KSBi-BIML 2023

Bioinformatics & Machine Learning(BIML)
Workshop for Life Scientists, Data Scientists,
and Bioinformatians

생물정보학&머신러닝 워크샵(온라인)

Introduction to genome-wide association studies

원홍희 _ 성균관대학교

SBi 한국생명정보학회
Korean Society for Bioinformatics

본 강의 자료는 한국생명정보학회가 주관하는 BIML 2023 워크샵 온라인 수업을 목적으로 제작된 것으로 해당 목적 이외의 다른 용도로 사용할 수 없음을 분명하게 알립니다.

이를 다른 사람과 공유하거나 복제, 배포, 전송할 수 없으며 만약 이러한 사항을 위반할 경우 발생하는 **모든 법적 책임은 전적으로 불법 행위자 본인에게 있음을 경고**합니다.

KSBi-BIML 2023

Bioinformatics & Machine Learning (BIML)
Workshop for Life Scientists, Data Scientists, and Bioinformatians

안녕하십니까?

한국생명정보학회가 개최하는 동계 교육 워크샵인 BIML-2023에 여러분을 초대합니다. 생명정보학 분야의 연구자들에게 최신 동향의 데이터 분석기술을 이론과 실습을 겸비해 전달하고자 도입한 전문 교육 프로그램인 BIML 워크샵은 2015년에 시작하여 올해로 9차를 맞이하게 되었습니다. 지난 2년간은 심각한 코로나 대유행으로 인해 아쉽게도 모든 강의가 온라인으로 진행되어 현장 강의에서만 가능한 강의자와 수강생 사이에 다양한 소통의 기회가 없음에 대한 아쉬움이 있었습니다. 다행히도 최근 사회적 거리두기 완화로 현장 강의가 가능해져 올해는 현장 강의를 재개함으로써 온라인과 현장 강의의 장점을 모두 갖춘 프로그램을 구성할 수 있게 되었습니다.

BIML 워크샵은 전통적으로 크게 인공지능과 생명정보분석 두 개의 분야로 구성되었습니다. 올해 AI 분야에서는 최근 생명정보 분석에서도 응용이 확대되고 있는 다양한 심층학습(Deep learning) 기법들에 대한 현장 강의가 진행될 예정이며, 관련하여 심층학습을 이용한 단백질구조예측, 유전체 분석, 신약개발에 대한 이론과 실습 강의가 함께 제공될 예정입니다. 또한 싱글셀오믹스 분석과 메타유전체분석 현장 강의는 많은 연구자의 연구 수월성 확보에 큰 도움을 줄 것으로 기대하고 있습니다. 이외에 다양한 생명정보학 분야에 대하여 30개 이상의 온라인 강좌가 개설되어 제공되며 온라인 강의의 한계를 극복하기 위해서 실시간 Q&A 세션 또한 마련했습니다. 특히 BIML은 각 분야 국내 최고 전문가들의 강의로 구성되어 해당 분야의 기초부터 최신 연구 동향까지 포함하는 수준 높은 내용의 강의가 될 것입니다.

이번 BIML-2023을 준비하기까지 너무나 많은 수고를 해주신 BIML-2023 운영위원회의 남진우, 우현구, 백대현, 정성원, 정인경, 장혜식, 박종은 교수님과 KOBIC 이병욱 박사님께 커다란 감사를 드립니다. 마지막으로 부족한 시간에도 불구하고 강의 부탁을 흔쾌히 허락하시고 훌륭한 현장 강의와 온라인 강의를 준비하시는데 노고를 아끼지 않으신 모든 연사분께 깊은 감사를 드립니다.

2023년 2월

한국생명정보학회장 이 인 석

강의개요

Introduction to genome-wide association studies

전장유전체연관분석(GWAS, genome-wide association studies)은 인간 질병이나 형질과 연관된 유전변이를 발굴하고 유전적 조성을 규명하는 대표적인 연구 방법론이다. 그 동안 전세계에서 진행된 대규모 GWAS 연구들은 다양한 형질과 연관된 유전 변이를 발굴하였고 이러한 변이들은 형질의 유전력을 상당 부분 설명하게 되었다. 나아가, 대규모 GWAS 분석 결과(GWAS summary statistics)가 공유됨에 따라, 유전력(heritability), 질병 간 유전적 상관성(genetic correlation), 다인자유전점수(polygenic risk score), 멘델리안 무작위법(Mendelian randomization) 등 여러 post-GWAS 분석이 가능하게 되었고 질병의 유전적 조성을 이해하는데 핵심적인 정보를 제공하고 있다.

본 강의에서는 GWAS를 중심으로 한 유전체 분석의 배경, 이론 및 분석 방법론 등을 소개하고, 복합 질환에서 최근 GWAS 연구 결과를 소개하고자 한다. 이를 통해 GWAS 기반의 연구를 해석하기 위한 기초 지식을 쌓고, 나아가 GWAS 분석 및 GWAS 결과의 응용 연구를 위한 핵심 역량을 갖추는 것을 목표로 한다.

강의는 다음의 내용을 포함한다:

- 유전체 분석을 위한 개념
- GWAS 분석의 이론과 방법론
- Post-GWAS 분석의 이론과 방법론
- 대표적인 연구 결과의 소개

* 참고강의교재:

  Tam et al. Benefits and limitations of genome-wide association studies, Nature Reviews Genetics, 20:467-484, 2019.

  Balding. A tutorial on statistical methods for population association studies, Nature Reviews Genetics, 7:781-791, 2006.

  이종극, 질병 유전체 분석법 3판

* 강의 난이도: 초급

* 강의: 원홍희 교수 (성균관대학교 삼성융합의과학원)

Curriculum Vitae

Speaker Name: Hong-Hee Won, Ph.D.

▶ **Personal Info**

Name	Hong-Hee Won
Title	Associate Professor
Affiliation	Sungkyunkwan University

▶ **Contact Information**

Address	81, Irwon-Ro, Gangnam-Gu, Seoul, 06351
Email	wonhh@skku.edu
Phone Number	010-6326-3452

Research Interest

Population genomics, genome-wide association study, polygenic risk score

Educational Experience

2002	B.S. in Computer Science, Yonsei University, Korea
2004	M.S. in Computer Science, Yonsei University, Korea
2011	Ph.D. in Bioinformatics, KAIST, Korea

Professional Experience

2004-2012	Research Scientist, Samsung Biomedical Research Institute and Samsung Medical Center, Korea
2012-2015	Research Fellow, Massachusetts General Hospital, Harvard Medical School, and Broad Institute of MIT and Harvard, USA
2016-2020	Assistant Professor, Sungkyunkwan University, Samsung Medical Center, Korea
2020-	Associate Professor, Sungkyunkwan University, Samsung Medical Center, Korea

Selected Publications (5 maximum)

1. Kim S, et al. Shared genetic architectures of subjective well-being in East Asian and European ancestry populations, Nature Human Behaviour, 6(7):1014-1026, 2022.
2. Kim M, et al. Association between adiposity and cardiovascular outcomes: an umbrella review and meta-analysis of observational and Mendelian randomization studies, European Heart Journal, 42(34):3388-3403, 2021.
3. Khera AV, et al. Association of rare and common variation in the lipoprotein lipase gene with coronary artery disease, Journal of the American Medical Association JAMA, 317(9):937-46, 2017.
4. Do R, et al. Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction, Nature 518:102-106, 2015.
5. Stitziel NO, et al. Inactivating mutations in NPC1L1 and protection from coronary heart disease, New England journal of medicine NEJM, 371(22):2072-2082, 2014.

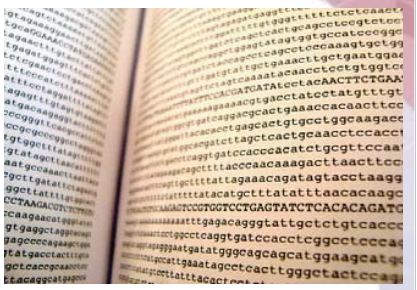
SBi 한국생명정보학회 Korean Society for Bioinformatics

Introduction to genome-wide association studies

원홍희, Ph.D.
honghee.won@gmail.com
성균관대학교 삼성융합의과학원
삼성서울병원

The Human Genome

- Instruction manual for human cells
- A book with 3.2 billion letters in 23 chapters or chromosomes



- 20,000 genes, exome (1% of the genome)
- 99.9% identical, 4 million letters are different
  - Variation, variant, mutation, polymorphism

Genetic variation affects phenotype (and risk for disease)

- Genetic variants
  - Pathogenic variants
    - Disease-causing, deleterious, damaging
    - Usually rare (<1%)
    - Often, referred to as “Mutations”
  - Neutral variants
    - Non-disease causing, but may affect disease susceptibility
    - Usually common (>5%)
    - Often, referred to as “Polymorphisms”
      - SNP (single nucleotide polymorphism)

누구나 수백만의 germline 유전 변이를 갖고 있다.

- Single nucleotide variants (SNV)
  - 단일 염기 변이 : 4백만개/사람

일생동안 변하지않는다

- Multi-nucleotide variants
  - Small insertions/deletions (indels) : 50만개/사람
  - Large copy number variants (CNVs)
  - Inversions
  - Translocations
  - Aneuploidy

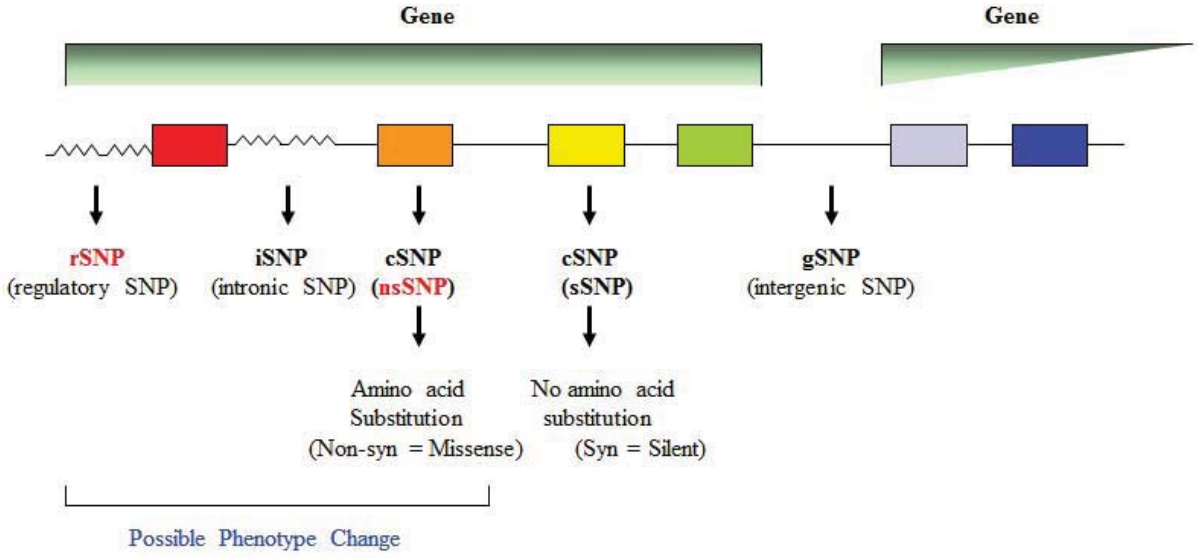
AAATAGC**A**CCGTTAGC
AAATAGC**C**CCGTTAGC

SNV 예

AAATAGCACCGTTAGC
AAATA-----GTTAGC

indels 예

Classification of SNP by Location

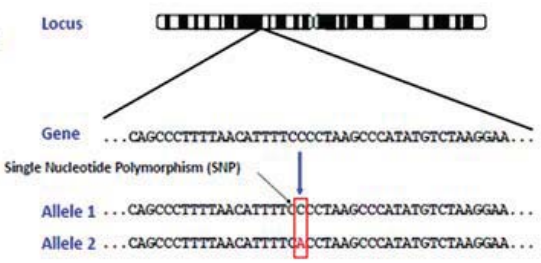


Definitions: Genes and Genotypes

Gene: **A unit of inheritance that is transmitted from parents to offspring, or**
A region of DNA that codes for a specific product (protein or RNA)

Locus (loci): **The place where a particular gene resides on a chromosome**
Gene ≈ Locus (interchangeable)

Allele: **Different forms of a gene, or**
Variants of DNA at a given locus



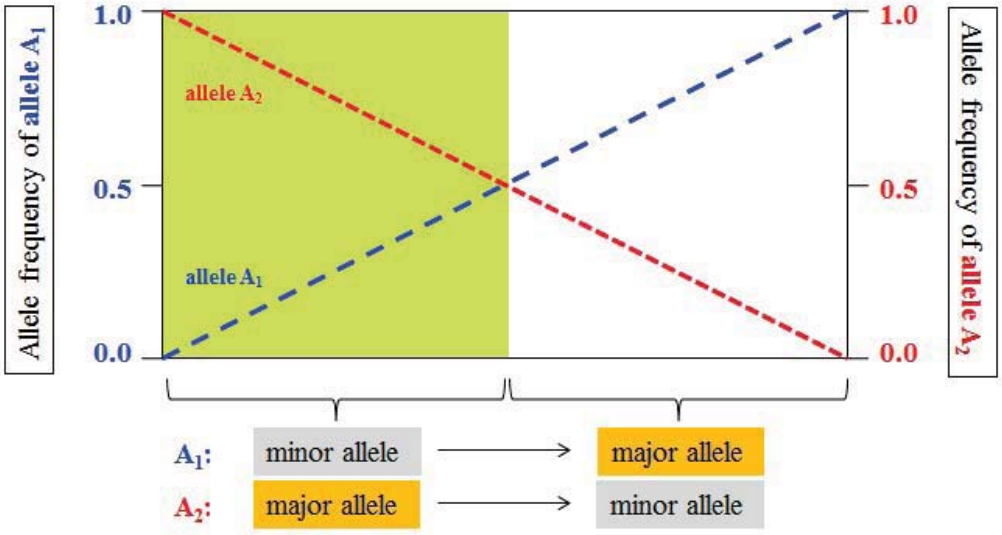
Genotype: **The diploid (pair) of alleles (except the X chromosome in males)**

In diploid individual, homozygote (wild type)= 2 identical (wild type) alleles
heterozygote (wild + mutant type)= 2 different alleles
homozygote (mutant type)= 2 identical (mutant type) alleles

In haploid (egg or sperm cells), cells have only one copy of a gene

Minor Allele Frequency (MAF)

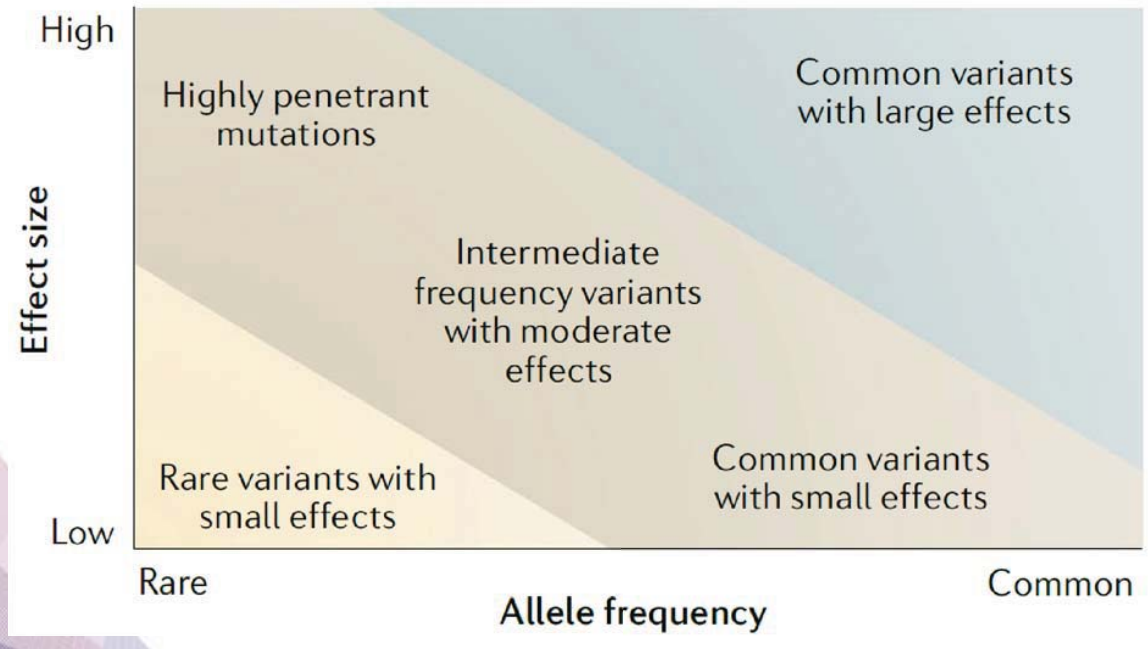
SNP : bi-allele (A_1/A_2) DNA marker



• **Minor Allele Frequency** (MAF)= a standard index for genetic diversity of a SNP marker.
* Maximum MAF= 0.5

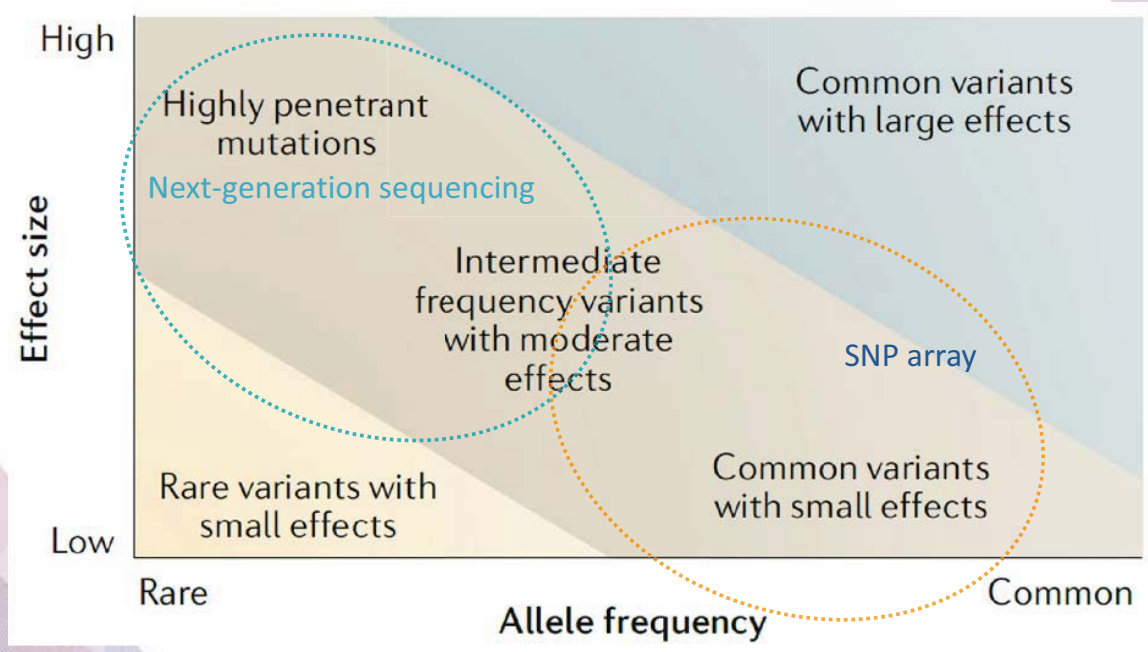
질병유전체분석법3

Variants by frequency and effect size



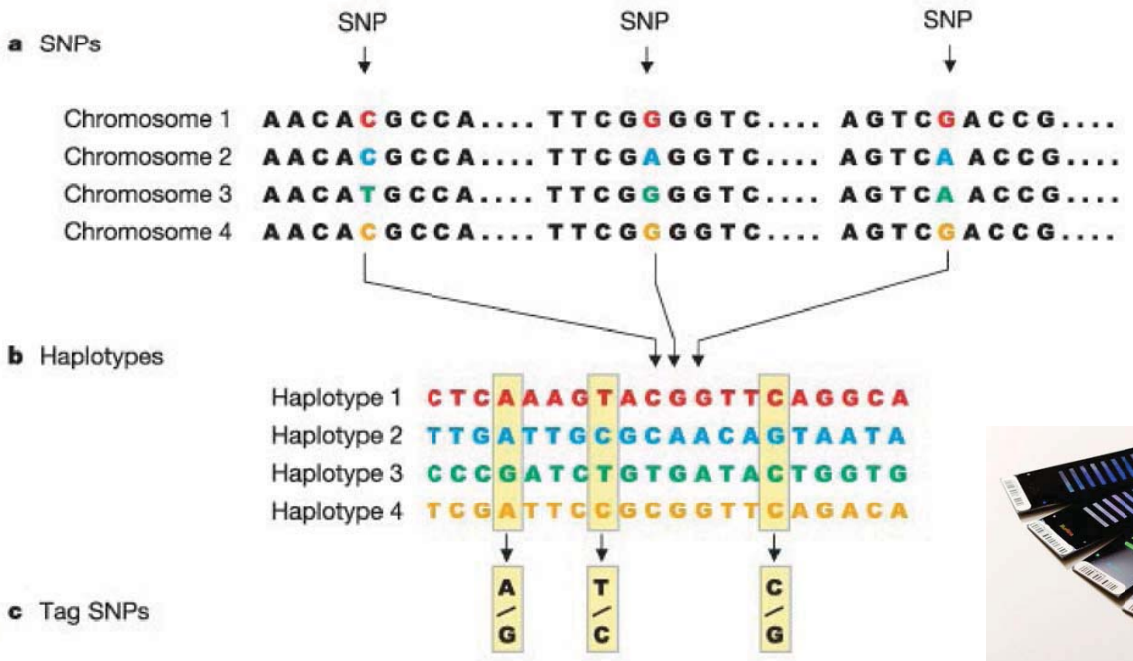
8

Variants by frequency and effect size



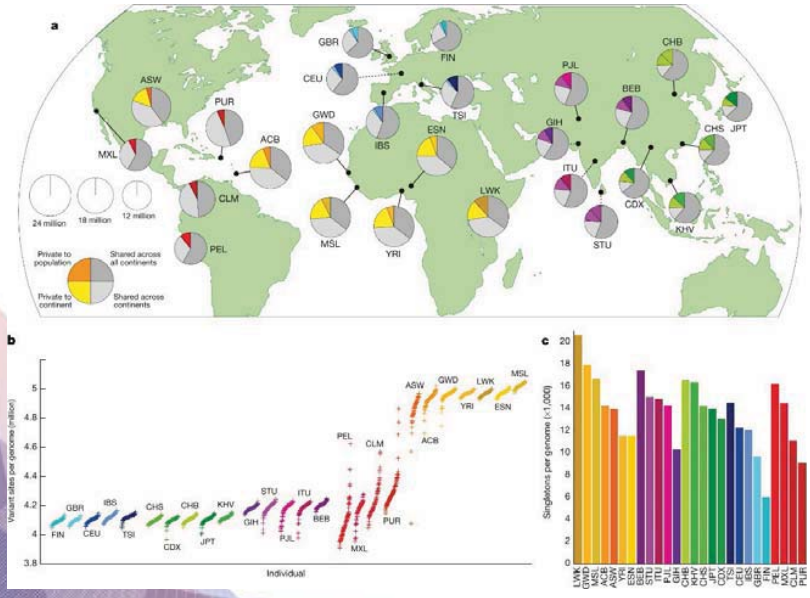
9

Int. HapMap Project (2002-2009)

