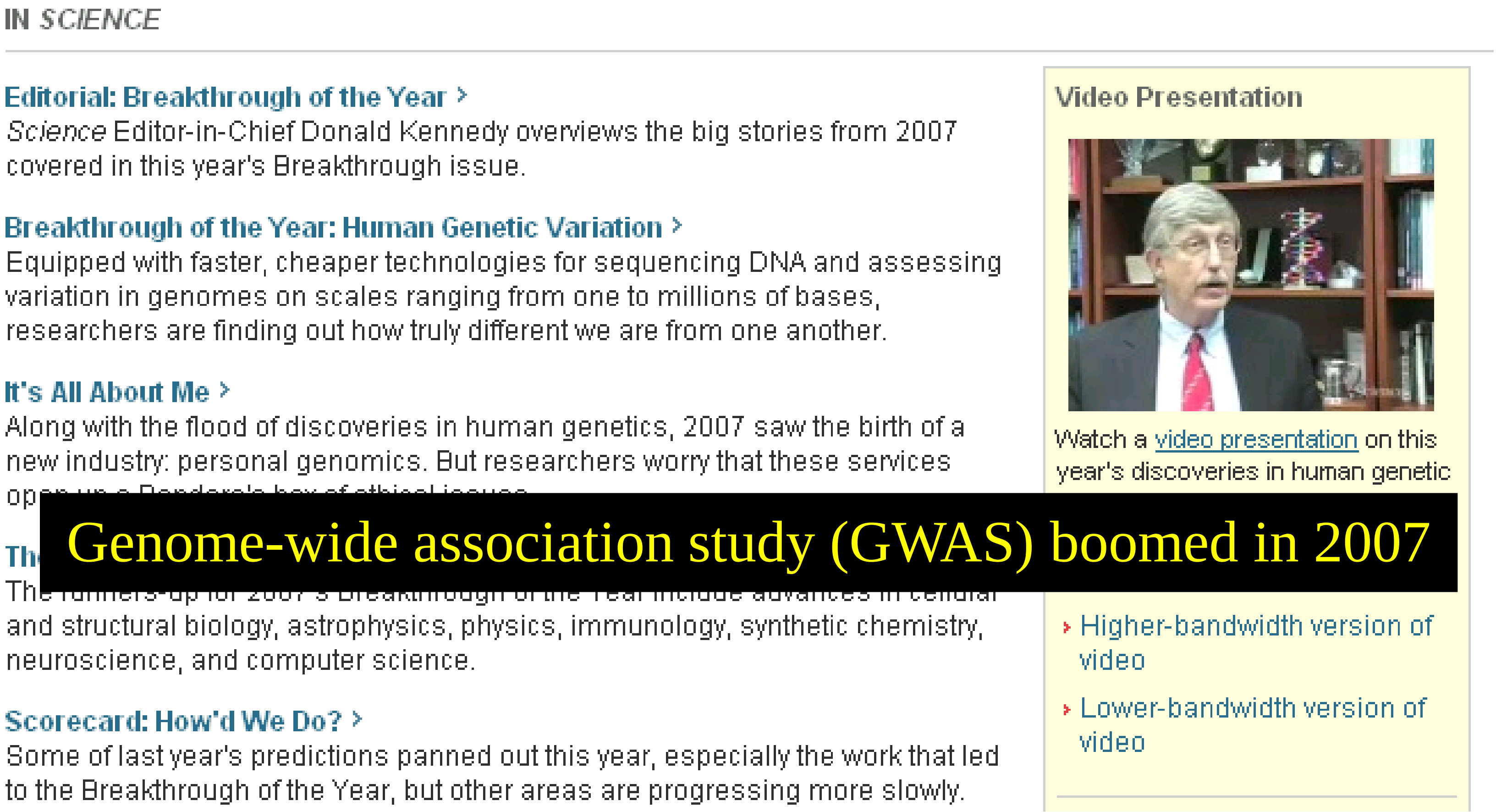
Statistical challenges to genome-wide association study

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.

Th
The runners-up for 2007's Breakthrough of the Year include advances in cellular and structural biology, astrophysics, physics, immunology, synthetic chemistry, neuroscience, and computer science.

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.

Video Presentation

Watch a video presentation on this year's discoveries in human genetic

- Higher-bandwidth version of video
- Lower-bandwidth version of video

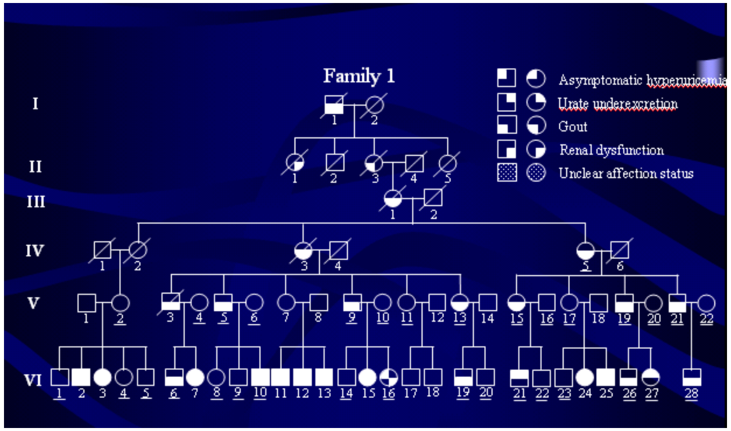
Genome-wide association study (GWAS) boomed in 2007

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

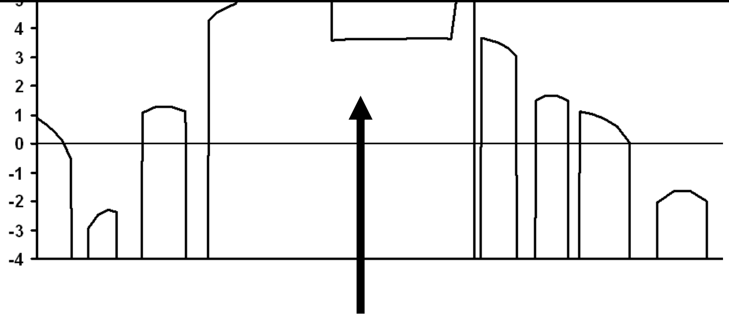


300 – 500 markers can cover the whole genome

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.

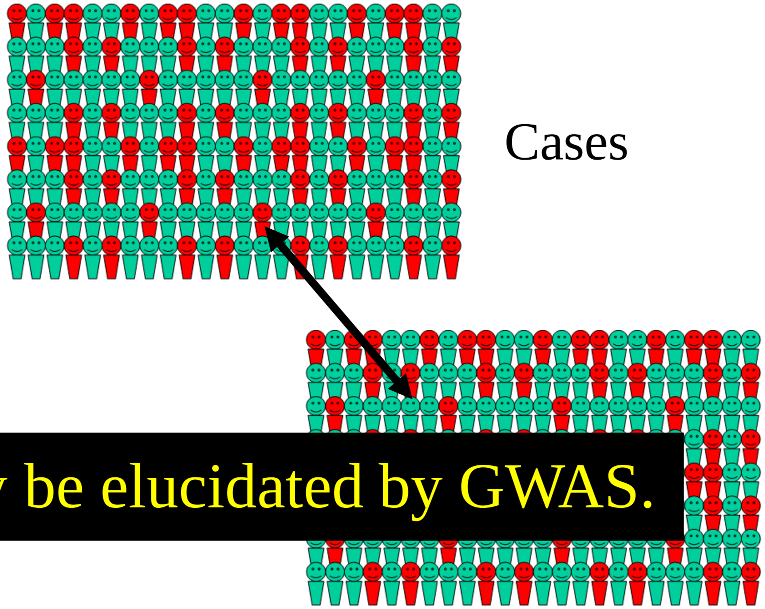


The phenotype-associated locus is here!

Can we identify disease-associated genes on the genome-wide basis,

even if the effect size is small or family data are unavailable?

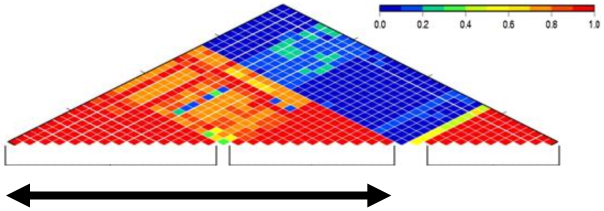
Yes, by the GWAS
(genome-wide association study)



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



range of linkage disequilibrium

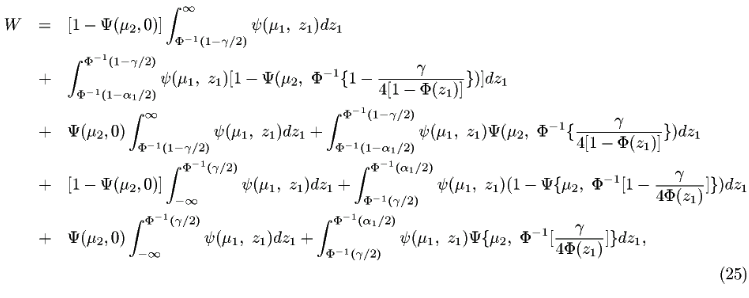
Era of GWAS has come!

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.



$$\begin{aligned} W &= [1-\Psi(\mu_2,0)]\int_{\Phi^{-1}(1-\gamma/2)}^{\infty}\psi(\mu_1,\ z_1)dz_1 \\ &+ \int_{\Phi^{-1}(1-\alpha_1/2)}^{\Phi^{-1}(1-\gamma/2)}\psi(\mu_1,\ z_1)[1-\Psi(\mu_2,\ \Phi^{-1}\{1-\frac{\gamma}{4[1-\Phi(z_1)]}\})]dz_1 \\ &+ \Psi(\mu_2,0)\int_{\Phi^{-1}(1-\gamma/2)}^{\infty}\psi(\mu_1,\ z_1)dz_1 + \int_{\Phi^{-1}(1-\alpha_1/2)}^{\Phi^{-1}(1-\gamma/2)}\psi(\mu_1,\ z_1)\Psi(\mu_2,\ \Phi^{-1}\{\frac{\gamma}{4[1-\Phi(z_1)]}\})dz_1 \\ &+ [1-\Psi(\mu_2,0)]\int_{-\infty}^{\Phi^{-1}(\gamma/2)}\psi(\mu_1,\ z_1)dz_1 + \int_{\Phi^{-1}(\gamma/2)}^{\Phi^{-1}(\alpha_1/2)}\psi(\mu_1,\ z_1)(1-\Psi\{\mu_2,\ \Phi^{-1}[1-\frac{\gamma}{4\Phi(z_1)}]\})dz_1 \\ &+ \Psi(\mu_2,0)\int_{-\infty}^{\Phi^{-1}(\gamma/2)}\psi(\mu_1,\ z_1)dz_1 + \int_{\Phi^{-1}(\gamma/2)}^{\Phi^{-1}(\alpha_1/2)}\psi(\mu_1,\ z_1)\Psi\{\mu_2,\ \Phi^{-1}[\frac{\gamma}{4\Phi(z_1)}]\}dz_1, \end{aligned} \tag{25}$$

Reports of GWAS in 2007

Samani NJ et al. Genomewide association analysis of coronary artery disease. N Engl J Med 357, 443, 2007

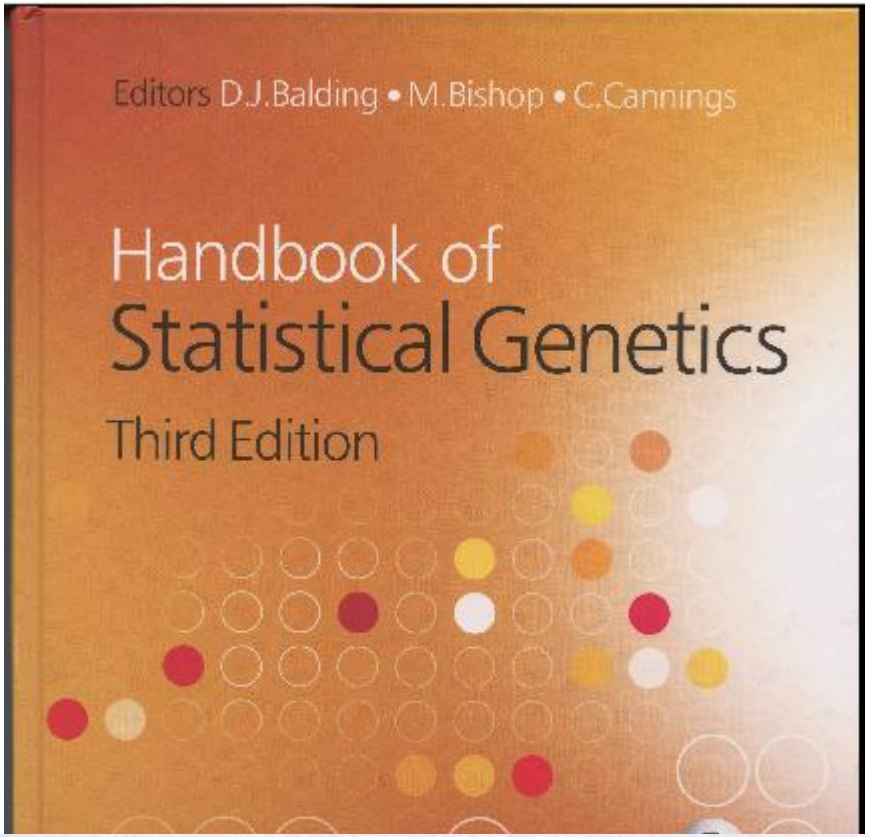
Stefansson H. et al. A genetic risk factor for periodic limb movements in sleep. N Engl J Med 357, 639, 2007

Dunckley et al. Whole-genome analysis of sporadic amyotrophic lateral sclerosis. N Engl J Med 357, 775, 2007

The international multiple sclerosis genetics consorsium. Risk alleles for multiple sclerosis identified by a genomewide study. N Engl J Med 357, 851, 2007

Plenge RM et al. TRAF1-C5 as a risk locus for rheumatoid arthritis--a genomewide study. N Engl J Med. 357, 1199, 2007

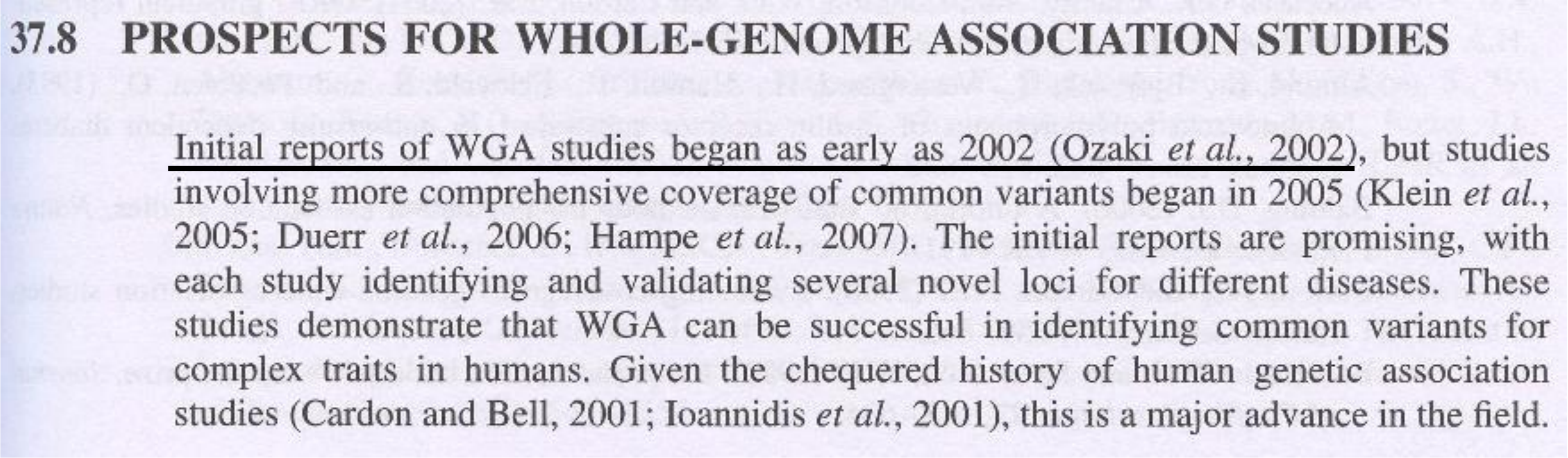
Majority of genetic causes of major diseases will be elucidated within a few years!!



Editors D.J.Balding • M.Bishop • C.Cannings

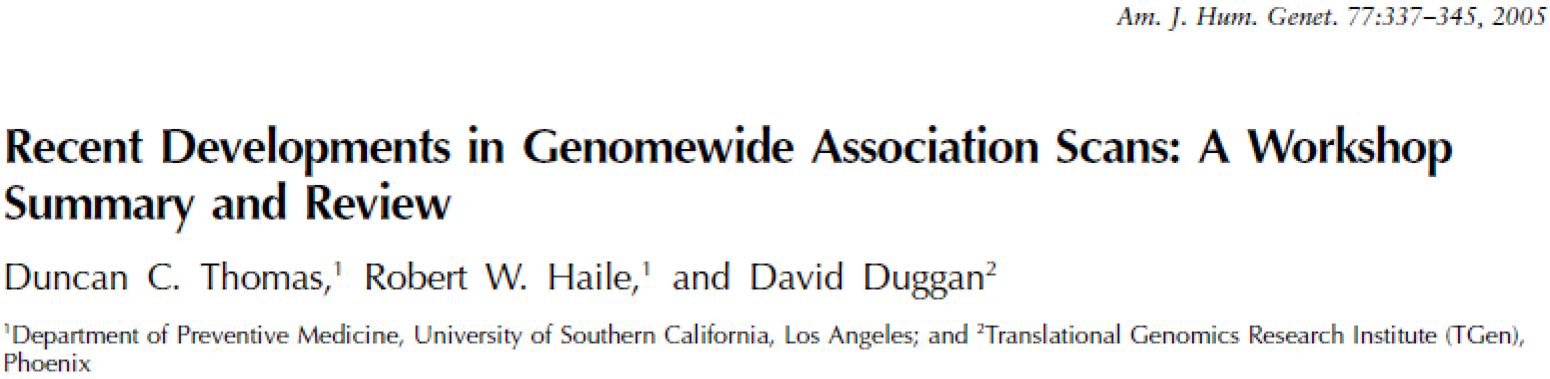
Handbook of Statistical Genetics

Third Edition



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.

