An Introduction to Genome Wide Association Studies (GWAS)

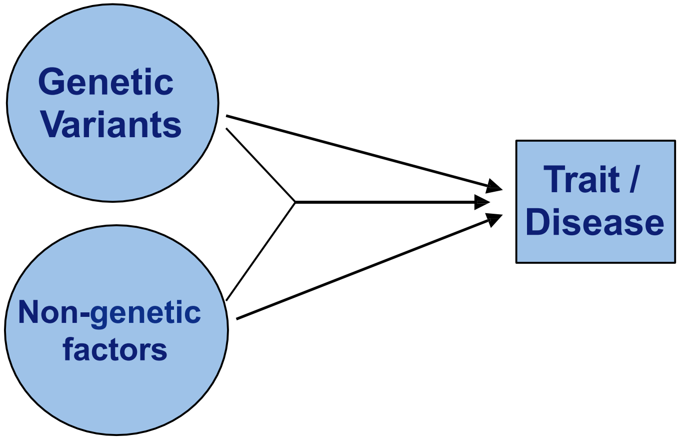
Evangelos Evangelou

✉ vangelis@cc.uoi.gr

eevangelou

Complex Traits: Multifactorial Inheritance

- Complex traits/disorders vs. Mendelian inherited disorders
- Complex disorders:
  - No Mendelian mode of inheritance
  - Multiple susceptibility loci
  - Incomplete penetrance
  - Major environmental risk factors
- Public health importance

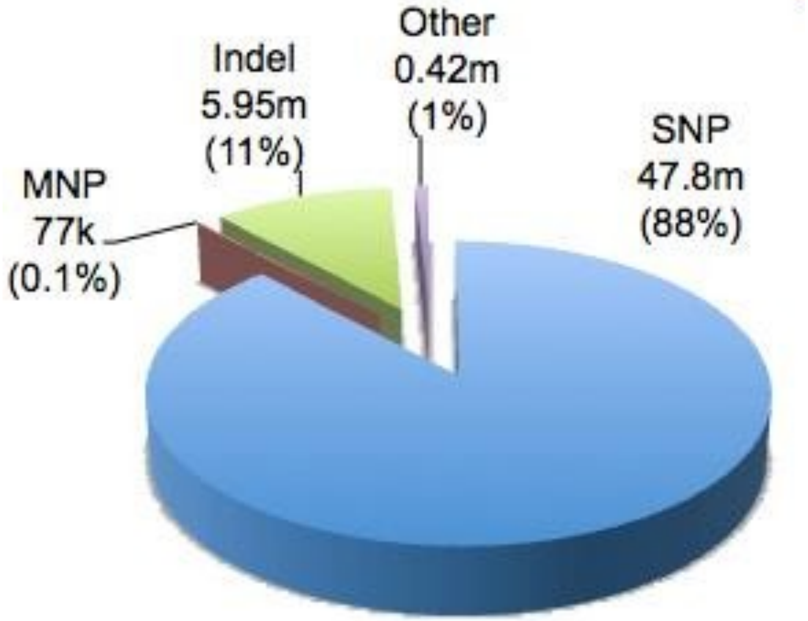


Types of genetic variations

- Copy number variations (CNVs): Interindividual variations in the number of copies of a specific gene or chromosomal region.

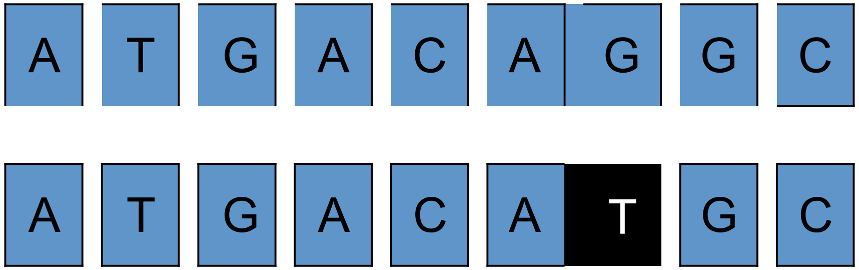
- Insertions and deletions (Indels): Regions of DNA that are either inserted into or deleted from the genome.

- Single nucleotide polymorphisms (SNPs): Single base pair changes in the genome in a population.



Single Nucleotide Polymorphisms: SNPs

- SNPs – DNA sequence variations that occur when a single nucleotide is changed



A T G A C A G G C

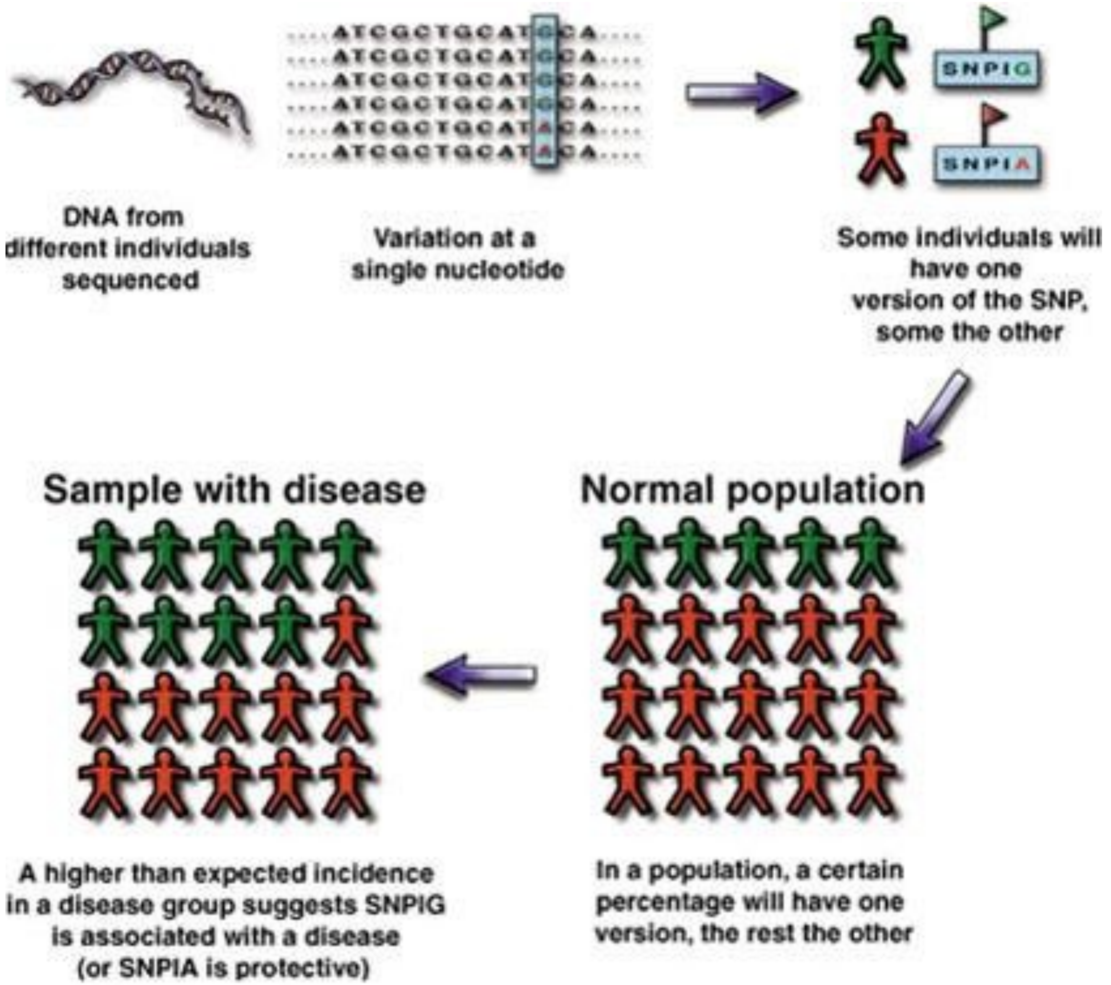
A T G A C A T G C

- Alleles at this SNP are “G” and “T”

- SNPs are the most common form of variation in the human genome

- SNPs are catalogued in several databases

Using SNPs to Track Predisposition to Disease



DNA from different individuals sequenced

Variation at a single nucleotide

Some individuals will have one version of the SNP, some the other

Sample with disease

Normal population

A higher than expected incidence in a disease group suggests SNPIG is associated with a disease (or SNPIA is protective)

In a population, a certain percentage will have one version, the rest the other

© Gibson & Muse, A Primer of Genome Science

© Francis Collins, 2008

Scope of a Genetic Association Study

- Candidate gene
  - Known functional variants
  - Variants with unknown function in exons, regulatory regions

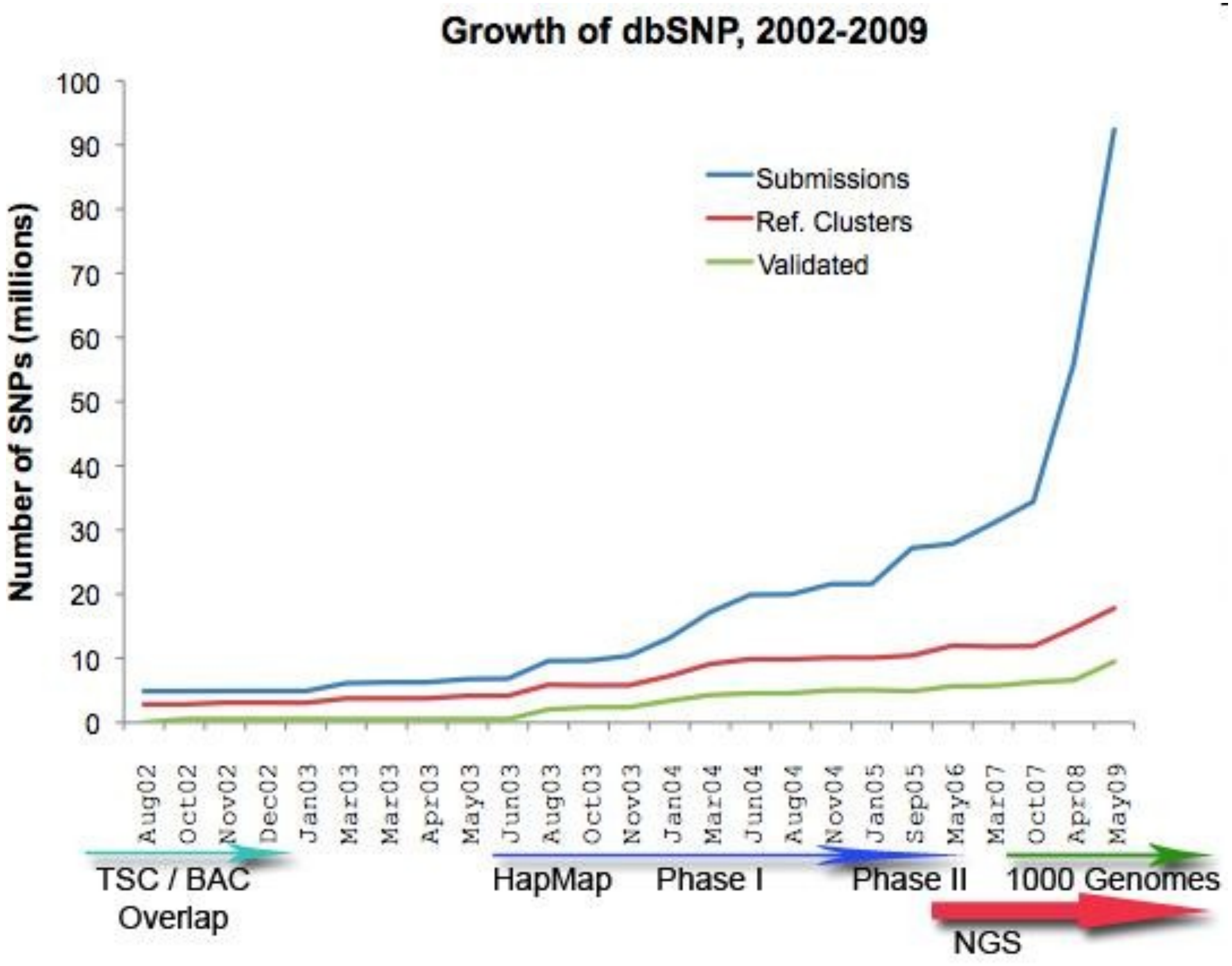
- Genome-wide
  - Test for association with hundreds of thousands (millions) of SNPs spread across the entire genome.
  - Many design strategies possible for distributing markers

Genome-Wide Association Studies

- Candidate-gene association
  - Greater power to identify smaller genetic effects
  - Rely on a priori knowledge about disease etiology
  - Low replication rate.

- Genome-wide association studies
  - Agnostic search
  - Needs large sample size
  - Robust findings

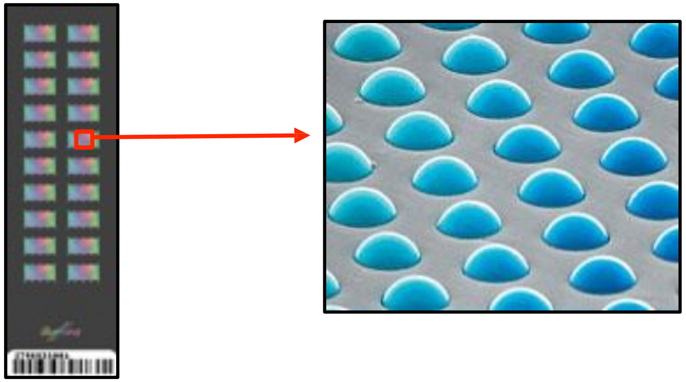
How many SNPs should be studied?



Costs of a Genome-wide association study in 2,000 individuals

Year	Number of SNPs	Costs per SNP	Total costs
2001	10,000,000	$ 1.00	$20,000,000,000

Microarray technology



clinical
research
facility
EDINBURGH

SNP Chips: Number and Placement of SNPs

- A “typical” SNP chip has at least 300,000 SNPs distributed across the genome. Nowadays even >1 million.

- The new chips can also measure some types of copy number variation.

Coverage and efficiency in current SNP chips

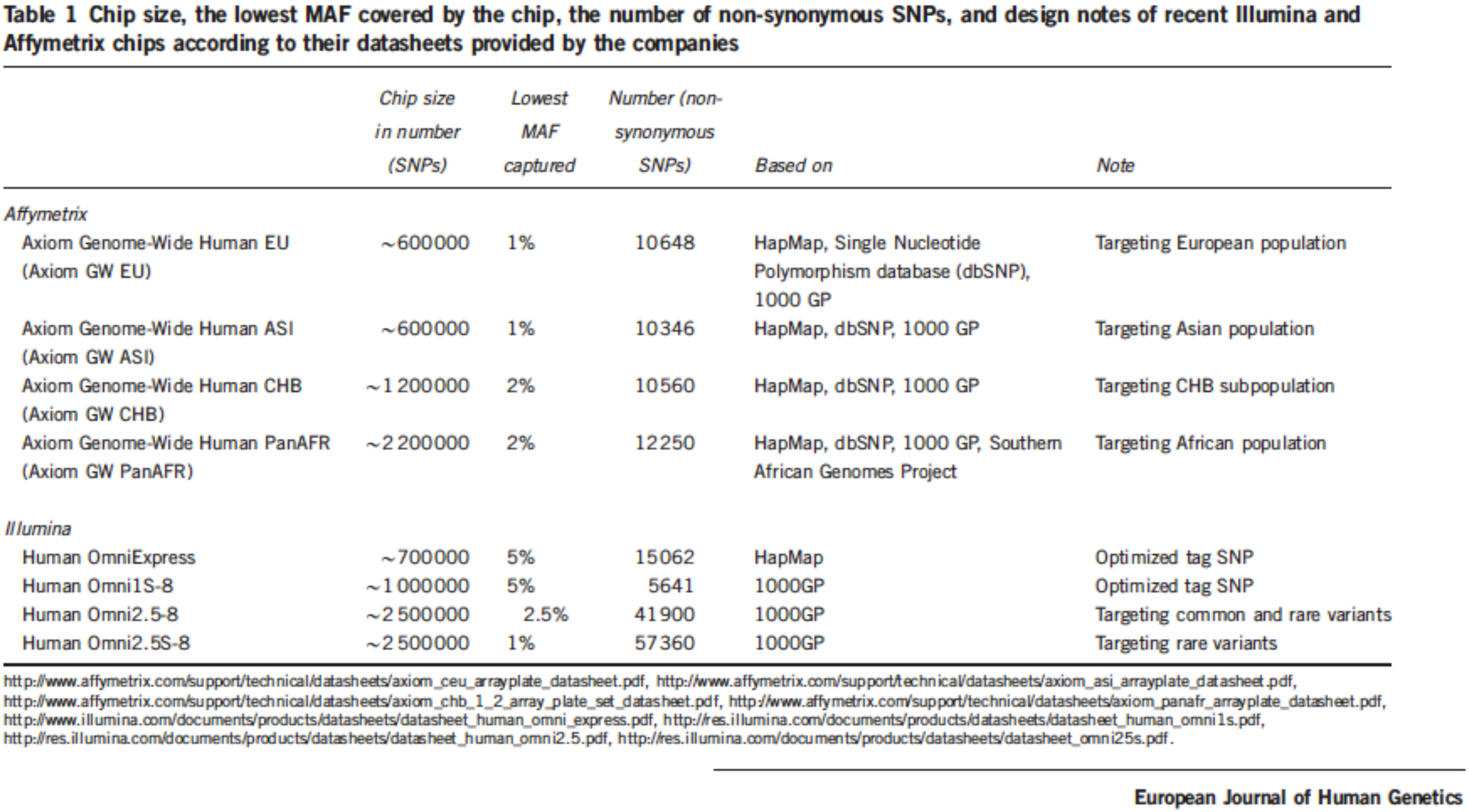
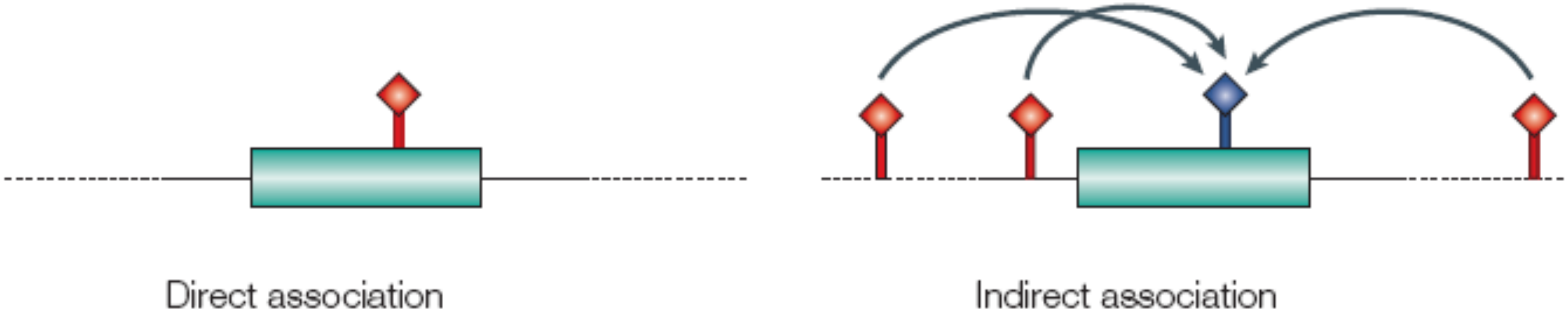


Table 1 Chip size, the lowest MAF covered by the chip, the number of non-synonymous SNPs, and design notes of recent Illumina and Affymetrix chips according to their datasheets provided by the companies

	Chip size in number (SNPs)	*Lowest MAF captured*	*Number (non-synonymous SNPs)*	*Based on*	*Note*
Affymetrix					
Axiom Genome-Wide Human EU (Axiom GW EU)	~600000	1%	10648	HapMap, Single Nucleotide Polymorphism database (dbSNP), 1000 GP	Targeting European population
Axiom Genome-Wide Human ASI (Axiom GW ASI)	~600000	1%	10346	HapMap, dbSNP, 1000 GP	Targeting Asian population
Axiom Genome-Wide Human CHB (Axiom GW CHB)	~1200000	2%	10560	HapMap, dbSNP, 1000 GP	Targeting CHB subpopulation
Axiom Genome-Wide Human PanAFR (Axiom GW PanAFR)	~2200000	2%	12250	HapMap, dbSNP, 1000 GP, Southern African Genomes Project	Targeting African population
Illumina					
Human OmniExpress	~700000	5%	15062	HapMap	Optimized tag SNP
Human Omni1S-8	~1000000	5%	5641	1000GP	Optimized tag SNP
Human Omni2.5-8	~2500000	2.5%	41900	1000GP	Targeting common and rare variants
Human Omni2.5S-8	~2500000	1%	57360	1000GP	Targeting rare variants

http://www.affymetrix.com/support/technical/datasheets/axiom_ceu_arrayplate_datasheet.pdf, http://www.affymetrix.com/support/technical/datasheets/axiom_asi_arrayplate_datasheet.pdf, http://www.affymetrix.com/support/technical/datasheets/axiom_chb_1_2_array_plate_set_datasheet.pdf, http://www.affymetrix.com/support/technical/datasheets/axiom_panafr_arrayplate_datasheet.pdf, http://www.illumina.com/documents/products/datasheets/datasheet_human_omni_express.pdf, http://res.illumina.com/documents/products/datasheets/datasheet_human_omni1s.pdf, http://res.illumina.com/documents/products/datasheets/datasheet_human_omni2.5.pdf, http://res.illumina.com/documents/products/datasheets/datasheet_omni25s.pdf.

Can we skip some of the SNPs?