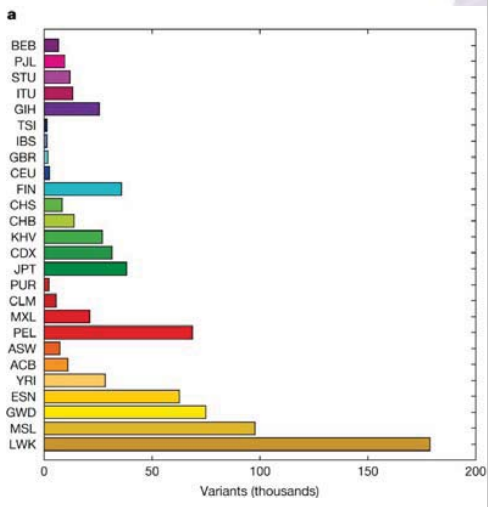


1000 Genomes Project

- Sequenced the genomes of 2,504 individuals from 26 populations in Africa (AFR), East Asia (EAS), Europe (EUR), South Asia (SAS), and the Americas (AMR)



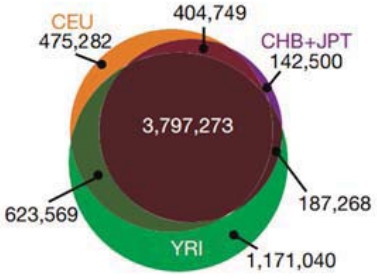
Variants found to be rare (<0.5%) within the global sample but common within a population



A Auton *et al*. *Nature* **526**, 68-74 (2015) doi:10.1038/nature15393

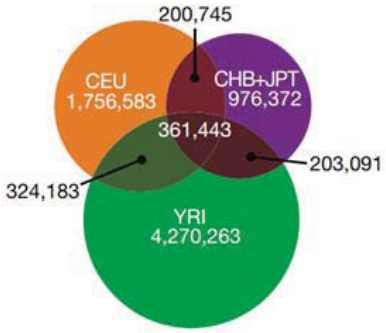
Rare SNVs are Much More Likely to be Population-Specific

Common SNVs



Average % of private SNPs between populations: **10.79%**

New Rare SNVs identified by sequencing

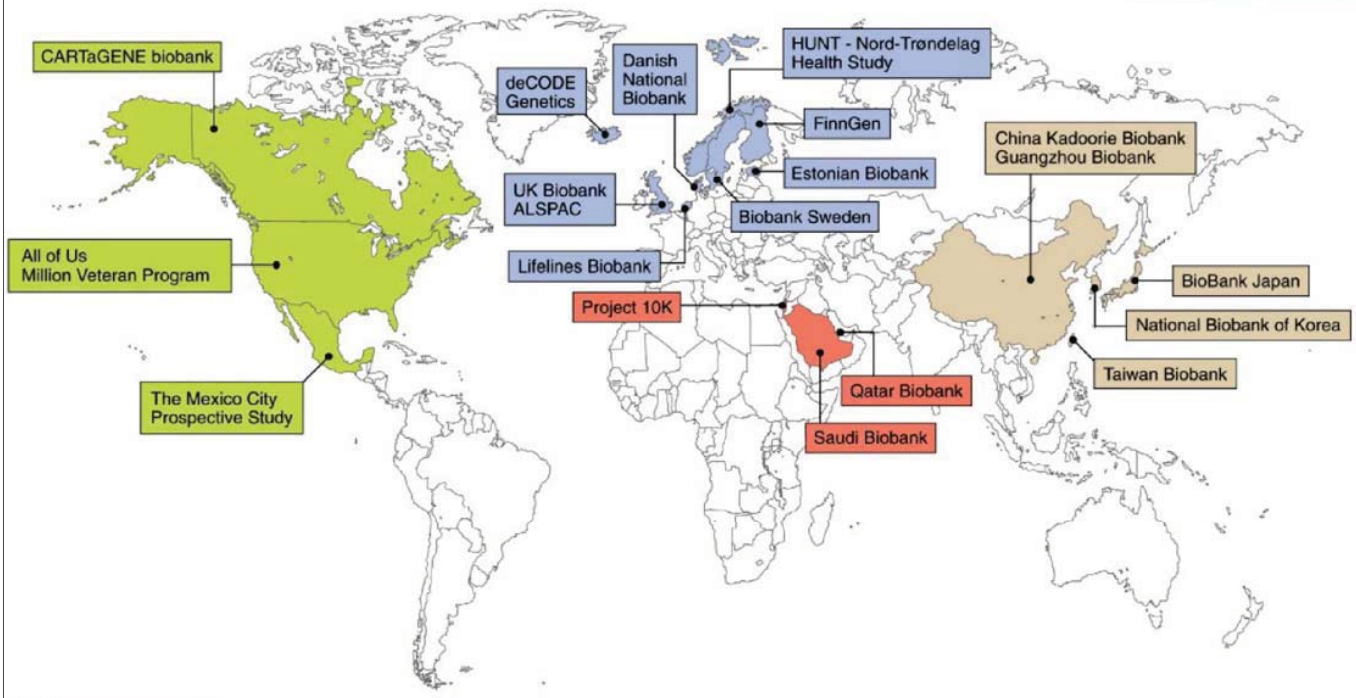


Average % of private SNPs between populations: **68.43%**

The 1000 genomes project consortium. Nature 467:1061-1074 (2010)

질병유전체분석법3

Global efforts made for genomic data – biobank



Nature Medicine 26, 29–38 (2020)

Genome-wide association study (GWAS)

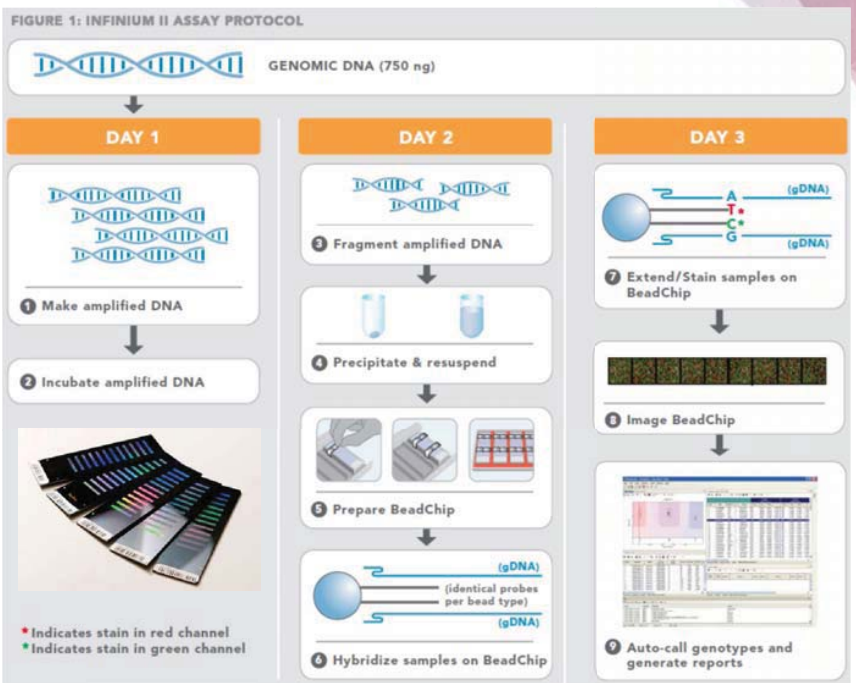
Summary of GWAS analysis and tools

- Quality control
  - Sample QC: PLINK
  - Variant QC: PLINK
  - Related samples: KING
  - PCA of genetic ancestry: EIGENSTRAT(smartpca)
- Imputation
  - Haplotype Reference Consortium: Michigan Imputation Server
  - TOPMed Imputation Server
- Association analysis
  - Logistic/linear regression (unrelated): PLINK
  - Mixed effects regression (including related): SAIGE, BOLT-LMM, REGENIE
- Visualization
  - QQ plot: CM-PLOT
  - Manhattan plot: CM-PLOT

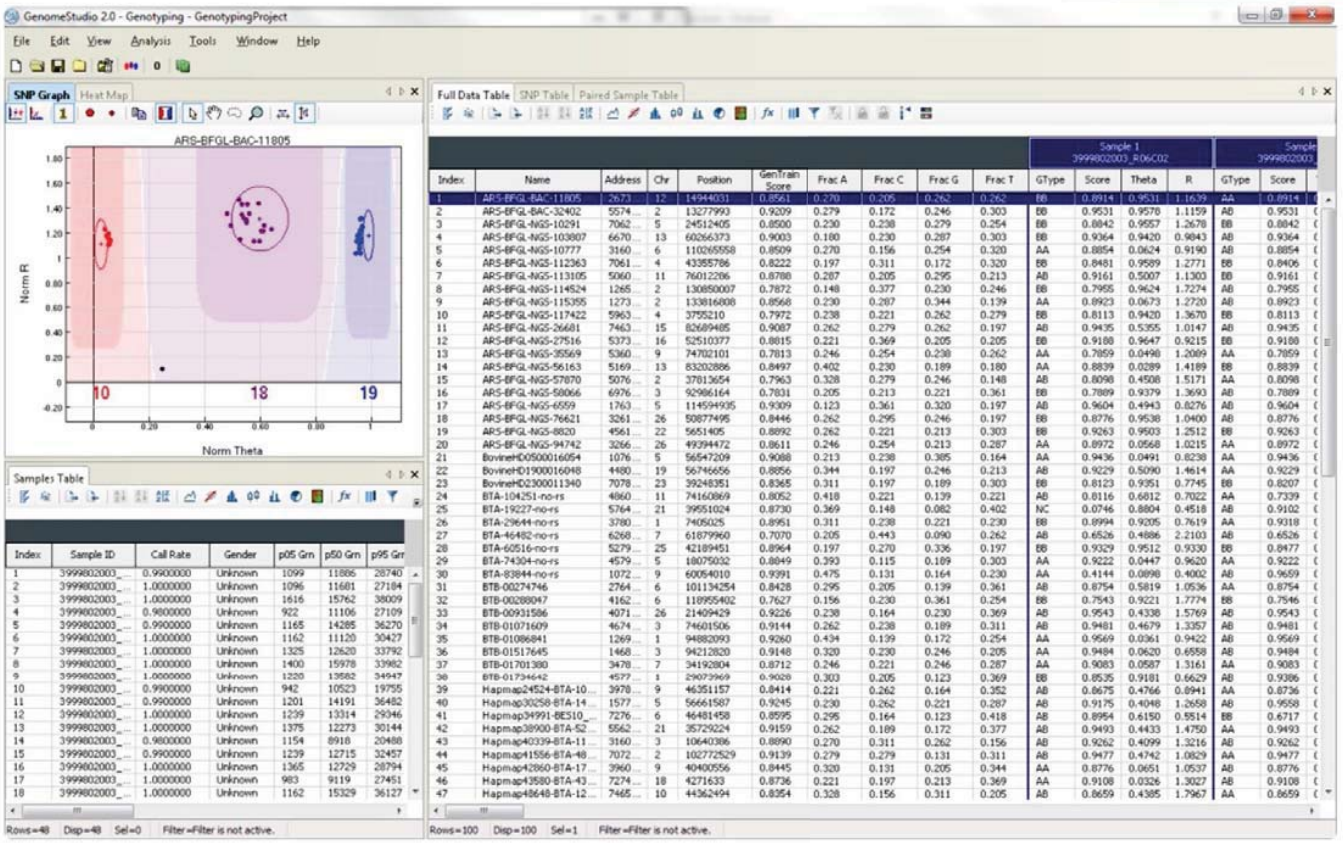
SNP arrays provide fast and accurate genotyping of about a million of genetic variants

Themo Fisher (Affymetrix)

Illumina

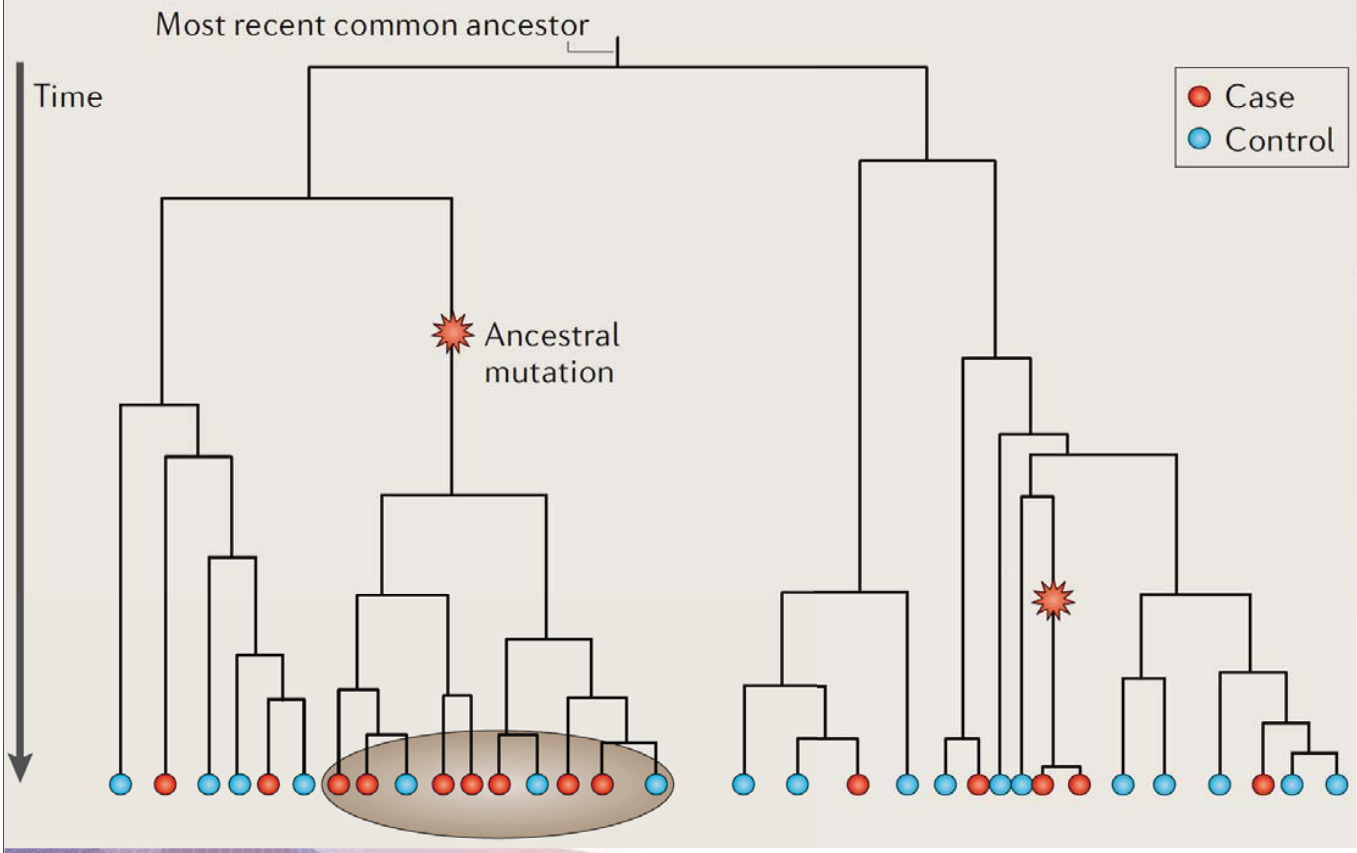


Illumina genotyping software

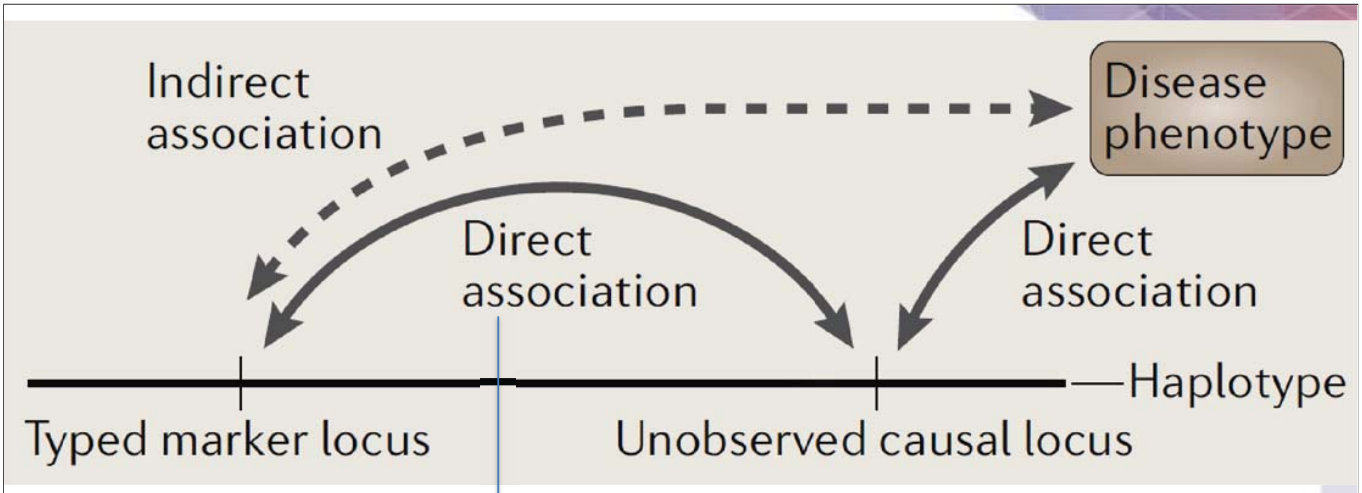


Genotypes are called for each sample (dot) by their signal intensity (Norm R) and Allele Frequency (Norm Theta) relative to canonical cluster positions (dark shading) for a given SNP marker.

Box 1 | **Rationale for association studies**



Ref: Balding, A tutorial on statistical methods for population association studies, *Nature Reviews Genetics*, 2006.



High linkage disequilibrium (LD)

- the degree to which alleles at two loci are associated
- correlation between two variants
- $r^2 = \dfrac{D^2}{p_1 * p_2 * q_1 * q_2}$
- $D = x_{11} - p_1 * q_1$
- $r^2 > 0.8$ considered "high" and $r^2 = 1$ "perfect LD"

Haplotype	Frequency
A_1B_1	x_{11}
A_1B_2	x_{12}
A_2B_1	x_{21}
A_2B_2	x_{22}

Allele	Frequency
A_1	$p_1 = x_{11} + x_{12}$
A_2	$p_2 = x_{21} + x_{22}$
B_1	$q_1 = x_{11} + x_{21}$
B_2	$q_2 = x_{12} = x_{22}$

	A_1	A_2	Total
B_1	$x_{11} = p_1q_1 + D$	$x_{21} = p_2q_1 - D$	q_1
$B2$	$x_{12} = p_1q_2 - D$	$x_{22} = p_2q_2 + D$	q_2
Total	p_1	p_2	

If two loci are in linkage equilibrium, D=0.
If two loci are in linkage disequilibrium, D>0.

Ref: Balding, A tutorial on statistical methods for population association studies, *Nature Reviews Genetics*, 2006.

Quality control of data is very important

- Sample QC
- Variant QC
- Population structure

PROTOCOL

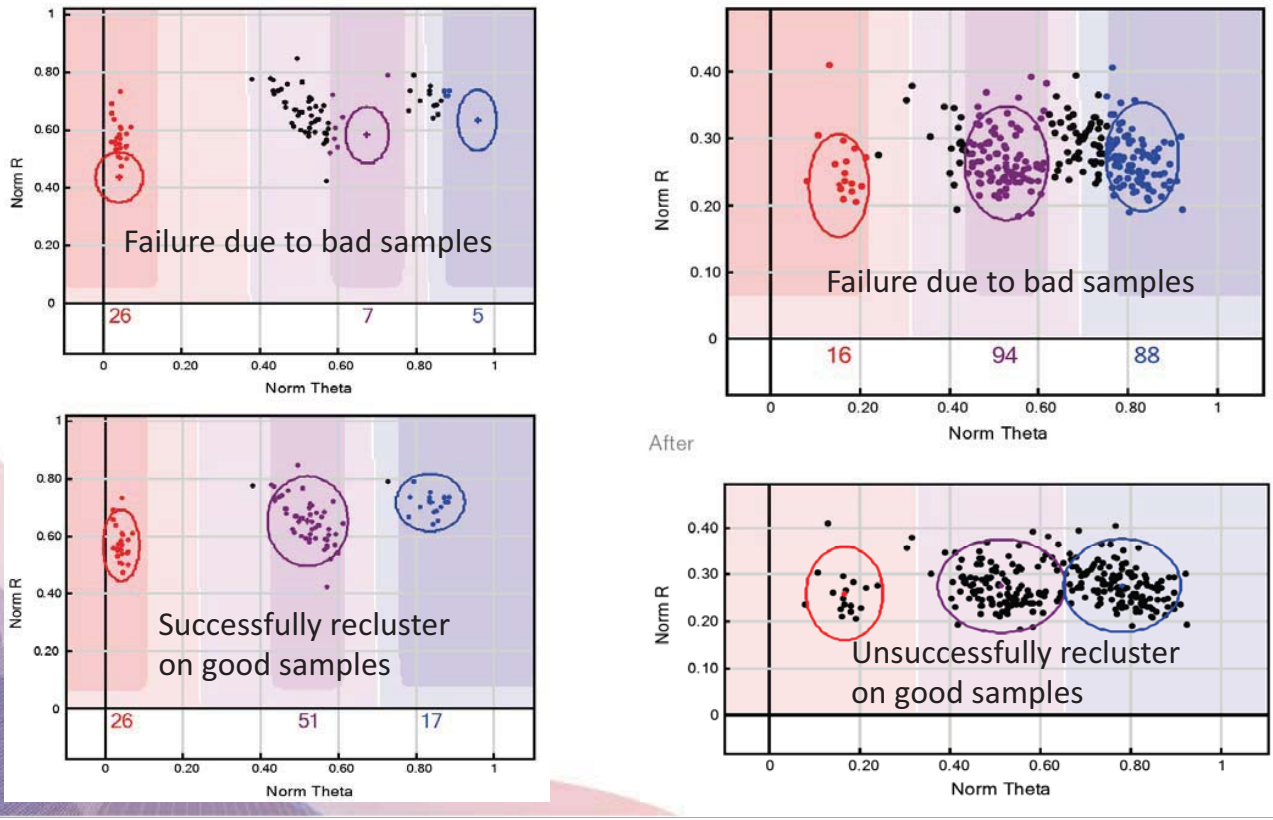
Data quality control in genetic case-control association studies

Carl A Anderson[1,2], Fredrik H Pettersson[1], Geraldine M Clarke[1], Lon R Cardon[3], Andrew P Morris[1] & Krina T Zondervan[1]

[1]Genetic and Genomic Epidemiology Unit, Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [2]Statistical Genetics, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. [3]GlaxoSmithKline, King of Prussia, Pennsylvania, USA. Correspondence should be addressed to C.A.A. (carl.anderson@sanger.ac.uk) or K.T.Z. (krinaz@well.ox.ac.uk).