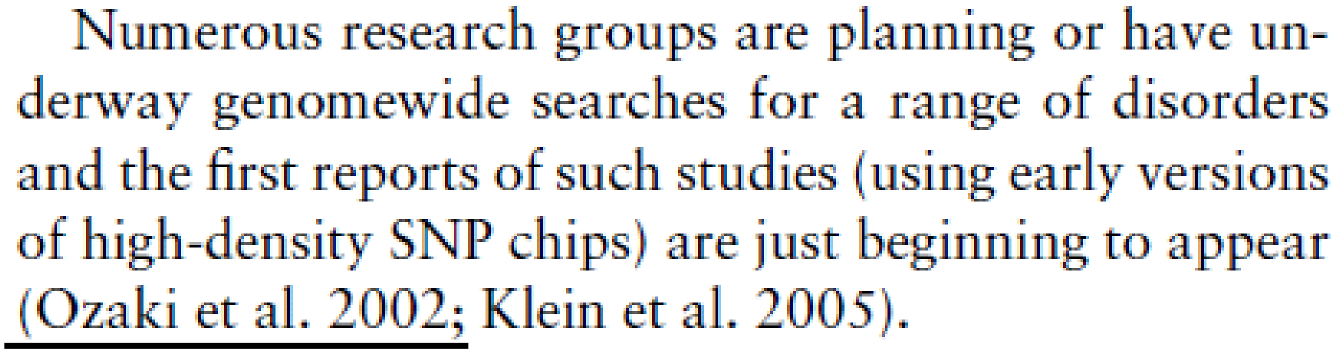


Am. J. Hum. Genet. 77:337–345, 2005

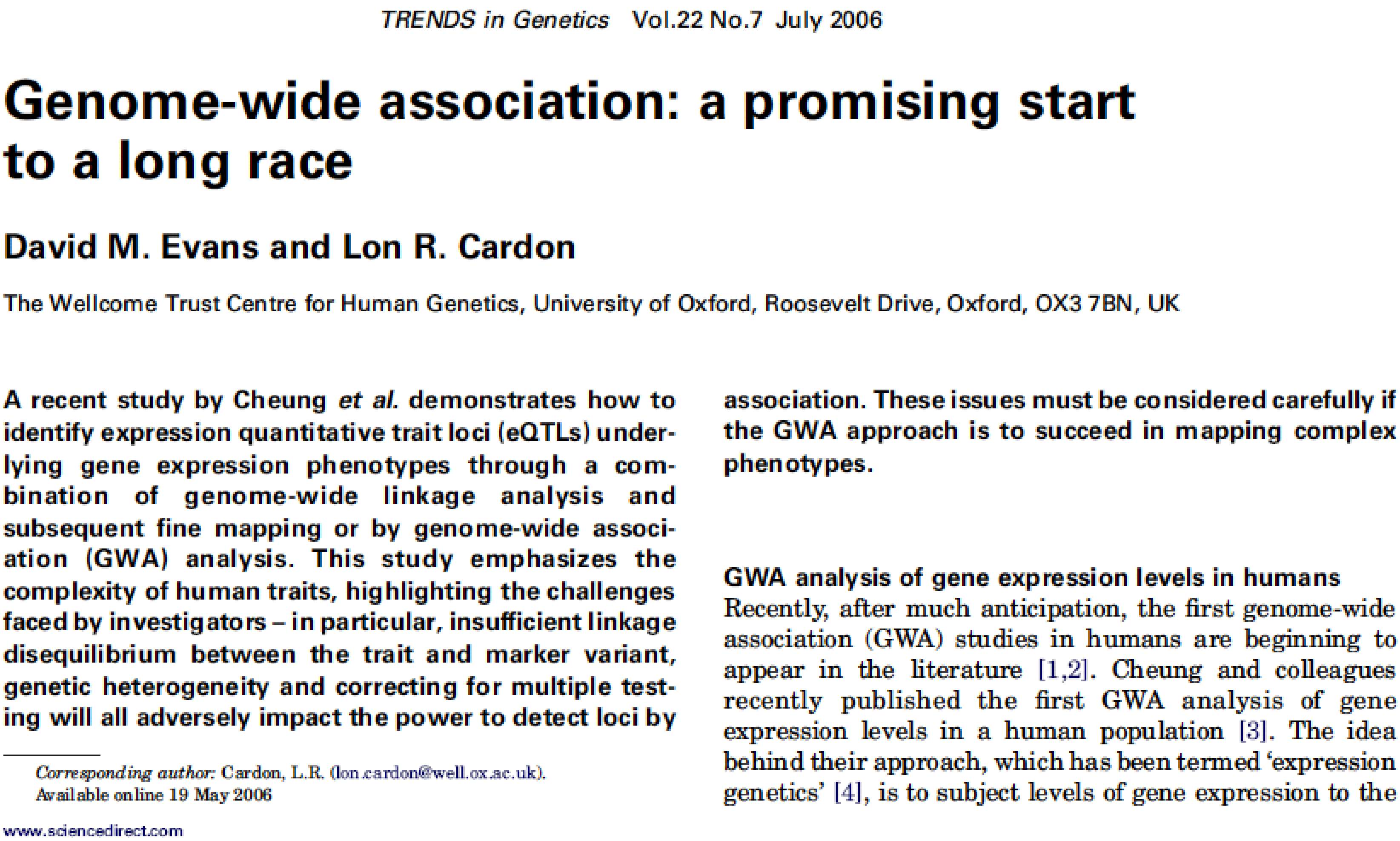
Recent Developments in Genomewide Association Scans: A Workshop Summary and Review

Duncan C. Thomas,[1] Robert W. Haile,[1] and David Duggan[2]

[1]Department of Preventive Medicine, University of Southern California, Los Angeles; and [2]Translational 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).



TRENDS in Genetics Vol.22 No.7 July 2006

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

Corresponding author: Cardon, L.R. (lon.cardon@well.ox.ac.uk).
Available online 19 May 2006

www.sciencedirect.com

References

1 Klein, R.J. *et al*. (2005) Complement factor H polymorphism in age-related macular degeneration. *Science* 308, 385–389

2 Ozaki, K. *et al*. (2002) Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. *Nat. Genet.* 32, 650–654

GWAS（genome-wide association) reports from RIKEN

1. Ozaki K et al. Functional SNPs in the lymphotoxin−alpha gene that are associated with susceptibility to myocardial infarction. Nat Genet. 2002 Dec;32(4):650−4.

2. Suzuki A et al. Functional haplotypes of PADI4, encoding citrullinating enzyme peptidylarginine deiminase 4, are associated with rheumatoid arthritis. Nat Genet. 2003 Aug;34(4):395−402.

3. Tokuhiro S et al.. An intronic SNP in a RUNX1 binding site of SLC22A4, encoding an organic cation transporter, is associated with rheumatoid arthritis. Nat Genet. 2003 Dec;35(4):341−8.

4. Ozaki K et al. Functional variation in LGALS2 confers risk of myocardial infarction and regulates lymphotoxin−alpha secretion in vitro. Nature. 2004 May 6;429(6987):72−5.

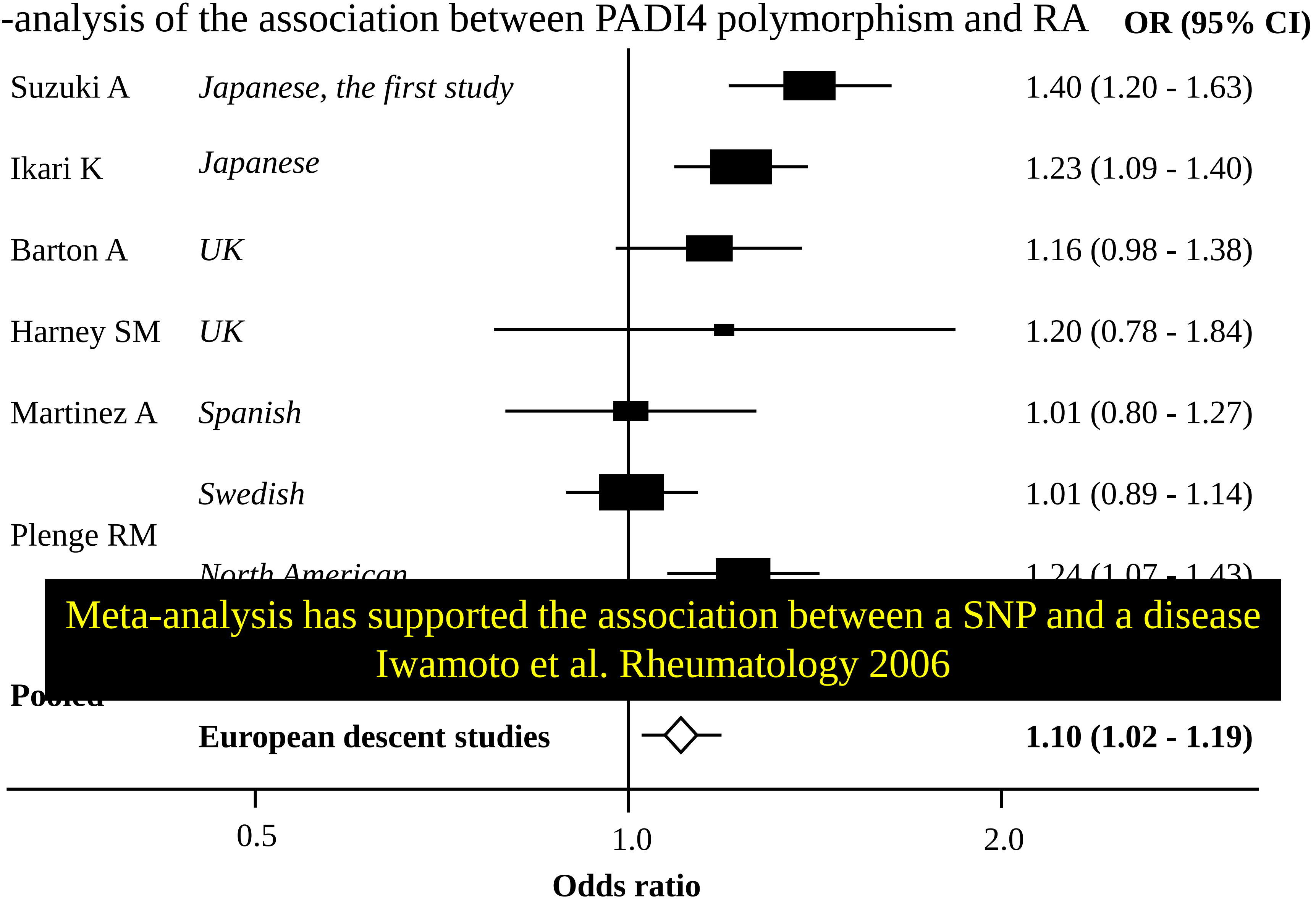
5: Kizawa H et al. An aspartic acid repeat polymorphism in asporin inhibits chondrogenesis and increases susceptibility to osteoarthritis. Nat Genet. 2005 Feb;37(2):138−44.

6. Kochi Y et al. A functional variant in FCRL3, encoding Fc receptor−like 3, is associated with rheumatoid arthritis and several autoimmunities. Nat Genet. 2005 May;37(5):478−85.

7. Seki S et al. A functional SNP in CILP, encoding cartilage intermediate layer protein, is associated with susceptibility to lumbar disc disease. Nat Genet. 2005 Jun;37(6):607−12.

8. Ozaki K et al. A functional SNP in PSMA6 confers risk of myocardial infarction in the Japanese population. Nat Genet. 2006 Aug;38(8):921−5.

Meta-analysis of the association between PADI4 polymorphism and RA



Meta-analysis has supported the association between a SNP and a disease
Iwamoto et al. Rheumatology 2006

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.

QC of large quantity of data

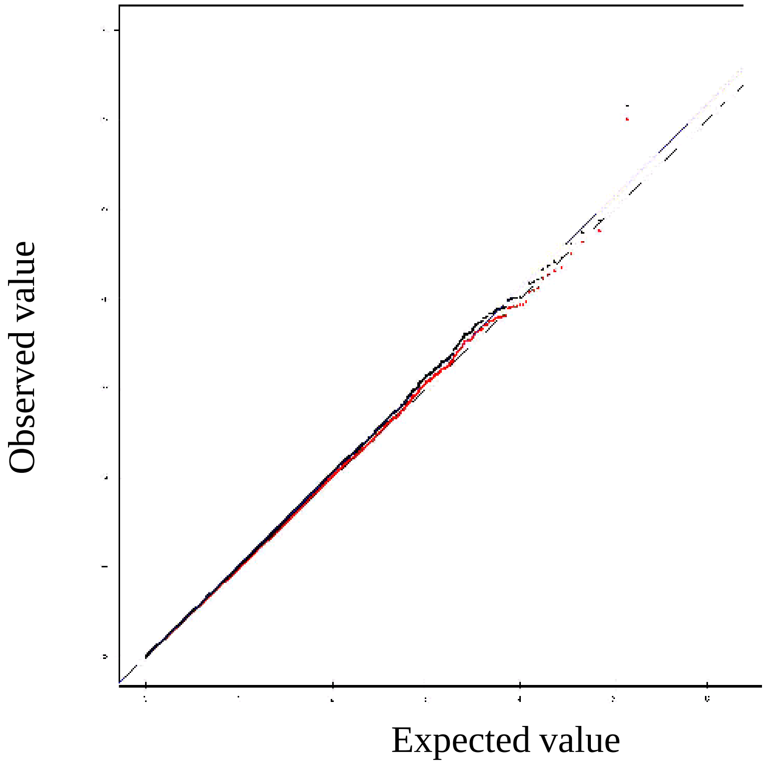
(500,000 SNP genotypes from > 10,000 subjects)

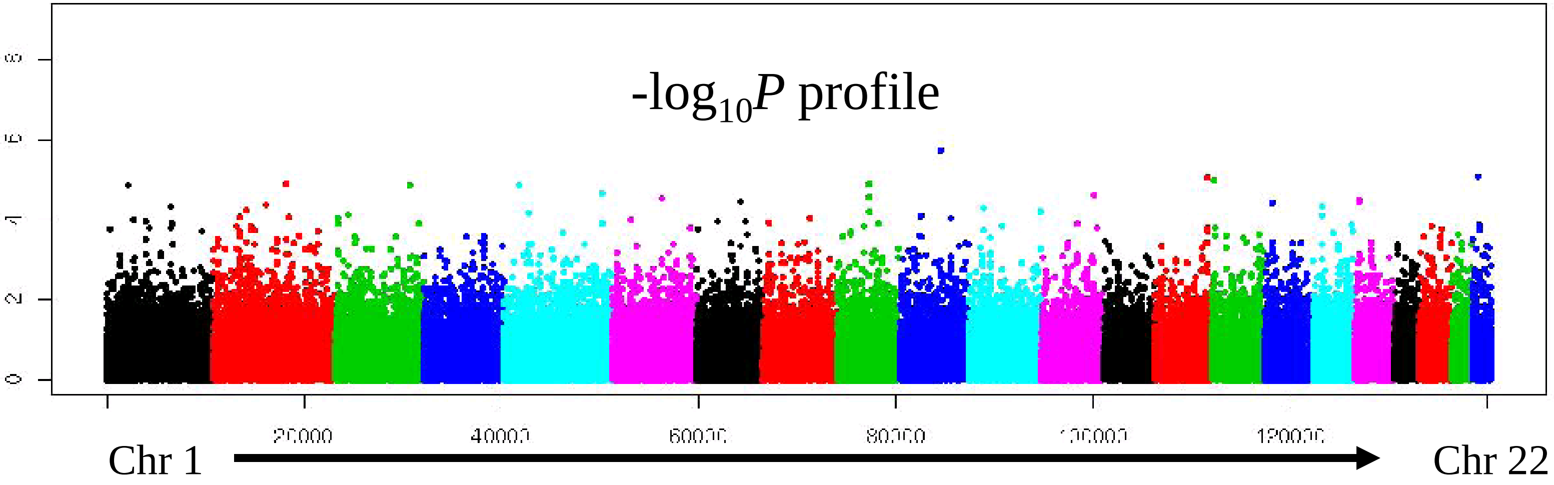
1. QC (quality control) is extremely laborious.
2. Mistypes lead to false significance.
3. We can use both genetics and statistics –based methods for QC.
4. Reliable conclusion from GWAS is dependent on sophisticated QC filter.

More than 10^{10} data points.

Report of the results of an association study

QQ plot





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

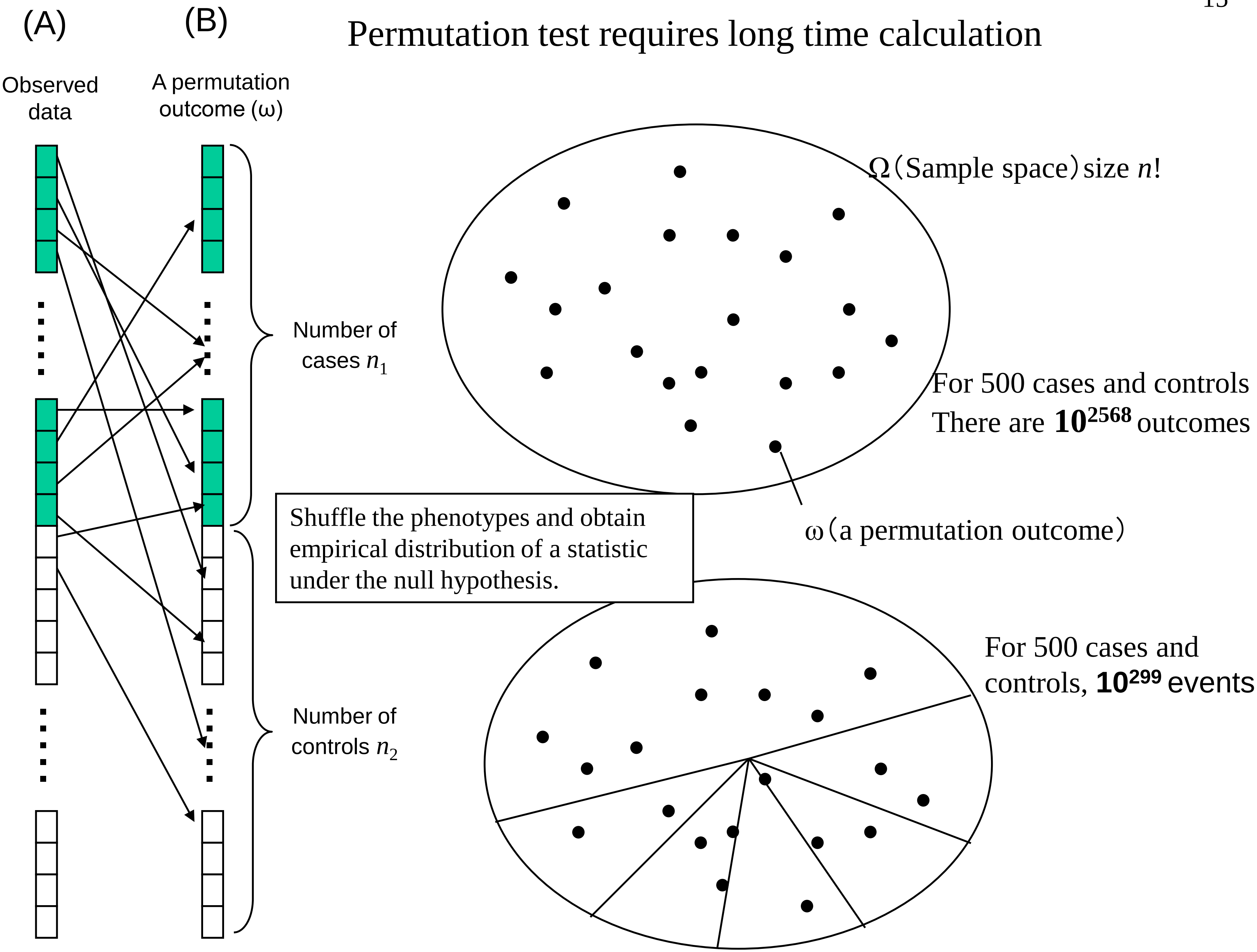
   (a) Use of the concept FDR (false-discovery rate)

