Statistical challenges to genome-wide association study

Naoyuki Kamatani, M.D., Ph.D.

1. Director and Professor, Institute of Rheumatology, Tokyo Women’s Medical University

2. Director, Medical Informatics Group, SNP Research Center, RIKEN
Scientific breakthrough of the year 2007

IN SCIENCE

Editorial: Breakthrough of the Year
Science Editor-in-Chief Donald Kennedy overviews the big stories from 2007 covered in this year's Breakthrough issue.

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.

It's All About Me
Along with the flood of discoveries in human genetics, 2007 saw the birth of a new industry: personal genomics. But researchers worry that these services open up a Pandora's box of ethical issues.

The Scorecard: How'd We Do?
Some of last year's predictions panned out this year, especially the work that led to the Breakthrough of the Year, but other areas are progressing more slowly.

Genome-wide association study (GWAS) boomed in 2007

Human genetic variation

Proof of the Poincaré Conjecture for 2006
Can we identify disease-associated genes on the genome-wide basis?

Yes, by the Linkage Analysis

300 – 500 markers can cover the whole genome

Causes of the majority of Mendelian diseases have been elucidated.

10^7 base pair sequence is transmitted together to the next generation

However, the effect size should be large and family data are necessary.

The phenotype-associated locus is here!
Can we identify disease-associated genes on the genome-wide basis, even if the effect size is small or family data are unavailable?

Yes, by the GWAS (genome-wide association study)

100,000 – 1,000,000 markers can cover the whole genome

Causes of complex diseases may be elucidated by GWAS.

10^4-10^5 base pair sequence is associated with each other
Era of GWAS has come!

GWAS: Genome-wide association study

1. A list of SNPs covering the whole genome was made (HapMap)

2. Chips and Beads used for the genotyping for 100,000 – 1,000,000 individual SNPs are now commercially available.

3. Methods for analyzing the large size genotyping data are available.
Reports of GWAS in 2007


Majority of genetic causes of major diseases will be elucidated within a few years!!
37.8 PROSPECTS FOR WHOLE-GENOME ASSOCIATION STUDIES

Initial reports of WGA studies began as early as 2002 (Ozaki et al., 2002), but studies involving more comprehensive coverage of common variants began in 2005 (Klein et al., 2005; Duerr et al., 2006; Hampe et al., 2007). The initial reports are promising, with each study identifying and validating several novel loci for different diseases. These studies demonstrate that WGA can be successful in identifying common variants for complex traits in humans. Given the chequered history of human genetic association studies (Cardon and Bell, 2001; Ioannidis et al., 2001), this is a major advance in the field.
Recent Developments in Genomewide Association Scans: A Workshop Summary and Review

Duncan C. Thomas,¹ Robert W. Haile,¹ and David Duggan²

¹Department of Preventive Medicine, University of Southern California, Los Angeles; and ²Translational Genomics Research Institute (TGen), Phoenix

Numerous research groups are planning or have underway genomewide searches for a range of disorders and the first reports of such studies (using early versions of high-density SNP chips) are just beginning to appear (Ozaki et al. 2002; Klein et al. 2005).
Genome-wide association: a promising start to a long race

David M. Evans and Lon R. Cardon

The Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford, OX3 7BN, UK

A recent study by Cheung et al. demonstrates how to identify expression quantitative trait loci (eQTLs) underlying gene expression phenotypes through a combination of genome-wide linkage analysis and subsequent fine mapping or by genome-wide association (GWA) analysis. This study emphasizes the complexity of human traits, highlighting the challenges faced by investigators – in particular, insufficient linkage disequilibrium between the trait and marker variant, genetic heterogeneity and correcting for multiple testing will all adversely impact the power to detect loci by association. These issues must be considered carefully if the GWA approach is to succeed in mapping complex phenotypes.

GWA analysis of gene expression levels in humans

Recently, after much anticipation, the first genome-wide association (GWA) studies in humans are beginning to appear in the literature [1,2]. Cheung and colleagues recently published the first GWA analysis of gene expression levels in a human population [3]. The idea behind their approach, which has been termed ‘expression genetics’ [4], is to subject levels of gene expression to the

References

GWAS (genome-wide association) reports from RIKEN


ORs (proportional to sample size) with 95% CIs from each study testing the association of RA with the risk allele of PADI4 gene. The pooled ORs with 95% CI for overall analysis and subgroup analysis in populations of European descent were calculated with the Mantel–Haenszel method (diamonds). The first study by Suzuki et al. [6] is shown for reference only and was not included in the meta-analysis.