Published online 26 August 2010; doi:10.1038/nprot.2010.116

Because clustering is not perfect for many reasons…



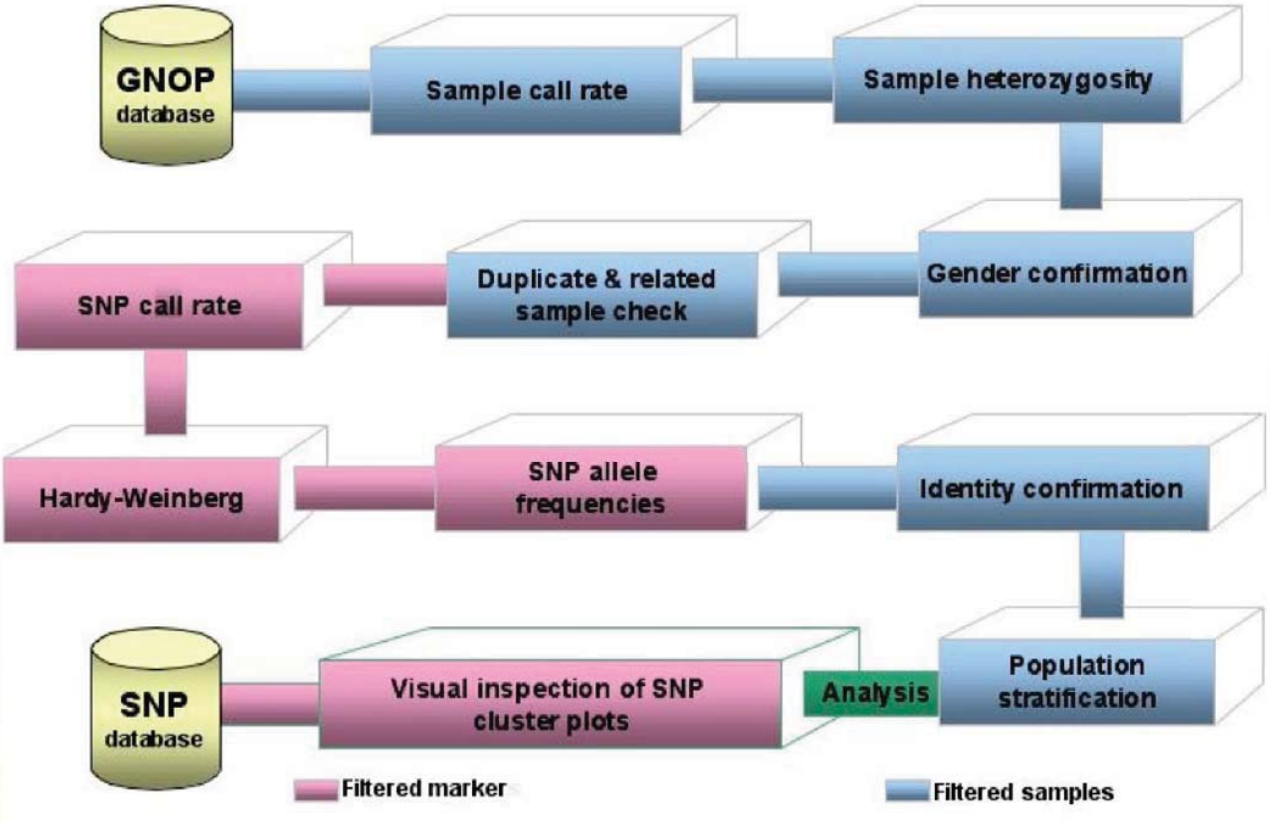
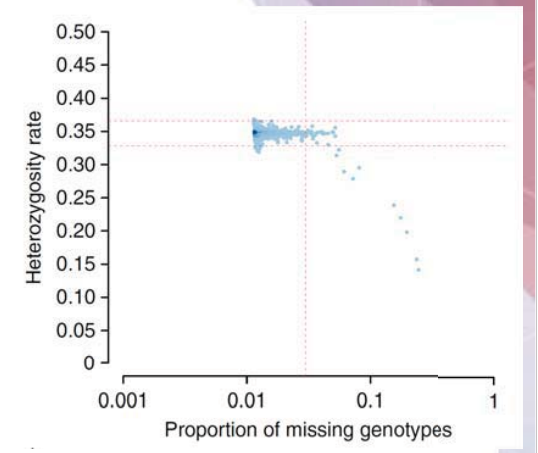
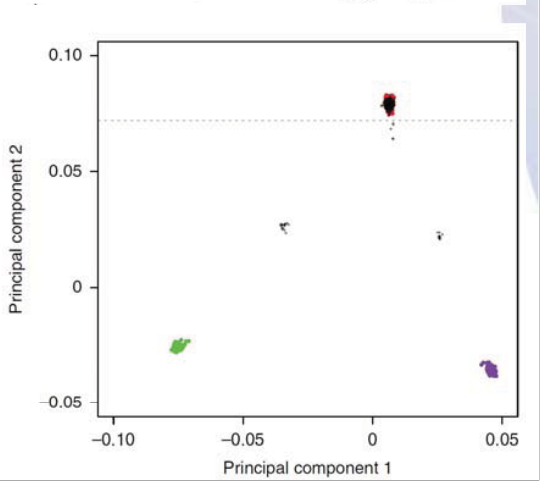


Figure 2: A chronological overview of the pipeline

Sample QC

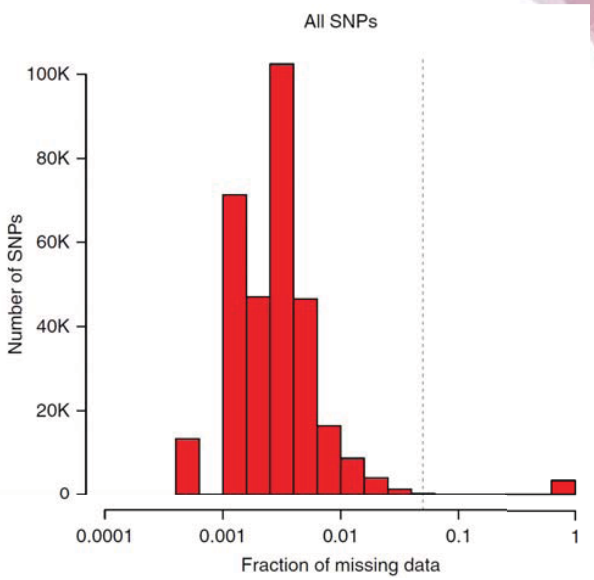
- High batch effects
- Low call rate (<95%)
- Excess of heterozygosity (>mean +5sd or <-5sd)
- Sex mismatch (btw. reported and estimated)
- Population stratification (sub-structure)
- (Hidden) familial relationships





Variant QC

- High batch effects
- Low call rate (<98%)
- Hardy-Weinberg equilibrium (HWE) p<1e-06
- Low minor allele frequency (MAF) <1%



Hardy-Weinberg equilibrium (HWE) test

- Test whether observed genotype counts are deviated from expectations (Hardy-Weinberg equilibrium)
  - Deviations indicate genotyping error, (non-random mating, genetic drift, natural selection, etc.)

Phenotype	White-spotted (AA)	Intermediate (Aa)	Little spotting (aa)	Total
Number	1469	138	5	1612

$$p = \frac{2 \times \text{obs}(AA) + \text{obs}(Aa)}{2 \times (\text{obs}(AA) + \text{obs}(Aa) + \text{obs}(aa))} = \frac{1469 \times 2 + 138}{2 \times (1469 + 138 + 5)} = 0.954$$

$$\begin{aligned} q &= 1 - p \\ &= 1 - 0.954 \\ &= 0.046 \end{aligned}$$

$$\begin{aligned} \text{Exp}(AA) &= p^2 n = 0.954^2 \times 1612 = 1467.4 \\ \text{Exp}(Aa) &= 2pqn = 2 \times 0.954 \times 0.046 \times 1612 = 141.2 \\ \text{Exp}(aa) &= q^2 n = 0.046^2 \times 1612 = 3.4 \end{aligned}$$

$$\begin{aligned} \chi^2 &= \sum \frac{(O-E)^2}{E} \\ &= \frac{(1469 - 1467.4)^2}{1467.4} + \frac{(138 - 141.2)^2}{141.2} + \frac{(5 - 3.4)^2}{3.4} \end{aligned}$$

Ref: HW principal from Wikipedia

QC using PLINK

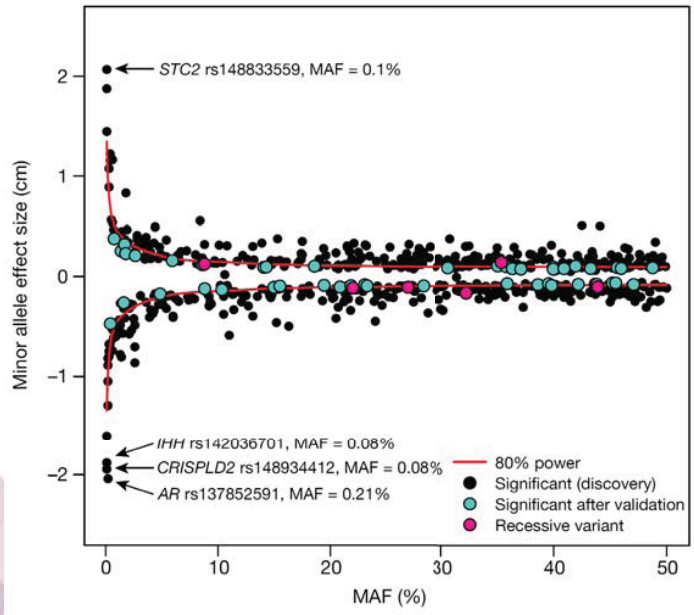
- remove SNPs with MAF <0.01 : **--maf 0.01**
- remove SNPs with missingness rate ≥0.02 (call rate <0.98) : **--geno 0.02**
- remove SNPs with HWE test P-value <1e-06 : **--hwe 1e-06**
- remove samples with missingness rate ≥0.05 (call rate <0.95) : **--mind 0.05**

```
plink --bfile gwas --maf 0.01 --mind 0.05 --geno 0.02 --hwe 1e-06 --make-bed --out QC/gwas.1
```

Feature	As summary	As inclusion criteria
Missingness per individual	--missing	--mind N
Missingness per marker	--missing	--geno N
Allele frequency	--freq	--maf N
Hardy-Weinberg equilibrium	--hardy	--hwe N

https://zzz.bwh.harvard.edu/plink/

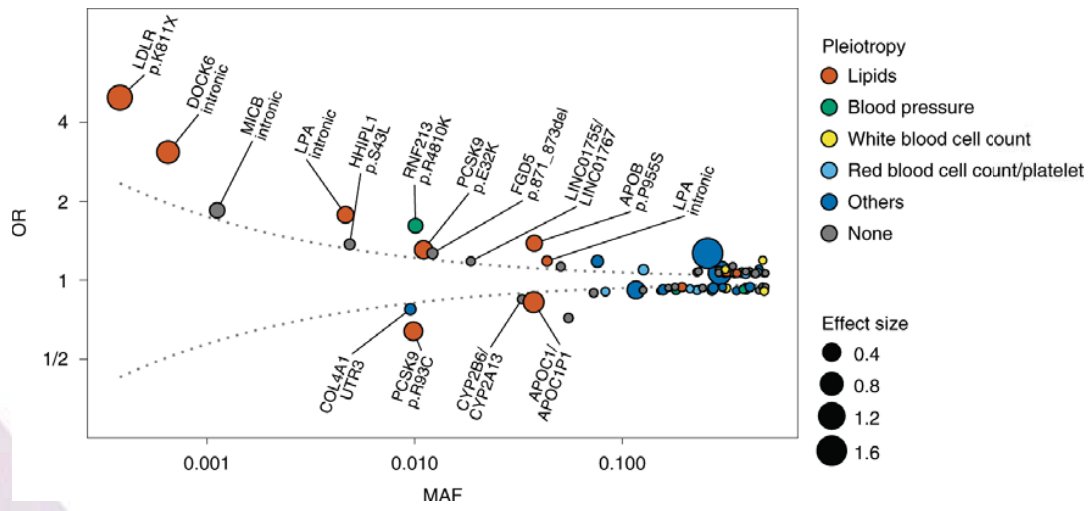
Effects of rare and low-frequency variants on height



- 458,927 individuals
- 697 known loci explained **23.3%** of height heritability
- New loci explained additional **4.1%**
- Rare variants give an increase of **1-2 cm** per allele

Nature 2017 Feb

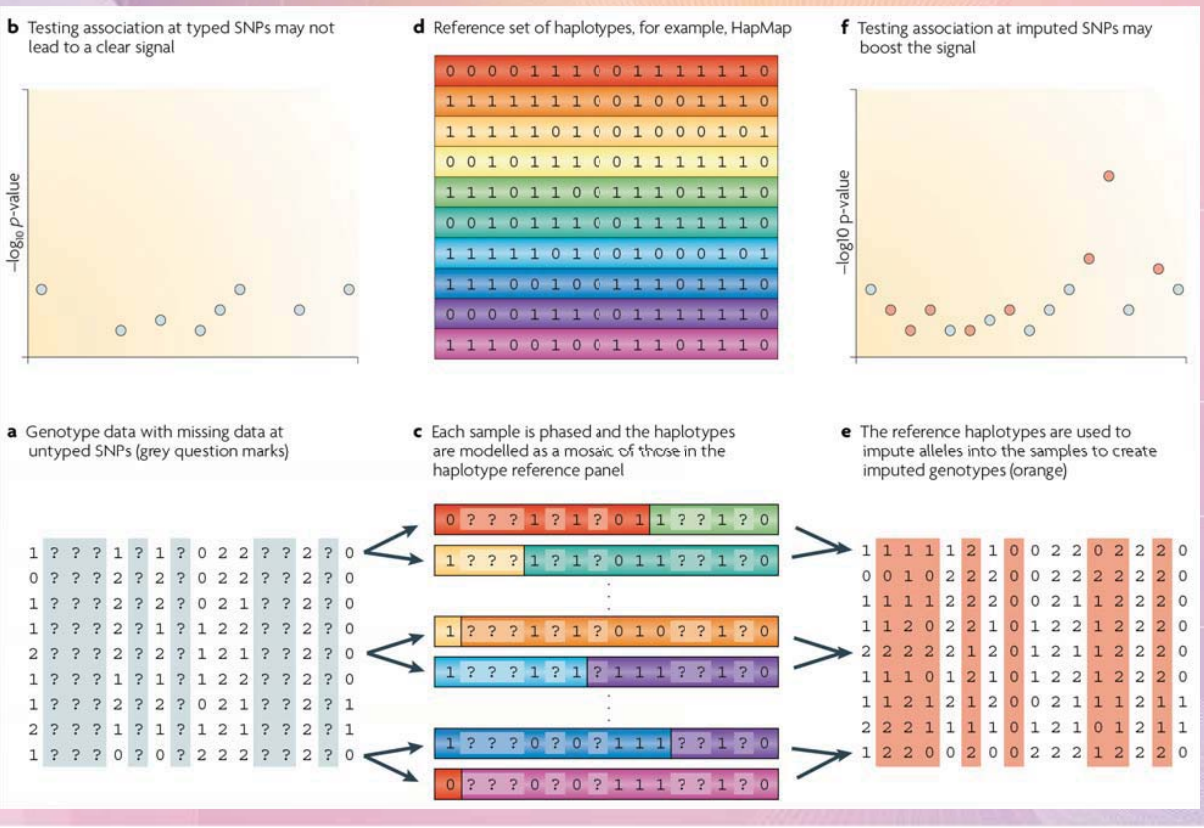
Effects of rare and low-frequency variants on coronary artery disease



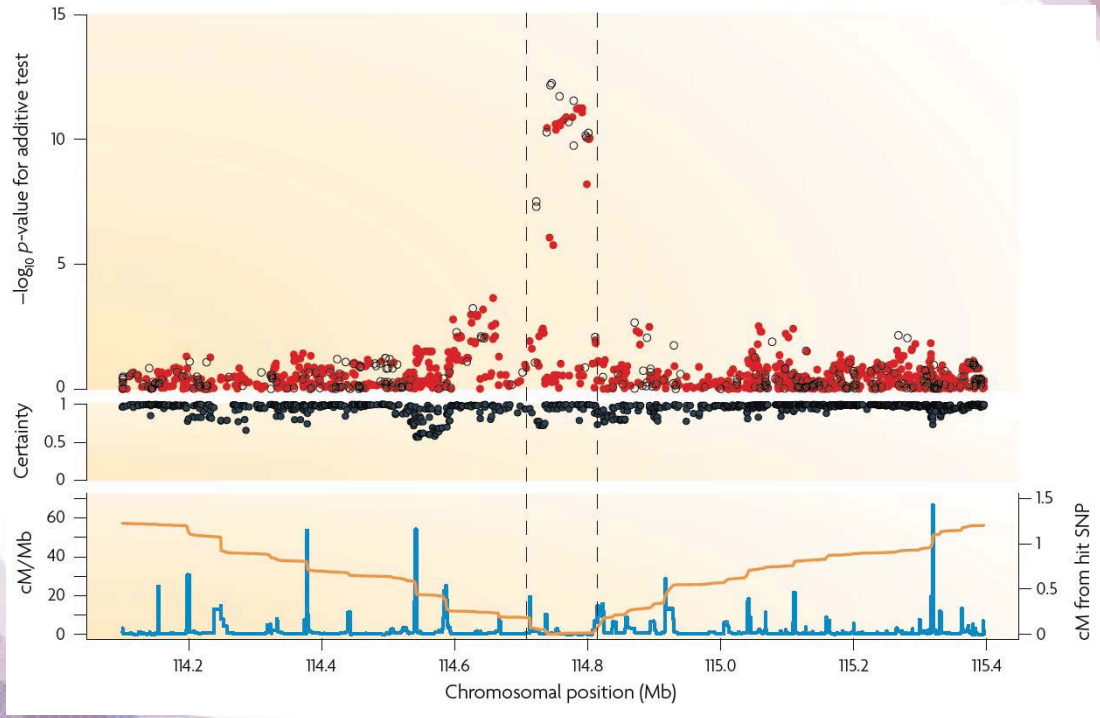
Nature Genetics 52, 1169–1177 (2020)

1 of 2 approaches for low-frequency variants:

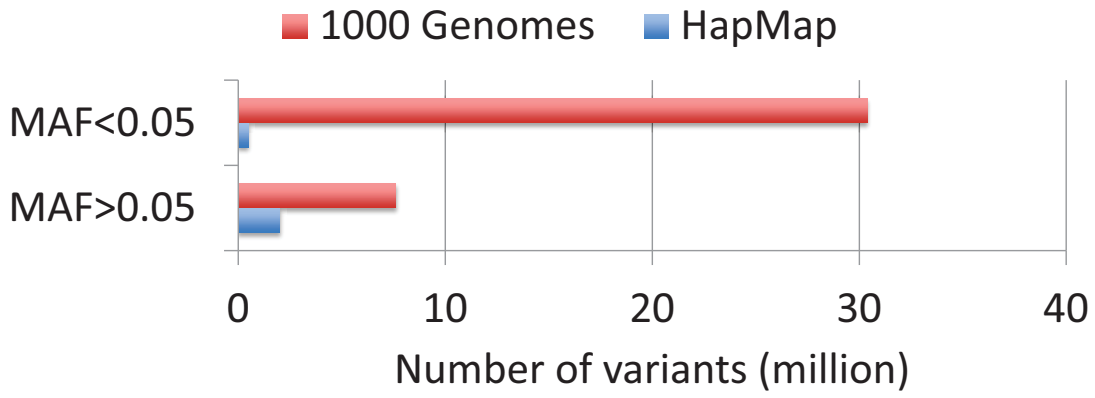
Statistical imputation : more variants

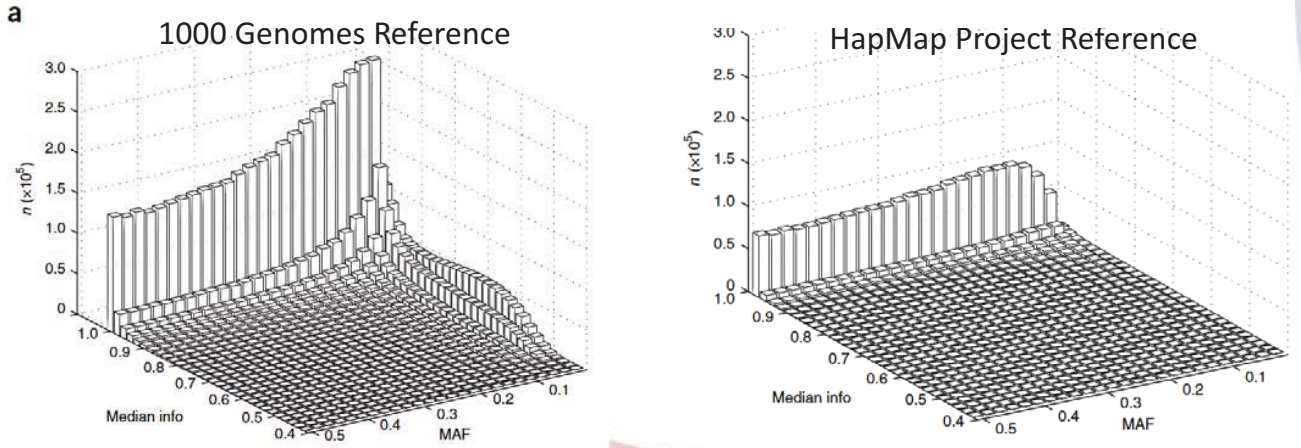


Imputed SNPs : good candidates for replication



1000 Genomes based imputation





Imputation reference

The Haplotype Reference Consortium

OVERVIEW | PARTICIPATING COHORTS | USING THE RESOURCE | CONTACT | SITE LIST

Overview

Goal The Haplotype Reference Consortium (HRC) will create a large reference panel of human haplotypes by combining together sequencing data from multiple cohorts.

