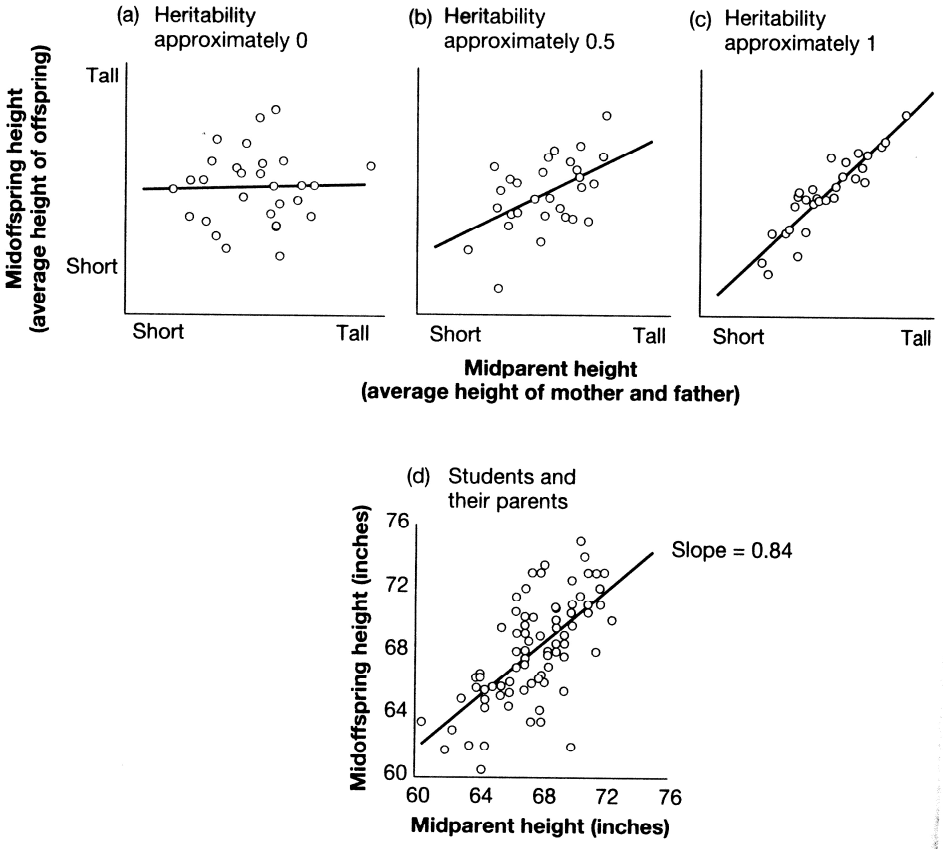
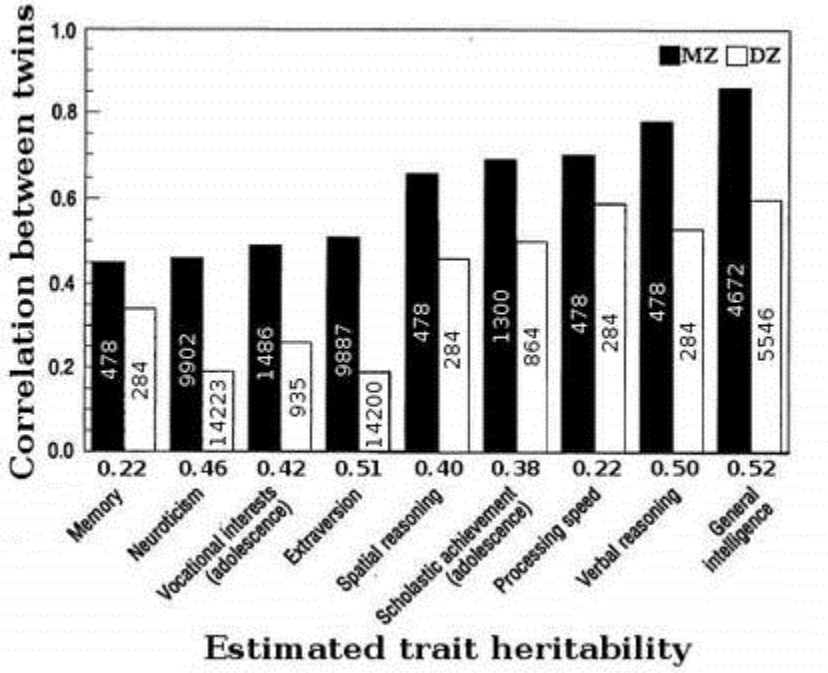
Gene-set analyses of genome-wide association studies

Sangsoo Kim
Dept. of Bioinformatics
Soongsil University, Seoul, KOREA

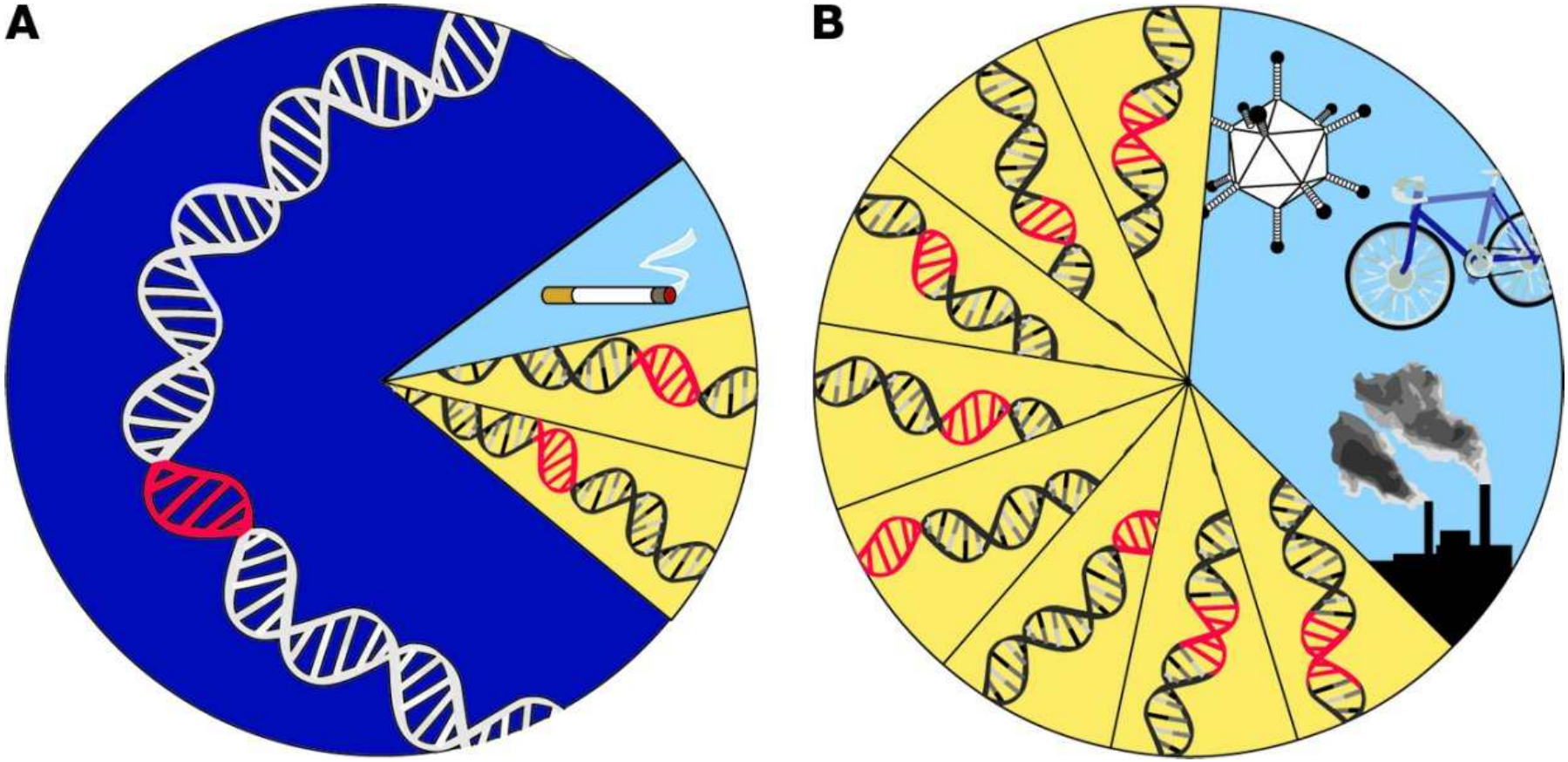
Phenotype (P) = Genotype (G) + Environment (E)
Var(P) = Var(G) + Var(E) + 2 Cov(G,E)

Heritability, Var(G)/Var(P), is often estimated from family data



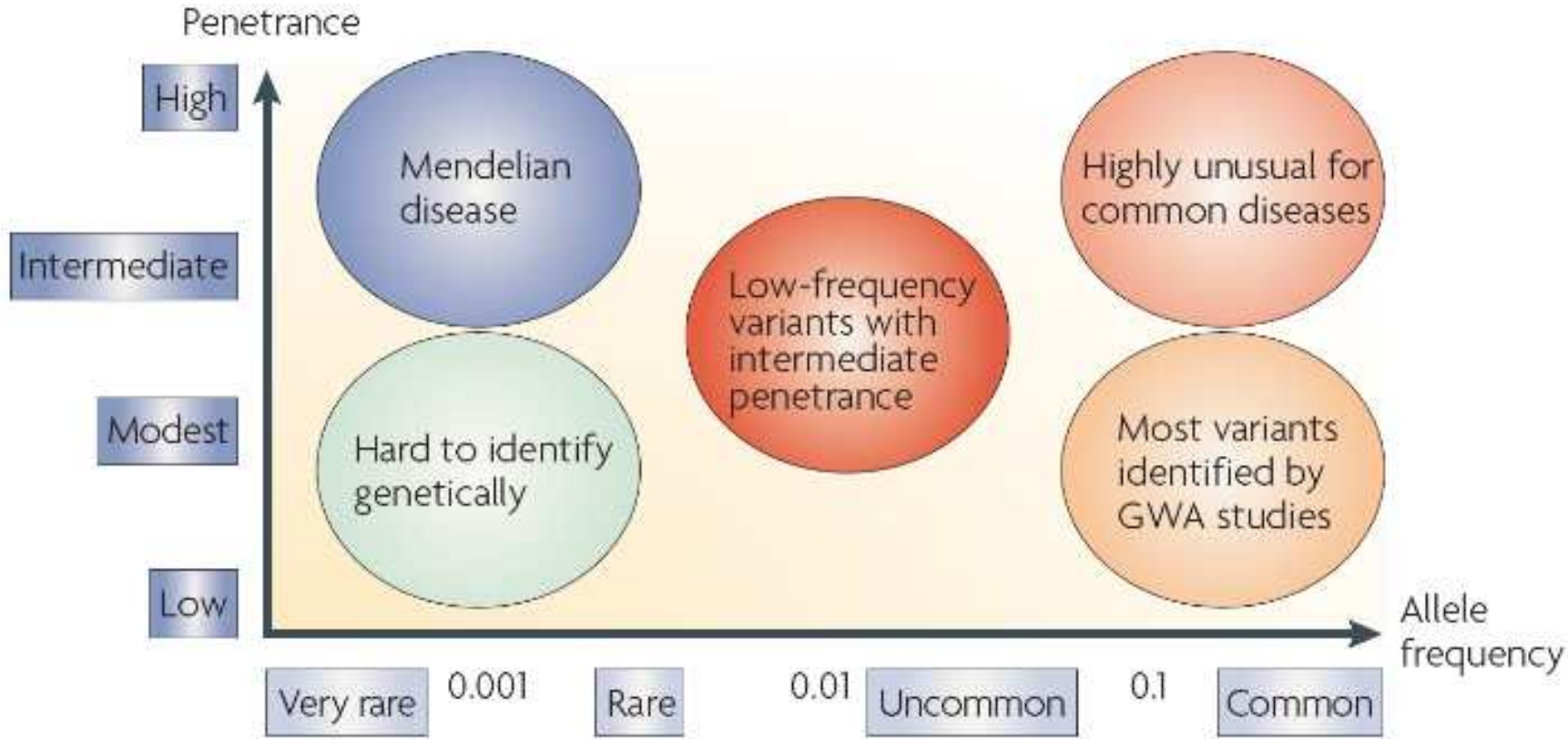


Genetic and environmental contributions to (A) monogenic and (B) complex disorders



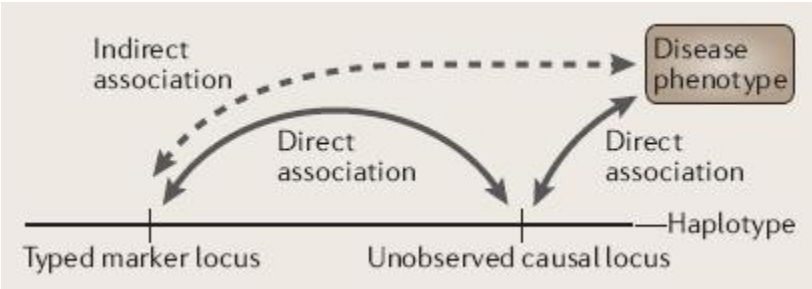
A HapMap harvest of insights into the genetics of common disease
J. Clin. Invest. 118:5 doi:10.1172/JCI34772

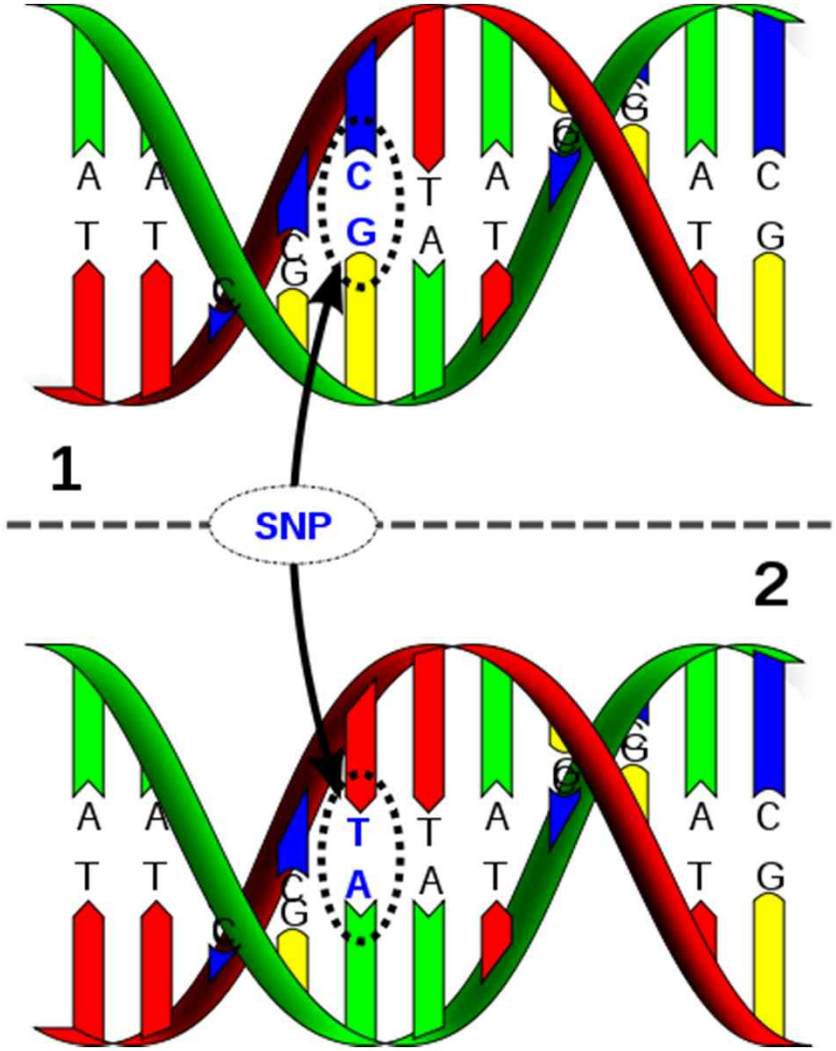
Low-frequency variants and disease susceptibility



Single Nucleotide Polymorphisms

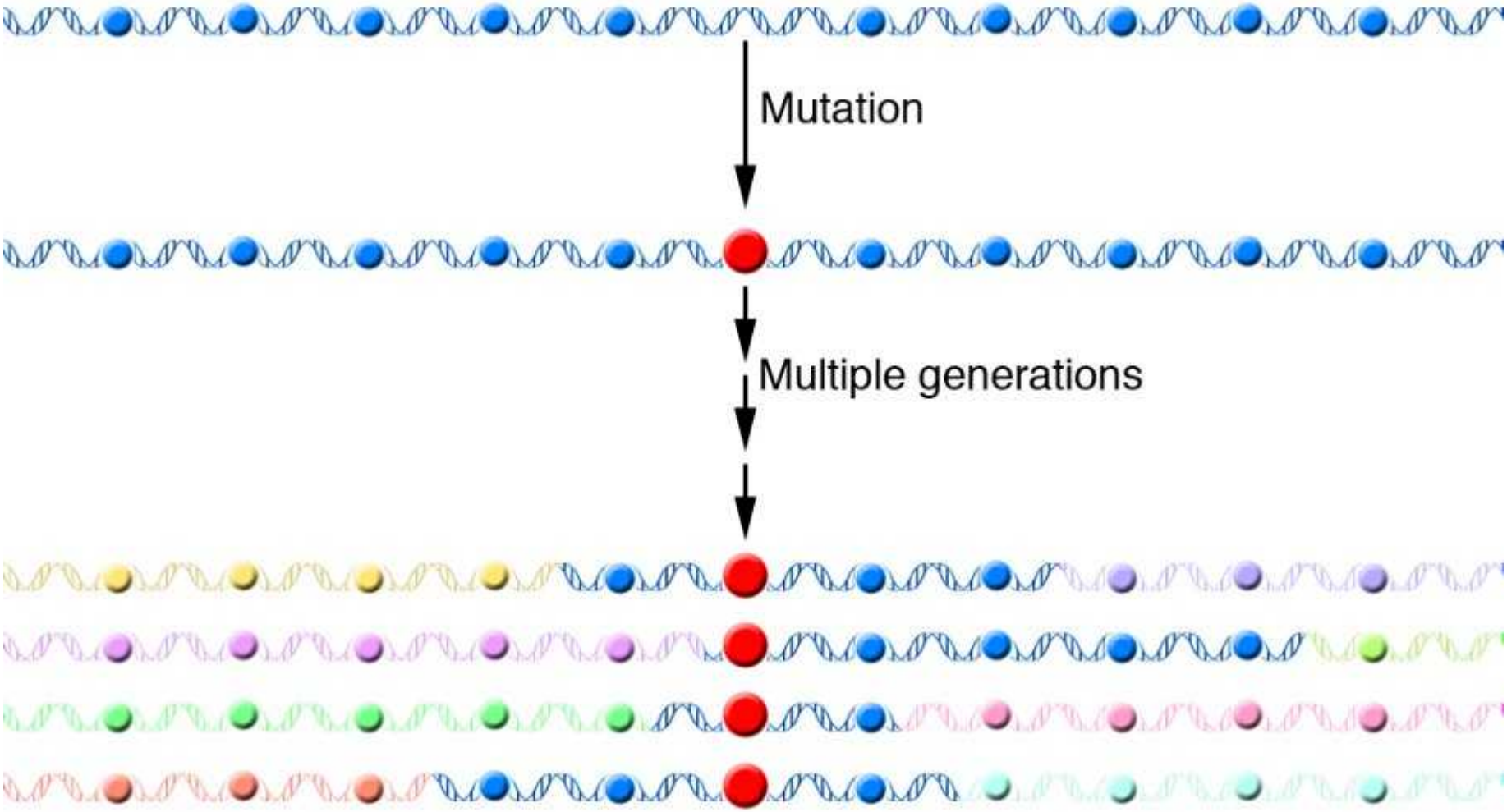
- Imply common variations (minor allele frequency >1%)
- ~18 million RefSNPs in dbSNP (Build 130)
  - 9.5 million validated
- Most dense genetic marker
- Useful in mapping diseases
  - Directly
  - Indirectly