Hirschhorn & Daly, Nat Rev Genet 2005

Linkage Disequilibrium (LD)

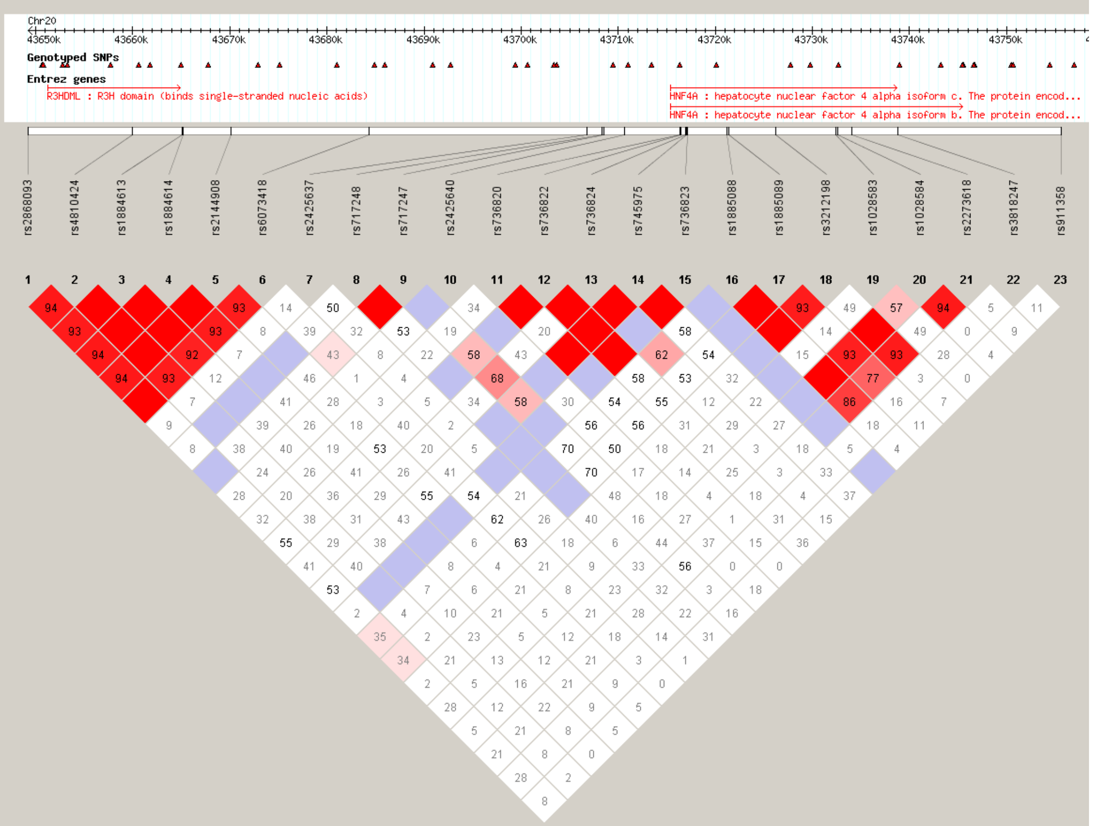
- LD is the correlation between SNPs

- LD is observed in various regions of the genome, not only nearby the genes causing the diseases or in coding regions

- Measure of LD: ***r^2*, *D'***
- ***r^2*** gets values from 0 to 1; 0 denotes independent variants whereas 1 denotes that variants are in total LD

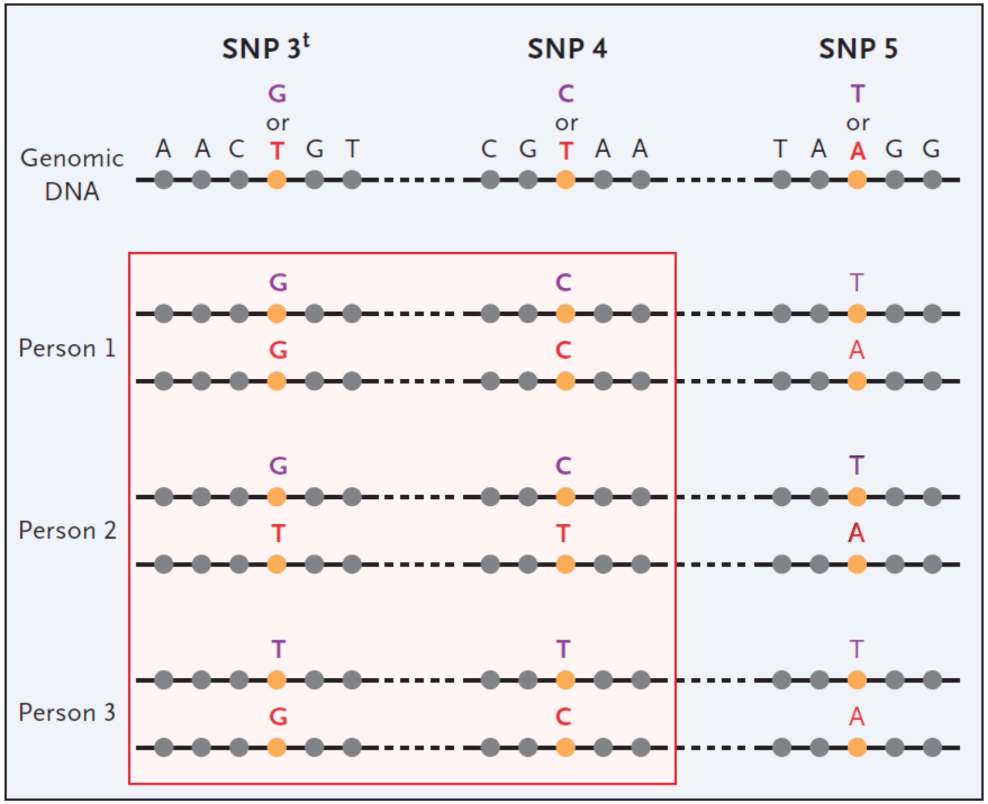
Linkage Disequilibrium (LD)

- LD varies depending on region of genome
- LD between two SNPs decreases with distance



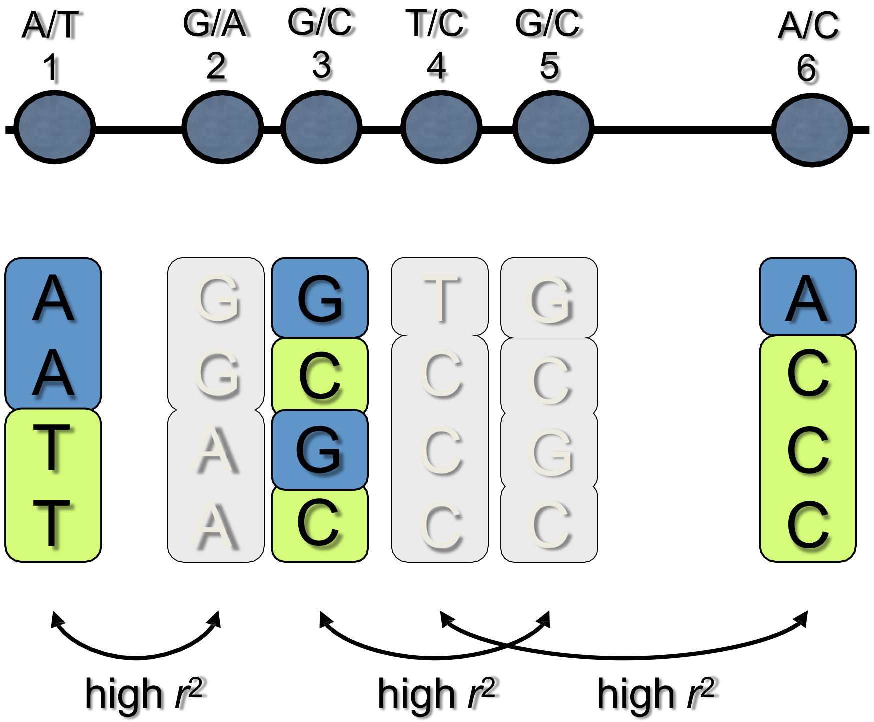
LD and Proxy

- Due to LD, one SNP may serve as proxy for others



Christensen and Murray, N Engl J Med 2007; 356:1094-1097

Can one SNP tag others?



Tags:

SNP 1
SNP 3
SNP 6

3 in total

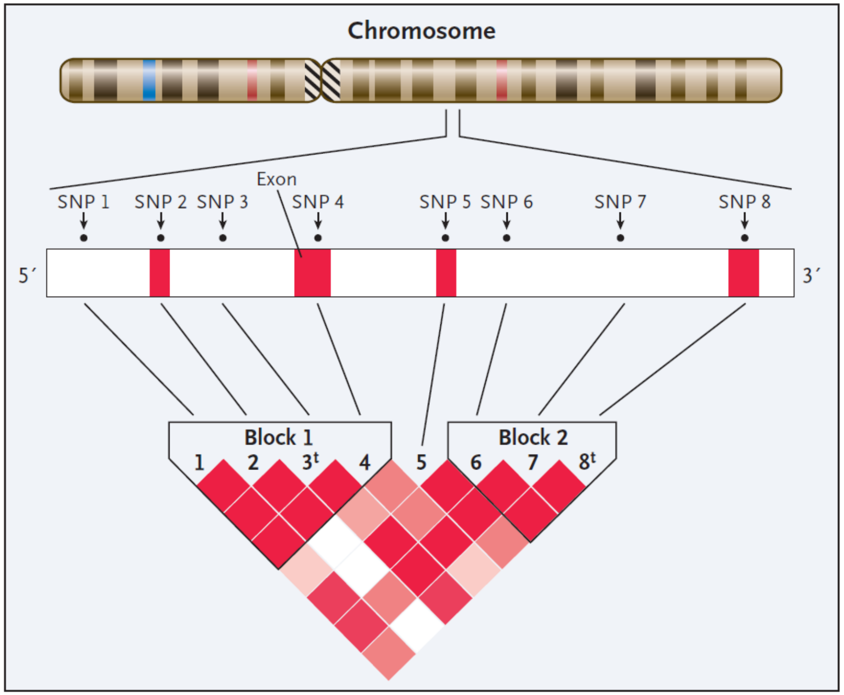
Test for association:

SNP 1
SNP 3
SNP 6

After Carlson *et al.* (2004) *AJHG* **74**:106

Map of the tagging SNPs

- Map of the relationships among SNPs is useful
- Such a map varies by ethnic groups



Christensen and Murray, N Engl J Med 2007; 356:1094-1097

www.hapmap.org

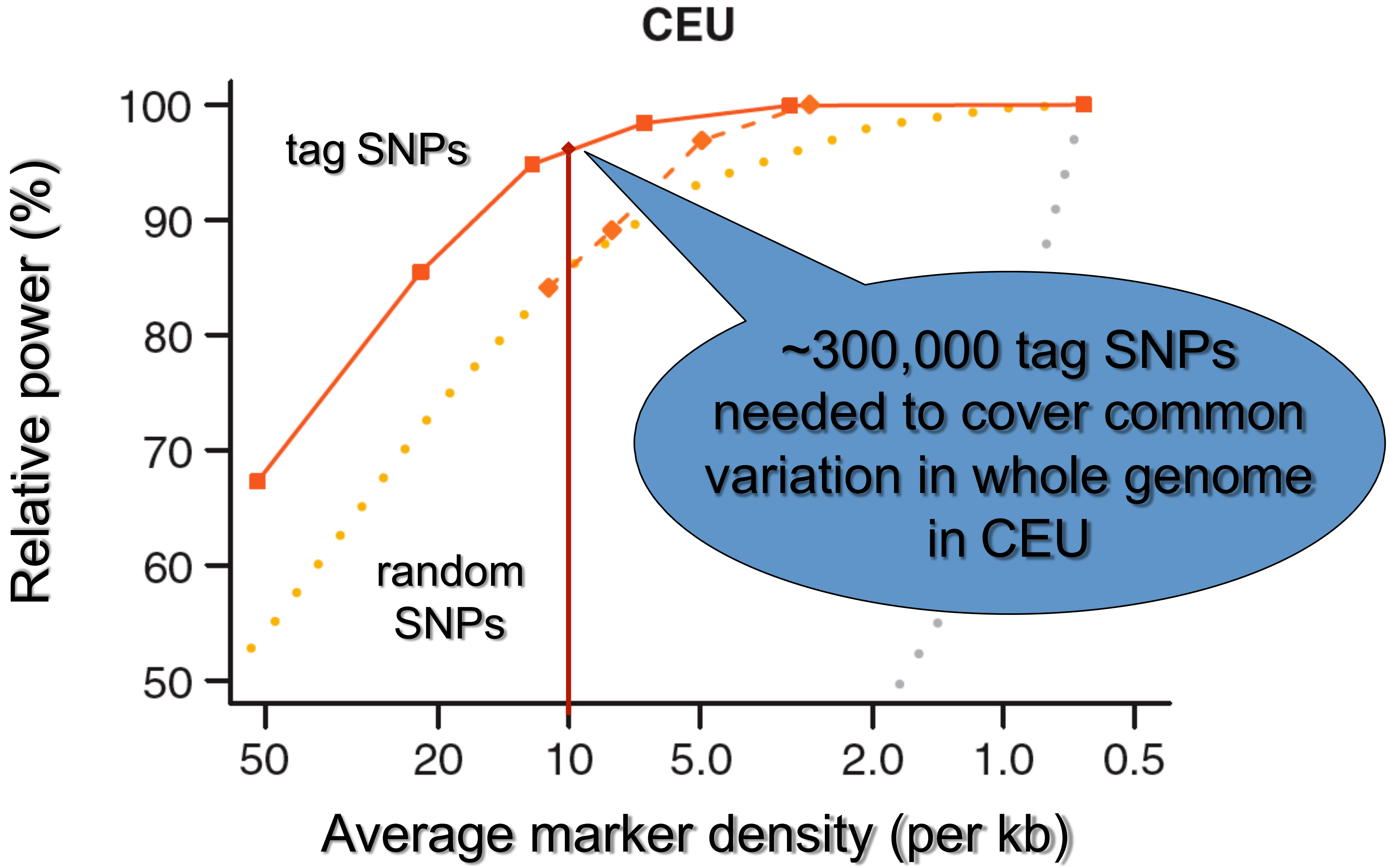
Genomic information in mapping complex disease genes

2001

2009

2015

Efficiency and power



~300,000 tag SNPs needed to cover common variation in whole genome in CEU

P.I.W. de Bakker *et al.* (2005) *Nat Genet*

Why are They Possible Now?

- Genotyping Technology:
  - Now have ability to type hundreds of thousands (or millions) of SNPs in one reaction on a “SNP chip.”
  - The cost can be as low as £19 per person.
- Design and analysis:
  - Availability of SNP databases, HapMap, 1000KG and other resources to identify the SNPs and design SNP chips.
  - Faster computers to carry out the millions of calculations make implementation possible.

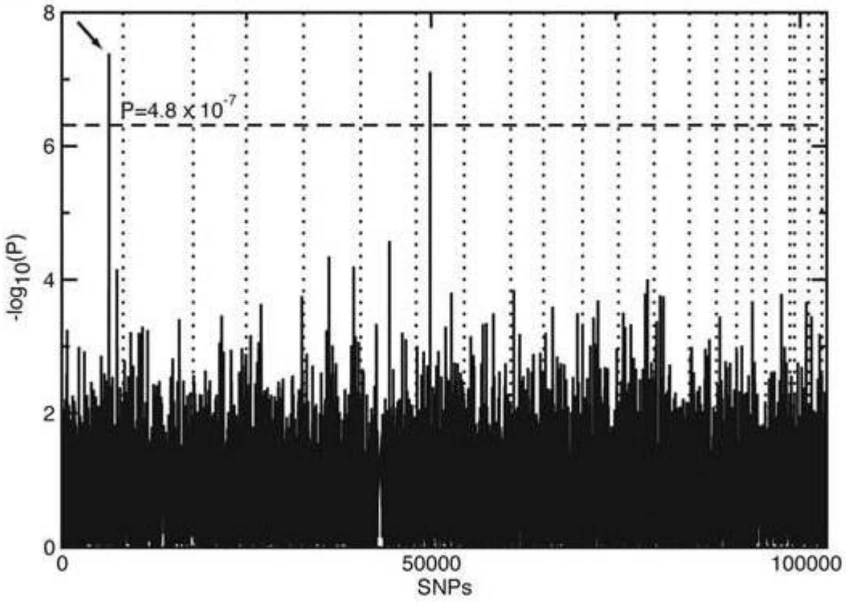
Costs of a Genome-wide association study in 2,000 individuals

Year	Number of SNPs	Costs per SNP	Total costs
2001	10,000,000	$ 1.00	$20,000,000,000
2007	500,000	$ 0.001	$1,000,000

First GWAs in 2005

- The first successful GWA study was published in 2005

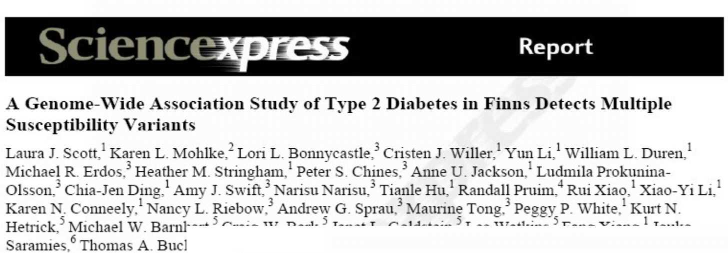
- Among 116,204 single-nucleotide polymorphisms genotyped, an intronic and common variant in the complement factor H gene (CFH) was strongly associated with AMD



Klein Rj et al. Science 2005

T2D GWA studies in 2007

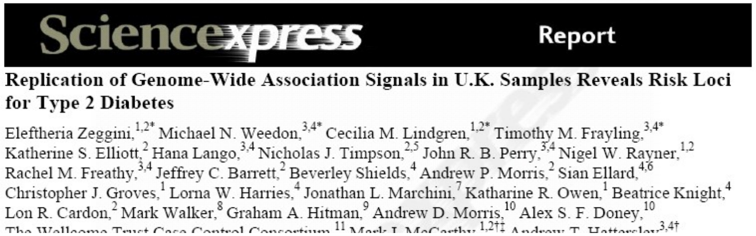
- By the end of 2007 from a total of 9 genes 6 were described in GWA studies for T2D



Sciencexpress Report

A Genome-Wide Association Study of Type 2 Diabetes in Finns Detects Multiple Susceptibility Variants

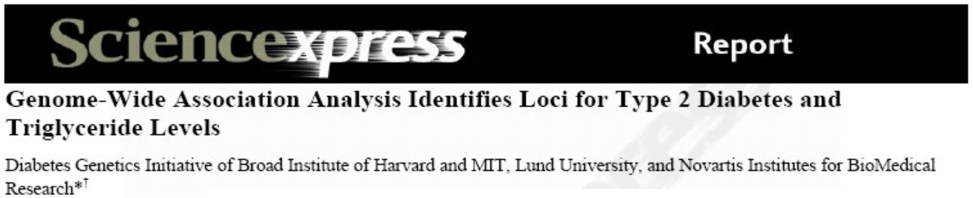
Laura J. Scott,[1] Karen L. Mohlke,[2] Lori L. Bonnycastle,[3] Cristen J. Willer,[1] Yun Li,[1] William L. Duren,[1] Michael R. Erdos,[3] Heather M. Stringham,[1] Peter S. Chines,[3] Anne U. Jackson,[1] Ludmila Prokunina-Olsson,[3] Chia-Jen Ding,[1] Amy J. Swift,[3] Narisu Narisu,[3] Tianle Hu,[1] Randall Pruim,[4] Rui Xiao,[1] Xiao-Yi Li,[1] Karen N. Conneely,[1] Nancy L. Riebow,[3] Andrew G. Sprau,[3] Maurine Tong,[3] Peggy P. White,[1] Kurt N. Hetrick,[5] Michael W. Barnhart,[5] Craig W. Bark,[5] Janet L. Goldstein,[5] Lee Watkins,[5] Fang Xiang,[1] Jouko Saramies,[6] Thomas A. Buc
R. Abecasis,[1] Elizabeth W.
Francis S. Collins,[3]* Micha



Sciencexpress Report

Replication of Genome-Wide Association Signals in U.K. Samples Reveals Risk Loci for Type 2 Diabetes

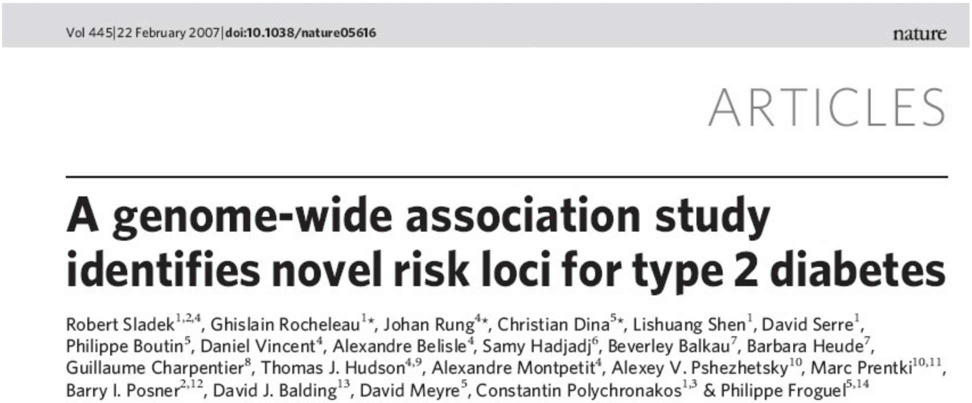
Eleftheria Zeggini,[1,2*] Michael N. Weedon,[3,4*] Cecilia M. Lindgren,[1,2*] Timothy M. Frayling,[3,4*] Katherine S. Elliott,[2] Hana Lango,[3,4] Nicholas J. Timpson,[2,5] John R. B. Perry,[3,4] Nigel W. Rayner,[1,2] Rachel M. Freathy,[3,4] Jeffrey C. Barrett,[2] Beverley Shields,[4] Andrew P. Morris,[2] Sian Ellard,[4,6] Christopher J. Groves,[1] Lorna W. Harries,[4] Jonathan L. Marchini,[7] Katharine R. Owen,[1] Beatrice Knight,[4] Lon R. Cardon,[2] Mark Walker,[8] Graham A. Hitman,[9] Andrew D. Morris,[10] Alex S. F. Doney,[10] The Wellcome Trust Case Control Consortium,[11] Mark I. McCarthy,[1,2†‡] Andrew T. Hattersley[3,4†]



Sciencexpress Report

Genome-Wide Association Analysis Identifies Loci for Type 2 Diabetes and Triglyceride Levels

Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes for BioMedical Research*[†]

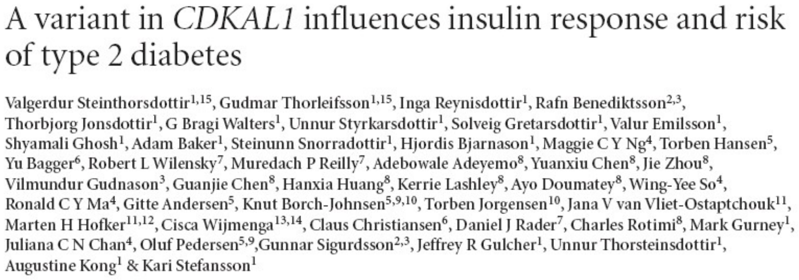


Vol 445|22 February 2007|**doi:10.1038/nature05616** nature

ARTICLES

A genome-wide association study identifies novel risk loci for type 2 diabetes

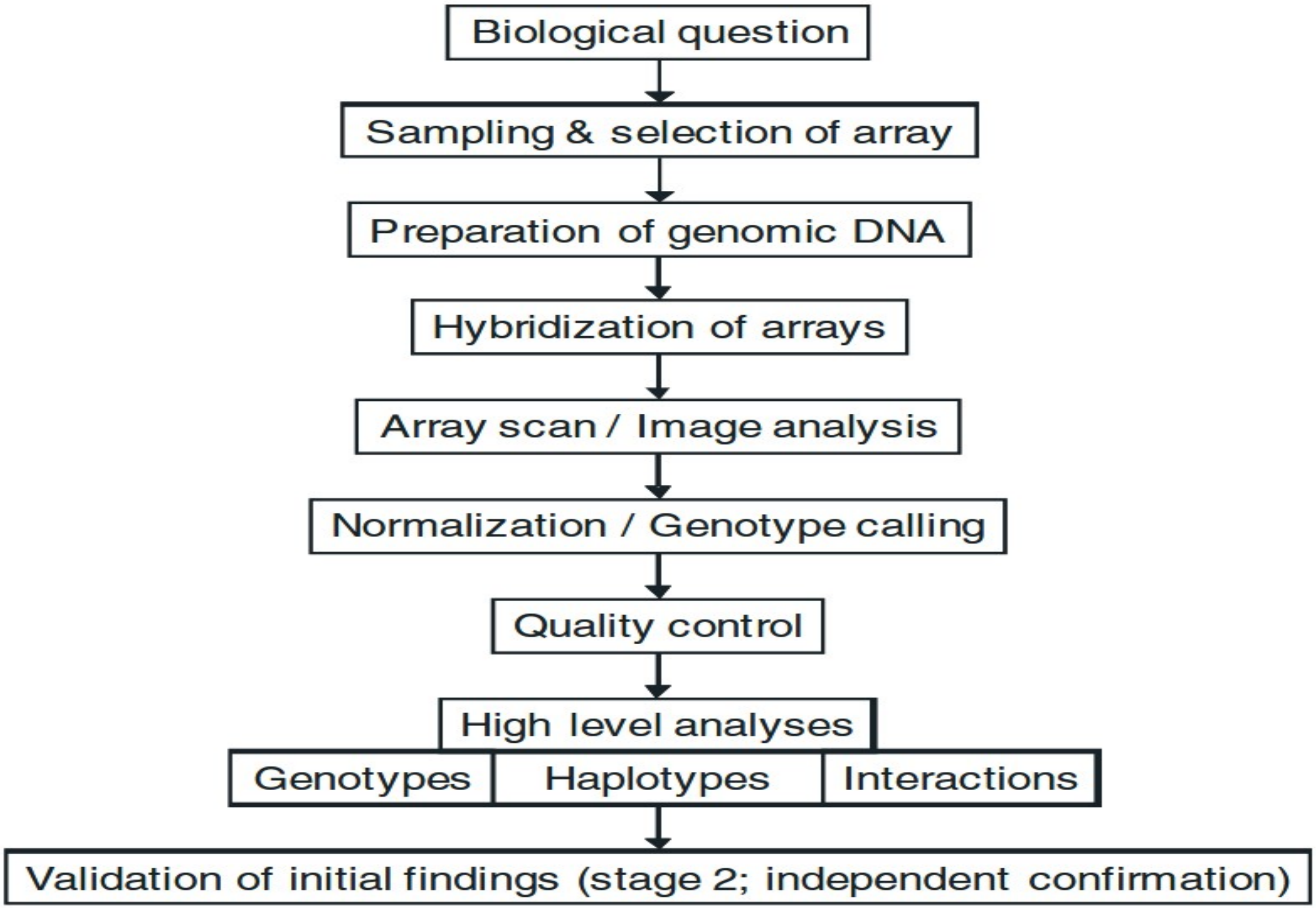
Robert Sladek[1,2,4], Ghislain Rocheleau[1]*, Johan Rung[4]*, Christian Dina[5]*, Lishuang Shen[1], David Serre[1], Philippe Boutin[5], Daniel Vincent[4], Alexandre Belisle[4], Samy Hadjadj[6], Beverley Balkau[7], Barbara Heude[7], Guillaume Charpentier[8], Thomas J. Hudson[4,9], Alexandre Montpetit[4], Alexey V. Pshezhetsky[10], Marc Prentki[10,11], Barry I. Posner[2,12], David J. Balding[13], David Meyre[5], Constantin Polychronakos[1,3] & Philippe Froguel[5,14]