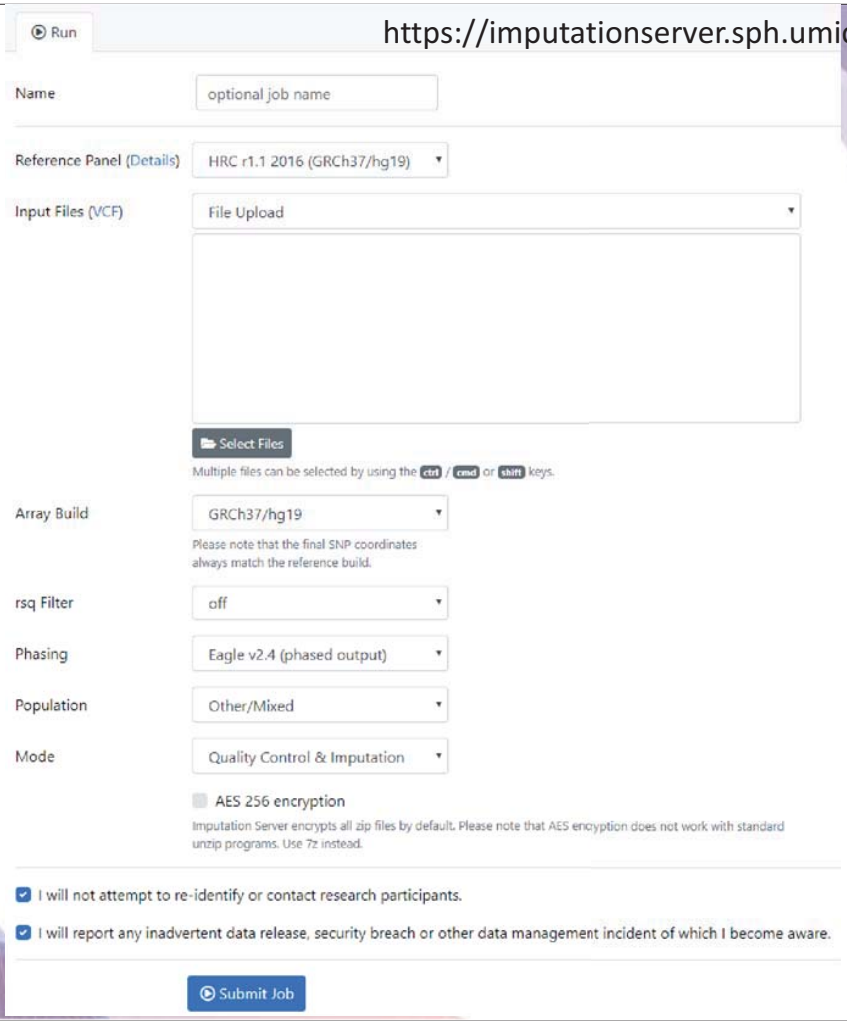
Uses The reference panel will be used for genotype imputation and phasing in other cohorts, typically genome-wide association studies (GWAS), where genotypes are available from genome-wide SNP microarrays.

Benefits By combining together multiple cohorts, the reference panels produced by the project will be as large as possible in terms of both number of haplotypes, and numbers of variants. This will increase the accuracy of the genotype imputation, especially at low-frequency variants, and the number of imputable variants, thus increasing the power of GWAS.

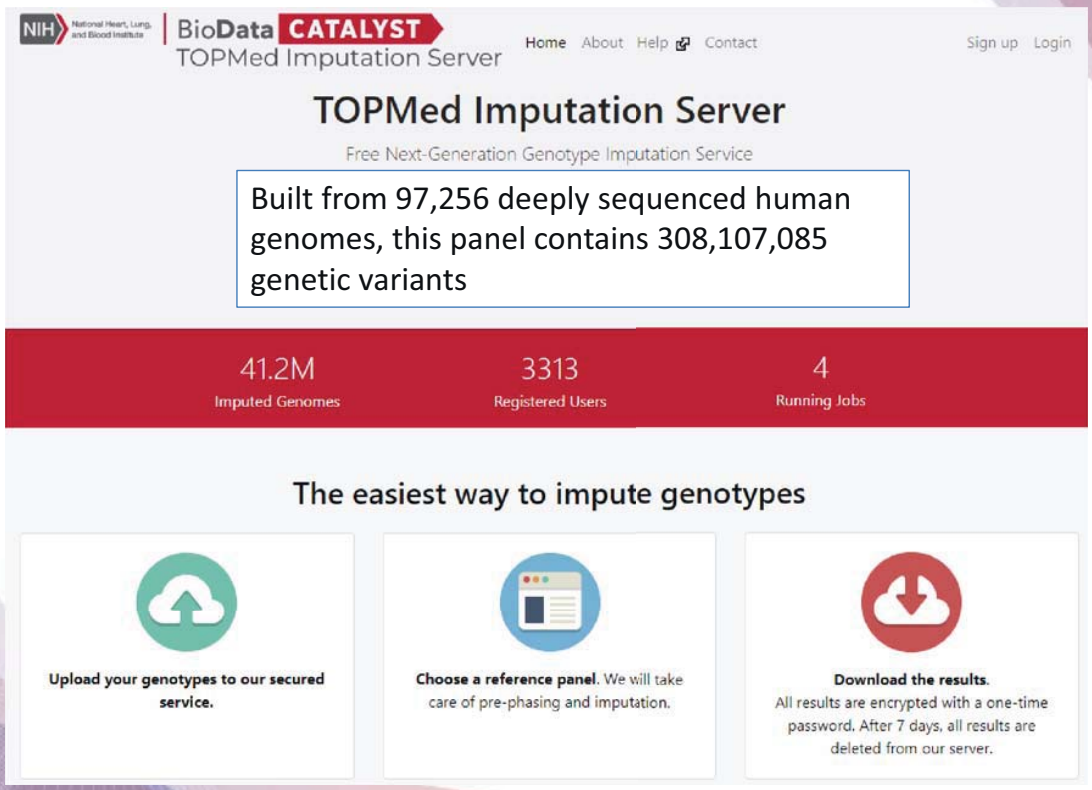
Ancestry Initially, the reference panel will contain haplotypes from individuals with predominantly European ancestry, although the HRC will include the 1000 Genomes Project data. In the future, we envisage the reference panel increasing in size and consisting of samples from a more diverse set of world-wide populations.

Timelines The first release will consist of 64,976 haplotypes at 39,235,157 SNPs, all with an estimated minor allele count of >= 5. The first release will become accessible early summer 2015.

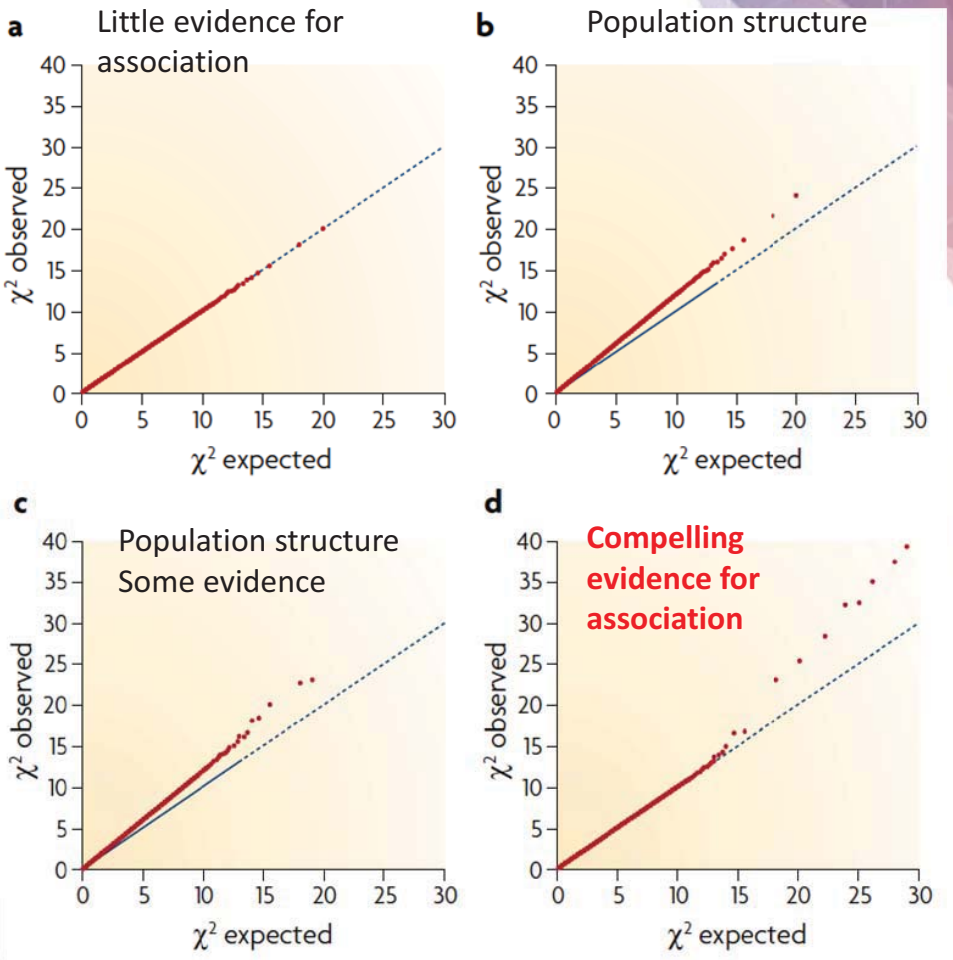
https://imputationserver.sph.umich.edu/index.html



TOPMed Imputation server

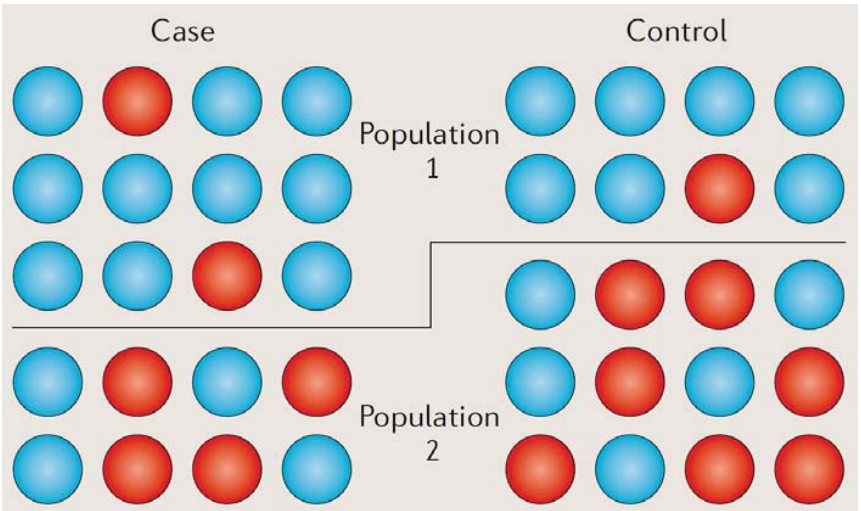


Quantile -quantile plot



Ref: Nat Rev Genet

Spurious associations due to population structure



2/12 = 17% (4% higher freq.)

1/8 = 13%

4/8 = 50% (8% lower freq.)

7/12 = 58%

6/20 = 30% (10% lower freq.)

8/20 = 40%

Ref: Nat Rev Genet

Genetics of chopstic use

successful-use-of-selected-hand instruments gene' (SUSHI)

Sample 3: Americans + Chinese

$\chi^2 = 34.2 \quad P = 4.9 \times 10^{-9}$

	Use of chopsticks		
Allele	**Yes**	**No**	**Total**
A1	640	340	980
A2	400	100	500
Total	1040	440	1480

Sample 1: Americans

$\chi^2 = 0 \quad P = 1$

	Use of chopsticks		
Allele	**Yes**	**No**	**Total**
A1	320	320	640
A2	80	80	160
Total	400	400	800

Sample 2: Chinese

$\chi^2 = 0 \quad P = 1$

	Use of chopsticks		
Allele	**Yes**	**No**	**Total**
A1	320	20	340
A2	320	20	340
Total	640	40	680

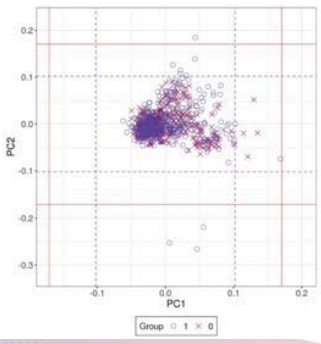
Ref: Taru Tukiainen

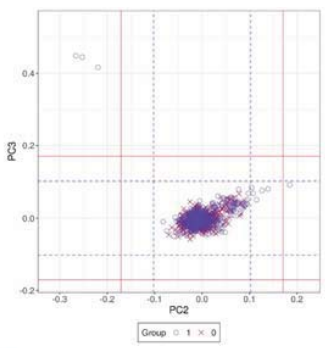
Principal component analysis

- Objective
  - Detect sub-population and any individuals of different ancestry
- Tools
  - smartpca tool of EIGENSOFT software (or using PLINK)

https://www.hsph.harvard.edu/alkes-price/software/

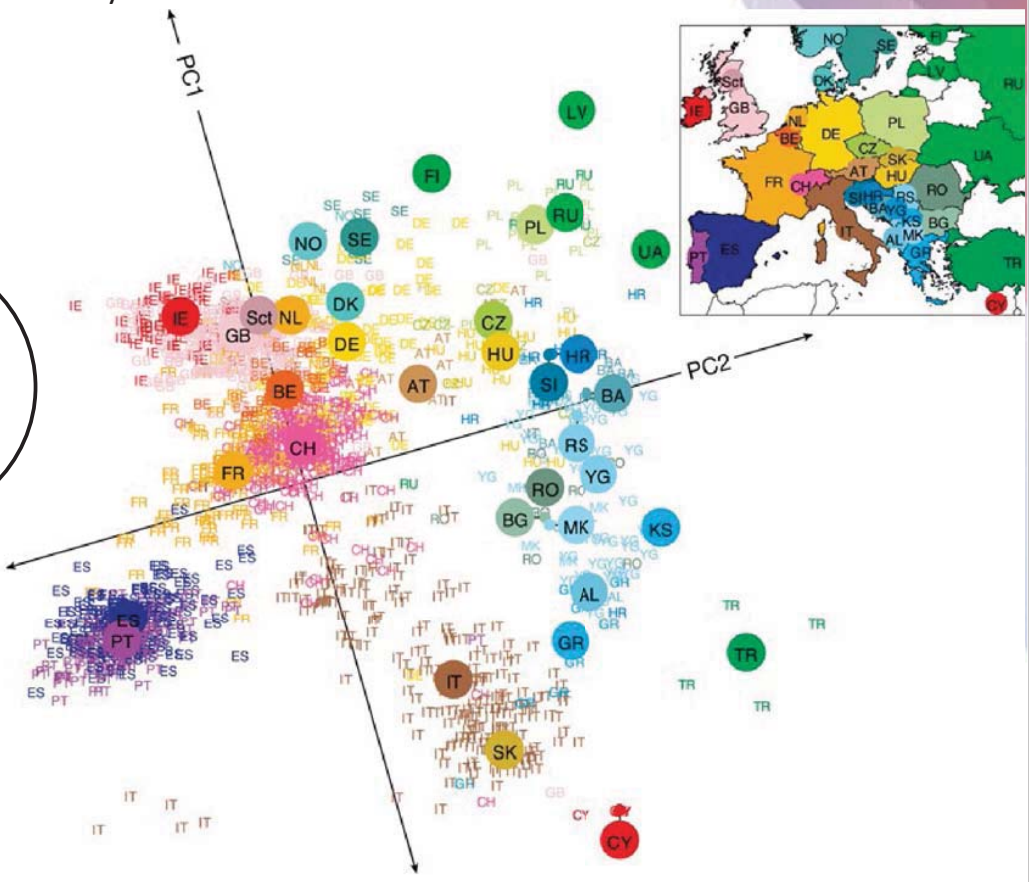
- Solution
  - Check if cases and controls are well overlaid. If not, systematic or technical differences between cases and controls might exist
  - Remove outliers (e.g. >|5σ|) or include 10 or 20 principal components as covariates in GWAS analyses





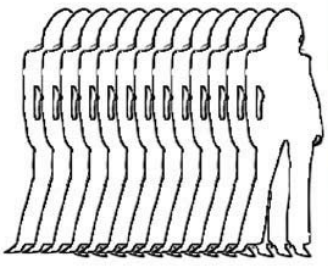
Principle component analysis

PCA plot



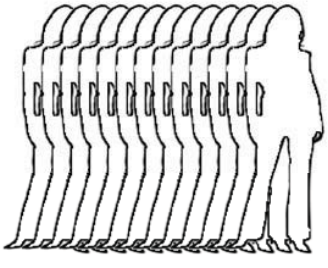
Genome-wide association study

Patients



GC CC GG GC CC GC GC
GG CC GC GG GC GG

Controls



GC CC GC GC GG CC CC
CC GC GC GG GC GG

SNP1	*SNP2*	*SNP ...*
Cases Count of G: 2104 of 4000	**Cases** Count of G: 1648 of 4000	*Repeat for all SNPs*
Frequency of G: 52.6%	Frequency of G: 41.2%	
Controls Count of G: 2676 of 6000	**Controls** Count of G: 2532 of 6000	
Frequency of G: 44.6%	Frequency of G: 42.2%	
P-value: $5.0 \cdot 10^{-15}$	**P-value:** 0.33	

Basic association test

- H_0: Frequency of 'A1' is *independent* of case/control status.

	A1	A2	
Cases	w	x	
Controls	y	z	

Odds Ratio (OR): Odds of Allele occurring in cases to the odds of Allele occurring in controls:

$$\frac{w/x}{y/z} = \frac{wz}{xy}$$

$$\chi^2 = (O-E)^2/E$$

[Pearson's chi-Square]

In PLINK, OR > 1 implies A1 is at higher frequency in cases relative to controls.

Note that this is not uniform across all analytical platforms.

Ref: Chris Cotsapas

Regression anlaysis

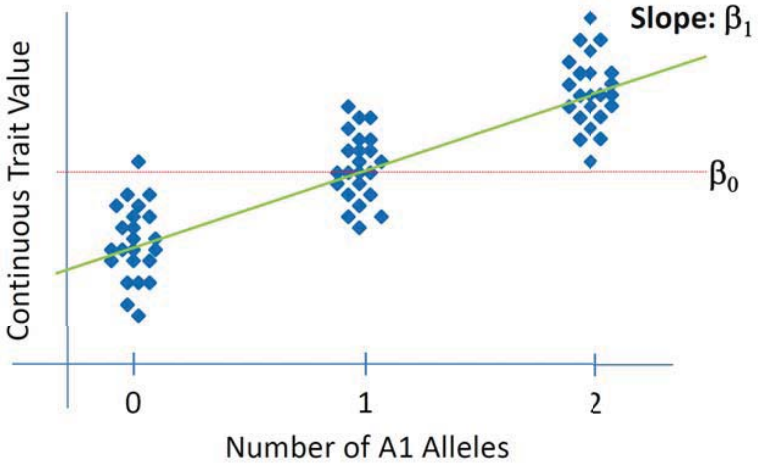
- Analysis of the relationship between a dependent or outcome variable (phenotype) with one or more independent or predictor variables (SNP genotype)

Linear Regression Equation

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Logistic Regression Equation

$$\ln\left(\frac{p_i}{(1-p_i)}\right) = \beta_0 + \beta_1 X_i + \varepsilon_i$$



Ref: Chris Cotsapas

PLINK software (A to Z)

https://zzz.bwh.harvard.edu/plink/

- Google PLINK
- Quality control
- Data management
  - .ped / .map
- Summary stats
- Population stratification
- Association tests
  - Regression, Dominant/Recessive/Trend, Fisher's Exact
- Etc.

```
FAM001  1  0 0  1  2  A A  G G  A C
FAM001  2  0 0  1  2  A A  A G  0 0
...
```

```
Family ID
Individual ID
Paternal ID
Maternal ID
Sex (1=male; 2=female; other=unknown)
Phenotype
```

```
1  rs123456  0  1234555
1  rs234567  0  1237793
1  rs224534  0  -1237697
1  rs233556  0  1337456
...
```

Mixed effects model

- Mixed effects model
  - Y = SNP + sex + age + PCs + *Kinship* + e
- Fixed effects
  - SNP, sex, age, PCs
- Random effects
  - Kinship matrix (due to relatedness)

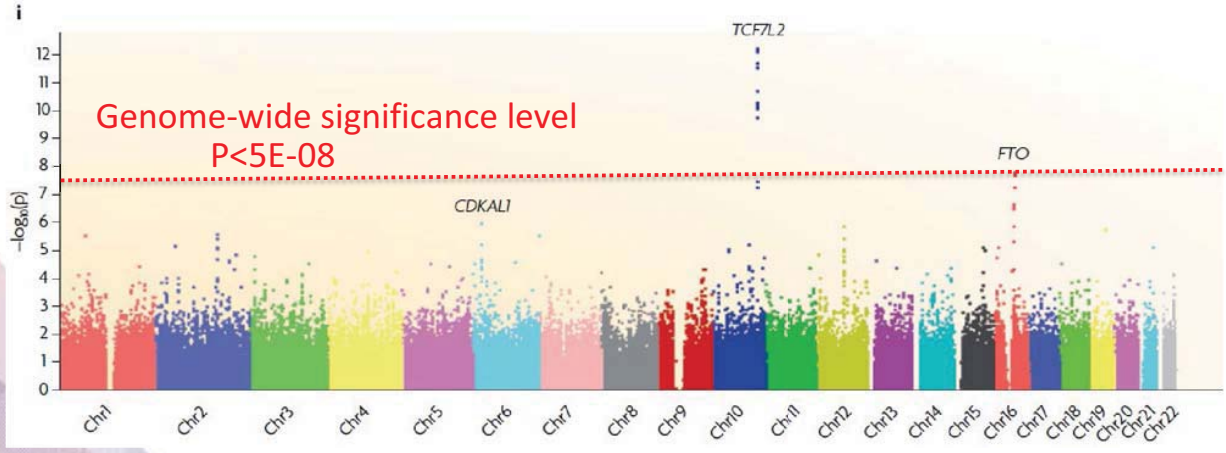
- Tools
  - Binary (disease): SAIGE, REGENIE
  - Continuous (BMI, blood pressure et al.): BOLT-LMM, REGENIE

Genome-wide significance level

- Multiple-testing (comparisons) problem
  - the problem that arises when many null hypotheses are tested; some significant results are likely even if all the hypotheses are false
- Bonferroni's correction (more stringent method)
  - 0.05 / # of tested variants (usually assuming 1M)
  - 0.05 / 1,000,000 = 5E-08
- False discovery rate (less stringent method)