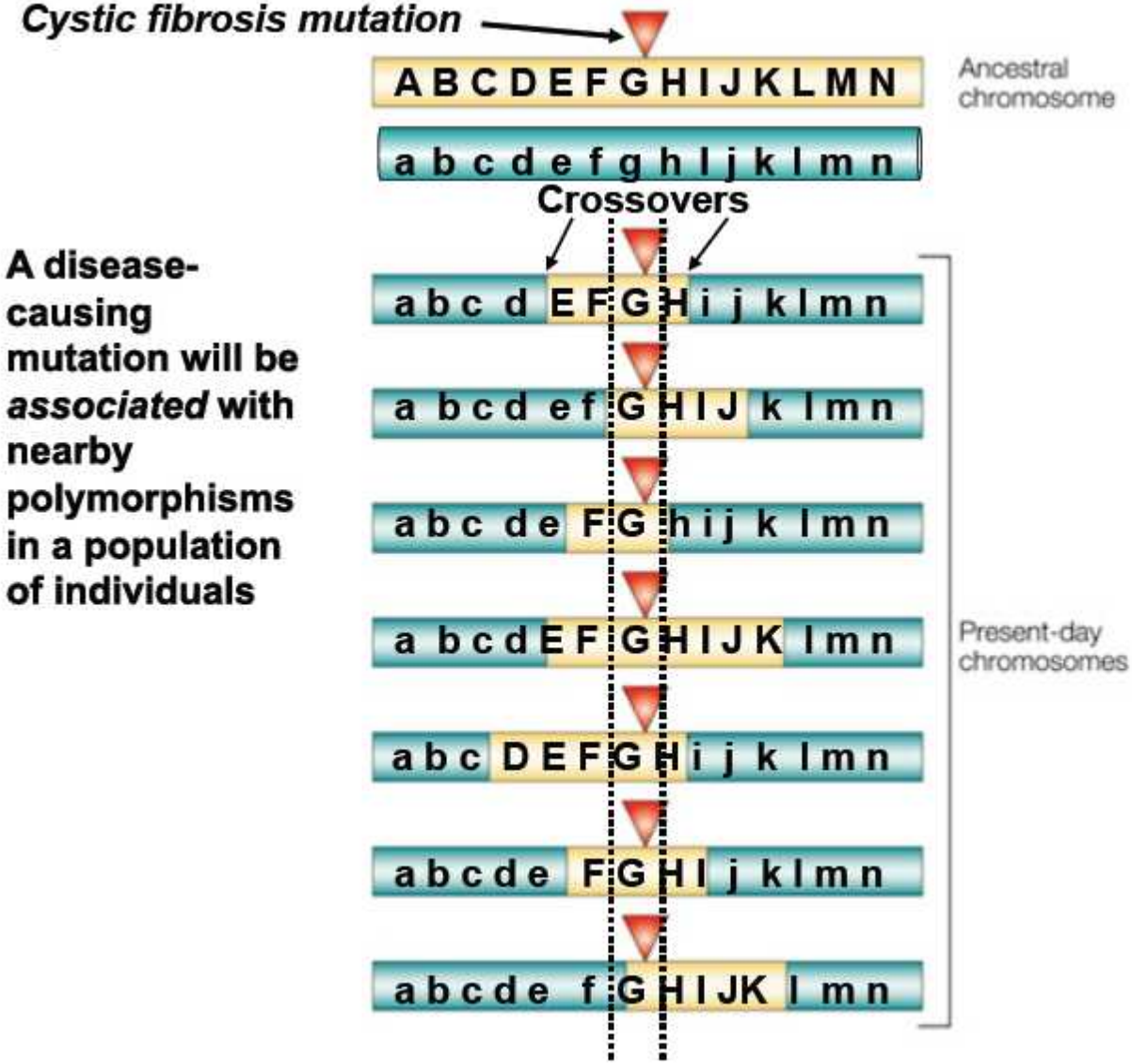


http://en.wikipedia.org/wiki/Single-nucleotide_polymorphism

Breakdown of LD around a new SNP



A HapMap harvest of insights into the genetics of common disease
J. Clin. Invest. 118:5 doi:10.1172/JCI34772


Cystic fibrosis mutation
A B C D E F G H I J K L M N
Ancestral chromosome
a b c d e f g h I j k l m n
Crossovers
A disease-causing mutation will be *associated* with nearby polymorphisms in a population of individuals
a b c d E F G H i j k l m n
a b c d e f G H I J k l m n
a b c d e F G h i j k l m n
a b c d E F G H I J K l m n
a b c D E F G H i j k l m n
a b c d e F G H I j k l m n
a b c d e f G H I JK l m n
Present-day chromosomes

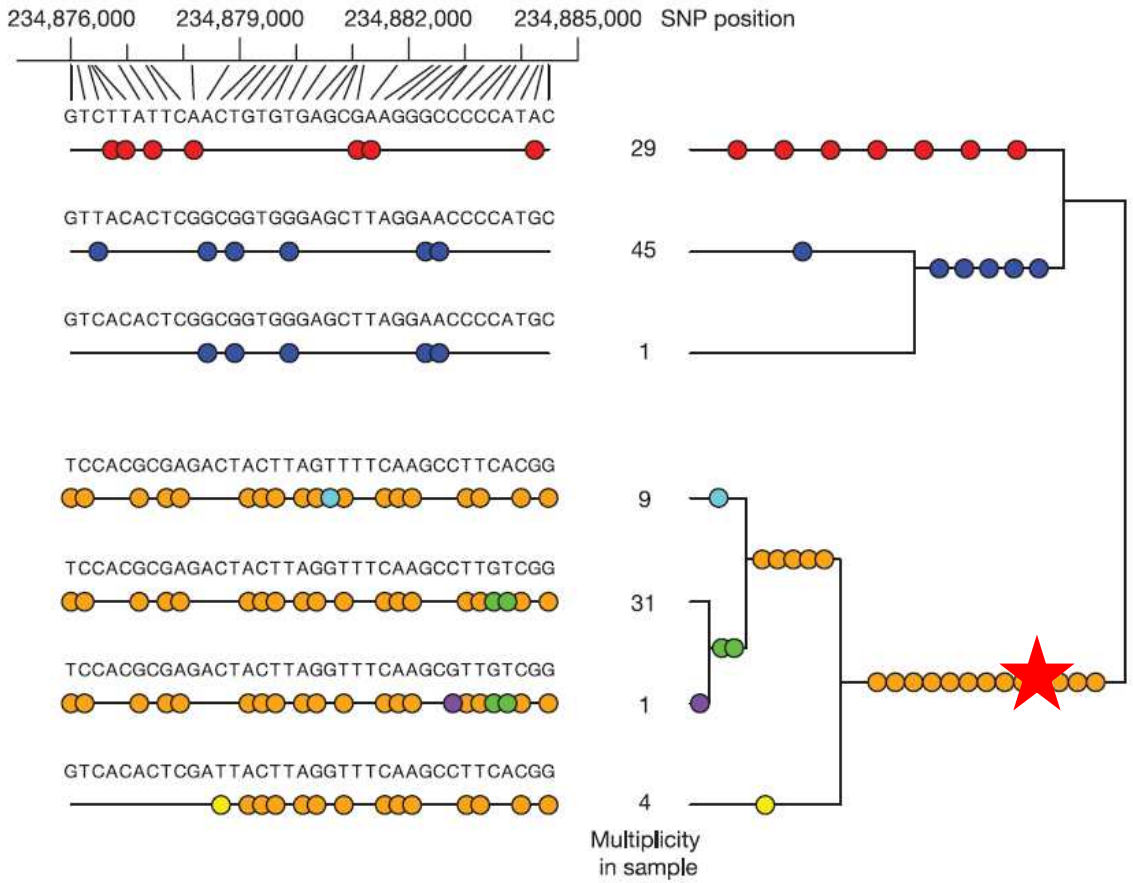
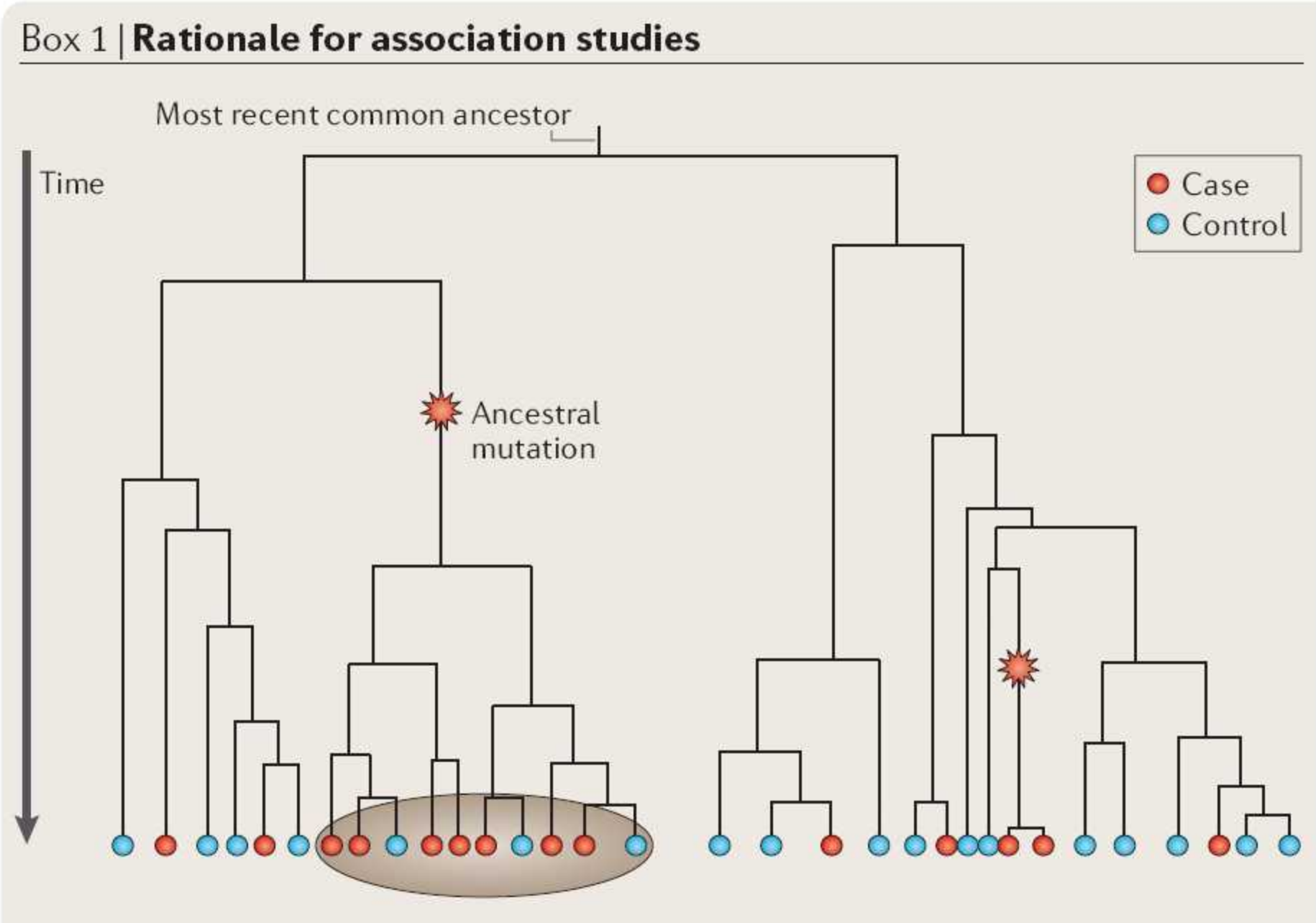


Figure 7 | Genealogical relationships among haplotypes and r^2 values in a region without obligate recombination events. The region of chromosome 2 (234,876,004–234,884,481 bp; NCBI build 34) within ENr131.2q37 contains 36 SNPs, with zero obligate recombination events in the CEU samples. The left part of the plot shows the seven different haplotypes observed over this region (alleles are indicated only at SNPs), with their respective counts in the data. Underneath each of these haplotypes is a binary representation of the same data, with coloured circles at SNP positions where a haplotype has the less common allele at that site. Groups of SNPs all captured by a single tag SNP (with $r^2 \geq 0.8$) using a pairwise tagging algorithm[53,54] have the same colour. Seven tag SNPs corresponding to the seven different colours capture all the SNPs in this region. On the right these SNPs are mapped to the genealogical tree relating the seven haplotypes for the data in this region.



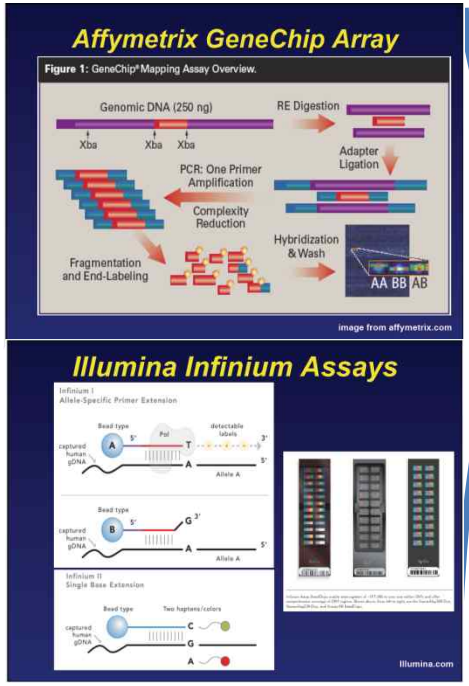
A tutorial on statistical methods for population association studies
NATURE REVIEWS GENETICS VOLUME 7 | OCTOBER 2006 | 781

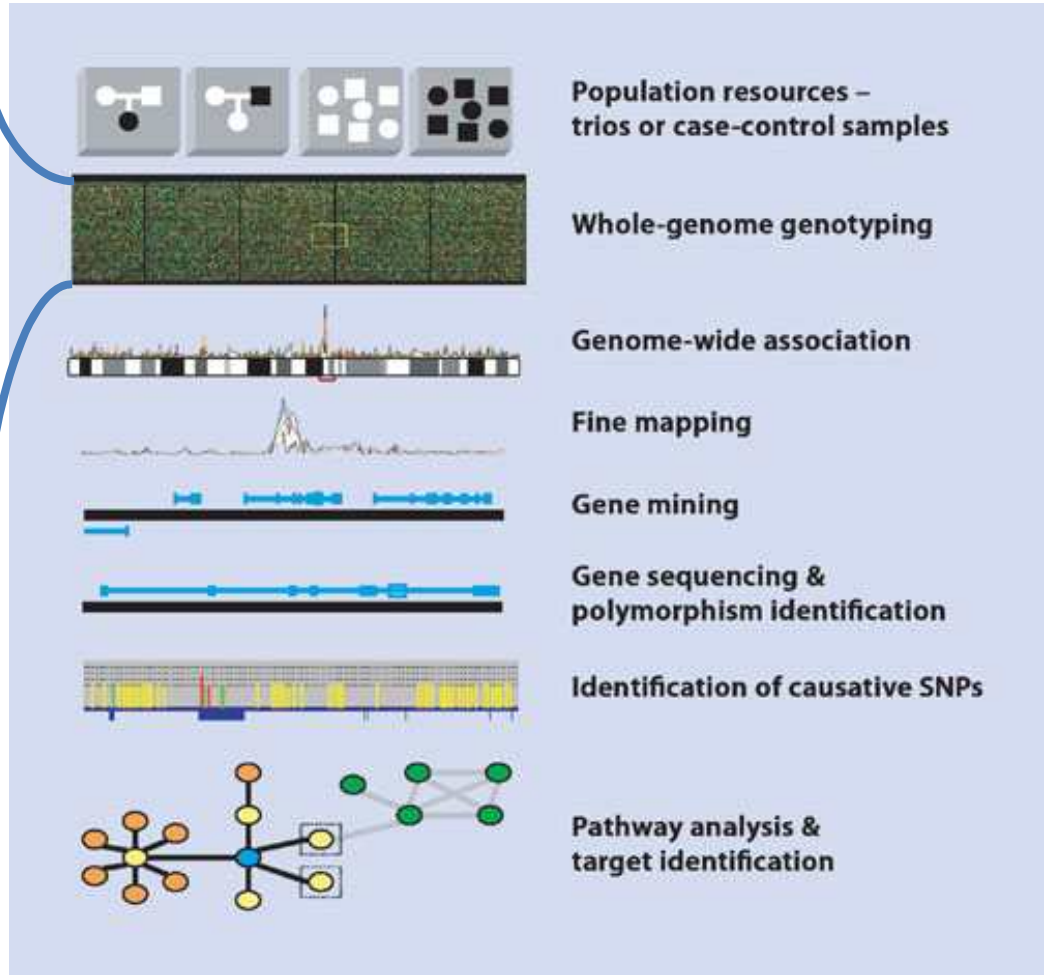
What is GWAS?

Genome-wide Association Study

- An examination of genetic variation across a given genome
- Designed to identify genetic associations with observable traits
  - Such as blood pressure or weight,
  - or why some people get a disease or condition
- Hypothesis-free approach
  - Candidate gene approach

Overview of GWAS

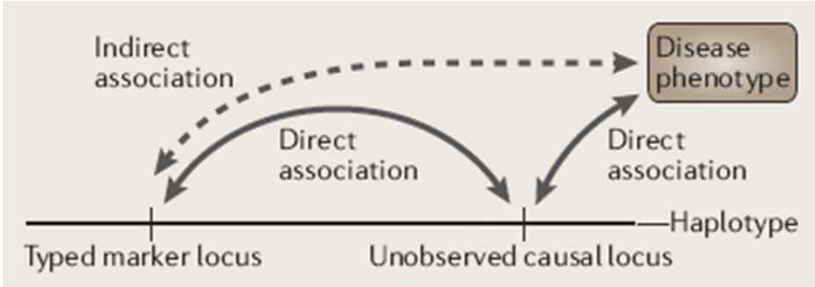




http://www.genengnews.com/gen-articles/human-genome-wide-association-studies/1970/

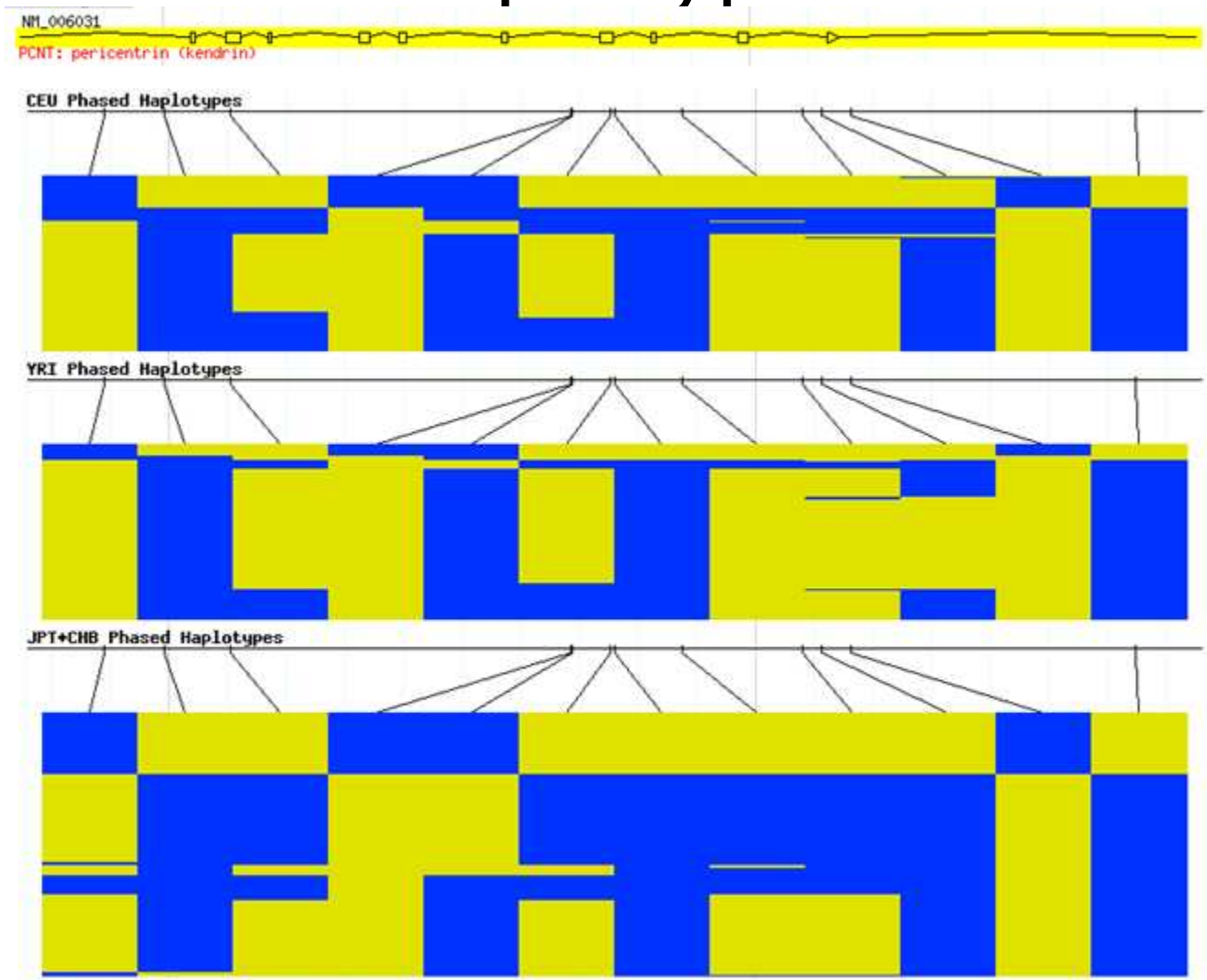
Assumptions in GWAS

- Bi-allelic SNPs
- Common ancestors (human effective population sizes are small)
- Linkage disequilibrium and haplotypes



- Common disease-common variant
  - disease-predisposing might have been advantageous in the past
  - selection pressure is weak on late-onset diseases and on variants that contribute only a small risk

Linkage Disequilibrium & Haplotypes



International HapMap Project

The International HapMap Project is a partnership of scientists and funding agencies from Canada, China, Japan, Nigeria, the United Kingdom and the United States to develop that will help researchers find genes associated with human disease and response to pharmaceuticals. See "**About the International HapMap Project**" for more information.

