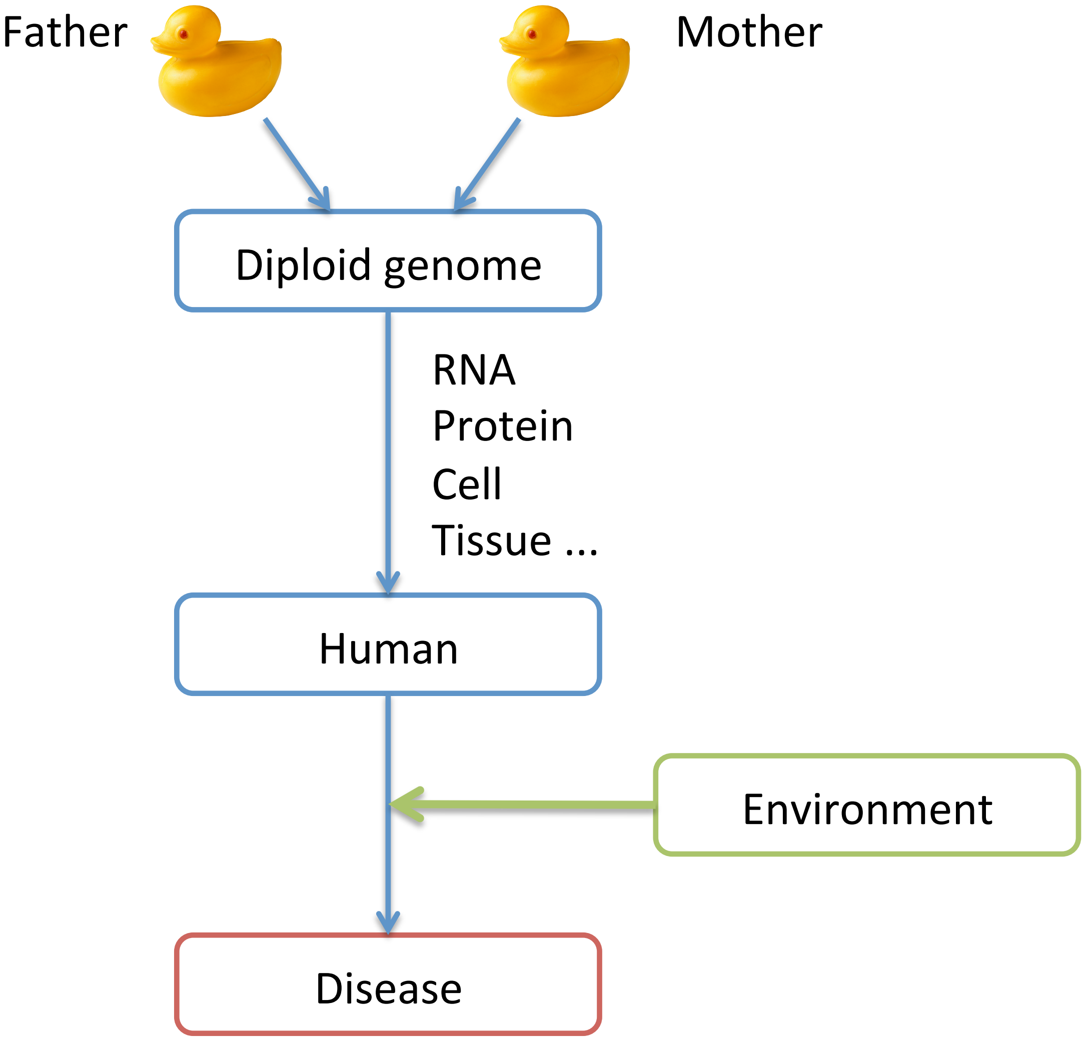
理化学研究所

Statistical analyses used for gene mapping of human diseases

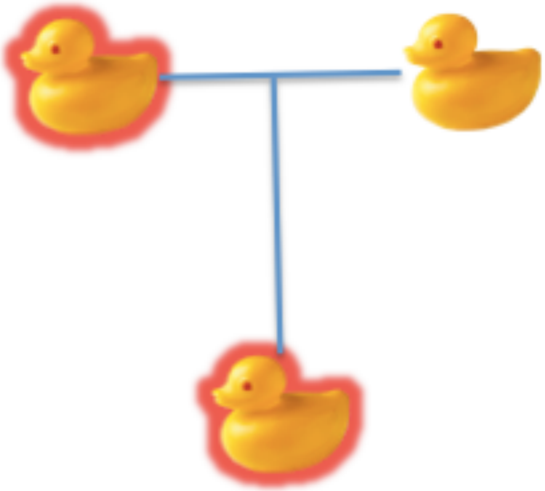
Yoichiro Kamatani
Laboratory for Statistical Analysis
RIKEN center for Integrative Medical Sciences

Human disease genetic mapping



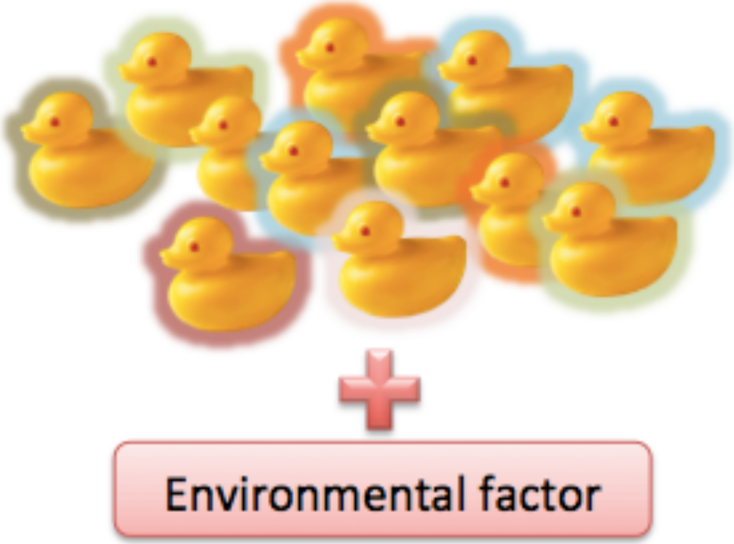
Human disease genetic mapping

Mendelian disease



Linkage analysis

Complex disease

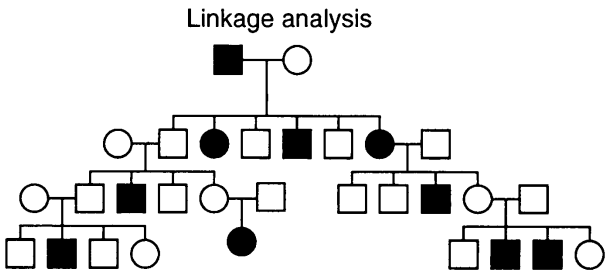


Environmental factor

Genetic association analysis

Linkage analysis

- Linkage analysis



Linkage analysis

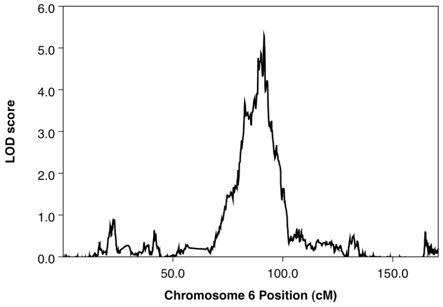
Model parameters explicitly, estimate them, and the lod score, a kind of likelihood ratio, is evaluated.

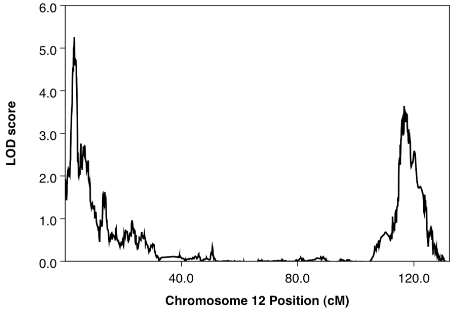
Likelihood calculation

Elston-Stewart algorithm

Lander-Green algorithm

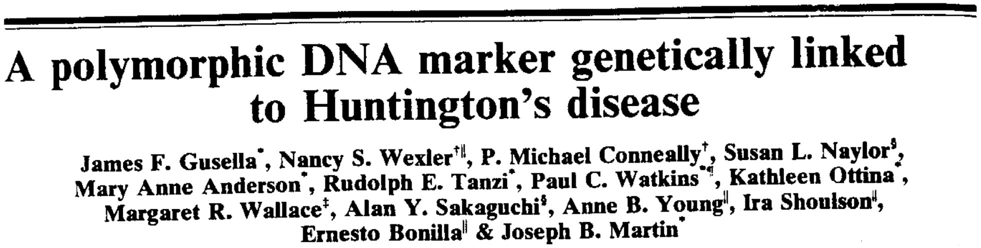
MCMC based algorithm





Linkage analysis

- Successes of linkage analysis



A polymorphic DNA marker genetically linked to Huntington's disease

James F. Gusella[*], Nancy S. Wexler[†‖], P. Michael Conneally[†], Susan L. Naylor[§], Mary Anne Anderson[*], Rudolph E. Tanzi[*], Paul C. Watkins[*¶], Kathleen Ottina[*], Margaret R. Wallace[‡], Alan Y. Sakaguchi[§], Anne B. Young[‖], Ira Shoulson[‖], Ernesto Bonilla[‖] & Joseph B. Martin[*]

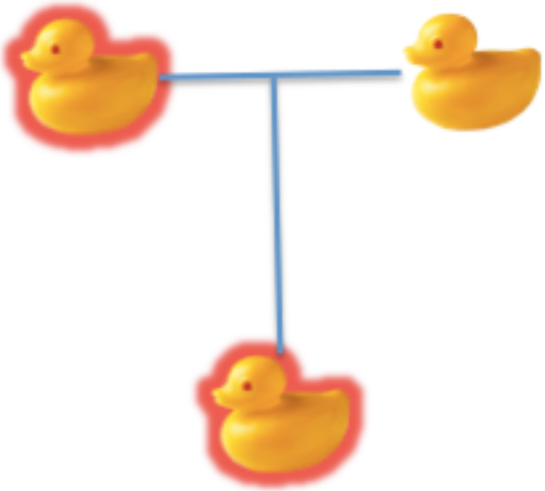
Gusella et al, Nature 1983
→ Huntingtin gene

  - Cystic fibrosis
  - Familial breast cancer (BRCA1 / BRCA2)
  - possibly, Familial hypercholesterolemia (LDLR)