A variant in *CDKAL1* influences insulin response and risk of type 2 diabetes

Valgerdur Steinthorsdottir[1,15], Gudmar Thorleifsson[1,15], Inga Reynisdottir[1], Rafn Benediktsson[2,3], Thorbjorg Jonsdottir[1], G Bragi Walters[1], Unnur Styrkarsdottir[1], Solveig Gretarsdottir[1], Valur Emilsson[1], Shyamali Ghosh[1], Adam Baker[1], Steinunn Snorradottir[1], Hjordis Bjarnason[1], Maggie C Y Ng[4], Torben Hansen[5], Yu Bagger[6], Robert L Wilensky[7], Muredach P Reilly[7], Adebowale Adeyemo[8], Yuanxiu Chen[8], Jie Zhou[8], Vilmundur Gudnason[3], Guanjie Chen[8], Hanxia Huang[8], Kerrie Lashley[8], Ayo Doumatey[8], Wing-Yee So[4], Ronald C Y Ma[4], Gitte Andersen[5], Knut Borch-Johnsen[5,9,10], Torben Jorgensen[10], Jana V van Vliet-Ostaptchouk[11], Marten H Hofker[11,12], Cisca Wijmenga[13,14], Claus Christiansen[6], Daniel J Rader[7], Charles Rotimi[8], Mark Gurney[1], Juliana C N Chan[4], Oluf Pedersen[5,9],Gunnar Sigurdsson[2,3], Jeffrey R Gulcher[1], Unnur Thorsteinsdottir[1], Augustine Kong[1] & Kari Stefansson[1]

Steps for conducting a GWAS



Biological question

Sampling & selection of array

Preparation of genomic DNA

Hybridization of arrays

Array scan / Image analysis

Normalization / Genotype calling

Quality control

High level analyses

Genotypes

Haplotypes

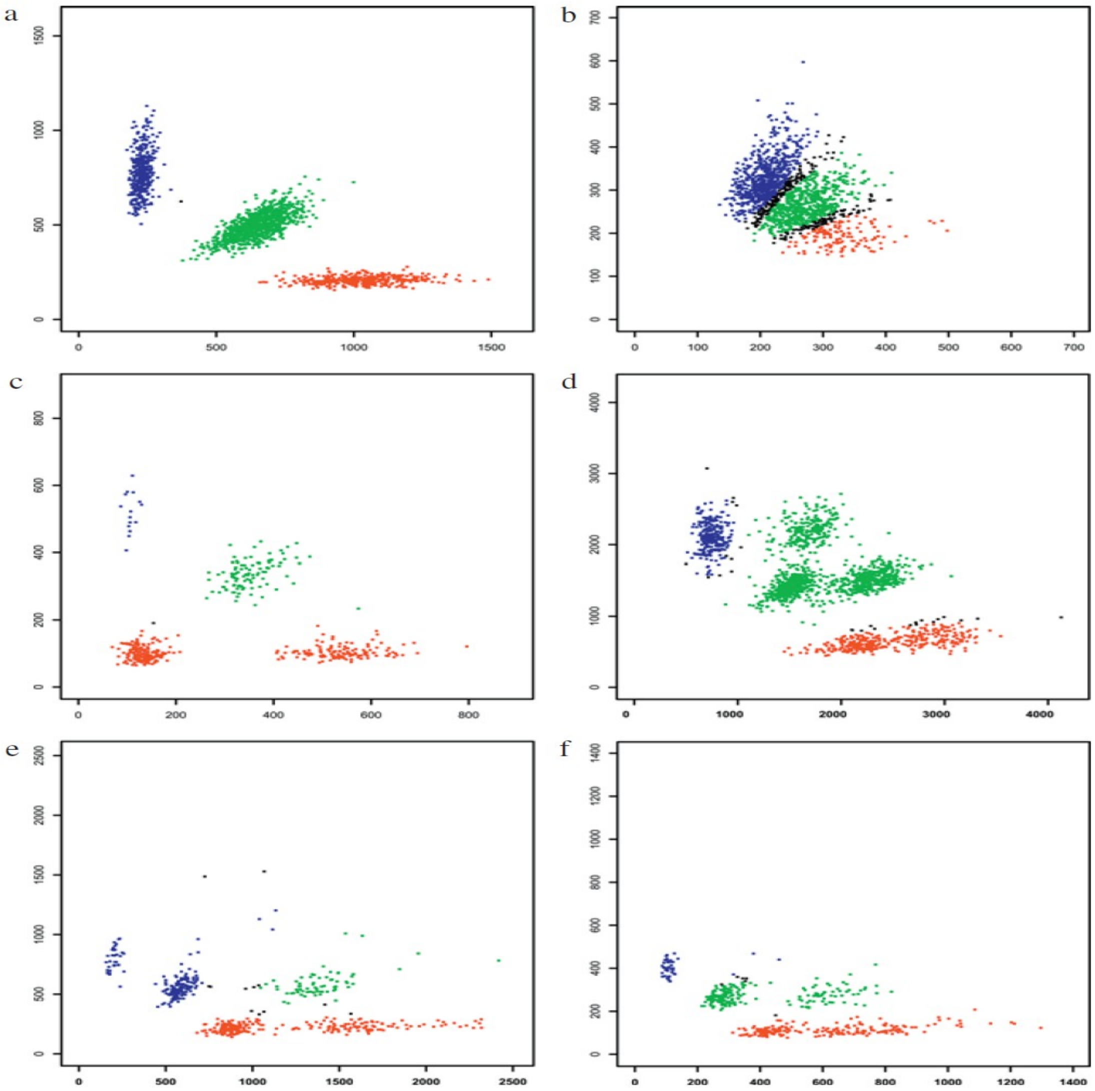
Interactions

Validation of initial findings (stage 2; independent confirmation)

Figure 1 Succession of design, experimental and data analysis steps in a genome-wide association study.

Ziegler A, et al. Biometrical J 2008;8-

Calling of genotypes



Ziegler A, et al. Biometrical J 2008;8-28

GWA QC procedure steps (1)

- Genotype call rate (i.e., assignment of genotypes to subjects): 95% cut-off for missing data for each SNP

- Reproducibility across genotyping platforms and technologies: 99% within platform, 95% across platforms

- MAF: threholds based on interest, imputation quality etc

- HWE in the controls: exclude SNPs if P<10-6

GWA QC procedure steps (2)

- Sample call rate: exclude subjects with many SNPs missing (e.g., >10%)
- Autosomal heterozygosity
- Relatedness check
- Gender check

Imputation of SNPs

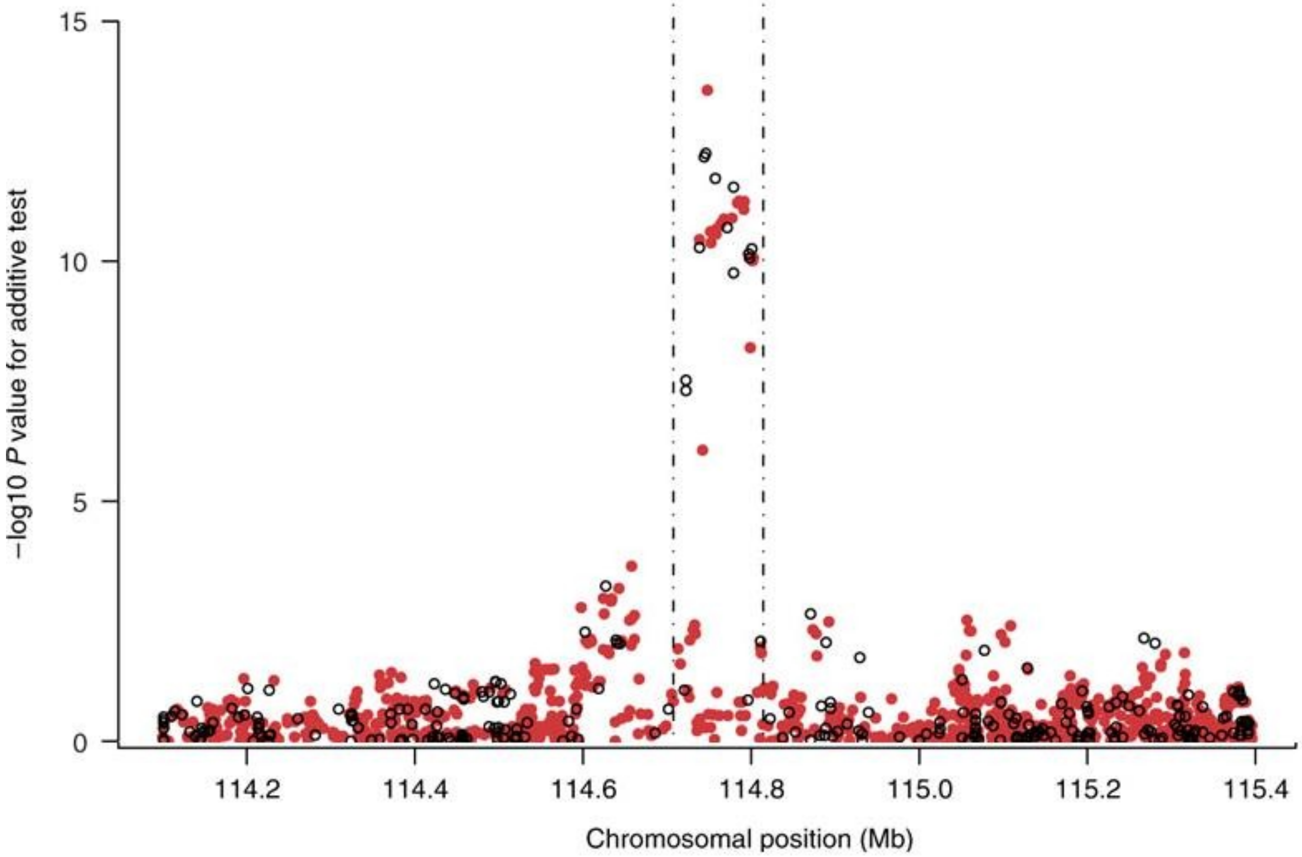
- Genotyping arrays include a limited number of SNPs

- Imputation is to estimate the unmeasured or missing SNPs

- Estimation is based on measured SNPs and external info

- Why imputation?
  - Increase GWAS power
  - Improves fine-mapping
  - Imputes Indels
  - Allow for combining data across different platforms (e.g., Affy & Illumina) (for replication / meta-analysis)

Imputation increases the power



TCF7L2 gene region & T2D from the WTCCC data

Imputation Example

Observed Genotypes

.	.	.	.	A	.	.	.	.	.	.	.	A	.	.	.	.	A	.	.	.
.	.	.	.	G	.	.	.	.	.	.	.	C	.	.	.	.	A	.	.	.

Study Sample

Reference Haplotypes

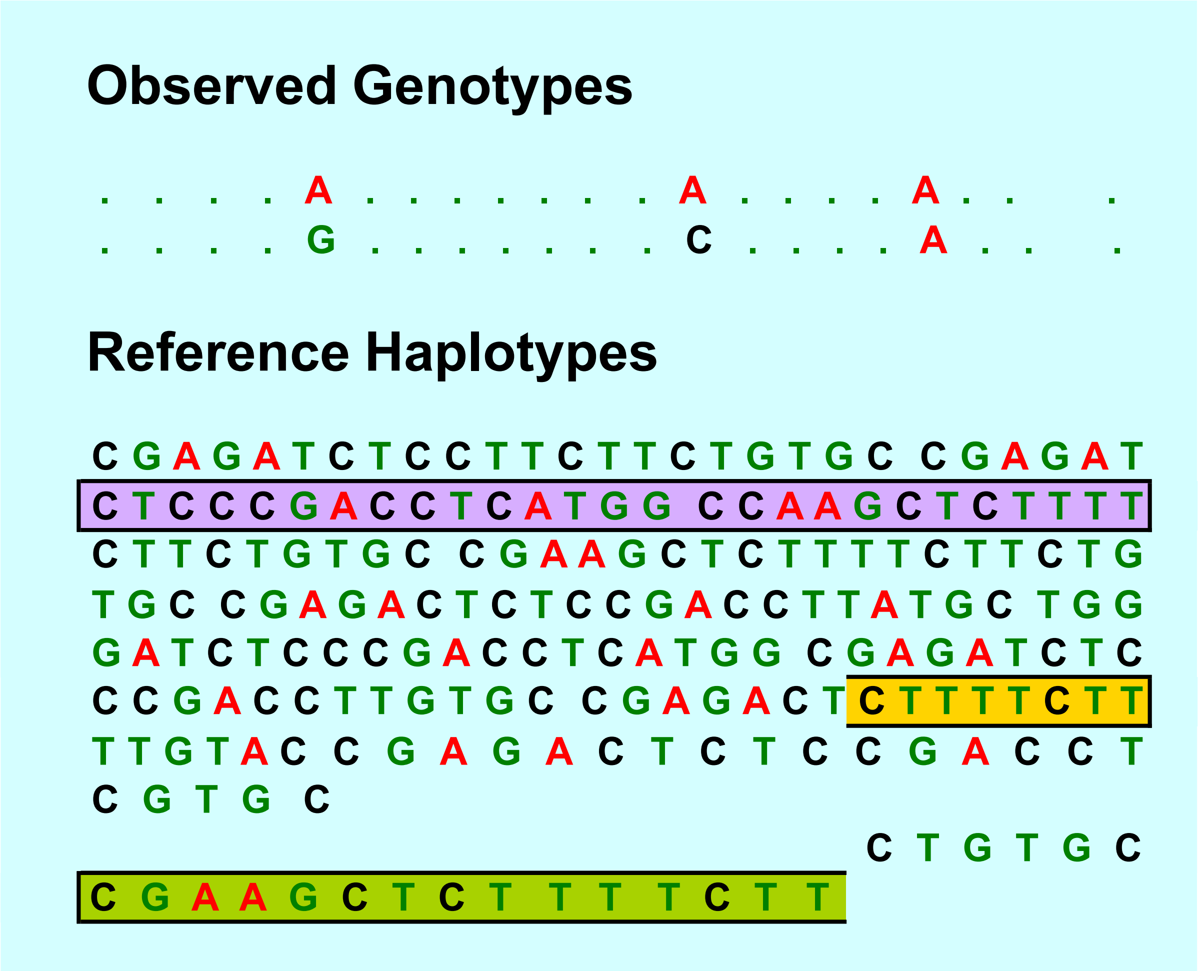
C	G	A	G	A	T	C	T	C	C	T	T	C	T	T	C	T	G	T	G	C
C	G	A	G	A	T	C	T	C	C	C	G	A	C	C	T	C	A	T	G	G
C	C	A	A	G	C	T	C	T	T	T	T	C	T	T	C	T	G	T	G	C
C	G	A	A	G	C	T	C	T	T	T	T	C	T	T	C	T	G	T	G	C
C	G	A	G	A	C	T	C	T	C	C	G	A	C	C	T	T	A	T	G	C
T	G	G	G	A	T	C	T	C	C	C	G	A	C	C	T	C	A	T	G	G
C	G	A	G	A	T	C	T	C	C	C	G	A	C	C	T	T	G	T	G	C
C	G	A	G	A	C	T	C	T	T	T	T	C	T	T	T	T	G	T	A	C
C	G	A	G	A	C	T	C	T	C	C	G	A	C	C	T	C	G	T	G	C
C	G	A	A	G	C	T	C	T	T	T	T	C	T	T	C	T	G	T	G	C

Reference Panel

Gonçalo Abecasis

http://www.sph.umich.edu/csg/abecasis/MACH

Identify Match with Reference



Observed Genotypes

. . . . A . . . . A . . . . A . . .

. . . . G . . . . C . . . . A . . .

Reference Haplotypes

C G A G A T C T C C T T C T T C T G T G C C G A G A T

C T C C C G A C C T C A T G G C C A A G C T C T T T T

C T T C T G T G C C G A A G C T C T T T T C T T C T G

T G C C G A G A C T C T C C G A C C T T A T G C T G G

G A T C T C C C G A C C T C A T G G C G A G A T C T C

C C G A C C T T G T G C C G A G A C T C T T T T C T T

T T G T A C C G A G A C T C T C C G A C C T

C G T G C

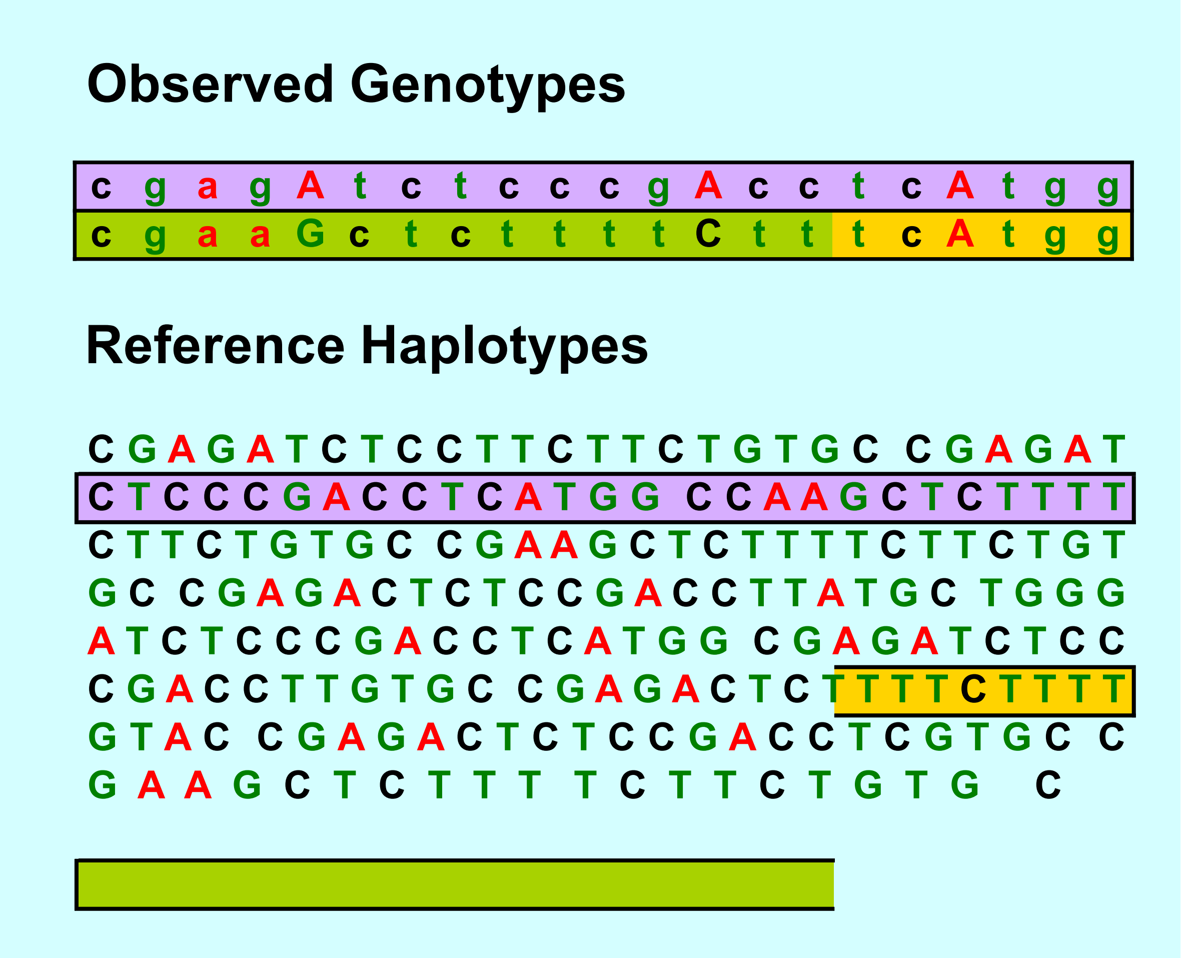
C T G T G C

C G A A G C T C T T T T C T T

Gonçalo Abecasis

http://www.sph.umich.edu/csg/abecasis/MACH

Phase chromosomes, impute missing genotypes



Observed Genotypes

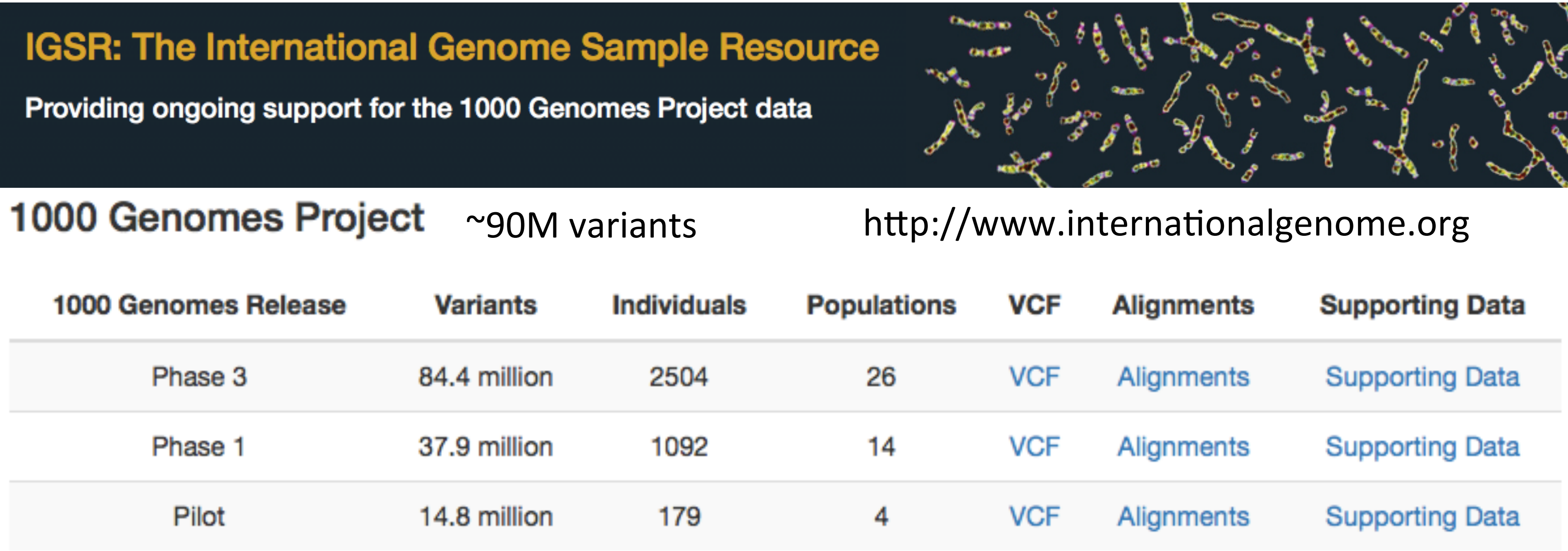
c g a g A t c t c c c g A c c t c A t g g
c g a a G c t c t t t t C t t t c A t g g

Reference Haplotypes

C G A G A T C T C C T T C T T C T G T G C C G A G A T
C T C C C G A C C T C A T G G C C A A G C T C T T T T
C T T C T G T G C C G A A G C T C T T T T C T T C T G T
G C C G A G A C T C T C C G A C C T T A T G C T G G G
A T C T C C C G A C C T C A T G G C G A G A T C T C C
C G A C C T T G T G C C G A G A C T C T T T T C T T T T
G T A C C G A G A C T C T C C G A C C T C G T G C C
G A A G C T C T T T T C T T C T G T G C

Gonçalo Abecasis http://www.sph.umich.edu/csg/abecasis/MACH

Reference Panels



IGSR: The International Genome Sample Resource

Providing ongoing support for the 1000 Genomes Project data

1000 Genomes Project ~90M variants http://www.internationalgenome.org

1000 Genomes Release	Variants	Individuals	Populations	VCF	Alignments	Supporting Data
Phase 3	84.4 million	2504	26	VCF	Alignments	Supporting Data
Phase 1	37.9 million	1092	14	VCF	Alignments	Supporting Data
Pilot	14.8 million	179	4	VCF	Alignments	Supporting Data