Meta-analysis has supported the association between a SNP and a disease.

Iwamoto et al. Rheumatology 2006
QC of large quantity of data

(500,000 SNP genotypes from > 10,000 subjects)

1. QC (quality control) is extremely laborious.

2. Mistypes lead to false significance.

3. We can use both genetics and statistics–based methods for QC.

4. Reliable conclusion from GWAS is dependent on sophisticated QC filter.

More than $10^{10}$ data points.
QQ plot

Report of the results of an association study

-\log_{10}P\ profile
Multiple-comparison problem

1. If a test of independence is performed for 500,000 SNPs with a significance level of 0.05, about 2,500 SNPs will become false positive.

2. Since many SNPs are associated with each other, Bonferroni’s correction is too conservative.

3. Several correction methods have been proposed
   (a) Use of the concept FDR (false-discovery rate)
   (b) Permutation test
   (c) Exact calculation of type 1 error rate
   (d) Bayesian method (FPRP)
Observed data

A permutation outcome ($\omega$)

Number of cases $n_1$

Number of controls $n_2$

$\Omega$ (Sample space) size $n!$

For 500 cases and controls, $10^{2568}$ outcomes

For 500 cases and controls, $10^{299}$ events

Permutation test requires long time calculation

Shuffle the phenotypes and obtain empirical distribution of a statistic under the null hypothesis.
Problem of population structuring

1. A large sample size is necessary to identity a SNP with a small effect size.

2. If the sample size is large, however, the problem of population structuring emerges.
Inflation of type I error rate by mixing two different subpopulations

If allele or genotype frequencies are different between subpopulations, and the prevalence of a disease is different between the subpopulations,

\[
OR = \frac{q_1 p_1 + (1 - q_1) p_2}{(1 - p_1) q_1 + (1 - p_2) (1 - q_1)} \times \frac{(1 - p_1) q_2 + (1 - q_2) (1 - p_2)}{p_1 q_2 + (1 - q_2) p_2}
\]

\(OR\) is 1 when \(p_1 = p_2\), or \(q_1 = q_2\) then, false-positive associations will occur.
To avoid false positive associations, we may use a clustering technique.
Principle component analysis

Draw a line in the space with 140,000 dimension so that the variance of the projections of the points to the line becomes the largest.

7,000 points in a 140,000 dimension space

Use of the projections of the points to separate subjects

This is impossible because the calculation of covariance matrix for 140,000 x 140,000 matrix is impossible.
Principle component analysis (implemented in EIGENSTRAT)

Draw a line in the space with 7,000 dimension so that the variance of the projections of the points to the line becomes the largest.

140,000 points in a 7,000 dimension space

Use of factors of Eigenvectors to separate subjects
Genotype data from 7,000 subjects with 140,000 SNPs

<table>
<thead>
<tr>
<th>Subjects</th>
<th>7,000 lines</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNP</td>
<td>140,000 rows</td>
</tr>
</tbody>
</table>

Normalize for each SNP (mean $p$, variance $2p(1-p)$)
## Normalized genotype data

<table>
<thead>
<tr>
<th>SNP</th>
<th>140,000 rows</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>-0.694811</td>
<td>-0.288294</td>
</tr>
<tr>
<td>-0.694811</td>
<td>-0.288294</td>
</tr>
<tr>
<td>-0.694811</td>
<td>-1.72976</td>
</tr>
<tr>
<td>1.04222</td>
<td>-0.288294</td>
</tr>
<tr>
<td>-0.694811</td>
<td>-0.288294</td>
</tr>
<tr>
<td>-0.694811</td>
<td>-0.288294</td>
</tr>
<tr>
<td>-0.694811</td>
<td>-1.72976</td>
</tr>
<tr>
<td>1.04222</td>
<td>1.15318</td>
</tr>
<tr>
<td>-0.694811</td>
<td>-0.288294</td>
</tr>
<tr>
<td>-0.694811</td>
<td>-0.288294</td>
</tr>
<tr>
<td>-0.694811</td>
<td>-0.288294</td>
</tr>
<tr>
<td>1.04222</td>
<td>1.15318</td>
</tr>
<tr>
<td>1.04222</td>
<td>1.15318</td>
</tr>
<tr>
<td>1.04222</td>
<td>1.15318</td>
</tr>
<tr>
<td>1.04222</td>
<td>1.15318</td>
</tr>
<tr>
<td>1.04222</td>
<td>1.15318</td>
</tr>
<tr>
<td>-0.694811</td>
<td>-0.288294</td>
</tr>
<tr>
<td>-0.694811</td>
<td>-0.288294</td>
</tr>
<tr>
<td>1.04222</td>
<td>1.15318</td>
</tr>
<tr>
<td>-0.694811</td>
<td>-0.288294</td>
</tr>
<tr>
<td>1.04222</td>
<td>1.15318</td>
</tr>
<tr>
<td>-0.694811</td>
<td>-0.288294</td>
</tr>
<tr>
<td>-0.694811</td>
<td>-1.72976</td>
</tr>
<tr>
<td>-0.694811</td>
<td>-0.172976</td>
</tr>
</tbody>
</table>