They are Mendelian diseases

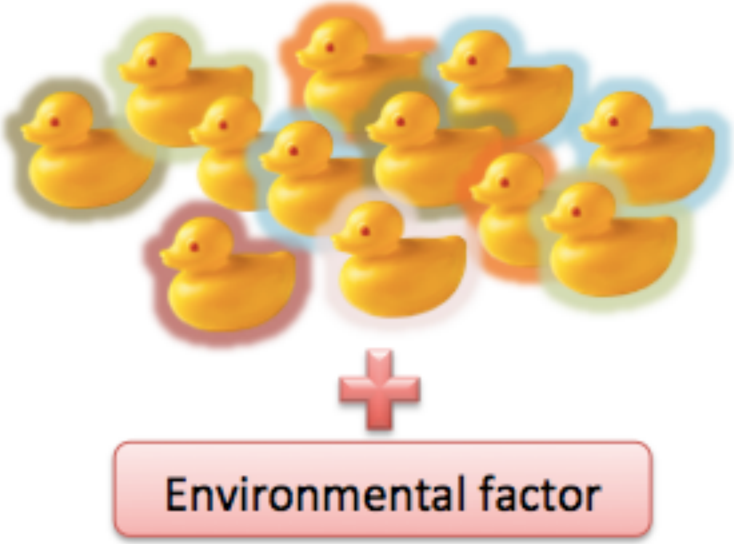
Gene mapping study

Mendelian disease



Linkage analysis

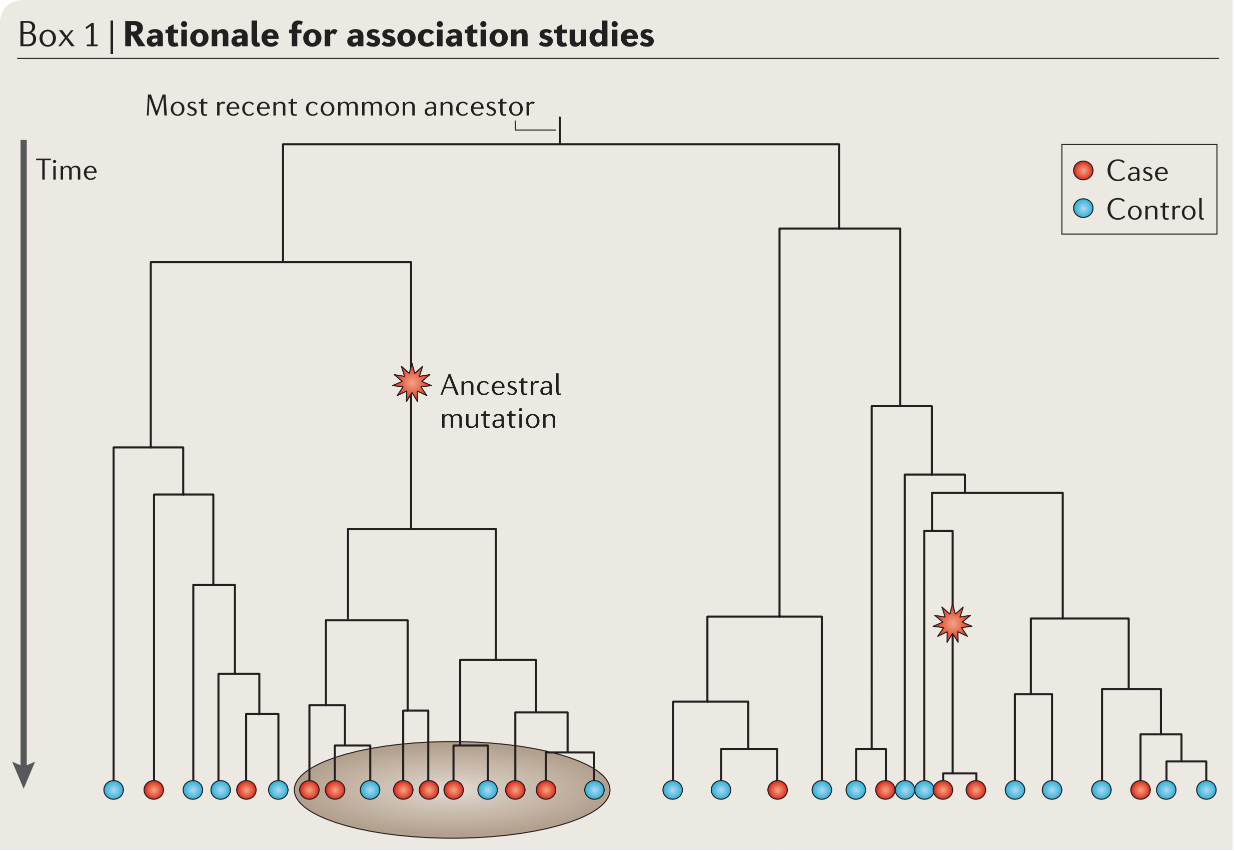
Complex disease



Environmental factor

Genetic association analysis

Genetic association study

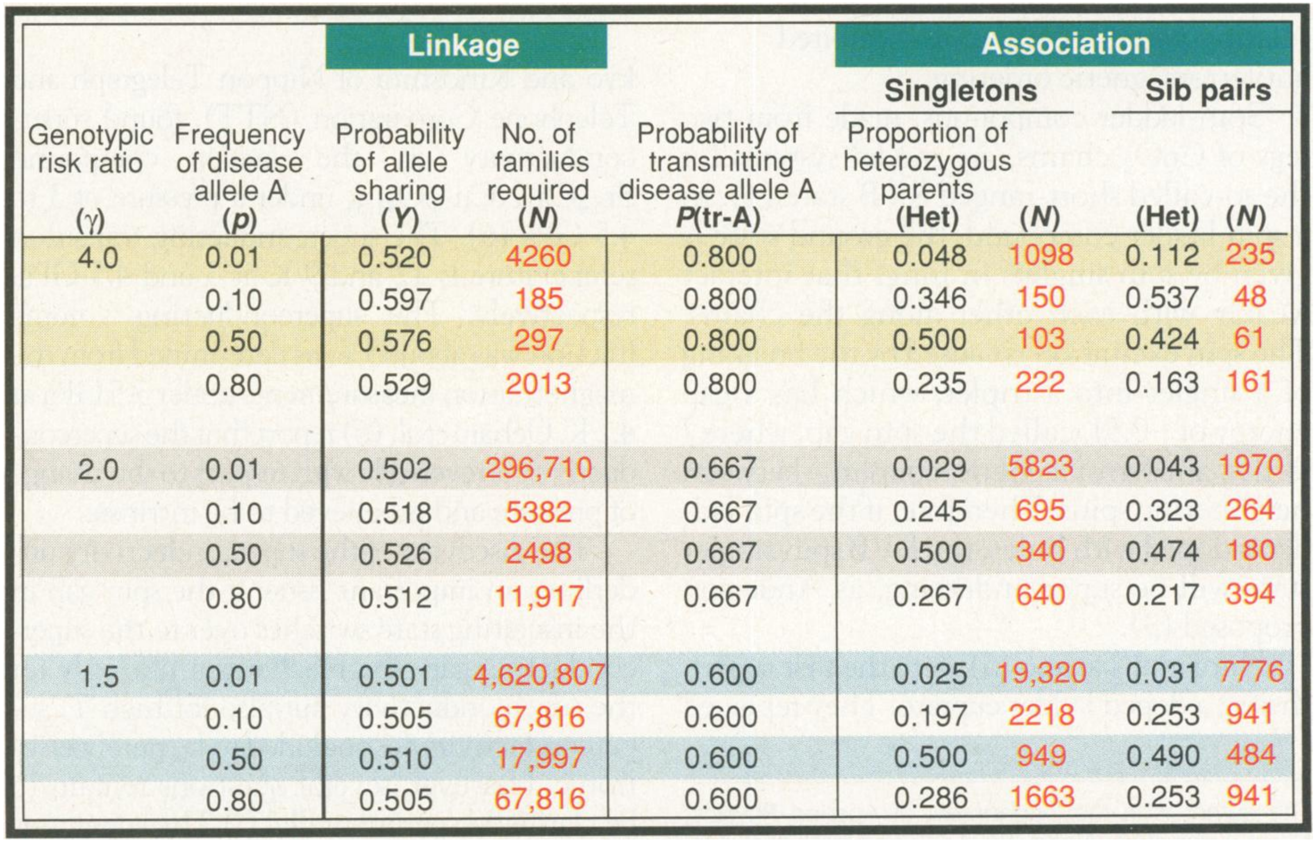


Box 1 | Rationale for association studies
Most recent common ancestor
Time
Ancestral mutation
Case
Control

(Balding DJ. Nat Rev Genet 2006; 7:781-91.)

Allele frequency might differ between Case and Control
Detect it by using association testing

Do association test … why?



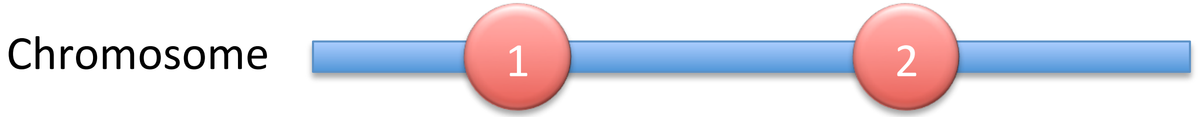
		Linkage			Association			
					Singletons		Sib pairs	
Genotypic risk ratio (γ)	Frequency of disease allele A (*p*)	Probability of allele sharing (*Y*)	No. of families required (*N*)	Probability of transmitting disease allele A *P*(tr-A)	Proportion of heterozygous parents (Het)	(*N*)	(Het)	(*N*)
4.0	0.01	0.520	4260	0.800	0.048	1098	0.112	235
	0.10	0.597	185	0.800	0.346	150	0.537	48
	0.50	0.576	297	0.800	0.500	103	0.424	61
	0.80	0.529	2013	0.800	0.235	222	0.163	161
2.0	0.01	0.502	296,710	0.667	0.029	5823	0.043	1970
	0.10	0.518	5382	0.667	0.245	695	0.323	264
	0.50	0.526	2498	0.667	0.500	340	0.474	180
	0.80	0.512	11,917	0.667	0.267	640	0.217	394
1.5	0.01	0.501	4,620,807	0.600	0.025	19,320	0.031	7776
	0.10	0.505	67,816	0.600	0.197	2218	0.253	941
	0.50	0.510	17,997	0.600	0.500	949	0.490	484
	0.80	0.505	67,816	0.600	0.286	1663	0.253	941

Comparison of linkage and association studies. Number of families needed for identification of a disease gene.

Risch N and Merikangas K. Science 1996; 273: 1516.

Linkage disequilibrium

- Suppose two genetic loci



- Alleles at these loci are independent if ...
  - these two loci locate on different chromosomes because of Mendel's law of segregation
  - these two loci locate on the same chromosome, but their distance is long enough to become independent because of repetitive meiotic recombination

- Otherwise they are associated, in other words, **in linkage disequilibrium (LD)**. If locus 1 and 2 are in LD, and locus 1 is the causative locus, then locus 2 would also show association.

Linkage disequilibrium

- LD is defined as "non-random sharing of combinations of variants"

