A short history of genome-wide association study (GWAS) and statistical challenges.

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

1. Medical Informatics Group, Center for Genomic Medicine, RIKEN
2. Institute of Data Analysis, StaGen Co. Ltd

Organization of CGM (Center for Genomic Medicine), RIKEN

Organization

- Center for Genomic Medicine
  Director : Yusuke Nakamura
- Research Group for Genotyping: Michiaki Kubo
  - Laboratory for Genotyping Development
    Team Leader:Michiaki Kubo
  - Laboratory for Biomarker Development
    Team Leader:Hideaki Nakagawa
- Research Group for Medical Informatics
  Group Director:Noyuki Kamatani

  - Laboratory for Statistical Analysis
    Team Leader:Noyuki Kamatani
  - Laboratory for Medical Informatics
    Team Leader:Tatsuhiko Tsunoda
- Research Group for Pharmacogenomics
  Group Director:Yusuke Nakamura
  - Laboratory for Pharmacogenetics
    Team Leader: Taisei Mushiroda
  - Laboratory for International Alliance
    Team Leader :Yusuke Nakamura
- Research Group for Disease-causing Mechanism
  Group Director:Toshihiro Tanaka
  - Laboratory for Cardiovascular Diseases
    Team Leader:Toshihiro Tanaka
  - Laboratory for Autoimmune Diseases
    Team Leader:Kazuhiko Yamamoto
  - Laboratory for Bone and Joint Diseases
    Team Leader: Shiro Ikegawa
  - Laboratory for Respiratory Diseases
    Team Leader:Mayumi Tamari
  - Laboratory for Endocrinology and Metabolism
    Team Leader:Shiro Maeda
  - Laboratory for Liver Diseases
    Team Leader :Kazuaki Chayama

What is GWAS (genome-wide association study)?

The Genomics Gold Rush

Eric J. Topol, MD
Sarah S. Murray, PhD
Kelly A. Frazer, PhD

IN RECENT WEEKS THERE HAS BEEN AN UNPRECEDENTED chain of discoveries in the genomics of complex traits.[1-14] The studies identified DNA markers associated with susceptibility to many of the most common diseases, ranging from acute lymphoblastic leukemia, the most important pediatric cancer, to obesity, type 2 diabetes mellitus, and coronary heart disease, which collectively affect nearly a billion individuals worldwide. The breakneck pace of discovery will be continuing in the months ahead, with anticipated findings for many cancers, cardiovascular diseases, and neurological diseases. In aggregate, these studies have the potential to radically change medicine. This Commentary is intended to provide perspective for the medical community, to understand the limitations of the work that has thus far been completed, and to outline the challenges that lie ahead.

The Era of Genome-Wide Association Studies

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 Runners-Up ›
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 variation, featuring Francis Collins, David Altshuler, and *Science* news writer Liz Pennisi.

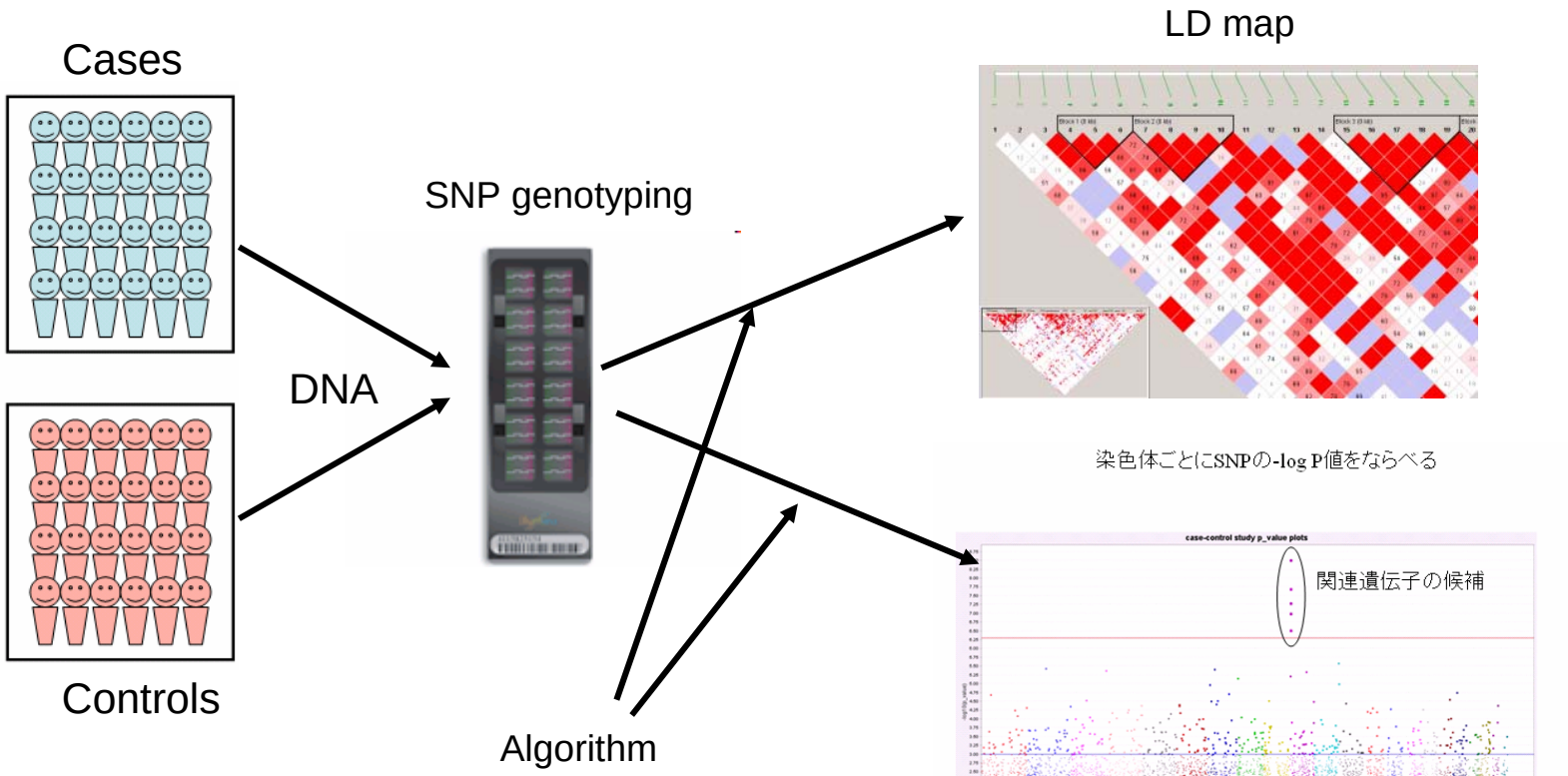
› Higher-bandwidth version of video

› Lower-bandwidth version of video

Human genetic variation

GWAS; genome-wide association study

Majority of the genetic causes of diseases will be elucidated.



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

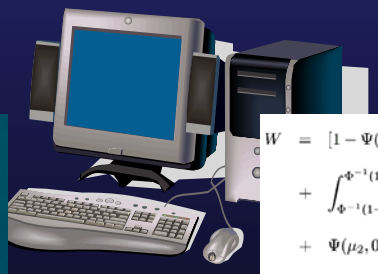
Three elements contributing to GWAS

GWA: Genome-wide association study

1. List of SNPs covering the whole genome（HapMap）

2. Technologies for SNP genotypeing
（Invader, Illumina, Affymetrix）

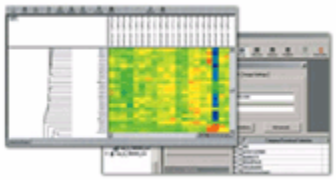
3. Technologies for data analysis and statistical genetics



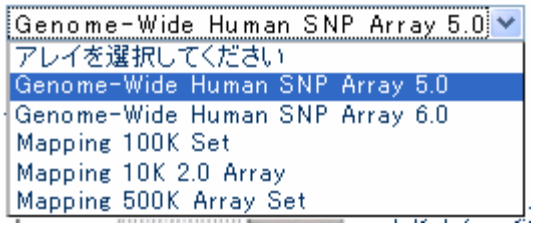
$$
\begin{aligned}
W &= [1-\Psi(\mu_2,0)]\int_{\Phi^{-1}(1-\gamma/2)}^{\infty}\psi(\mu_1,\,z_1)dz_1 \\
&+ \int_{\Phi^{-1}(1-\alpha_1/2)}^{\Phi^{-1}(1-\gamma/2)}\psi(\mu_1,\,z_1)[1-\Psi(\mu_2,\,\Phi^{-1}\{1-\frac{\gamma}{4[1-\Phi(z_1)]}\})]dz_1 \\
&+ \Psi(\mu_2,0)\int_{\Phi^{-1}(1-\gamma/2)}^{\infty}\psi(\mu_1,\,z_1)dz_1+\int_{\Phi^{-1}(1-\alpha_1/2)}^{\Phi^{-1}(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

Commercial SNP genotyping systems

Affymetrix



GeneChip Mapping Array



Illumina

Illumina BeadChip

Genotype data for 300,000 – 1,000,000 SNP genotypes per person

Analysis of genomic and phenotypic data to identify causes of diseases and drug reactions

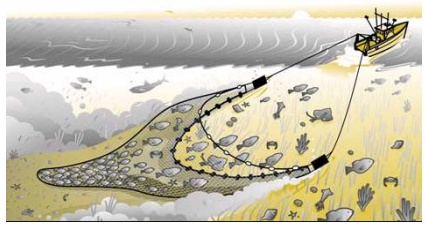
1. Genetic factors

   Candidate gene approach (biochemistry, molecular biology)