Project Information

About the Project
HapMap Publications
HapMap Tutorial
HapMap Mailing List
HapMap Project Participants
HapMap Mirror Site in Japan

Project Data

HapMap Genome Browser (Phase 1, 2 & 3 - merged genotypes & frequencies)
HapMap Genome Browser (Phase 3 - genotypes, frequencies & LD)
HapMap Genome Browser (Phase 1 & 2 - full dataset)
GWAs Karyogram
HapMart
Bulk Data Download
Data Freezes for Publication
ENCODE Project
Guidelines For Data Use

Useful Links

TSC SNP Downloads
HapMap Samples at Coriell Institute

News

- 2009-04-02: **HapMap3 CEL files available**

  Raw signal intensity data from HapMap3 genotypes on the Genome-Wide Human SNP Array 6.0 are now **available fo**

- 2009-02-09: **HapMap3 Phased Haplotypes available**

  Phased haplotypes for consensus HapMap3 release 2 data has been phased for autosomes are now **available for bul**

- 2009-02-06: **HapMap Public Release #27 (merged II+III)**

  Genotypes and frequency data for the three phases of the project (I+II: rel #24 and III: release #2), were combined in NC (dbSNP b126) coordinates. Data is **available for downloading** and also **available for browsing**. Click here to read **notes**.

- 2009-01-07: **HapMap Phase 3 draft 2 release available for download**

  Genotypes and frequency data for phase 3 (NCBI build 36, dbSNP b126) of the HapMap are **available for bulk downl** will subsequently be merged with phase I+II data, and once merged, the complete dataset will be made available in the browser and HapMart utility. Here are some **notes and SNP counts** for this dataset.

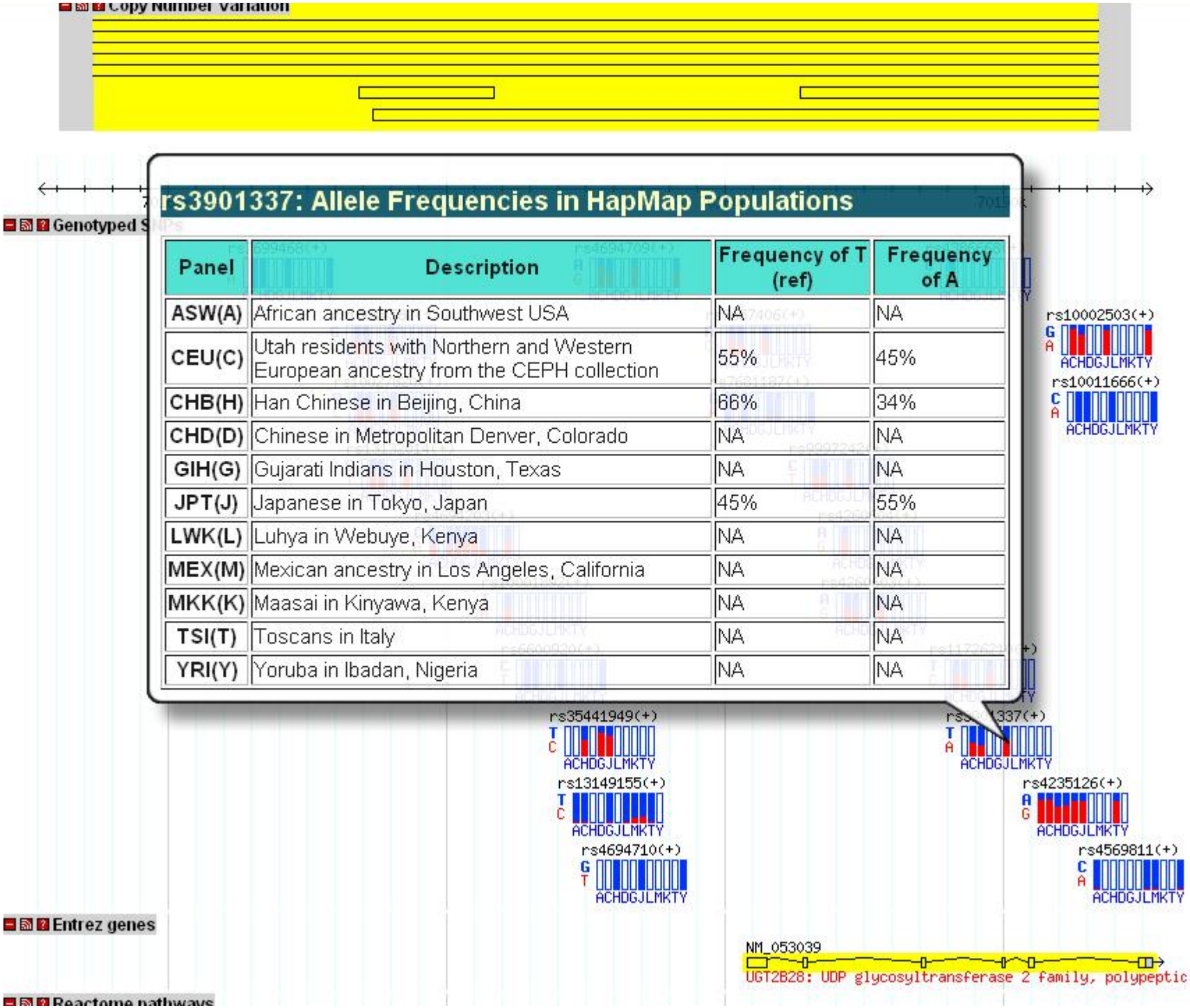
- 2008-11-26: **HapMap Public Release #26 (merged II+III)**

  Genotypes and frequency data for phases I+II (rel #24) and III (draft #1) of the project is available in NCBI build 36 (dbSN coordinates. Data is now **available for downloading** and also available for browsing. Click here to read this **release** This release is no longer available for browsing, instead please use the latest merged data **release #27**.]

- 2008-11-26: **HapMap Public Release #24 on genome browser**

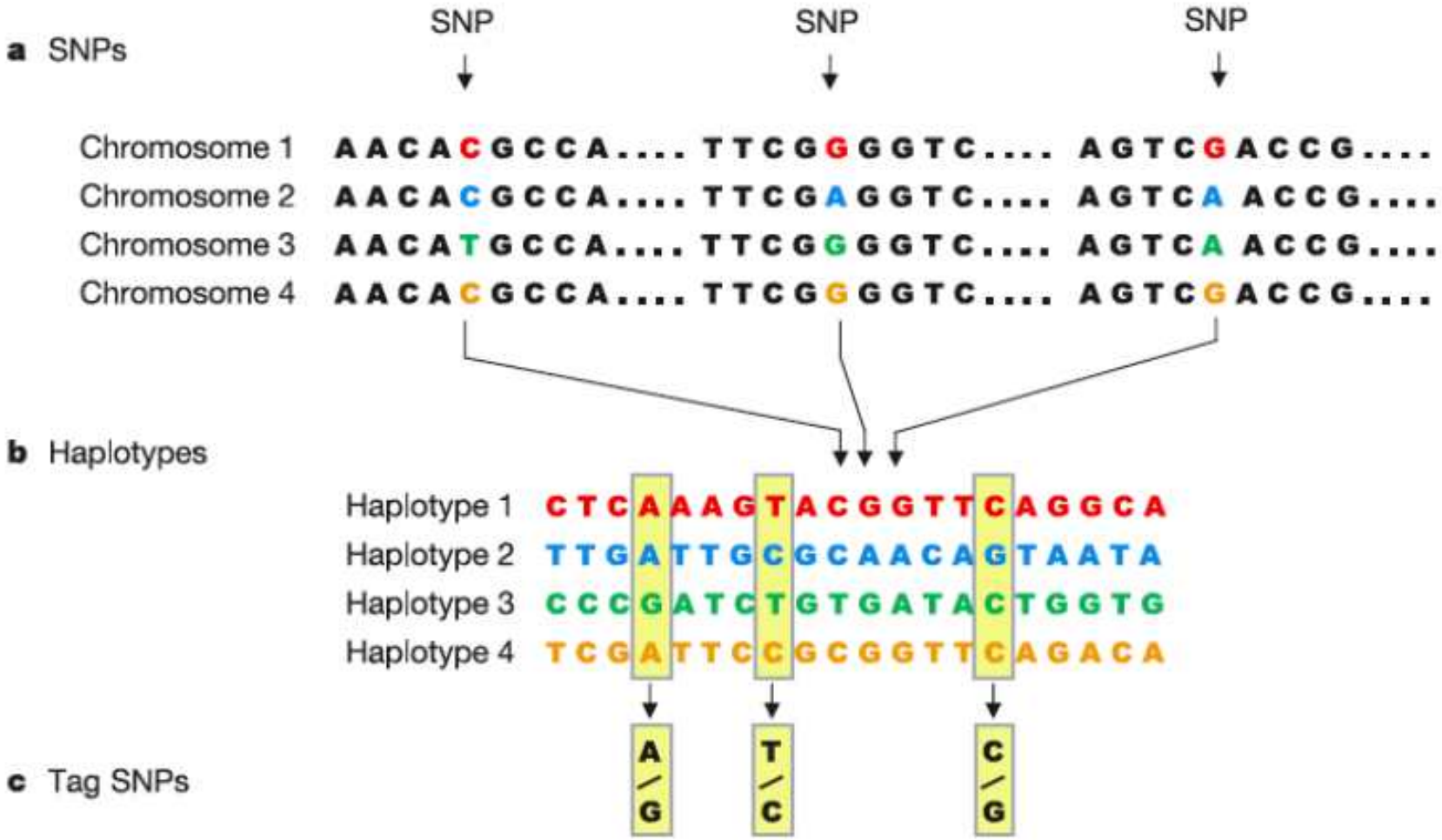
  Data for phases I+II of the project is now **available for browsing** in NCBI build 36 (dbSNP b126) coordinates.

- 2008-10-12: **HapMap Public Release #24 (phase II) available for download**



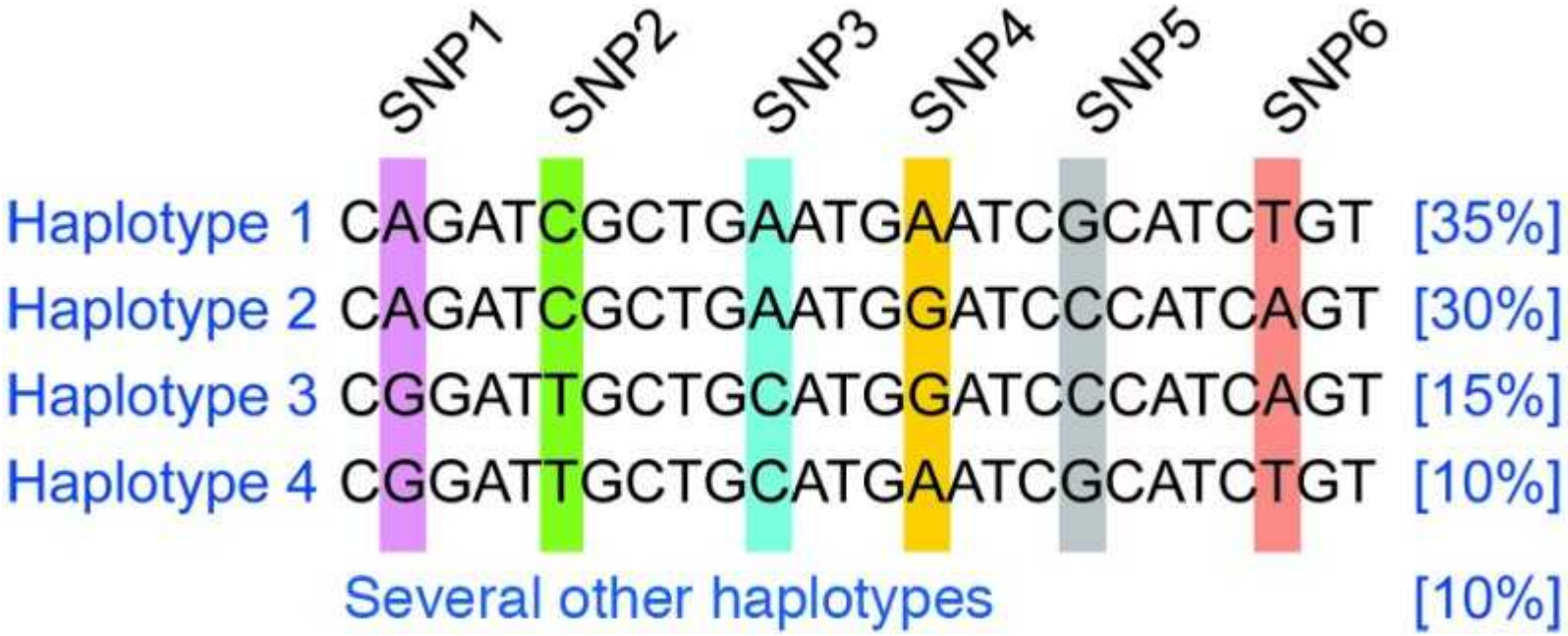
Panel	Description	Frequency of T (ref)	Frequency of A
ASW(A)	African ancestry in Southwest USA	NA	NA
CEU(C)	Utah residents with Northern and Western European ancestry from the CEPH collection	55%	45%
CHB(H)	Han Chinese in Beijing, China	66%	34%
CHD(D)	Chinese in Metropolitan Denver, Colorado	NA	NA
GIH(G)	Gujarati Indians in Houston, Texas	NA	NA
JPT(J)	Japanese in Tokyo, Japan	45%	55%
LWK(L)	Luhya in Webuye, Kenya	NA	NA
MEX(M)	Mexican ancestry in Los Angeles, California	NA	NA
MKK(K)	Maasai in Kinyawa, Kenya	NA	NA
TSI(T)	Toscans in Italy	NA	NA
YRI(Y)	Yoruba in Ibadan, Nigeria	NA	NA

Selecting 'haplotype tag' SNPs



International HapMap Consortium (2003) Nature 426:789

Tag SNPs can define common haplotypes



A HapMap harvest of insights into the genetics of common disease
J. Clin. Invest. 118:5 doi:10.1172/JCI34772

Samples

- Matched case-control samples on age, sex, demographics
- Case: more severely affected individuals
- Control: low risk of disease, rather than population-based samples
- Common population structure
  - Population stratification

Statistical tests

- Case-control
  - Allelic chisq test
  - Cochran-Armitage trend test
  - Logistic regression (http://www.well.ox.ac.uk/rmott/LECTURES/LOGISTIC_REGRESSION/Logistic%20Regression%20using%20R.ppt)
- Quantitative traits
  - Linear regression
- Covariate interations
  - Age, sex etc

Case-control association test
Chi square & OR

Genotype	aa	aA	AA	Total
Case	542	2062	2033	4637
Control	514	1905	1786	4205
Total	1056	3967	3819	8842

aa	aA	AA	Total
0	292	4345	4637
0	381	3824	4205
0	673	8169	8842

Allele	a	A	Total
Case	3146	6128	9274
Control	2933	5477	8410
Total	6079	11605	17684

a	A	Total
292	8982	9274
381	8029	8410
673	17011	17684

Odds (case)	3146/6128=0.513	292/8982=0.0325
Odds (control)	2933/5477=0.5355	381/8029=0.04745
Odds ratio	0.513/0.5355=0.959	0.0325/0.04745=0.685
P (χ^2)	0.183	1.619e-06

Cochran-Armitage Trend Test

Genotype	aa	aA	AA	Sum
Cases	r_0	r_1	r_2	R
Contorls	s_0	s_1	s_2	S
Sum	n_0	n_1	n_2	N

aa	aA	AA	Sum
542	2062	2033	4637
514	1905	1786	4205
1056	3967	3819	8842

$$T := \sum_{i=0}^{2} t_i(r_i S - s_i R),$$

$$\Pr(\text{Case}|\text{Genotype } i) = \Pr(\text{Case}|\text{Genotype } j) = n_i/N$$

- a dominant over A
  t = (1,1,0)
- a recessive to A
  t = (0,1,1)
- a and A additive
  t = (0,1,2)

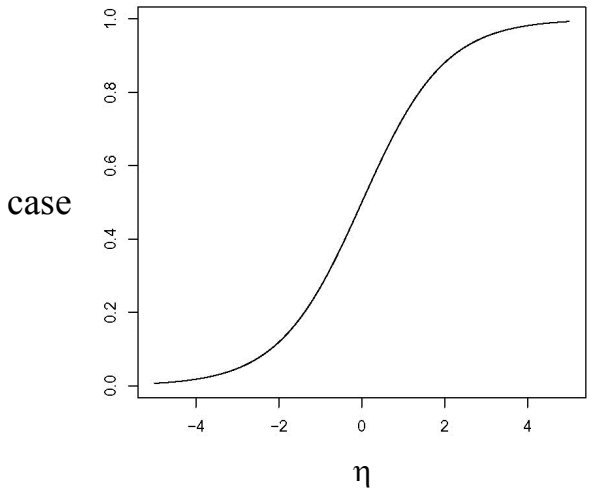
- Additive P = 0.1842
- Dominant P = 0.1941
- Recessive P = 0.4386

Covariate adjustment

- Case-control
  - Logistic regression

$$\eta = \text{genotype} + \text{sex} + \text{age} + \varepsilon$$

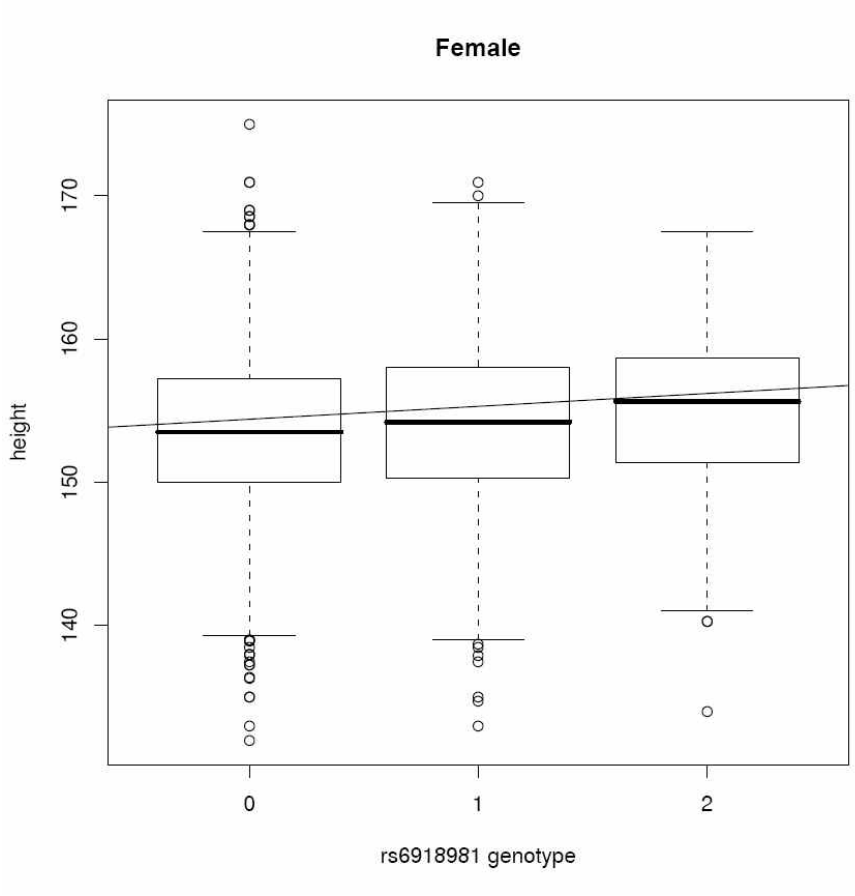
$$\text{case} \sim \exp(\eta)/(1+\exp(\eta))$$