When f_A = 0.1 and f_B = 0.4

Random sharing

	B	**b**
A	0.04	0.06
a	0.36	0.54

$$f_{AB} = f_A f_B$$

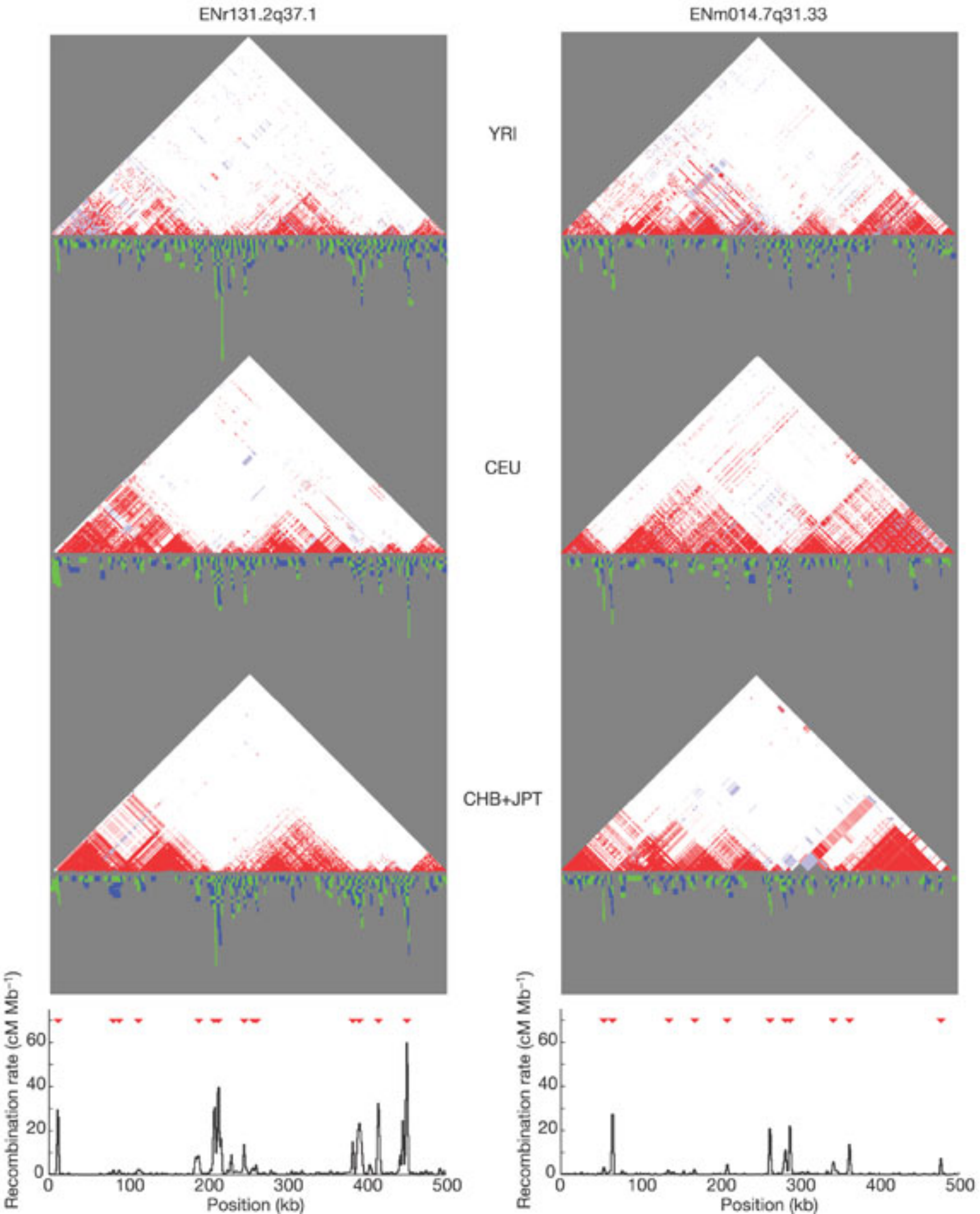
Non-random sharing

	B	**b**
A	0.00	0.10
a	0.40	0.50

$$f_{AB} \neq f_A f_B$$

$$D_{AB} = f_{AB} - f_A f_B$$

$$r^2 = \frac{D^2}{f_A f_a f_B f_b}$$



No need to genotype all the variants in a region ... it is enough to select some SNPs which are in LD with other variants.

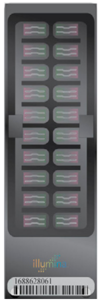
In general, a SNP set which have r^2 value above 0.8 with the other SNPs is called “tagSNPs”.

Simple statistical genetics calculation can show that N/r^2 sample size is needed to achieve the same power to detect association with the “tagSNPs”.

The International HapMap consortium. Nature 2005.

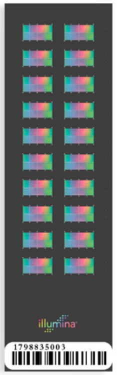
Commercial SNP arrays based on HapMap tag SNPs

~ 2,500,000 SNPs
discovered by HapMap2



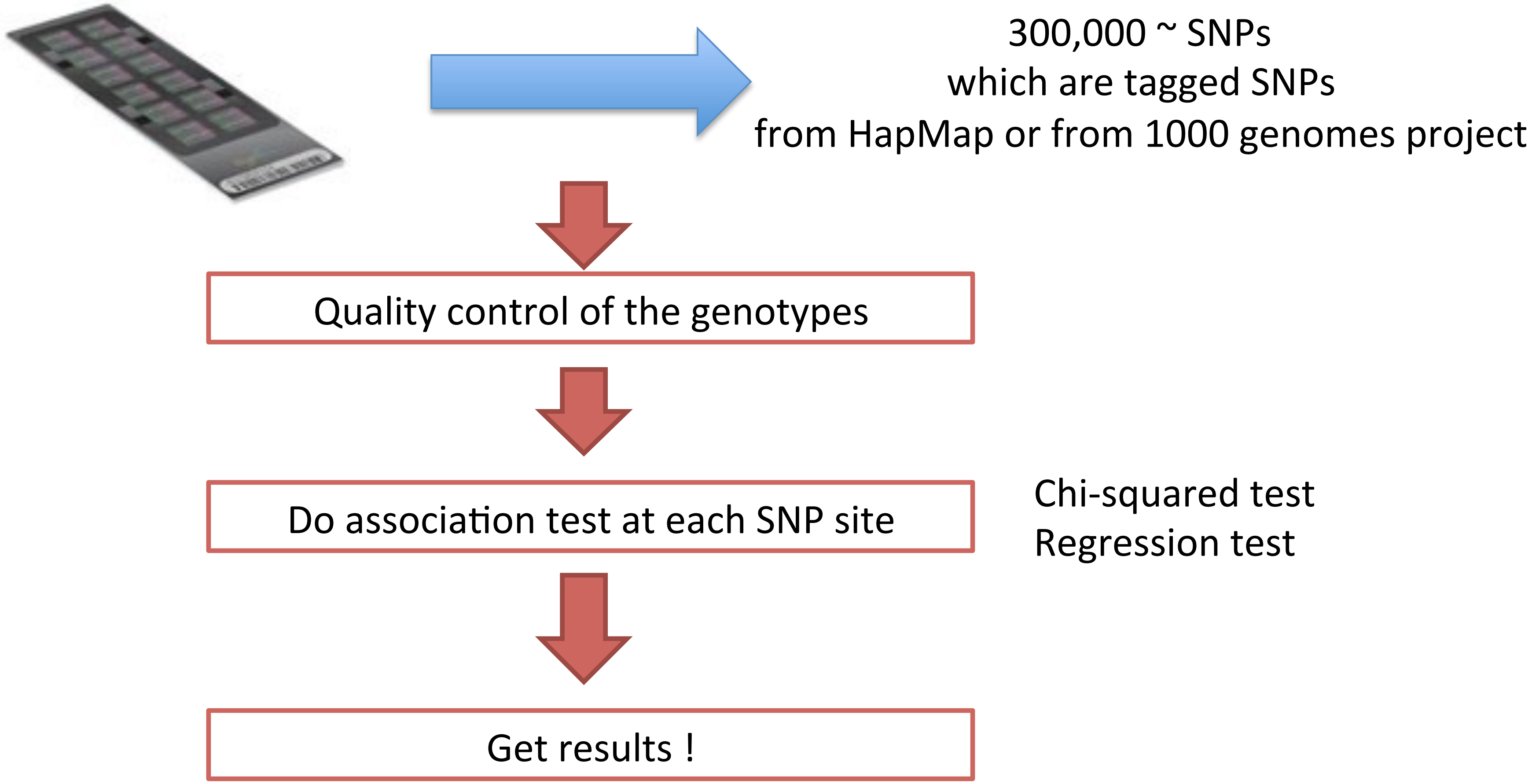
~ 300,000 tagSNPs
representing SNPs
in Caucasian

~ 500,000 tagSNPs
representing SNPs
in Caucasian + Asian



~ 1,000,000 tagSNPs
representing SNPs in
Caucasian + Asian + African

Genome-wide association study (GWAS)



300,000 ~ SNPs
which are tagged SNPs
from HapMap or from 1000 genomes project

Quality control of the genotypes

Do association test at each SNP site

Chi-squared test
Regression test

Get results !

Q-Q plot for quality control

- GWAS uses hundreds of thousands of SNP results, and the quality must be assured by Quality Control processes. And this can be evaluated by Q-Q plot.

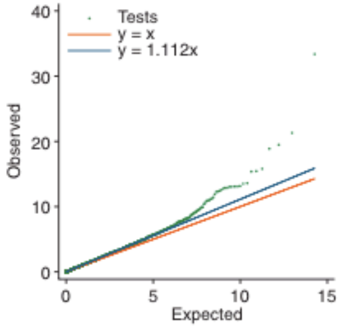


Figure 2 Quantile-quantile plots of Cochran-Armitage test statistics. The ranked, observed values for 6,322 nsSNPs are plotted against the values expected for sampling from a χ^2 distribution with one degree of freedom (the distribution expected under the null hypothesis).

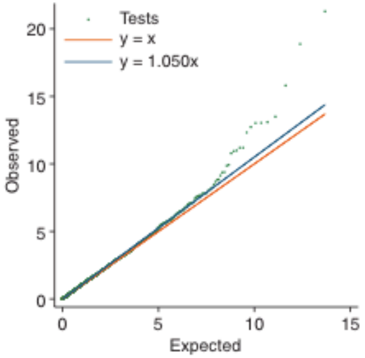
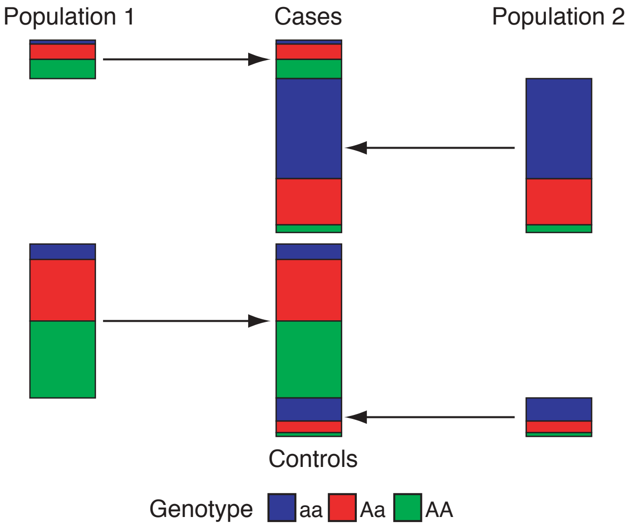


Figure 3 Quantile-quantile plots of Cochran-Armitage test statistics of 4,629 nsSNPs with half-call rates <0.5% and a difference in call rates between cases and controls of no more than 5%.

Clayton D et al. Nat Genet 2005; 37: 1423.

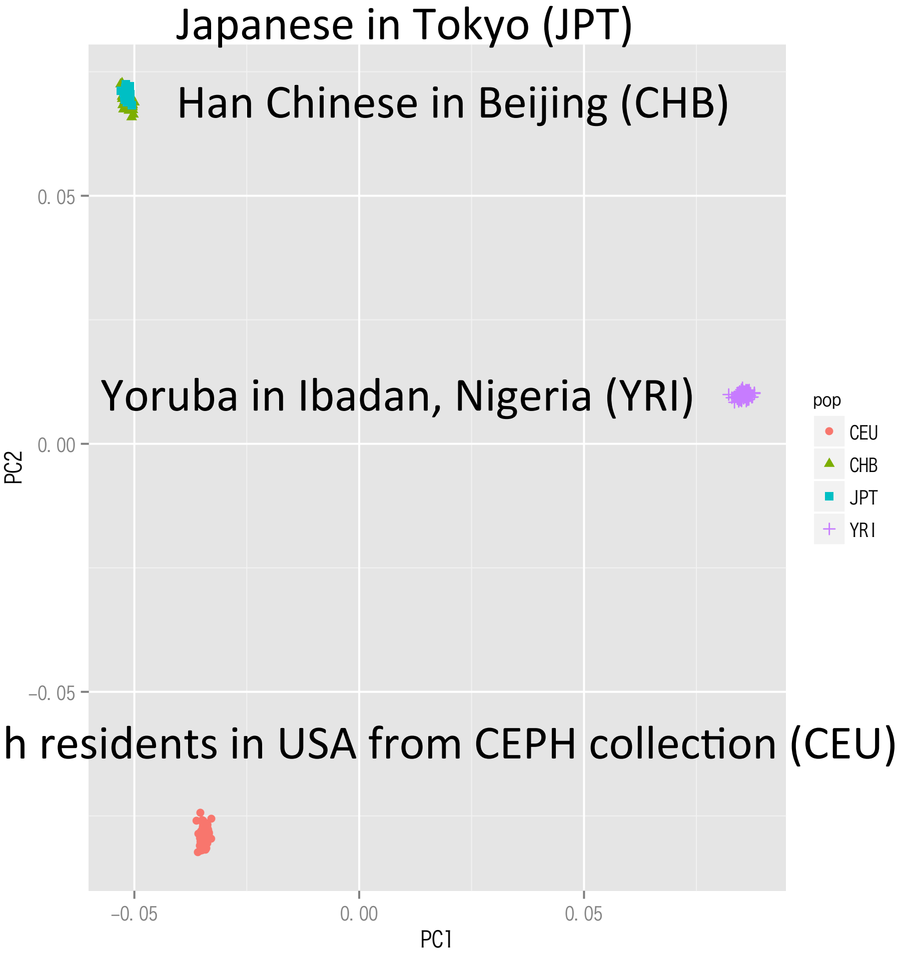
Population stratificaiton

- Population stratification can cause false positives



Marchini J et al. Nat Genet 2004; 36: 512.