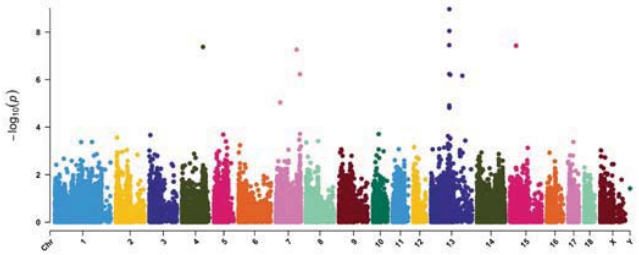
Manhattan plot

http://pipoli.com

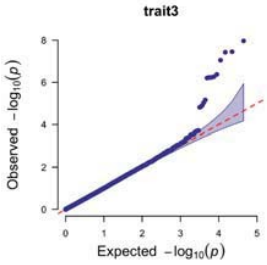


A high-quality drawing tool designed for Manhattan plot of genomic analysis

Circular-Manhattan plot



Rectangular-Manhattan plot



Q-Q plot

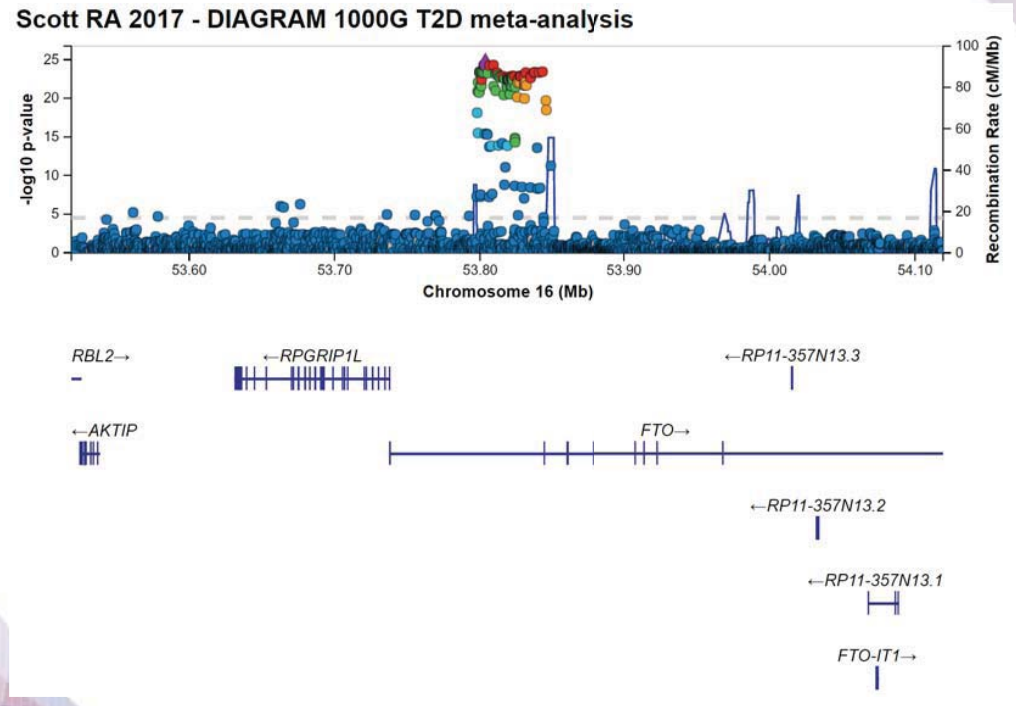
```
> install.packages("CMplot")
> library("CMplot")

## Rectangular Manhattan plot
> cmplot(pig60K, type="p", plot.type="m", LOG10=TRUE,
threshold=NULL, file="jpg", memo="", dpi=300, file.output=TRUE,
verbose=TRUE, width=14, height=6, 考.labels.angle=45)

## QQ plot
> cmplot(pig60K, plot.type="q", box=FALSE, file="jpg",memo="",
dpi=300, conf.int=TRUE, conf.int.col=NULL, threshold.col="red",
threshold.lty=2, file.output=TRUE, verbose=TRUE, width=5,
height=5)
```

https://github.com/YinLiLin/CMplot

Regional plot of association



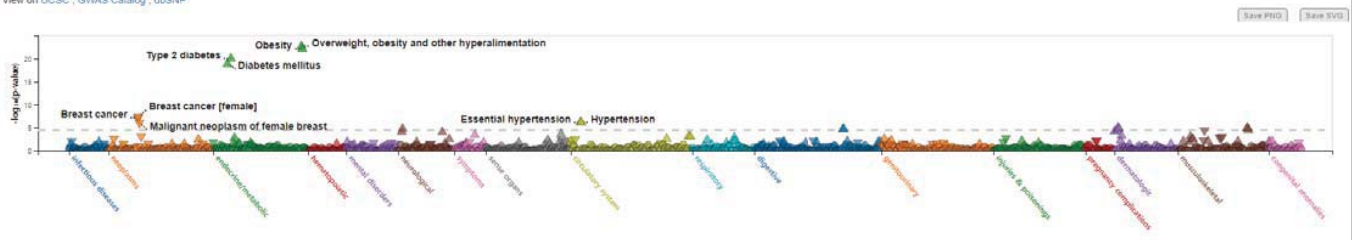
LocusZoom, a visualization tool of GWAS results (http://locuszoom.org/)

Phenome-wide analysis (PheWAS)

https://pheweb.org/UKB-SAIGE/

16 : 53,821,125 A / G (rs17817712)

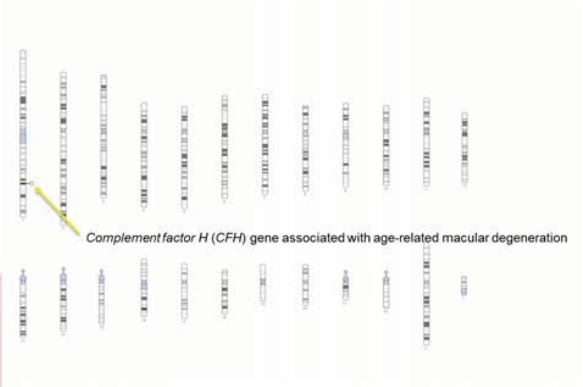
Nearest gene: *FTO*
AF: 0.39
View on UCSC , GWAS Catalog , dbSNP



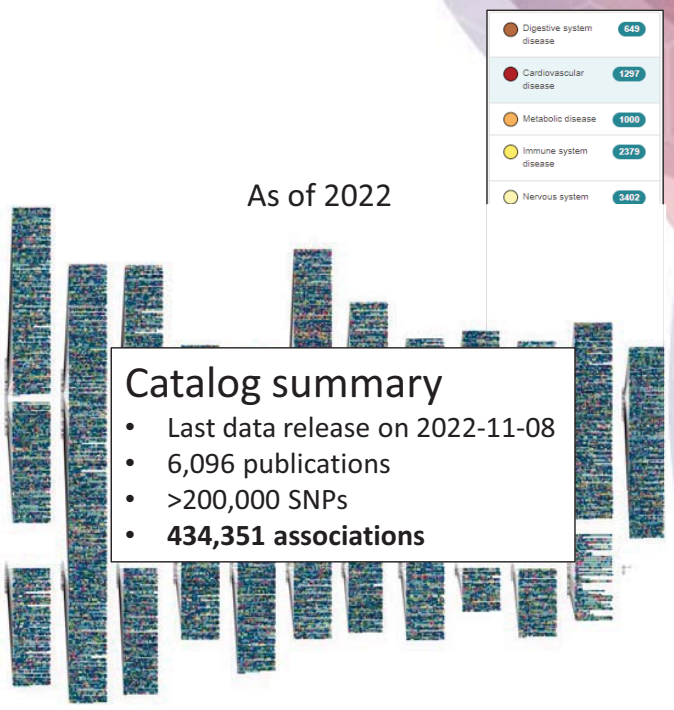
Category	Phenotype	P-value	Effect Size (se)	Number of samples
endocrine/metabolic	Overweight, obesity and other hyperalimentation	1.9e-23	0.14 (0.014)	10968 / 397993
endocrine/metabolic	Obesity	7.3e-23	0.14 (0.015)	10799 / 397993
endocrine/metabolic	Type 2 diabetes	9.6e-21	0.11 (0.011)	18945 / 388756
endocrine/metabolic	Diabetes mellitus	1.4e-19	0.10 (0.011)	20203 / 388756
neoplasms	Breast cancer [female]	4.3e-8	-0.074 (0.014)	12671 / 388549
neoplasms	Breast cancer	7.1e-8	-0.073 (0.013)	12898 / 388549
circulatory system	Hypertension	7.1e-7	0.033 (0.0066)	77977 / 330366
circulatory system	Essential hypertension	7.3e-7	0.033 (0.0066)	77723 / 330366

GWAS loci at p<5E-8

As of 2005



As of 2022

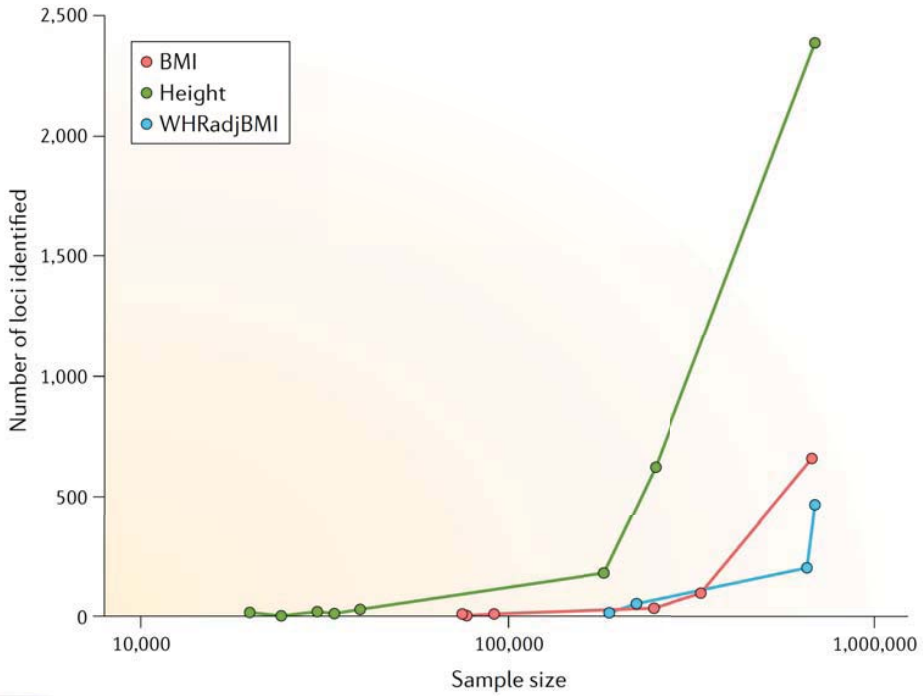


Catalog summary

- Last data release on 2022-11-08
- 6,096 publications
- >200,000 SNPs
- **434,351 associations**

https://www.ebi.ac.uk/gwas

Number of loci identified increases as a function of GWAS sample size



Summary of GWAS analysis and tools

- Quality control
  - Sample QC: PLINK
  - Variant QC: PLINK
  - Related samples: KING
  - PCA of genetic ancestry: EIGENSTRAT(smartpca)
- Imputation
  - Haplotype Reference Consortium: Michigan Imputation Server
  - TOPMed Imputation Server
- Association analysis
  - Logistic/linear regression (unrelated): PLINK
  - Mixed effects regression (including related): SAIGE, BOLT-LMM, REGENIE
- Visualization
  - QQ plot: CM-PLOT
  - Manhattan plot: CM-PLOT

Summary

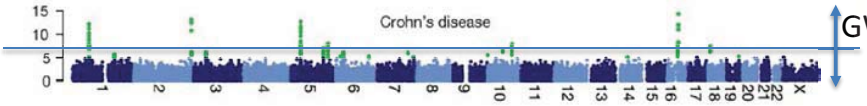
- SNP arrays and statistical imputation provide fast and accurate genotyping of about a million of genetic variants
- Sample-level and variant-level quality control is very important to remove technical errors and false positive findings
- GWAS have identified >200,000 variants associated with various human traits/diseases

Post-GWAS analysis

Summary of post-GWAS analysis and tools

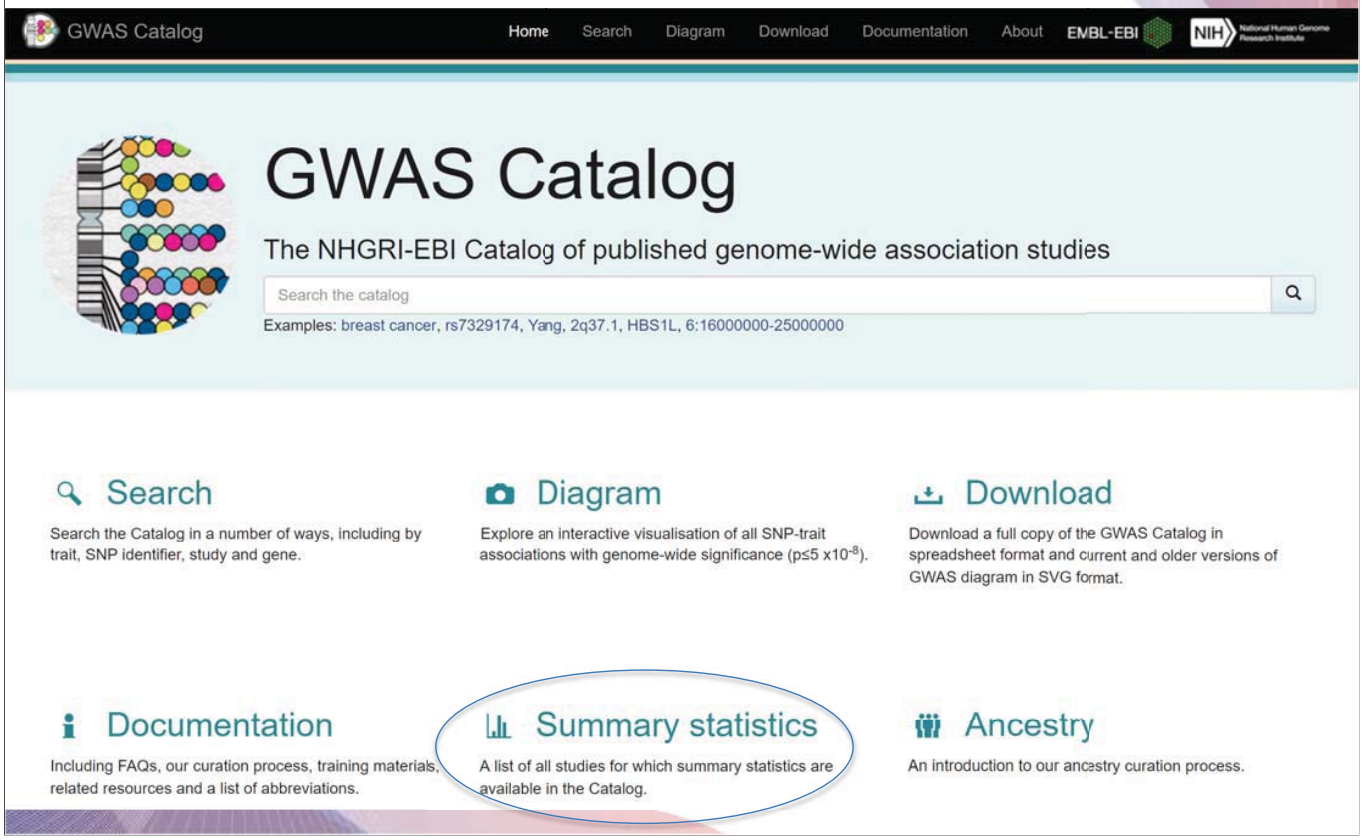
- Understanding genetic architecture
  - SNP-based heritability: LDSC, *GCTA (if genotype available)*
  - Genetic correlation: LDSC (same ancestry), POPCORN or S-LDXR (transancestry)
  - SNP heritability in specific tissues or cells: LDSC-SEG
- Finding causal variants, genes, and pathways
  - Fine-mapping (causal variants): CAVIAR, FINEMAP, PAINTOR, SUSIE
  - eQTL and colocalization analysis (genes): COLOC2
  - Pathway enrichment analysis (pathways or gene sets): MAGMA
- *Identifying individuals at high genetic risk (genotype required)*
  - Polygenic risk score: PRSICE-2, LDPRED, PRS-CS
- Inferring causality between traits
  - Mendelian randomization: MR-BASE, TwoSampleMR (R package)

GWAS summary statistics are publicly available



- Detailed GWAS results of all variants
  - SNP(rsID), effect allele, OR or beta, SE, P value, etc.
- GWAS Catalog
  - https://www.ebi.ac.uk/gwas
- GWAS Atlas
  - https://atlas.ctglab.nl
- UK Biobank
  - https://github.com/weizhouUMICH/SAIGE
- Consortium websites
  - CARDIoGRAMpluC4D http://www.cardiogramplusc4d.org/data-downloads
  - Diabetes DIAGRAM Consortium http://diagram-consortium.org/downloads.html

https://www.ebi.ac.uk/gwas/



Summary statistics contain most GWAS results

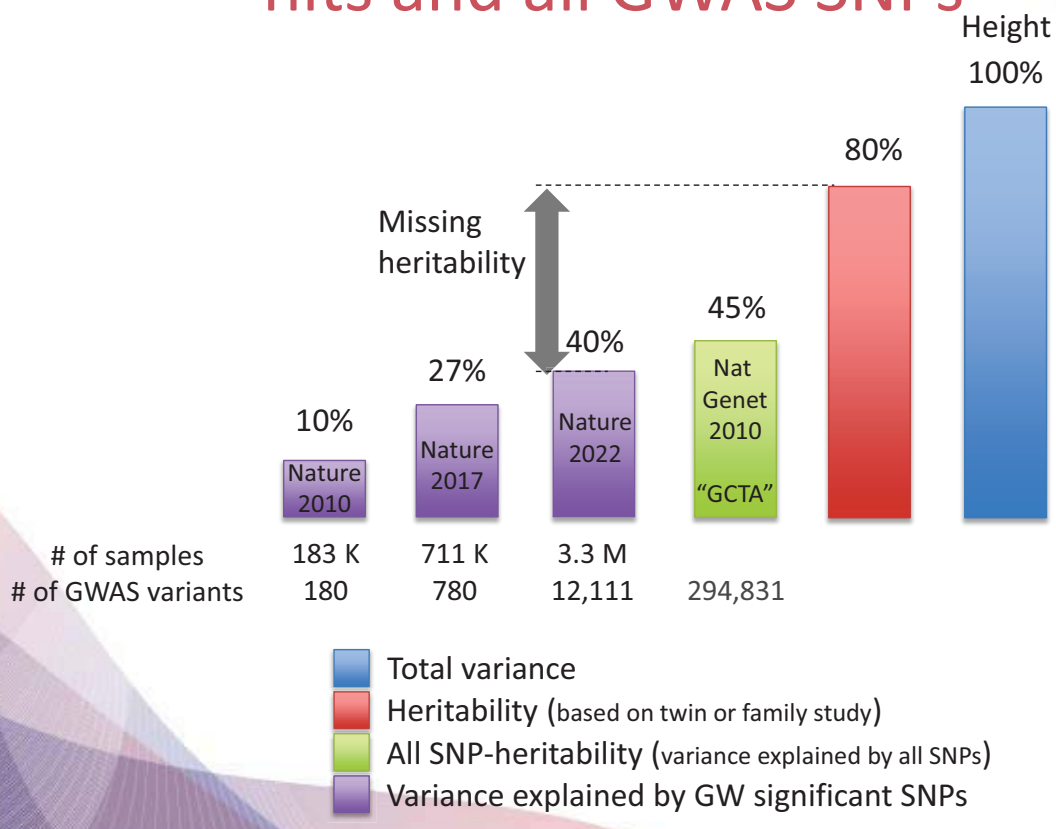
chrom	pos	snpid	ref	alt	ac	af	num_cases	num_controls	beta	sebeta	Tstat	pval
1	16071	rs541172944	G	A	39.843	5.00E-05	650	399970	-2.62	7.55	-0.046	7.28E-01
1	16280	rs866639523	T	C	124	0.000155	650	399970	-2.99	4.06	-0.182	4.61E-01
1	49298	rs10399793	T	C	499790.227	0.623771	650	399970	-0.0468	0.0984	-23.4	6.34E-01
1	54353	rs140052487	C	A	285.302	0.000356	650	399970	-1.22	3.19	-0.12	7.03E-01
1	54564	rs558796213	G	T	121.776	0.000152	650	399970	-0.224	2.89	-0.0269	9.38E-01
1	54591	rs561234294	A	G	79.153	9.90E-05	650	399970	-2.91	6.75	-0.064	6.66E-01
1	54676	rs2462492	C	T	321190.055	0.400866	650	399970	0.039	0.0975	16.9	6.89E-01
1	55326	rs3107975	T	C	6698.62	0.00836	650	399970	-1	0.552	-3.29	6.89E-02

Full information for all the variants (~ several millions)

What can we do with summary statistics?

- Estimate the heritability of traits
- Estimate the genetic correlations among traits
- Test associations between genes and traits
- Infer causality between two traits using MR
- Use for weights of SNPs for disease prediction using polygenic risk score (PRS)
- And more..