   Genome-wide approach（genetics, statistics, informatics）

Rod and line fishing

Trawl fishing



2. Environmental factors

3. Gene-environment interactions

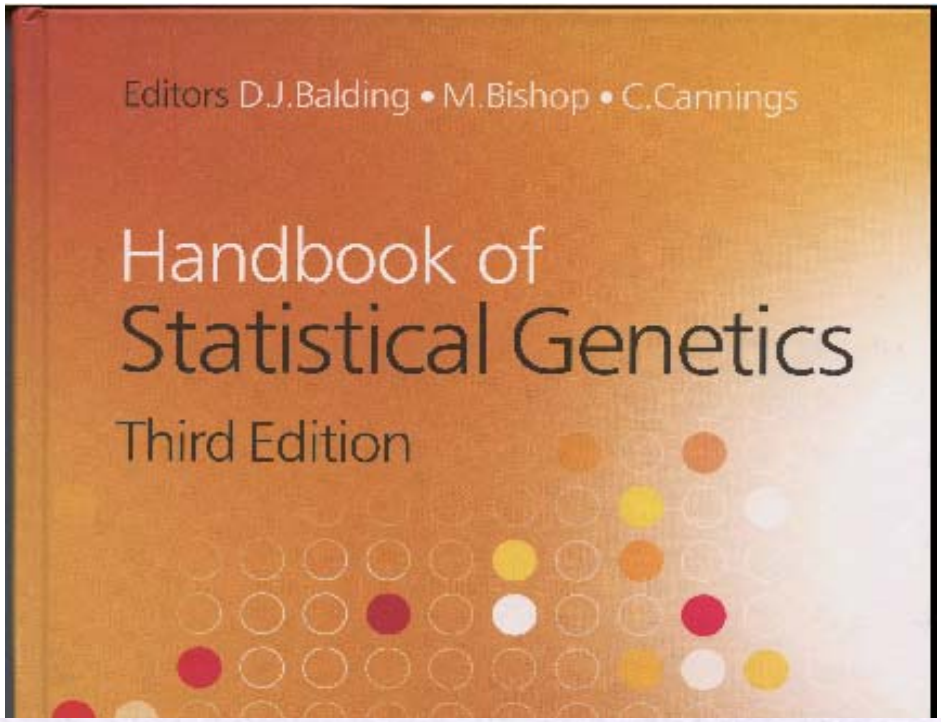
Genotypic value　G-E interaction

$$\Phi = G + E + (G, E)$$

Phenotypic value　Environmental value

$$\log (P/(1\text{-}P)) = G + E + (G, E)$$

Who did GWAS for the first time in the world?



37.8 PROSPECTS FOR WHOLE-GENOME ASSOCIATION STUDIES

Initial reports of WGA studies began as early as 2002 (Ozaki *et al.*, 2002)

Initial reports of GWAS from CGM, RIKEN

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

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

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;35:341-8.

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

Subsequent reports from CGM, RIKEN

Myocardial infarction
PSMA6: Nature Genetics, 2006
MIAT (functional RNA): J Hum Genet, 2007
BRAP: Nature Genetics, 2009

Cerebral thrombosis
PRKCH: Nature Genetics, 2007
AGTRL1: Hum Mol Genet, 2007

Rheumatoid arthritis
FcRH3: Nature Genetics, 2005
CD244: Nature Genetics, 2008

Osteoarthritis
Asporin：Nature Genetics, 2005
Calmodulin 1： Hum Mol. Genet, 2005
GDF5: Nature Genetics, 2007
COL11A1 : Am J Hum Genet, 2007
SLC35D1: Nature Medicine, 2007
DVWA: Nature Genetics, 2008

Lumber disc herniation
CILP: Nature Genetics, 2005
COL9A2: J Hum Genet, 2007
THBS2、MMP9: Am J Hum Genet, 2008

Osteoarthritic disorders
SLC35D1 : Nature Medicine , 2007

Statistical genetics
Clustering: Am J Hum Genet. 2008
Genotyping: Bioinformatics. 2007

Bronchial asthma
Tenascin-C Fn-III-D: Hum Mol Genet, 2005
IL13: Int Arch Allergy Immunol, 2006

Diabetes/Diabetic nephropathy
ELMO1： Diabetes, 2005
KLF9: Mol Endocrinol, 2006
Renin-Angiotensin: J Hum Genet,2007
Neurocalcin: Human Genet, 2007
KCNQ1: Nature Genetics, 2008

Obesity
SCG3 : J Clin Endocrinol, 2007

Crohn's disease
TNFSF15: Hum Mol Genet, 2005

Kawasaki disease
ITPKC: Nature Genetics, 2008

Effect of tamoxifen
CYP2D6: Cancer Science, 2008

HapMap project
HapMap 1M: Nature, 2005
HapMap 3M： Nature, 2007
HapMap ： Nature, 2007

Method for CNV detection
RETINA法: Hum Mutation, 2007

Pharmacogenomics
Warfarin: N Engl J Med, 2009

QC (quality control) of giga genotype data

QC filter

1. Call rate check (use of SNPs and subjects with high call rates)
2. Remove individuals who are close relatives (IBS and IBD check)
3. Analyze population structure, remove outliers and estimate the inflation factor λ.
4. Test of the accordance with HWE（Hardy-Weinberg equilibrium）to remove mistyped SNPs, CNV and in/dels.
5. Visual inspection of 2-dimensional graph for clustering.

Removal of highly related individuals

	offspring	sister brother	niece nephew	cousin	grandson
kinship coefficient	1/4	1/4	1/8	1/16	1/8

Kinship coefficient is the probability that two alleles from a pair are identical by descent

XV.—**The Correlation between Relatives on the Supposition of Mendelian Inheritance.** By **R. A. Fisher**, B.A. *Communicated by* Professor J. Arthur Thomson. (With Four Figures in Text.)

Generations.	Half 2nd Cousin.	Half 1st Cousin.	Half Brother.	Ancestral Line.	Brother.	1st Cousin.	2nd Cousin.
Own	$1/64$	$1/16$	$1/4$	1	$1/2$	$1/8$	$1/32$
Father's	$1/128$	$1/32$	$1/8$	$1/2$	$1/4$	$1/16$	$1/64$
Grandfather's	$1/256$	$1/64$	$1/16$	$1/4$	$1/8$	$1/32$	$1/128$
Great-grandfather's	$1/512$	$1/128$	$1/32$	$1/8$	$1/16$	$1/64$	$1/256$
Great-great-grandfather's	$1/1024$	$1/256$	$1/64$	$1/16$	$1/32$	$1/128$	$1/512$

Note that, in genetics, probabilities are exact!

HWE（Hardy-Weinberg equilibrium）test

Hardy-Weinberg's law

Wilhelm Weinberg

JULY 10, 1908 SCIENCE N. S. Vol. XXVIII: 49–50

DISCUSSION AND CORRESPONDENCE

Mendelian Proportions in a Mixed Population

To The Editor of Science: I am reluctant to intrude in a discussion concerning matters of which I have no expert knowledge, and I should have expected the very simple point which I wish to make to have been familiar to biologists. However, some remarks of Mr. Udny Yule, to which Mr. R. C. Punnett has called my attention, suggest that it may still be worth making.

In the *Proceedings of the Royal Society of Medicine* (Vol I., p. 165) Mr. Yule is reported to have suggested, as a criticism of the Mendelian position, that if brachydactyly is dominant "in the course of time one would expect, in the absence of counteracting factors, to get three brachydactylous persons to one normal."

It is not difficult to prove, however, that such an expectation would be quite groundless. Suppose that *Aa* is a pair of Mendelian characters, *A* being dominant, and that in any given generation the numbers of pure dominants (*AA*), heterozygotes (*Aa*), and pure recessives (*aa*) are as $p:2q:r$. Finally, suppose that the numbers are fairly large, so that the mating may be regarded as random, that the sexes are evenly distributed among the three varieties, and that all are equally fertile. A little mathematics of the multiplication-table type is enough to show that in the next generation the numbers will be as

$$(p+q)^2 : 2(p+q)(q+r) : (q+r)^2,$$

or as $p_1:2q_1:r_1$, say.

The interesting question is – in what circumstances will this distribution be the same as that in the generation before? It is easy to see that the condition for this is $q^2 = pr$. And since $q_1^2 = p_1r_1$, whatever the values of p, q, and r may be, the distribution will in any case continue unchanged after the second generation.

Suppose, to take a definite instance, that *A* is brachydactyly, and that we start from a population of pure brachydactylous and pure normal persons, say in the ratio of 1:10,000. Then $p = 1$, $q = 0$, $r = 10{,}000$ and $p_1 = 1$, $q_1 = 10{,}000$, $r_1 = 100{,}000{,}000$. If brachydactyly is dominant, the proportion of brachydactylous persons in the second generation is 20,001:100,020,001, or practically 2:10,000, twice that in the first generation; and this proportion will afterwards have no tendency whatever to increase. If, on the other hand, brachydactyly were recessive, the proportion in the second generation would be 1:100,020,001, or practically 1:100,000,000, and this proportion would afterwards have no tendency to decrease.

In a word, there is not the slightest foundation for the idea that a dominant character should show a tendency to spread over a whole population, or that a recessive should tend to die out.

I ought perhaps to add a few words on the effect of the small deviations from the theoretical proportions which will, of course, occur in every generation. Such a distribution as $p_1:2q_1:r_1$, which satisfies the condition $q_1^2 = p_1r_1$, we may call a *stable* distribution. In actual fact we shall obtain in the second generation not $p_1:2q_1:r_1$ but a slightly different distribution $p:2q:r$, which is not "stable." This should, according to theory, give us in the third generation a "stable" distribution $p_2:2q_2:r_2$, also differing from $p_1:2q_1:r_1$; and so on. The sense in which the distribution $p_1:2q_1:r_1$ is "stable" is this, that if we allow for the effects of casual deviations in any subsequent generation, we should, according to theory, obtain at the next generation a new "stable" distribution differing but slightly from the original distribution.