The Haplotype Reference Consortium

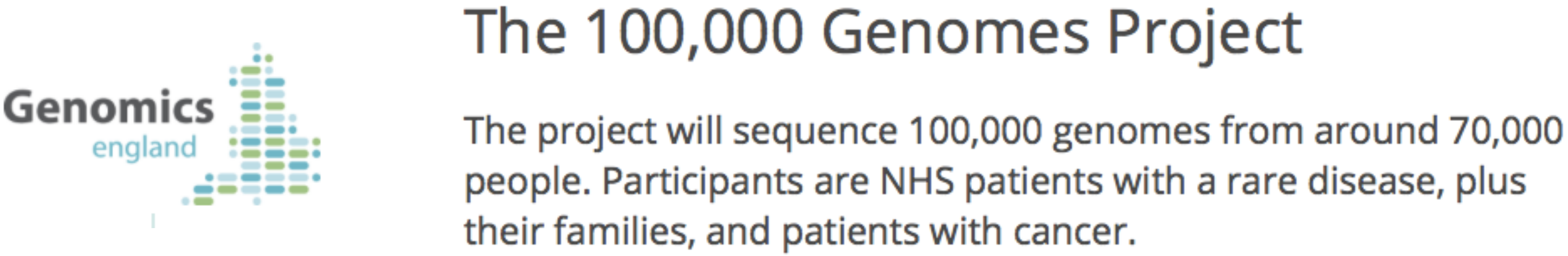
~39M variants

http://www.haplotype-reference-consortium.org

Other Reference Panels

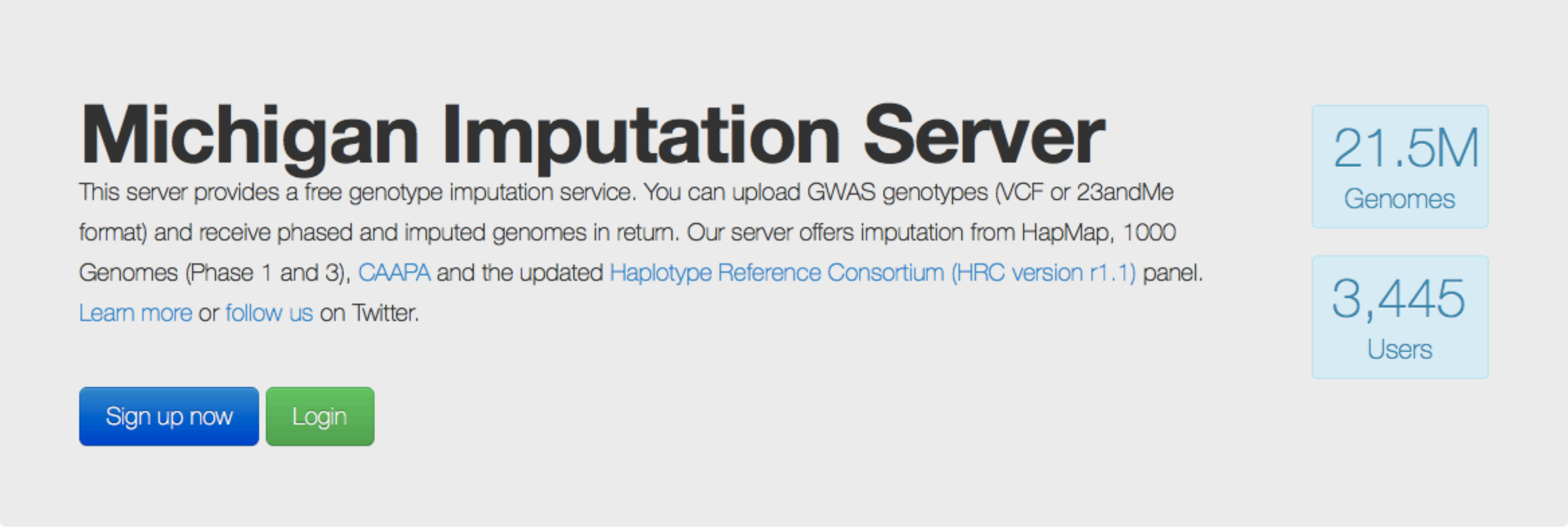
UK10K

Rare Genetic Variants in Health and Disease


The 100,000 Genomes Project

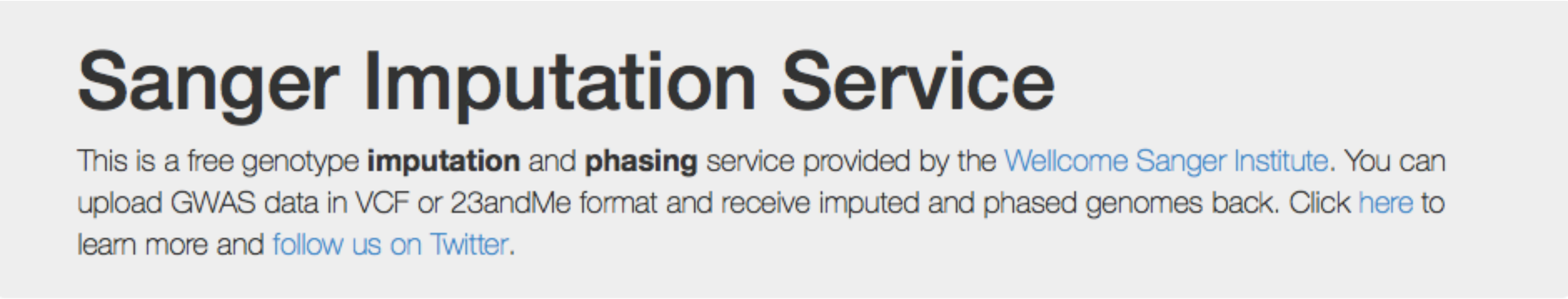
The project will sequence 100,000 genomes from around 70,000 people. Participants are NHS patients with a rare disease, plus their families, and patients with cancer.

Imputation Servers



Michigan Imputation Server

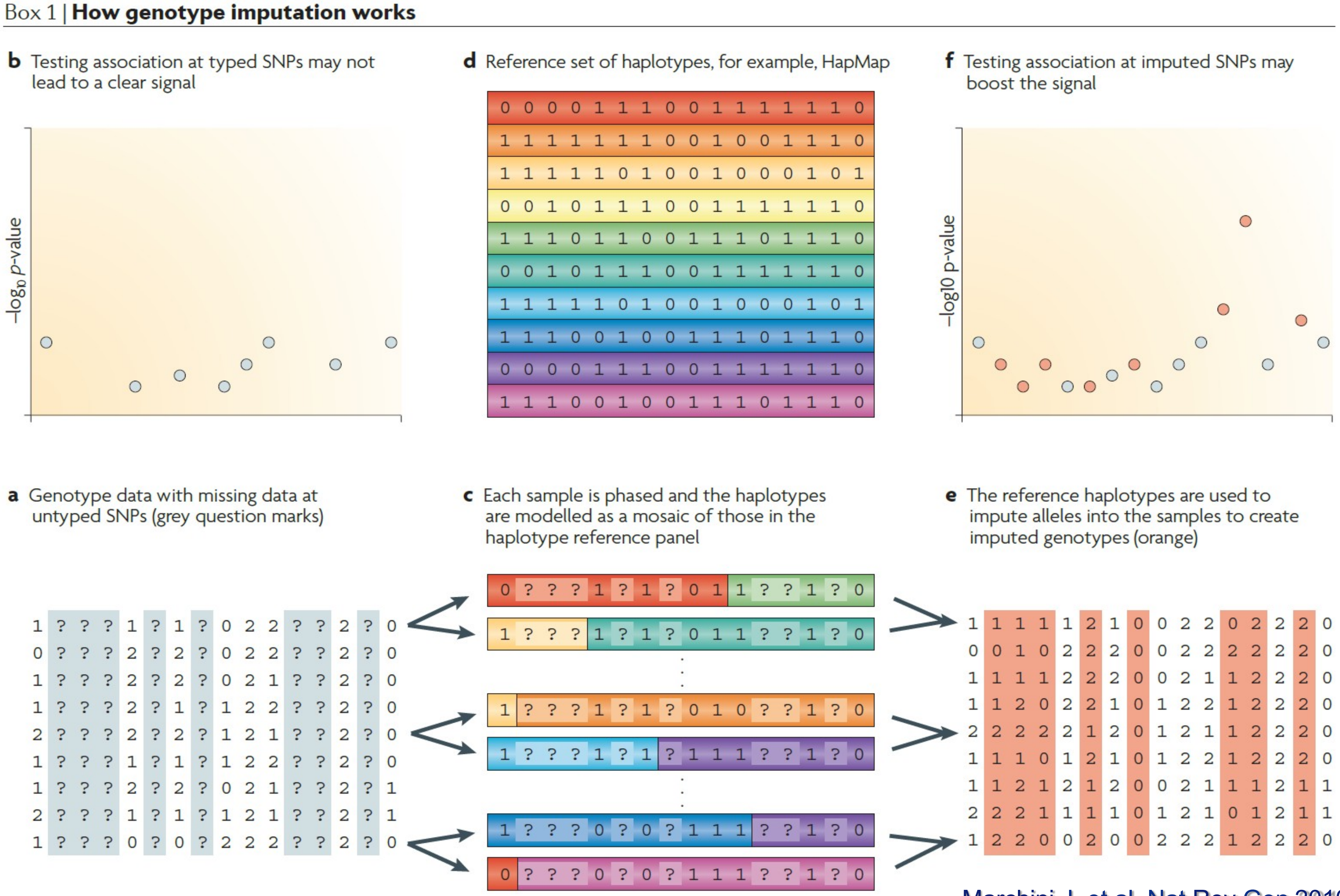
This server provides a free genotype imputation service. You can upload GWAS genotypes (VCF or 23andMe format) and receive phased and imputed genomes in return. Our server offers imputation from HapMap, 1000 Genomes (Phase 1 and 3), CAAPA and the updated Haplotype Reference Consortium (HRC version r1.1) panel. Learn more or follow us on Twitter.



Sanger Imputation Service

This is a free genotype **imputation** and **phasing** service provided by the Wellcome Sanger Institute. You can upload GWAS data in VCF or 23andMe format and receive imputed and phased genomes back. Click here to learn more and follow us on Twitter.

Genotype imputation in GWAS



Box 1 | **How genotype imputation works**

b Testing association at typed SNPs may not lead to a clear signal

d Reference set of haplotypes, for example, HapMap

```
0 0 0 0 1 1 1 0 0 1 1 1 1 1 1 0
1 1 1 1 1 1 1 0 0 1 0 0 1 1 1 0
1 1 1 1 1 0 1 0 0 1 0 0 0 1 0 1
0 0 1 0 1 1 1 0 0 1 1 1 1 1 1 0
1 1 1 0 1 1 0 0 1 1 1 0 1 1 1 0
0 0 1 0 1 1 1 0 0 1 1 1 1 1 1 0
1 1 1 1 1 0 1 0 0 1 0 0 0 1 0 1
1 1 1 0 0 1 0 0 1 1 1 0 1 1 1 0
0 0 0 0 1 1 1 0 0 1 1 1 1 1 1 0
1 1 1 0 0 1 0 0 1 1 1 0 1 1 1 0
```

f Testing association at imputed SNPs may boost the signal

a Genotype data with missing data at untyped SNPs (grey question marks)

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

c Each sample is phased and the haplotypes are modelled as a mosaic of those in the haplotype reference panel

```
0 ? ? ? 1 ? 1 ? 0 1 1 ? ? 1 ? 0
1 ? ? ? 1 ? 1 ? 0 1 1 ? ? 1 ? 0
               .
               .
               .
1 ? ? ? 1 ? 1 ? 0 1 0 ? ? 1 ? 0
1 ? ? ? 1 ? 1 ? 1 1 1 ? ? 1 ? 0
               .
               .
               .
1 ? ? ? 0 ? 0 ? 1 1 1 ? ? 1 ? 0
0 ? ? ? 0 ? 0 ? 1 1 1 ? ? 1 ? 0
```

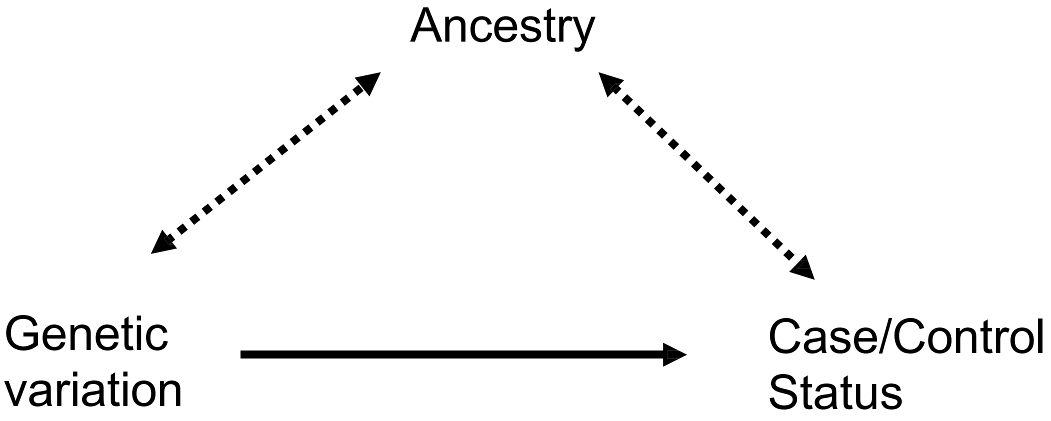
e The reference haplotypes are used to impute alleles into the samples to create imputed genotypes (orange)

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

Marchini J, et al. Nat Rev Gen 2010;499-5

Confounding by Ancestry (Population Stratification)

- Distortion of the relationship between the genetic risk factor and the outcome of interest due to ancestry that is related to both the frequency of the putative genetic risk factor and whether or not subject is a case or a control.



Ancestry

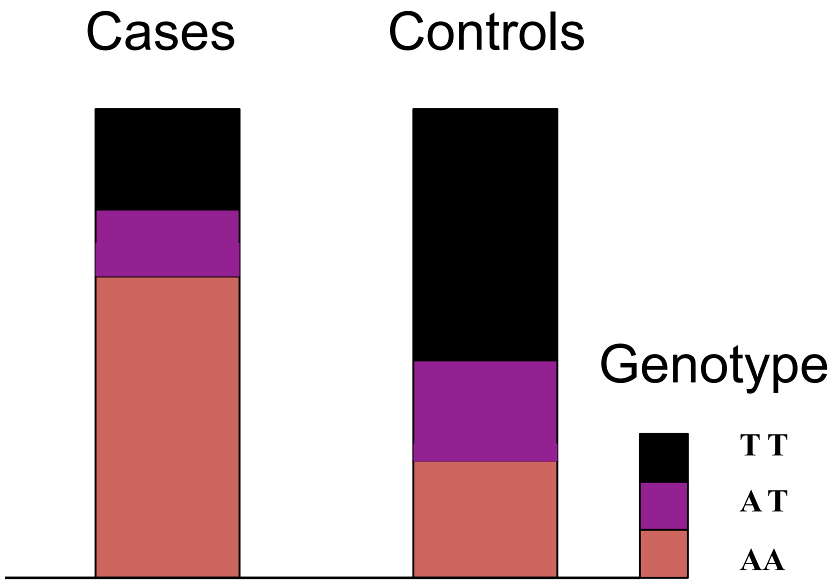
Genetic variation

Case/Control Status

Spurious association due to population stratification

- Distribution of genotypes differs between cases and controls

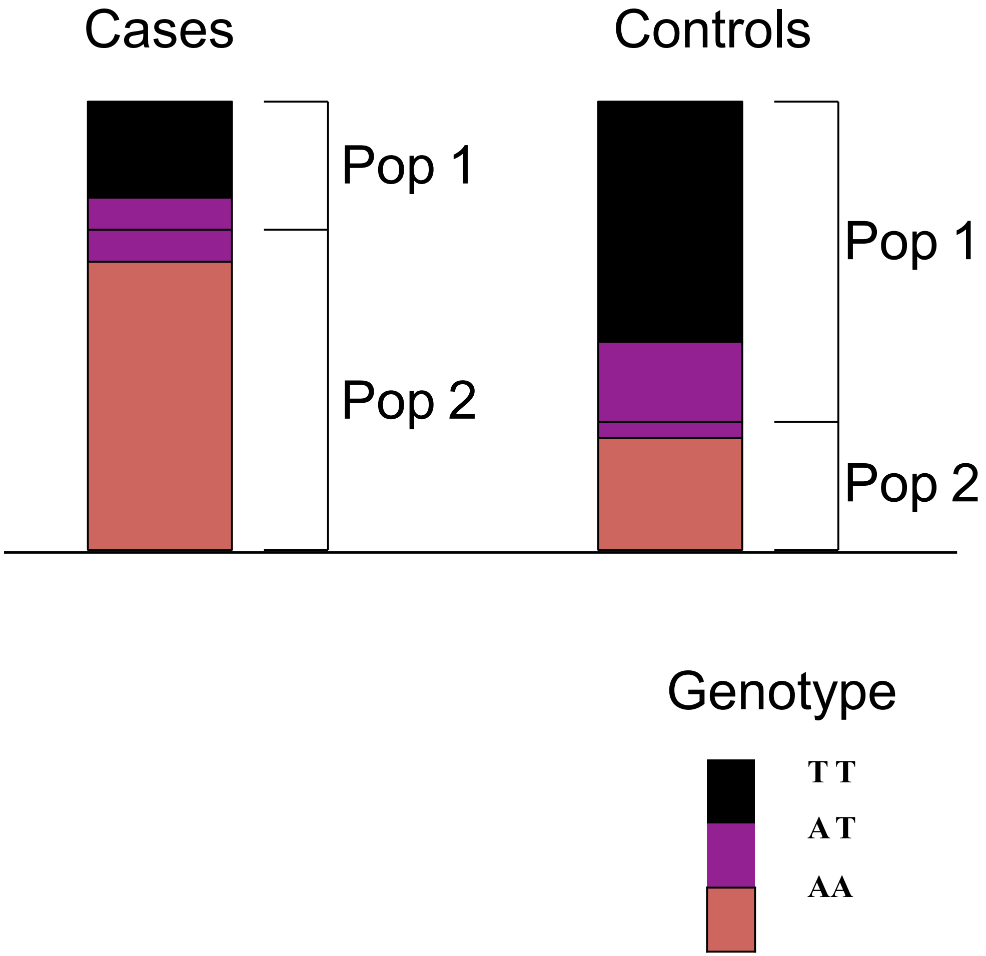
- Might conclude that allele A (or genotype AA) related to disease



Population Stratification

- Unequal distribution of non-disease-related alleles between cases and controls

- Any allele more common in population with increased risk of disease may appear to be associated with disease



Using the GWA Data to Avoid Population Stratification

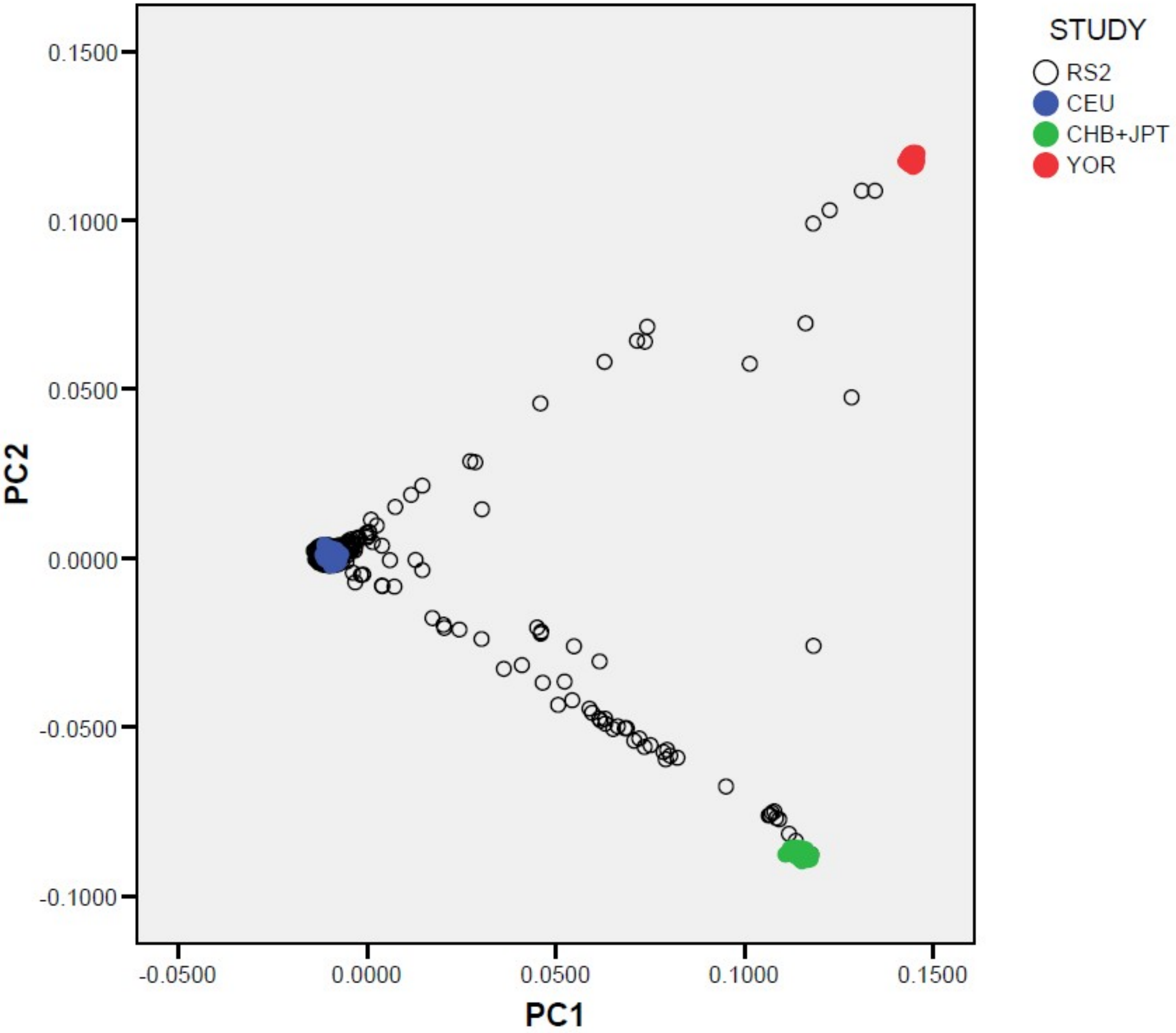
- The information on the genome-wide markers could be used to:
  - Estimate ancestry groups and remove extreme outliers
  - Estimate inflation of test statistic and adjust for it

Identifying and correcting for population stratification

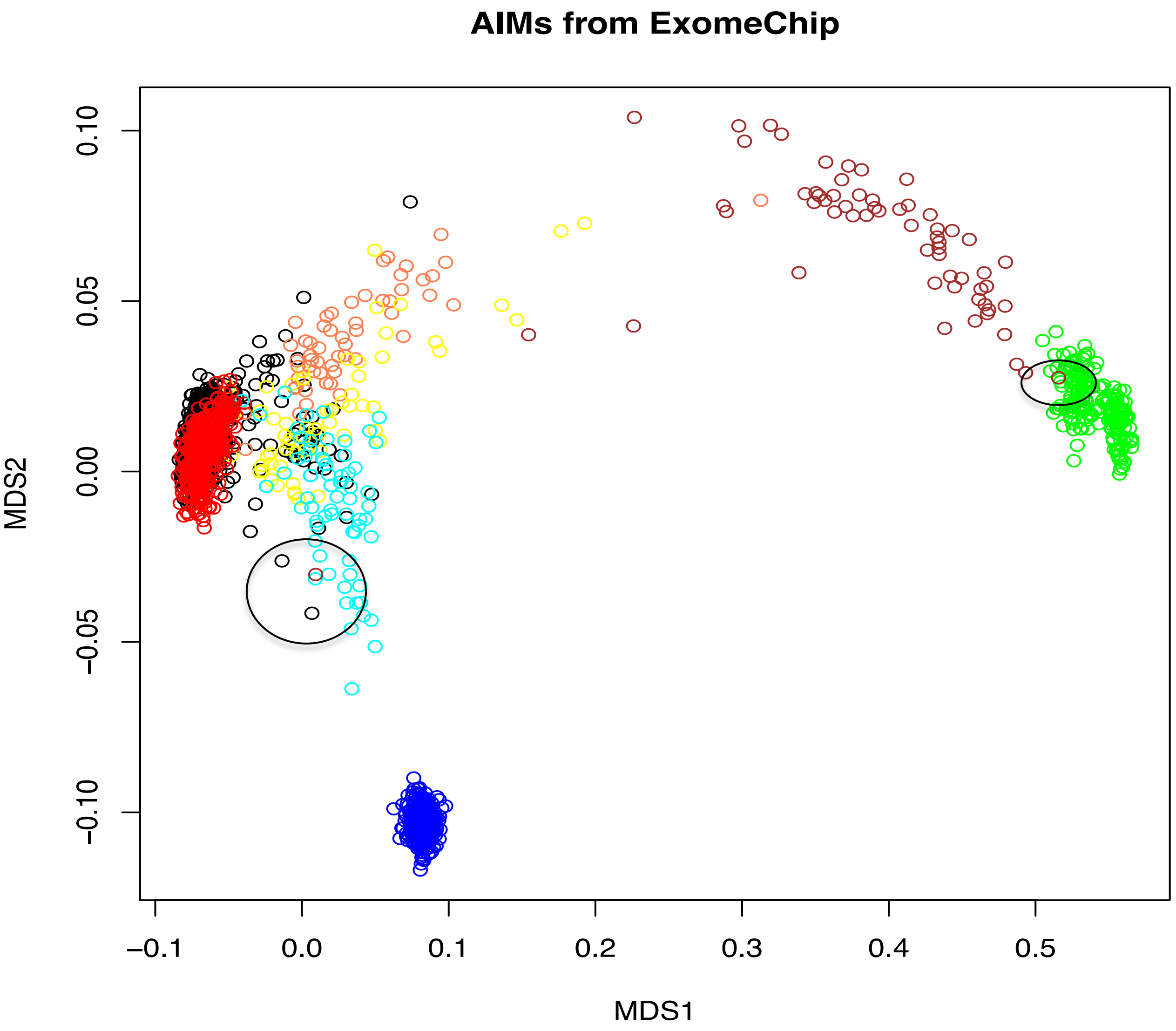
- Genomic control
  - SNPs are used to calculate background inflation in test statistic (due to population stratification)
  - Significant associations are excluded
  - Diminishes the statistical power