Heritability is explained in part by GWAS hits and all GWAS SNPs

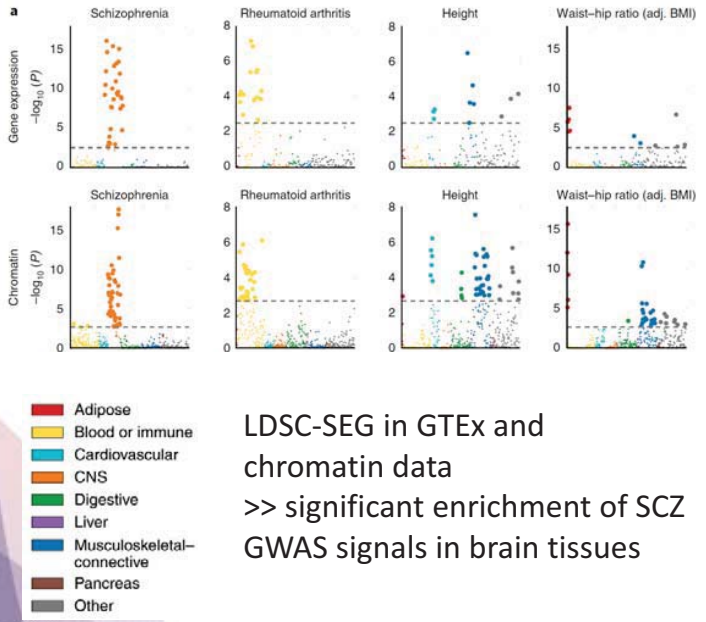


Heritability of GWAS hits and all GWAS SNPs

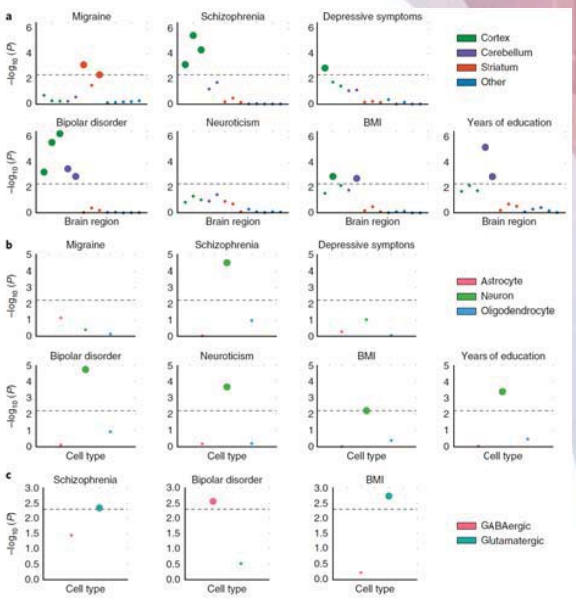
Trait or Disease	h^2 Pedigree Studies	h^2 GWAS Hits[a]	h^2 All GWAS SNPs[b]
Type 1 diabetes	0.9[98]	0.6[99,c]	0.3[12]
Type 2 diabetes	0.3–0.6[100]	0.05-0.10[34]	
Obesity (BMI)	0.4–0.6[101,102]	0.01-0.02[36]	0.2[14]
Crohn's disease	0.6–0.8[103]	0.1[11]	0.4[12]
Ulcerative colitis	0.5[103]	0.05[12]	
Multiple sclerosis	0.3–0.8[104]	0.1[45]	
Ankylosing spondylitis	>0.90[105]	0.2[106]	
Rheumatoid arthritis	0.6[107]		
Schizophrenia	0.7–0.8[108]	0.01[79]	0.3[109]
Bipolar disorder	0.6–0.7[108]	0.02[79]	0.4[12]
Breast cancer	0.3[110]	0.08[111]	
Von Willebrand factor	0.66–0.75[112,113]	0.13[114]	0.25[14]
Height	0.8[115,116]	0.1[13]	0.5[13,14]
Bone mineral density	0.6-0.8[117]	0.05[118]	
QT interval	0.37–0.60[119,120]	0.07[121]	0.2[14]
HDL cholesterol	0.5[122]	0.1[57]	
Platelet count	0.8[123]	0.05–0.1[58]	

The American Journal of Human Genetics *90*, 7–24, January 13, 2012

Heritability enrichment of specifically expressed genes in tissues and cell types



LDSC-SEG in GTEx and chromatin data
>> significant enrichment of SCZ GWAS signals in brain tissues

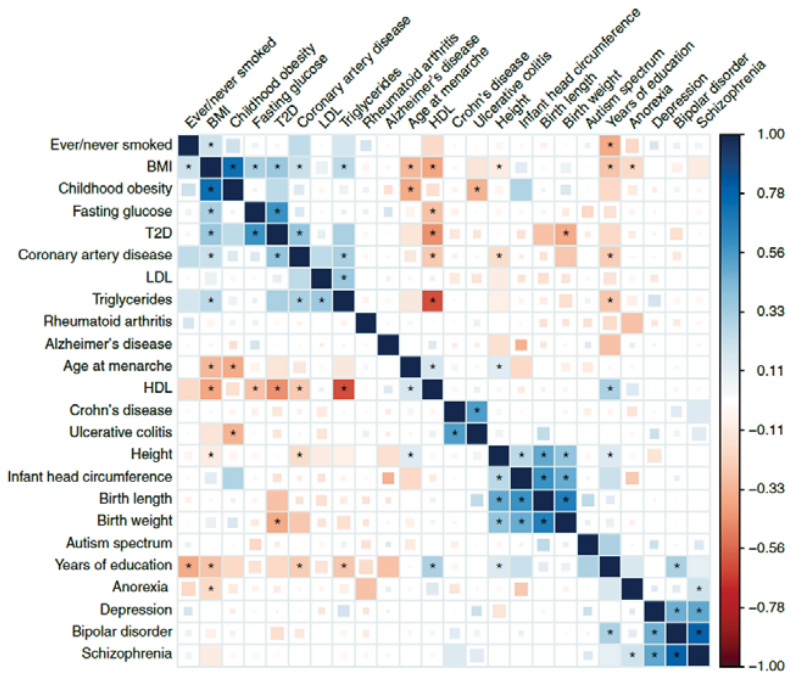


Significant enrichment in glutamatergic neurons in cortex

https://github.com/bulik/ldsc/wiki/Cell-type-specific-analyses

Nature Genetics 50, 621–629 (2018)

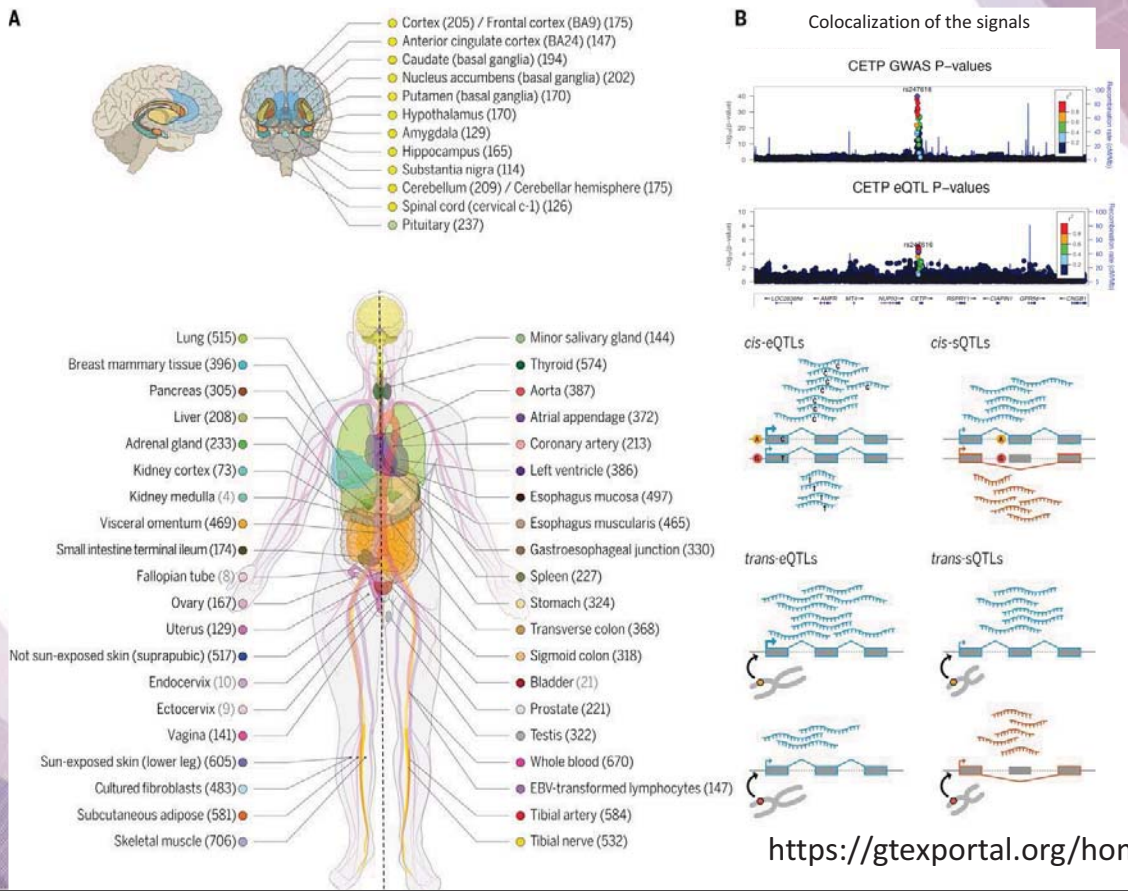
Genetic correlation among diseases



https://github.com/bulik/ldsc/wiki/Heritability-and-Genetic-Correlation

Nat Genet. 2015 Nov; 47(11): 1236–1241.

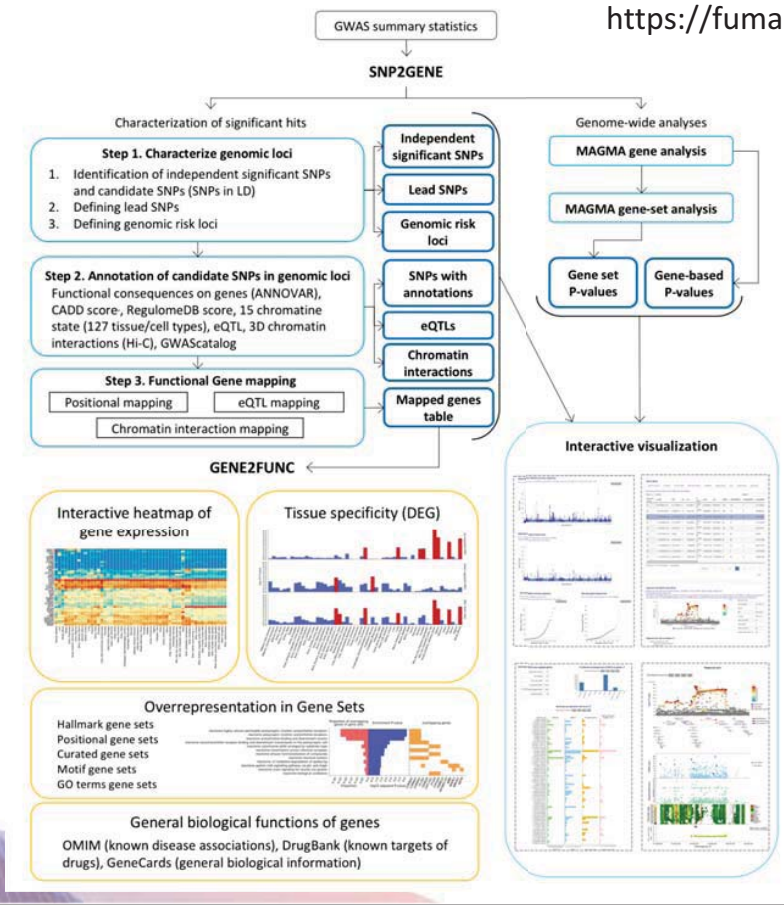
GTEx (genotype-tissue expression)



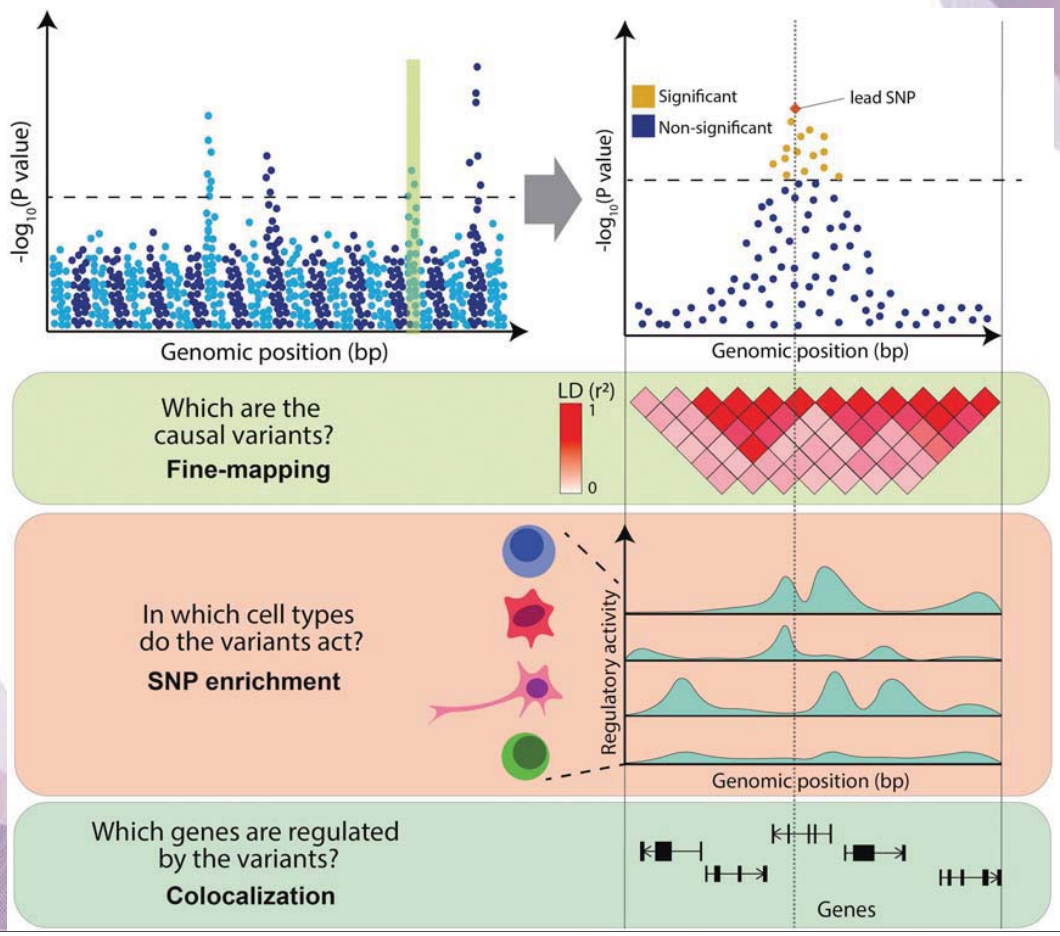
https://gtexportal.org/home/

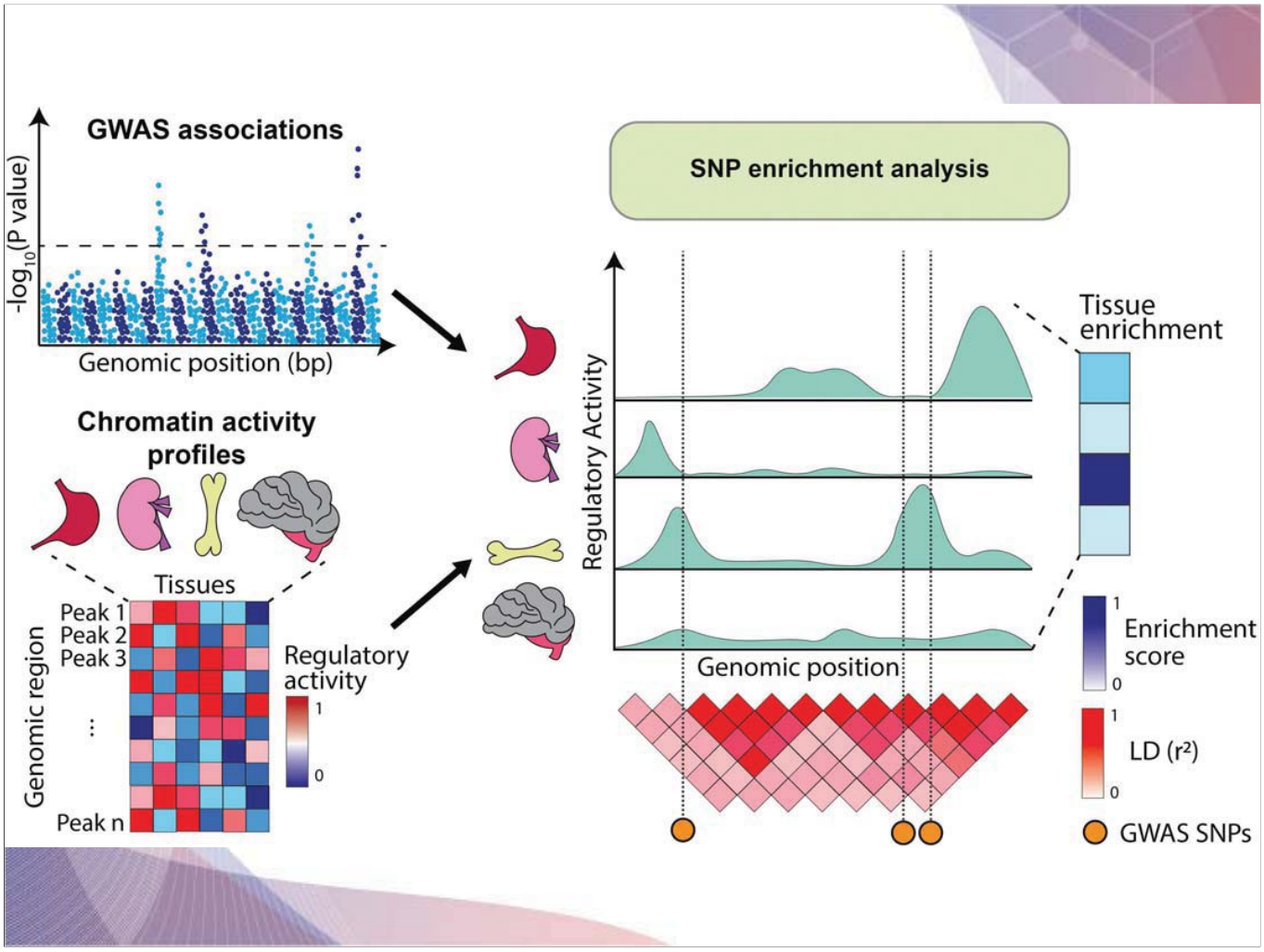
FUMA for various post-GWAS analyses

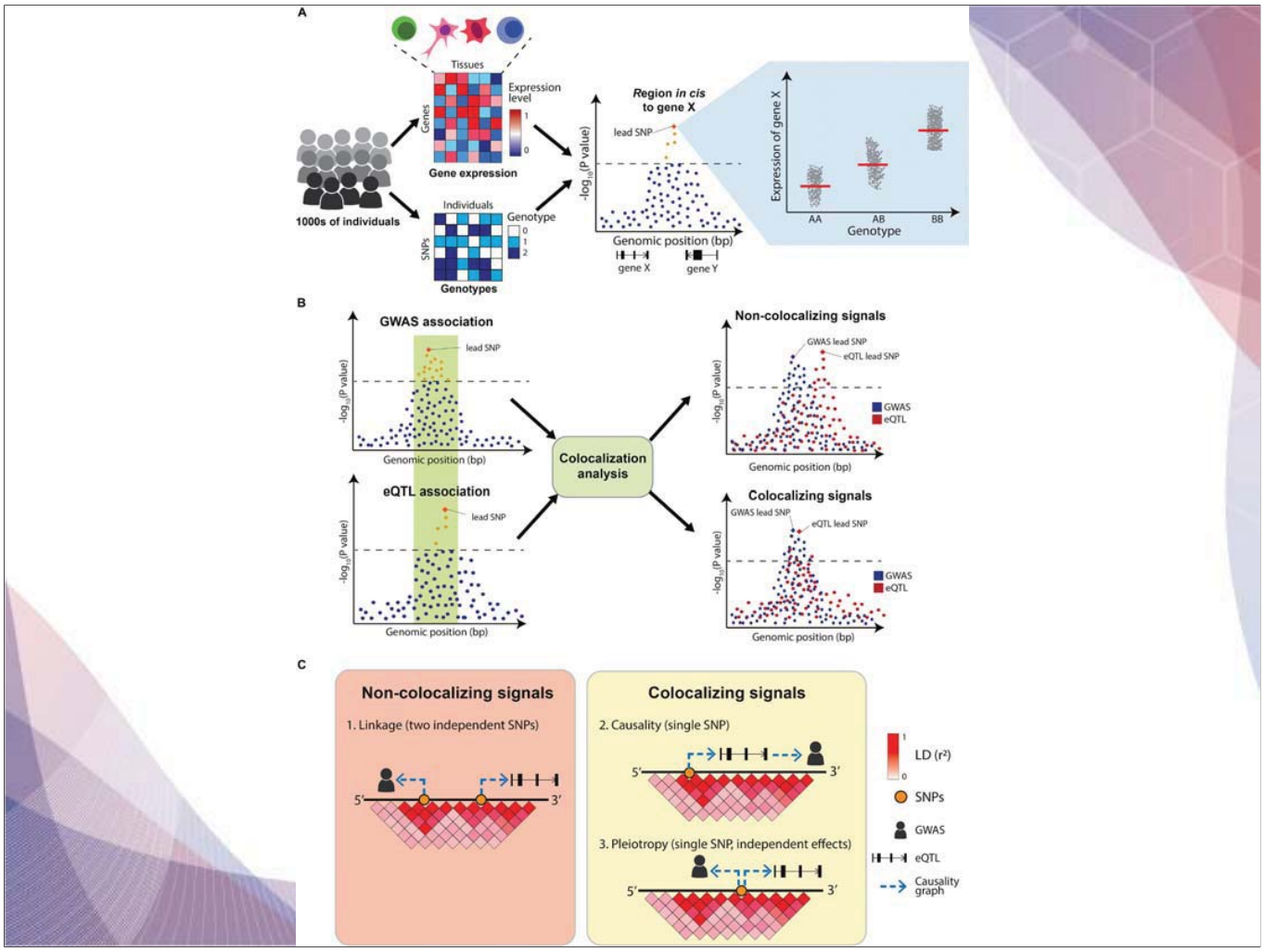
https://fuma.ctglab.nl/tutorial



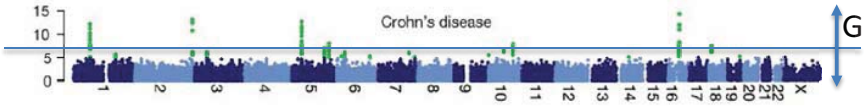
Review paper: https://www.frontiersin.org/articles/10.3389/fgene.2020.00424/full



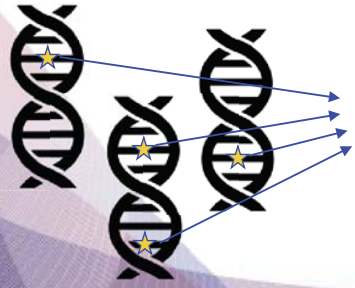

GWAS associations
-log10(P value)
Genomic position (bp)
Chromatin activity profiles
Tissues
Peak 1
Peak 2
Peak 3
Peak n
Genomic region
Regulatory activity
1
0
SNP enrichment analysis
Regulatory Activity
Genomic position
Tissue enrichment
1
Enrichment score
0
1
LD (r²)
0
GWAS SNPs


A
Tissues
Expression level
Genes
Gene expression
1000s of individuals
Individuals
Genotype
SNPs
Genotypes
Region in cis to gene X
lead SNP
-log10(P value)
Genomic position (bp)
gene X
gene Y
Expression of gene X
AA
AB
BB
Genotype
B
GWAS association
lead SNP
-log10(P value)
Genomic position (bp)
eQTL association
lead SNP
Colocalization analysis
Non-colocalizing signals
GWAS lead SNP
eQTL lead SNP
GWAS
eQTL
Colocalizing signals
C
Non-colocalizing signals
1. Linkage (two independent SNPs)
5′
3′
Colocalizing signals
2. Causality (single SNP)
3. Pleiotropy (single SNP, independent effects)
LD (r²)
SNPs
GWAS
eQTL
Causality graph

Polygenic risk score: high risk if one has many risk variants



- PRS based on GWAS
  - Sum of risk allele counts across GWAS variants
  - Weighted (effect size) sum of risk allele counts across GWAS variants



Complex disease/phenotypes

"Polygenicity"

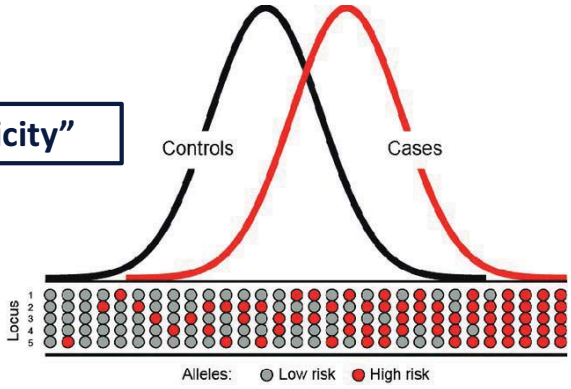


Image Ref: 2014.Genes_ Whiffin and Houlston

PRS (Pruning and Thresholding)

$$PRS(P_T) = \sum_{i=1}^{M} \mathbb{1}_{\{P_i < P_T\}} \tilde{\beta}_i g_i$$

M: # of SNPs remaining after LD-clumping

$\tilde{\beta}_i$: normalized marginal effect size estimates

P_T: P-value threshold

Summary statistics from GWAS (independent SNP list)

	Effect_Allele	beta	se	P-value
SNP1	A	1.7	0.26	3.11E-11
SNP2	G	1.3	0.25	9.96E-08
SNP3	A	-0.7	0.15	1.53E-06
SNP4	C	-1.2	0.27	4.41E-06
SNP5	G	1.8	0.82	1.41E-02
SNP6	T	0.4	0.41	1.65E-01
...				
...				