I have, of course, considered only the very simplest hypotheses possible. Hypotheses other that [*sic*] that of purely random mating will give different results, and, of course, if, as appears to be the case sometimes, the character is not independent of that of sex, or has an influence on fertility, the whole question may be greatly complicated. But such complications seem to be irrelevant to the simple issue raised by Mr. Yule's remarks.

G. H. Hardy
Trinity College, Cambridge,
April 5, 1908

Godfrey Harold Hardy

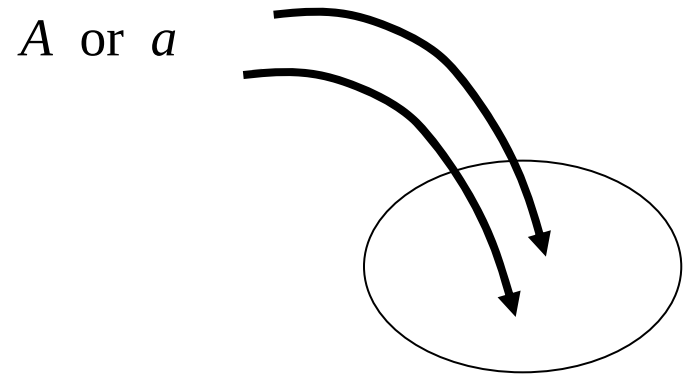
"I have never done anything 'useful'. No discovery of mine has made, or is likely to make, directly or indirectly, for good or ill, the least difference to the amenity of the world."

Hardy-Weinberg's law（HWE）

p : frequency of allele a

Genotype frequencies AA: p^2、Aa: $2p(1-p)$、aa: $(1-p)^2$

Independent sampling of two alleles for each subject



Sample space

A,*A*　*A*,*a*

a,*A*　*a*,*a*

Test of HWE

Let n_1, n_2, and n_3 denote the numbers of *AA*, *Aa*, and *aa* genotypes in a sample.

1. Pearson' chi square method

$$\hat{p} = \frac{2n_1 + n_2}{2n} \qquad C_1 = \frac{(n_1 - n\hat{p}^2)^2}{n\hat{p}^2} + \frac{[n_2 - 2n\hat{p}(1-\hat{p})]^2}{2n\hat{p}(1-\hat{p})} + \frac{[n_3 - n(1-\hat{p})^2]^2}{n(1-\hat{p})^2}$$

2. Likelihood ratio method

$$l_0 = n_1 \log p^2 + n_2 \log[2p(1-p)] + n_3 \log[(1-p)^2]$$

$$l_{max} = n_1 \log \frac{n_1}{n} + n_2 \log \frac{n_2}{n} + n_3 \log \frac{n_3}{n}$$

$$C_2 = -2 \log \frac{L_{0max}}{L_{max}} = -2(l_{0max} - l_{max})$$

3. Exact test

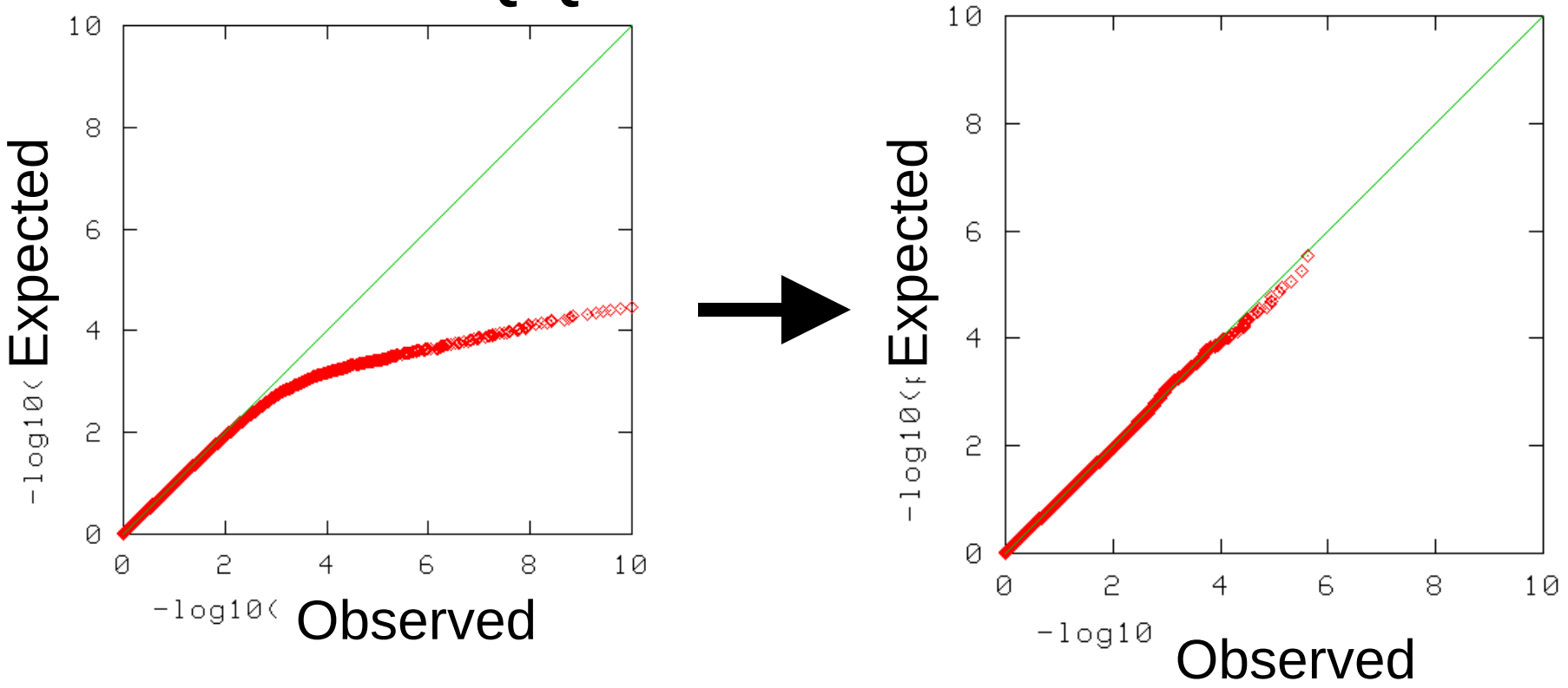
$$\sum_{x_1;[x_1 \geq 0,\ n_1 - n_3 \leq x_1 \leq (2n_1+n_2)/2,\ C_1(x_1) \geq C_1(n_1)]} P(X_1 = x_1) = \frac{2^{2n_1+n_2-2x_1}(2n_1+n_2)!(2n_3+n_2)!n!}{(2n)!x_1!(n_3-n_1+x_1)!(2n_1+n_2-2x_1)!}.$$

Before and after HWE check

Procedure:

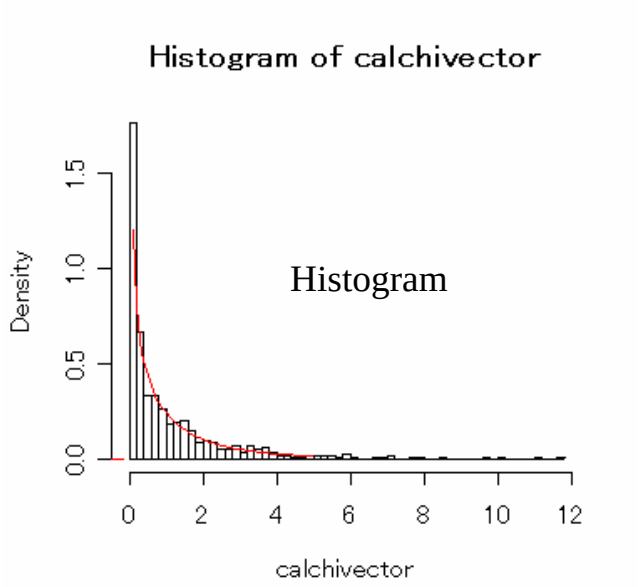
- **SNPs with pHWE < 1.0E-6 were removed**
- **SNPs with pHWE < 1.0E-4 and "undetermined samples >1" were removed**