   (b) Permutation test

   (c) Exact calculation of type 1 error rate

   (d) Bayesian method (FPRP)

Permutation test requires long time calculation



(A)

Observed data

(B)

A permutation outcome (ω)

Number of cases n_1

Number of controls n_2

Ω（Sample space）size $n!$

For 500 cases and controls There are **10^{2568}** outcomes

ω（a permutation outcome）

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

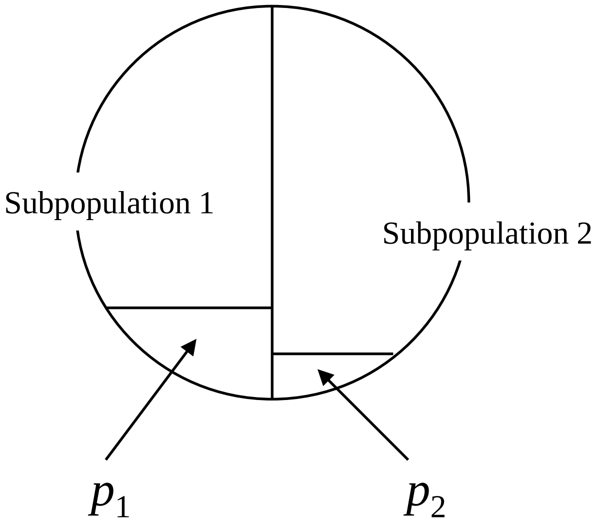
For 500 cases and controls, **10^{299}** events

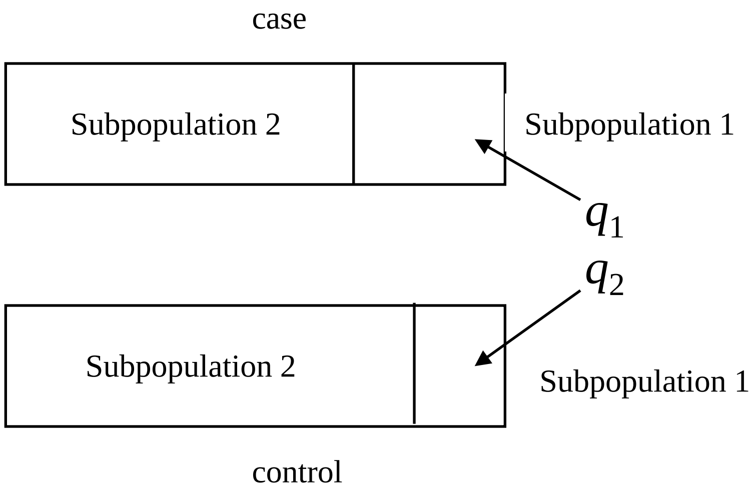
Problem of population structuring

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





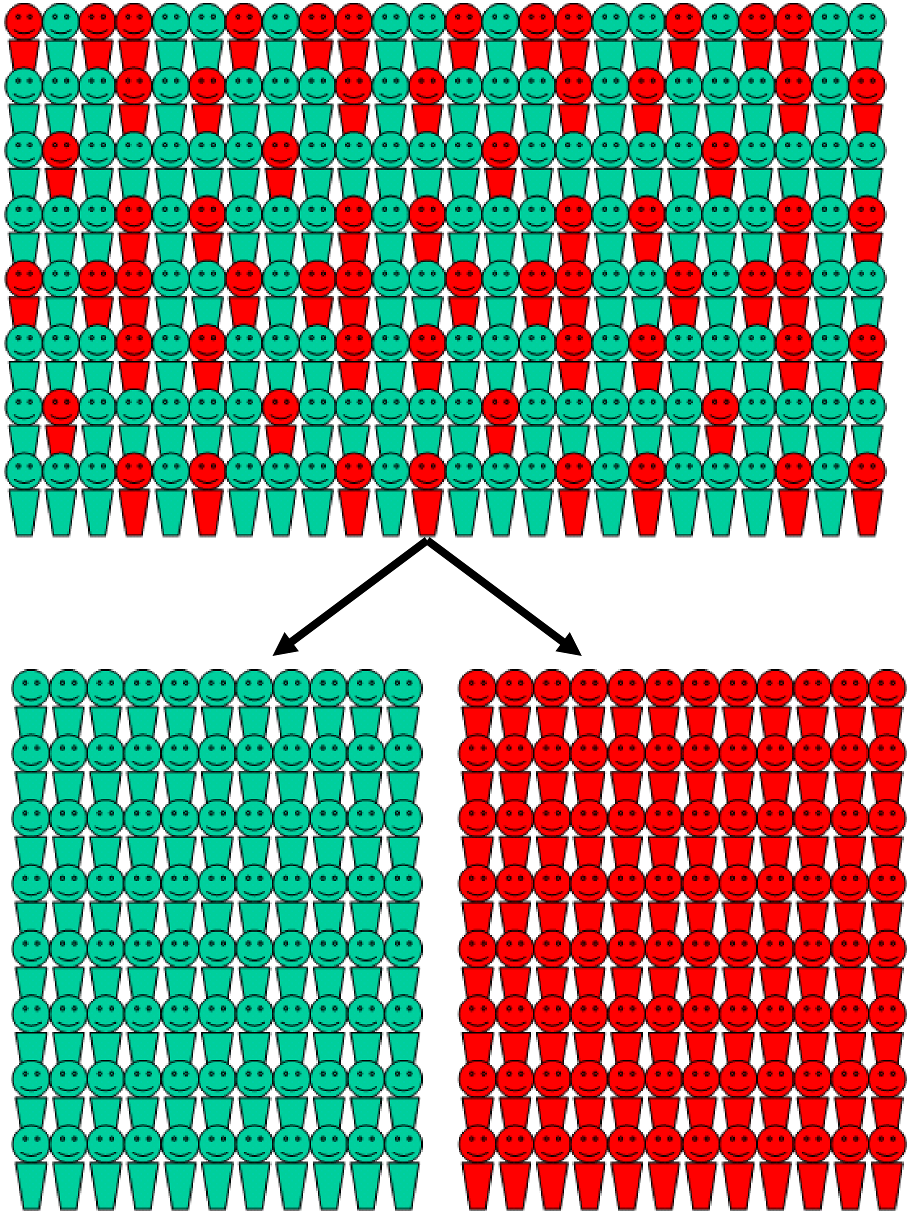
If allele or genotype frequencies are different between subpopulations, → and the prevalence of a disease is different between the subpopulations,

$$OR = \frac{q_1p_1 + (1-q_1)p_2}{(1-p_1)q_1 + (1-p_2)(1-q_1)} \times \frac{(1-p_1)q_2 + (1-q_2)(1-p_2)}{p_1q_2 + (1-q_2)p_2}$$

OR is 1 when $p_1=p_2$, or $q_1 = q_2$

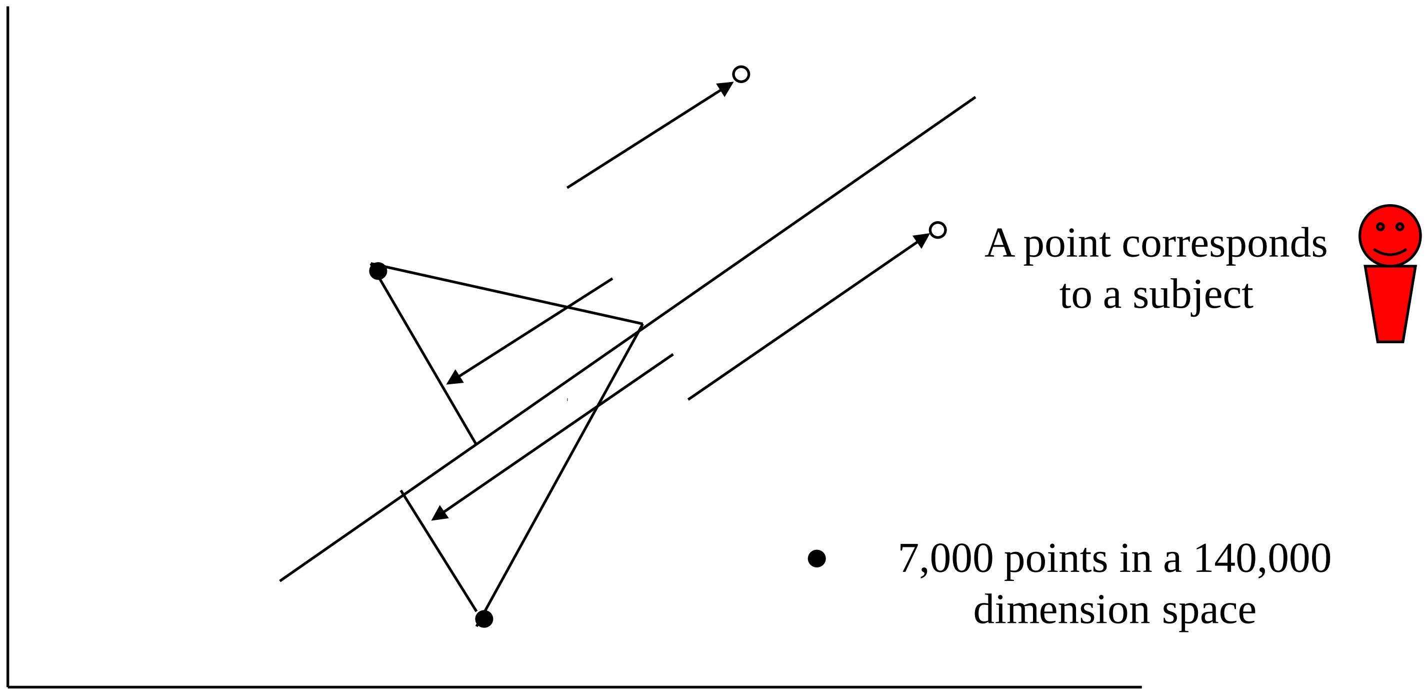
then, false-positive associations will occur.

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.



A point corresponds to a subject

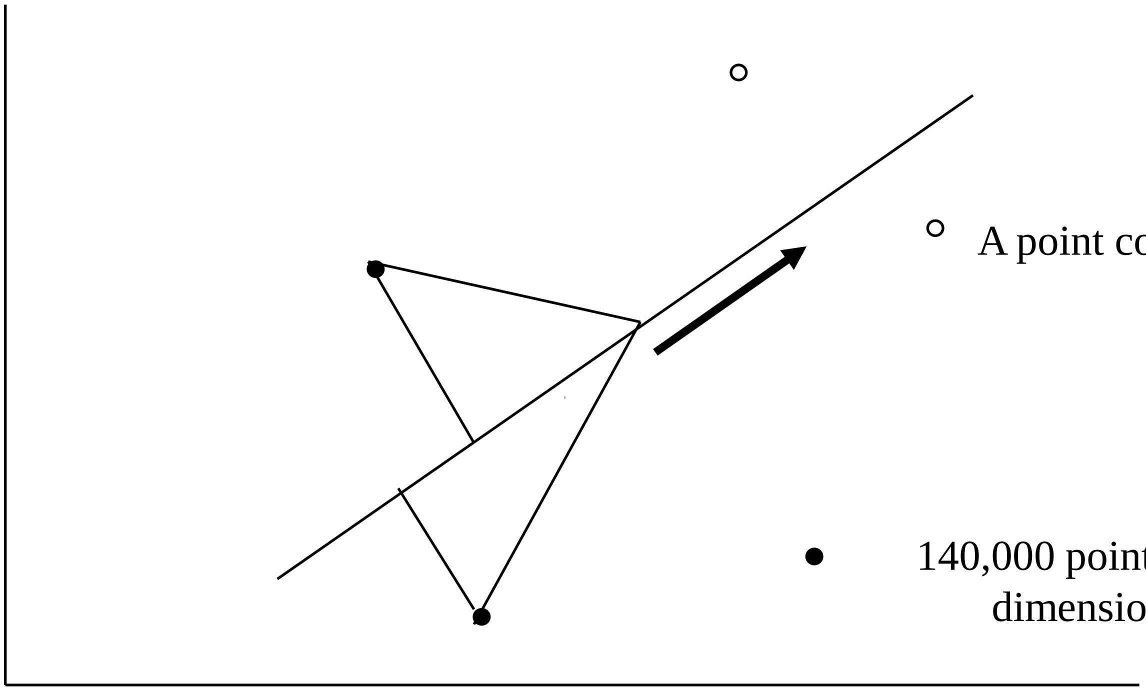
● 7,000 points in a 140,000 dimension space

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.