- Adjustment for population stratification
  - Principal components analysis, adjustment/matching for top PC’s

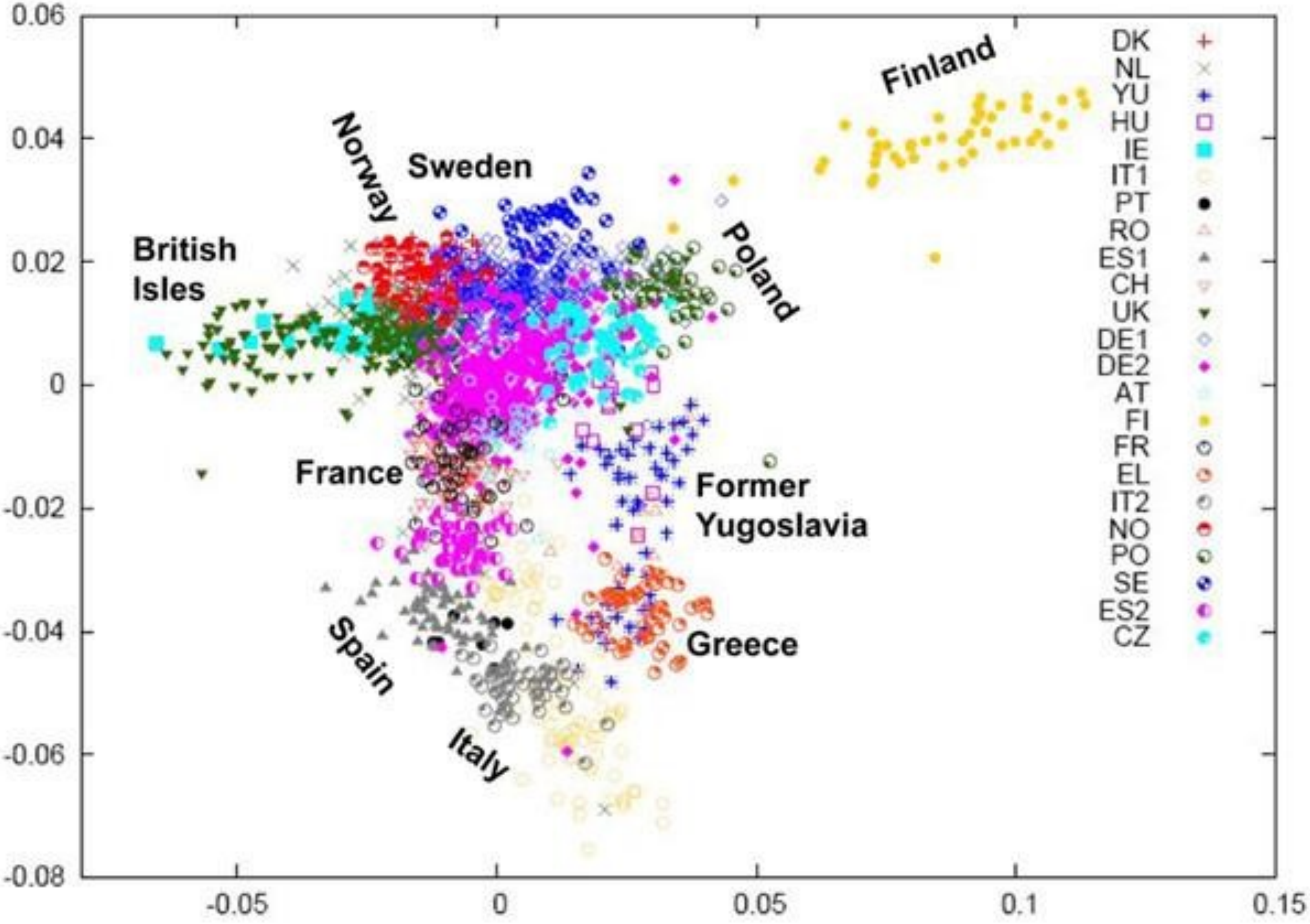
Estimating the ancestry groups



Identifying outliers



PCs pick up fine population structure



Razib, Current Biology 2008

Run GWAS analysis

- Running GWAS is actually repeatedly running a regression model for all SNPs
- Normally a QQ-plot is used to check the distribution of the pvalues
- Manhattan plots are used to get an overview of the findings
- Regional plots are used to take a close look at every locus

Models of inheritance

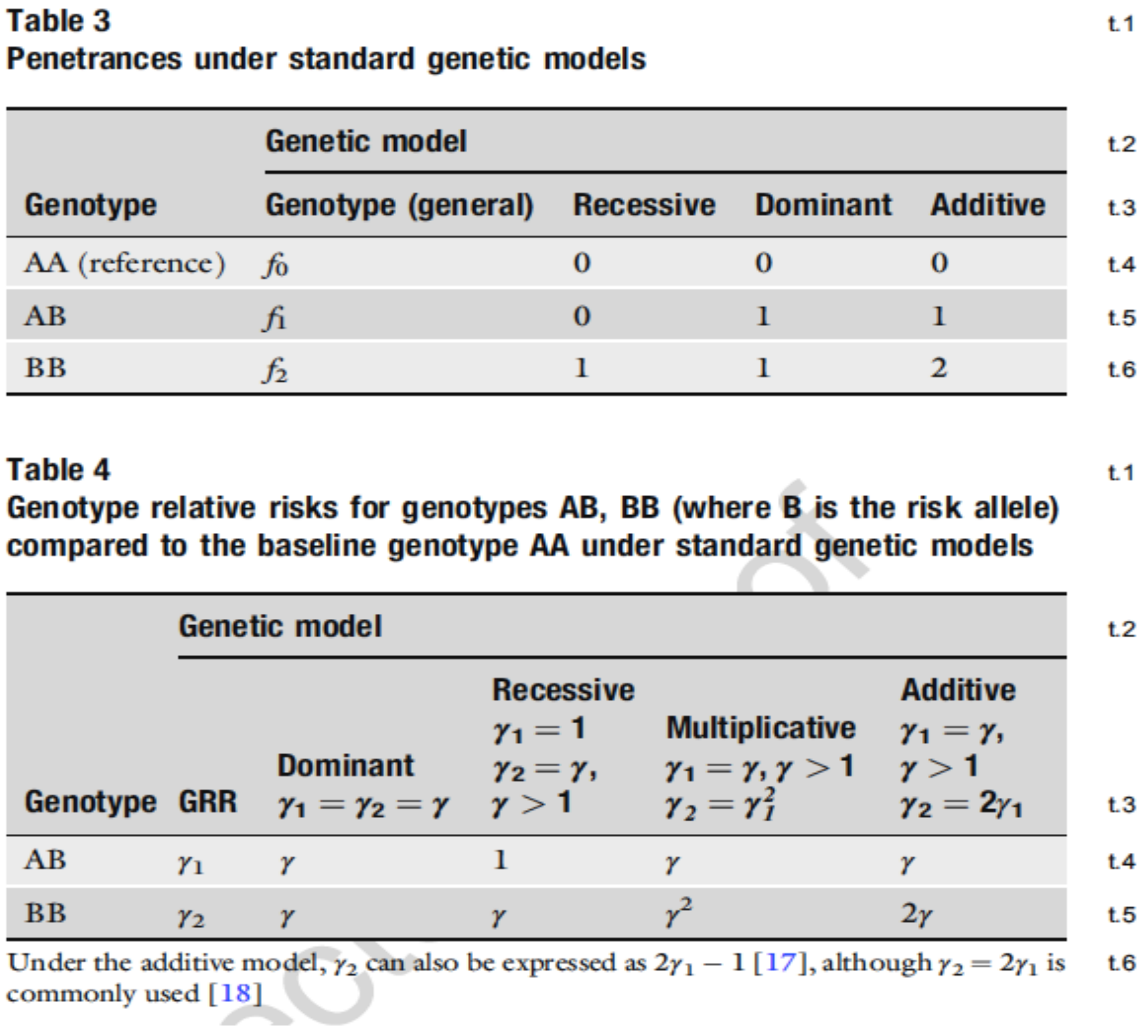


Table 3
Penetrances under standard genetic models

	Genetic model			
Genotype	**Genotype (general)**	**Recessive**	**Dominant**	**Additive**
AA (reference)	f_0	0	0	0
AB	f_1	0	1	1
BB	f_2	1	1	2

Table 4
Genotype relative risks for genotypes AB, BB (where B is the risk allele) compared to the baseline genotype AA under standard genetic models

	Genetic model				
Genotype	**GRR**	**Dominant** $\gamma_1 = \gamma_2 = \gamma$	**Recessive** $\gamma_1 = 1$ $\gamma_2 = \gamma,$ $\gamma > 1$	**Multiplicative** $\gamma_1 = \gamma, \gamma > 1$ $\gamma_2 = \gamma_1^2$	**Additive** $\gamma_1 = \gamma,$ $\gamma > 1$ $\gamma_2 = 2\gamma_1$
AB	γ_1	γ	1	γ	γ
BB	γ_2	γ	γ	γ^2	2γ

Under the additive model, γ_2 can also be expressed as $2\gamma_1 - 1$ [17], although $\gamma_2 = 2\gamma_1$ is commonly used [18]

Genetic Epidemiology: Methods and Protocols. Evangelou (ed)
Ch. 2: Key concepts in genetic epidemiology

Estimate inflation of test statistic

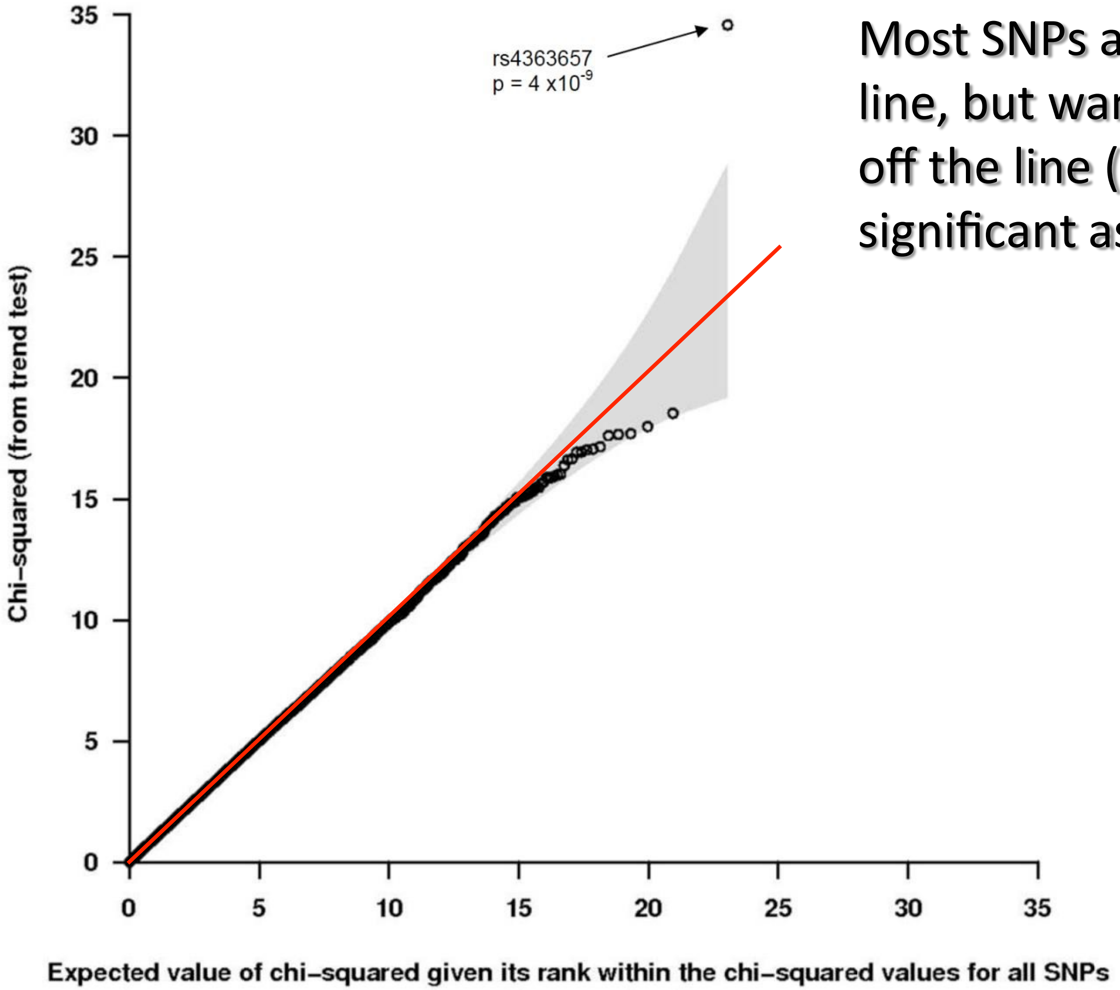
- Q-Q plot is a plot to compare two probability distributions

- In GWA studies, QQ-plots compare the distribution of p-values of GWAS with a distribution when no associations

- When no real association is found the two distributions are similar and the points will lie on the identity line (y = x)

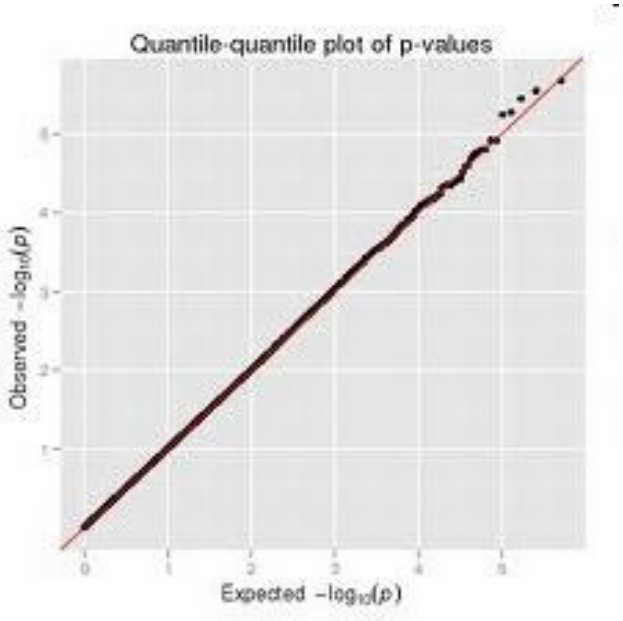
- Deviations from the identity line could be due to:
  - True associations
  - Population stratification

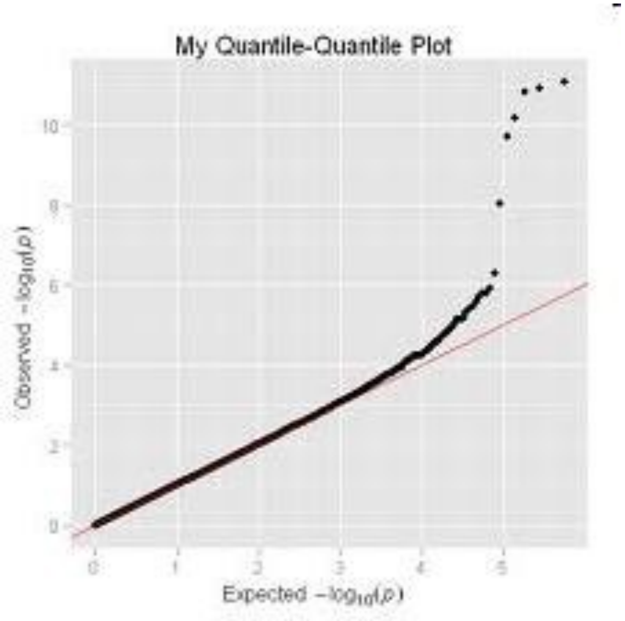
Quantile-quantile (QQ) plot

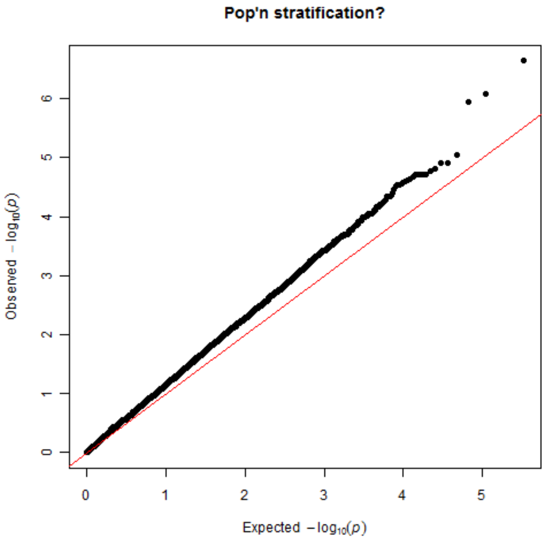


Most SNPs are on the line, but want a few hits off the line (true significant associations!)

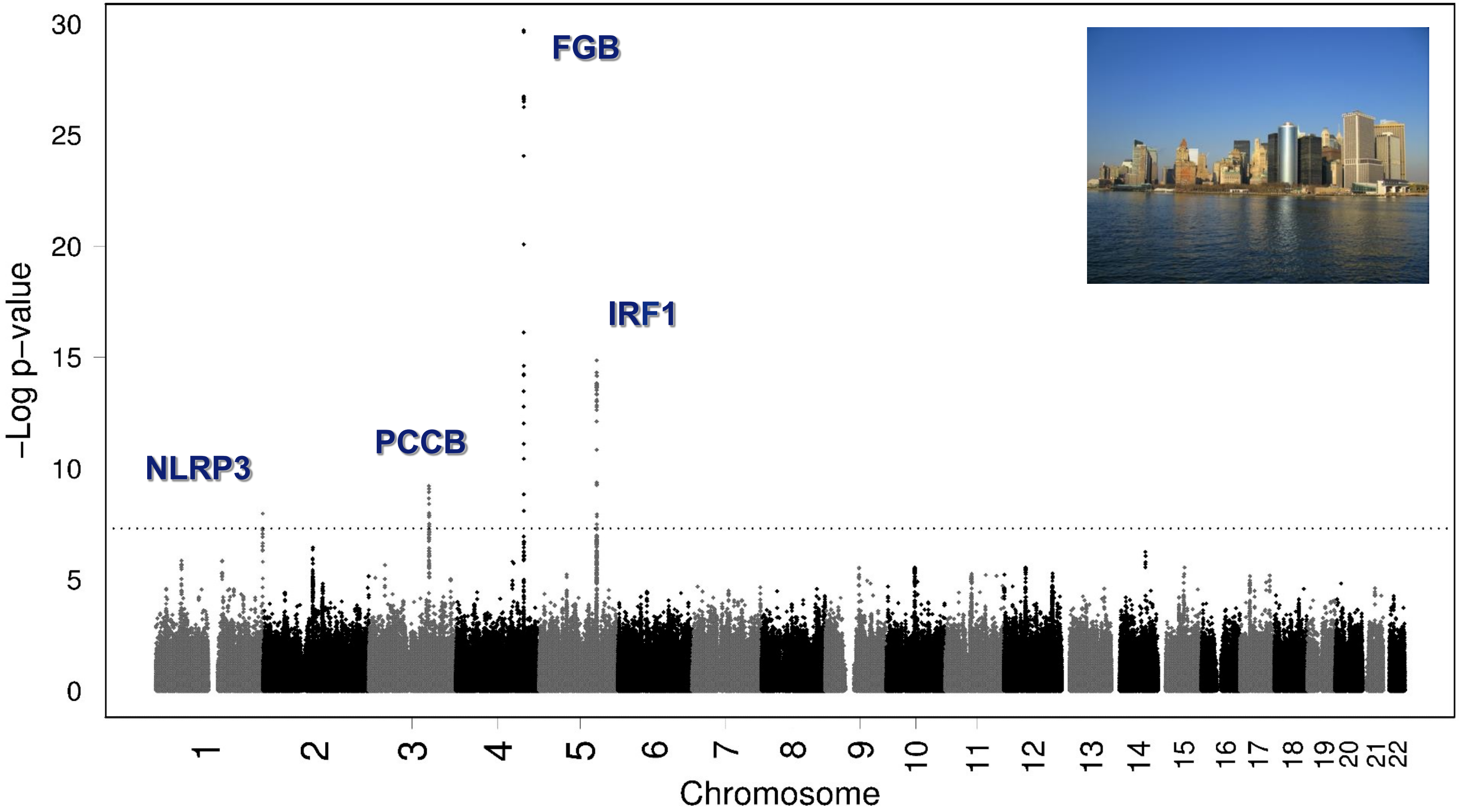
QQ-plots in GWAS





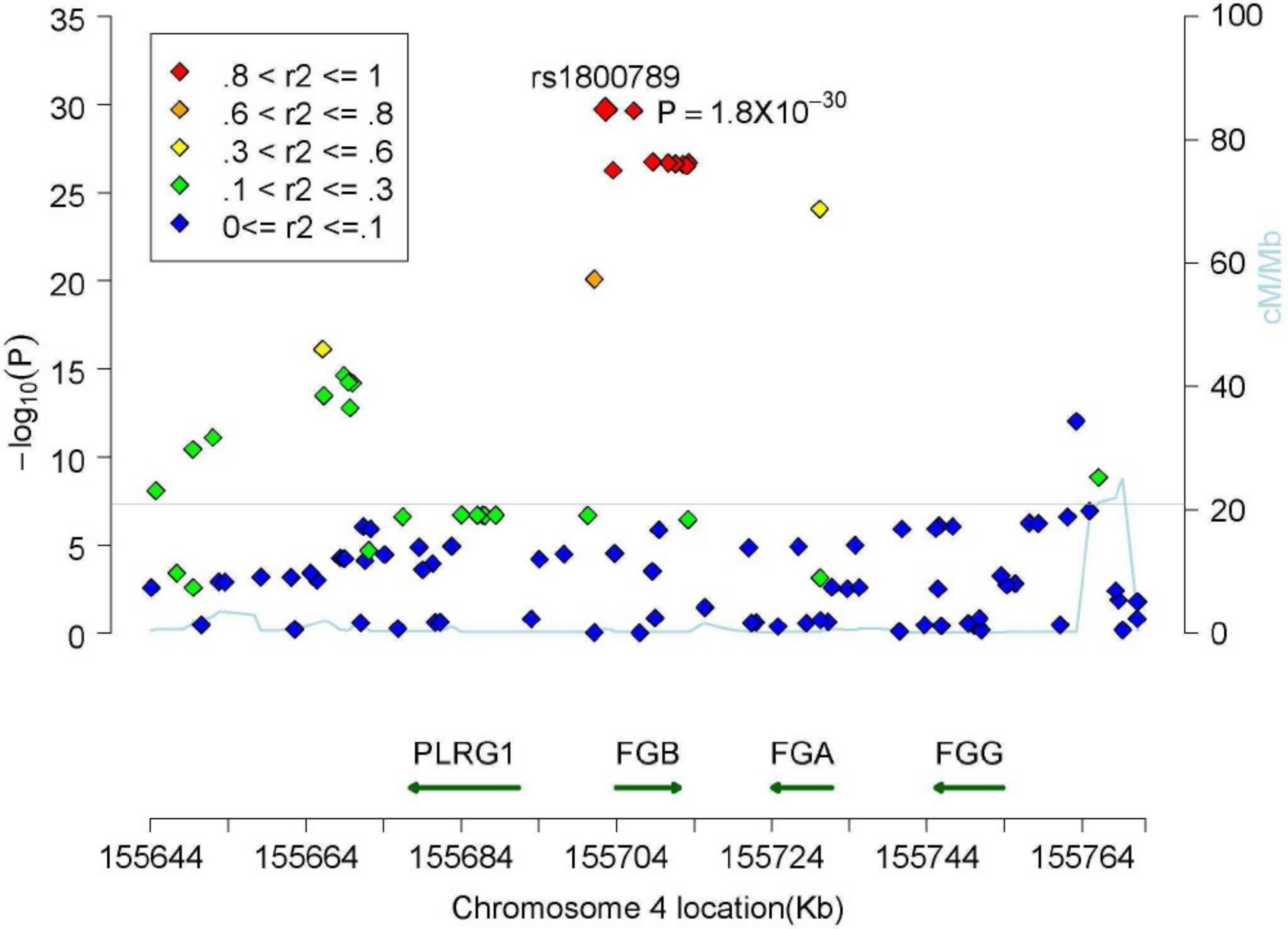


-log plot (Manhattan plot)



Dehghan et al. Circ Cardiovasc Genet. 2009

Regional plot



Dehghan et al. Circ Cardiovasc Genet. 2009

Multiple Testing Issue

Bonferroni correction

1. Assume all tests performed are independent
2. Estimate number of independent polymorphisms in genome
3. Threshold often considered appropriate: $5x10^{-8}$
4. Recently more conservative thresholds are used such as $1x10^{-8}$ or $1x10^{-9}$

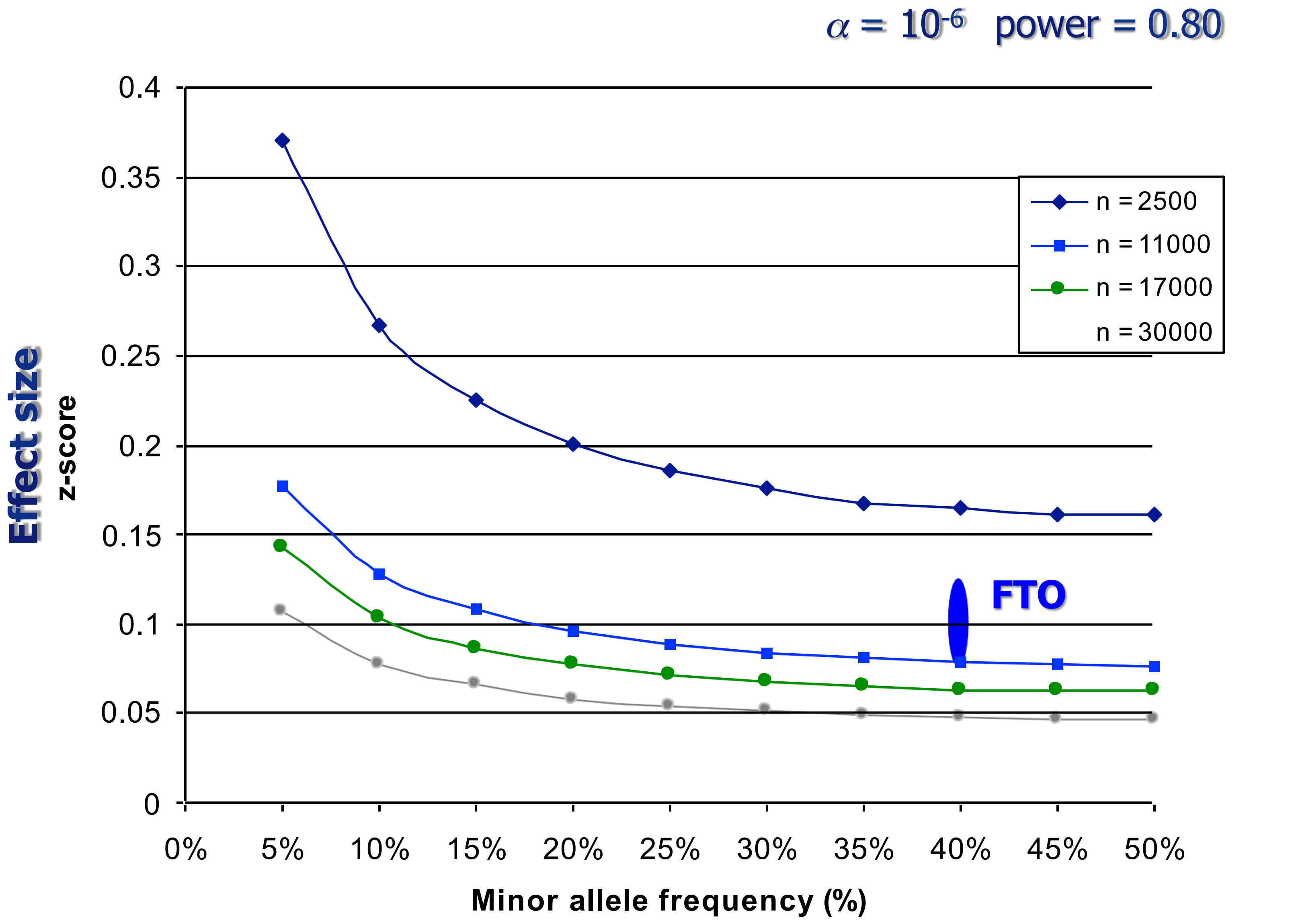
Multiple Testing Issue

Permutation

- Permute case and control status, perform all tests, record the most significant p-value among those tests and then re-permute case-control status and test again. Repeat many times.