- Quantitative traits
  - Linear regression

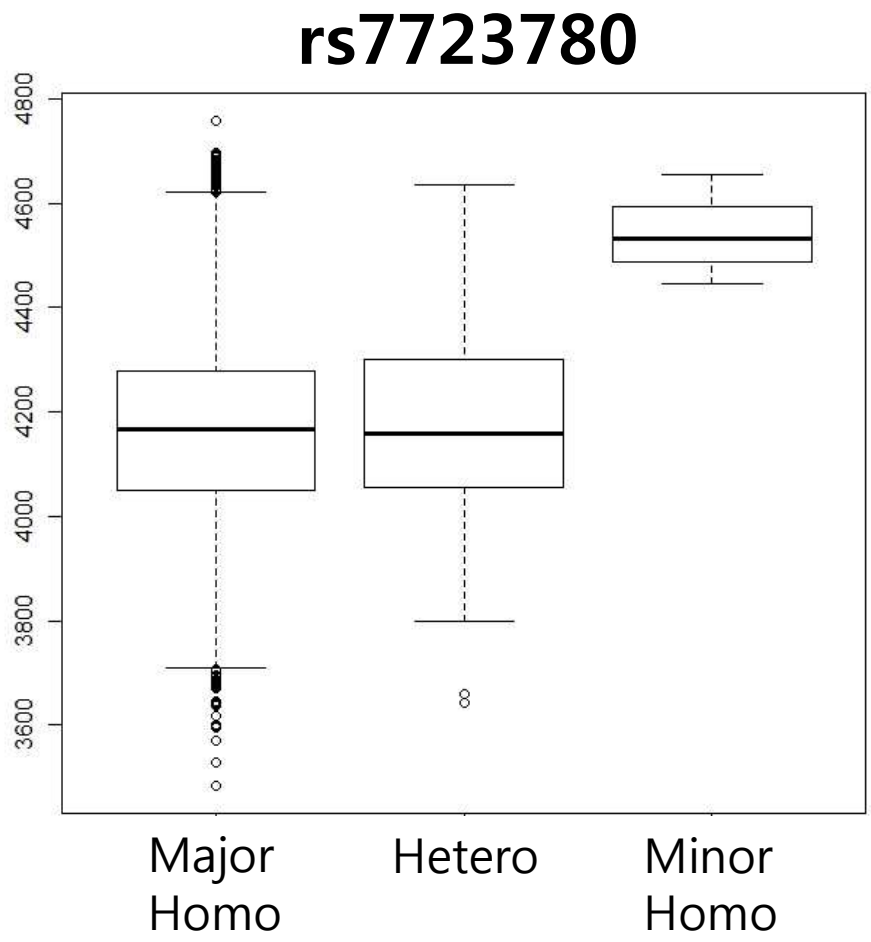
$$\text{height} \sim \text{genotype} + \text{sex} + \text{age} + \varepsilon$$

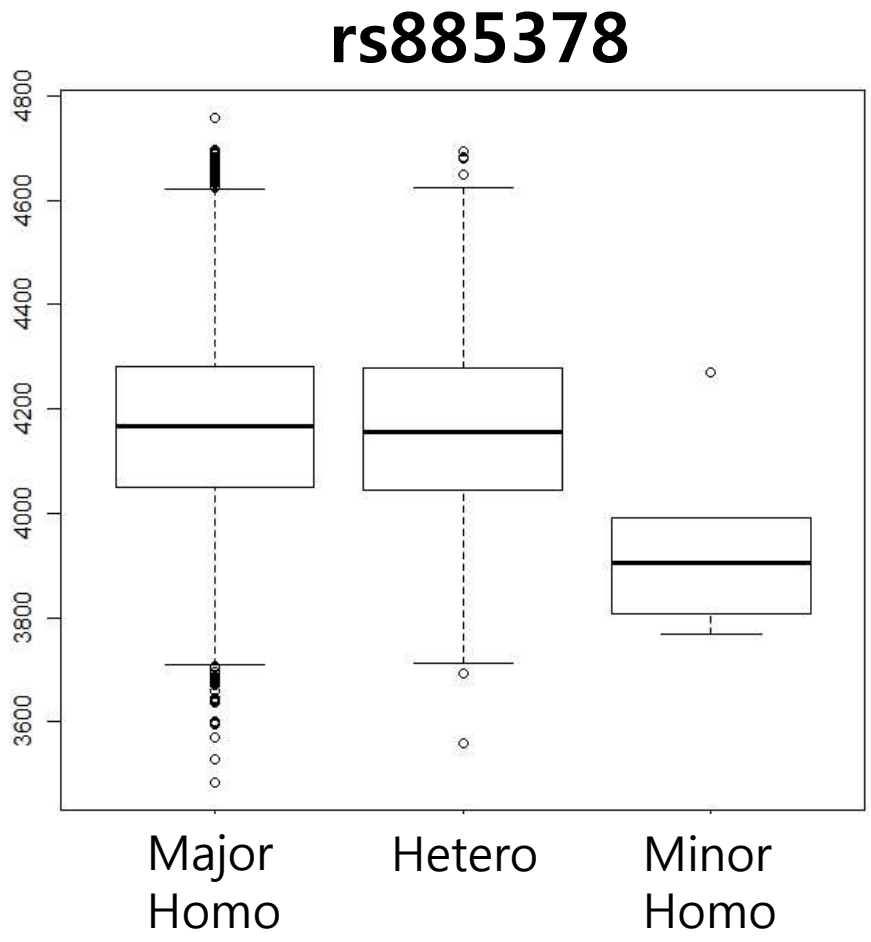
Quantitative traits



- Genotypes coded (additive mode)
  - 0 major homozygotes
  - 1 heterozygotes
  - 2 minor homozygotes
- Linear regression
  - Intercept = 153.54
  - Slope = 0.6086
  - P value = 2.05e-05

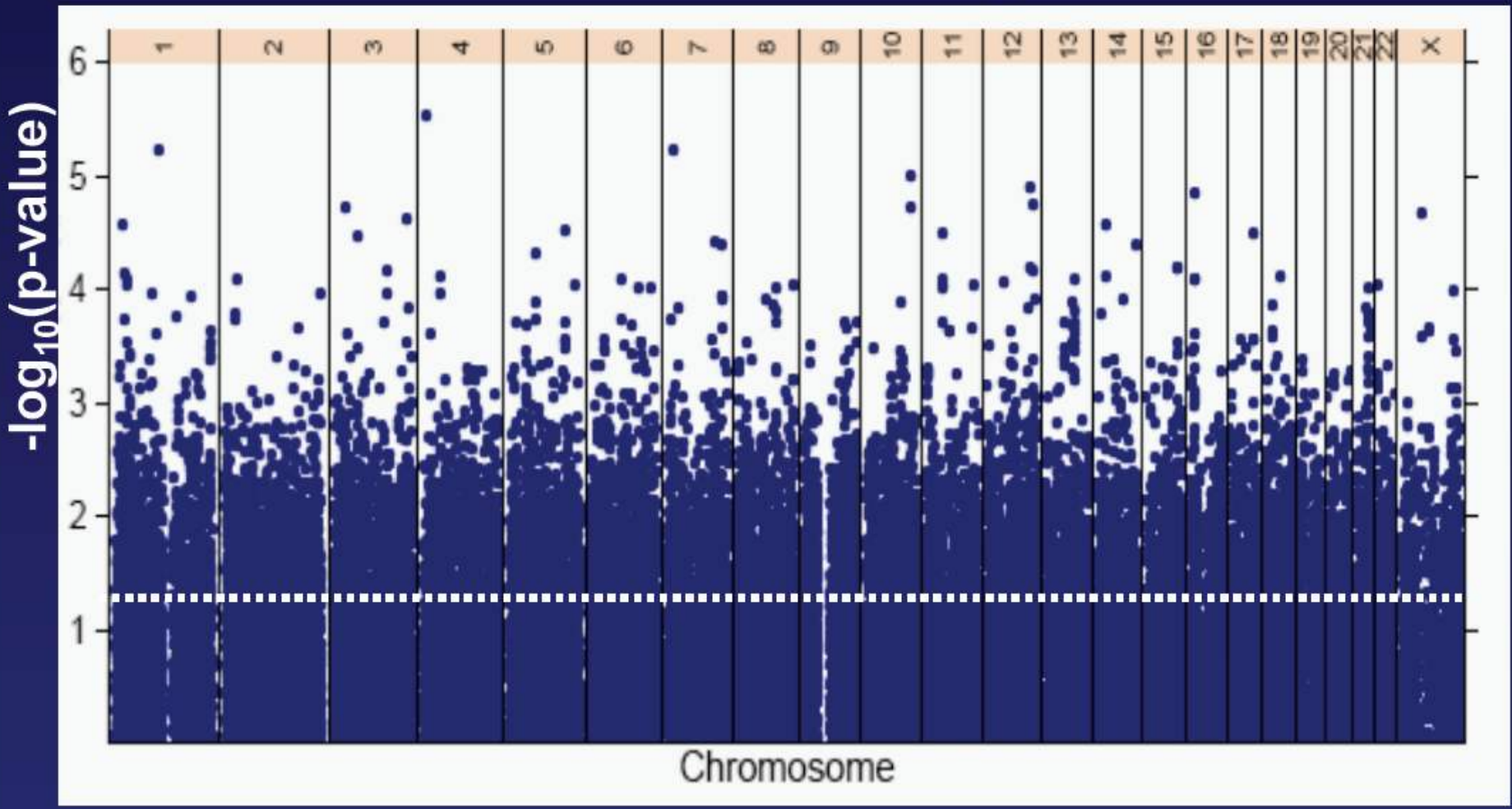
Recessive alleles, protective or risk



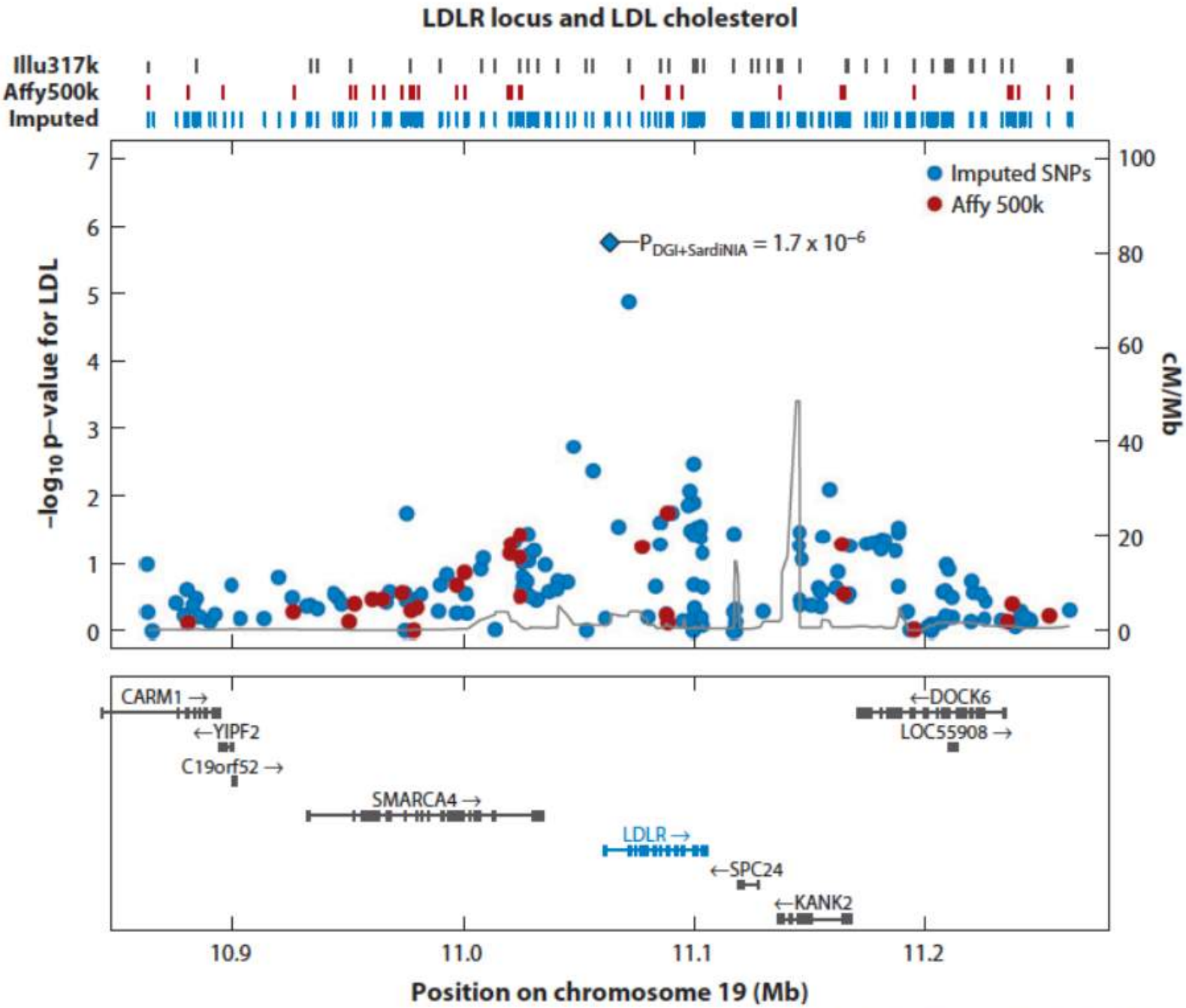


Type 2 diabetes association results

1161 Finnish T2D cases + 1174 Finnish normal glucose tolerant controls



Logistic regression using additive model adjusted for age, gender, birth province

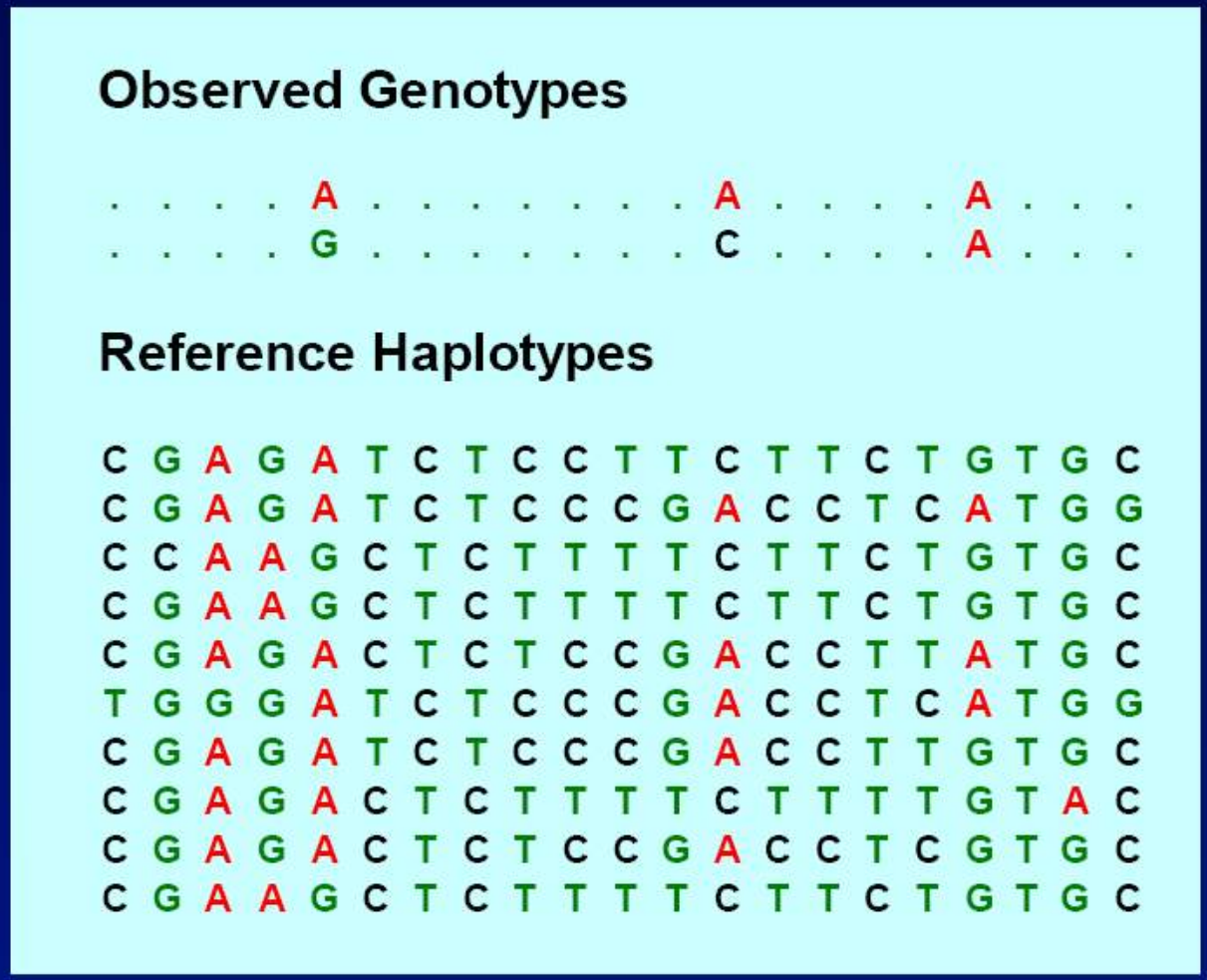


Li (2009) Ann Rev Genomics Hum Genet 10:387

Imputation

- Genotypes not measured with SNP chips can be inferred by referencing HapMap haplotypes
- Increases marker density; helps define signal boundaries
- Facilitates merging datasets from different platforms; critical for meta analysis