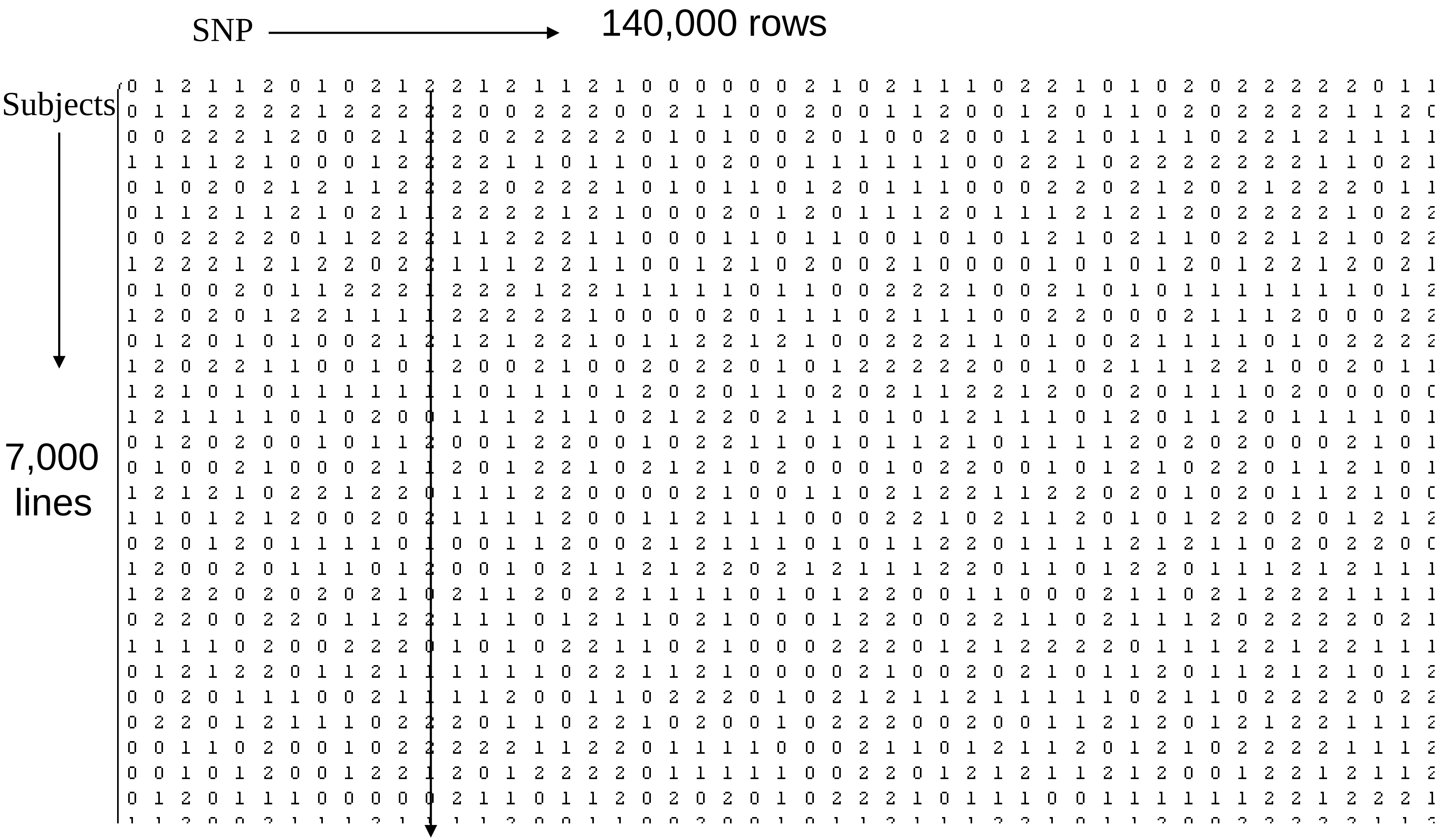
○ A point corresponds to a SNP

A/G

● 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



SNP

140,000 rows

Subjects

7,000 lines

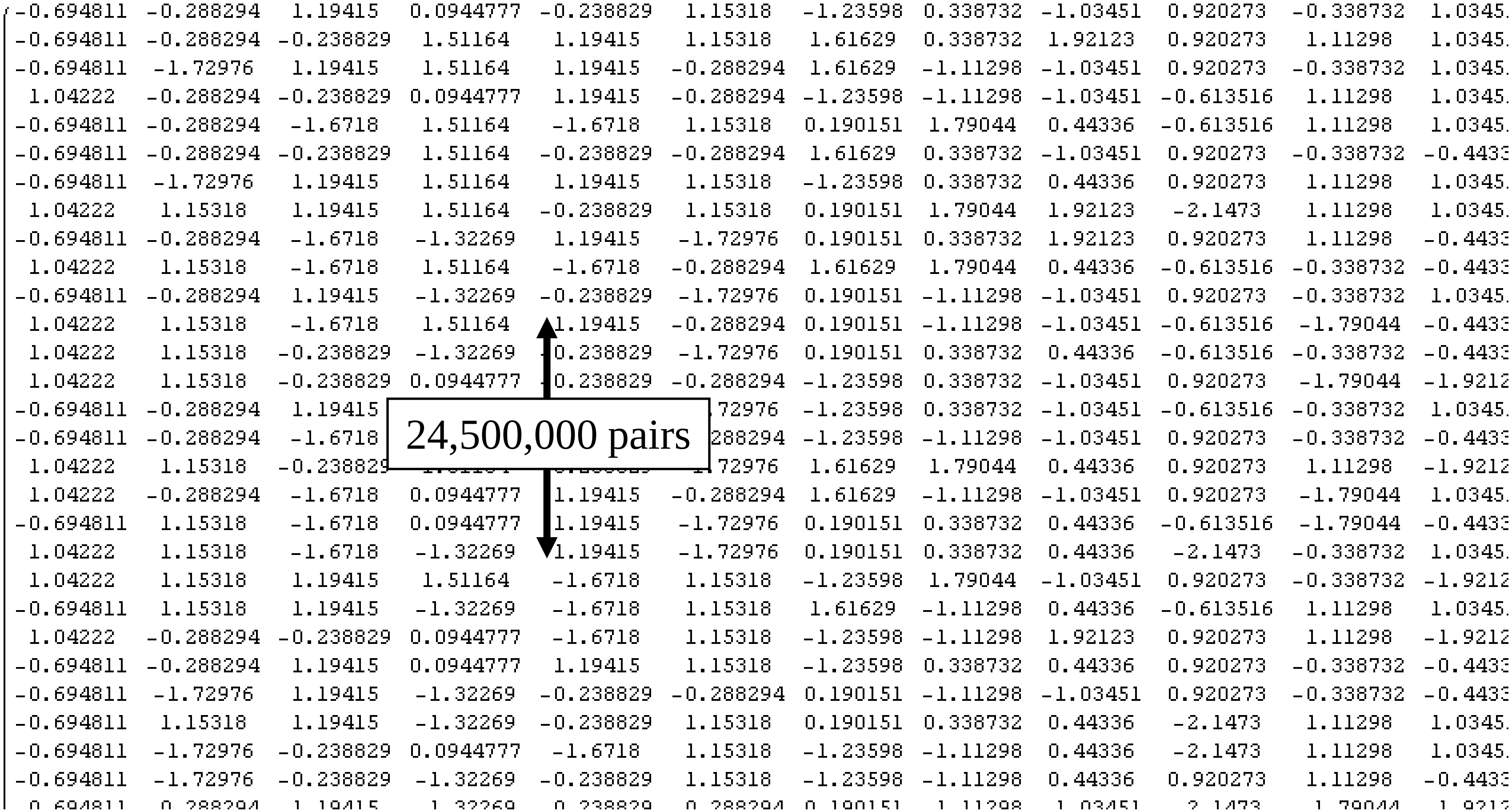
```
0 1 2 1 1 2 0 1 0 2 1 2 2 1 2 1 1 2 1 0 0 0 0 0 0 2 1 0 2 1 1 1 0 2 2 1 0 1 0 2 0 2 2 2 2 2 0 1 1
0 1 1 2 2 2 2 1 2 2 2 2 2 0 0 2 2 2 0 0 2 1 1 0 0 2 0 0 1 1 2 0 0 1 2 0 1 1 0 2 0 2 2 2 2 1 1 2 0
0 0 2 2 2 1 2 0 0 2 1 2 2 0 2 2 2 2 2 0 1 0 1 0 0 2 0 1 0 0 2 0 0 1 2 1 0 1 1 1 0 2 2 1 2 1 1 1 1
1 1 1 1 2 1 0 0 0 1 2 2 2 2 1 1 0 1 1 0 1 0 2 0 0 1 1 1 1 1 1 0 0 2 2 1 0 2 2 2 2 2 2 2 1 1 0 2 1
0 1 0 2 0 2 1 2 1 1 2 2 2 2 0 2 2 2 1 0 1 0 1 1 0 1 2 0 1 1 1 0 0 0 2 2 0 2 1 2 0 2 1 2 2 2 0 1 1
0 1 1 2 1 1 2 1 0 2 1 1 2 2 2 2 1 2 1 0 0 0 2 0 1 2 0 1 1 1 2 0 1 1 1 2 1 2 1 2 0 2 2 2 2 1 0 2 2
0 0 2 2 2 2 0 1 1 2 2 2 1 1 2 2 2 1 1 0 0 0 1 1 0 1 1 0 0 1 0 1 0 1 2 1 0 2 1 1 0 2 2 1 2 1 0 2 2
1 2 2 2 1 2 1 2 2 0 2 2 1 1 1 2 2 1 1 0 0 1 2 1 0 2 0 0 2 1 0 0 0 0 1 0 1 0 1 2 0 1 2 2 1 2 0 2 1
0 1 0 0 2 0 1 1 2 2 2 1 2 2 2 1 2 2 1 1 1 1 1 0 1 1 0 0 2 2 2 1 0 0 2 1 0 1 0 1 1 1 1 1 1 1 0 1 2
1 2 0 2 0 1 2 2 1 1 1 1 2 2 2 2 2 1 0 0 0 0 2 0 1 1 1 0 2 1 1 1 0 0 2 2 0 0 0 2 1 1 1 2 0 0 0 2 2
0 1 2 0 1 0 1 0 0 2 1 2 1 2 1 2 2 1 0 1 1 2 2 1 2 1 0 0 2 2 2 1 1 0 1 0 0 2 1 1 1 1 0 1 0 2 2 2 2
1 2 0 2 2 1 1 0 0 1 0 1 2 0 0 2 1 0 0 2 0 2 2 0 1 0 1 2 2 2 2 2 0 0 1 0 2 1 1 1 2 2 1 0 0 2 0 1 1
1 2 1 0 1 0 1 1 1 1 1 1 1 0 1 1 1 0 1 2 0 2 0 1 1 0 2 0 2 1 1 2 2 1 2 0 0 2 0 1 1 1 0 2 0 0 0 0 0
1 2 1 1 1 1 0 1 0 2 0 0 1 1 1 2 1 1 0 2 1 2 2 0 2 1 1 0 1 0 1 2 1 1 1 0 1 2 0 1 1 2 0 1 1 1 1 0 1
0 1 2 0 2 0 0 1 0 1 1 2 0 0 1 2 2 0 0 1 0 2 2 1 1 0 1 0 1 1 2 1 0 1 1 1 1 2 0 2 0 2 0 0 0 2 1 0 1
0 1 0 0 2 1 0 0 0 2 1 1 2 0 1 2 2 1 0 2 1 2 1 0 2 0 0 0 1 0 2 2 0 0 1 0 1 2 1 0 2 2 0 1 1 2 1 0 1
1 2 1 2 1 0 2 2 1 2 2 0 1 1 1 2 2 0 0 0 0 2 1 0 0 1 1 0 2 1 2 2 1 1 2 2 0 2 0 1 0 2 0 1 1 2 1 0 0
1 1 0 1 2 1 2 0 0 2 0 2 1 1 1 1 2 0 0 1 1 2 1 1 1 0 0 0 2 2 1 0 2 1 1 2 0 1 0 1 2 2 0 2 0 1 2 1 2
0 2 0 1 2 0 1 1 1 1 0 1 0 0 1 1 2 0 0 2 1 2 1 1 1 0 1 0 1 1 2 2 0 1 1 1 1 2 1 2 1 1 0 2 0 2 2 0 0
1 2 0 0 2 0 1 1 1 0 1 2 0 0 1 0 2 1 1 2 1 2 2 0 2 1 2 1 1 1 2 2 0 1 1 0 1 2 2 0 1 1 1 2 1 2 1 1 1
1 2 2 2 0 2 0 2 0 2 1 0 2 1 1 2 0 2 2 1 1 1 1 0 1 0 1 2 2 0 0 1 1 0 0 0 2 1 1 0 2 1 2 2 2 1 1 1 1
0 2 2 0 0 2 2 0 1 1 2 2 1 1 1 0 1 2 1 1 0 2 1 0 0 0 1 2 2 0 0 2 2 1 1 0 2 1 1 1 2 0 2 2 2 2 0 2 1
1 1 1 1 0 2 0 0 2 2 2 0 1 0 1 0 2 2 1 1 0 2 1 0 0 0 2 2 2 0 1 2 1 2 2 2 2 0 1 1 1 2 2 1 2 2 1 1 1
0 1 2 1 2 2 0 1 1 2 1 1 1 1 1 1 0 2 2 1 1 2 1 0 0 0 0 2 1 0 0 2 0 2 1 0 1 1 2 0 1 1 2 1 2 1 0 1 2
0 0 2 0 1 1 1 0 0 2 1 1 1 1 2 0 0 1 1 0 2 2 2 0 1 0 2 1 2 1 1 2 1 1 1 1 1 0 2 1 1 0 2 2 2 2 0 2 2
0 2 2 0 1 2 1 1 1 0 2 2 2 0 1 1 0 2 2 1 0 2 0 0 1 0 2 2 2 0 0 2 0 0 1 1 2 1 2 0 1 2 1 2 2 1 1 1 2
0 0 1 1 0 2 0 0 1 0 2 2 2 2 2 1 1 2 2 0 1 1 1 1 0 0 0 2 1 1 0 1 2 1 1 2 0 1 2 1 0 2 2 2 2 1 1 1 2
0 0 1 0 1 2 0 0 1 2 2 1 2 0 1 2 2 2 2 0 1 1 1 1 1 0 0 2 2 0 1 2 1 2 1 1 2 1 2 0 0 1 2 2 1 2 1 1 2
0 1 2 0 1 1 1 0 0 0 0 0 2 1 1 0 1 1 2 0 2 0 2 0 1 0 2 2 2 1 0 1 1 1 0 0 1 1 1 1 1 1 2 2 1 2 2 2 1
1 1 2 0 0 2 1 1 1 2 1 1 1 1 2 0 0 1 1 0 0 2 0 0 1 0 1 1 2 1 1 1 2 2 1 0 1 1 2 0 0 2 2 2 2 2 1 1 2
```

Normalize for each SNP (mean *p*, variance 2 *p* (1-*p*))

Normalized genotype data

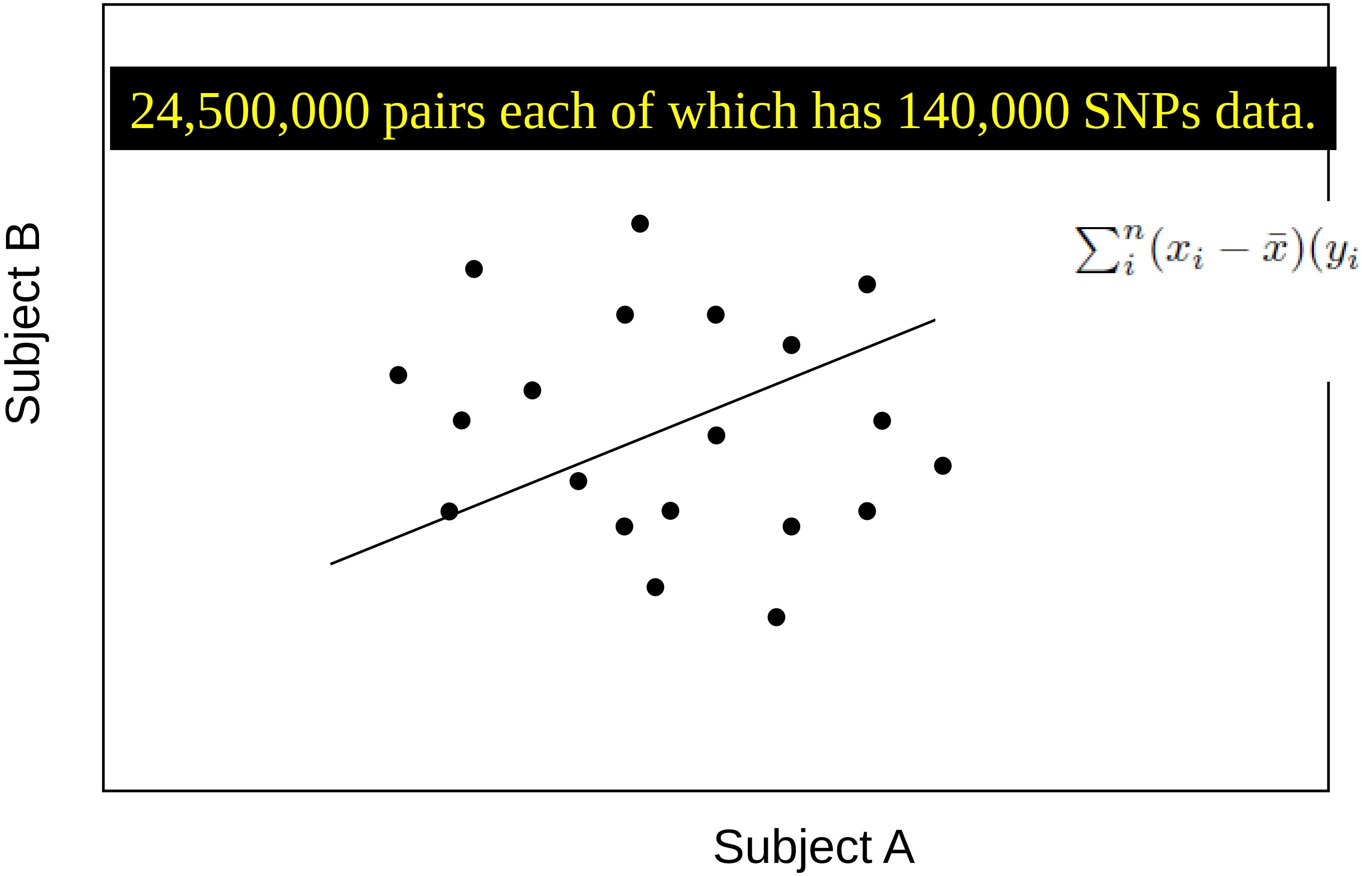
SNP → 140,000 rows

Subjects ↓ 7,000 lines



```
-0.694811  -0.288294   1.19415    0.0944777  -0.238829   1.15318   -1.23598   0.338732  -1.03451   0.920273  -0.338732  1.0345
-0.694811  -0.288294  -0.238829   1.51164     1.19415    1.15318    1.61629   0.338732   1.92123   0.920273   1.11298   1.0345
-0.694811  -1.72976    1.19415    1.51164     1.19415   -0.288294   1.61629  -1.11298   -1.03451   0.920273  -0.338732  1.0345
 1.04222   -0.288294  -0.238829   0.0944777   1.19415   -0.288294  -1.23598  -1.11298   -1.03451  -0.613516   1.11298   1.0345
-0.694811  -0.288294  -1.6718     1.51164    -1.6718     1.15318    0.190151  1.79044    0.44336  -0.613516   1.11298   1.0345
-0.694811  -0.288294  -0.238829   1.51164    -0.238829  -0.288294   1.61629   0.338732  -1.03451   0.920273  -0.338732 -0.4433
-0.694811  -1.72976    1.19415    1.51164     1.19415    1.15318   -1.23598   0.338732   0.44336   0.920273   1.11298   1.0345
 1.04222    1.15318    1.19415    1.51164    -0.238829   1.15318    0.190151  1.79044    1.92123  -2.1473     1.11298   1.0345
-0.694811  -0.288294  -1.6718    -1.32269     1.19415   -1.72976    0.190151  0.338732   1.92123   0.920273   1.11298  -0.4433
 1.04222    1.15318   -1.6718     1.51164    -1.6718    -0.288294   1.61629   1.79044    0.44336  -0.613516  -0.338732 -0.4433
-0.694811  -0.288294   1.19415   -1.32269    -0.238829  -1.72976    0.190151 -1.11298   -1.03451   0.920273  -0.338732  1.0345
 1.04222    1.15318   -1.6718     1.51164     1.19415   -0.288294   0.190151 -1.11298   -1.03451  -0.613516  -1.79044  -0.4433
 1.04222    1.15318   -0.238829  -1.32269    -0.238829  -1.72976    0.190151  0.338732   0.44336  -0.613516  -0.338732 -0.4433
 1.04222    1.15318   -0.238829   0.0944777  -0.238829  -0.288294  -1.23598   0.338732  -1.03451   0.920273  -1.79044  -1.9212
-0.694811  -0.288294   1.19415                           .72976   -1.23598   0.338732  -1.03451  -0.613516  -0.338732  1.0345
-0.694811  -0.288294  -1.6718                            288294   -1.23598  -1.11298   -1.03451   0.920273  -0.338732 -0.4433
 1.04222    1.15318   -0.238829                          .72976    1.61629   1.79044    0.44336   0.920273   1.11298  -1.9212
 1.04222   -0.288294  -1.6718     0.0944777   1.19415   -0.288294   1.61629  -1.11298   -1.03451   0.920273  -1.79044   1.0345
-0.694811   1.15318   -1.6718     0.0944777   1.19415   -1.72976    0.190151  0.338732   0.44336  -0.613516  -1.79044  -0.4433
 1.04222    1.15318   -1.6718    -1.32269     1.19415   -1.72976    0.190151  0.338732   0.44336  -2.1473    -0.338732  1.0345
 1.04222    1.15318    1.19415    1.51164    -1.6718     1.15318   -1.23598   1.79044   -1.03451   0.920273  -0.338732 -1.9212
-0.694811   1.15318    1.19415   -1.32269    -1.6718     1.15318    1.61629  -1.11298    0.44336  -0.613516   1.11298   1.0345
 1.04222   -0.288294  -0.238829   0.0944777  -1.6718     1.15318   -1.23598  -1.11298    1.92123   0.920273   1.11298  -1.9212
-0.694811  -0.288294   1.19415    0.0944777   1.19415    1.15318   -1.23598   0.338732   0.44336   0.920273  -0.338732 -0.4433
-0.694811  -1.72976    1.19415   -1.32269    -0.238829  -0.288294   0.190151 -1.11298   -1.03451   0.920273  -0.338732 -0.4433
-0.694811   1.15318    1.19415   -1.32269    -0.238829   1.15318    0.190151  0.338732   0.44336  -2.1473     1.11298   1.0345
-0.694811  -1.72976   -0.238829   0.0944777  -1.6718     1.15318   -1.23598  -1.11298    0.44336  -2.1473     1.11298   1.0345
-0.694811  -1.72976   -0.238829  -1.32269    -0.238829   1.15318   -1.23598  -1.11298    0.44336   0.920273   1.11298  -0.4433
 0.694811   0.288294   1.19415    1.32269     0.238829   0.288294   0.190151  1.11298    1.03451   2.1473     1.79044   1.9212
```

24,500,000 pairs



24,500,000 pairs each of which has 140,000 SNPs data.

