(\hat{e}_{X_1} - \hat{e}_{X_2}) = \left( (p_1(1-p_1)+p_2(1-p_2))CV \right)^2 \]


---

Pool Construction vs Genotyping Errors

Multiple Duplicate Pools

\[ \text{Simultaneous} \]

Single Pool

\[ \Rightarrow \text{Pool construction error } \ll \text{Pool measurement error} \]

Robert Plomin. personal communication

---
Use of DNA Pooling in Large-Scale Association Studies
Prof. Pak Sham

Experimental Variance: Pool Construction

Theoretical: \( \text{Var}(e_C) = \frac{P(1-P)}{2n} \frac{CV^2}{2n} \)


Empirical: Average Variance estimates of

- 0.00002 (Shifman et al (2002) Molecular and Cellular Proteins 16: 429-434)

Measurement Variance of 22 Markers

[Graph showing variance of 22 markers]

Robert Fung, personal communication.

Measurement Errors

- Experimental Errors
  - Systematic
  - Random
    - Differential Amplification
    - Pool Construction
    - Genotyping

- 0.5 - 2.0
- 0.00001 - 0.0001
- 0.0001 - 0.001

Each laboratory needs to optimise own procedures and to monitor quality.
Use of DNA Pooling in Large-Scale Association Studies
Prof. Pak Sham

Design Parameters

Cases

 Controls

 POOLS

 How many pools?

 Allele Frequency Measurements

Optimal Design

Sampling Variance ($V_s$) \hspace{1cm} Measurement Variance ($V_M$)

$$\frac{p(1-p)}{2n} \hspace{1cm} \frac{E^2}{r}$$

Information

$$E = \frac{V_s}{V_s + V_M}$$

Number of repeated measurements

$$r = \frac{E E^2 2n}{(1-E)p(1-p)}$$

If $n = 100$

$E^2 = 0.0004$

$p = 0.2$

Then $r = 2$

Optimal Design

$n = 50, r = 1$ \hspace{1cm} Barrett et al. (2003) Annals of Human Genetics 66:333-405

$n = 100, r = 2$ \hspace{1cm} May offer some advantages

separation of errors

quality control
Family Designs

Parental Controls

Control pool

Case pool

Sibling Controls

Control pool

Case pool

Avoid false positives from hidden population stratification

Unrelated versus Parental Controls

Non-centrality Parameter

(Expected chi-squared statistic - 1)

Case-controls

\[ \frac{(\hat{p}_1 - p) \hat{p}}{2n} \left( \frac{p_1 (1 - p)}{2} + \frac{p_2 (1 - p)}{2} \right) \frac{2 \sigma^2}{r} \]

Unrelated versus Parental Controls

Non-centrality Parameter

Case-parents

\[ \frac{(\hat{p}_1 - p_1) \hat{p}_1}{2n} \left( \frac{p_1 (1 - p_1)}{2} + \frac{p_2 (1 - p_1)}{2} \right) \frac{2 \sigma^2}{r} \]

David Clayton, Personal Communication
Use of DNA Pooling in Large-Scale Association Studies
Prof. Pak Sham

Stratified Designs

<table>
<thead>
<tr>
<th>Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>High risk</td>
<td></td>
</tr>
<tr>
<td>Low risk</td>
<td></td>
</tr>
</tbody>
</table>

Power to detect gene effect and gene-environment interaction

Design for Quantitative Traits

Optimal pooling fraction (no experimental error)

Design for Quantitative Traits

Optimal pooling fraction (no experimental error)

Impact of Measurement Error (σ)

σ → + information + optimal pool fraction

Juszczak et al. (2002) ERBB

The screen versions of these slides have full details of copyright and acknowledgements
Use of DNA Pooling in Large-Scale Association Studies
Prof. Pak Sham

**Statistical Analysis of Pooled DNA Data**

- Chi-Square Approach
  - Pool allele frequencies
  - Expected allele counts
  - Pearson chi-squared statistics
  - Problem: liberal p-values

**Statistical Analysis of Pooled DNA Data**

- Meta-Regression
  - Study \(\rightarrow\) Pool
  - Effect Size \(\rightarrow\) Allele Frequency
  - Study Variance \(\rightarrow\) Sampling Variance + Measurement Variance
  - Predictors \(\rightarrow\) Case-Control Status, Covariates

**Statistical Analysis of Pooled DNA Data**

- Likelihood Approaches
  - Model
    - Frequencies
      - Alleles
      - Haplotypes
    - Genotype-Phenotype Relationship
      - Disease Traits
      - Quantitative Traits
      - Penetrances
      - Additive effects
      - Dominance
    - Fit model to “observed” allele frequencies in pools
Use of DNA Pooling in Large-Scale Association Studies

Prof. Pak Sham

ADHD & Noradrenergic Transporter Gene

Background
The noradrenergic system is known to play a role in attentional mechanisms and is thought to be important in ADHD.

Samples
- 2 case pools (n=90, n=90)
- 4 control pools (n=90, n=88, n=77, n=79)

Lab-work
- SNPs were selected from the public databases.
- 21 SNPs with minor allele frequencies >10% were typed in pools.
- Positive associations (p<.05) from the DNA pooling were followed by individual genotyping.

Significance levels of 6 markers

IQ & Non-Synonymous SNPs

Samples
- Random sample of 1000 individuals
- Ranks by IQ
- Pools of the 5 quartiles (200 individuals each)

Lab-work
- Non-synonymous SNPs were selected from the public databases.
- Each SNP is measured 3 times in each DNA pool.
Use of DNA Pooling in Large-Scale Association Studies
Prof. Pak Sham

IQ Data: Likelihood vs. Meta-regression

Haplotype Analysis
Impossible in Pooling?
Almost, but not quite .... Especially if

<table>
<thead>
<tr>
<th>haplotype</th>
<th>h12</th>
<th>h21</th>
</tr>
</thead>
<tbody>
<tr>
<td>h21</td>
<td>0</td>
<td>p1</td>
</tr>
<tr>
<td>q1</td>
<td>q2</td>
<td>1</td>
</tr>
</tbody>
</table>

Many Poools
Allele frequencies correlate
Yang et al (2003) PNAS
Barrett et al (2002) AHS

Recommendations
- DNA pooling is an efficient method for conducting a first screen in large-scale association studies
- Need for optimal laboratory procedures and quality control to minimize experimental errors
- Careful consideration of design issues
- Meta-regression / likelihood methods of analysis
Use of DNA Pooling in Large-Scale Association Studies
Prof. Pak Sham

Acknowledgements

* Funding Sources: Medical Research Council, Wellcome Trust, National Eye Institute, National Institute of Mental Health

* Colleagues / Collaborators

Statistical: Molecula: Phenotypes
Joel Bader: Ian Craig: Robert Plomin
Aruna Bansal: Mike Owen: Phil Asherson
Shaun Purcell: Mick O’Donovan
Ansar Jawaid: Linzy Hill
Jo Knight: Farzana Hoda: Xieohui Xu