- P-value for most significant test is the proportion of permutations that had a “best” p-value as small or smaller than the one you observe with the observed data (the data with the right case and control labels).

Effect size – MAF - Power



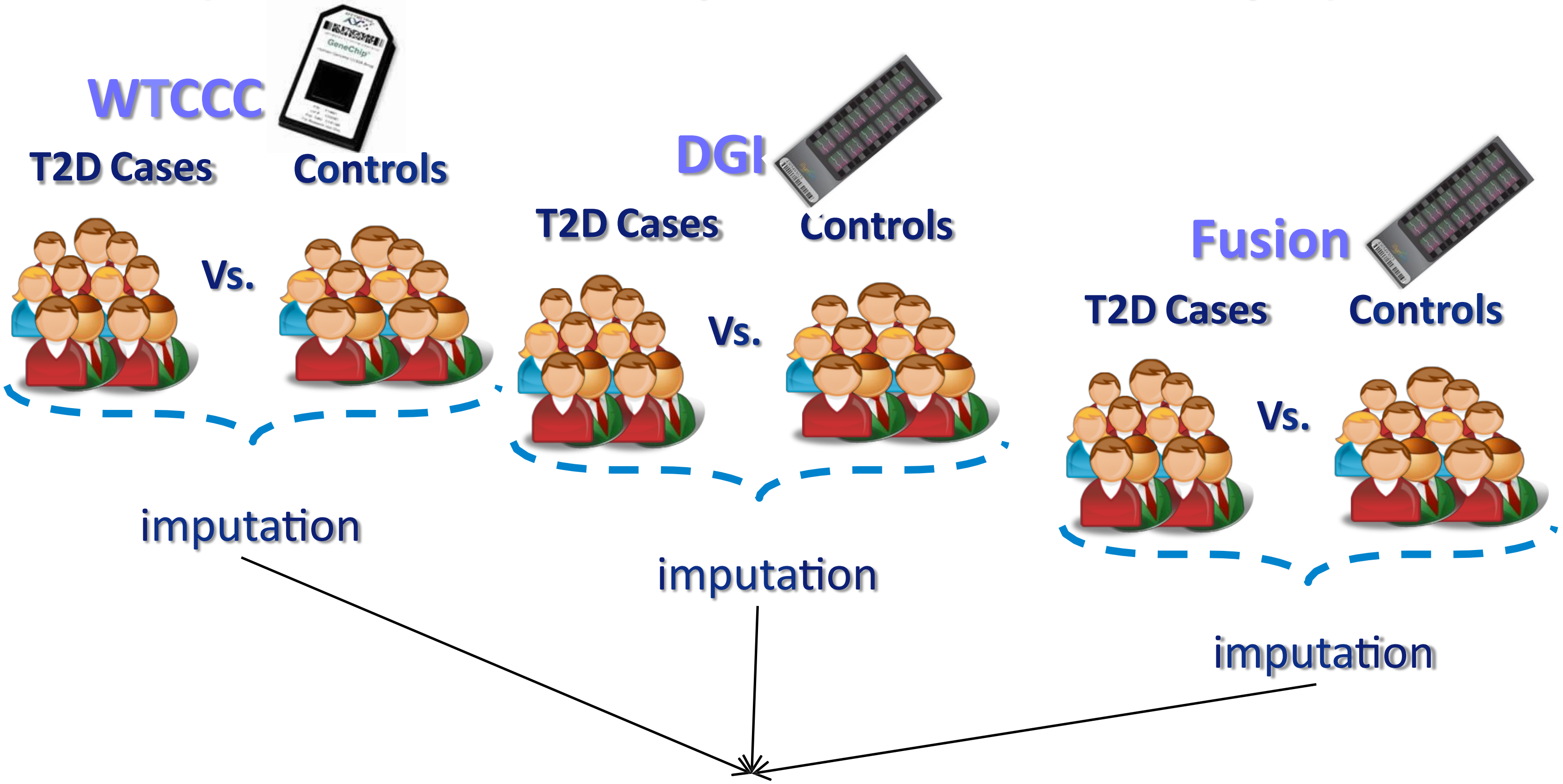
Loos, Lindgren, Li, unpublished data

Meta-analysis

- Large sample sizes are needed
- Combine multiple studies to increase power
- Either combine p-values (Fisher’s test), or coefficient estimates + standard error (better)

Meta-analysis of genome-wide association studies

DIAGRAM (DIAbetes Genetics Replication And Meta-analysis)consortium



WTCCC

T2D Cases Vs. Controls

imputation

DGI

T2D Cases Vs. Controls

imputation

Fusion

T2D Cases Vs. Controls

imputation

Meta-analysis

Interpreting the Statistical Results

- If you identify a SNP that is significantly associated with disease, there are three possibilities:
  - There is a causal relationship between SNP and disease
  - The marker is in linkage disequilibrium with a causal locus
  - False positive
- Many potential sources of systematic errors that might lead to false positive results.
  - Genotyping quality control issues particularly important
  - Population stratification

False positive

- This may occur specially when the pvalue is borderline.

- Most of the highly significant findings are true

- Heterogeneity should be considered

Replication

- GWA studies are hypothesis-generating (agnostic approach)

- The hypothesis should be tested in an independent sample

- When you have not reached genome-wide significance level

Replication

- To replicate:
  - Significance threshold = 0.05/#of SNPs (??)
  - Same genetic model (e.g. additive, dominant)
  - Same direction
  - Sufficient sample size for replication
  - Control for population stratification in replication samples

Non-replications

- Not necessarily a false positive
  - Underpowered (Winner's curse)
  - Ethnic background (LD structures)
  - Phenotype definition (subphenotype/phenotype)
  - Population stratification
  - Different covariates
  - .
  - .
  - False positive!

Replication challenges and solutions

- Providing enough sample size is challenging
- Harmonized phenotyping is not always possible
- Split the sample?

One-stage designs

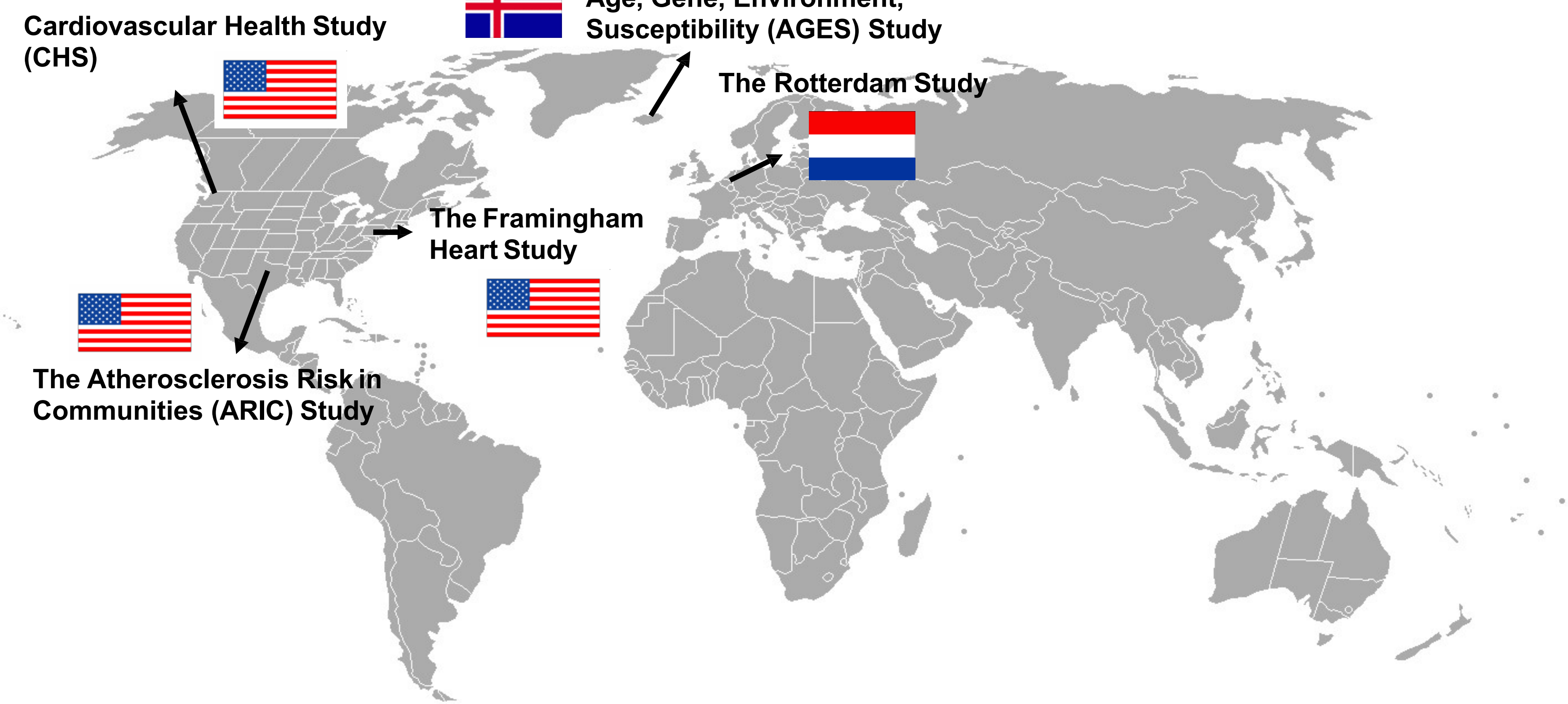
- Increase power by combining all available resources
- Replication sample may not have enough power to replicate signals
- P-value threshold?

Collaboration is the key to successful GWAS

- Large consortia were formed to provide the infrastructure for replication and pooling the data
- Building trust and agreeable regulations were the initial challenges
- Different genotyping platforms and measurement methods are still a challenge is all collaborative projects

General consortia

The Cohorts for Heart and Aging Research in Genomic Epidemiology



Cardiovascular Health Study (CHS)

Age, Gene, Environment, Susceptibility (AGES) Study

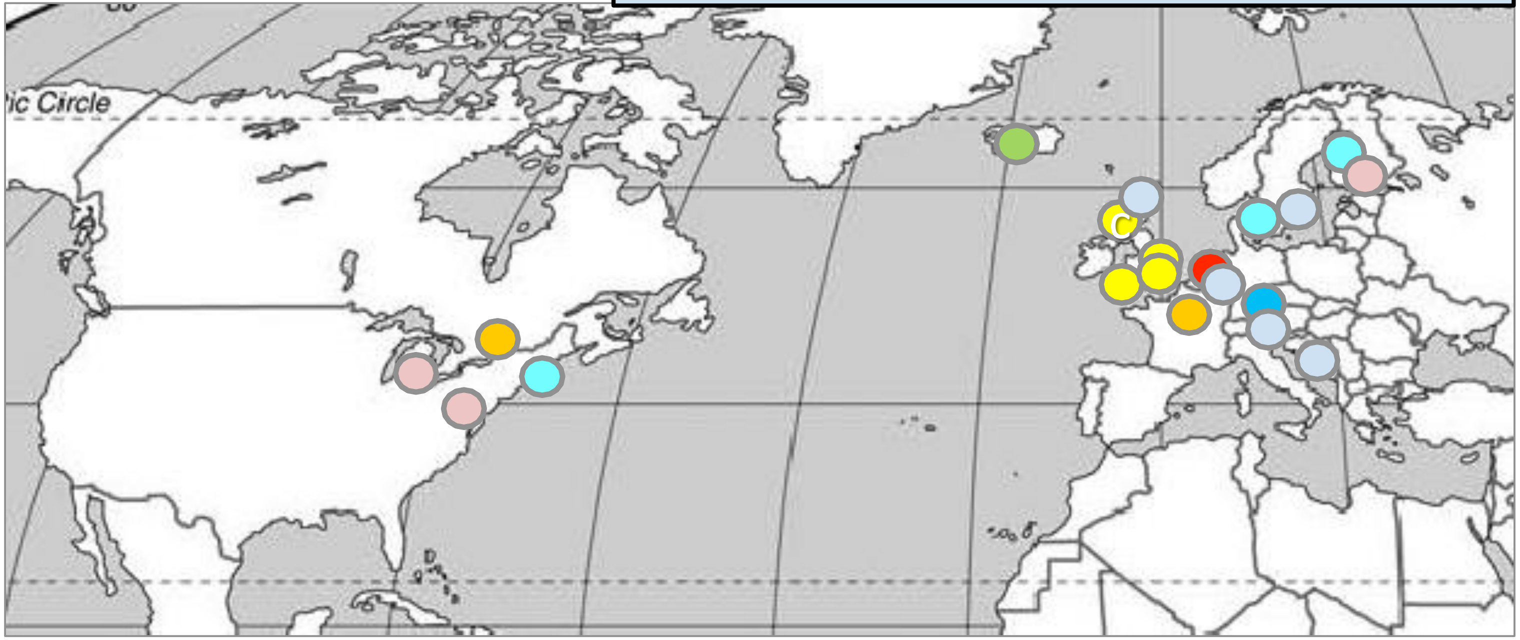
The Rotterdam Study

The Framingham Heart Study

The Atherosclerosis Risk in Communities (ARIC) Study

Disease Consortia

The DIAGRAM+ consortium



● UK

● FUSION (US/Finland)

● DGI (US/Sweden/Finland)

● DeCODE

● KORA

● Rotterdam

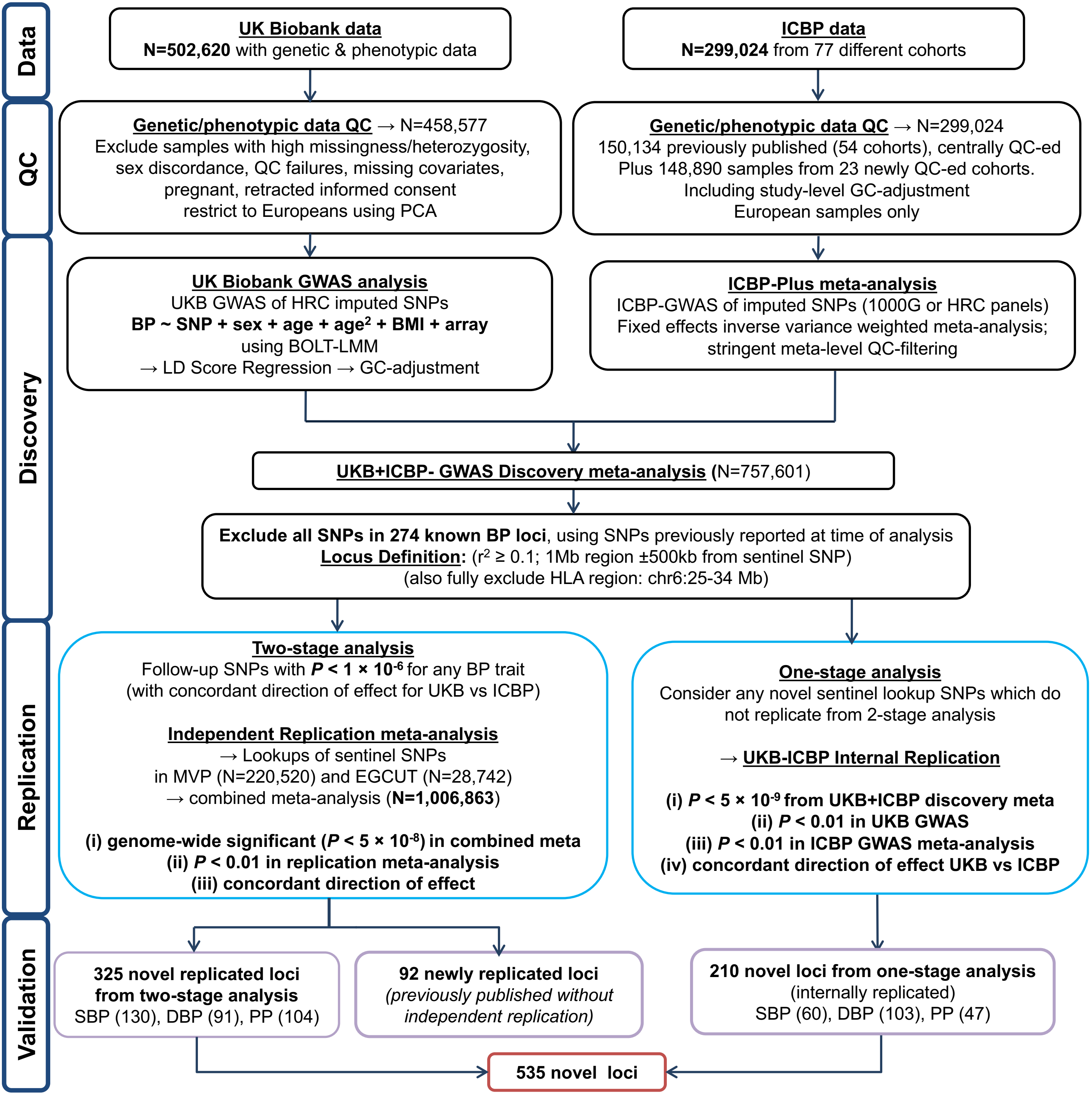
● DGDG (France/Canada)

● EUROSPAN

Genetic analysis of over one million people identifies 535 novel loci associated with blood pressure traits

Going beyond the 1 M participants

Study design



Data

UK Biobank data
N=502,620 with genetic & phenotypic data

ICBP data
N=299,024 from 77 different cohorts

QC

Genetic/phenotypic data QC → N=458,577
Exclude samples with high missingness/heterozygosity, sex discordance, QC failures, missing covariates, pregnant, retracted informed consent
restrict to Europeans using PCA

Genetic/phenotypic data QC → N=299,024
150,134 previously published (54 cohorts), centrally QC-ed
Plus 148,890 samples from 23 newly QC-ed cohorts.
Including study-level GC-adjustment
European samples only

Discovery

UK Biobank GWAS analysis
UKB GWAS of HRC imputed SNPs
BP ~ SNP + sex + age + age^2 + BMI + array
using BOLT-LMM
→ LD Score Regression → GC-adjustment

ICBP-Plus meta-analysis
ICBP-GWAS of imputed SNPs (1000G or HRC panels)
Fixed effects inverse variance weighted meta-analysis; stringent meta-level QC-filtering

UKB+ICBP- GWAS Discovery meta-analysis (N=757,601)

Exclude all SNPs in 274 known BP loci, using SNPs previously reported at time of analysis
Locus Definition: ($r^2 \geq 0.1$; 1Mb region ±500kb from sentinel SNP)
(also fully exclude HLA region: chr6:25-34 Mb)

Replication

Two-stage analysis
Follow-up SNPs with $\boldsymbol{P < 1 \times 10^{-6}}$ for any BP trait
(with concordant direction of effect for UKB vs ICBP)

Independent Replication meta-analysis
→ Lookups of sentinel SNPs
in MVP (N=220,520) and EGCUT (N=28,742)
→ combined meta-analysis (**N=1,006,863**)

(i) genome-wide significant ($P < 5 \times 10^{-8}$) in combined meta
(ii) $P < 0.01$ in replication meta-analysis
(iii) concordant direction of effect

One-stage analysis
Consider any novel sentinel lookup SNPs which do not replicate from 2-stage analysis

→ **UKB-ICBP Internal Replication**

(i) $P < 5 \times 10^{-9}$ from UKB+ICBP discovery meta
(ii) $P < 0.01$ in UKB GWAS
(iii) $P < 0.01$ in ICBP GWAS meta-analysis
(iv) concordant direction of effect UKB vs ICBP

Validation

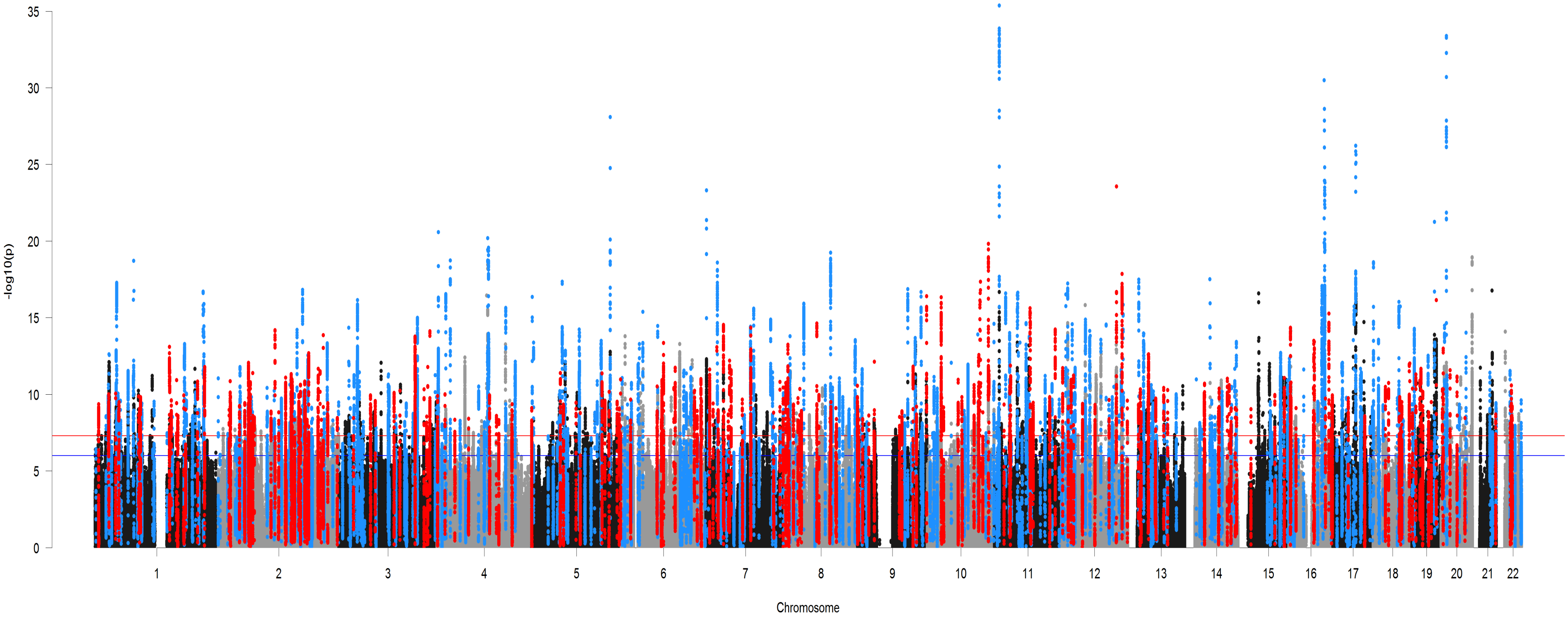
325 novel replicated loci from two-stage analysis
SBP (130), DBP (91), PP (104)