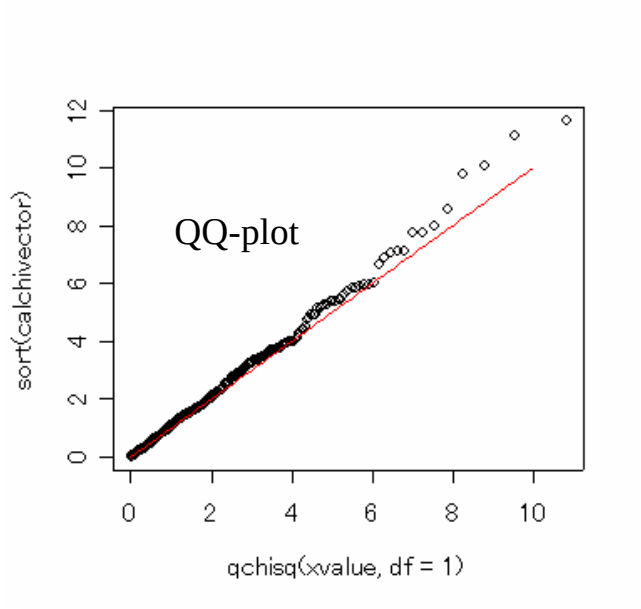
Q-Q Plot for HWE test

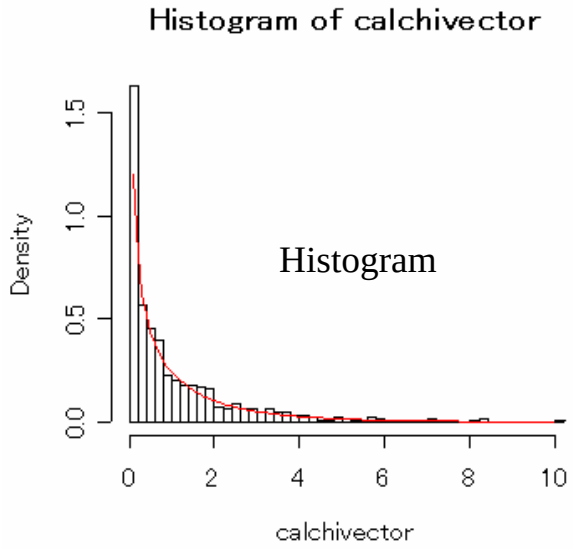


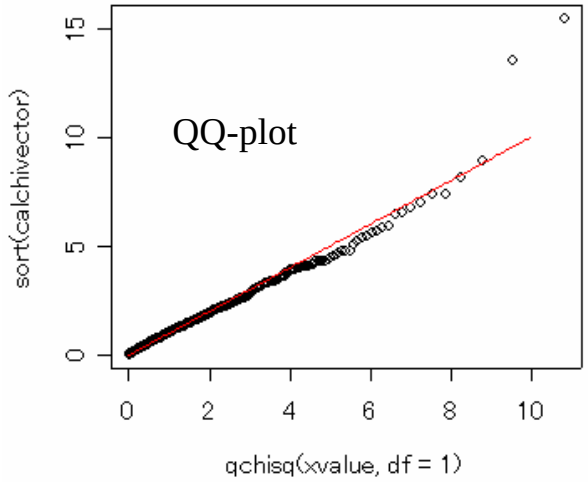
408 SNPs were removed

Histogram and QQ-plot for HWE test







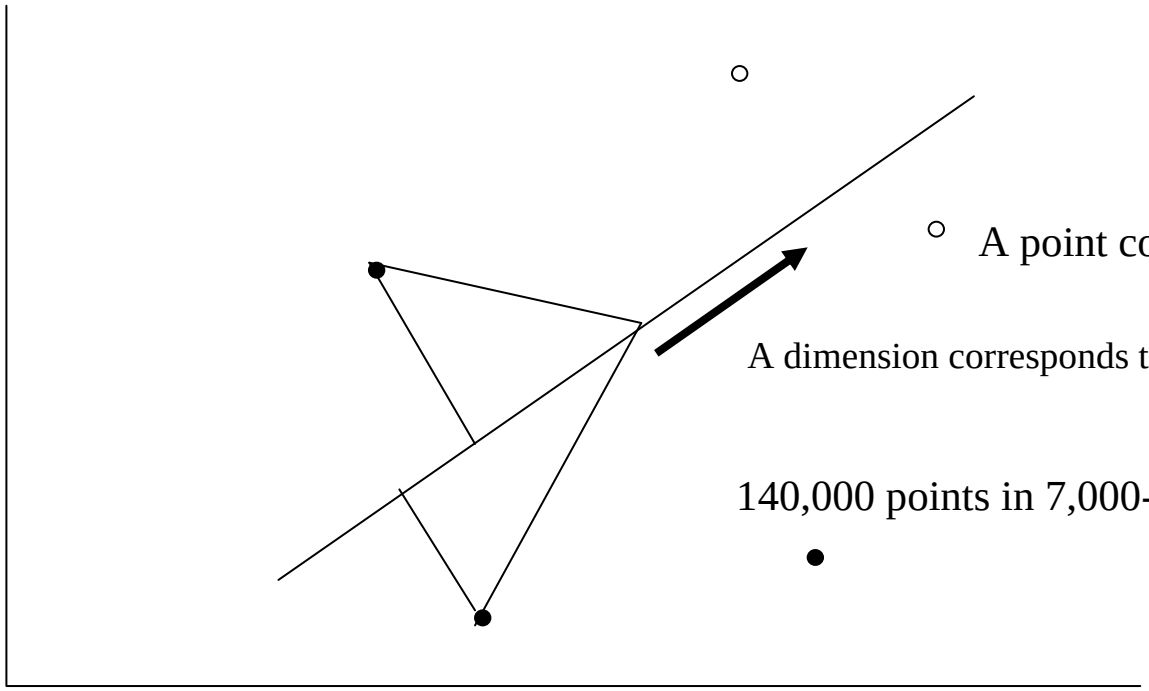


Population structuring problem

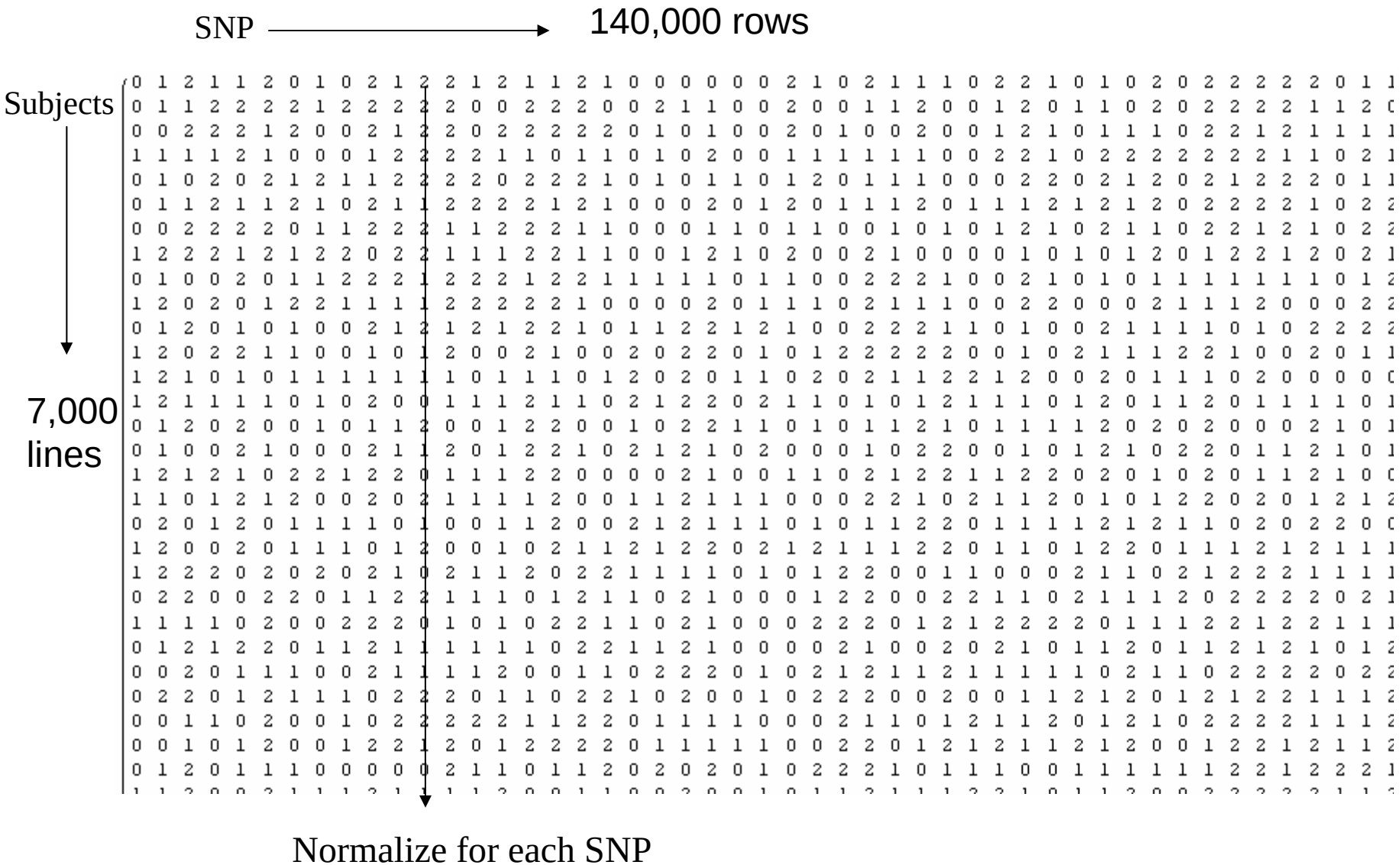
Population structure analysis (Yamaguchi-Kabata)