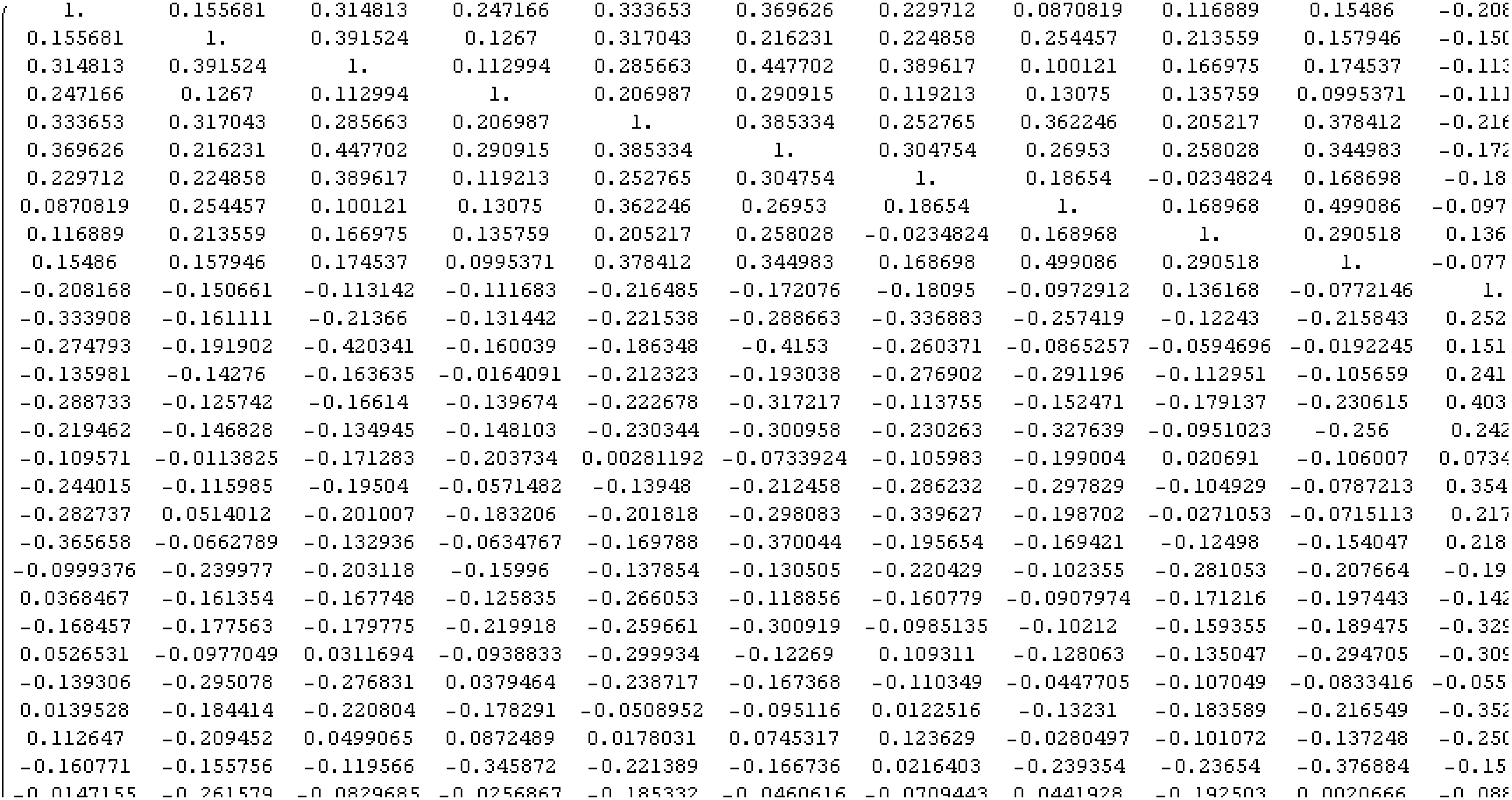
$$\sum_i^n (x_i - \bar{x})(y_i - \bar{y})$$

Covariance is calculated for each of 24,500,000 pairs

Covariance matrix

Subjects 7,000 →

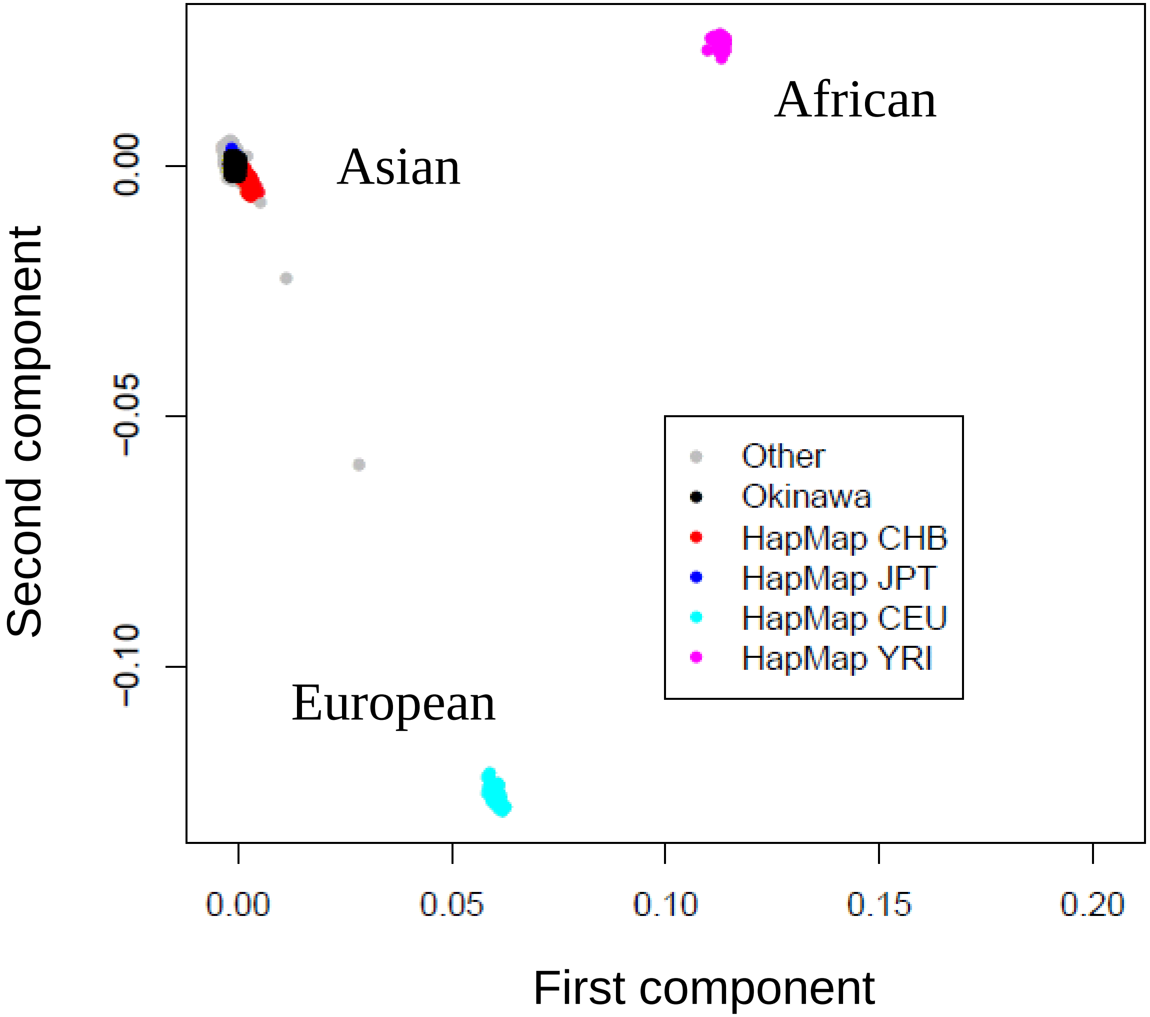
Subjects ↓



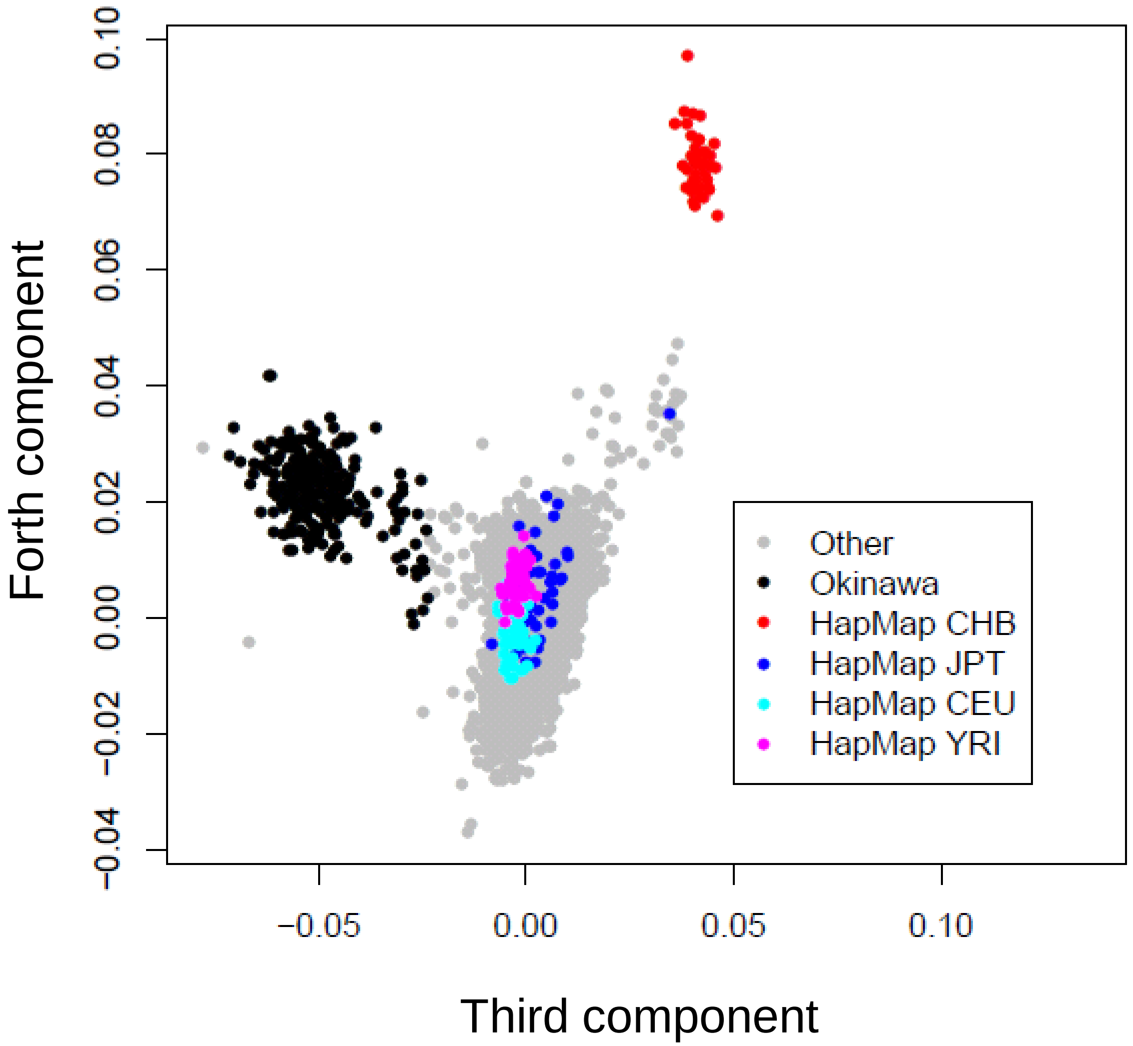
1.	0.155681	0.314813	0.247166	0.333653	0.369626	0.229712	0.0870819	0.116889	0.15486	-0.208
0.155681	1.	0.391524	0.1267	0.317043	0.216231	0.224858	0.254457	0.213559	0.157946	-0.150
0.314813	0.391524	1.	0.112994	0.285663	0.447702	0.389617	0.100121	0.166975	0.174537	-0.113
0.247166	0.1267	0.112994	1.	0.206987	0.290915	0.119213	0.13075	0.135759	0.0995371	-0.111
0.333653	0.317043	0.285663	0.206987	1.	0.385334	0.252765	0.362246	0.205217	0.378412	-0.216
0.369626	0.216231	0.447702	0.290915	0.385334	1.	0.304754	0.26953	0.258028	0.344983	-0.172
0.229712	0.224858	0.389617	0.119213	0.252765	0.304754	1.	0.18654	-0.0234824	0.168698	-0.18
0.0870819	0.254457	0.100121	0.13075	0.362246	0.26953	0.18654	1.	0.168968	0.499086	-0.097
0.116889	0.213559	0.166975	0.135759	0.205217	0.258028	-0.0234824	0.168968	1.	0.290518	0.136
0.15486	0.157946	0.174537	0.0995371	0.378412	0.344983	0.168698	0.499086	0.290518	1.	-0.077
-0.208168	-0.150661	-0.113142	-0.111683	-0.216485	-0.172076	-0.18095	-0.0972912	0.136168	-0.0772146	1.
-0.333908	-0.161111	-0.21366	-0.131442	-0.221538	-0.288663	-0.336883	-0.257419	-0.12243	-0.215843	0.252
-0.274793	-0.191902	-0.420341	-0.160039	-0.186348	-0.4153	-0.260371	-0.0865257	-0.0594696	-0.0192245	0.151
-0.135981	-0.14276	-0.163635	-0.0164091	-0.212323	-0.193038	-0.276902	-0.291196	-0.112951	-0.105659	0.241
-0.288733	-0.125742	-0.16614	-0.139674	-0.222678	-0.317217	-0.113755	-0.152471	-0.179137	-0.230615	0.403
-0.219462	-0.146828	-0.134945	-0.148103	-0.230344	-0.300958	-0.230263	-0.327639	-0.0951023	-0.256	0.242
-0.109571	-0.0113825	-0.171283	-0.203734	0.00281192	-0.0733924	-0.105983	-0.199004	0.020691	-0.106007	0.0734
-0.244015	-0.115985	-0.19504	-0.0571482	-0.13948	-0.212458	-0.286232	-0.297829	-0.104929	-0.0787213	0.354
-0.282737	0.0514012	-0.201007	-0.183206	-0.201818	-0.298083	-0.339627	-0.198702	-0.0271053	-0.0715113	0.217
-0.365658	-0.0662789	-0.132936	-0.0634767	-0.169788	-0.370044	-0.195654	-0.169421	-0.12498	-0.154047	0.218
-0.0999376	-0.239977	-0.203118	-0.15996	-0.137854	-0.130505	-0.220429	-0.102355	-0.281053	-0.207664	-0.19
0.0368467	-0.161354	-0.167748	-0.125835	-0.266053	-0.118856	-0.160779	-0.0907974	-0.171216	-0.197443	-0.142
-0.168457	-0.177563	-0.179775	-0.219918	-0.259661	-0.300919	-0.0985135	-0.10212	-0.159355	-0.189475	-0.329
0.0526531	-0.0977049	0.0311694	-0.0938833	-0.299934	-0.12269	0.109311	-0.128063	-0.135047	-0.294705	-0.309
-0.139306	-0.295078	-0.276831	0.0379464	-0.238717	-0.167368	-0.110349	-0.0447705	-0.107049	-0.0833416	-0.055
0.0139528	-0.184414	-0.220804	-0.178291	-0.0508952	-0.095116	0.0122516	-0.13231	-0.183589	-0.216549	-0.352
0.112647	-0.209452	0.0499065	0.0872489	0.0178031	0.0745317	0.123629	-0.0280497	-0.101072	-0.137248	-0.250
-0.160771	-0.155756	-0.119566	-0.345872	-0.221389	-0.166736	0.0216403	-0.239354	-0.23654	-0.376884	-0.15
-0.0147155	-0.261579	-0.0829685	-0.0256867	-0.185332	-0.0460616	-0.0709443	0.0441928	-0.192503	0.0020666	-0.088

Calculate Eigenvectors for this table

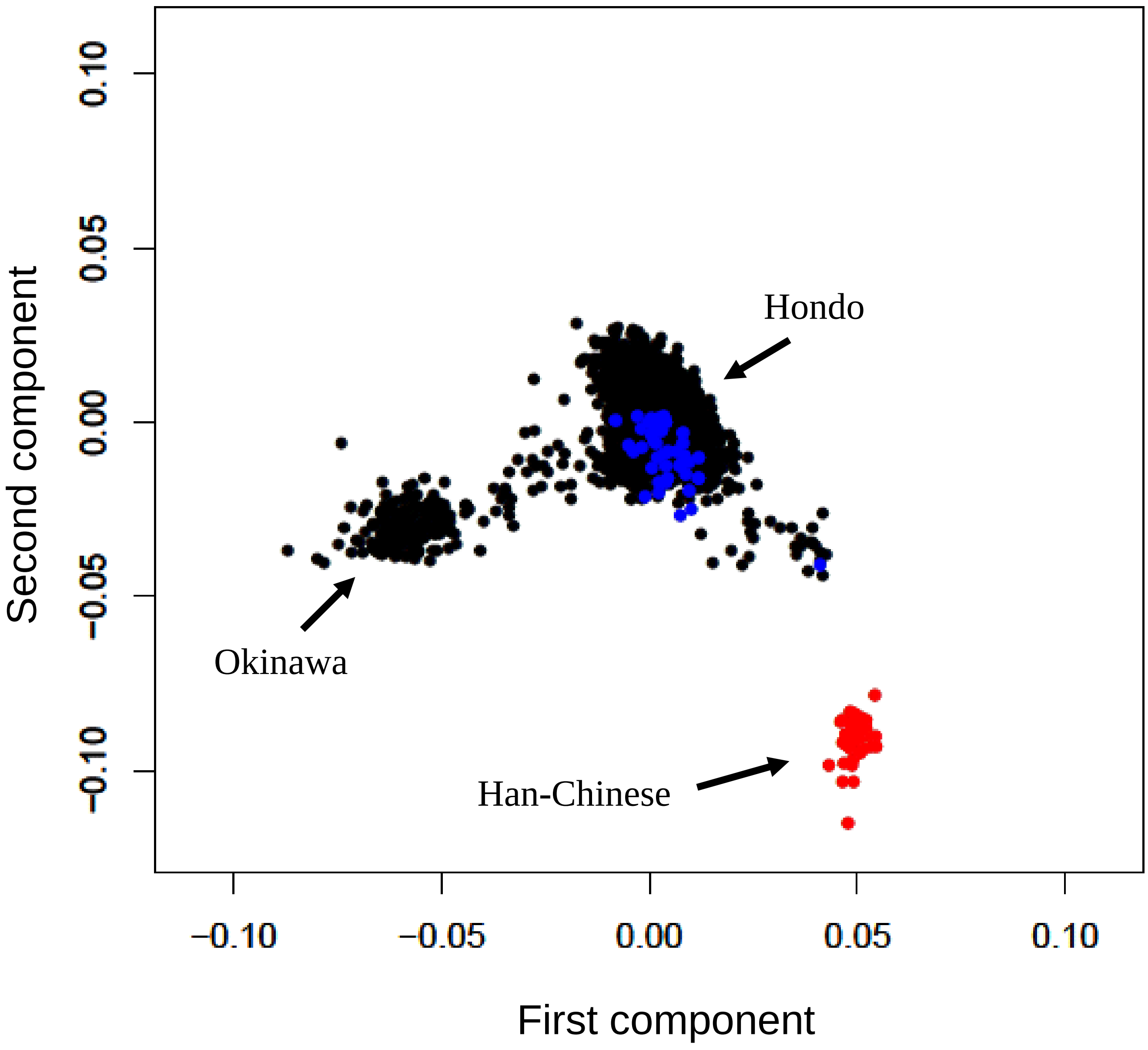
PCA analysis for African, European and Asian subjects