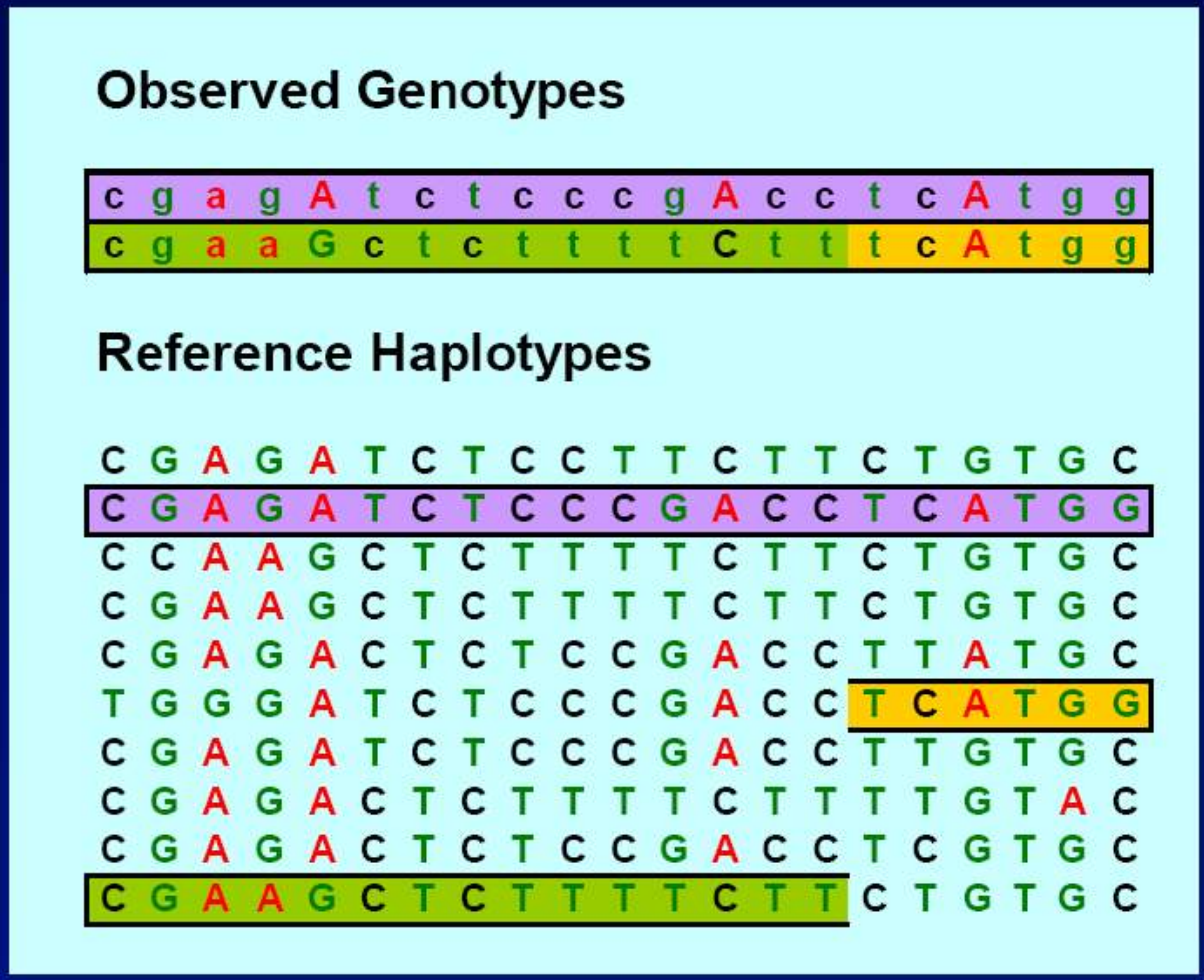
Imputation: Observed genotypes



Li (2009) Ann Rev Genomics Hum Genet 10:387

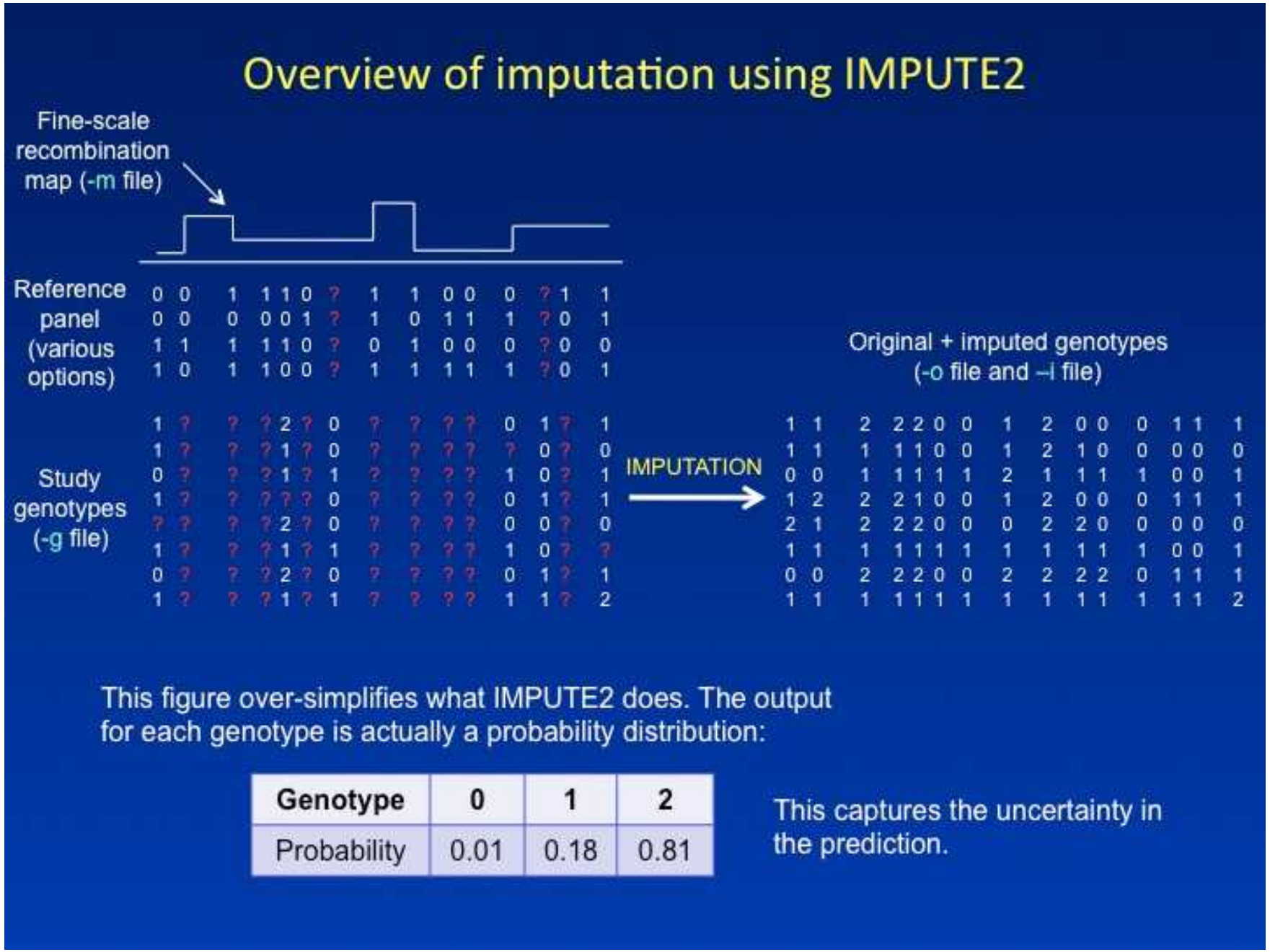
Gonçalo Abecasis

Phase chromosomes, impute missing genotypes

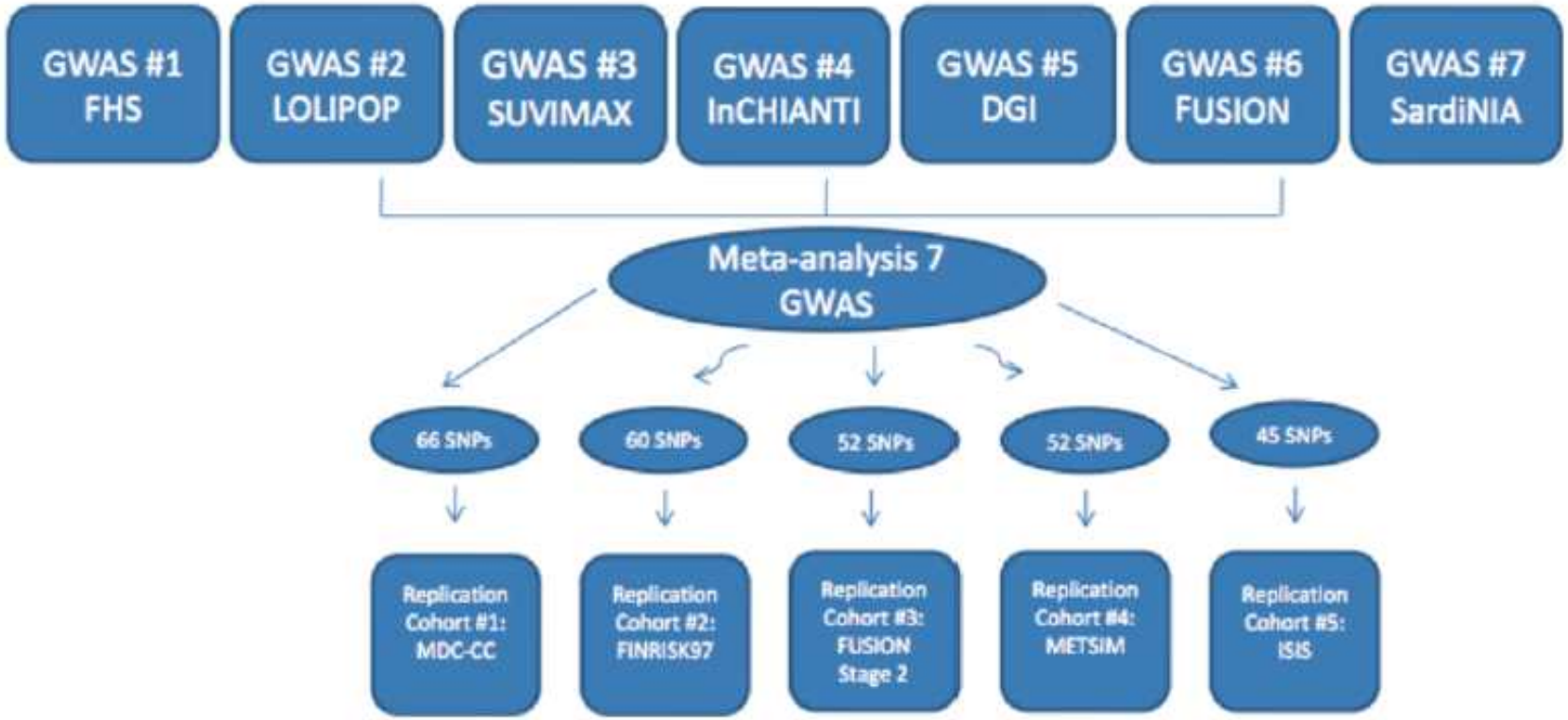


Li (2009) Ann Rev Genomics Hum Genet 10:387

Gonçalo Abecasis



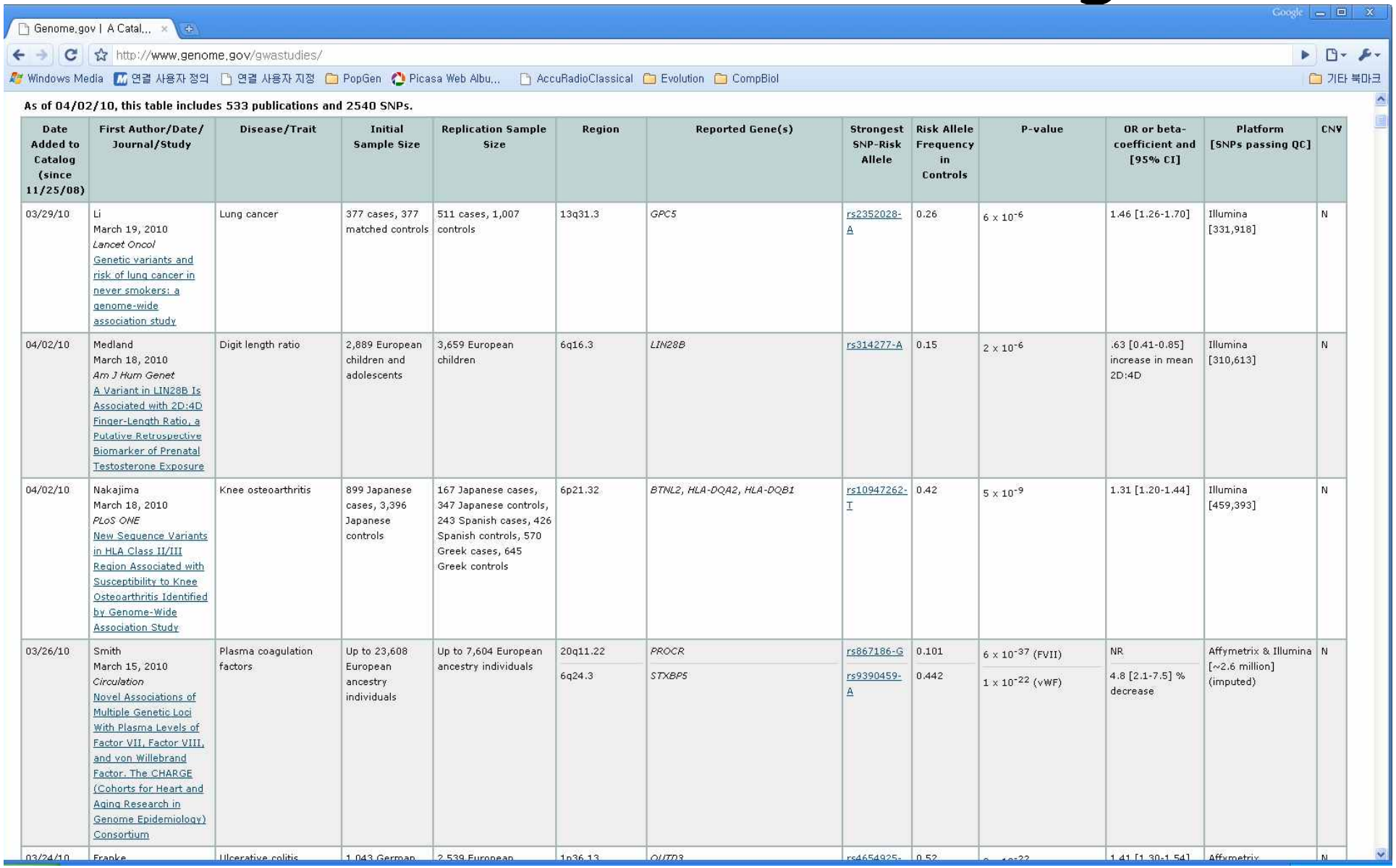
https://mathgen.stats.ox.ac.uk/impute/impute_v2.html



GWA in ~19,840 individuals
Follow-up in ~20,623 individuals

Kathiresan (2009) Nat Gen 41:56

NHGRI GWAS Catalog

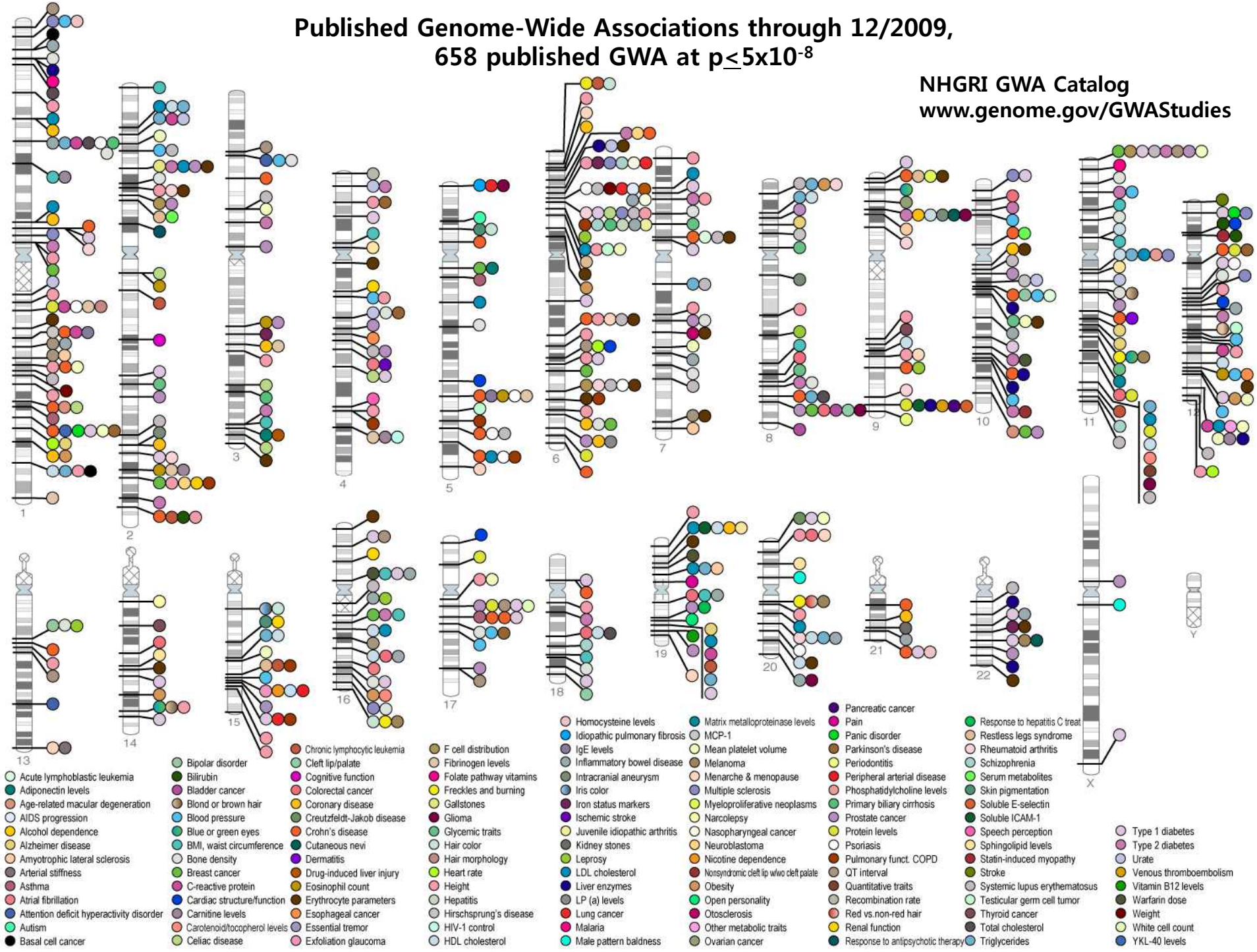


As of 04/02/10, this table includes 533 publications and 2540 SNPs.

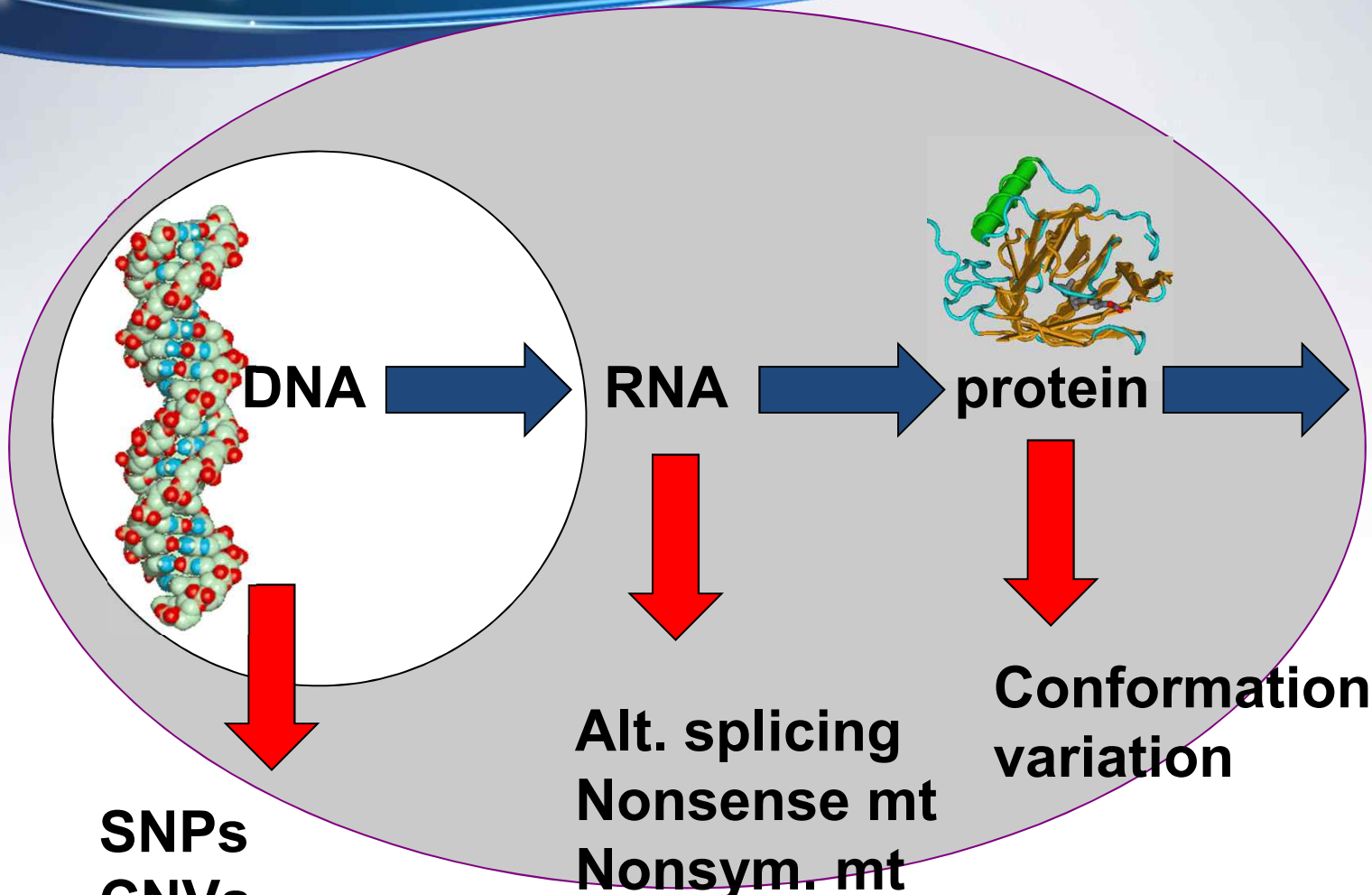
Date Added to Catalog (since 11/25/08)	First Author/Date/ Journal/Study	Disease/Trait	Initial Sample Size	Replication Sample Size	Region	Reported Gene(s)	Strongest SNP-Risk Allele	Risk Allele Frequency in Controls	P-value	OR or beta-coefficient and [95% CI]	Platform [SNPs passing QC]	CNV
03/29/10	Li March 19, 2010 *Lancet Oncol* Genetic variants and risk of lung cancer in never smokers: a genome-wide association study	Lung cancer	377 cases, 377 matched controls	511 cases, 1,007 controls	13q31.3	*GPC5*	rs2352028-A	0.26	6×10^{-6}	1.46 [1.26-1.70]	Illumina [331,918]	N
04/02/10	Medland March 18, 2010 *Am J Hum Genet* A Variant in LIN28B Is Associated with 2D:4D Finger-Length Ratio, a Putative Retrospective Biomarker of Prenatal Testosterone Exposure	Digit length ratio	2,889 European children and adolescents	3,659 European children	6q16.3	*LIN28B*	rs314277-A	0.15	2×10^{-6}	.63 [0.41-0.85] increase in mean 2D:4D	Illumina [310,613]	N
04/02/10	Nakajima March 18, 2010 *PLoS ONE* New Sequence Variants in HLA Class II/III Region Associated with Susceptibility to Knee Osteoarthritis Identified by Genome-Wide Association Study	Knee osteoarthritis	899 Japanese cases, 3,396 Japanese controls	167 Japanese cases, 347 Japanese controls, 243 Spanish cases, 426 Spanish controls, 570 Greek cases, 645 Greek controls	6p21.32	*BTNL2, HLA-DQA2, HLA-DQB1*	rs10947262-T	0.42	5×10^{-9}	1.31 [1.20-1.44]	Illumina [459,393]	N
03/26/10	Smith March 15, 2010 *Circulation* Novel Associations of Multiple Genetic Loci With Plasma Levels of Factor VII, Factor VIII, and von Willebrand Factor. The CHARGE (Cohorts for Heart and Aging Research in Genome Epidemiology) Consortium	Plasma coagulation factors	Up to 23,608 European ancestry individuals	Up to 7,604 European ancestry individuals	20q11.22 6q24.3	*PROCR* *STXBP5*	rs867186-G rs9390459-A	0.101 0.442	6×10^{-37} (FVII) 1×10^{-22} (vWF)	NR 4.8 [2.1-7.5] % decrease	Affymetrix & Illumina [~2.6 million] (imputed)	N
03/24/10	Franke	Ulcerative colitis	1,043 German	2,539 European	1p36.13	*ZUTD3*	rs4654925-	0.52	[illegible]	1.41 [1.30-1.54]	Affymetrix	N

Published Genome-Wide Associations through 12/2009, 658 published GWA at $p \leq 5 \times 10^{-8}$

NHGRI GWA Catalog
www.genome.gov/GWAStudies

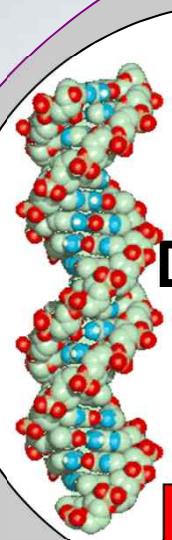


Central Dogma

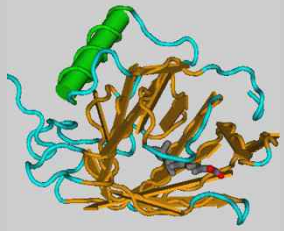


Adapted from Pevsner 2003

Genome-wide Association Study (GWAS)



DNA → RNA → protein → phenotype



SNPs
CNVs
Structural var.