**Subjects**  
**7,000 lines**  
**24,500,000 pairs**
Covariance is calculated for each of 24,500,000 pairs each of which has 140,000 SNPs data.

\[ \sum_{i}^{n} (x_i - \bar{x})(y_i - \bar{y}) \]
## Covariance matrix

Subjects 7,000

<p>| | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.00000</td>
<td>0.57756</td>
<td>0.58845</td>
<td>0.33203</td>
<td>0.30350</td>
<td>0.29527</td>
<td>0.11765</td>
<td>0.18763</td>
<td>0.32503</td>
<td>0.35023</td>
</tr>
<tr>
<td>1.05568</td>
<td>0.57756</td>
<td>1.00000</td>
<td>0.58845</td>
<td>0.33203</td>
<td>0.30350</td>
<td>0.29527</td>
<td>0.11765</td>
<td>0.18763</td>
<td>0.32503</td>
<td>0.35023</td>
</tr>
<tr>
<td>0.31481</td>
<td>0.58845</td>
<td>0.58845</td>
<td>1.00000</td>
<td>0.58845</td>
<td>0.33203</td>
<td>0.30350</td>
<td>0.29527</td>
<td>0.11765</td>
<td>0.18763</td>
<td>0.32503</td>
</tr>
<tr>
<td>0.24716</td>
<td>0.33203</td>
<td>0.33203</td>
<td>0.33203</td>
<td>1.00000</td>
<td>0.58845</td>
<td>0.30350</td>
<td>0.29527</td>
<td>0.11765</td>
<td>0.18763</td>
<td>0.32503</td>
</tr>
<tr>
<td>0.33653</td>
<td>0.30350</td>
<td>0.30350</td>
<td>0.30350</td>
<td>0.58845</td>
<td>1.00000</td>
<td>0.58845</td>
<td>0.33203</td>
<td>0.30350</td>
<td>0.29527</td>
<td>0.11765</td>
</tr>
<tr>
<td>0.36962</td>
<td>0.29527</td>
<td>0.29527</td>
<td>0.29527</td>
<td>0.33203</td>
<td>0.58845</td>
<td>1.00000</td>
<td>0.58845</td>
<td>0.33203</td>
<td>0.30350</td>
<td>0.29527</td>
</tr>
<tr>
<td>0.22971</td>
<td>0.11765</td>
<td>0.11765</td>
<td>0.11765</td>
<td>0.29527</td>
<td>0.33203</td>
<td>0.58845</td>
<td>1.00000</td>
<td>0.58845</td>
<td>0.33203</td>
<td>0.30350</td>
</tr>
<tr>
<td>0.08708</td>
<td>0.18763</td>
<td>0.18763</td>
<td>0.18763</td>
<td>0.18763</td>
<td>0.18763</td>
<td>0.18763</td>
<td>1.00000</td>
<td>0.58845</td>
<td>0.33203</td>
<td>0.30350</td>
</tr>
<tr>
<td>0.11689</td>
<td>0.32503</td>
<td>0.32503</td>
<td>0.32503</td>
<td>0.32503</td>
<td>0.32503</td>
<td>0.32503</td>
<td>0.32503</td>
<td>1.00000</td>
<td>0.58845</td>
<td>0.33203</td>
</tr>
<tr>
<td>0.15486</td>
<td>0.35023</td>
<td>0.35023</td>
<td>0.35023</td>
<td>0.35023</td>
<td>0.35023</td>
<td>0.35023</td>
<td>0.35023</td>
<td>0.35023</td>
<td>1.00000</td>
<td>0.58845</td>
</tr>
</tbody>
</table>

Calculate Eigenvectors for this table
PCA analysis for African, European and Asian subjects

![PCA plot showing the distribution of African, European, and Asian subjects across the first and second components.](image-url)
PCA analysis for African, European and Asian subjects
PCA analysis for Asian subjects