P-value < 1e-05

Genotype data

	SNP1	SNP2	SNP3	SNP4	...
indiv1	AT	GG	AC	AA	...
indiv2	AT	TT	AC	CC	...
indiv3	AA	GG	AC	AA	...
indiv4	TT	GT	CC	AC	...
indiv5	AT	TT	AC	AC	...

→

	SNP1	SNP2	SNP3	SNP4	...
indiv1	1	2	1	0	...
indiv2	1	0	1	2	...
indiv3	2	2	1	0	...
indiv4	0	1	0	1	...
indiv5	1	0	1	1	...

70

PRS (P+T)

$$PRS(P_T) = \sum_{i=1}^{M} \mathbb{1}_{\{P_i < P_T\}} \tilde{\beta}_i g_i$$

- M: # of SNPs remaining after LD-clumping
- P_T: P-value threshold
- $\tilde{\beta}_i$: normalized marginal effect size estimates

Summary statistics from GWAS (independent SNP list)

	Effect_Allele	beta	se	P-value
SNP1	A	1.7	0.26	3.11E-11
SNP2	G	1.3	0.25	9.96E-08
SNP3	A	-0.7	0.15	1.53E-06
SNP4	C	-1.2	0.27	4.41E-06
SNP5	G	1.8	0.82	1.41E-02
SNP6	T	0.4	0.41	1.65E-01
...				
...				

P-value < 1e-05

PRS for indiv1

$= g_1*\beta_1 + g_2*\beta_2 + g_3*\beta_3 + g_4*\beta_4$

$= 1*1.7 + 2*1.3 + 1*(-0.7) + 0*(-1.2)$

$= 3.6$

Genotype data

	SNP1	SNP2	SNP3	SNP4	...
indiv1	1	2	1	0	...
indiv2	1	0	1	2	...
indiv3	2	2	1	0	...
indiv4	0	1	0	1	...
indiv5	1	0	1	1	...

	PRS
indiv1	3.6
indiv2	-1.4
indiv3	5.3
indiv4	0.1
indiv5	-0.2

71

nature genetics

LETTERS

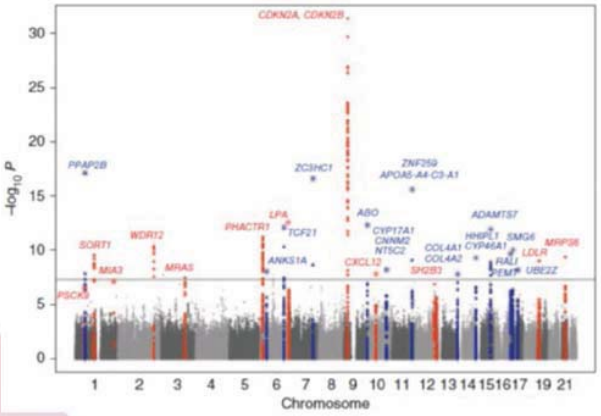
https://doi.org/10.1038/s41588-018-0183-z

Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations

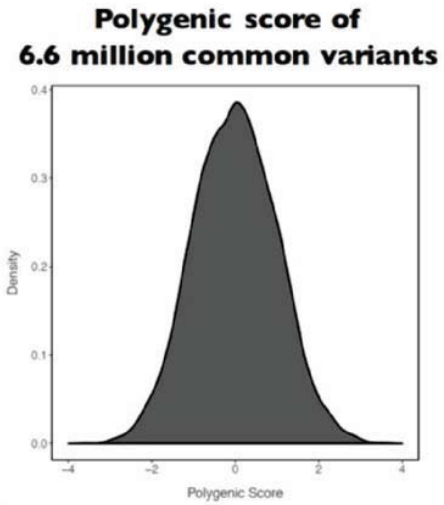
Amit V. Khera[1,2,3,4,5], Mark Chaffin[4,5], Krishna G. Aragam[1,2,3,4], Mary E. Haas[4], Carolina Roselli[4], Seung Hoan Choi[4], Pradeep Natarajan[2,3,4], Eric S. Lander[4], Steven A. Lubitz[2,3,4], Patrick T. Ellinor[2,3,4] and Sekar Kathiresan[1,2,3,4*]

Khera et al. Nat Genet 2018

Polygenic risk score for CAD using 6 million variants is normally distributed

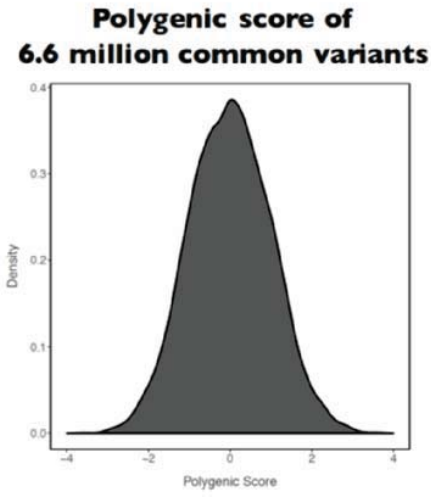


Summary statistics of 6 million variants from a previous large GWAS for coronary artery disease (CAD)



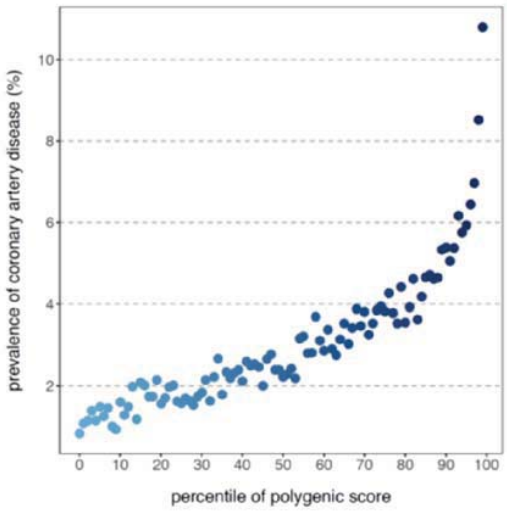
Calculating PRS in independent, testing samples

The empirical risk of CAD rising sharply in the right tail of the distribution



Calculating PRS in independent, testing samples

Percentile



Prevalence of CAD according to the percentile of the GPS$_{CAD}$.

The genetic prediction accuracy was far lower for other populations than for European populations

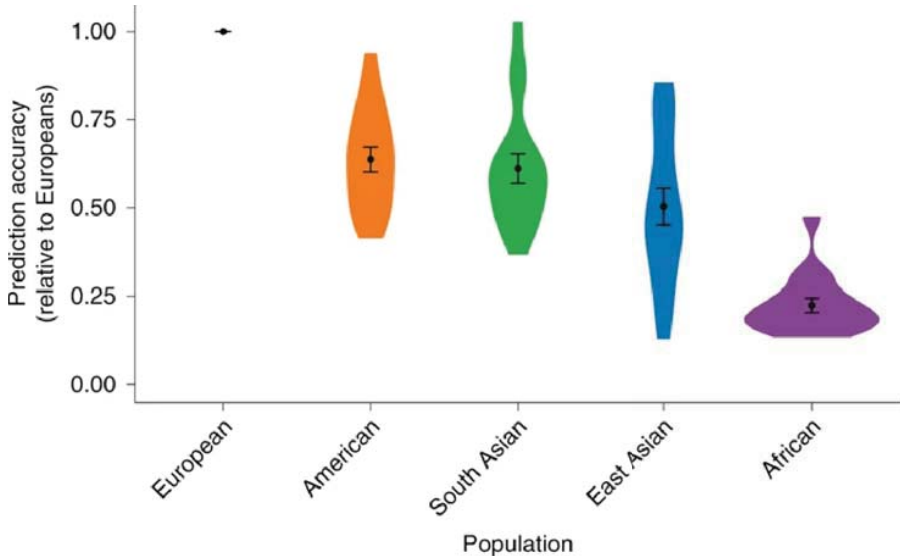
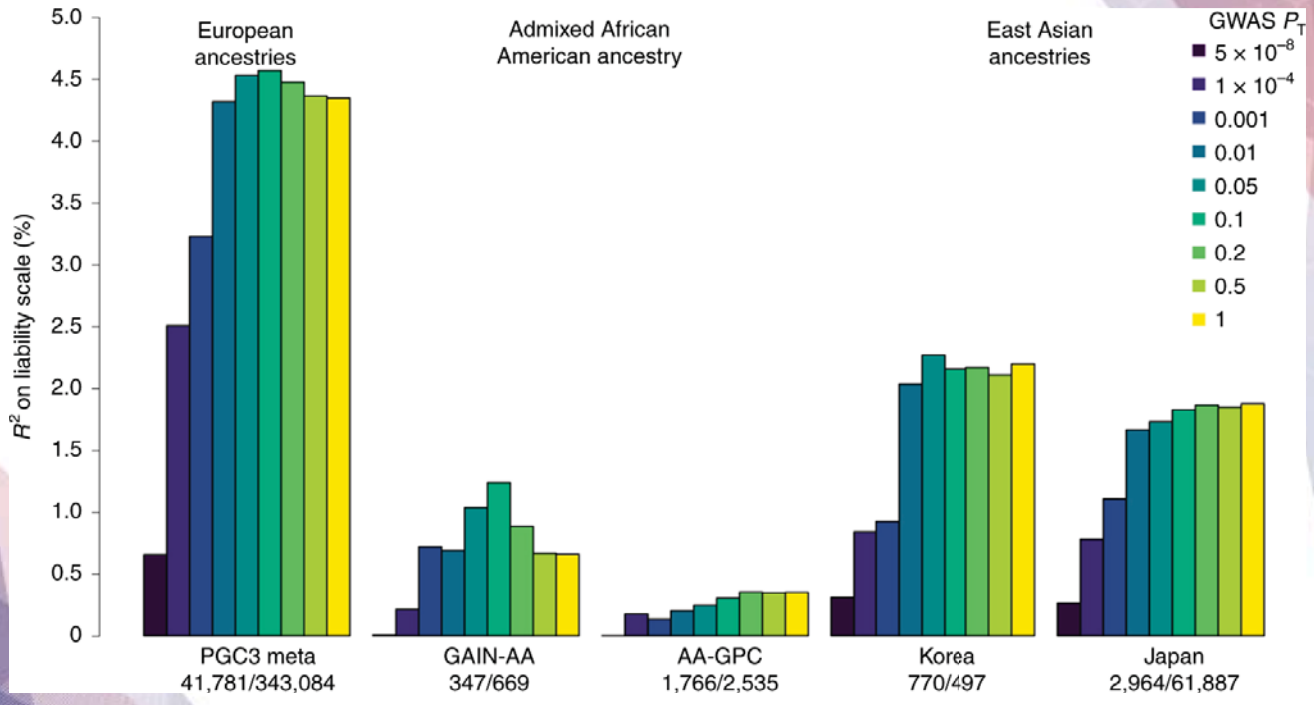


Fig. 3: Prediction accuracy relative to European-ancestry individuals across 17 quantitative traits and 5 continental populations in the UKBB.

Martin et al. Nat Genet (2019)

PRS for bipolar disorder across ancestries

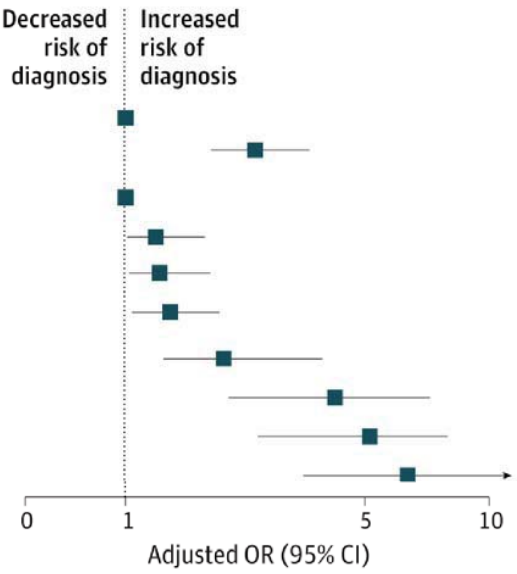


Nature Genetics 53, 817–829 (2021)

PRS for Alzheimer's disease for Koreans

A AD dementia diagnosis

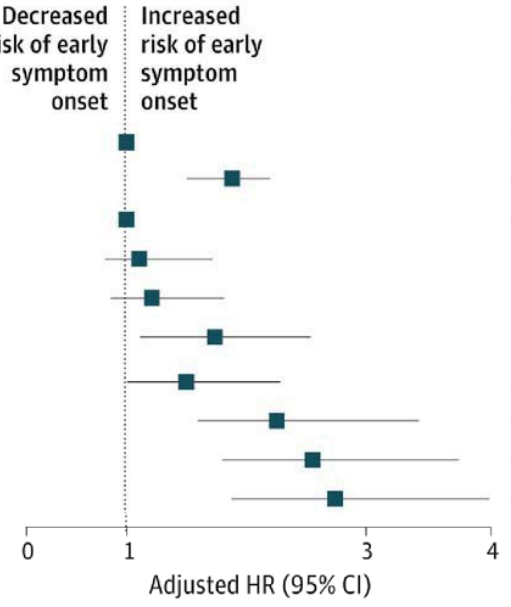
APOE ε4 status	PRS for AD dementia	No. participants with AD dementia/ total No.	Adjusted OR (95% CI)	*P* value	*P* value for trend
Noncarrier	NA	240/579	1 [Reference]	NA	
carrier	NA	314/432	3.16 (2.46-4.09)	<.001	
Noncarrier	Low	54/144	1 [Reference]	NA	
	Intermediate	62/148	1.52 (1.01-2.29)	.04	
	High	64/148	1.55 (1.03-2.34)	.04	<.001
	Very high	60/139	1.70 (1.13-2.58)	.01	
Carrier	Low	59/97	2.63 (1.63-4.29)	<.001	
	Intermediate	78/107	4.49 (2.74-7.54)	<.001	
	High	85/113	5.11 (3.10-8.62)	<.001	
	Very high	92/115	6.73 (3.99-11.75)	<.001	



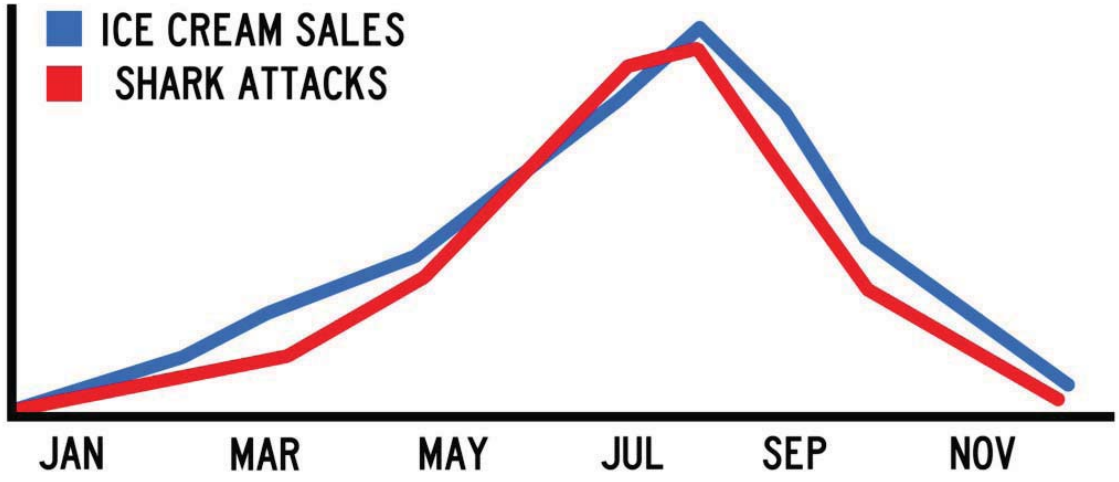
PRS for Alzheimer's disease for Koreans

B AD dementia onset age

APOE ε4 status	PRS for AD dementia	No. participants with AD dementia/ total No.	Adjusted HR (95% CI)	*P* value	*P* value for trend
Noncarrier	NA	209/414	1 [Reference]	NA	
carrier	NA	273/344	1.82 (1.51-2.18)	<.001	
Noncarrier	Low	45/102	1 [Reference]	NA	
	Intermediate	51/104	1.16 (0.78-1.74)	.47	
	High	54/108	1.21 (0.81-1.80)	.35	<.001
	Very high	59/100	1.71 (1.16-2.53)	.007	
Carrier	Low	50/75	1.51 (1.00-2.27)	.05	
	Intermediate	70/87	2.23 (1.59-3.41)	<.001	
	High	76/93	2.59 (1.78-3.76)	<.001	
	Very high	77/89	2.74 (1.88-4.00)	<.001	



CORRELATION IS NOT CAUSATION!

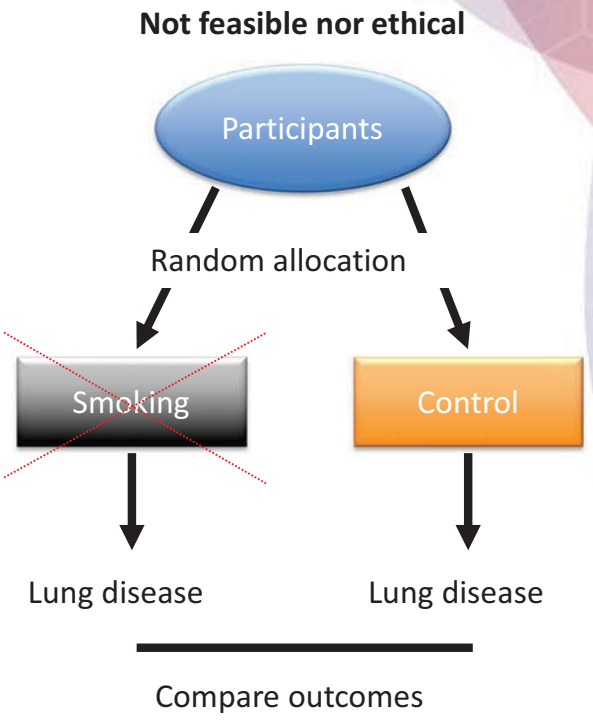


Both ice cream sales and shark attacks increase when the weather is hot and sunny, but they are not caused by each other (they are caused by good weather, with lots of people at the beach, both eating ice cream and having a swim in the sea)

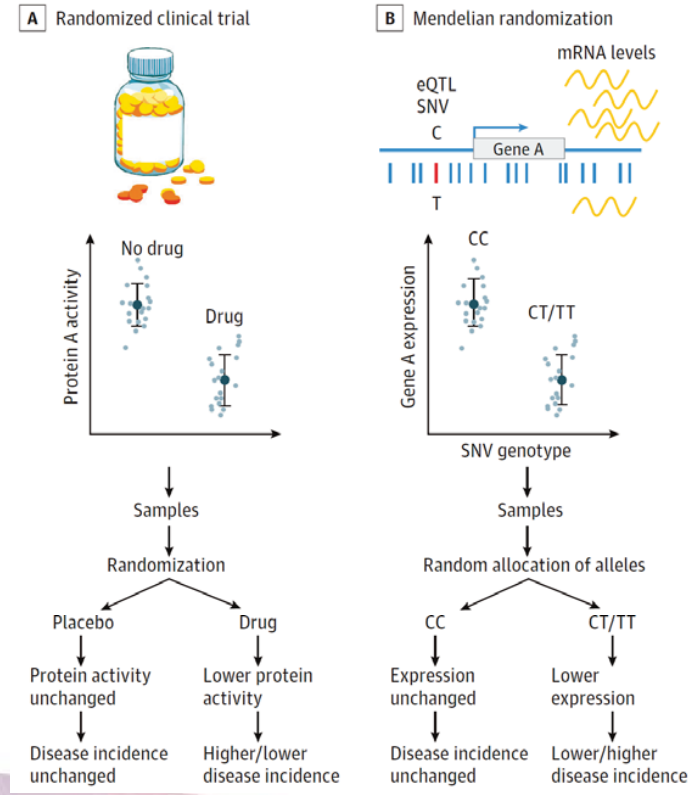
https://idatassist.com/why-journalists-love-causation-and-how-statisticians-can-help/

RCT is the gold-standard design to infer causality, but

- Exceedingly expensive and time-consuming efforts
- High failure rates (>50% fail owing to lack of efficacy)
- Not always feasible or ethical to conduct

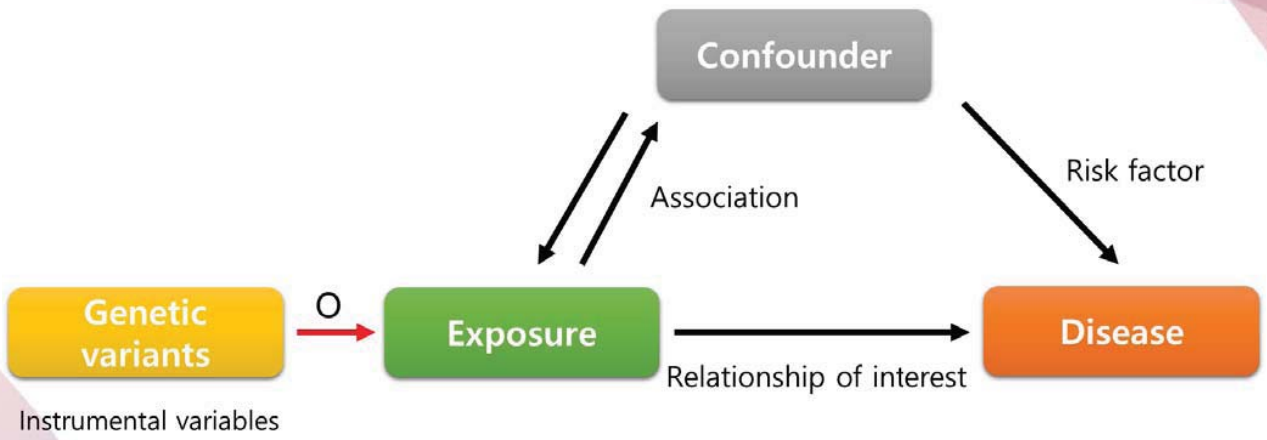


Analogy between RCT and MR



JAMA Psychiatry 2021;78(6):623-631

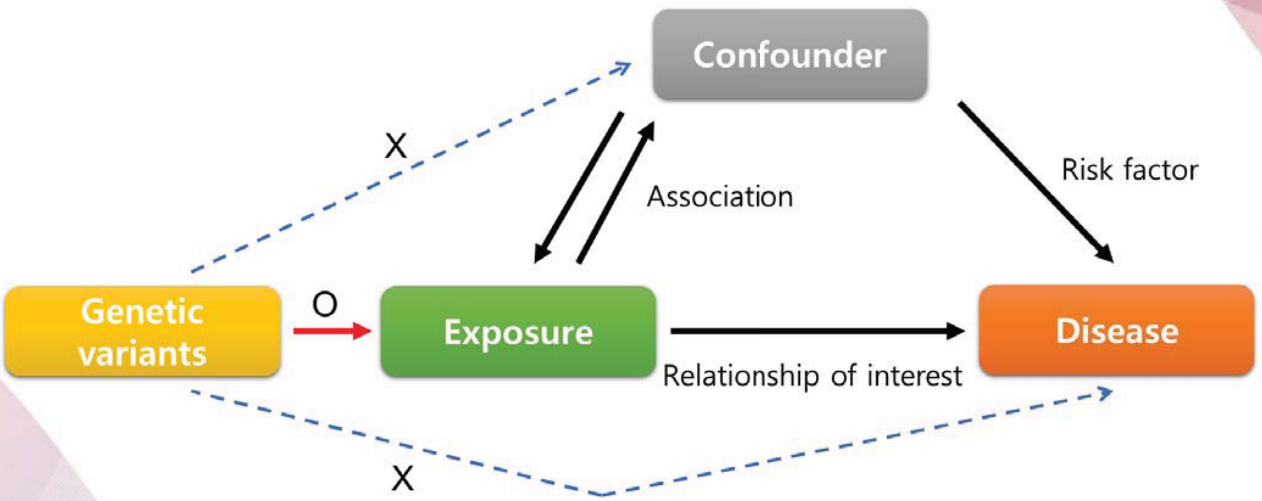
Mendelian randomization (MR)