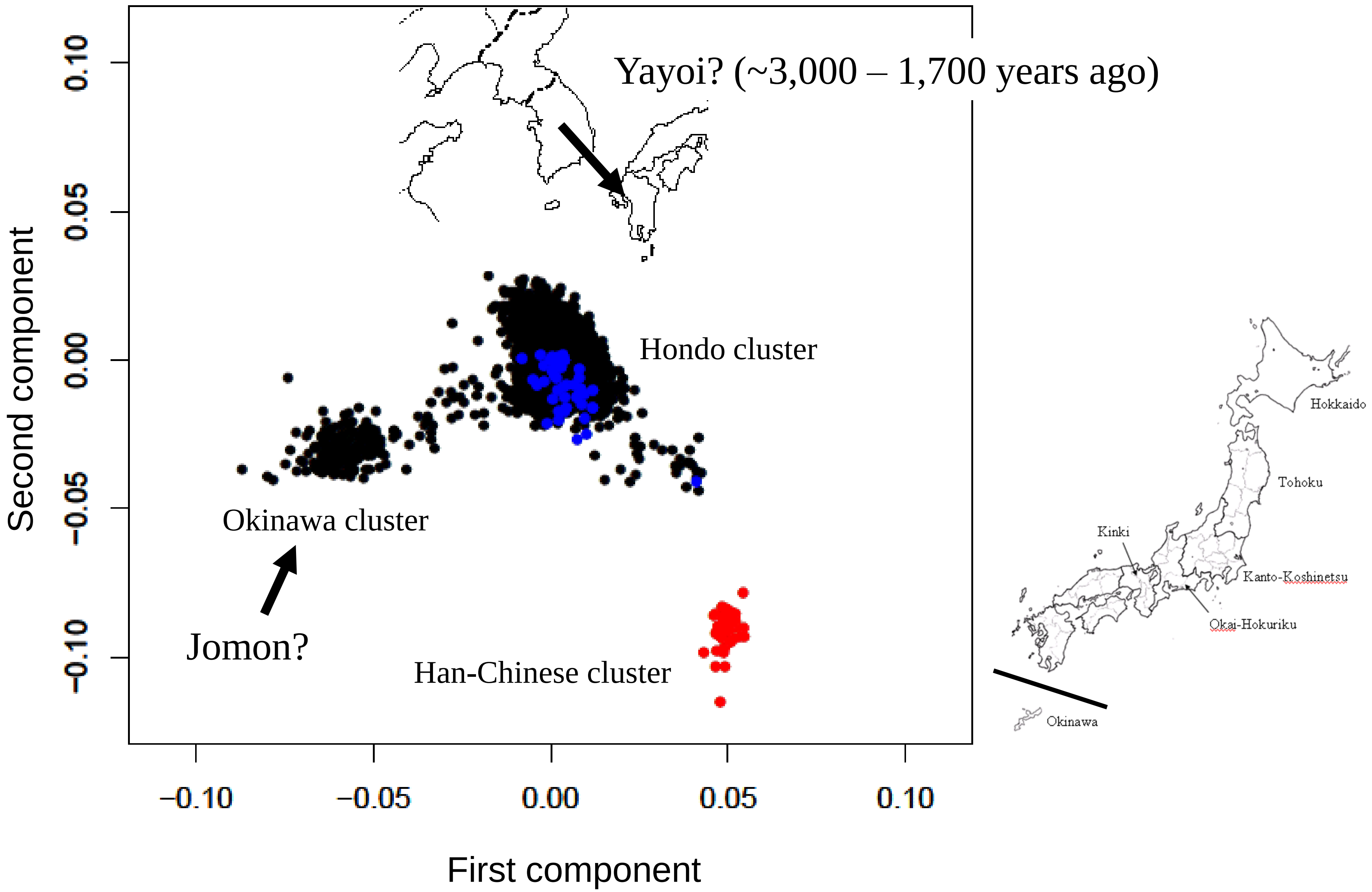
PCA analysis for African, European and Asian subjects

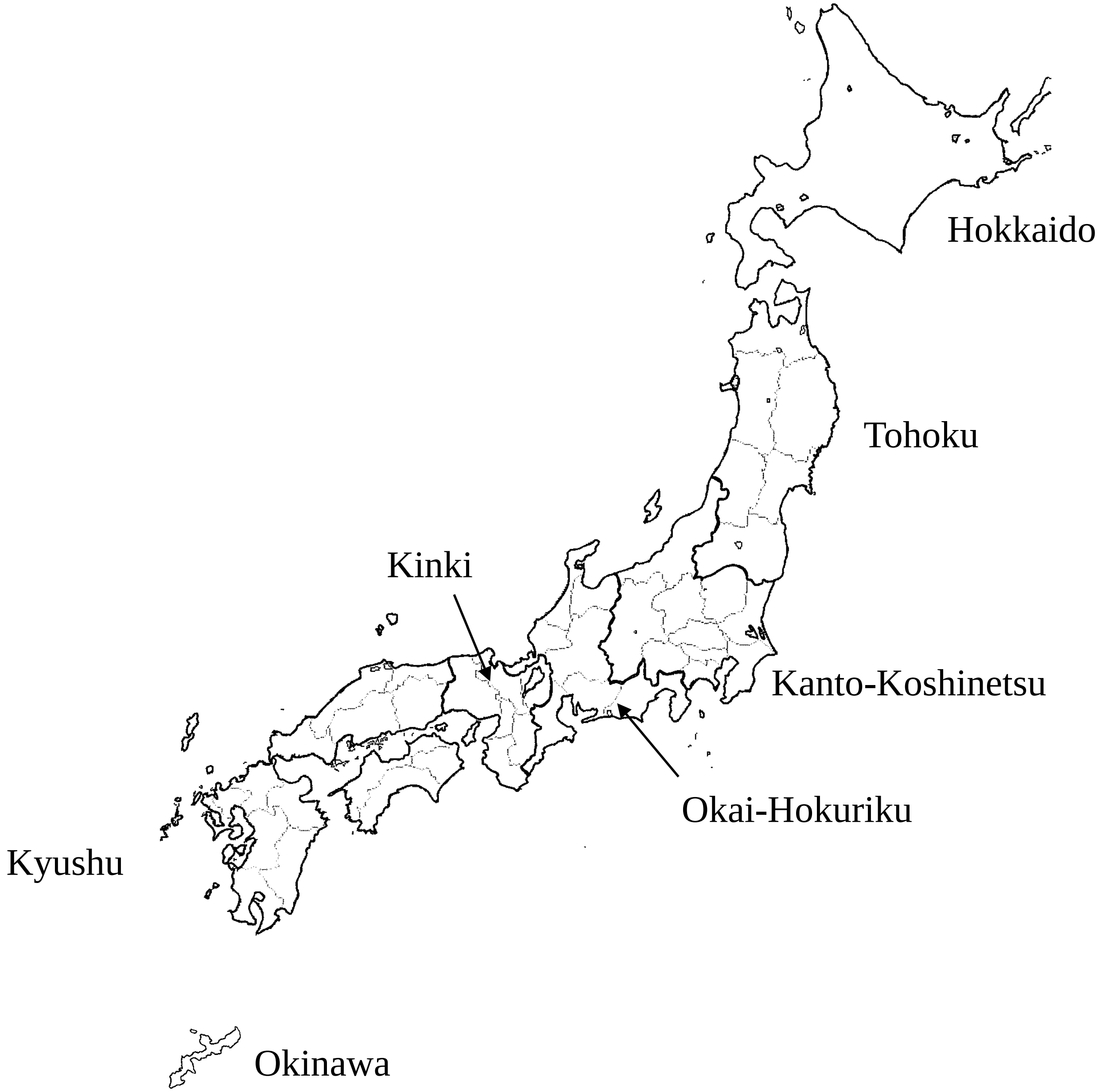


PCA analysis for Asian subjects

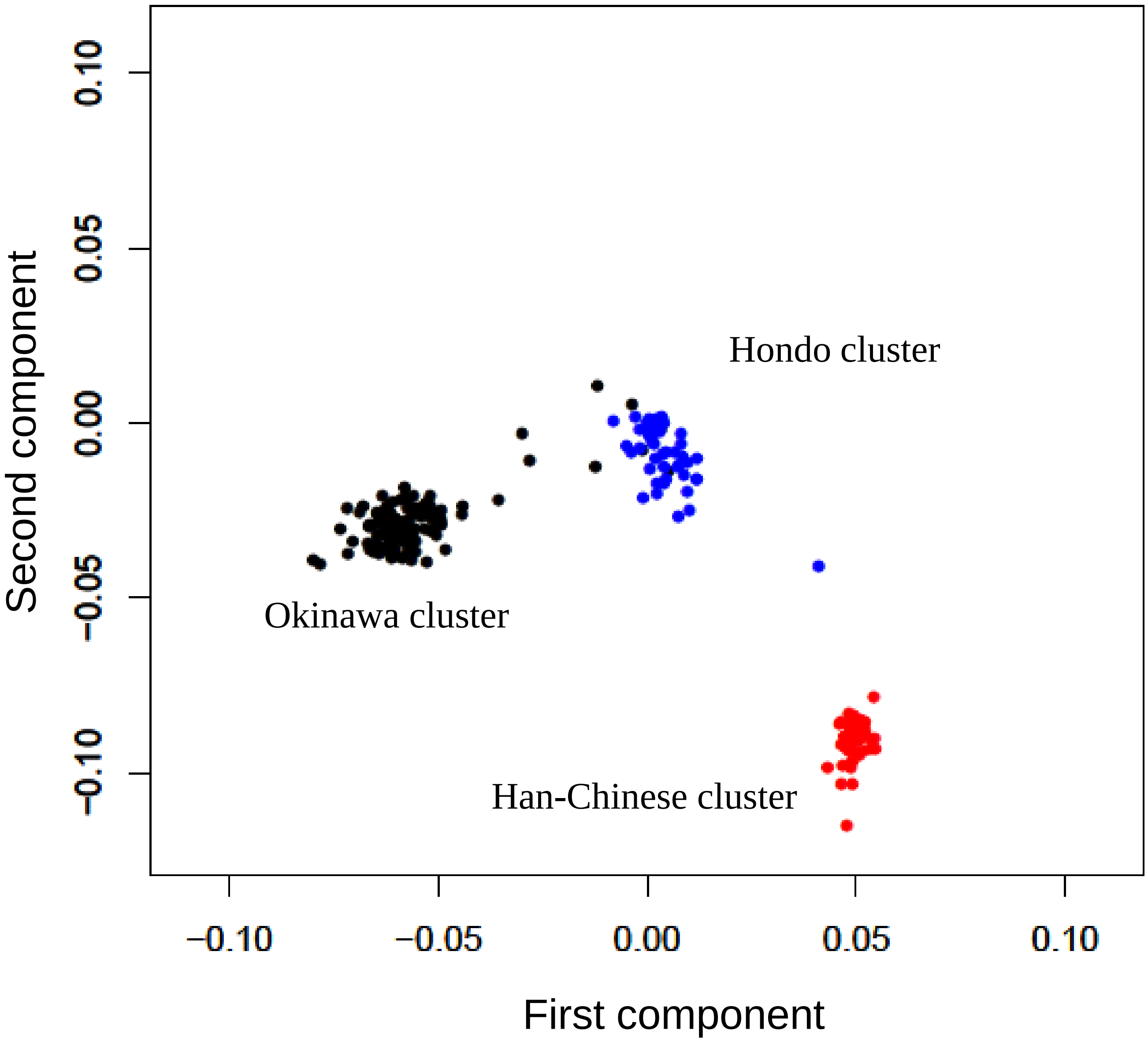


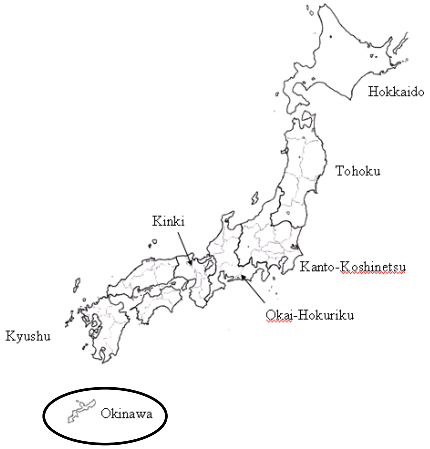
All Japanese samples + HapMap Han-Chinese and Japanese samples