Principal component analysis



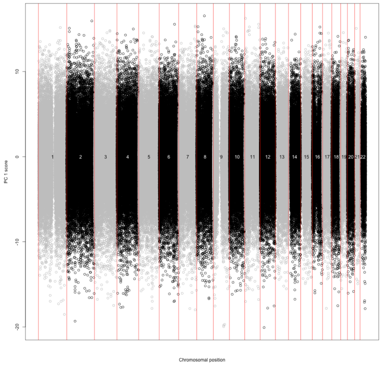
Japanese in Tokyo (JPT)

Han Chinese in Beijing (CHB)

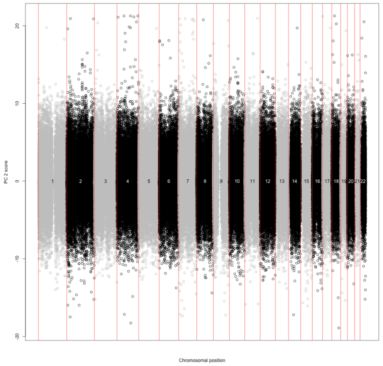
Yoruba in Ibadan, Nigeria (YRI)

Utah residents in USA from CEPH collection (CEU)

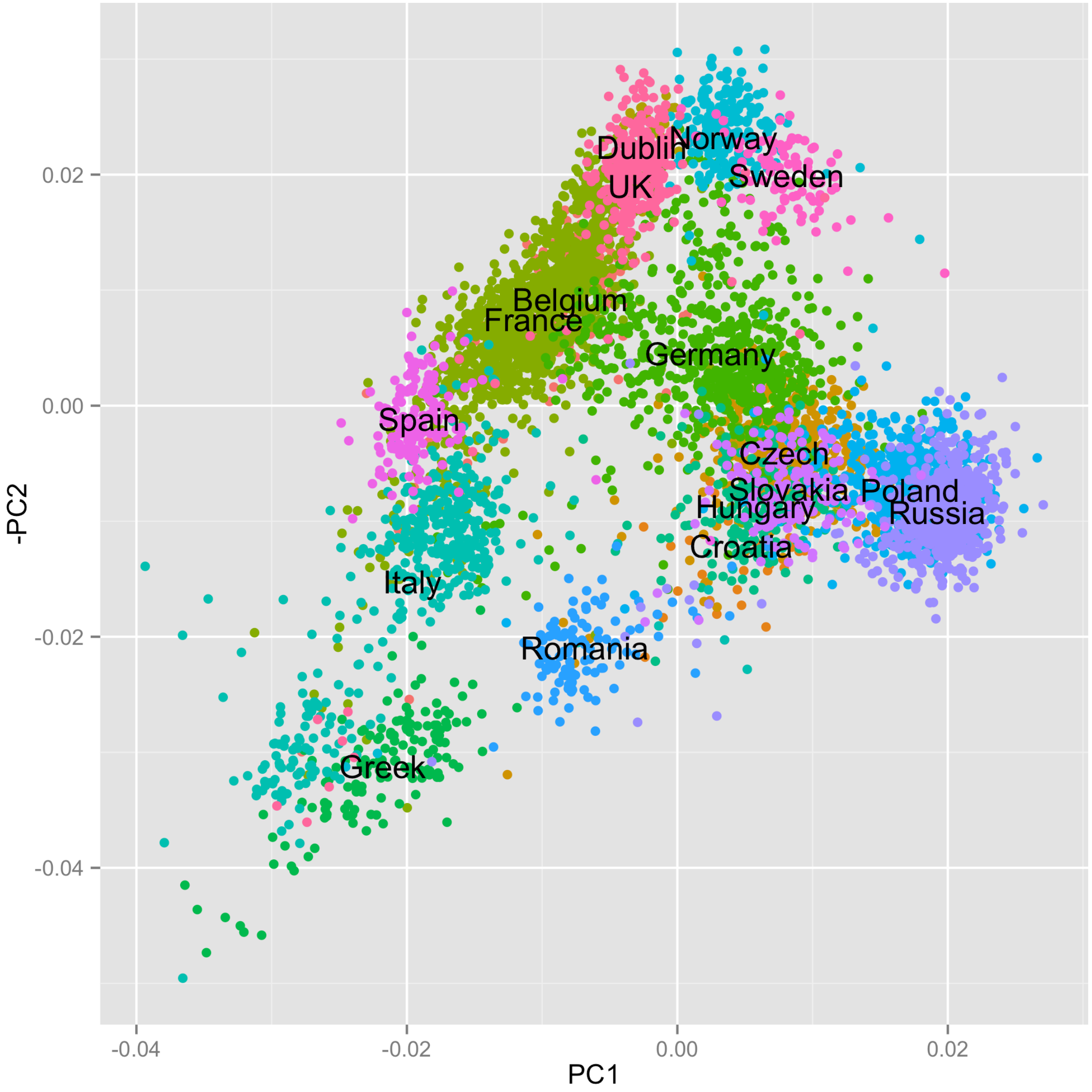
Score for PC1

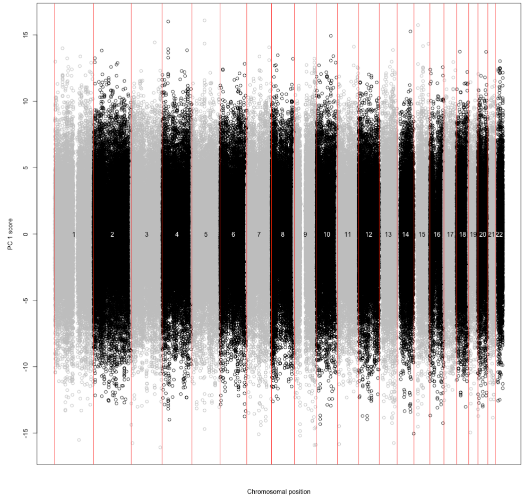


Score for PC2

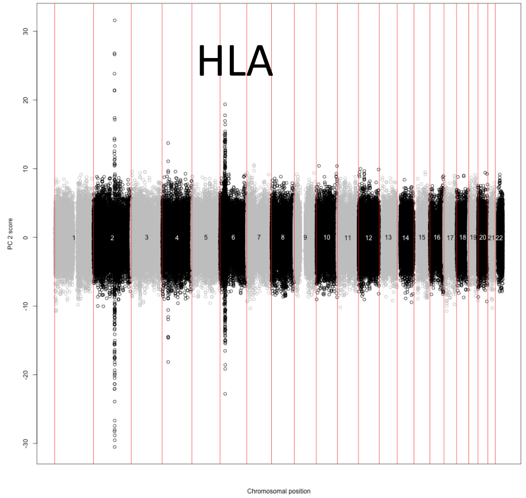


Principal component analysis





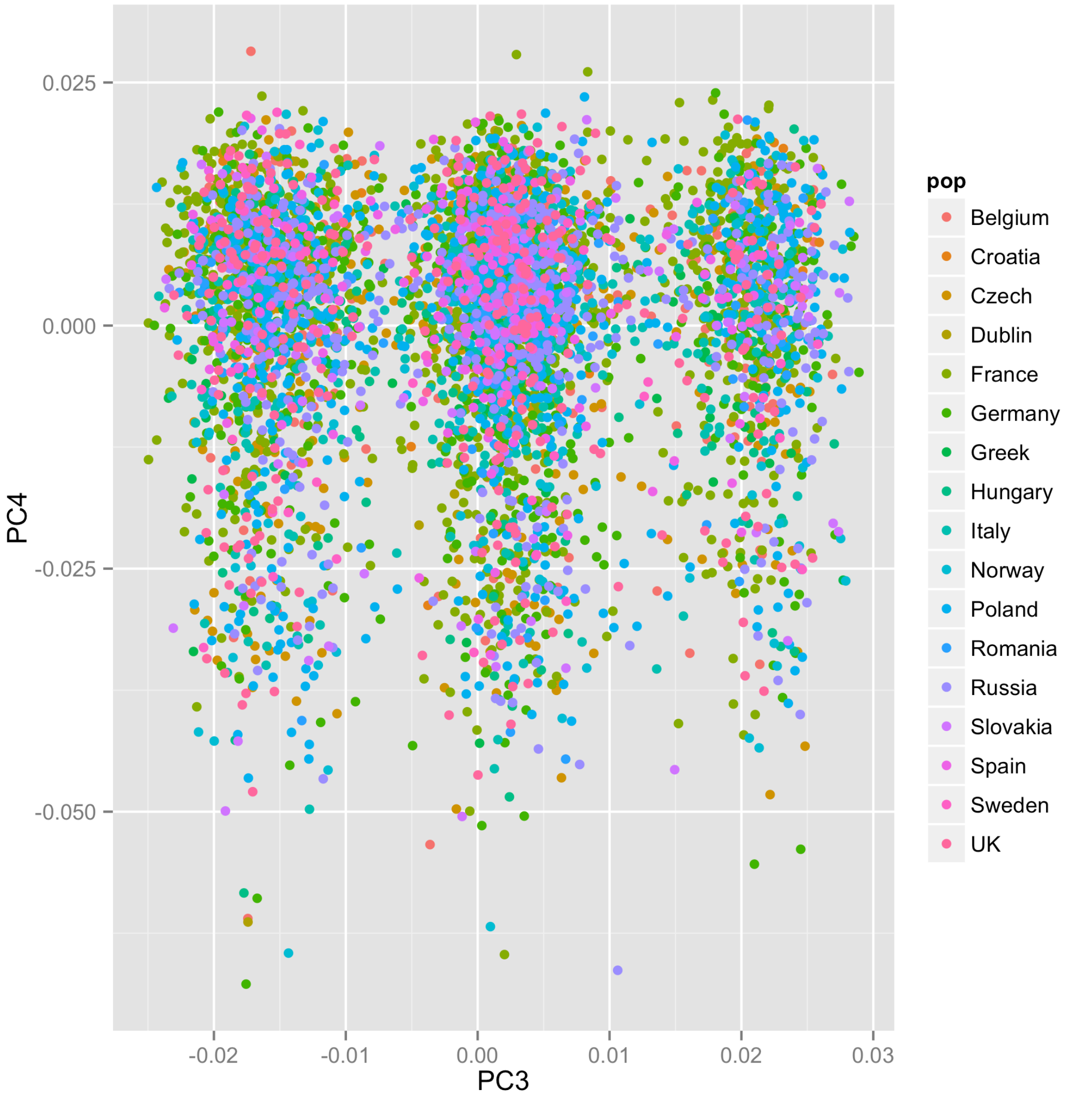
LCT

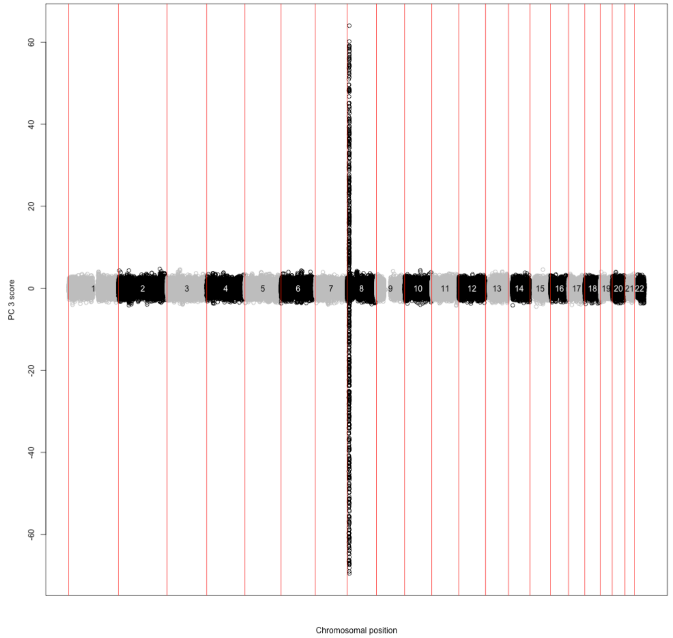


HLA

(Heath, SC et al. EJHG 2008; 16: 1413.)