Statistical association btw
genotype & phenotype

GWAS may have low power due to

- Multiple intermediate steps such as epigenetic & transcriptional regulations
- Multiple DNA variants may contribute to the same phenotype
- These factors usually form a complex network of interactions

Most GWAS hits are

- not non-synonymous substitutions

- found in non-exonic regions
  - likely to regulate gene expression

➢ Gene expression difference btw individuals may be molecular and intermediate phenotypes
  - inducing changes in higher-order disease traits (Schadt et al. *PLoS Biol*. 2008)

Hundreds of GWAS applications tells us

- Many common variants of highly significant disease association have been found
- They confer relatively small increments in risk (1.0~1.5 fold)
- They explain only a small portion of heritability
  - Human height is estimated to have 80% heritability
  - About 5% of phenotype variance is explained based on $>10^4$ people

Manolio et al., Finding the missing heritability of complex diseases. Nature 2009

Excuses for the missing heritability

- Large numbers of variants of smaller effect yet to be found
- Rarer variants (possibly with larger effect)
- Structural variants poorly captured by existing arrays
- Low power to detect gene-gene interactions
- Inadequate accounting for shared environment among relatives

Manolio et al., Finding the missing heritability of complex diseases. Nature 2009

NATURE GENETICS VOLUME 42 | NUMBER 7 | JULY 2010

Common SNPs explain a large proportion of the heritability for human height

SNPs discovered by genome-wide association studies (GWASs) account for only a small fraction of the genetic variation of complex traits in human populations. Where is the remaining heritability? We estimated the proportion of variance for human height explained by 294,831 SNPs genotyped on 3,925 unrelated individuals using a linear model analysis, and validated the estimation method with simulations based on the observed genotype data. We show that 45% of variance can be explained by considering all SNPs simultaneously. Thus, most of the heritability is not missing but has not previously been detected because the individual effects are too small to pass stringent significance tests. We provide evidence that the remaining heritability is due to incomplete linkage disequilibrium between causal variants and genotyped SNPs, exacerbated by causal variants having lower minor allele frequency than the SNPs explored to date.

LETTER

832 | NATURE | VOL 467 | 14 OCTOBER 2010

doi:10.1038/nature09410

Hundreds of variants clustered in genomic loci and biological pathways affect human height

- Meta analysis of 46 data sets
- 133,653 European individuals
- Pathways found
  - Growth, kinase, development, insulin, bone etc

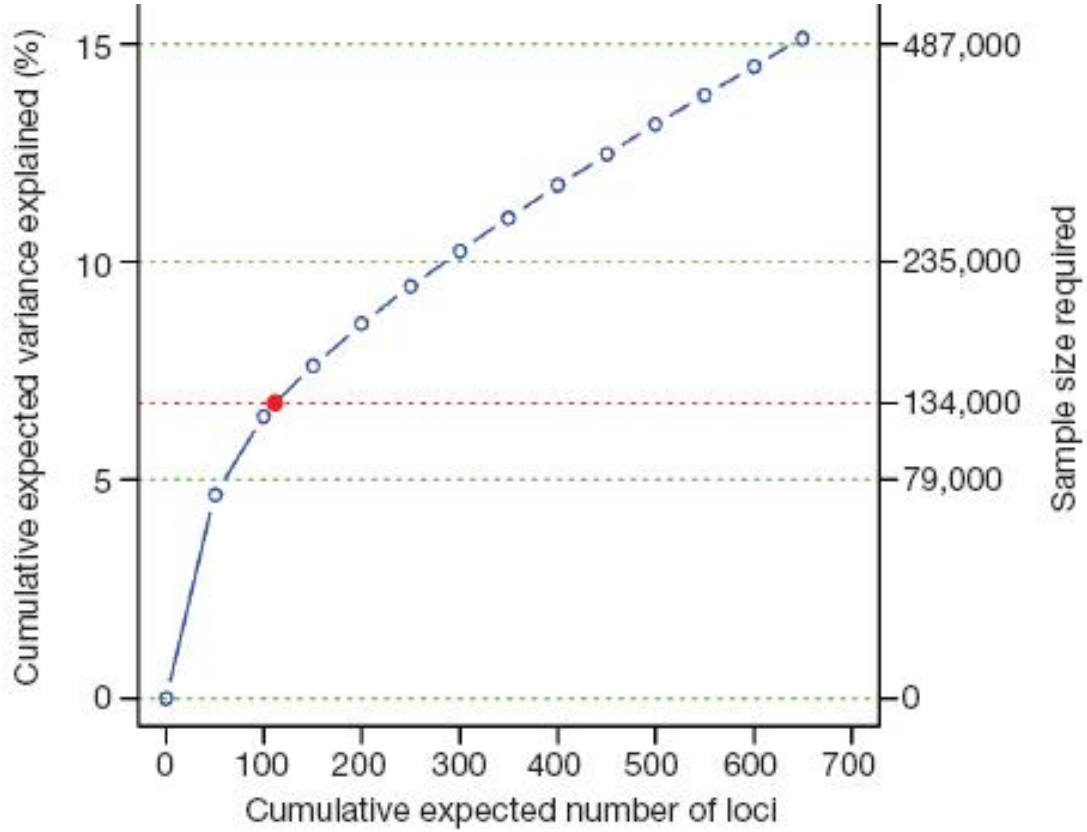


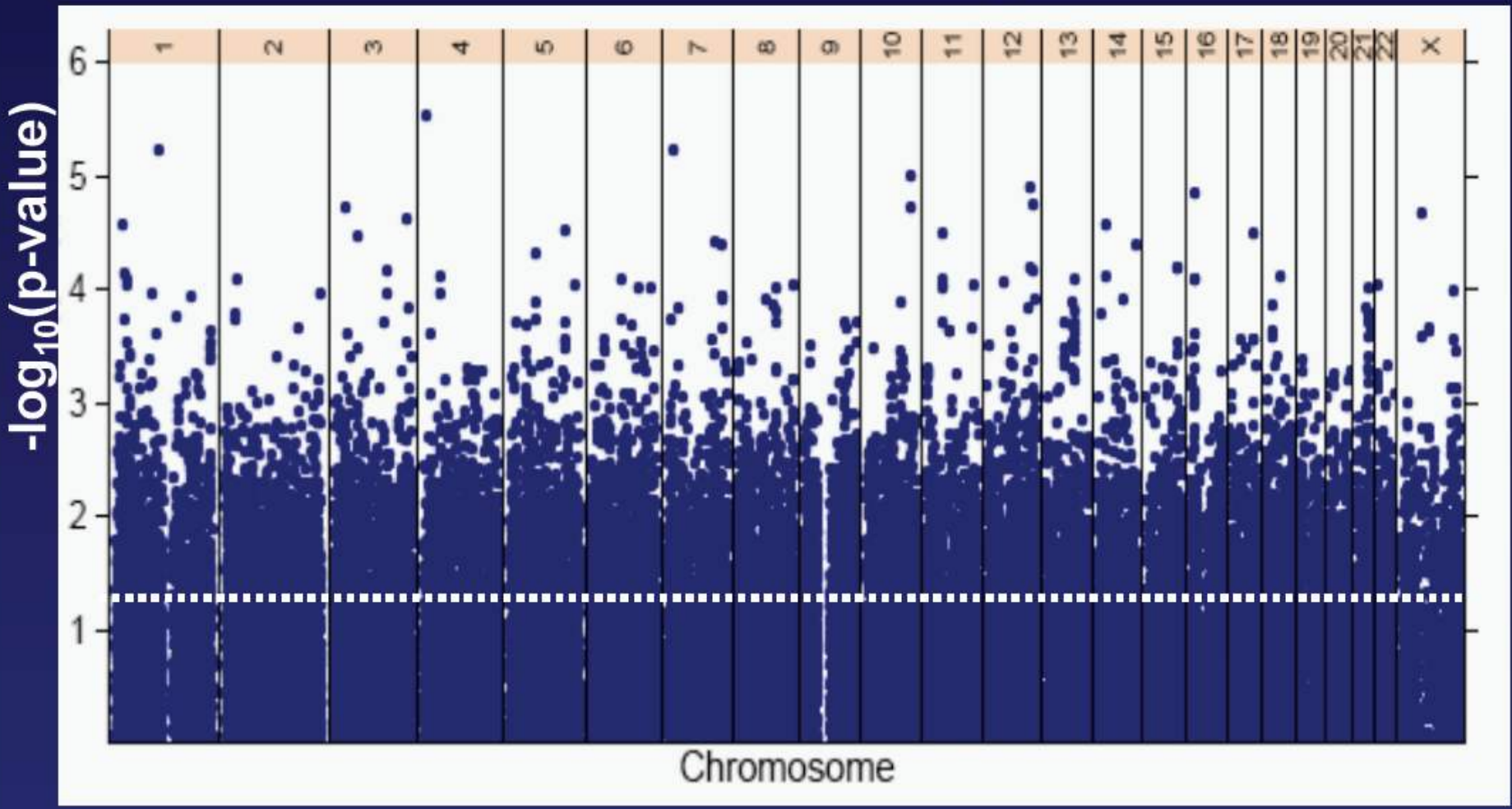
Figure 1 | Phenotypic variance explained by common variants.

Problems of GWAS

- Correcting p-values of a million of hypotheses
- A very stringent cutoff is used to yield only a small number of significant SNPs
- Many moderate associations below the cutoff is lost
- This is very ineffective and wasteful

Type 2 diabetes association results

1161 Finnish T2D cases + 1174 Finnish normal glucose tolerant controls

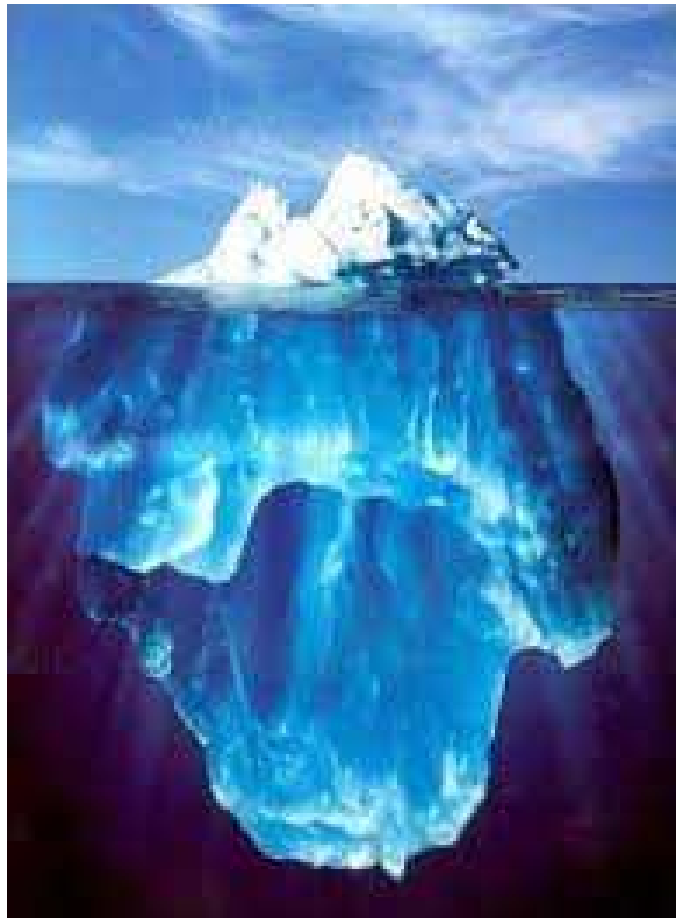


Logistic regression using additive model adjusted for age, gender, birth province

Highly significant signals are found, but difficult to discuss biology

Trait	RS ID	Class	Locus	Nearby genes[a]	Minor allele	MAF	GWAS ($n = 8,842$)
BMI	rs17178527	Unknown	6q24.1	*LOC729076*	A	0.25	1.2E–08
	rs9939609	Intron	16q12.2	*FTO*	A	0.13	1.7E–06
WHR	**rs2074356**	Intron	12q24.13	*C12orf51*	T	0.15	1.8E–07
	rs17089410	Unknown	13q21.33		T	0.14	6.1E–06
Height	rs6918981	Unknown	6p21.31	*HMGA1*	G	0.21	3.2E–08
	rs17038182	Unknown	1p12		C	0.42	4.3E–08
	rs10513137	Intron	3q23	*ZBTB38*	A	0.26	5.6E–08
	rs13273123	Intron	8q12.1	*PLAG1*	G	0.07	1.1E–06
	rs600130	Intron	9q22.32	*FBP2*	G	0.15	2.7E–06
	rs2079795	Unknown	17q23.2	*BCAS3, TBX2*	A	0.33	2.9E–06
	rs3791675	Intron	2p16.1	*EFEMP1*	G	0.22	3.6E–06
	rs41464348	Intron	2p22.3	*LTBP1*	T	0.35	7.4E–06
SBP	**rs17249754**	Unknown	12q21.33	*ATP2B1*	A	0.37	9.1E–07
	rs715987	Unknown	10p15.1		C	0.15	4.5E–06
DBP	rs17249754	Unknown	12q21.33	*ATP2B1*	A	0.37	1.2E–06
Pulse rate	**rs12731740**	Unknown	1q32.2	*CD46, LOC148696*	T	0.10	3.7E–07
	rs12110693	Unknown	6q22.31	*LOC644502*	A	0.49	1.3E–06
	rs11576175	Intron	1q21.2	*CTSS*	A	0.24	8.3E–06
BD-RT	**rs7776725**	Intron	7q31.31	*FAM3C*	C	0.13	1.0E–11
	rs9525667	Unknown	13q14.11		T	0.43	3.1E–06
BD-TT	**rs7776725**	Intron	7q31.31	*FAM3C*	C	0.13	1.6E–06
	rs1721400	Unknown	7p14.1	*TXNDC3, SFRP4, EPDR1*	T	0.17	1.4E–07
	rs550677	NearGene-5	12q24.31	*TMEM132B*	T	0.17	5.0E–06
	rs6974574	Unknown	7p14.1		A	0.30	7.9E–06

KARE results
Cho et al, 2009

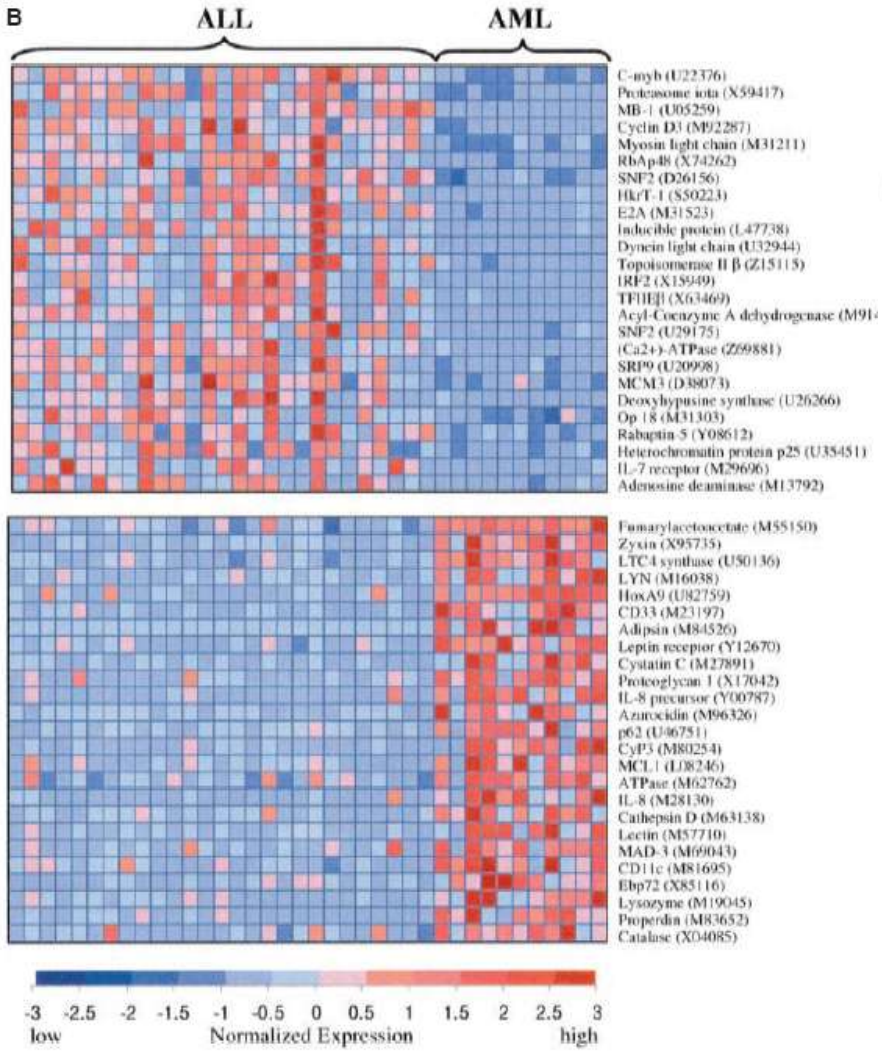


Gene-Set based approach

- Testing the association of **biologically pre-defined gene sets** instead of testing individual SNPs
- Gene sets are derived from Gene Ontology, KEGG pathways, molecular signatures, etc
- It aims to detect **moderate but coordinated** associations within a gene set (as well as strong signals)

Gene-Set based approach

- **Rationale**: Even if the members of a gene set are only moderately associated, such moderate signals taken together can represent a significant pattern
- Such set-wise association signals may be more **reproducible** among different cohorts



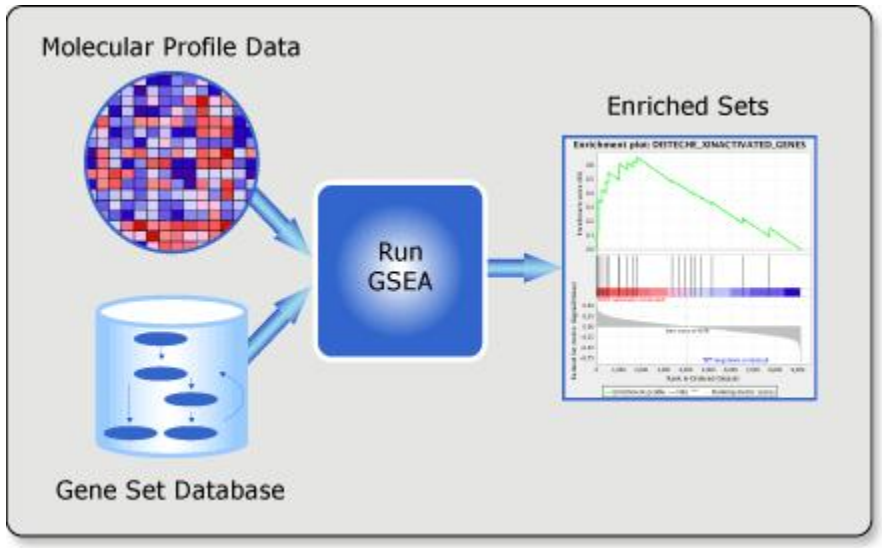
Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring

Science, 1999

Classification of two types of leukemia data using microarray

GSEA (Gene-set Enrichment Analysis)

Broad Inst.