Principle component analysis（EIGENSTRAT）

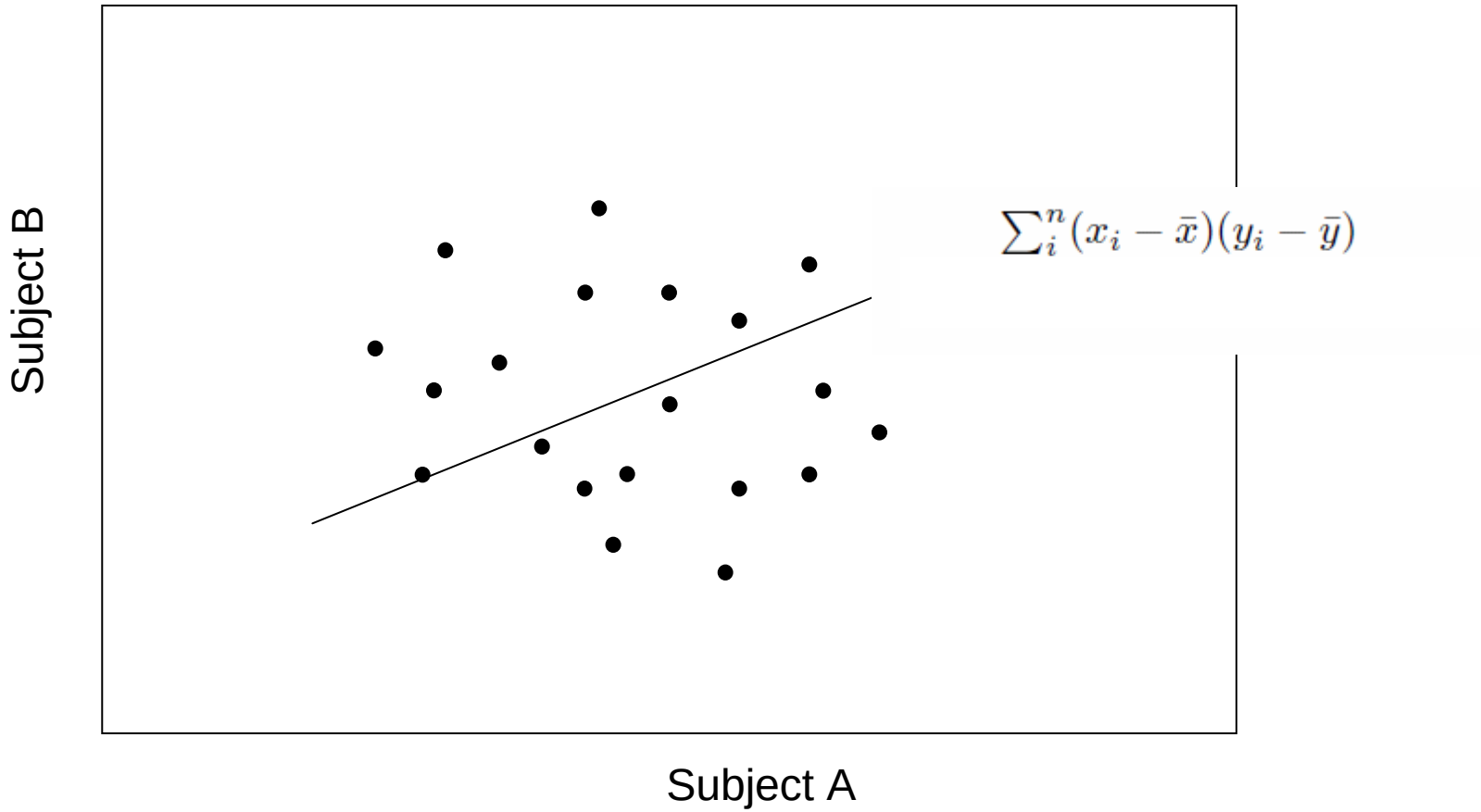
Draw a line in *n*-dimensional space that gives the largest variance.



Genotype data for 140,000 SNPs from 7,000 Japanese subjects



Calculation of covariance

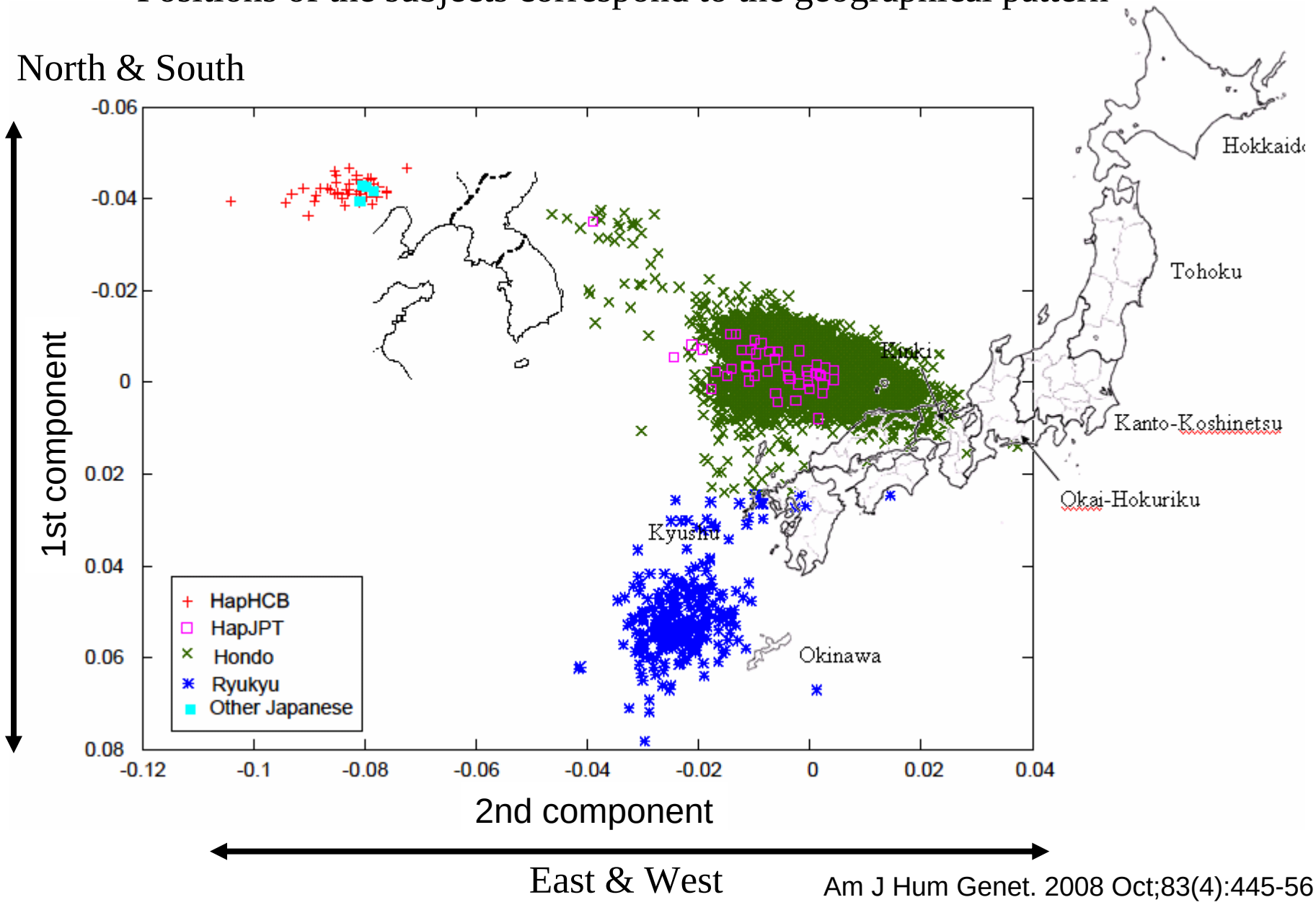


A graph contains 140,000 points from which covariance is determined.

There are 24,500,000 such graphs.

Eigenvectors are determined for the covariance matrix（7,000 x 7,000）

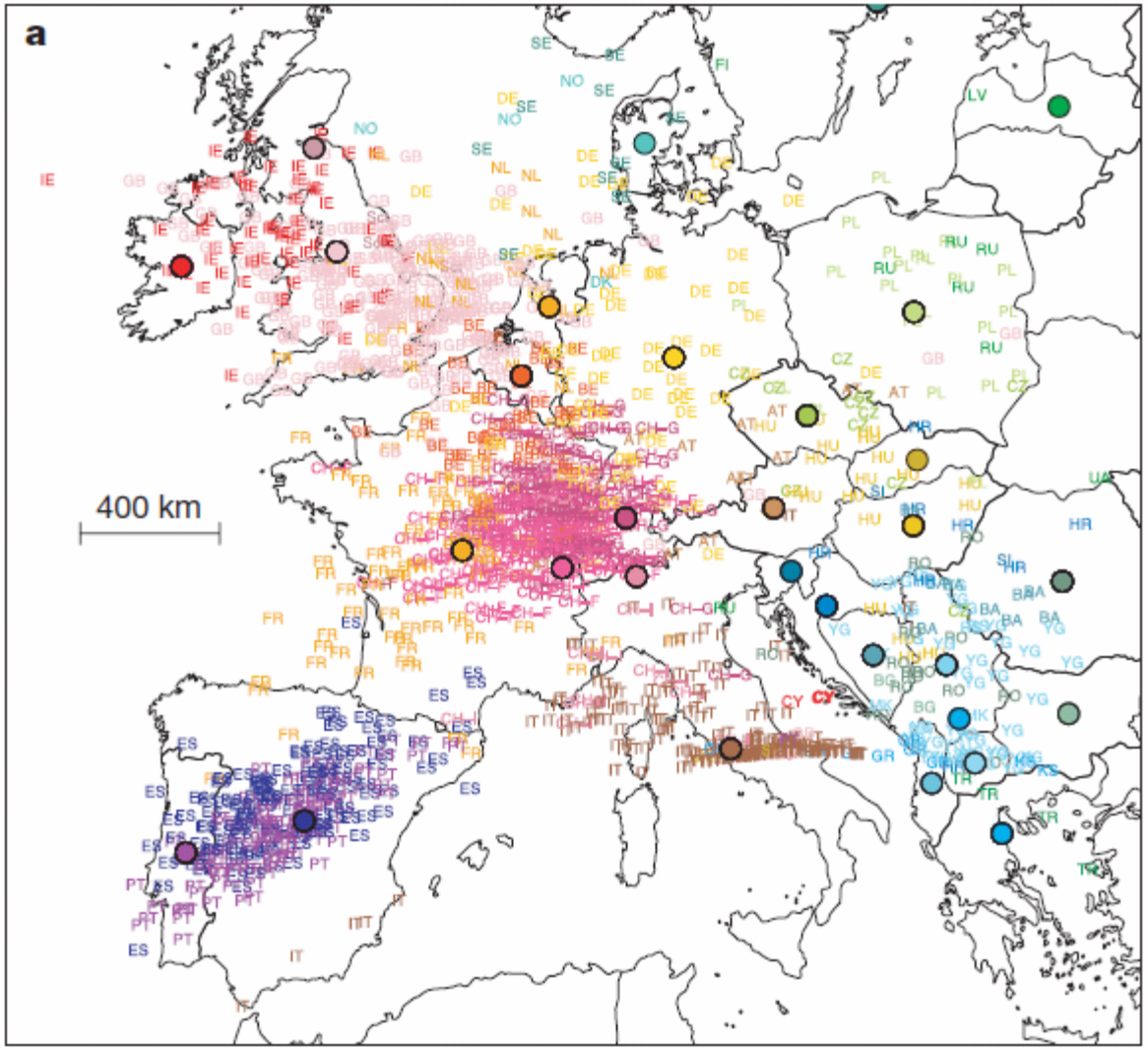
Positions of the subjects correspond to the geographical pattern