Principal component analysis





Common inversion polymorphism in chromosome 8 (Hevra R, de la Chapelle A. AJHG 1976; 28: 208.)

Mixed Linear Model Association

- Relatedness between individuals in case or in control could cause spurious association since it can increase / decrease allele frequency irrespective of disease status.

- Typically, sample filtration is performed to remove 1st and 2nd degree relatedness, and possibly more.

- Mixed Linear Model Association (MLMA) is a solution to adjust any levels of relatedness

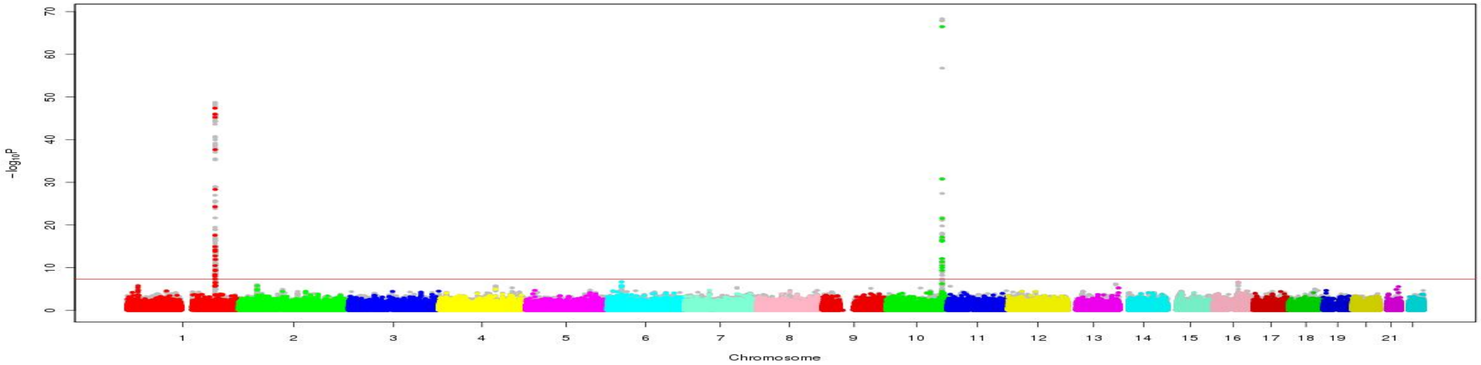
$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} + \boldsymbol{\epsilon}$$

$$\mathrm{Var}(\mathbf{u}) = \sigma_g^2 \mathbf{K}$$

K: matrix of pairwise genetic similarity

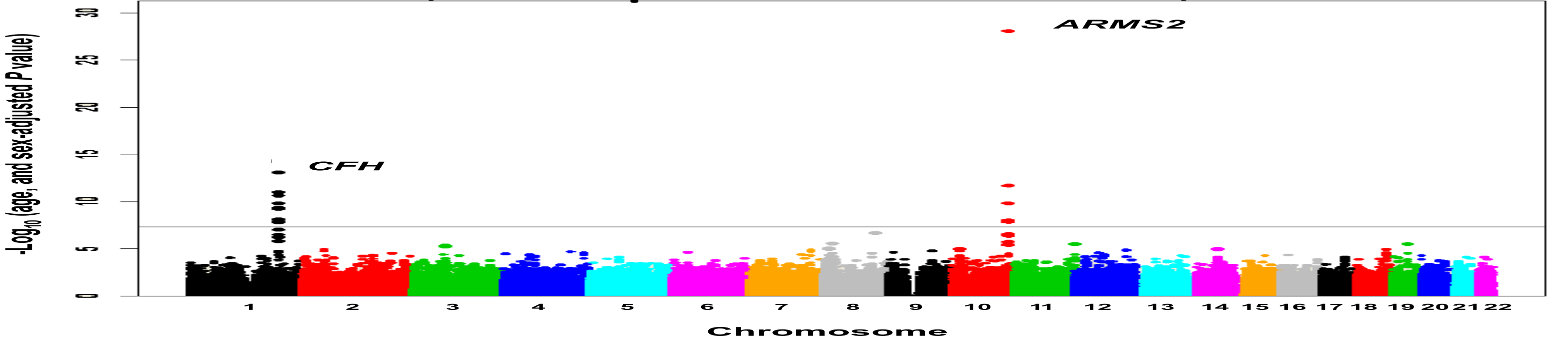
GWAS for Age-related Macular Degeneration (AMD)

European Cases ~ 1,000: European Controls ~ 4,000 -> 2 GW signals



(Leveillard T, Kamatani Y, Lathrop M et al. Unpublished data)

Japanese Cases ~ 1,500: Japanese Controls ~ 20,000 -> 2 GW signals



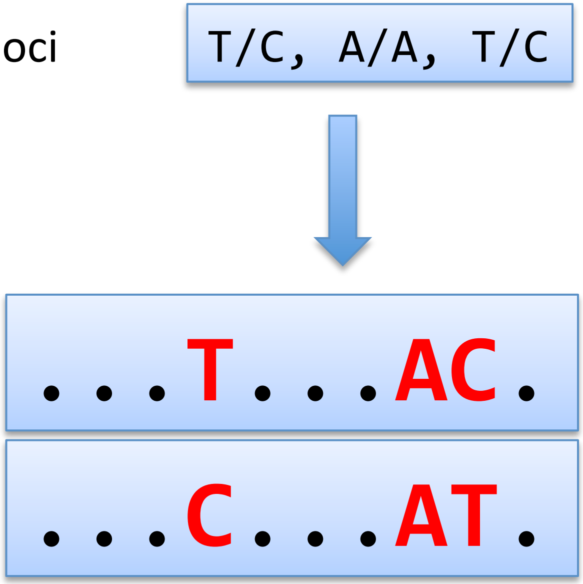
(Arakawa S et al. Nat Genet 2011;43:1001-5)

GWAS yields disease susceptible loci with confident associations and with robustness.

Imputation

Observed data at 3 genetic loci

Construct haplotypes

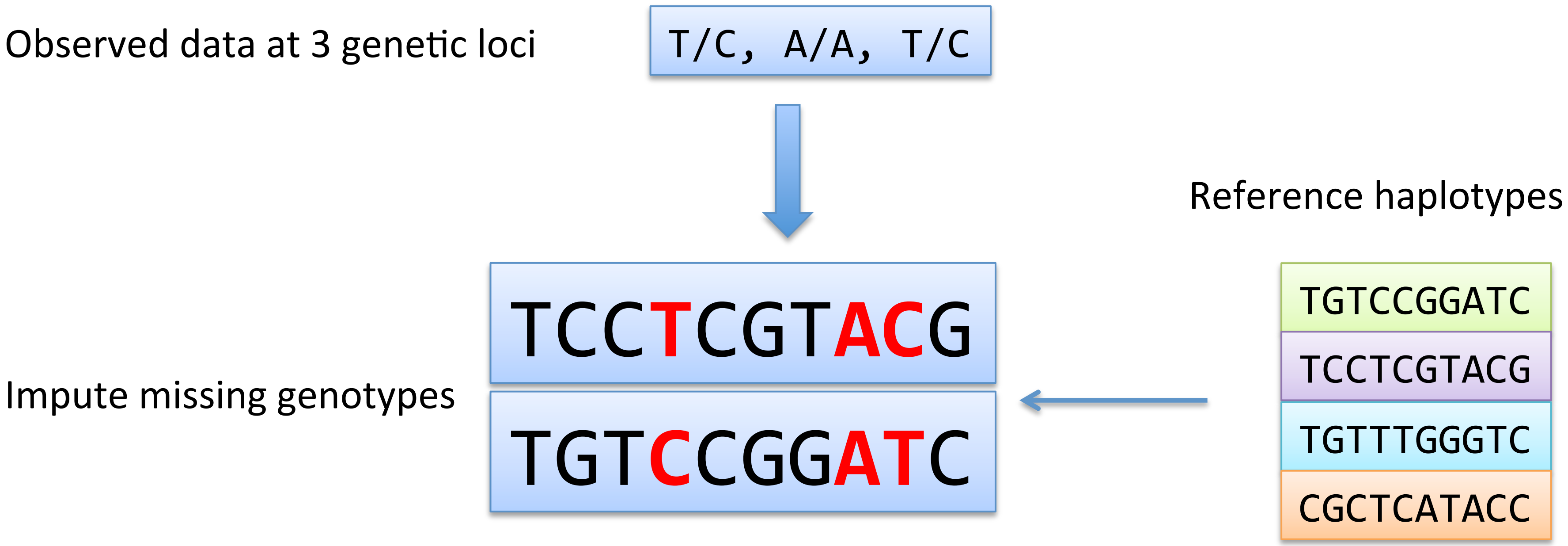


T/C, A/A, T/C

...T...AC.

...C...AT.

Imputation



Observed data at 3 genetic loci

T/C, A/A, T/C

Impute missing genotypes

TCCTCGTACG

TGTCCGGATC

Reference haplotypes

TGTCCGGATC

TCCTCGTACG

TGTTTGGGTC

CGCTCATACC

- The reference template is typically HapMap panel or 1000 genomes panel.
- Observed loci are typically from SNP arrays, of which loci are “tagSNPs” from HapMap or from 1000 genomes results.

Imputation (Marchini's model)

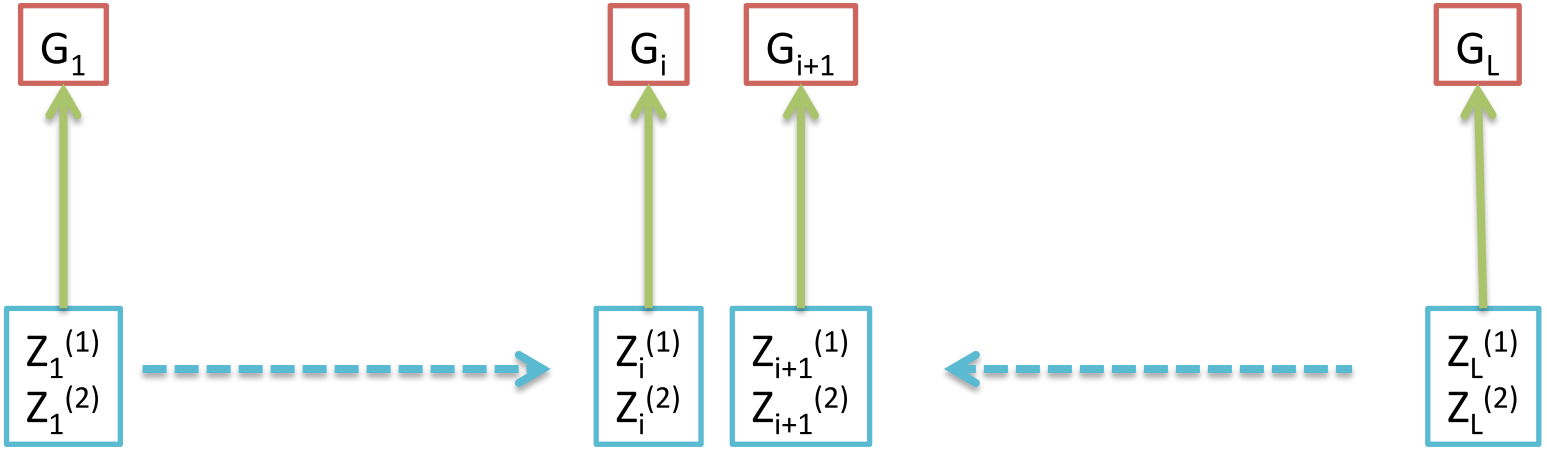
$$P(G_i|H,\theta,\rho)=\sum_{z}P(G_i|Z,\theta)P(Z|H,\rho)$$

G_i : vector of genotypes of individual i
H: population haplotypes
θ: other parameters
ρ: recombination map across the genome
Z: 2 copies of haplotypes from population, which form individual genotypes