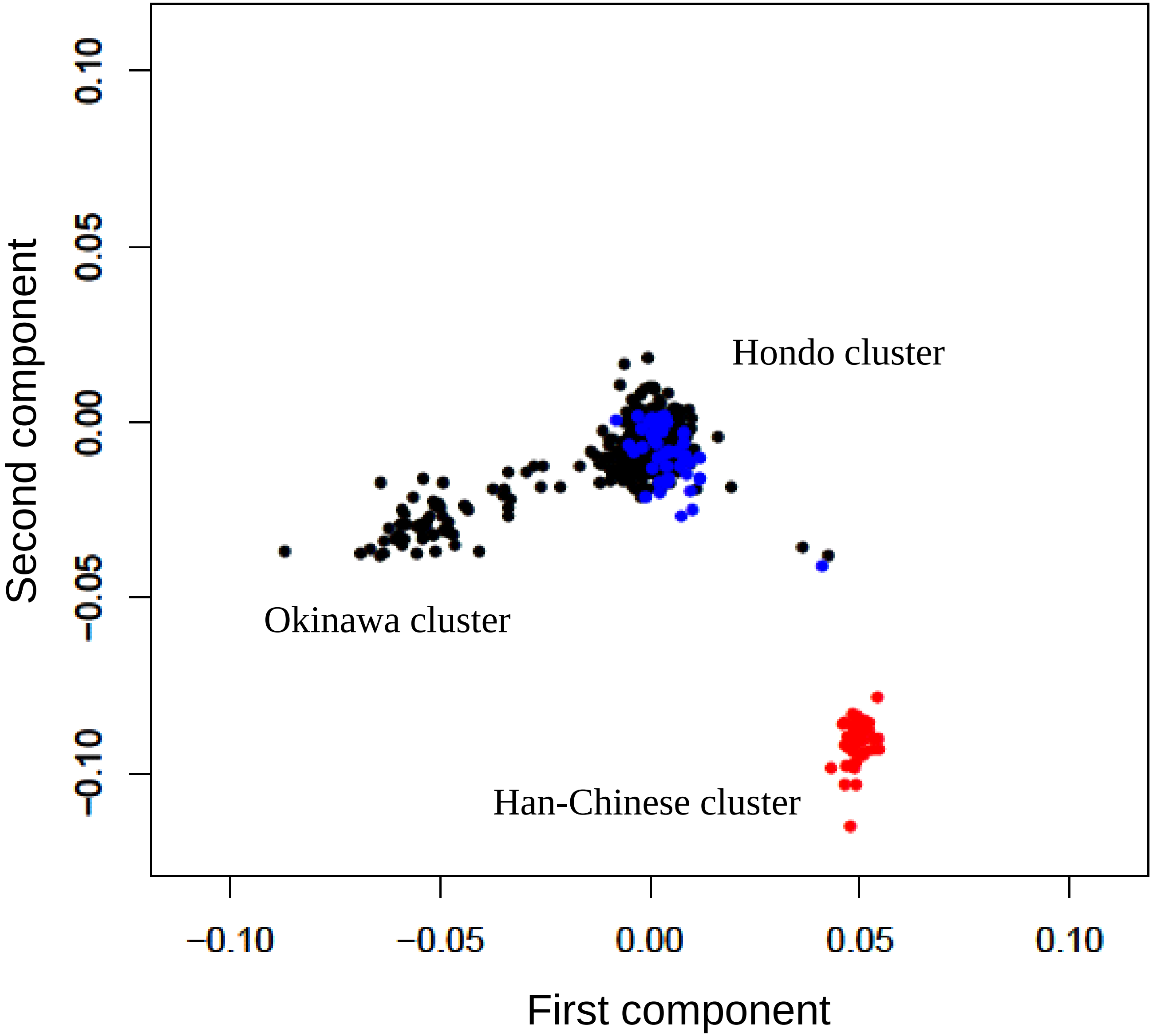

Hokkaido
Tohoku
Kinki
Kanto-Koshinetsu
Okai-Hokuriku
Kyushu
Okinawa

Samples from Okinawa



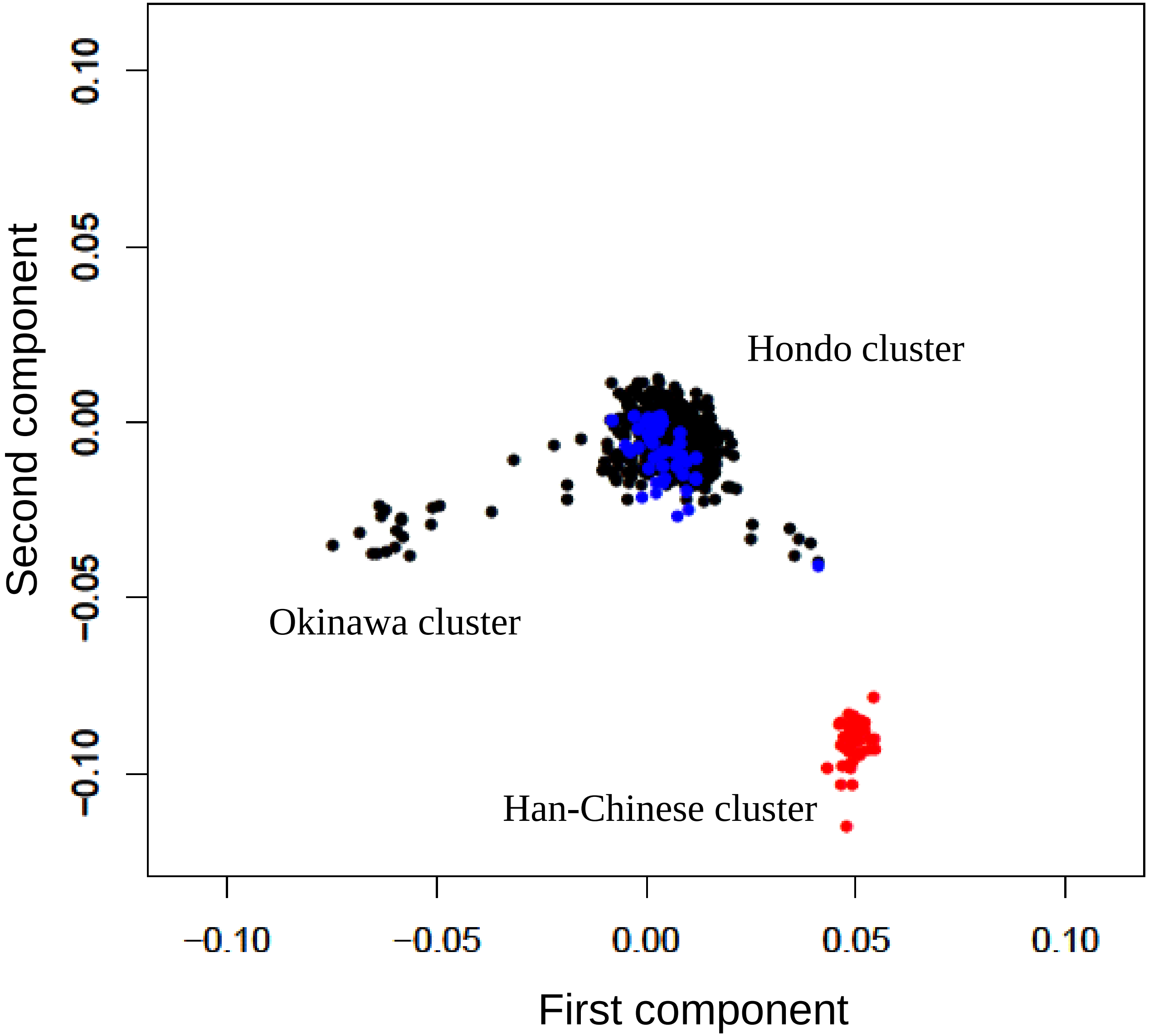


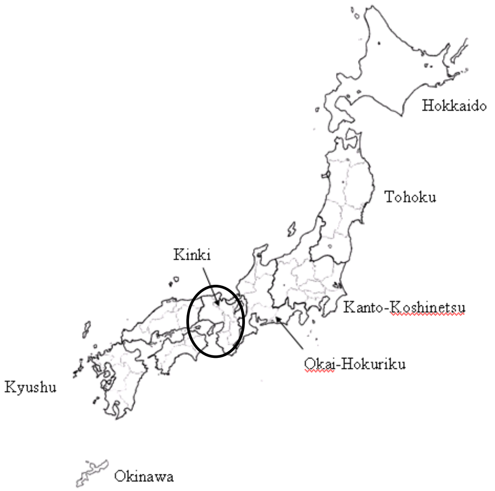
Samples from Kyushu



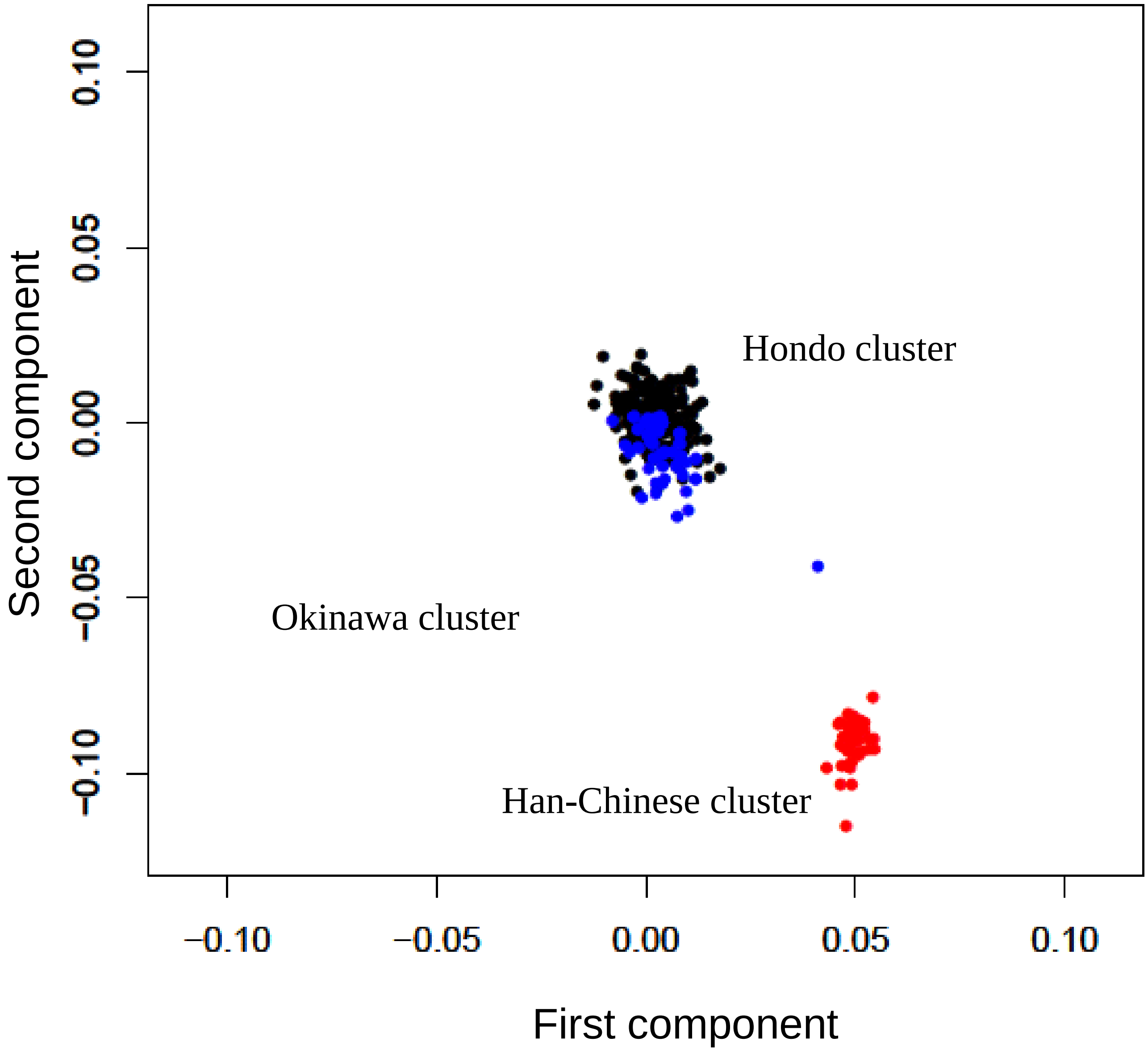
Hokkaido
Tohoku
Kinki
Kanto-Koshinetsu
Okai-Hokuriku
Kyushu
Okinawa

Samples from Kinki



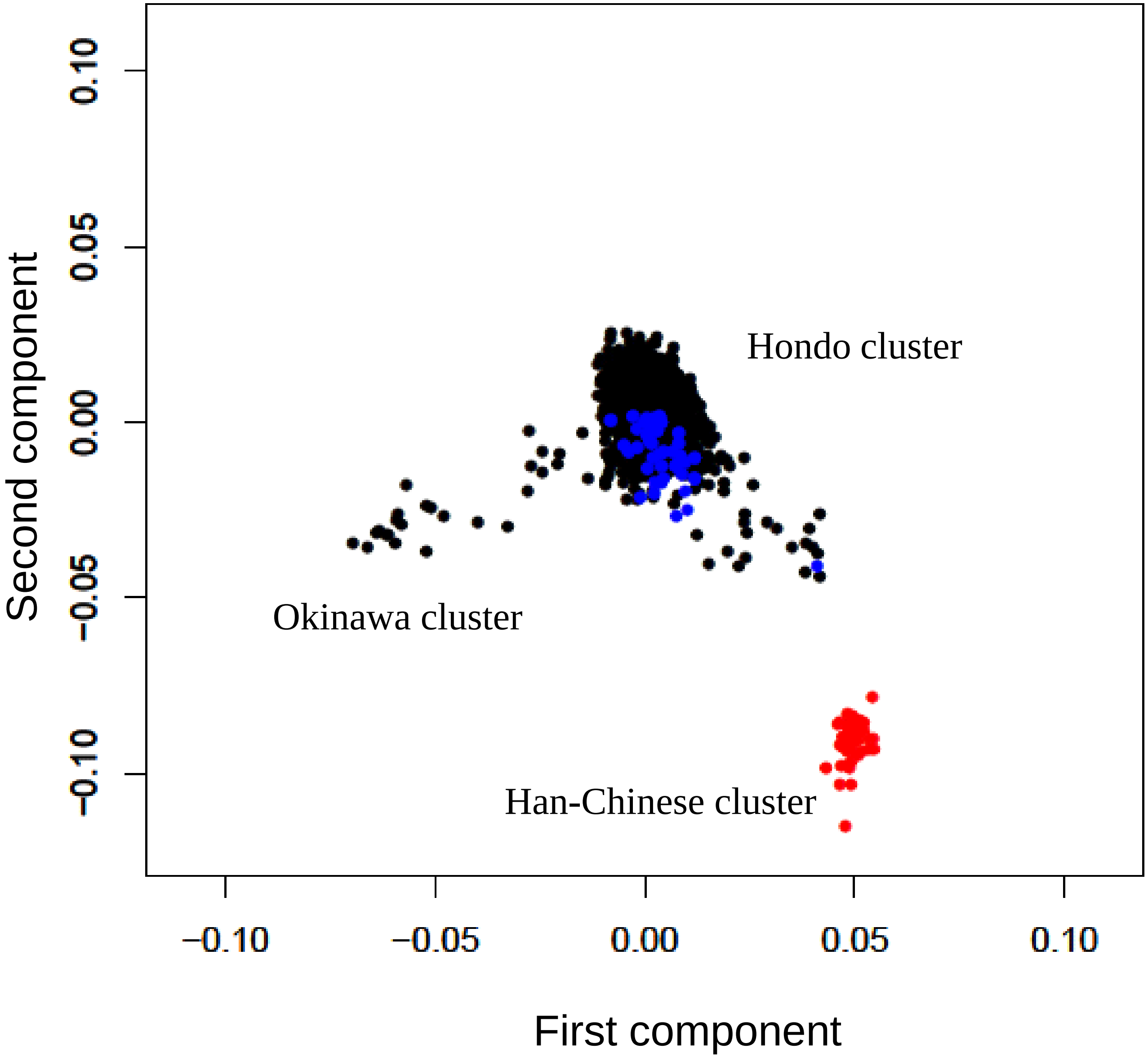


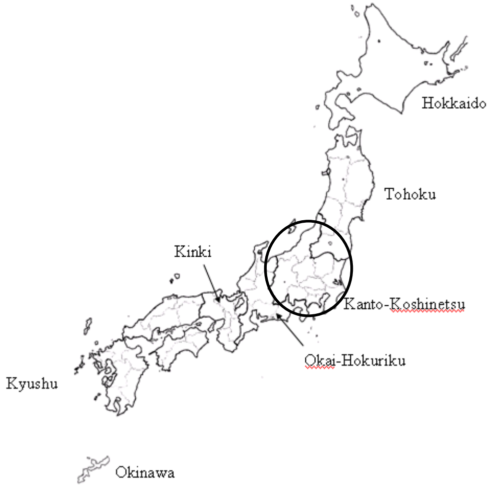
Samples from Tokai-Hokuriku



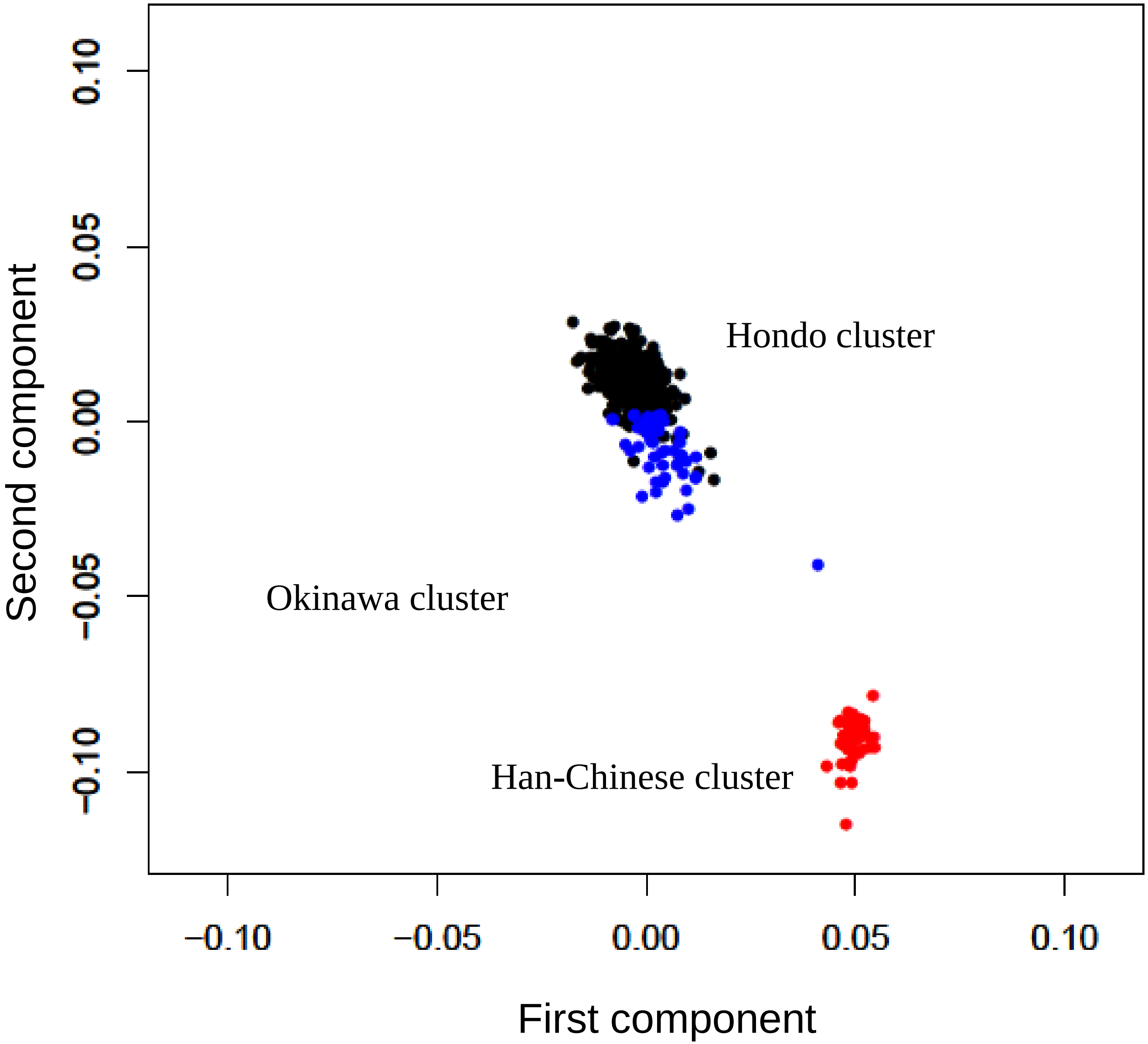
Hokkaido
Tohoku
Kinki
Kanto-Koshinetsu
Okai-Hokuriku
Kyushu
Okinawa