Am J Hum Genet. 2008 Oct;83(4):445-56

Population structure in Europe

North & South



1st component

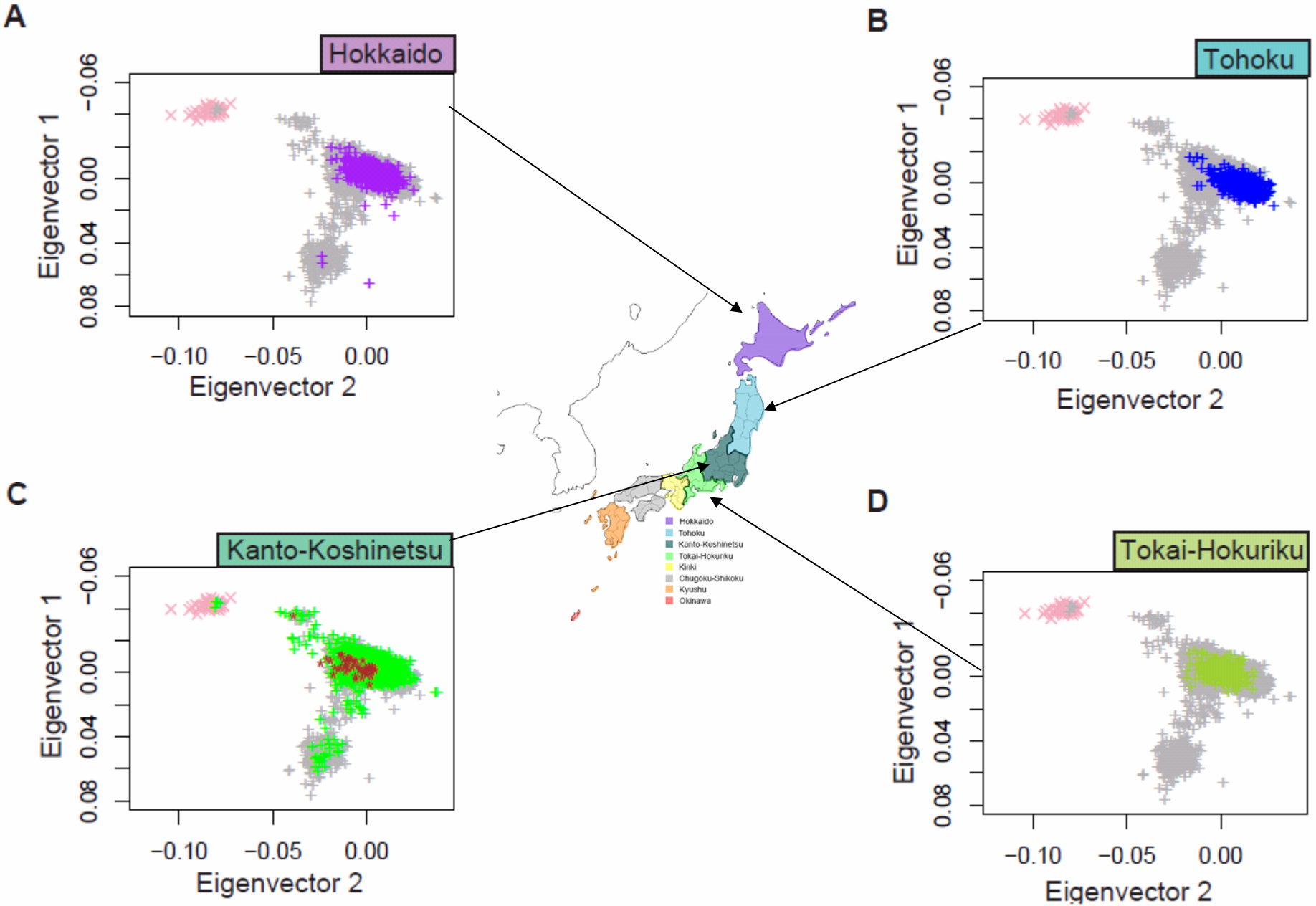
2nd component

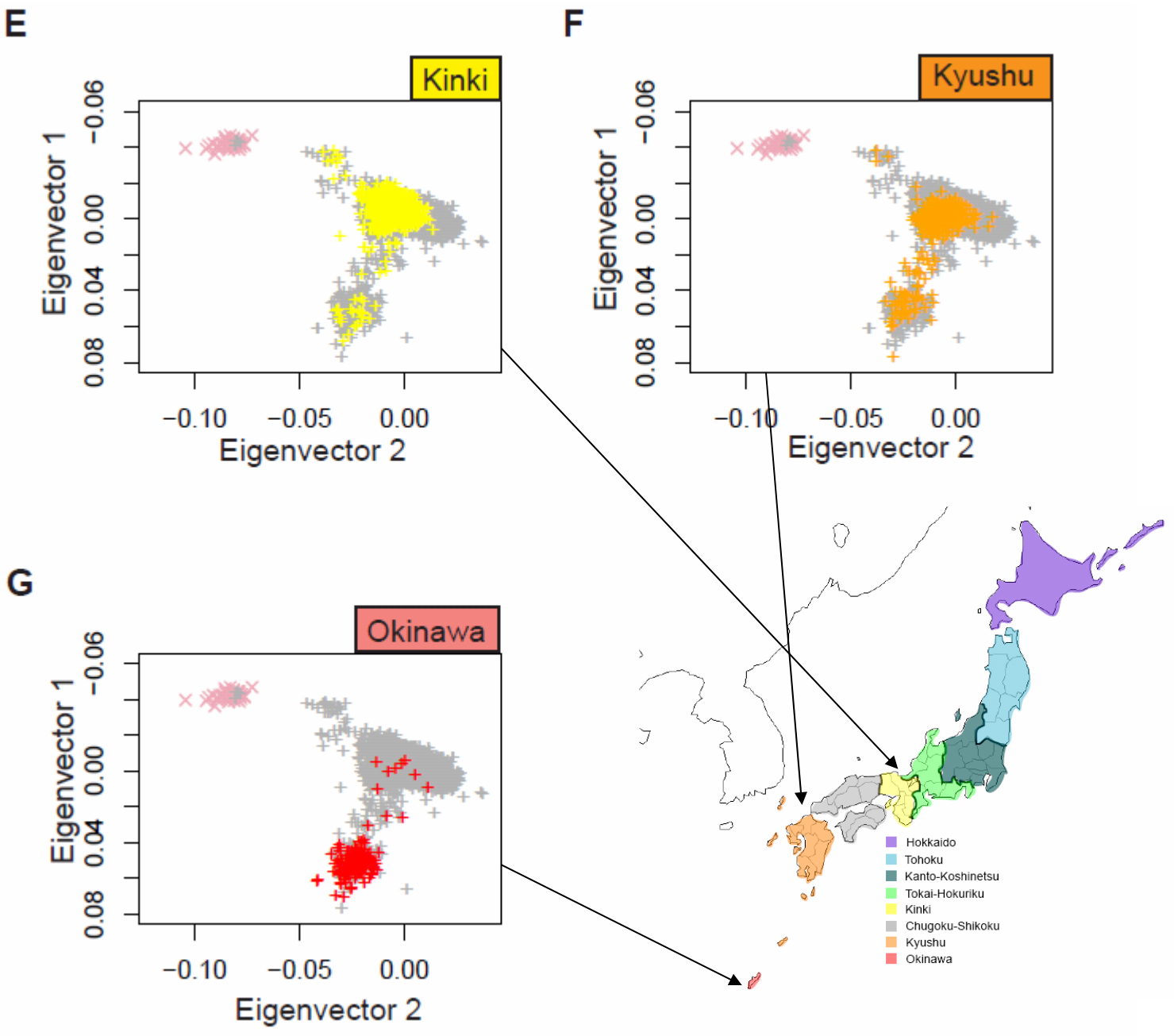
East & West

Novembre J et al. Nature. 2008;456(7218):98-101.

Why are the results of the genomics analysis so clean?

Because the probabilities of Mendelian laws are exact, and those for genomics data are stable.


A
Hokkaido
Eigenvector 1
−0.06
0.00
0.04
0.08
−0.10
−0.05
0.00
Eigenvector 2
B
Tohoku
Eigenvector 1
−0.06
0.00
0.04
0.08
−0.10
−0.05
0.00
Eigenvector 2
C
Kanto-Koshinetsu
Eigenvector 1
−0.06
0.00
0.04
0.08
−0.10
−0.05
0.00
Eigenvector 2
D
Tokai-Hokuriku
Eigenvector 1
−0.06
0.00
0.04
0.08
−0.10
−0.05
0.00
Eigenvector 2
Hokkaido
Tohoku
Kanto-Koshinetsu
Tokai-Hokuriku
Kinki
Chugoku-Shikoku
Kyushu
Okinawa

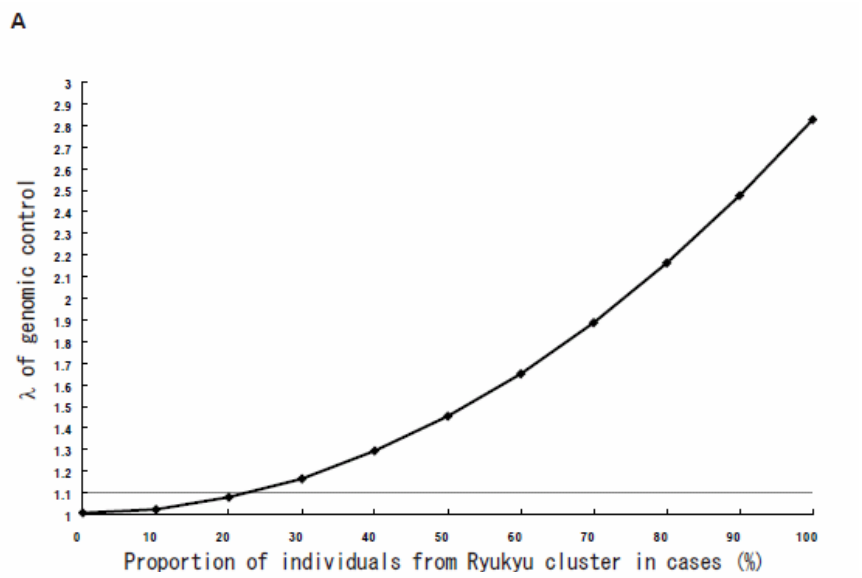

E
Kinki
F
Kyushu
G
Okinawa
Eigenvector 1
Eigenvector 2
−0.06
0.00
0.04
0.08
−0.10
−0.05
0.00
Hokkaido
Tohoku
Kanto-Koshinetsu
Tokai-Hokuriku
Kinki
Chugoku-Shikoku
Kyushu
Okinawa