Imputation (Marchini's model)

$$P(G_i|H,\theta,\rho) = \sum_{z} P(G_i|Z,\theta)P(Z|H,\rho)$$

Emission probability : governed by mutation rate



Transition probability : governed by recombination rate (ρ)

Imputation (Marchini's model)

$$P(G_i|H,\theta,\rho)=\sum_{z}P(G_i|Z,\theta)P(Z|H,\rho)$$

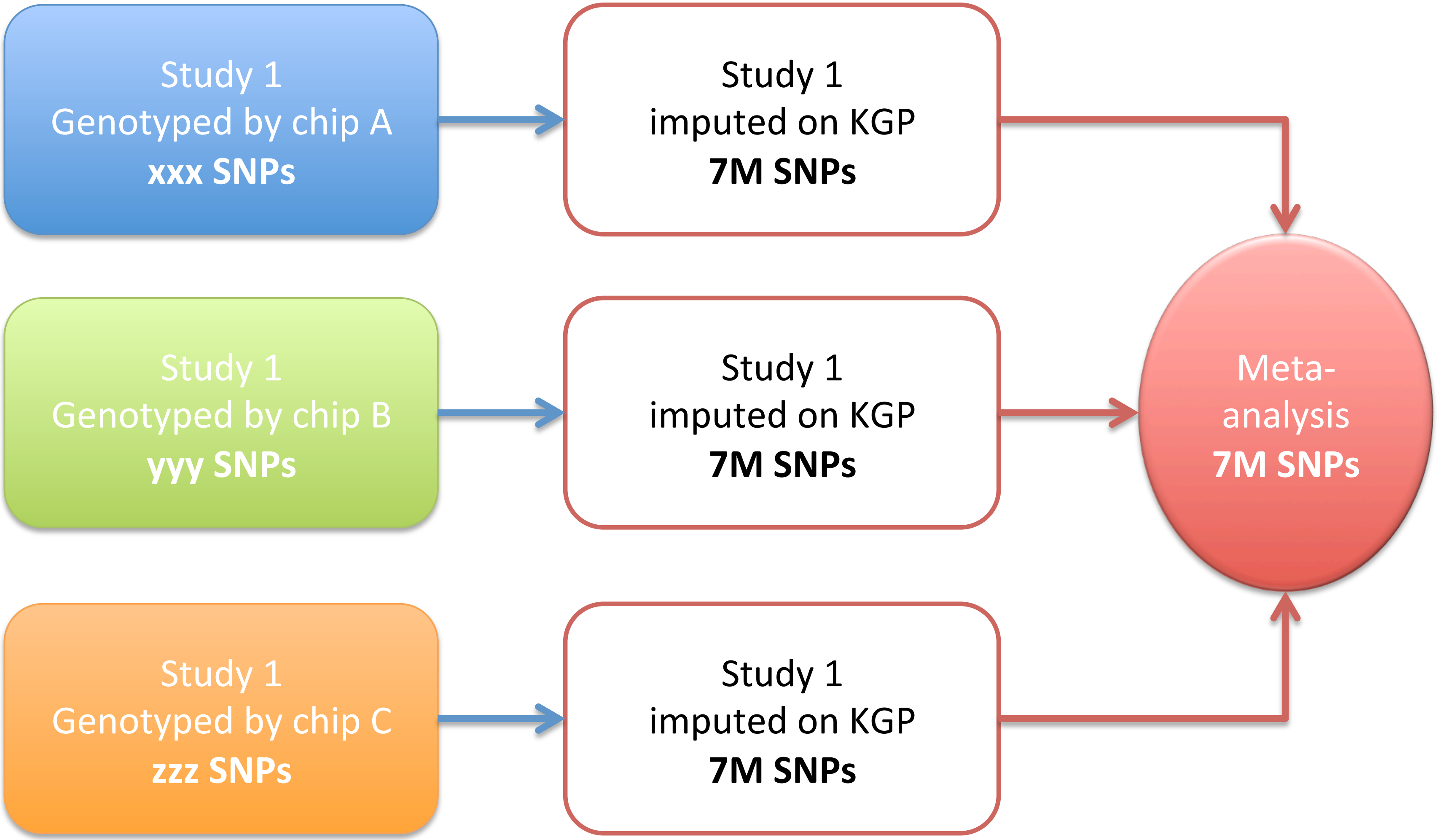
Transition probability

$$Pr(\{Z_{il}^{(1)},Z_{il}^{(2)}\}\rightarrow\{Z_{i(l+1)}^{(1)},Z_{i(l+1)}^{(2)}\}|H)=\begin{cases}\left(e^{-\frac{\rho_l}{N}}+\frac{1-e^{-\frac{\rho_l}{N}}}{N}\right)^2 & Z_{il}^{(1)}=Z_{i(l+1)}^{(1)},Z_{il}^{(2)}=Z_{i(l+1)}^{(2)}\\ \left(e^{-\frac{\rho_l}{N}}+\frac{1-e^{-\frac{\rho_l}{N}}}{N}\right)\left(\frac{1-e^{-\frac{\rho_l}{N}}}{N}\right) & Z_{il}^{(1)}=Z_{i(l+1)}^{(1)},Z_{il}^{(2)}\neq Z_{i(l+1)}^{(2)}\\ & Z_{il}^{(1)}\neq Z_{i(l+1)}^{(1)},Z_{il}^{(2)}=Z_{i(l+1)}^{(2)}\\ \left(\frac{1-e^{-\frac{\rho_l}{N}}}{N}\right)^2 & Z_{il}^{(1)}\neq Z_{i(l+1)}^{(1)},Z_{il}^{(2)}\neq Z_{i(l+1)}^{(2)}\end{cases}\quad(3)$$

Emission probability

		G_{il}		
		0	1	2
	0	$(1-\lambda)^2$	$2\lambda(1-\lambda)$	λ^2
$H_{Z_{il}^{(1)}l}+H_{Z_{il}^{(2)}l}$	1	$\lambda(1-\lambda)$	$\lambda^2+(1-\lambda)^2$	$\lambda(1-\lambda)$
	2	λ^2	$2\lambda(1-\lambda)$	$(1-\lambda)^2$

Application of imputation



Study 1
Genotyped by chip A
xxx SNPs

Study 1
imputed on KGP
7M SNPs

Study 1
Genotyped by chip B
yyy SNPs

Study 1
imputed on KGP
7M SNPs

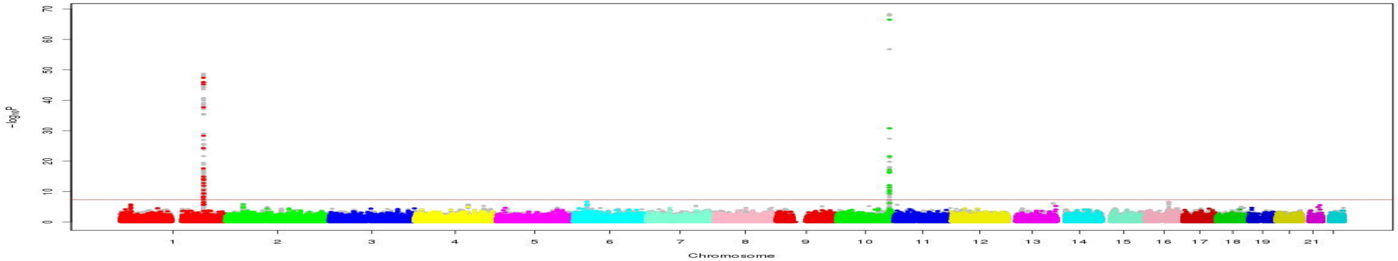
Study 1
Genotyped by chip C
zzz SNPs

Study 1
imputed on KGP
7M SNPs

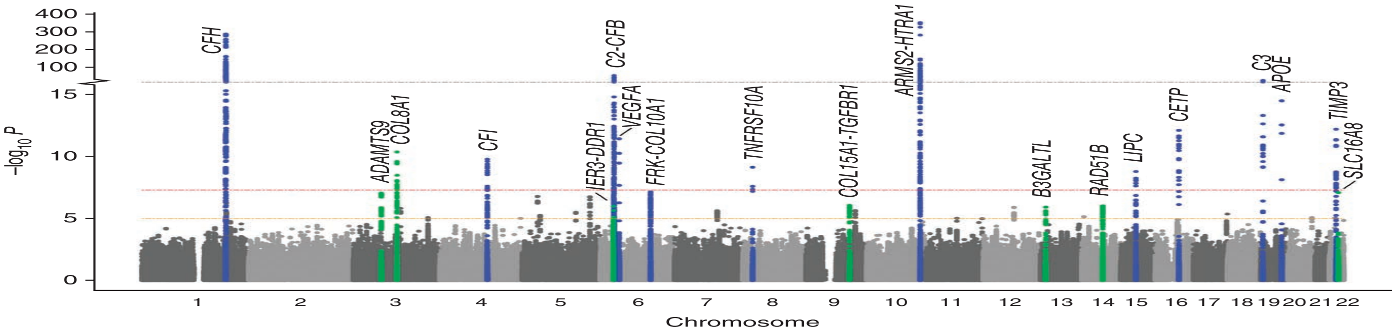
Meta-
analysis
7M SNPs

Meta-analysis of AMD GWAS

Cases ~ 1,000: Controls ~ 4,000 > 2 GW signals



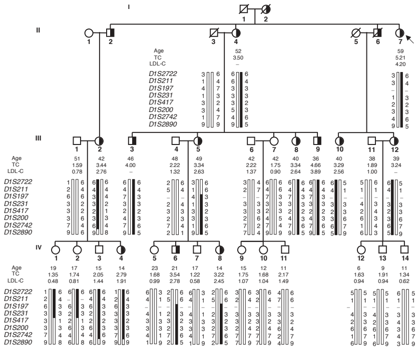
Cases ~ 17,000: Controls ~ 60,000 > 19 GW signals



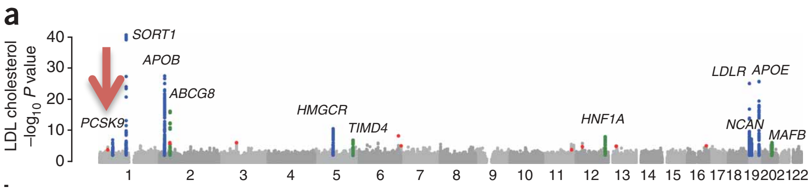
(The AMD Gene Consortium. Nat Genet 2013;45:433-9)

Genome-wide meta-analysis can increase statistical power, and enables us to identify tens of susceptible loci for a disease trait or a quantitative trait.

Creating drugs using gene mapping result

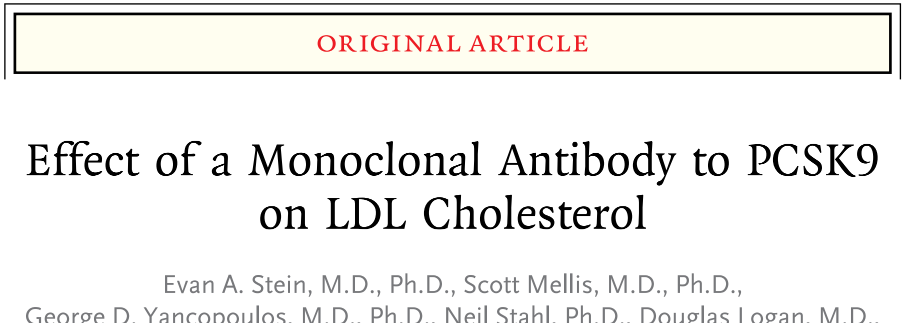


Linkage analysis and positional cloning identified ***PCSK9*** as a novel causative locus for autosomal dominant hypercholesterolemia (Abifadel M et al. Nat Genet 2003;34:154)



GWAS confirmed that ***PCSK9*** was also associated with LDL cholesterol level in general population (Global Lipids Genetics Consortium. Nat Genet 2013; 45: 1274, and several other reports.)

Functional role of **PCSK9 protein** was revealed...