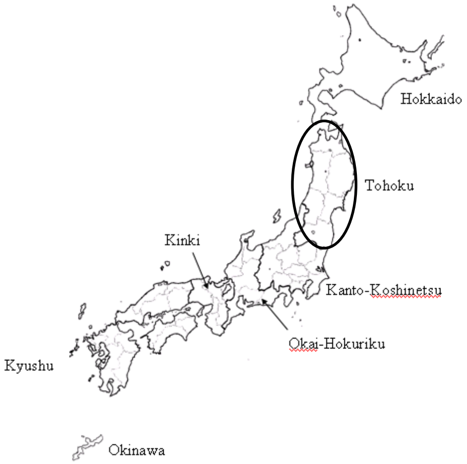
Samples from Kanto-Koshinetsu



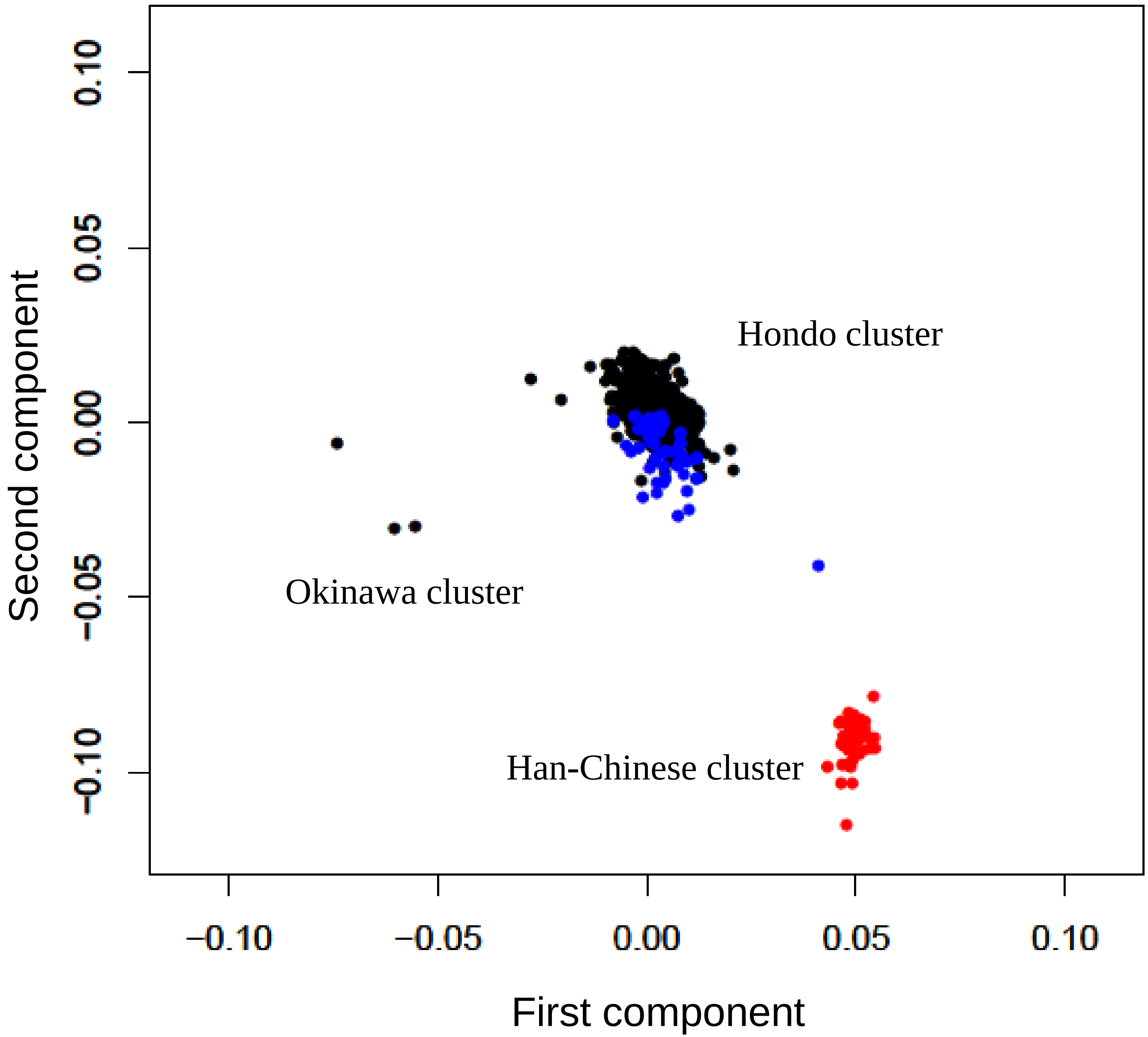


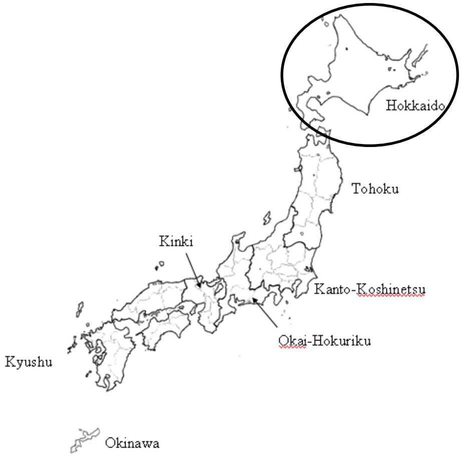
Samples from Tohoku



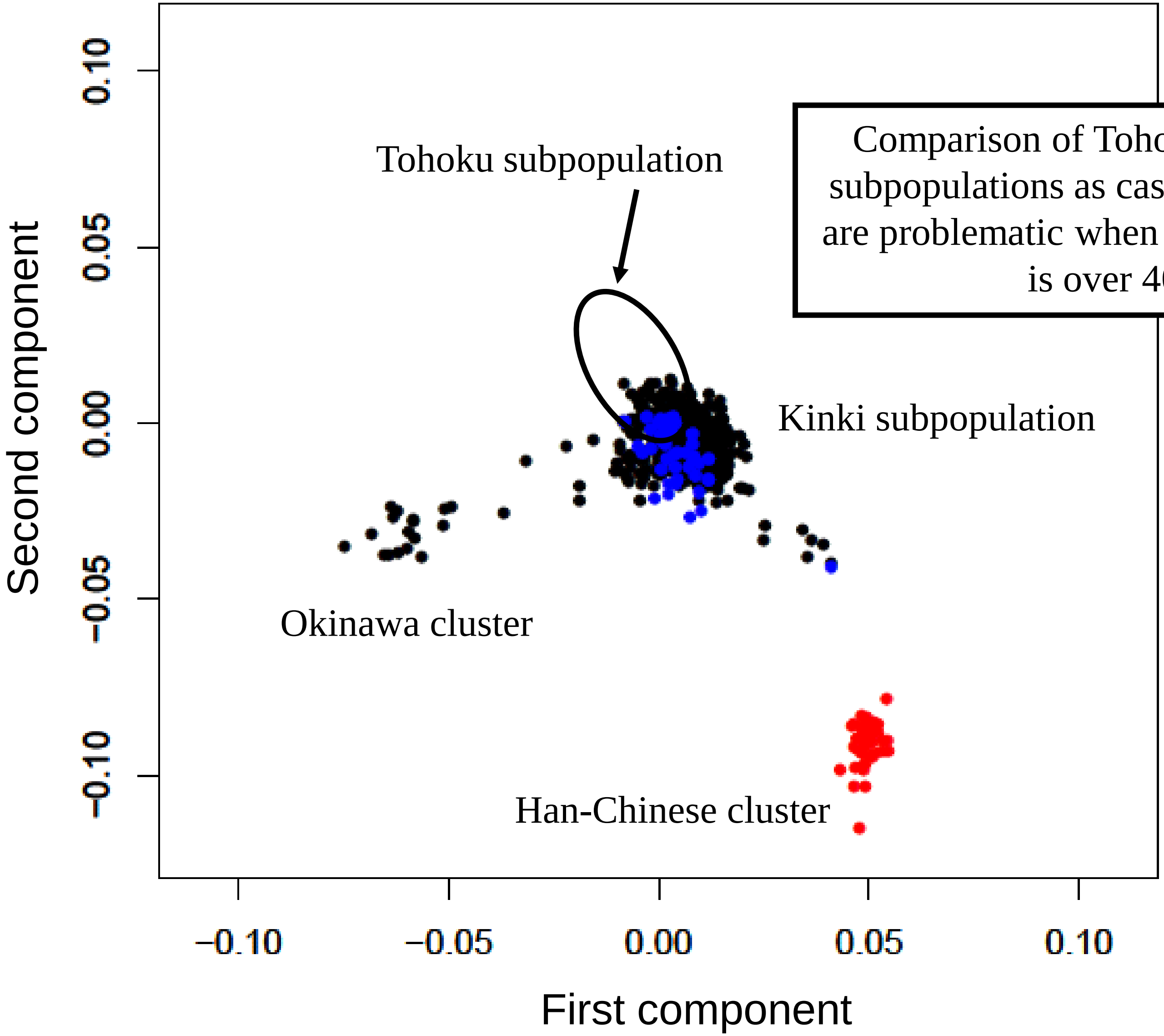


Samples from Hokkaido





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

Hokkaido
Tohoku
Kinki
Kanto-Koshinetsu
Okai-Hokuriku
Kyushu
Okinawa

Nonsynonymous SNPs ranked according to P values of Armitage test based on genotypes for Hondo and Okinawa clusters

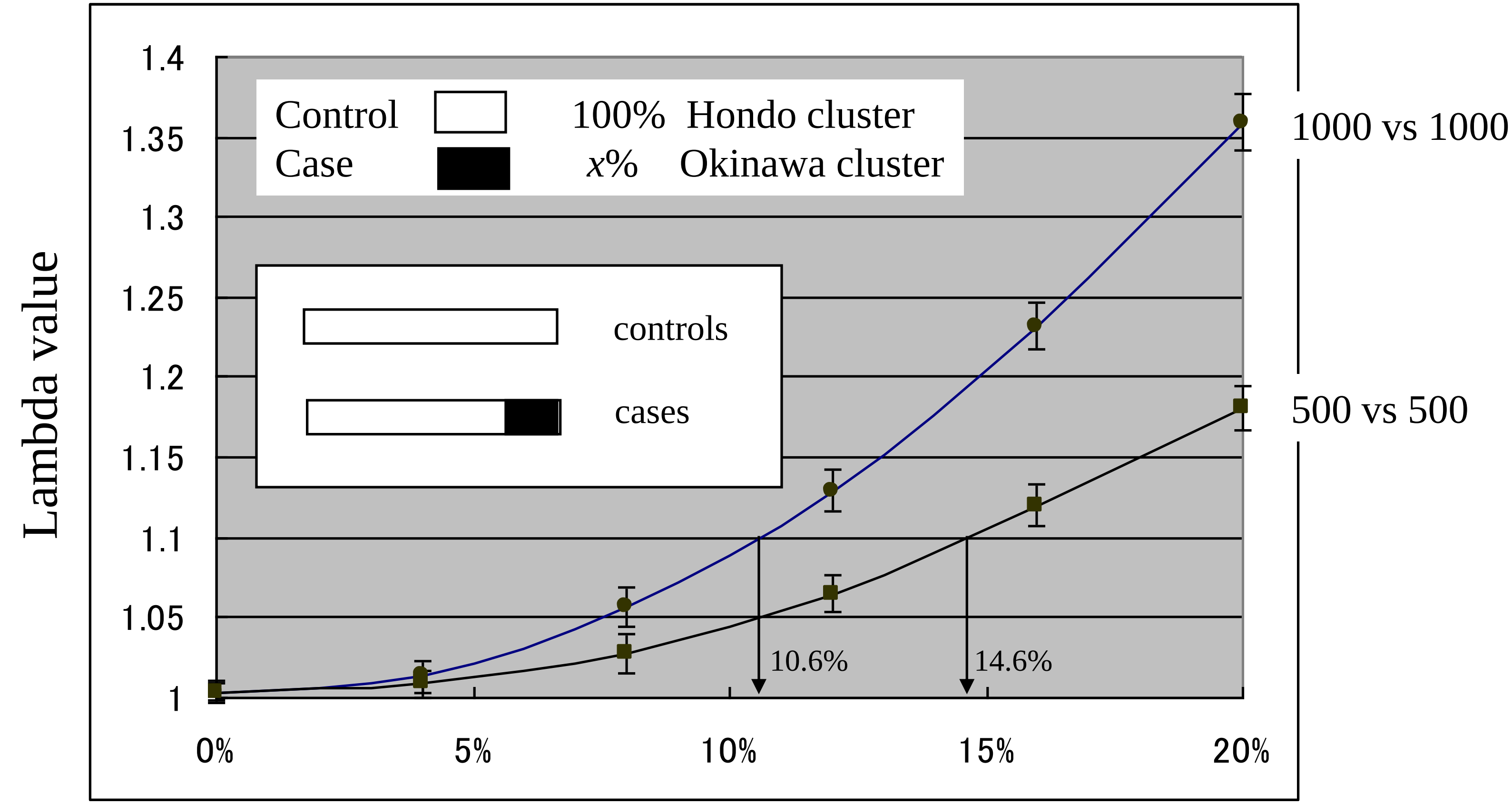
rs_number	chr	chr_pos	P value	gene
rs3827760	2	108880033	1.61E−20	EDAR
rs17822931	16	46815699	8.48E−20	ABCC11
rs4285045	4	144355168	1.23E−19	USP38
rs1799986	12	55821533	3.44E−17	LRP1
rs2274067	1	229443429	1.49E−15	C1orf131
rs2230611	19	5163482	3.49E−15	PTPRS
rs1872056	15	69827828	1.57E−14	FLJ13710
rs2298645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

Known to be associated with the thickness of the hair

Known to be associated with dry or wet ear wax

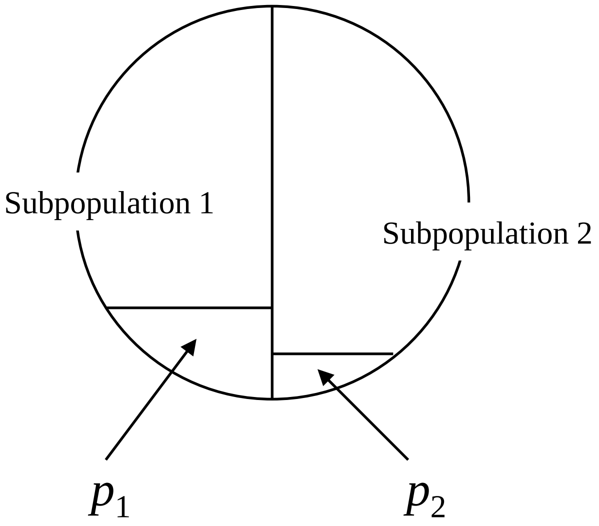
This method may be useful to identify genes that have been the targets of natural selection

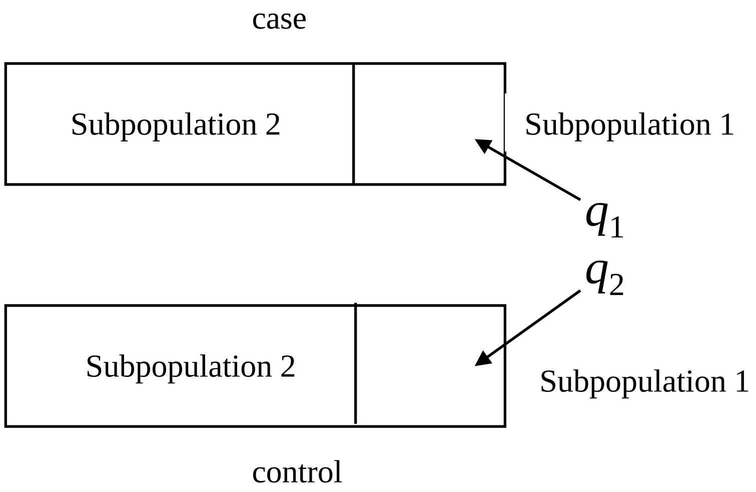
Inflation of type 1 error due to population structuring (expressed by lambda value for genomic control, mean and sd).



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



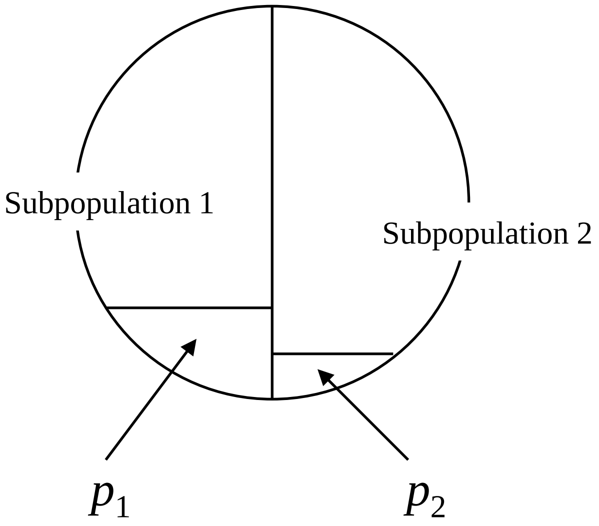


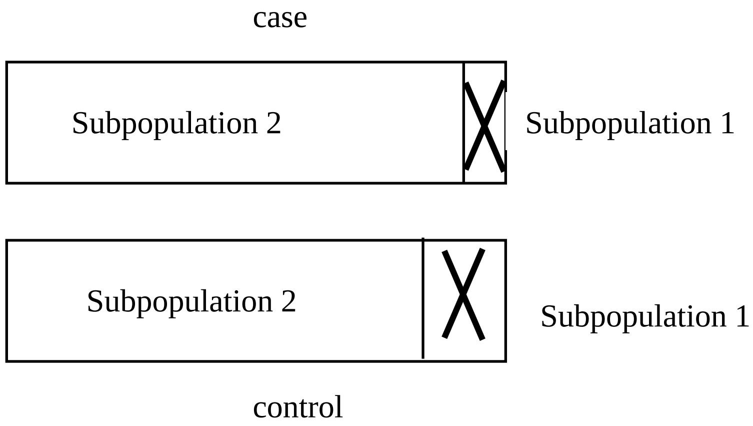
Adjust the proportion of subpopulation 1 so that $q_1 = q_2$ followed by a simple chi square test or by Mantel-Haenzel test

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

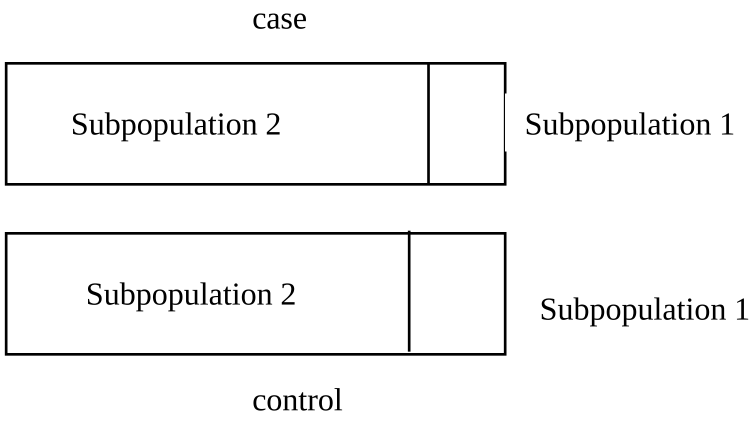
OR is 1 when $p_1 = p_2$, or $q_1 = q_2$

Method for avoiding the Inflation of type I error rate by mixing two different subpopulations



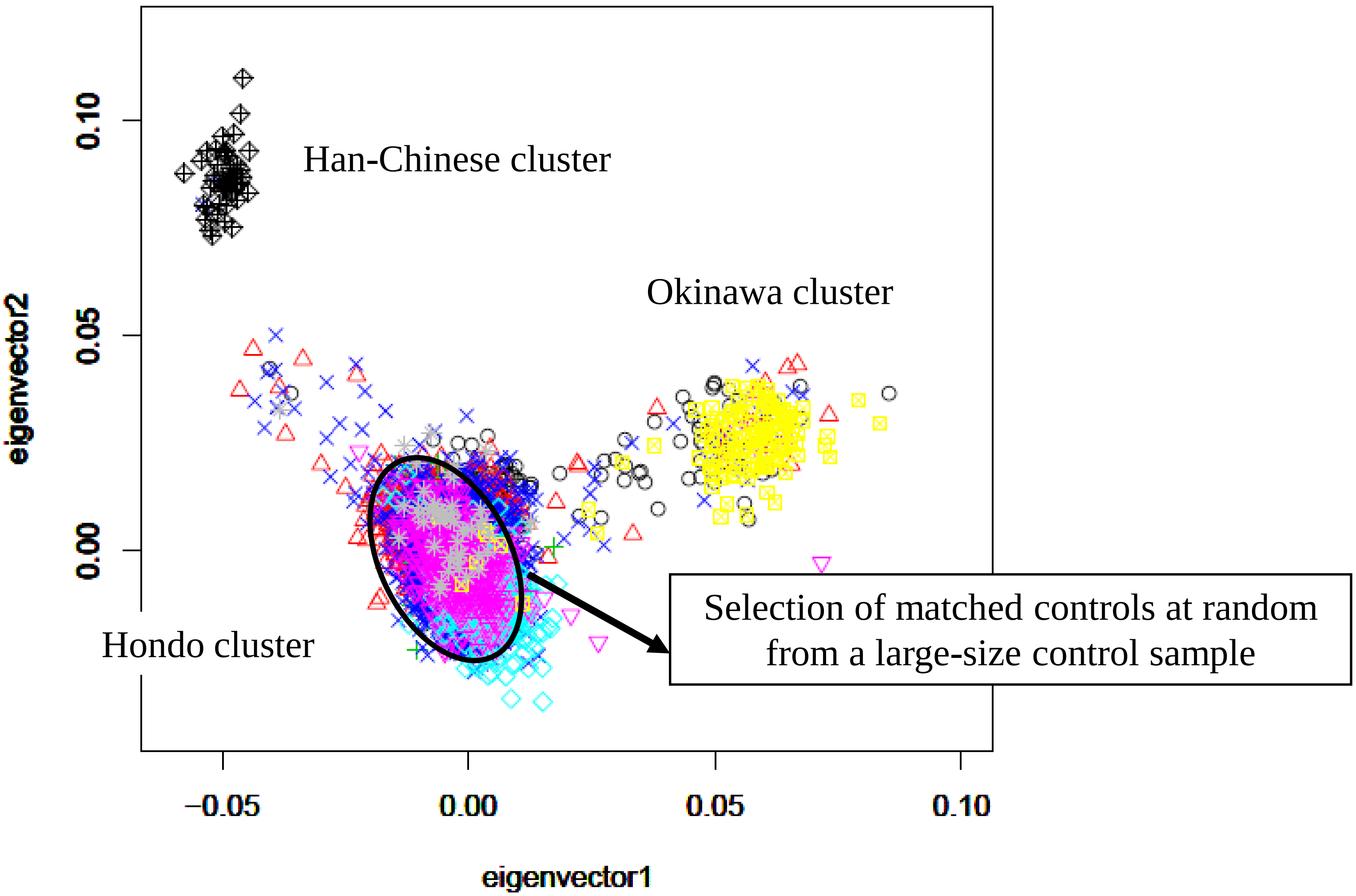


Simply exlcude subpopulation 1



Mantel-Haenzel test

Method for avoiding the Inflation of type I error rate by mixing two different subpopulations



Han-Chinese cluster

Okinawa cluster

Hondo cluster

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.