Missense substitutions whose allele frequencies are the most remarkably different between Ryukyu and Hondo clusters

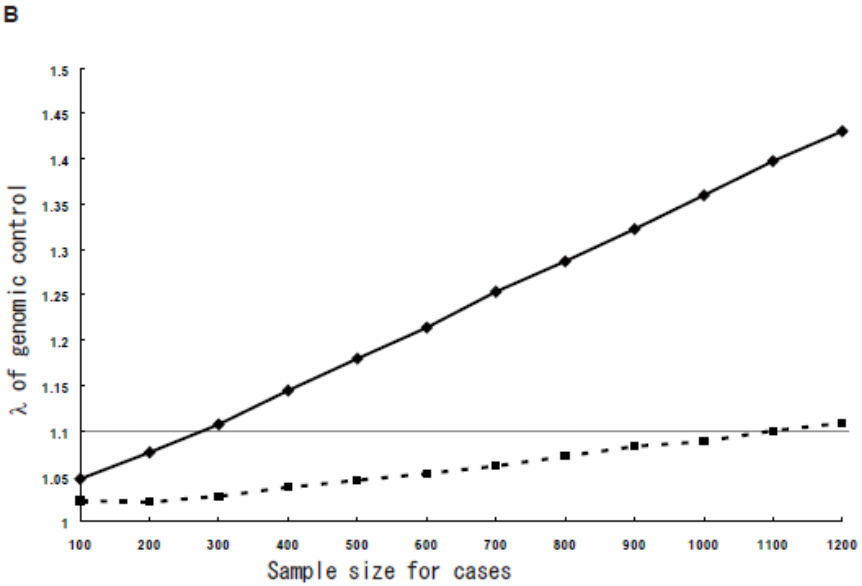
SNP	Chrom.	Position	Gene	Amino acid change	P-value*	
rs3827760	2	108880033	*EDAR*	370 Ala/Val	7.73×10^{-21}	Difference in hair thickness
rs17822931	16	46815699	*ABCC11*	180 Gly/Arg	1.63×10^{-20}	Dry and wet ear wax
rs2274067	1	229443429	*C1orf131*	28 Val/Leu	1.20×10^{-15}	
rs3744921	18	28121686	*FAM59A*	291 Lys/Arg	1.11×10^{-14}	
rs9932051	16	10482297	*ATF7IP2*	537 Thr/Ile	1.63×10^{-12}	
rs2465811	12	69276321	*PTPRB*	127 Gly/Ser	2.55×10^{-12}	
rs2589957	15	88704315	*MGC75360*	83 Asn/Ser	2.85×10^{-12}	
rs928302	21	42683153	*TMPRSS3*	53 Val/Ile	4.87×10^{-12}	
rs2273697	10	101553805	*ABCC2*	417 Ile/Val	1.18×10^{-11}	
rs3734166	5	137693222	*CDC25C*	70 Cys/Arg	1.01×10^{-10}	
rs3765534	13	94613416	*ABCC4*	757 Glu/Lys	1.54×10^{-10}	
rs2070235	20	41764871	*MYBL2*	427 Ser/Gly	4.81×10^{-10}	
rs2289178	15	46842064	*CEP152*	700 Ile/Ser	5.27×10^{-10}	
rs3778922	7	151433265	*GALNT11*	197 Tyr/Asp	9.64×10^{-10}	
rs10487075	7	88802957	*FLJ32110*	909 Lys/Glu	1.52×10^{-9}	
rs14103	1	35093829	*LOC113444*	14 Leu/Val	4.88×10^{-9}	
rs2275586	10	99230748	*MMS19L*	68 Gly/Ala	1.16×10^{-8}	
rs2228226	12	56152088	*GLI1*	1100 Gln/Glu	1.17×10^{-8}	
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	

Increase in λ value

Controls all belong to Hondo cluster 200 cases and 200 controls



Controls all belong to Hondo cluster. 10 or 20% of the cases belong to Ryukyu cluster