92 newly replicated loci
(previously published without independent replication)

210 novel loci from one-stage analysis
(internally replicated)
SBP (60), DBP (103), PP (47)

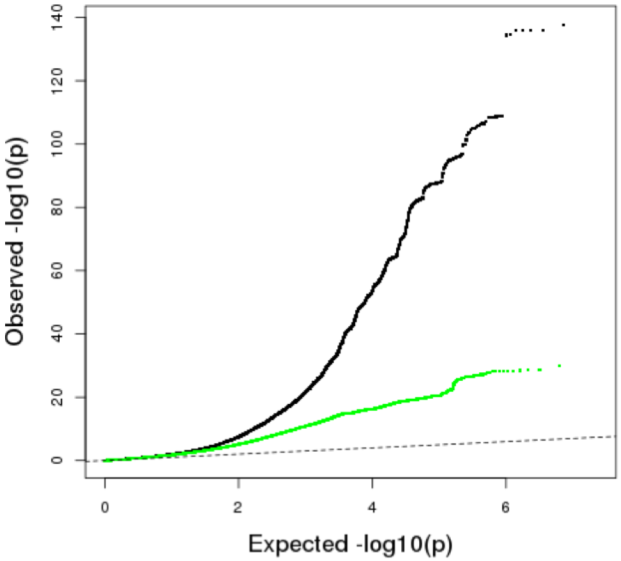
535 novel loci

Manhattan Plot excluding known variants

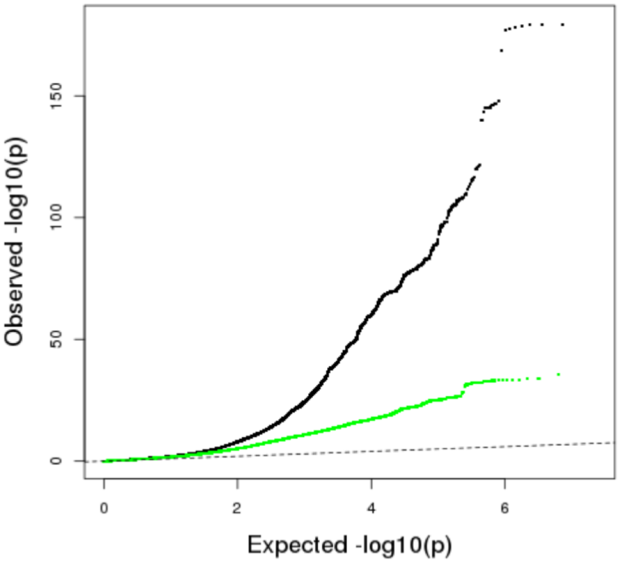


Early deviation due to large power

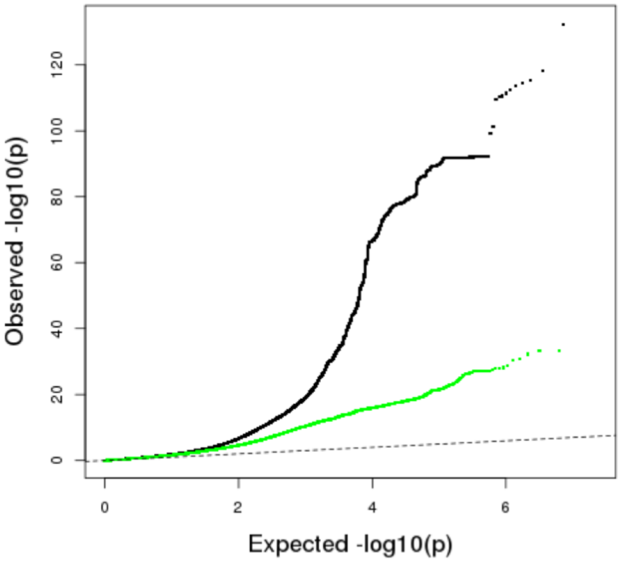
SBP



DBP



PP



How could the genetic data be used in clinic?

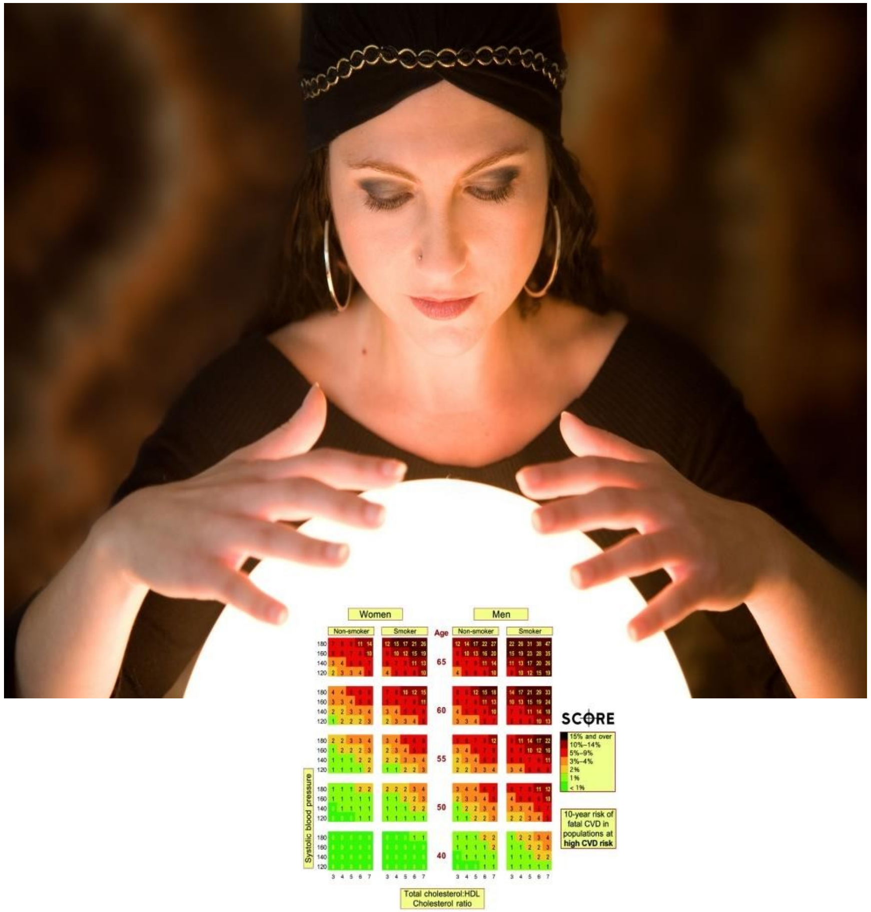
- Drug target

- Precision medicine

Risk prediction

- Risk prediction is widely used for clinical practice e.g. in cardiology

- Various risk scores have so far been developed



		Women										Men									
		Non-smoker					Smoker					Non-smoker					Smoker				
Age	Systolic blood pressure	3	4	5	6	7	3	4	5	6	7	3	4	5	6	7	3	4	5	6	7
65	180	[illegible]	[illegible]	[illegible]	11	14	12	15	17	21	26	12	14	17	22	27	22	26	31	38	47
	160	[illegible]	[illegible]	[illegible]	[illegible]	10	[illegible]	10	12	15	19	[illegible]	10	13	16	20	15	19	23	28	35
	140	3	4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	11	13	[illegible]	[illegible]	[illegible]	11	14	11	13	17	20	26
	120	2	3	3	4	[illegible]	4	[illegible]	[illegible]	[illegible]	10	4	[illegible]	[illegible]	[illegible]	10	[illegible]	10	12	15	19
60	180	4	4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	10	12	15	[illegible]	[illegible]	12	15	18	14	17	21	29	33
	160	3	3	4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	11	[illegible]	[illegible]	[illegible]	11	13	10	13	15	19	24
	140	2	2	3	3	4	3	4	[illegible]	[illegible]	[illegible]	4	[illegible]	[illegible]	[illegible]	10	[illegible]	[illegible]	11	14	18
	120	1	2	2	2	3	2	3	3	4	[illegible]	3	3	4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	10	13
55	180	2	2	3	3	4	3	4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	12	[illegible]	11	14	17	22
	160	1	2	2	2	3	2	3	4	[illegible]	[illegible]	3	4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	10	12	16
	140	1	1	1	2	2	2	2	3	3	4	2	3	4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	11
	120	1	1	1	1	2	1	1	2	2	3	2	2	3	3	4	3	4	[illegible]	[illegible]	[illegible]
50	180	1	1	1	2	2	2	2	2	3	4	3	4	4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	11	12
	160	1	1	1	1	1	1	1	2	2	3	2	3	3	4	[illegible]	4	[illegible]	[illegible]	[illegible]	10
	140	0	1	1	1	1	1	1	1	2	2	1	2	2	3	4	3	3	4	[illegible]	[illegible]
	120	0	0	0	1	1	1	1	1	1	1	1	1	2	2	3	2	2	3	4	[illegible]
40	180	0	0	0	0	0	0	0	0	1	1	1	1	1	2	2	1	2	2	3	4
	160	0	0	0	0	0	0	0	0	0	0	1	1	1	1	2	1	1	2	2	3
	140	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	2	2
	120	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1

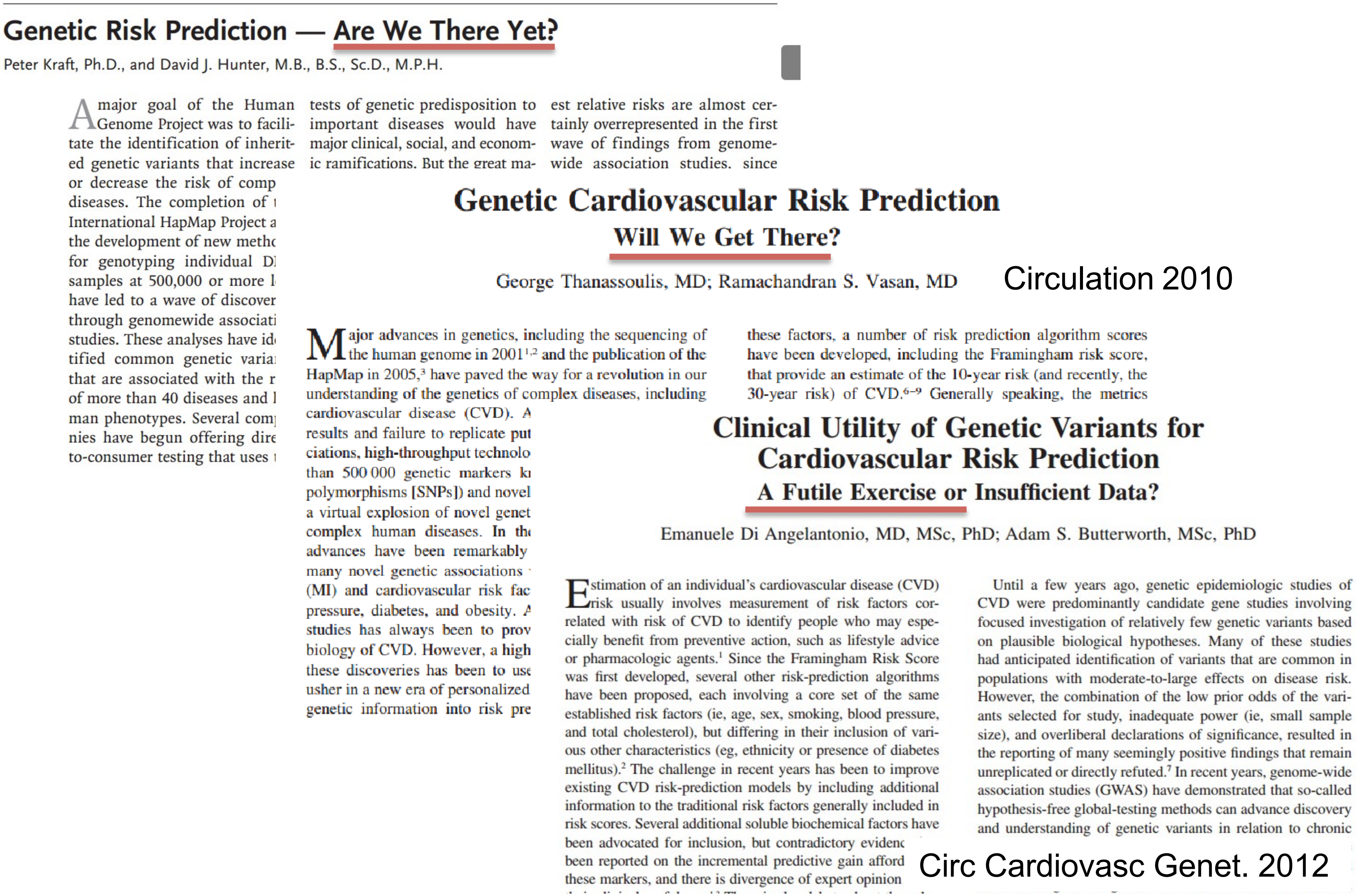
Total cholesterol:HDL Cholesterol ratio

SCORE

■	15% and over
■	10%–14%
🟥	5%–9%
🟧	3%–4%
🟨	2%
🟩	1%
🟩	< 1%

10-year risk of fatal CVD in populations at **high CVD risk**

New England Journal of Medicine 2009



Genetic Risk Prediction — Are We There Yet?

Peter Kraft, Ph.D., and David J. Hunter, M.B., B.S., Sc.D., M.P.H.

A major goal of the Human Genome Project was to facilitate the identification of inherited genetic variants that increase
or decrease the risk of comp
diseases. The completion of t
International HapMap Project a
the development of new metho
for genotyping individual DI
samples at 500,000 or more l
have led to a wave of discover
through genomewide associati
studies. These analyses have id
tified common genetic varia
that are associated with the r
of more than 40 diseases and l
man phenotypes. Several com
nies have begun offering dire
to-consumer testing that uses t

tests of genetic predisposition to important diseases would have major clinical, social, and economic ramifications. But the great ma-

est relative risks are almost certainly overrepresented in the first wave of findings from genome-wide association studies. since

Circulation 2010

Genetic Cardiovascular Risk Prediction

Will We Get There?

George Thanassoulis, MD; Ramachandran S. Vasan, MD

Major advances in genetics, including the sequencing of the human genome in 2001[1,2] and the publication of the HapMap in 2005,[3] have paved the way for a revolution in our understanding of the genetics of complex diseases, including cardiovascular disease (CVD). A
results and failure to replicate put
ciations, high-throughput technolo
than 500 000 genetic markers kı
polymorphisms [SNPs]) and novel
a virtual explosion of novel genet
complex human diseases. In the
advances have been remarkably
many novel genetic associations
(MI) and cardiovascular risk fac
pressure, diabetes, and obesity. A
studies has always been to prov
biology of CVD. However, a high
these discoveries has been to use
usher in a new era of personalized
genetic information into risk pre

these factors, a number of risk prediction algorithm scores have been developed, including the Framingham risk score, that provide an estimate of the 10-year risk (and recently, the 30-year risk) of CVD.[6–9] Generally speaking, the metrics

Circ Cardiovasc Genet. 2012

Clinical Utility of Genetic Variants for Cardiovascular Risk Prediction

A Futile Exercise or Insufficient Data?

Emanuele Di Angelantonio, MD, MSc, PhD; Adam S. Butterworth, MSc, PhD

Estimation of an individual's cardiovascular disease (CVD) risk usually involves measurement of risk factors correlated with risk of CVD to identify people who may especially benefit from preventive action, such as lifestyle advice or pharmacologic agents.[1] Since the Framingham Risk Score was first developed, several other risk-prediction algorithms have been proposed, each involving a core set of the same established risk factors (ie, age, sex, smoking, blood pressure, and total cholesterol), but differing in their inclusion of various other characteristics (eg, ethnicity or presence of diabetes mellitus).[2] The challenge in recent years has been to improve existing CVD risk-prediction models by including additional information to the traditional risk factors generally included in risk scores. Several additional soluble biochemical factors have been advocated for inclusion, but contradictory evidenc
been reported on the incremental predictive gain afford
these markers, and there is divergence of expert opinion

Until a few years ago, genetic epidemiologic studies of CVD were predominantly candidate gene studies involving focused investigation of relatively few genetic variants based on plausible biological hypotheses. Many of these studies had anticipated identification of variants that are common in populations with moderate-to-large effects on disease risk. However, the combination of the low prior odds of the variants selected for study, inadequate power (ie, small sample size), and overliberal declarations of significance, resulted in the reporting of many seemingly positive findings that remain unreplicated or directly refuted.[7] In recent years, genome-wide association studies (GWAS) have demonstrated that so-called hypothesis-free global-testing methods can advance discovery and understanding of genetic variants in relation to chronic

MIT Technology Review

VOL. 121 NO. 2 MARCH/APRIL 2018 US $6.99/CAN $7.99

3-D metal printing

Babel-fish earbuds

The sensing city

AI for everyone

Dueling neural networks

Materials' quantum leap

Zero-carbon natural gas

Perfect online privacy

Artificial embryos

and

Genetic fortune-telling

10 Breakthrough Technologies 2018

DISPLAY UNTIL 05/1/2018

$6.99US $7.99CAN

Forecasts of genetic fate just got a lot more accurate

DNA-based scores are getting better at predicting intelligence, risks for common diseases, and more.

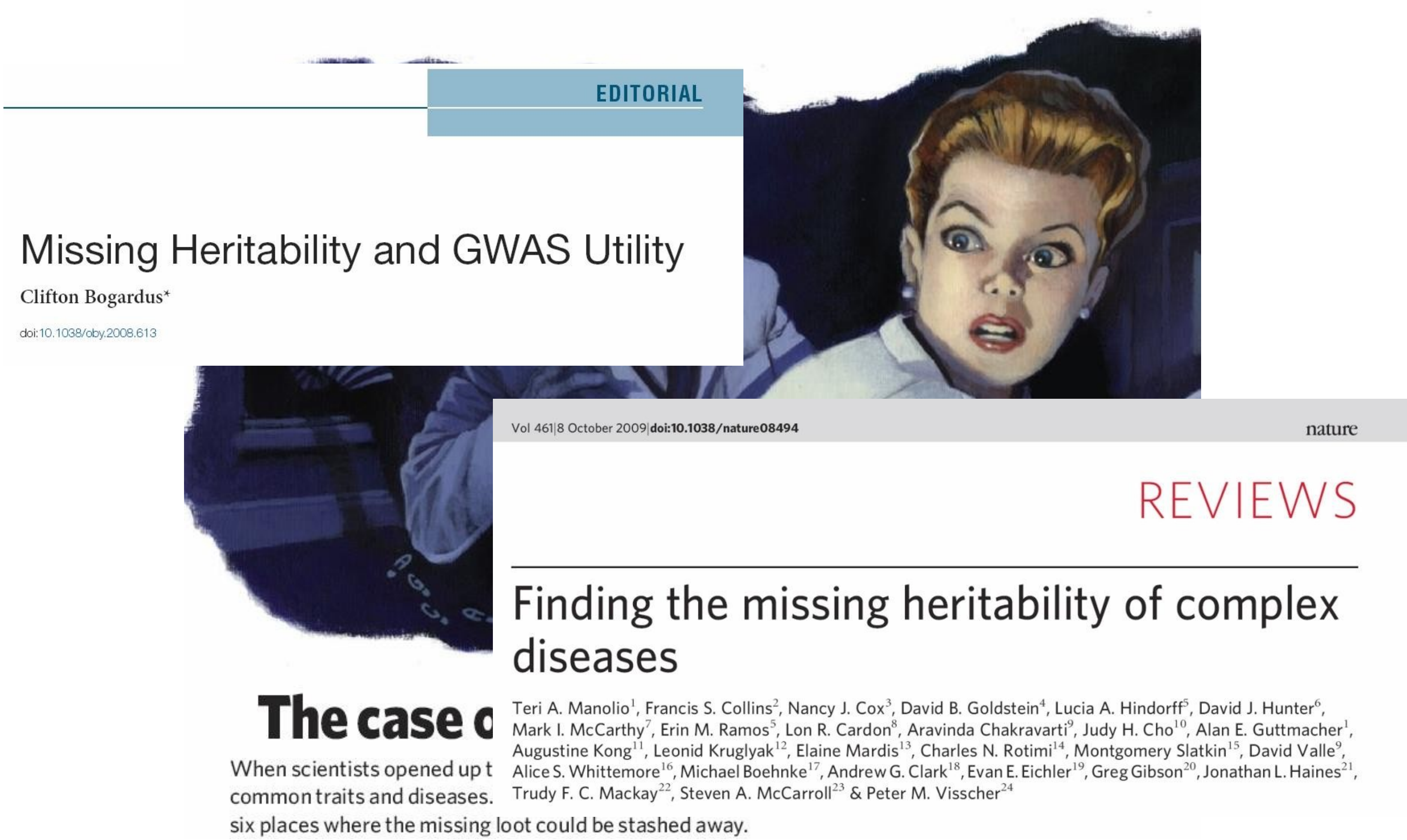
BY ANTONIO REGALADO

Missing Heritability?

The case of the missing heritability

When scientists opened up the human genome, they expected to find the genetic components of common traits and diseases. But they were nowhere to be seen. **Brendan Maher** shines a light on six places where the missing loot could be stashed away.

Missing Heritability?



EDITORIAL

Missing Heritability and GWAS Utility

Clifton Bogardus*

doi:10.1038/oby.2008.613

The case o

When scientists opened up t
common traits and diseases.
six places where the missing loot could be stashed away.

Vol 461|8 October 2009|**doi:10.1038/nature08494** nature

REVIEWS

Finding the missing heritability of complex diseases

Teri A. Manolio[1], Francis S. Collins[2], Nancy J. Cox[3], David B. Goldstein[4], Lucia A. Hindorff[5], David J. Hunter[6], Mark I. McCarthy[7], Erin M. Ramos[5], Lon R. Cardon[8], Aravinda Chakravarti[9], Judy H. Cho[10], Alan E. Guttmacher[1], Augustine Kong[11], Leonid Kruglyak[12], Elaine Mardis[13], Charles N. Rotimi[14], Montgomery Slatkin[15], David Valle[9], Alice S. Whittemore[16], Michael Boehnke[17], Andrew G. Clark[18], Evan E. Eichler[19], Greg Gibson[20], Jonathan L. Haines[21], Trudy F. C. Mackay[22], Steven A. McCarroll[23] & Peter M. Visscher[24]

Variation explained

- The variation explained is yet very small for many traits

- By doubling the sample size, the number of identified loci is more than double, however, the % variance explained is normally increased ~ 50% (rule of thumb!)