ORIGINAL ARTICLE

Effect of a Monoclonal Antibody to PCSK9 on LDL Cholesterol

Evan A. Stein, M.D., Ph.D., Scott Mellis, M.D., Ph.D.,
George D. Yancopoulos, M.D., Ph.D., Neil Stahl, Ph.D., Douglas Logan, M.D.,

And a new drug lowering LDL cholesterol is going to be approved

Estimation of heritability (GWAS)

- By using “genome-wide significant” SNPs,

$$h^2 = \sum_i 2f_i(1 - f_i)\beta_i^2$$

calculates aggregate contribution of significant SNPs under additive genetic model, when values {0,1,2} are given to each biallelic genotype (for example, A/A, A/a, and a/a) .β_i is an effect size at locus i.

- This should be equal to “narrow-sense heritability”

Estimation of heritability

- Polygenic model

$$P = G + E$$

$$h^2 = \frac{V_G}{V_P}$$

G: genetic effect
E: residuals (supposed to be environmental effect)
P: phenotypic value

Polygenic model

$$P = A + E$$

$$P = A + D + E$$

$$P = A + D + AA + E$$

$$P = A + D + AA + AD + E$$

$$P = A + D + AA + AD + AAA + E$$

Polygenic model

- Narrow-sense heritability

$$h^2 = \frac{V_A}{V_P}$$

- Broad-sense heritability

$$H^2 = \frac{V_G}{V_P} = \frac{V_A + V_D + V_{AA} + V_{AD} + V_{AAA} + \cdots}{V_P}$$

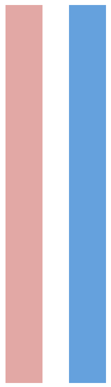
Estimation of heritability (twin study)

- Covariance of twins

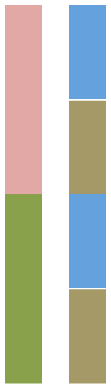
$$Cov_{mz} = V_A + V_{C,mz}$$

$$Cov_{dz} = \frac{1}{2}V_A + V_{C,dz}$$

Monozygotic twins



Dizygotic twins



Estimation of heritability (twin study)

- Covariance of twins

$$Cov_{mz} = V_A + V_{C,mz}$$

$$Cov_{dz} = \frac{1}{2} V_A + V_{C,dz}$$

$$\Downarrow$$

$$2(r_{mz} - r_{dz}) = \frac{V_A}{V_P} = h^2$$

Estimation of heritability (twin study)

- Covariance of twins under the existence of dominance and epistasis effects

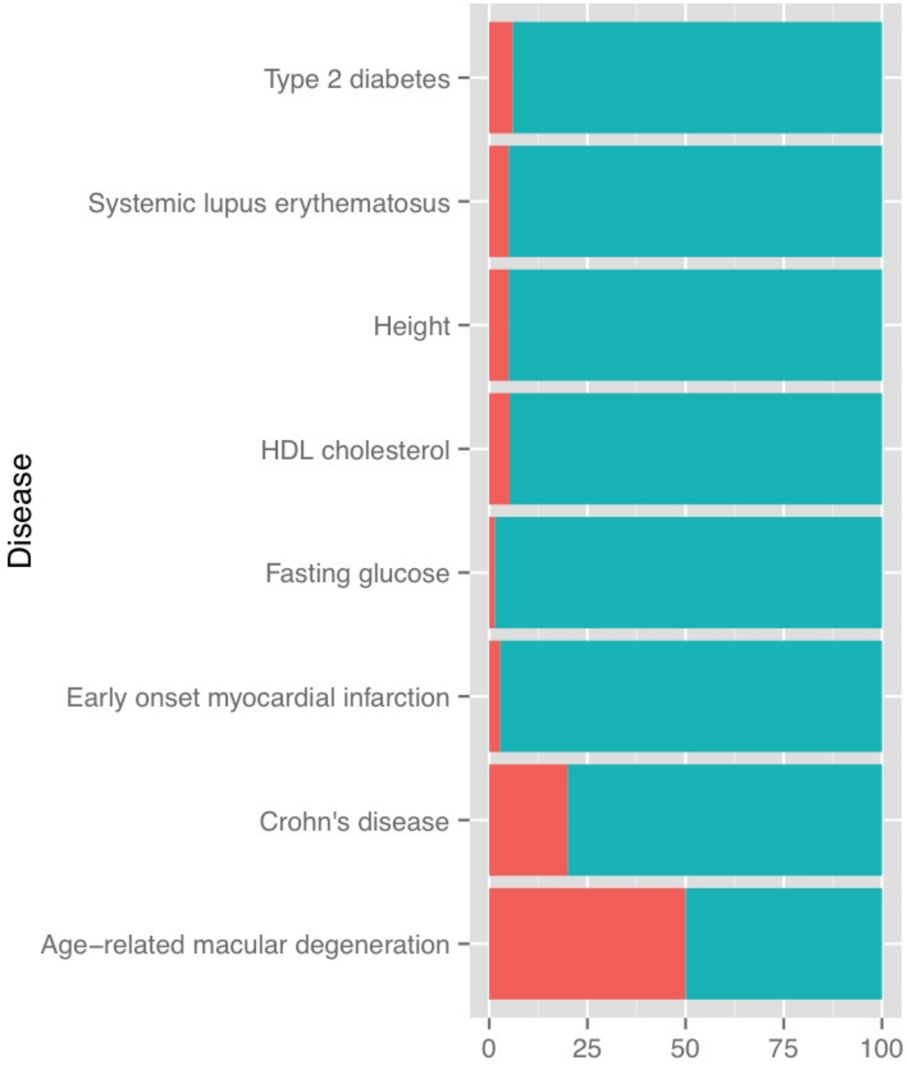
$$Cov_{mz} = V_A + V_D + V_{AA} + V_{AD} + V_{AAA} + \cdots + V_{C,mz}$$

$$Cov_{dz} = \frac{1}{2}V_A + \frac{1}{4}V_D + \frac{1}{4}V_{AA} + \frac{1}{8}V_{AD} + \frac{1}{8}V_{AAA} + \cdots + V_{C,dz}$$

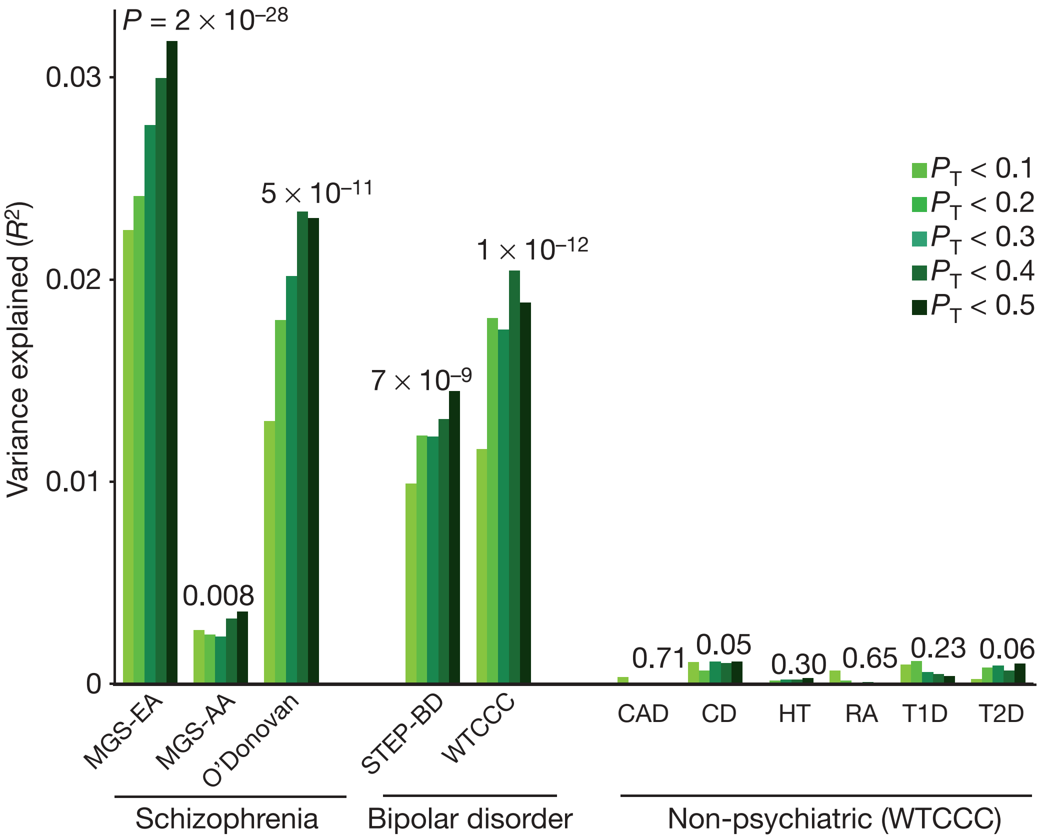
$$2(r_{mz} - r_{dz}) = \frac{V_A + \frac{3}{2}V_D + \frac{3}{2}V_{AA} + \frac{7}{4}V_{AD} + \frac{7}{4}V_{AAA} + \cdots}{V_P} > h^2$$

Missing heritability

The explained variance using genome-wide significant loci (red bars) are much smaller than the heritability estimates from twin studies, which are expressed as 100% in the right plot.



Polygenic score analysis



Purcell et al. did not find any GW significant Schizophrenia locus.

But they gave “**polygenic risk score**” to each individual including **more than thousands of genetic variants**, and tried to see predictive value of this.

They showed that polygenic risk scores could predict schizophrenia in an independent sample but not in non-psychiatric diseases.

Most notably they showed similar polygenic background behind schizophrenia and bipolar disorder.

Altogether, these indicate polygenic nature of complex disease genetics.

(The International Schizophrenia Consortium. Nature 2009; 460: 748.)

Estimation of SNP heritability

- Mixed model with total genotypic effects

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{u} + \epsilon \qquad \mathrm{var}(\mathbf{y}) = \mathbf{W}\mathbf{W}'\sigma_u^2 + \mathbf{I}\sigma_\epsilon^2$$

y: phenotype
β: fixed effects (age, sex, ...)
X: covariate values of fixed effect terms
W: standardized genotype matrix
u: SNP effects as random effects $\mathbf{u} \sim N\left(0, \mathbf{I}\sigma_u^2\right)$
$\boldsymbol{\epsilon} \sim N\left(0, \mathbf{I}\sigma_\epsilon^2\right)$

(Jian Yang et al. Nat Genet 2010; 42: 565.)

Estimation of SNP heritability

- Mixed model with total genotypic effects

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{W}\mathbf{u} + \epsilon \qquad \mathrm{var}(\mathbf{y}) = \mathbf{W}\mathbf{W}'\sigma_u^2 + \mathbf{I}\sigma_\epsilon^2$$

- By taking $\mathbf{A} = \mathbf{W}\mathbf{W}'/N$ and $\sigma_g^2 = N\sigma_u^2$

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{g} + \epsilon \qquad \mathrm{var}(\mathbf{y}) = \mathbf{A}\sigma_g^2 + \mathbf{I}\sigma_\epsilon^2$$

(Jian Yang et al. Nat Genet 2010; 42: 565.)