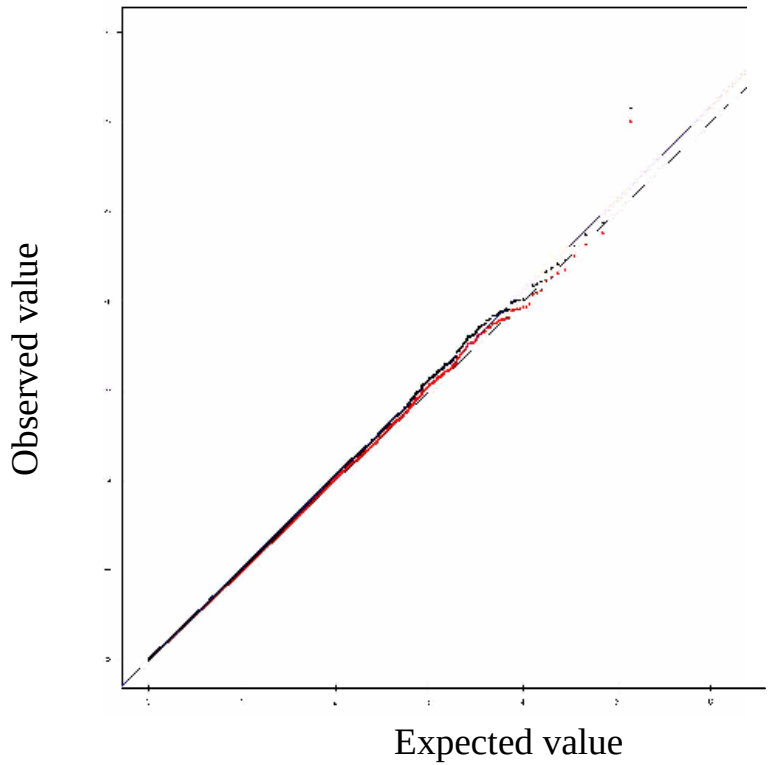


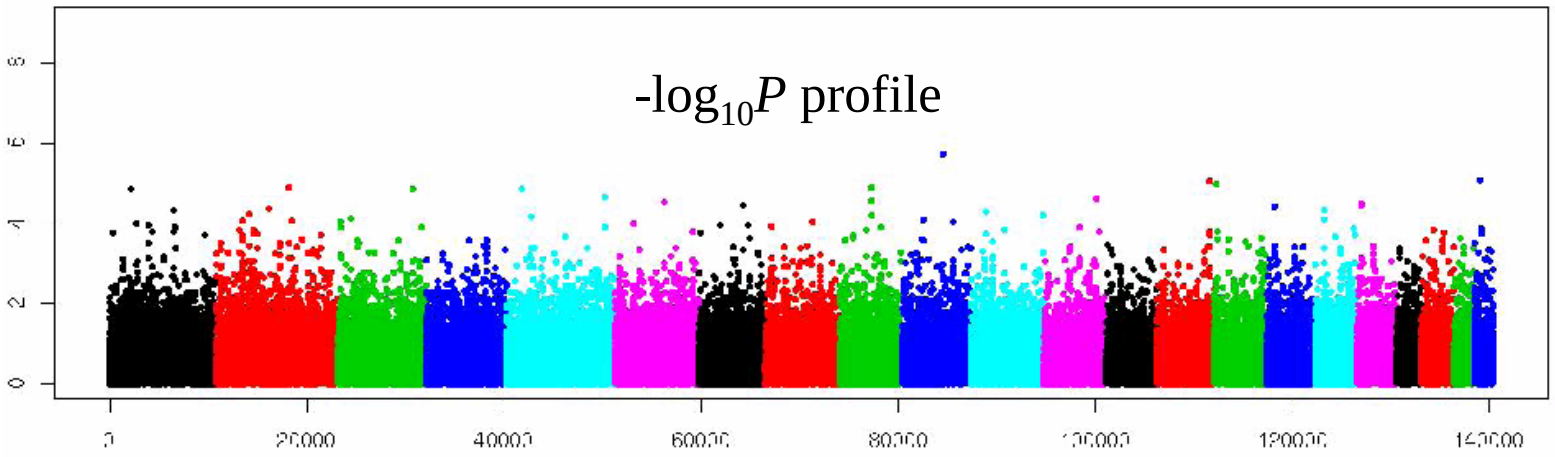
Solution to multiple-comparison problem

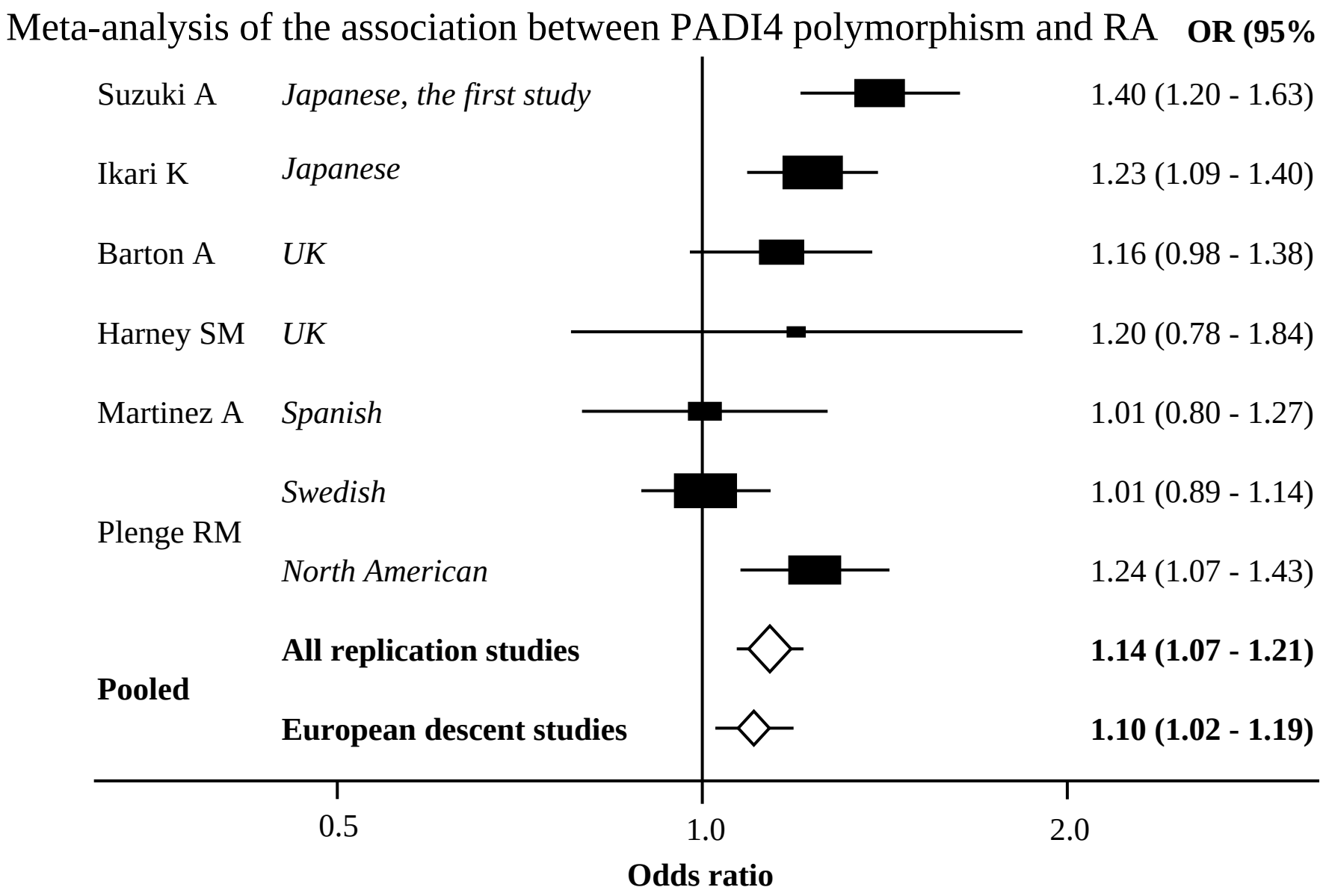
1. If 500,000 independent tests are performed at the significance level of alpha = 0.05, as many as 2,500 SNPs will exhibit significance.
2. Bonferroni's correction (alpha = 10^{-7}) is too conservative but desirable.
3. Other methods include, use of FDR concept, use of Bayes' method, and permutation test.
4. We have also propose an exact method (Misawa K et al. New correction algorithms for multiple comparisons in case-control multilocus association studies based on haplotypes and diplotype configurations. J Hum Genet. 2008;53(9):789-801).

Report of the results of an association study

QQ plot



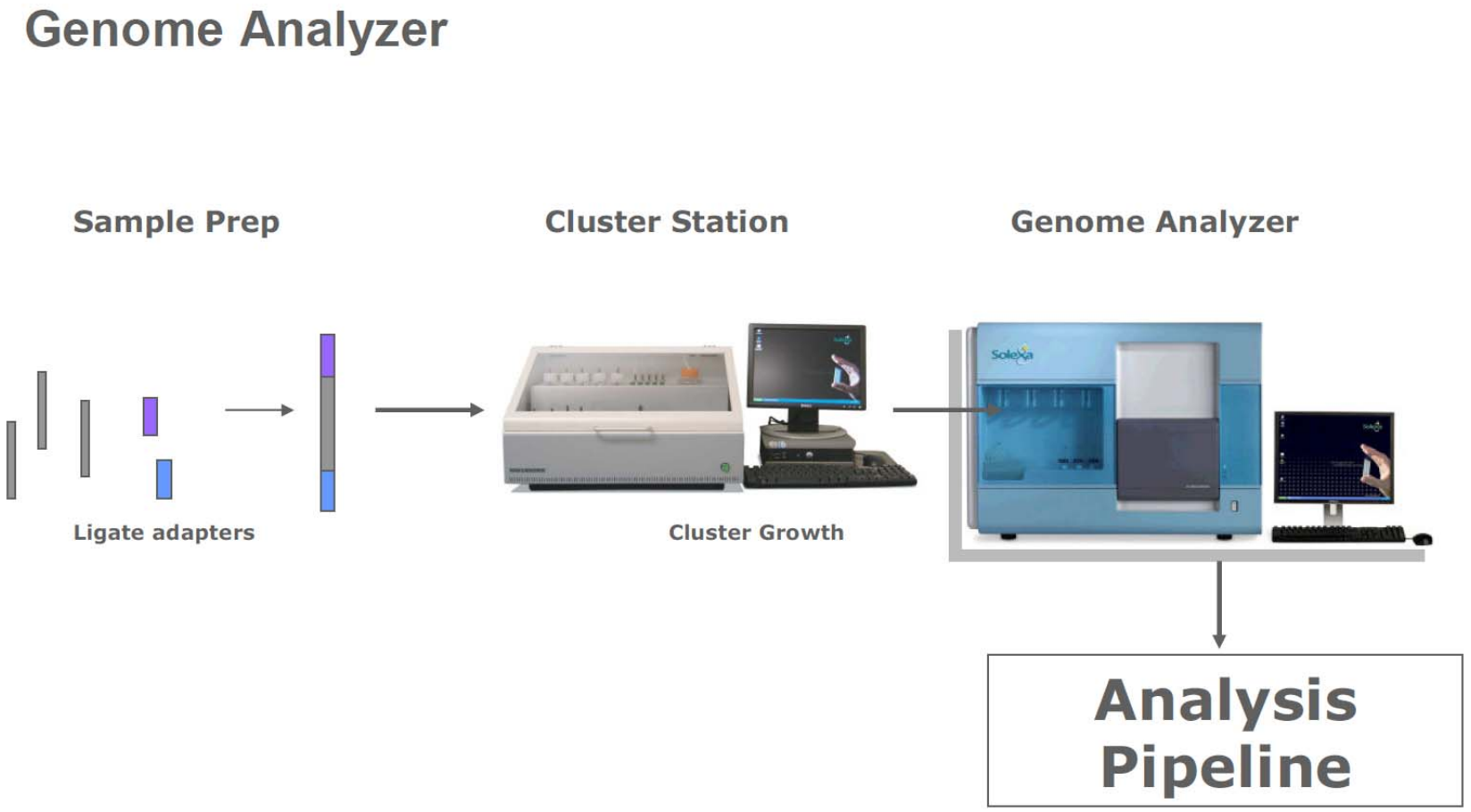




ORs (proportional to sample size) with 95% CIs from each study testing the association of RA with the risk allele of PADI4 gene. The pooled ORs with 95% CI for overall analysis and subgroup analysis in populations of European descent were calculated with the Mantel–Haenszel method (diamonds). The first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

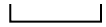
SNP genotype data are only the beginning of the large-scale genomics data

Giga sequencer (Next-generation sequencer)



Size of the data

ATCGGAT …..（If we write in this size）

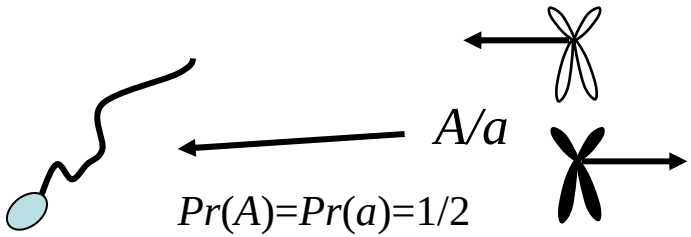


Data	size	length
Personal genome	3.2×10^9	From Tokyo to New York
Giga sequencer	3×10^{10}/run	2 fold circumstance of the earth
Whole genotypes in CGM	10^{11}	8.2 fold circumstance of the earth
Genome of whole cells in a subject	3.6×10^{23}	Our galaxy (100 thousand light years)
10 Peta flops computer	3.2×10^{23}/year	

Supercomputing is extremely useful when the size of the data is large and ……. ***the probabilities are stable***.

Probabilities in genetics are, unlike economics, very stable

Mendelian law of segregation



$Pr(A)=Pr(a)=1/2$

Stock price

$Pr(up)$= ???
$Pr(down)$= ???

Probabilities in Mendelian laws are exact as stated by RA Fisher

Stable probabilities were selected for in the evolution, and
are maintained by the molecular mechanism.

Subprime mortgage problem will not occur in biology.

By “stable probabilities”, I mean

The following limit theorems hold not only in theory but also in the **real world**.

Law of large numbers

Number

Sum

$$\lim_{n\to\infty} \frac{S_n}{n} = \mu$$

The larger the size of the data, the more accurate the estimation will be.

Central limit theorem

Number

Sum

$$\lim_{n\to\infty} \frac{S_n/n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

S.D.

Law of iterated logarithm

$$|\theta - \hat{\theta}_n| < C\sqrt{\frac{\log\log n}{n}}$$

Error of estimate

Supercomputing is extremely useful when the size of the data is large and ……. ***the probabilities are stable***.