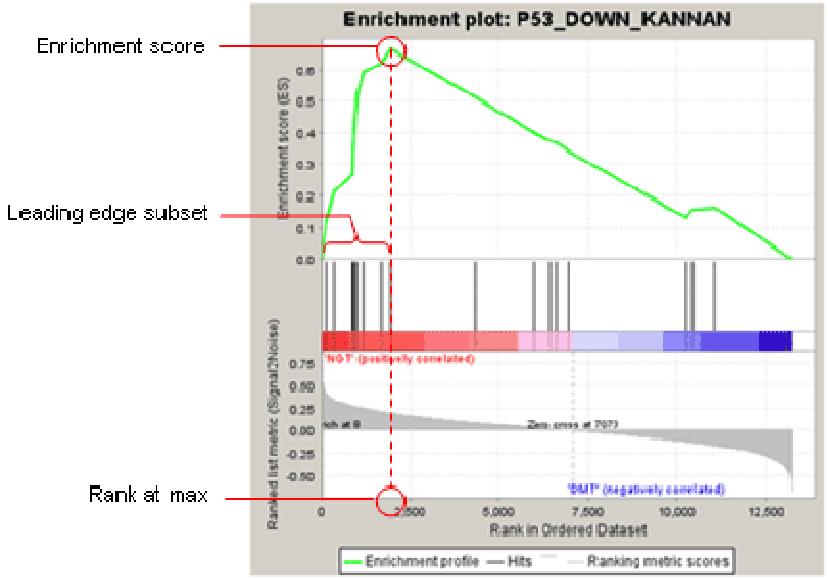
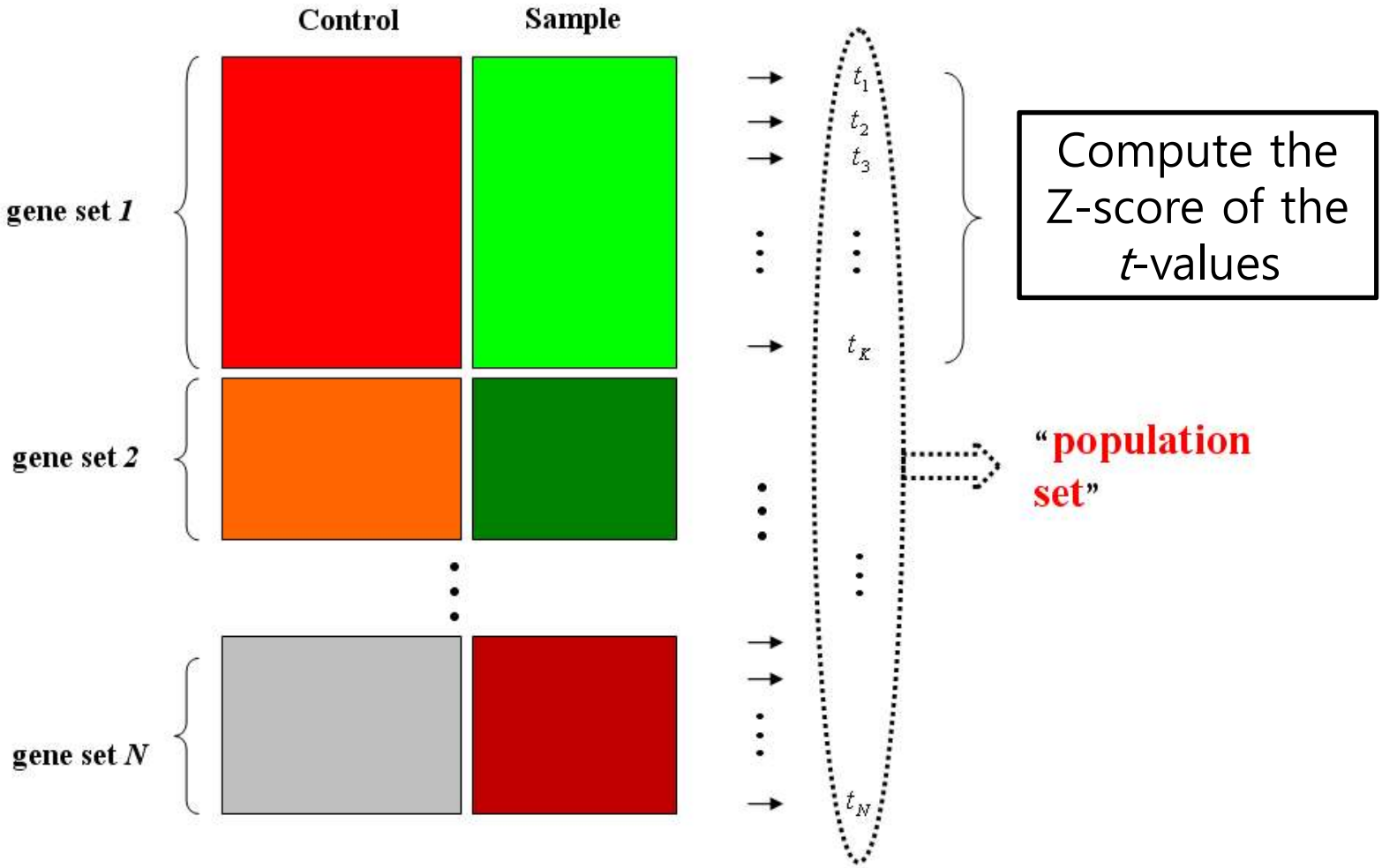


Fig 1: Enrichment plot: P53_DOWN_KANNAN
Profile of the Running ES Score & Positions of GeneSet Members on the Rank Ordered List

Z-statistic method

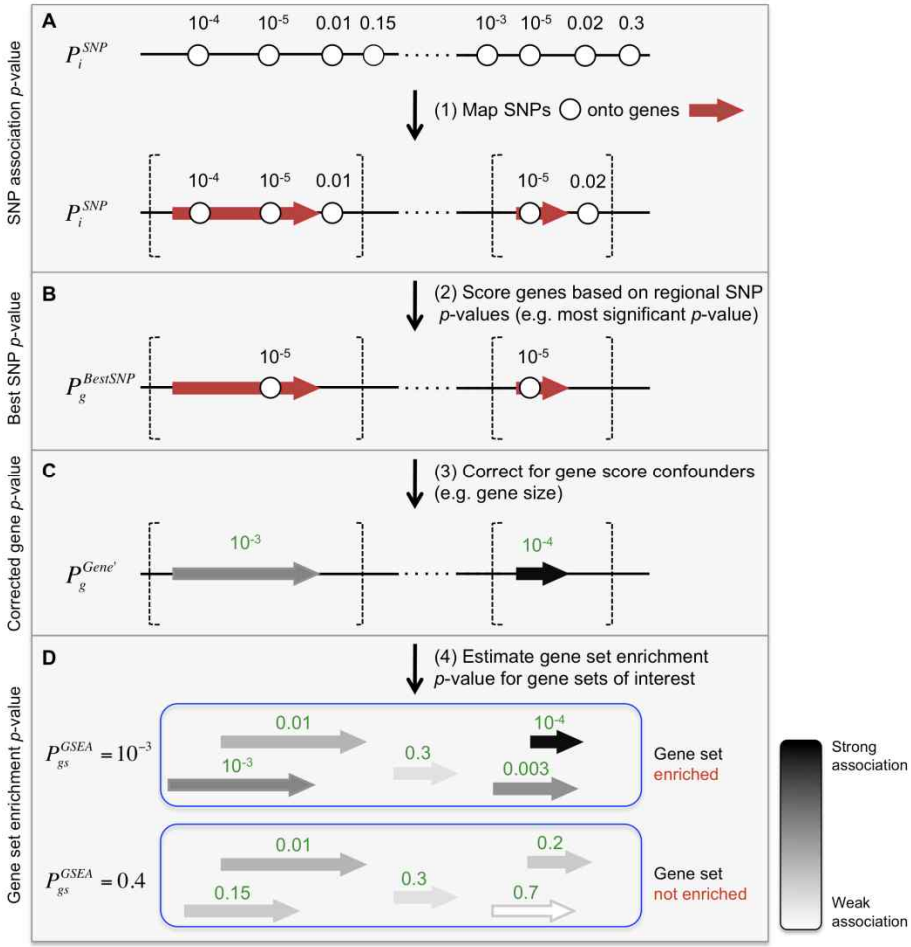
- In gene expression array



Gene-set Analysis of GWAS

- Compute association p-values for each SNP using a GWAS software
- Assign each SNP to the nearest gene: within some padding (eg 20k bp)
- Gene score: Choose the best p-values among the assigned SNPs
- Then, apply GSA on the gene scores

PLOS Gen. (2010) Segrè et al.



How to assess the significance of a gene-set

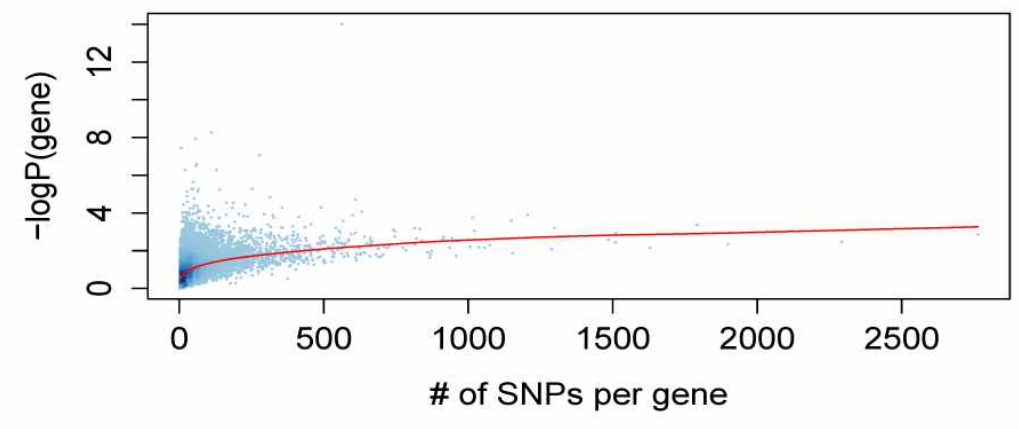
- If genotype data are available,
  - permute the phenotype labels and
  - do GWAS followed by GSA for each permutation
  - Count the permutations that exceed the original gene-set score
- The original GSEA implemented this approach

- Label permutation can remove most biases due to variable SNP density and gene-set size

➢ Most often the genotype data are not available but only the SNP P-values are available

Biases in Gene Scores

- Gene score is often assigned by the best SNP *P*-values
- This can be biased if the number of SNPs per gene is variable
  - The more one samples, the more extreme values are likely observed



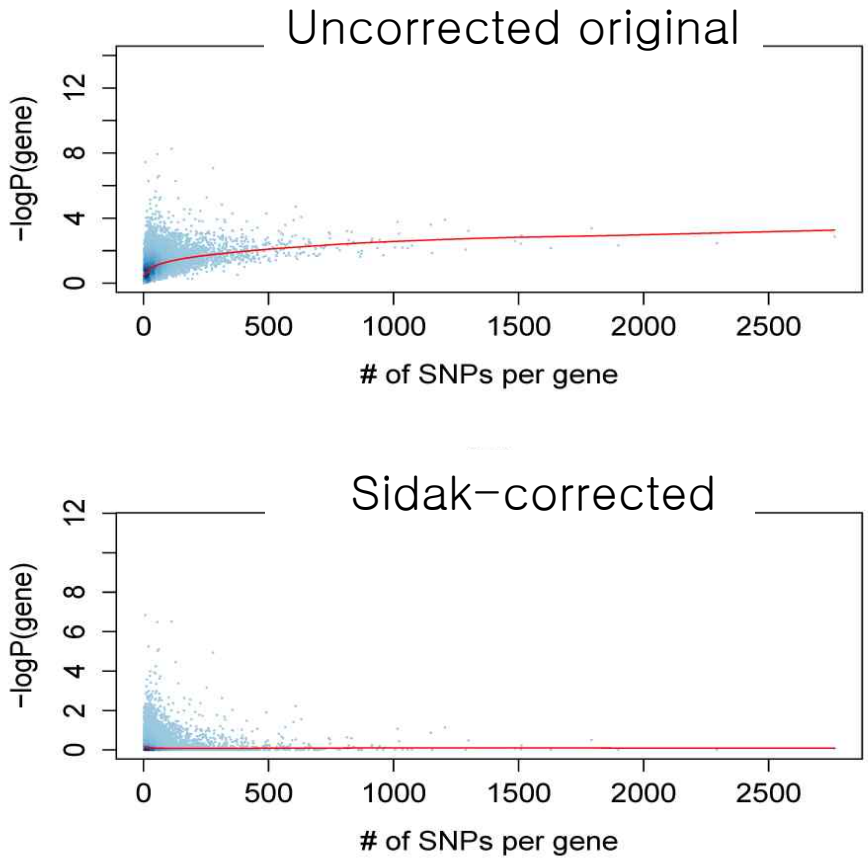
- Various corrections have been suggested
  - Analytical formula
  - Empirical regression
  - Simulation method

Gene score correction

- Šidák's multiple testing correction

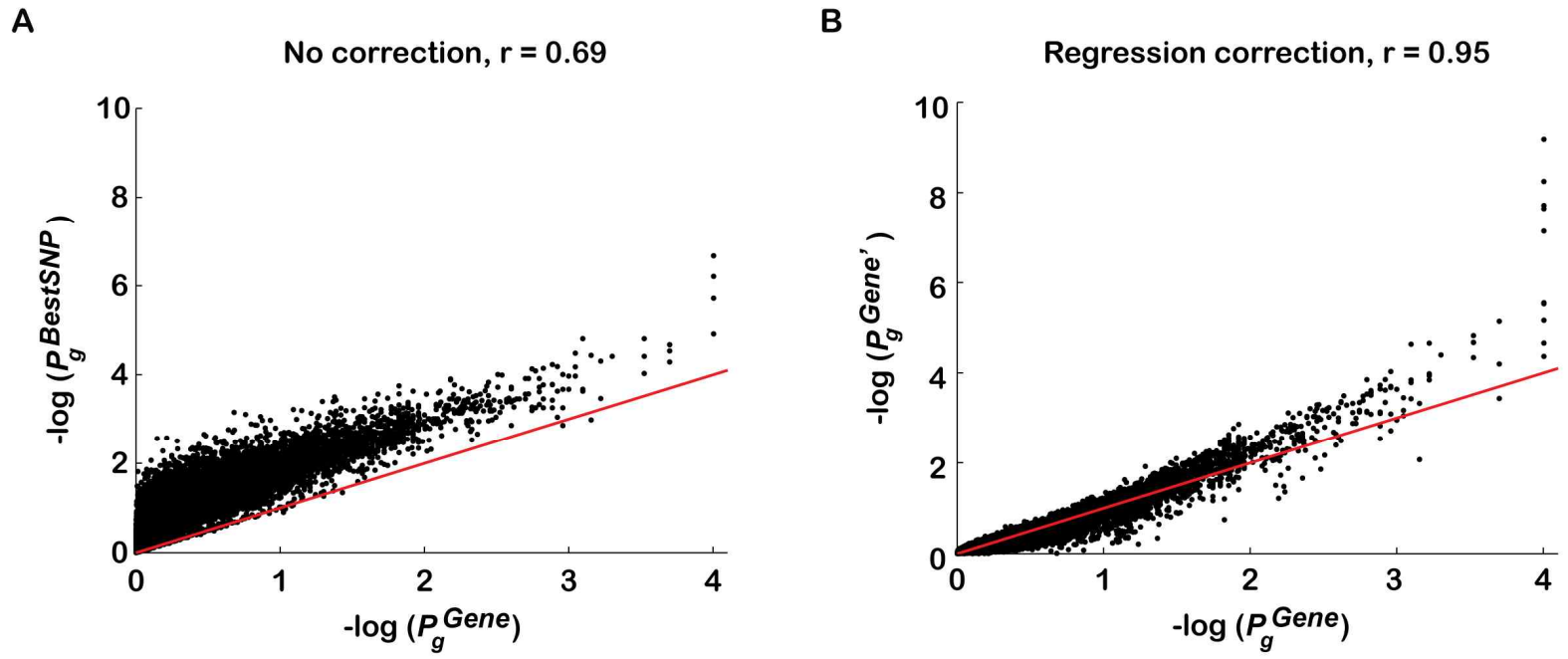
$$P' = 1 - (1-P)^{(N+1)/2}$$

  - N SNPs for a gene
  - About ½ of them are outside linkage disequilibrium (LD)



Empirical regression-based correction

MAGENTA @ Broad Inst.

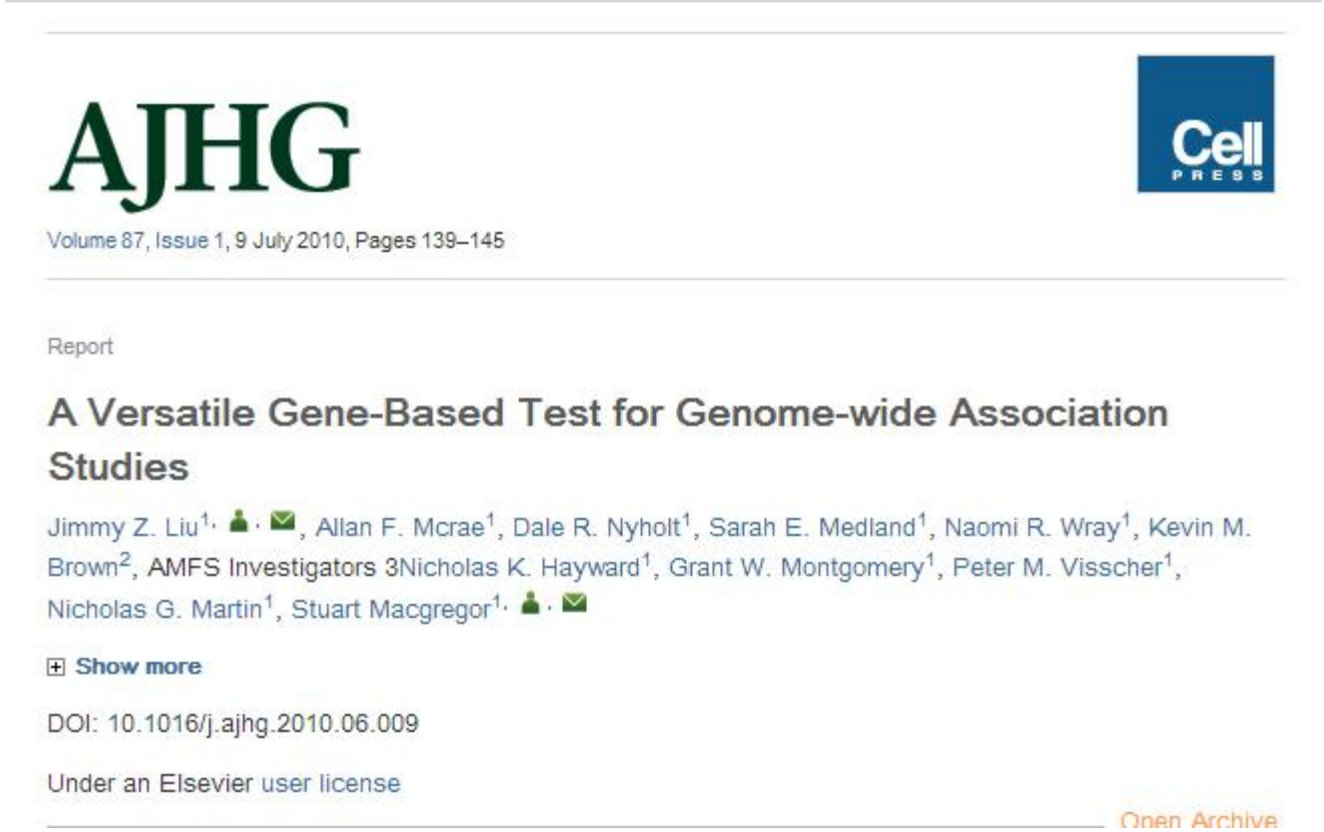


$$Z_g^{BestSNP} = \alpha \cdot d_g + \beta \cdot n_g + \delta \cdot u_g + \gamma \cdot h_g + \eta \cdot c_g + \kappa \cdot l_g + r_g$$

- Gene score is regressed by factors such as SNP density, recombination hotspots, LD block size etc

PLOS Gen. (2010) Segrè et al.