First component

Second component

Hondo

Okinawa

Han-Chinese
All Japanese samples + HapMap Han-Chinese and Japanese samples

Yayoi? (~3,000 – 1,700 years ago)

First component

Second component
Samples from Okinawa

- Okinawa cluster
- Hondo cluster
- Han-Chinese cluster
Samples from Kyushu

Hondo cluster

Okinawa cluster

Han-Chinese cluster
Samples from Kinki

- Hondo cluster
- Okinawa cluster
- Han-Chinese cluster
Samples from Tokai-Hokuriku

- Hondo cluster
- Okinawa cluster
- Han-Chinese cluster
Samples from Kanto-Koshinetsu

First component

Second component

Hondo cluster

Okinawa cluster

Han-Chinese cluster
Samples from Tohoku

- Han-Chinese cluster
- Hondo cluster
- Okinawa cluster
Samples from Hokkaido

First component

Second component

Hondo cluster

Okinawa cluster

Han-Chinese cluster
Comparison of samples from Kinki and Tohoku areas

Comparison of Tohoku and Kinki subpopulations as cases and controls are problematic when the sample size is over 400.
Nonsynonymous SNPs ranked according to P values of Armitage test based on genotypes for Hondo and Okinawa clusters

<table>
<thead>
<tr>
<th>rs_number</th>
<th>chr</th>
<th>chr_pos</th>
<th>P value</th>
<th>gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs3827760</td>
<td>2</td>
<td>108880033</td>
<td>1.61E-20</td>
<td>EDAR</td>
</tr>
<tr>
<td>rs17822931</td>
<td>16</td>
<td>46815699</td>
<td>8.48E-20</td>
<td>ABCC11</td>
</tr>
<tr>
<td>rs4285045</td>
<td>4</td>
<td>144355168</td>
<td>1.23E-19</td>
<td>USP38</td>
</tr>
<tr>
<td>rs1799986</td>
<td>12</td>
<td>55821533</td>
<td>3.44E-17</td>
<td>LRP1</td>
</tr>
<tr>
<td>rs2274067</td>
<td>1</td>
<td>229443429</td>
<td>1.49E-15</td>
<td>C1orf131</td>
</tr>
<tr>
<td>rs2230611</td>
<td>19</td>
<td>5163482</td>
<td>3.49E-15</td>
<td>PTPRS</td>
</tr>
<tr>
<td>rs1872056</td>
<td>15</td>
<td>69827828</td>
<td>1.57E-14</td>
<td>FLJ13710</td>
</tr>
<tr>
<td>rs2298645</td>
<td>18</td>
<td>75829123</td>
<td>1.97E-14</td>
<td>LOC440498</td>
</tr>
<tr>
<td>rs3744921</td>
<td>18</td>
<td>28121686</td>
<td>3.03E-14</td>
<td>FAM59A</td>
</tr>
<tr>
<td>rs631248</td>
<td>1</td>
<td>43843808</td>
<td>8.79E-14</td>
<td>PTPRF</td>
</tr>
<tr>
<td>rs9932051</td>
<td>16</td>
<td>10482297</td>
<td>7.99E-13</td>
<td>ATF7IP2</td>
</tr>
</tbody>
</table>

Known to be associated with the thickness of the hair

Known to be associated with dry or wet ear wax

This method may be useful to identify genes that have been the targets of natural selection
Inflation of type 1 error due to population structuring (expressed by lambda value for genomic control, mean and sd).

Subjects in Hondo and Okinawa clusters were mixed to construct a 500 or 1,000 size case group. Control group consisted of only the subjects from Hondo cluster.
Method for avoiding the Inflation of type I error rate by mixing two different subpopulations

Adjust the proportion of subpopulation 1 so that \( q_1 = q_2 \) followed by a simple chi square test or by Mantel-Haenzel test.

\[
OR = \frac{q_1 p_1 + (1 - q_1) p_2}{(1 - p_1) q_1 + (1 - p_2)(1 - q_1)} \times \frac{(1 - p_1) q_2 + (1 - q_2)(1 - p_2)}{p_1 q_2 + (1 - q_2) p_2}
\]

\( OR \) is 1 when \( p_1 = p_2 \), or \( q_1 = q_2 \).
Method for avoiding the Inflation of type I error rate by mixing two different subpopulations

Mantel-Haenzel test

Simply exclude subpopulation 1
Method for avoiding the Inflation of type I error rate by mixing two different subpopulations

Selection of matched controls at random from a large-size control sample
Conclusion

The data management and statistical analysis for millions or billions of individual genotypes in GWAS are extremely laborious; however, they are a challenging world for statisticians.