Estimation of SNP heritability

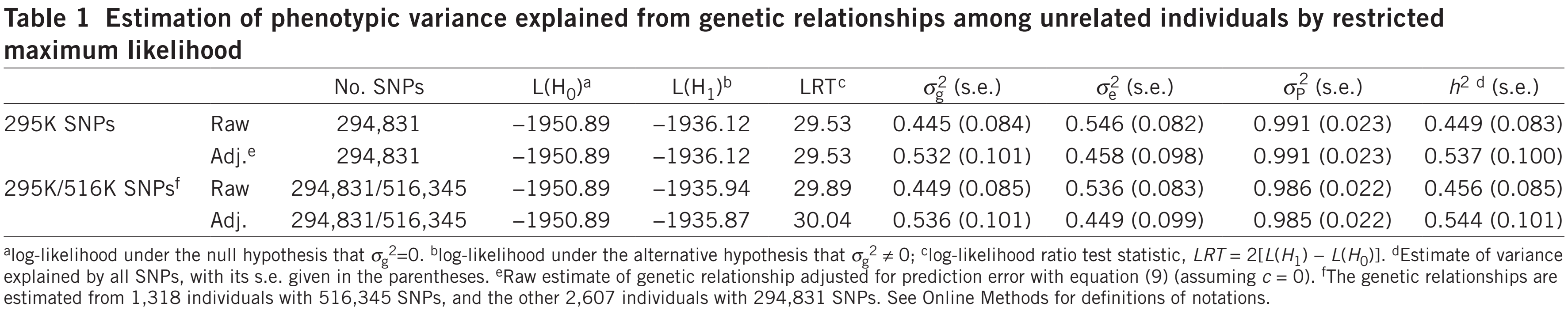


Table 1 Estimation of phenotypic variance explained from genetic relationships among unrelated individuals by restricted maximum likelihood

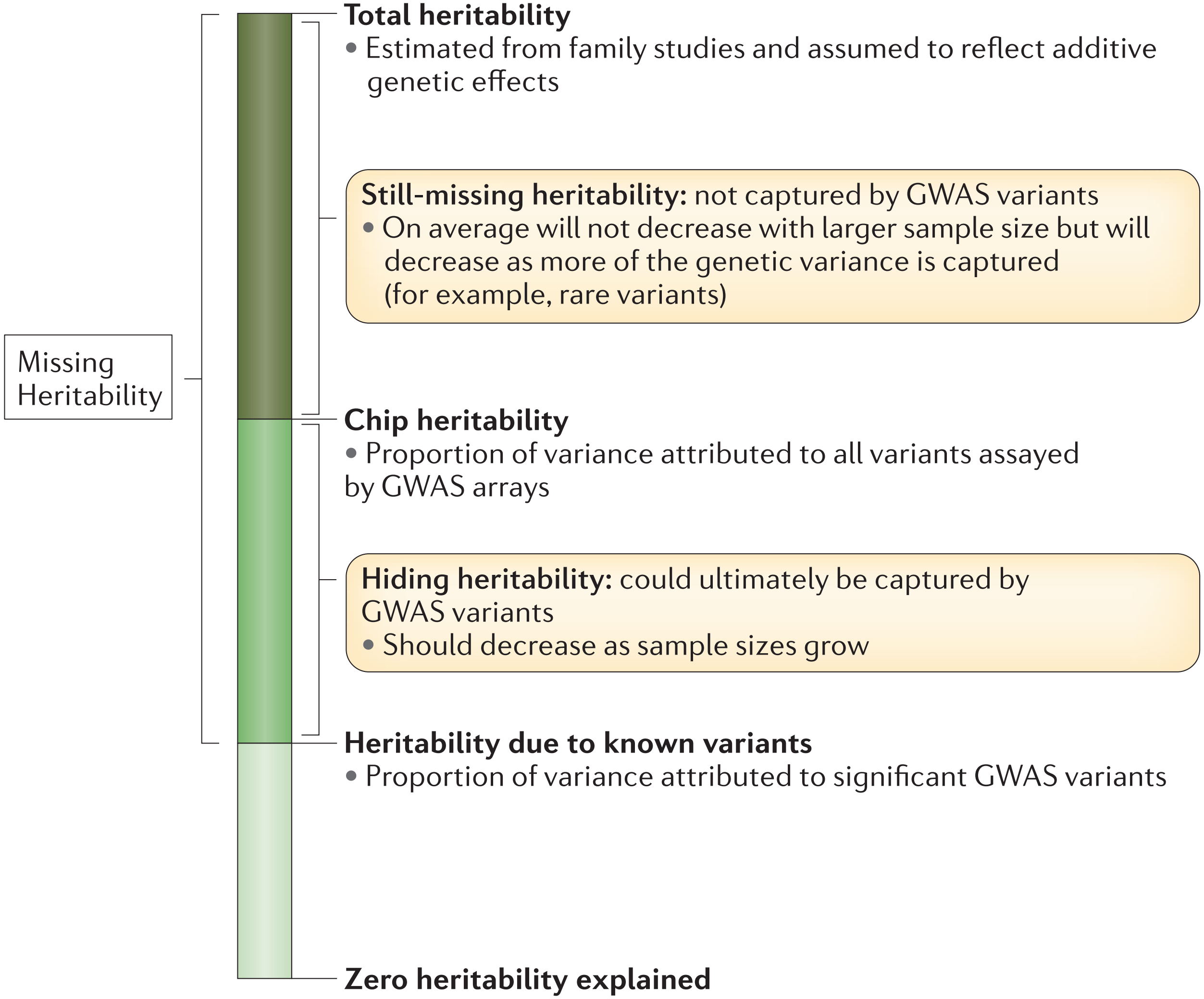
		No. SNPs	$L(H_0)$[a]	$L(H_1)$[b]	LRT[c]	σ_g^2 (s.e.)	σ_e^2 (s.e.)	σ_P^2 (s.e.)	h^2 [d] (s.e.)
295K SNPs	Raw	294,831	−1950.89	−1936.12	29.53	0.445 (0.084)	0.546 (0.082)	0.991 (0.023)	0.449 (0.083)
	Adj.[e]	294,831	−1950.89	−1936.12	29.53	0.532 (0.101)	0.458 (0.098)	0.991 (0.023)	0.537 (0.100)
295K/516K SNPs[f]	Raw	294,831/516,345	−1950.89	−1935.94	29.89	0.449 (0.085)	0.536 (0.083)	0.986 (0.022)	0.456 (0.085)
	Adj.	294,831/516,345	−1950.89	−1935.87	30.04	0.536 (0.101)	0.449 (0.099)	0.985 (0.022)	0.544 (0.101)

[a] log-likelihood under the null hypothesis that $\sigma_g^2=0$. [b] log-likelihood under the alternative hypothesis that $\sigma_g^2 \neq 0$; [c] log-likelihood ratio test statistic, $LRT = 2[L(H_1) - L(H_0)]$. [d] Estimate of variance explained by all SNPs, with its s.e. given in the parentheses. [e] Raw estimate of genetic relationship adjusted for prediction error with equation (9) (assuming $c = 0$). [f] The genetic relationships are estimated from 1,318 individuals with 516,345 SNPs, and the other 2,607 individuals with 294,831 SNPs. See Online Methods for definitions of notations.

- GW significant SNPs can only explain ~ 5% of height variance
- However, all SNPs could explain ~ 45%.
- This implicates that human height would be determined by hundreds or thousands of genetic variants, and most of them have not been discovered because of low statistical power.
- This explained variance is still lower than twin study's heritability (80-90%). It is suggested that "SNPs" act as markers, true causative variants (possibly low frequency) are more informative and may increase explained variance.

(Jian Yang et al. Nat Genet 2010; 42: 565.)

Current understanding of complex disease genetics



Missing Heritability

Total heritability
- Estimated from family studies and assumed to reflect additive genetic effects

Still-missing heritability: not captured by GWAS variants
- On average will not decrease with larger sample size but will decrease as more of the genetic variance is captured (for example, rare variants)

Chip heritability
- Proportion of variance attributed to all variants assayed by GWAS arrays

Hiding heritability: could ultimately be captured by GWAS variants
- Should decrease as sample sizes grow

Heritability due to known variants
- Proportion of variance attributed to significant GWAS variants

Zero heritability explained

(Witte JS et al. Nat Rev Genet 2014; 15: 765.)

Other estimation methods of SNP heritability

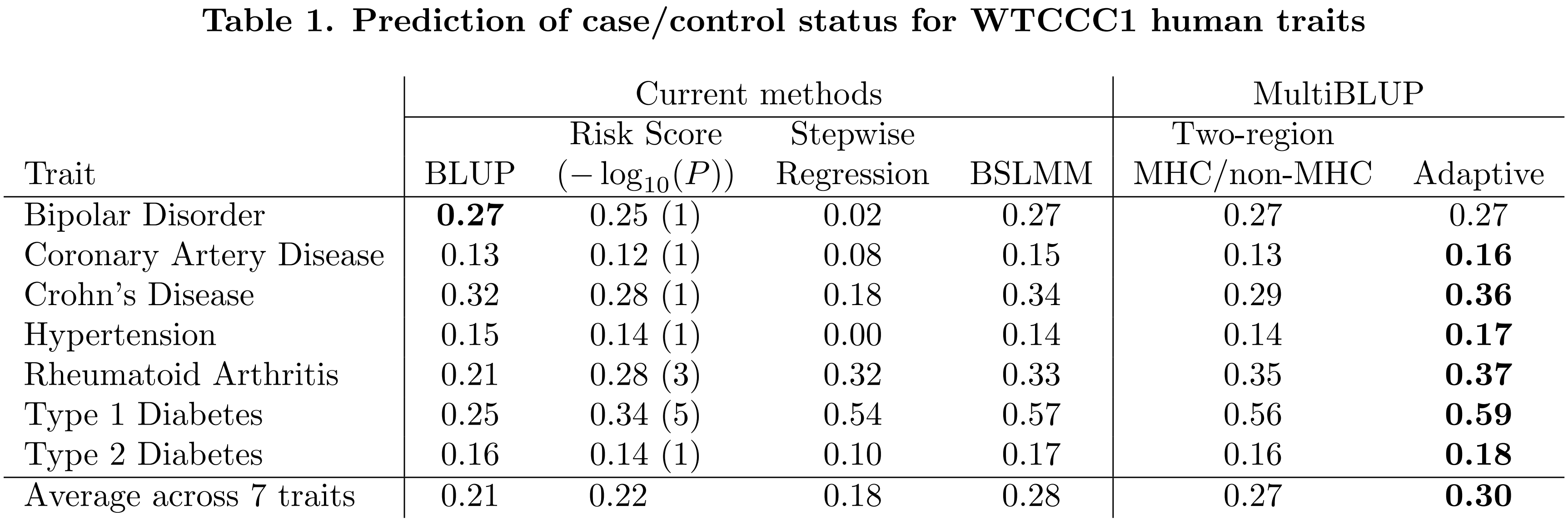


Table 1. Prediction of case/control status for WTCCC1 human traits

	Current methods				MultiBLUP	
		Risk Score	Stepwise		Two-region	
Trait	BLUP	$(-\log_{10}(P))$	Regression	BSLMM	MHC/non-MHC	Adaptive
Bipolar Disorder	**0.27**	0.25 (1)	0.02	0.27	0.27	0.27
Coronary Artery Disease	0.13	0.12 (1)	0.08	0.15	0.13	**0.16**
Crohn's Disease	0.32	0.28 (1)	0.18	0.34	0.29	**0.36**
Hypertension	0.15	0.14 (1)	0.00	0.14	0.14	**0.17**
Rheumatoid Arthritis	0.21	0.28 (3)	0.32	0.33	0.35	**0.37**
Type 1 Diabetes	0.25	0.34 (5)	0.54	0.57	0.56	**0.59**
Type 2 Diabetes	0.16	0.14 (1)	0.10	0.17	0.16	**0.18**
Average across 7 traits	0.21	0.22	0.18	0.28	0.27	**0.30**

(Speed D and Balding DJ. Genome Res 2014 published in advance.)

Current targets ...

- **Larger and larger GWAS**: to capture common variants with small effect sizes

- **Low frequency variants, structural variants**: some of them have not been captured by SNP array

- **Heritable epigenetic marks**: data not obtained by SNP array, but the existence of parent-of origin effect indicates its involvement

- **Epistasis (gene-gene interaction)**: could show heritability beyond additive effects. A few analyses succeeded to identify it, but not enough

- **Gene-environmental interaction**: sophisticated epidemiological sample would be necessary, and statistical geneticists typically do not have it

Closing remarks

- We are analyzing BioBankJapan samples; ~200,000 disease samples from 47 diseases and ~ 30,000 population controls, all of them are Japanese and have ~ 1,000,000 SNP genotype results.

- Our main aim at now is to find out low-frequency variants by combining this data with Whole Genome Sequencing results.

- We are welcome to collaborate with researchers who want to use our "big" data and apply statistically sophisticated analysis!