Simulation-based gene-scores



AJHG

Cell PRESS

Volume 87, Issue 1, 9 July 2010, Pages 139–145

Report

A Versatile Gene-Based Test for Genome-wide Association Studies

Jimmy Z. Liu[1], Allan F. Mcrae[1], Dale R. Nyholt[1], Sarah E. Medland[1], Naomi R. Wray[1], Kevin M. Brown[2], AMFS Investigators 3Nicholas K. Hayward[1], Grant W. Montgomery[1], Peter M. Visscher[1], Nicholas G. Martin[1], Stuart Macgregor[1]

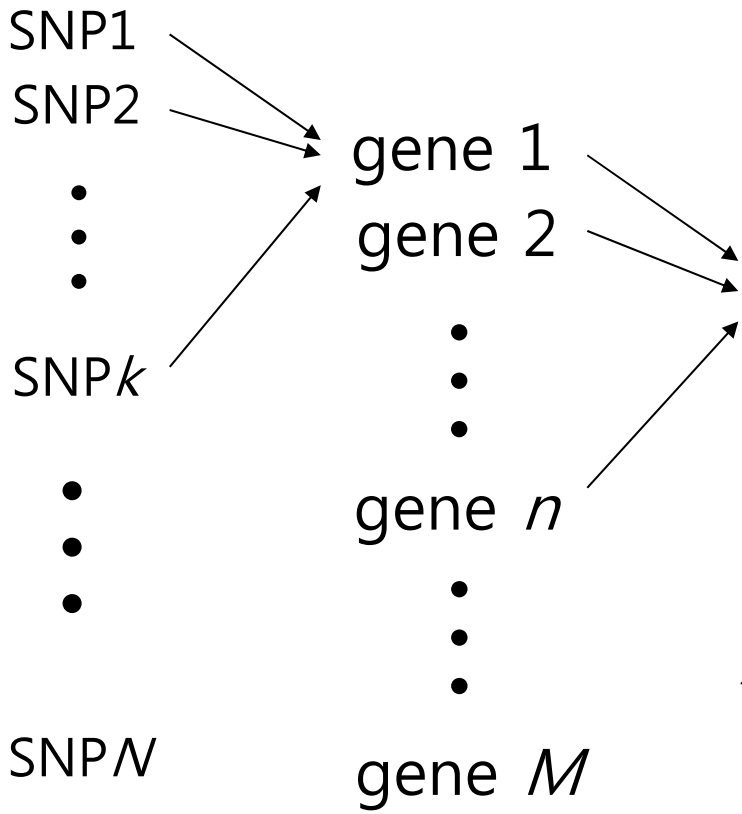
Show more

DOI: 10.1016/j.ajhg.2010.06.009

- VEGAS requires LD information of the population (usually from HapMap)

How to assess gene-set score

Z-statistics



gene set 1

$$Z(GS) = \frac{\overline{X} - m_0}{\sigma / \sqrt{n}}$$

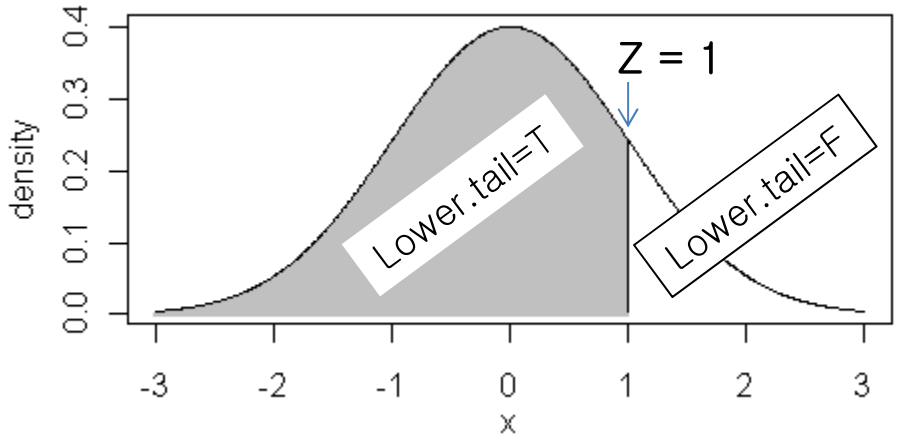
$\overline{X}$: mean of n gene scores
m_0 : mean of M gene scores
Σ : sd of M gene scores

Significance of gene-set scores

- Parametric *P*-value

  P = pnorm(*Z*,lower.tail=F)

- Permutation *P*-value
  - Random sample the same number of genes per gene-set
  - Calculate Z-scores for each permutation
  - Count the number of permutations exceeding the original Z-score



- Permutation approach replaces the density distribution function with the one empirically generated through permutation of the real gene scores

Statistics other than Z

- Two-sample Wilcoxon (Mann-Whitney) test

  wilcox.test(*gs*, *all*, alternative='gr')

  - *gs*: scores of gene-set member genes
  - *all*: scores of all genes

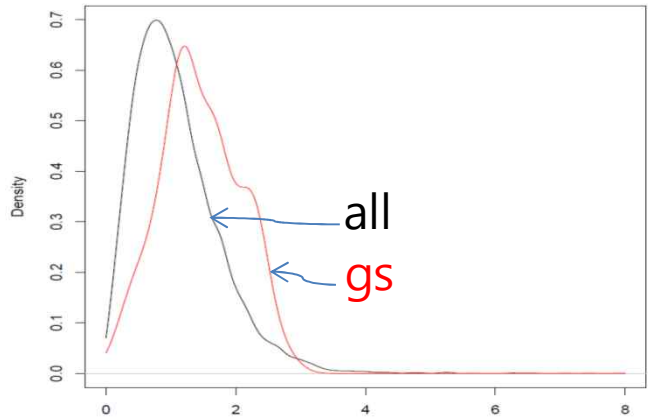
- Kolmogorov-smirnov test

  ks.test(*gs*, *all*, alternative='le')

- GSEA statistic

$$\mathrm{ES}(S) = \max_{1 \le j \le N} \left\{ \sum_{G_{j^*} \in S, j^* \le j} \frac{\left| r_{(j^*)} \right|^p}{N_R} - \sum_{G_{j^*} \notin S, j^* \le j} \frac{1}{N - N_H} \right\}$$

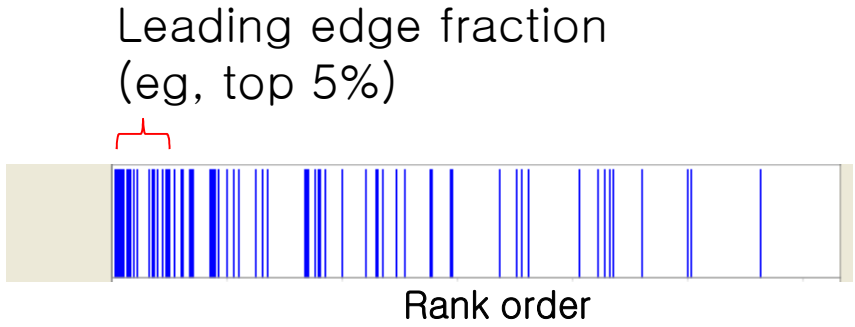
  - *r*: gene score $N_R = \sum_{G_{j^*} \in S} \left| r_{(j^*)} \right|^p$
  - N_H: gene-set size p = 1 (usually)



Counting leading edge fraction only

MAGENTA @ Broad Inst.

- Count the number of genes from a gene-set within the top ranked ones (eg, top 5%)
- Compare this with the permutations to assess significance
  - Detailed distribution of low ranking genes is immaterial, focusing only the strong signals
  - How to cutoff 'top' ranks?



- *MAGENTA* suggests to use lower cutoff for complex traits with many contributing genes

Weighting by leading edge fraction

i-GSEA4GWAS @ Chin. Acad. Sci.

- SNP permutation instead of phenotype permutation
- Otherwise, the same as GSEA4GWAS
- Weight gene-set scores by the leading edge fraction
  - Proportion of genes mapped by top 5% SNPs
  - Perhaps too sensitive(?)

Table 1. The number of gene set hits identified by gene set analyses

Software	Gene score[a]	GO		KEGG	
		Unimputed	Imputed	Unimputed	Imputed
i-GSEA4GWAS	Best	283	1,070	12	78
GSA-SNP	Best	61	27	14	9
	Second best	94	38	20	19

GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; GSA, gene set analysis; SNP, single nucleotide polymorphism.
[a]Either the best or second-best p-value of SNPs residing inside or within 20 kb of the gene boundary was assigned to each gene as the score. Unlike i-GSEA4GWAS, which assigns the best p-value, GSA-SNP has an option to assign the second-best p-value.

Multiple testing correction of gene-set analysis results

- Once P-value is calculated for each gene-set,
  - We want to report a list of gene-sets that are significantly associated
- This is a typical multiple testing problem as we have tested gene-sets on the order of
  - hundreds (KEGG pathways)
  - thousands (GO terms)

- Bonferroni correction

  $Q = P \times N$

  N : # of gene-sets tested
  - Perhaps too stringent

- Benjamini-Hochberg FDR

  $Q_k = P_k \times N \div k$

  P_k : sorted raw P-values in ascending order
  - Accept the largest k at the desired significance level

We are NOT the first advocating this strategy

- 9 different methods are available

ALIGATOR
i-GSEA4GWAS
GenGen
GESBAP
GRASS
GSA-SNP
GSEA-SNP
PLINK set-test
SNP ratio test

NATURE REVIEWS | GENETICS
VOLUME 11 | DECEMBER 2010 | 843

GENOME-WIDE ASSOCIATION STUDIES

Analysing biological pathways in genome-wide association studies

*Kai Wang**‡, *Mingyao Li*§ *and Hakon Hakonarson**||

Abstract | Genome-wide association (GWA) studies have typically focused on the analysis of single markers, which often lacks the power to uncover the relatively small effect sizes conferred by most genetic variants. Recently, pathway-based approaches have been developed, which use prior biological knowledge on gene function to facilitate more powerful analysis of GWA study data sets. These approaches typically examine whether a group of related genes in the same functional pathway are jointly associated with a trait of interest. Here we review the development of pathway-based approaches for GWA studies, discuss their practical use and caveats, and suggest that pathway-based approaches may also be useful for future GWA studies with sequencing data.

GSA-SNP software

- A Java based software for gene set analysis of SNP arrays
- Provides **three** widely used gene set analysis methods for SNPs: Z-statistic, Restandardization, and GSEA
- Based on p-values: Applicable to **both case-control and quantitative trait** data
- Quite **fast** and **easy to use**

Published online 25 May 2010 *Nucleic Acids Research, 2010, Vol. 38, Web Server issue* ***W749–W754***
doi:10.1093/nar/gkq428

GSA-SNP: a general approach for gene set analysis of polymorphisms

Dougu Nam[1], Jin Kim[2], Seon-Young Kim[3] and Sangsoo Kim[4,*]

[1]School of Nano-Biotech and Chemical Engineering, Ulsan National Institute of Science and Technology, Ulsan, 689-798, [2]School of Computer Science and Engineering, Seoul National University, Seoul, 151-742, [3]Medical Genomics Research Center, Korea Research Institute for Bioscience and Biotechnology, Daejeon, 305-806 and [4]Department of Bioinformatics and Life Science, Soongsil University, Seoul, Korea 156-743

Received February 7, 2010; Revised April 24, 2010; Accepted May 6, 2010

ABSTRACT

Genome-wide association (GWA) study aims to identify the genetic factors associated with the traits of interest. However, the power of GWA analysis has been seriously limited by the enormous number of markers tested. Recently, the gene set analysis (GSA) methods were introduced to GWA studies to address the association of gene sets that share common biological functions. GSA considerably increased the power of association analysis and successfully identified coordinated association patterns of gene sets. There have been several approaches in this direction with some limitations. Here, we present a general approach for GSA in GWA analysis and a stand-alone software GSA-SNP that implements three widely used GSA methods. GSA-SNP provides a fast computation and an easy-to-use interface. The software and test datasets are freely available at http://gsa.muldas.org. We provide an exemplary analysis on adult heights in a Korean population.

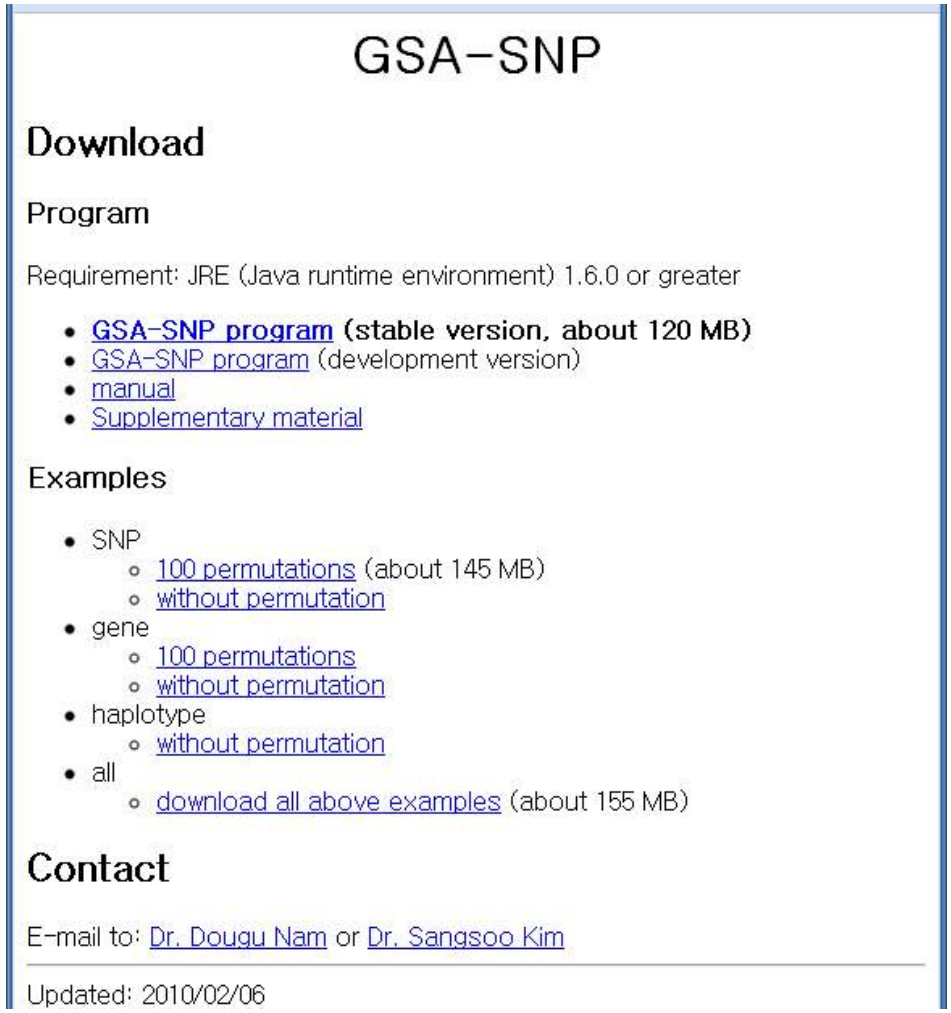
INTRODUCTION

Genome-wide association (GWA) study of a large population offers potential genetic causes of complex disease or the traits of interest (1,2). The typical approach assesses beyond individual markers or genes. Moreover, many of those prominent SNPs are not reproducible among independent experiments. Another important problem is that many moderate but meaningful associations are lost below the stringent cutoff. In recent years, the gene set analysis (GSA) methods were taken into account in GWA studies which may address these problems.

GSA methods were originally developed for a transcriptome analysis to assess the differential expression of pre-defined gene sets that share common biological functions. They exhibited stronger statistical power than the individual gene analysis, and have revealed many novel gene sets with 'subtle but coordinated' expression patterns (3–5). Given that the basic goal of GWA studies is to prioritize the biological networks or processes associated with the trait of interest, it may be reasonable to consider the pre-defined gene sets or pathways as the units of an association analysis. Indeed, by analyzing SNPs on the gene set level, GSA was able to reveal many coordinated association patterns that might be lost by the individual marker analysis.

Several case–control studies employed GSA methods. Wang *et al.* (6) devised a GSEA framework for SNP arrays. They assigned the most highly associated SNP (best SNP) to each gene to summarize the association of multiple SNPs in each gene. Using the method, they successfully identified the Parkinson's disease susceptibility pathways. Wang *et al.* (7) applied the same methods which implicated the molecular mechanism of autism

Program is freely downloadable from our web page



GSA-SNP

Download

Program

Requirement: JRE (Java runtime environment) 1.6.0 or greater

- GSA-SNP program (stable version, about 120 MB)
- GSA-SNP program (development version)
- manual
- Supplementary material

Examples

- SNP
  - 100 permutations (about 145 MB)
  - without permutation
- gene
  - 100 permutations
  - without permutation
- haplotype
  - without permutation
- all
  - download all above examples (about 155 MB)

Contact

E-mail to: Dr. Dougu Nam or Dr. Sangsoo Kim

Updated: 2010/02/06

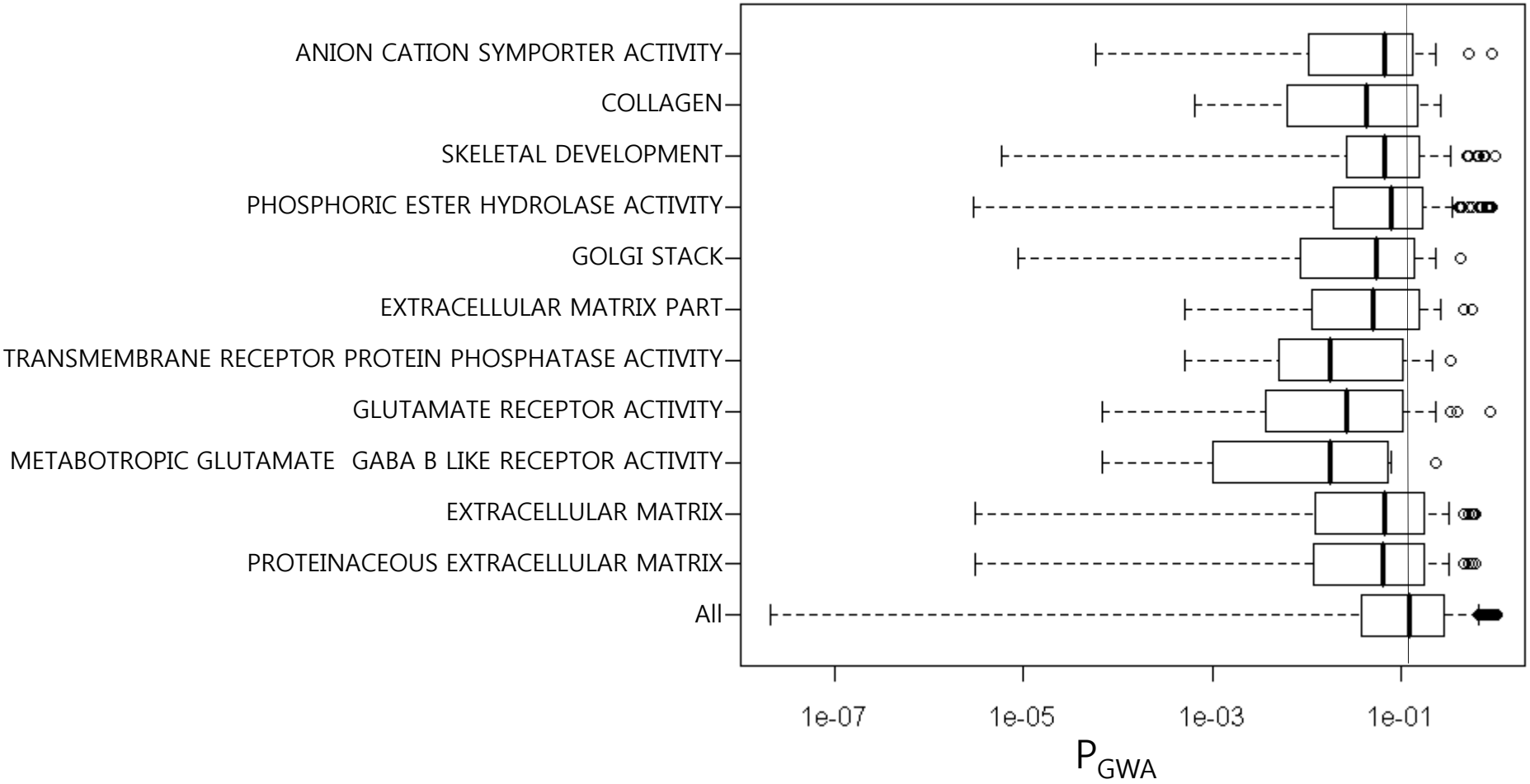
- Just type 'GSA-SNP' in google
- Program, tested data set, and user's manual are available

Korea Association Resource (KARE) project

- Affymetrix 5.0 genotypes on 10,004 individuals (ages 40~69)
- 352,228 SNPs passed QC
  - 38,364 markers violated HWE ($P < 10^{-6}$)
  - 17,926 genotype call rates < 95%
  - 92,050 MAF < 0.01
- 8,842 individuals passed QC
  - 11 sample contamination
  - 41 gender inconsistency
  - 608 cryptic relatedness
  - 101 serious concomitant illness

- Revisit by GSA-SNP

Moderate but consistent associations with *height* were detected in some Gene Ontology sets



Literature survey (height)

PROTEINACEOUS EXTRACELLULAR MATRIX	"A key biological function in height regulation" by Weedon et al. (2008) Genome -wide association analysis identifies 20 loci that influence adult height. Nat Gene t, 40, 575-583.
EXTRACELLULAR MATRIX	
METABOTROPIC GLUTAMATE GABA B LIKE RECEPTOR ACTIVITY	GRIA1, one of the members, was implicated near a loci associated with height in Croatian population. Endogenous activation of metabotropic glutamate receptor s is known to modulate GABAnergic transmission of gonadotropin-releasing hor mone (GnRH) neurons. Moreover, treatment with a GnRH agonist in short adoles cents increased adult height
GLUTAMATE RECEPTOR ACTIVITY	
TRANSMEMBRANE RECEPTOR PROTEIN PHOSPHATASE ACTIVITY	
EXTRACELLULAR MATRIX PART	Related to EXTRACELLULAR MATRIX
GOLGI STACK	
PHOSPHORIC ESTER HYDROLASE ACTIVITY	
SKELETAL DEVELOPMENT	Gudbjartsson et al. (2008) Many sequence variants affecting diversity of adult hu man height. Nat Genet, 40, 609-615.
COLLAGEN	The most abundant proteins in ECM
ANION CATION SYMPORTER ACTIVITY	

Acknowledgements

- Dougu Nam (UNIST)
  - Jin Kim (SNU)
  - Seon-Young Kim (KRIBB)
- Ji-sun Kwon (SSU)
- KARE Consortium
- **KCDC NIH GRC for providing KARE data**
- **KOBIC for a PC cluster and storage**
- **NRF for funding**