Reasons for missing heritability

- “Common disease, common variant” is incorrect – study rarer variants

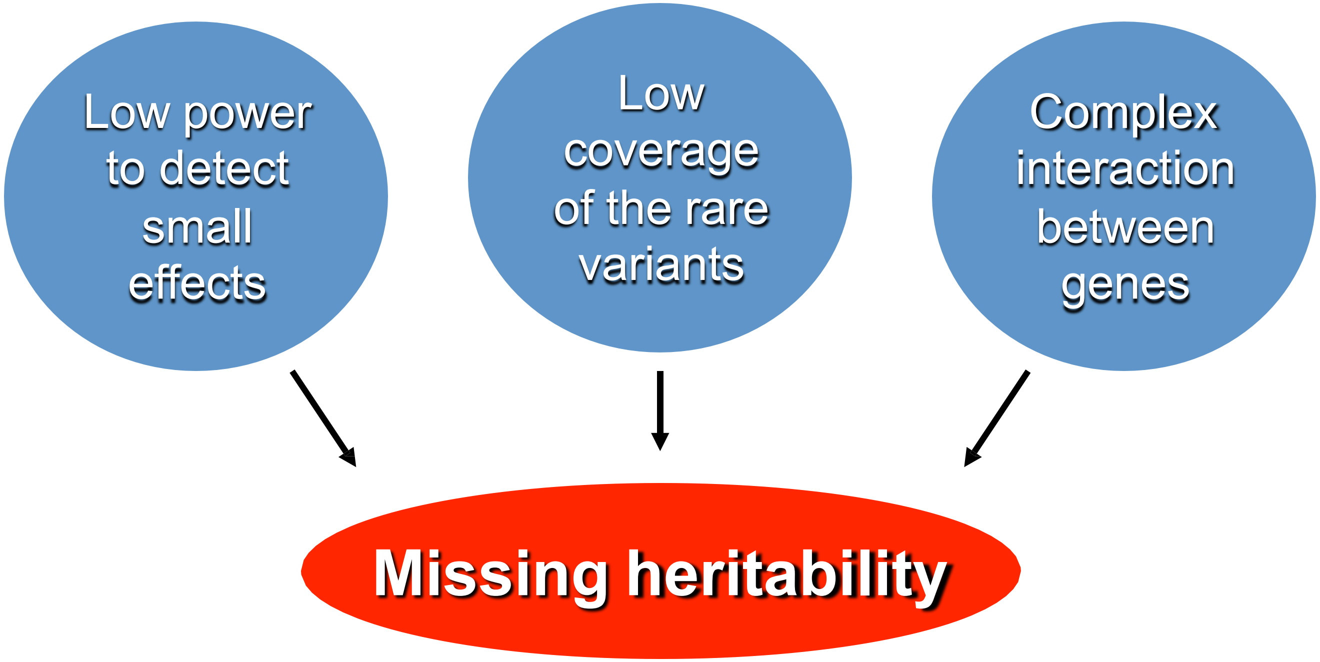
- Calculation of heritability effects is wrong?

- Not enough common variants of small effect detected

- Structural or other genomic variants more important

- Difficult to analyse gene-gene/gene-environment interactions and in general high-dimensional and systems biology data (i.e., combination of genomic, transcriptomic, proteomic, metabolomic data)

Reasons for missing heritability



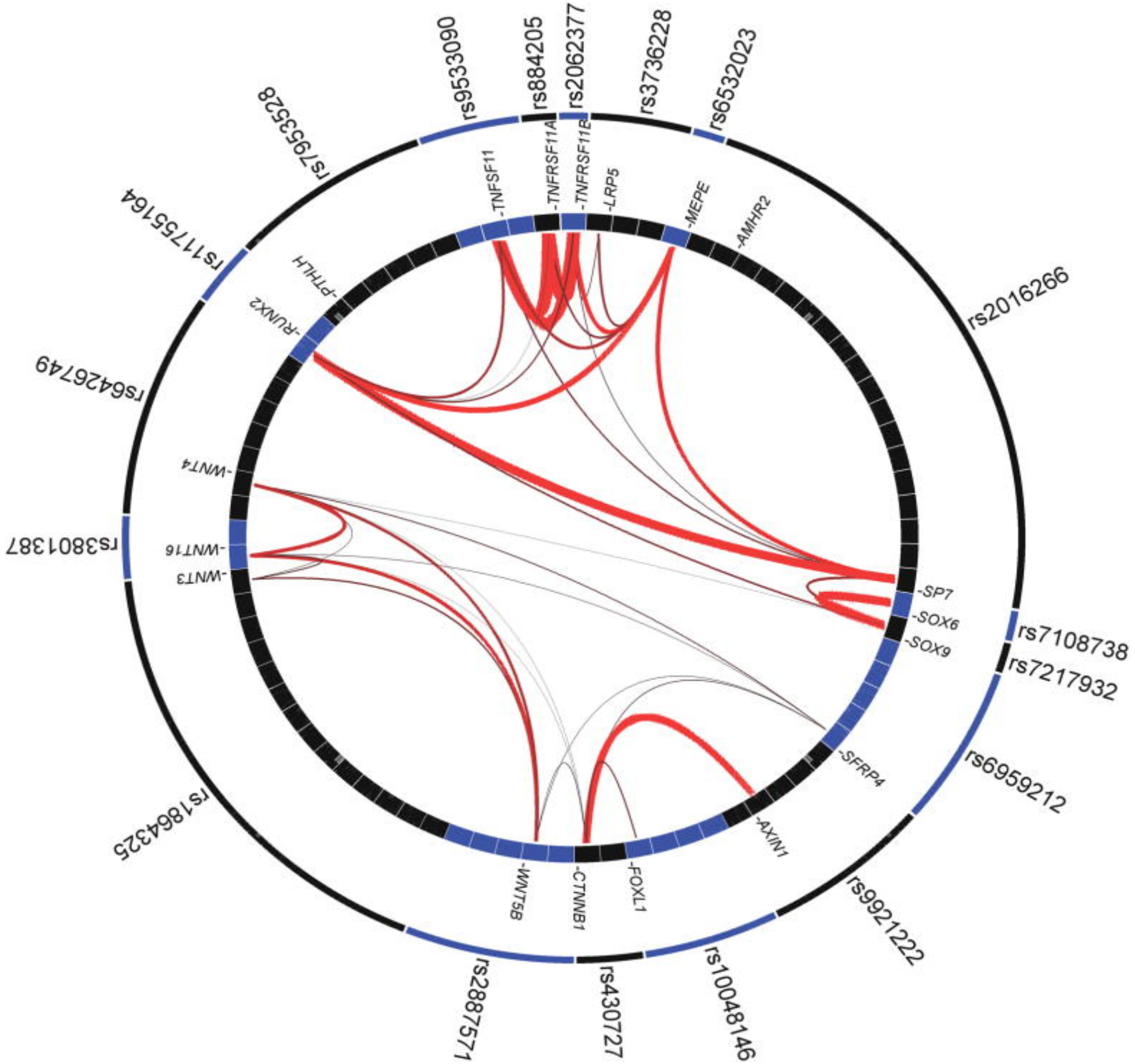
Low power to detect small effects

Low coverage of the rare variants

Complex interaction between genes

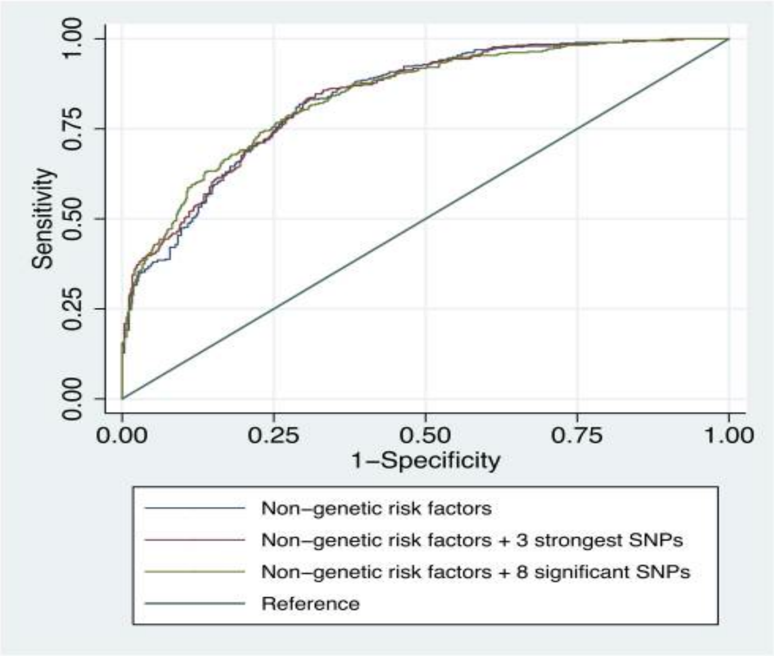
Missing heritability

Gene-Gene interactions



Large sample sizes are required to support evidence of gene-gene interactions

Gene-environment interactions



Stefanaki I et al. PLoS One; 2013

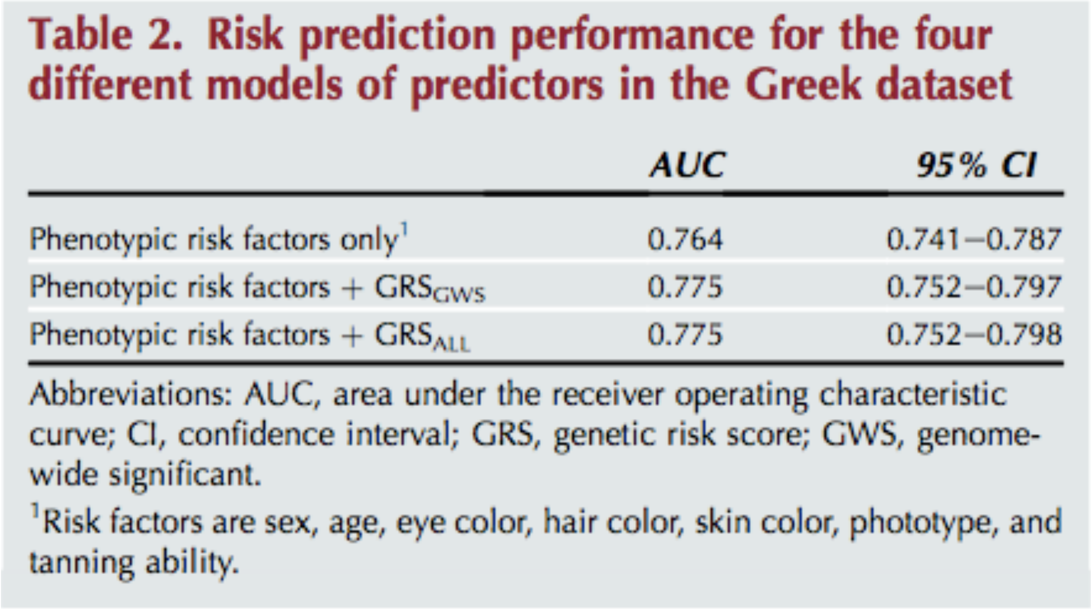


Table 2. Risk prediction performance for the four different models of predictors in the Greek dataset

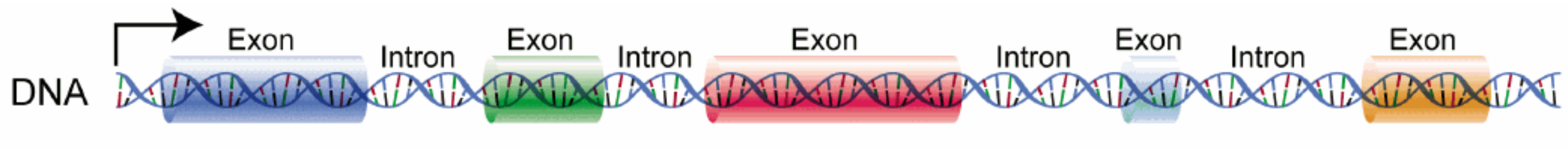
	AUC	*95% CI*
Phenotypic risk factors only[1]	0.764	0.741–0.787
Phenotypic risk factors + GRS_{GWS}	0.775	0.752–0.797
Phenotypic risk factors + GRS_{ALL}	0.775	0.752–0.798

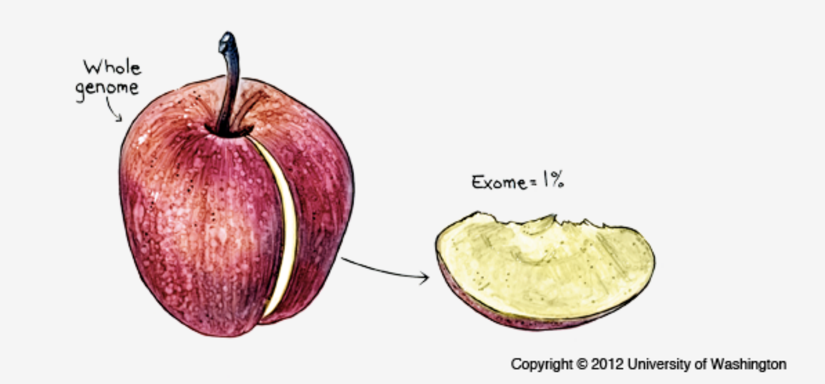
Abbreviations: AUC, area under the receiver operating characteristic curve; CI, confidence interval; GRS, genetic risk score; GWS, genome-wide significant.

[1] Risk factors are sex, age, eye color, hair color, skin color, phototype, and tanning ability.

Kypreou KP et al. J Invest Dermatol; 2016

Whole exome and whole genome sequencing





Copyright © 2012 University of Washington

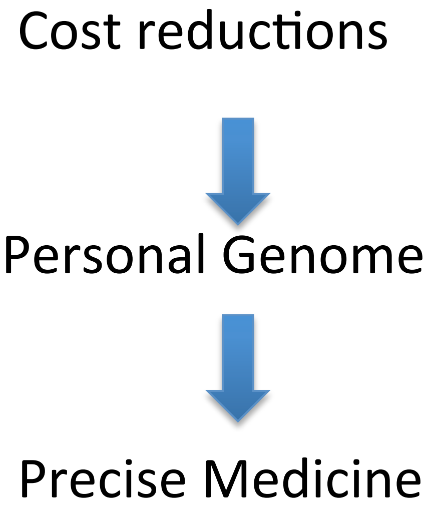
Human Genome Epidemiology (HuGE) Review

Genome-wide Significant Associations for Variants With Minor Allele Frequency of 5% or Less—An Overview: A HuGE Review

Orestis A. Panagiotou, Evangelos Evangelou, and John P. A. Ioannidis*

In the near future

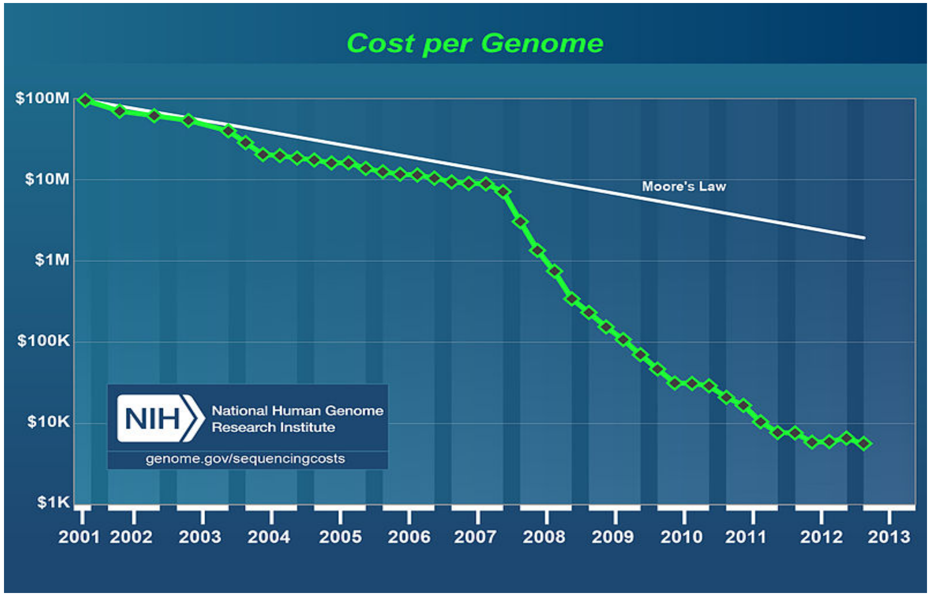
- Exome sequencing-Whole genome sequencing



Cost reductions

Personal Genome

Precise Medicine



Clinical assessment incorporating a personal genome

Euan A Ashley, Atul J Butte, Matthew T Wheeler, Rong Chen, Teri E Klein, Frederick E Dewey, Joel T Dudley, Kelly E Ormond, Aleksandra Pavlovic, Alexander A Morgan, Dmitry Pushkarev, Norma F Neff, Louanne Hudgins, Li Gong, Laura M Hodges, Dorit S Berlin, Caroline F Thorn, Katrin Sangkuhl, Joan M Hebert, Mark Woon, Hersh Sagreiya, Ryan Whaley, Joshua W Knowles, Michael F Chou, Joseph V Thakuria, Abraham M Rosenbaum, Alexander Wait Zaranek, George M Church, Henry T Greely, Stephen R Quake, Russ B Altman

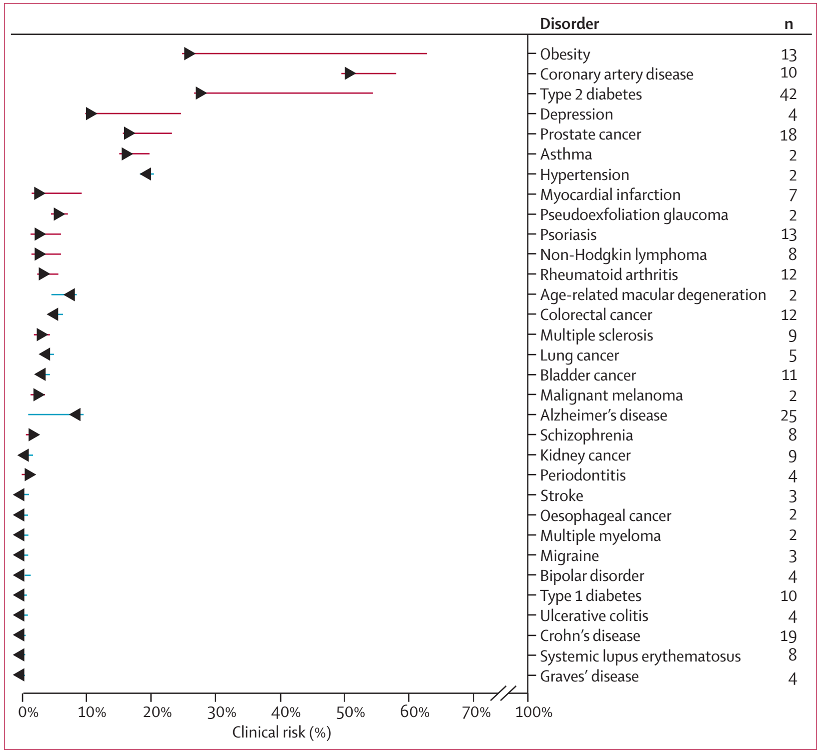


Figure 3: **Clinical risk incorporating genetic-risk estimates for major diseases**

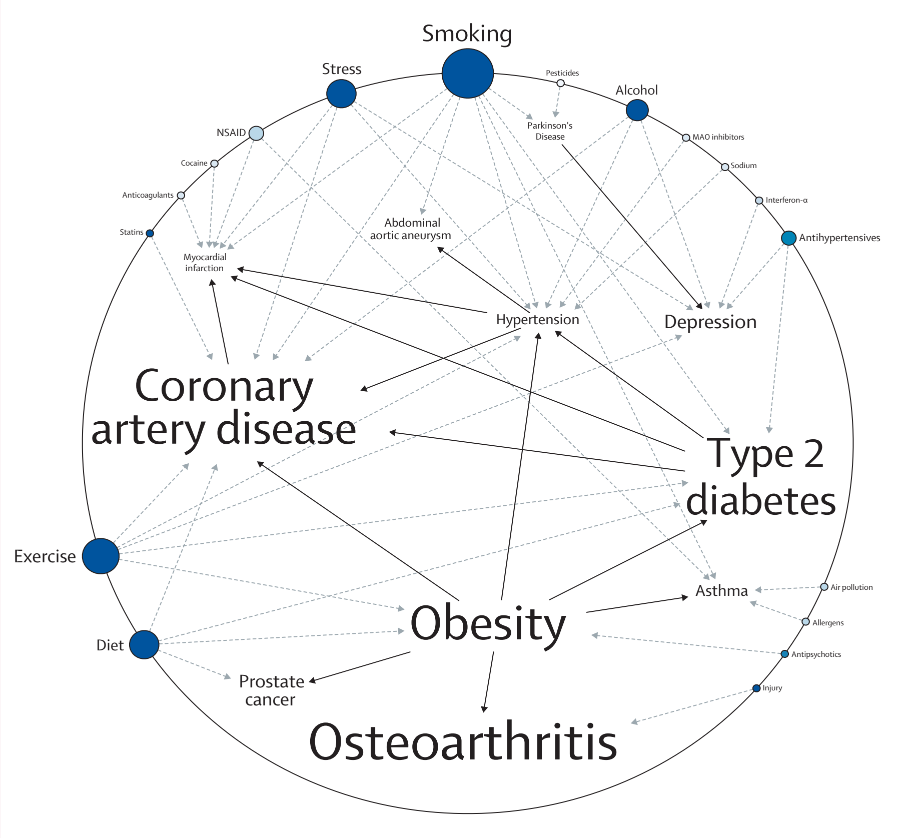
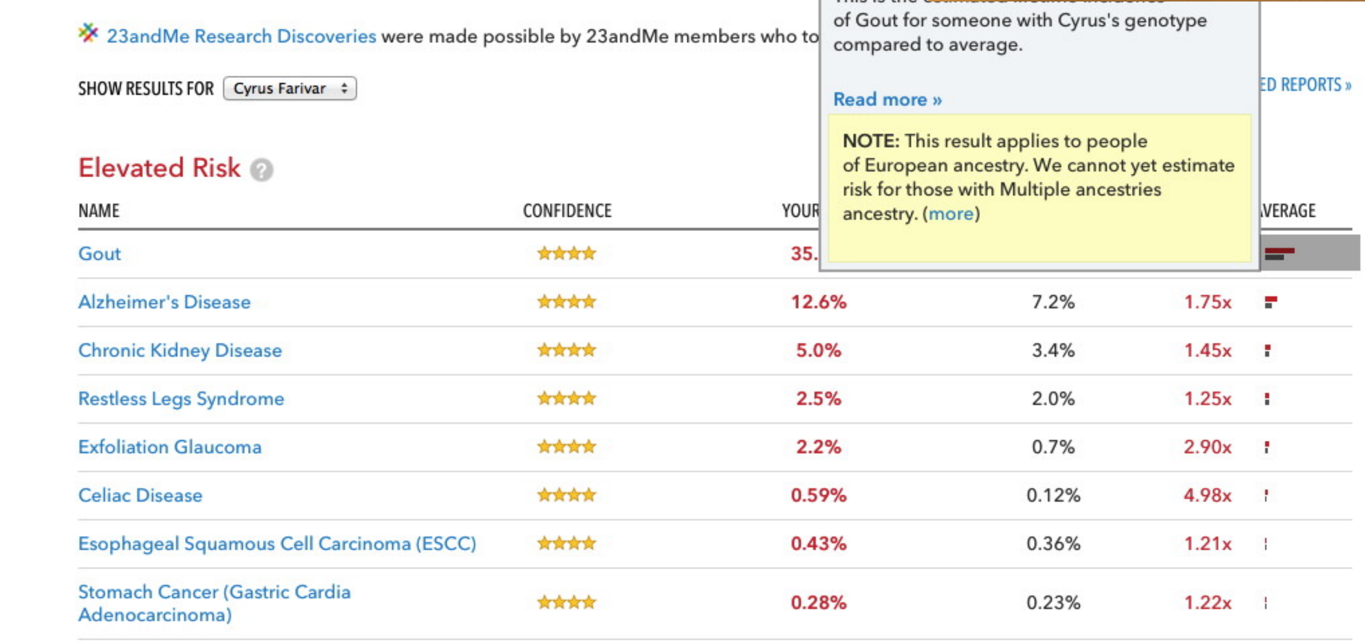


Figure 5: **Gene-environment interaction**

Before providing your sample, register your kit at:
www.23andme.com/start
Your sample will NOT be processed unless it is registered.
STOP
Register this kit now at:
www.23andme.com/start
Your sample will NOT be processed unless it is registered.
1 Register
2 Fill line
3 Close funnel
4 Remove funnel
5 Screw on cap
6 Shake 5 seconds
7 Seal in bag
8 Mail in box

HOME MY RESULTS FAMILY & FRIENDS RESEARCH & COMMUNITY

HEALTH RISKS



23andMe Research Discoveries were made possible by 23andMe members who to

SHOW RESULTS FOR Cyrus Farivar

Cyrus:

Average:

This is the e

of Gout for someone with Cyrus's genotype compared to average.

Read more »

NOTE: This result applies to people of European ancestry. We cannot yet estimate risk for those with Multiple ancestries ancestry. (more)

ED REPORTS »

Elevated Risk

NAME	CONFIDENCE	YOUR			VERAGE
Gout	★★★★	35.			▬
Alzheimer's Disease	★★★★	12.6%	7.2%	1.75x	▬
Chronic Kidney Disease	★★★★	5.0%	3.4%	1.45x	▬
Restless Legs Syndrome	★★★★	2.5%	2.0%	1.25x	▬
Exfoliation Glaucoma	★★★★	2.2%	0.7%	2.90x	▬
Celiac Disease	★★★★	0.59%	0.12%	4.98x	▬
Esophageal Squamous Cell Carcinoma (ESCC)	★★★★	0.43%	0.36%	1.21x	▬
Stomach Cancer (Gastric Cardia Adenocarcinoma)	★★★★	0.28%	0.23%	1.22x	▬

Ways forward…

- Further genetic discovery (larger sample size)
- Denser genotyping
- Whole genome sequencing
- Systems biology approaches
- Development of clinically useful risk prediction models
- Other translation