Assumption I: Genetic variants are robustly associated with the exposure

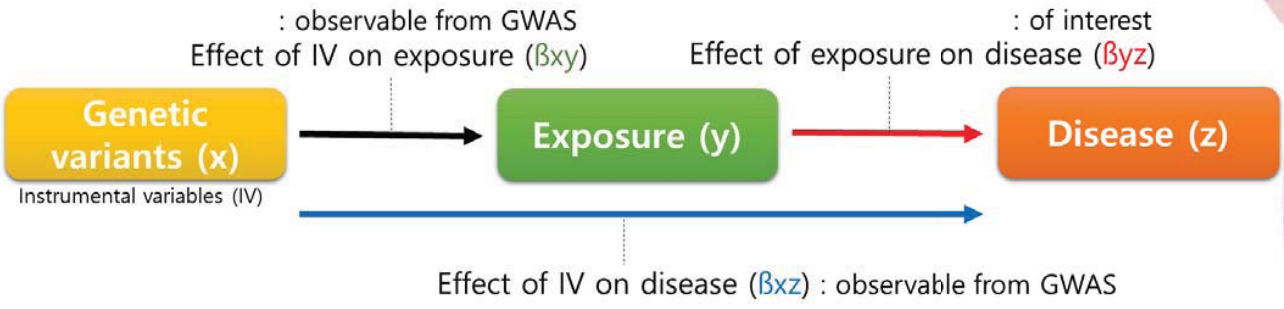
Mendelian randomization (MR)

Assumption II: Genetic variants are NOT associated with the confounder



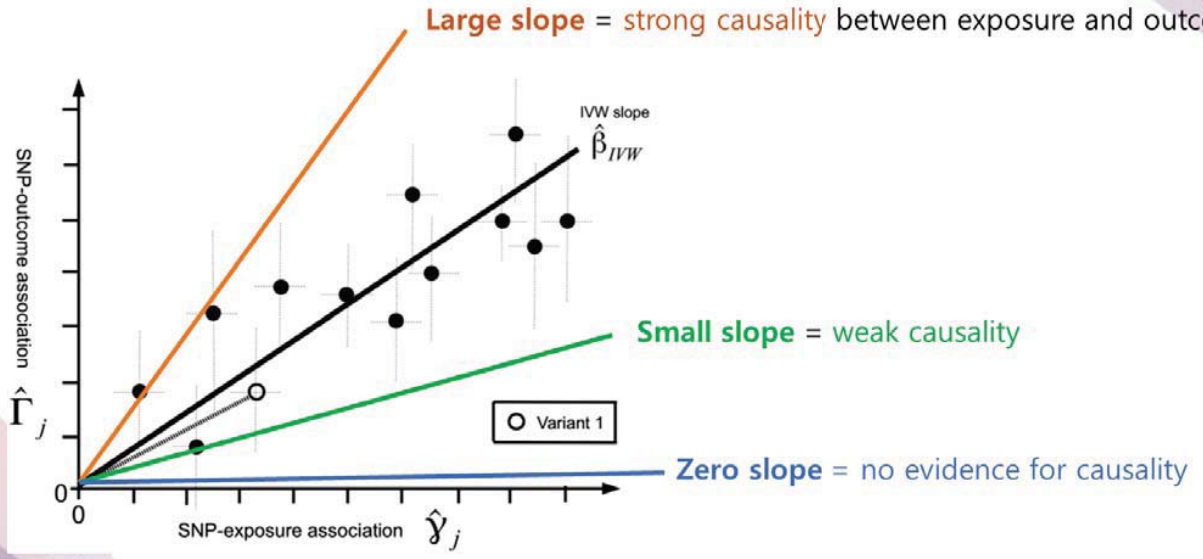
Assumption III: Genetic variants are associated with the outcome through the exposure (vertical pleiotropy)

Mendelian randomization (MR)

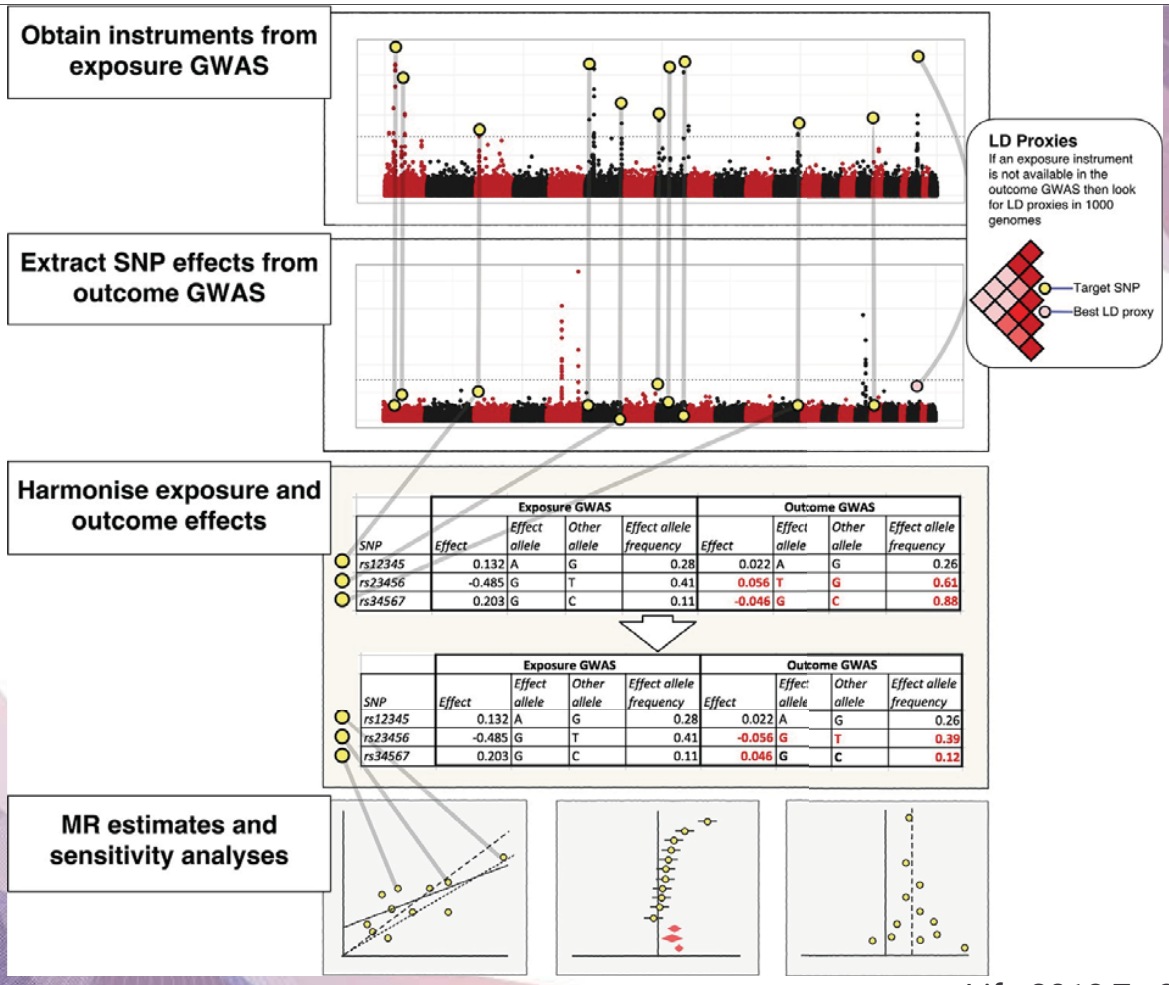


$$\text{Effect of exposure on disease } (\beta yz) = \frac{\text{Effect of IV on disease } (\beta xz)}{\text{Effect of IV on exposure } (\beta xy)}$$

Two-sample MR



Effect of exposure on disease (ßyz) = Effect of IV on disease (ßxz) / Effect of IV on exposure (ßxy)



eLife 2018;7:e34408

Which modifiable risks are causally associated with AD?

RESEARCH

OPEN ACCESS

Modifiable pathways in Alzheimer's disease: Mendelian randomisation analysis

Susanna C Larsson,[1,2] Matthew Traylor,[3] Rainer Malik,[3] Martin Dichgans,[3,4,5] Stephen Burgess,[6,7] Hugh S Markus,[3] for the CoSTREAM Consortium, on behalf of the International Genomics of Alzheimer's Project

ABSTRACT

OBJECTIVE
To determine which potentially modifiable risk factors, including socioeconomic, lifestyle/dietary, cardiometabolic, and inflammatory factors, are associated with Alzheimer's disease.

DESIGN
Mendelian randomisation study using genetic variants associated with the modifiable risk factors as instrumental variables.

SETTING
International Genomics of Alzheimer's Project.

PARTICIPANTS
17 008 cases of Alzheimer's disease and 37 154 controls.

MAIN OUTCOME MEASURES
Odds ratio of Alzheimer's per genetically predicted increase in each modifiable risk factor estimated with Mendelian randomisation analysis.

RESULTS
This study included analyses of 24 potentially modifiable risk factors. A Bonferroni corrected threshold of P=0.002 was considered to be significant, and P<0.05 was considered suggestive of evidence for a potential association. Genetically predicted educational attainment was significantly associated with Alzheimer's. The odds ratios were 0.89 (95% confidence interval 0.86 to 0.93, P=2.4×10⁻⁶) per year of education completed and 0.74 (0.63 to 0.86, P=8.0×10⁻⁵) per unit increase in log odds of having completed college/university. The correlated trait intelligence had a suggestive association with Alzheimer's (per genetically predicted 1 SD higher intelligence: 0.73, 0.57 to 0.93, P=0.01). There was suggestive evidence for potential associations between genetically predicted higher quantity of smoking (per 10 cigarettes a day: 0.69, 0.49 to 0.99, P=0.04) and 25-hydroxyvitamin D concentrations (per 20% higher levels: 0.92, 0.85 to 0.98, P=0.01) and lower odds of Alzheimer's and between higher coffee consumption (per one cup a day: 1.26, 1.05 to 1.51, P=0.01) and higher odds of Alzheimer's. Genetically predicted alcohol consumption, serum folate, serum vitamin B12, homocysteine, cardiometabolic factors, and C reactive protein were not associated with Alzheimer's disease.

CONCLUSION
These results provide support that higher educational attainment is associated with a reduced risk of Alzheimer's disease.

WHAT IS ALREADY KNOWN ON THIS TOPIC

Conventional observational studies have shown that educational attainment is associated with the risk of Alzheimer's disease

Evidence for the associations between lifestyle behaviours and cardiometabolic factors and risk of Alzheimer's disease is inconclusive

Available data on modifiable risk factors in relation to Alzheimer's disease are primarily from conventional observational studies, which are vulnerable to confounding and reverse causation bias

WHAT THIS STUDY ADDS

A Mendelian randomisation approach shows that a genetic predisposition towards longer education is associated with lower odds of Alzheimer's disease

This study found suggestive evidence of possible associations between higher intelligence, smoking, and concentrations of 25-hydroxyvitamin D and lower odds of Alzheimer's disease and between higher coffee consumption and higher odds of Alzheimer's disease

Introduction

Alzheimer's disease is the leading cause of dementia. The chief hallmarks are amyloid plaques and neurofibrillary tangles.[1] The amyloid cascade hypothesis implies that accumulation of amyloid β triggers neuronal dysfunction and cell death in the brain.[1] An alternative theory—the vascular hypothesis—implicates cerebral hypoperfusion as the primary trigger, this driven oxidative stress, deposition of amyloid β, neuroinflammation, blood-brain barrier breakdown, cognitive decline, and neurodegeneration.[2,3]

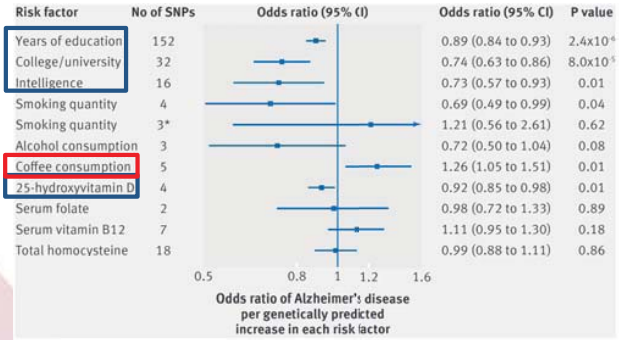
Apart from increasing age and the apolipoprotein E (APOE) ε4 allele, the causes of Alzheimer's disease are largely unknown, and treatment trials have been disappointing.[4] This has led to increasing interest in the potential for reducing Alzheimer's by targeting modifiable risk factors. Conventional observational studies have consistently shown that low educational attainment is associated with an increased risk,[5] and it has been estimated that 19% of cases are potentially attributable to low education.[6] Inconclusive evidence from conventional observational studies indicates that obesity, hypertension, and hypercholesterolaemia in midlife and diabetes, smoking, low vitamin D and folate concentrations, hyperhomocysteinaemia, and high C reactive protein concentrations are associated with increased risk, whereas physical activity, a healthy diet, moderate alcohol drinking, and coffee consumption are associated with decreased risk (table A in appendix 1).[7-10] A 2010 State of the Science report concluded that there was insufficient evidence to support the association with any modifiable factors with risk.[1] Available evidence is in large part

the bmj | BMJ 2017;359:j5375 | doi: 10.1136/bmj.j5375

- Modifiable risk factors
  - Selected for the most consistent evidence for an association with Alzheimer's disease in meta-analyses of prospective observational studies
  - 24 socioeconomic, lifestyle/dietary, cardiometabolic, and inflammatory factors were included

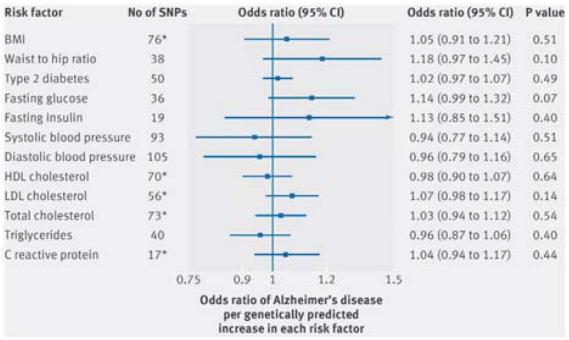
Which modifiable risks are causally associated with AD?

Educational attainment, intelligence, and lifestyle and dietary factors



Risk factor	No of SNPs	Odds ratio (95% CI)	P value
Years of education	152	0.89 (0.84 to 0.93)	2.4×10^{-6}
College/university	32	0.74 (0.63 to 0.86)	8.0×10^{-5}
Intelligence	16	0.73 (0.57 to 0.93)	0.01
Smoking quantity	4	0.69 (0.49 to 0.99)	0.04
Smoking quantity	3*	1.21 (0.56 to 2.61)	0.62
Alcohol consumption	3	0.72 (0.50 to 1.04)	0.08
Coffee consumption	5	1.26 (1.05 to 1.51)	0.01
25-hydroxyvitamin D	4	0.92 (0.85 to 0.98)	0.01
Serum folate	2	0.98 (0.72 to 1.33)	0.89
Serum vitamin B12	7	1.11 (0.95 to 1.30)	0.18
Total homocysteine	18	0.99 (0.88 to 1.11)	0.86

Cardiometabolic and inflammatory factors

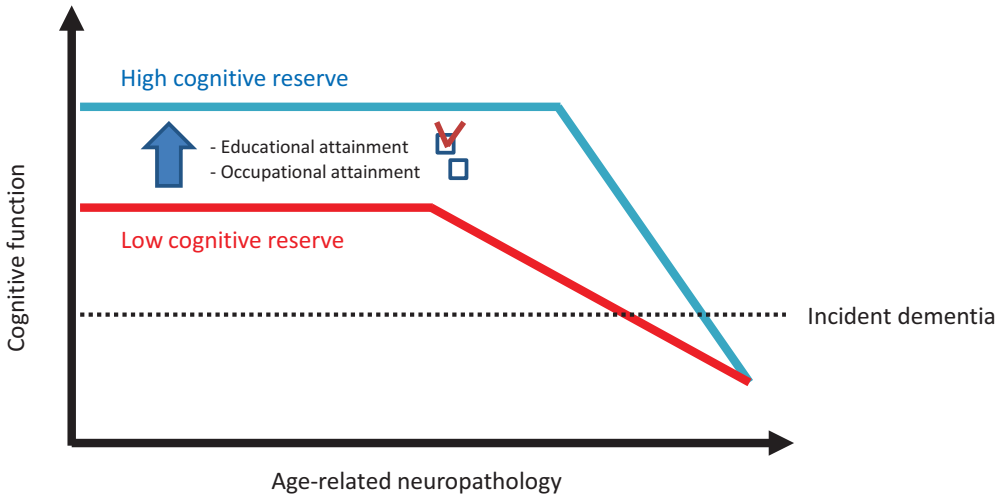


Risk factor	No of SNPs	Odds ratio (95% CI)	P value
BMI	76*	1.05 (0.91 to 1.21)	0.51
Waist to hip ratio	38	1.18 (0.97 to 1.45)	0.10
Type 2 diabetes	50	1.02 (0.97 to 1.07)	0.49
Fasting glucose	36	1.14 (0.99 to 1.32)	0.07
Fasting insulin	19	1.13 (0.85 to 1.51)	0.40
Systolic blood pressure	93	0.94 (0.77 to 1.14)	0.51
Diastolic blood pressure	105	0.96 (0.79 to 1.16)	0.65
HDL cholesterol	70*	0.98 (0.90 to 1.07)	0.64
LDL cholesterol	56*	1.07 (0.98 to 1.17)	0.14
Total cholesterol	73*	1.03 (0.94 to 1.12)	0.54
Triglycerides	40	0.96 (0.87 to 1.06)	0.40
C reactive protein	17*	1.04 (0.94 to 1.17)	0.44

CONCLUSION

These results provide support that higher educational attainment is associated with a reduced risk of Alzheimer's disease.

Cognitive reserve



Does occupational attainment also protect against AD?

BRAIN

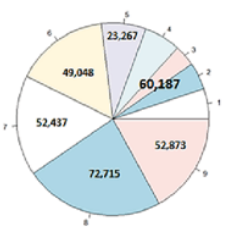
ACCEPTED MANUSCRIPT

Genome-wide association study of occupational attainment as a proxy for cognitive reserve

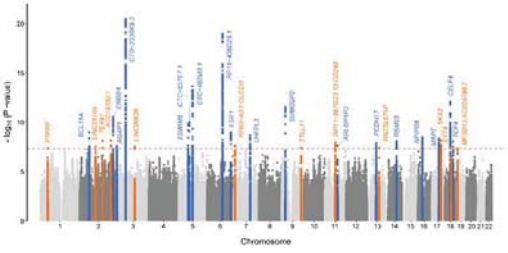
Hyunwoong Ko, Soyeon Kim, Kiwon Kim, Sang-Hyuk Jung, Injeong Shim, Soojin Cha, Hyewon Lee, Beomsu Kim, Joohyun Yoon, Tae Hyon Ha, Seyul Kwak, Jae Myeong Kang, Jun-Young Lee, Jinho Kim, Woong-Yang Park, Kwangsik Nho, Doh Kwan Kim, Woojae Myung ✉, Hong-Hee Won ✉

Brain, awab351, https://doi.org/10.1093/brain/awab351
Published: 06 October 2021 **Article history ▾**

Job Levels	
9. Managers and Senior Officials	52,873
8. Professional Occupations	72,715
7. Associate Professional and Technical Occupations	52,437
6. Administrative and Secretarial Occupations	49,048
5. Skilled Trades Occupations	23,267
4. Personal Service Occupations	18,974
3. Sales and Customer Service Occupations	11,077
2. Process, Plant and Machine Operatives	14,179
1. Elementary Occupations	15,957

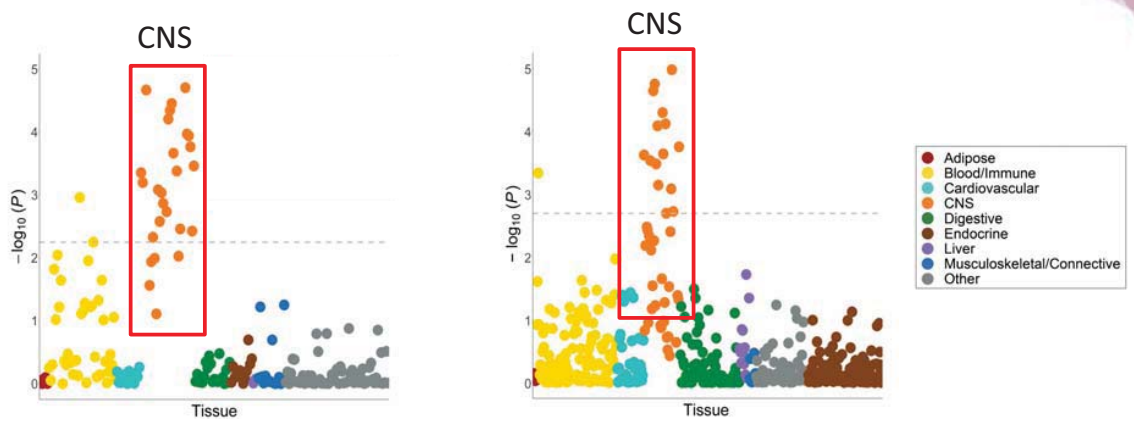


Genome-wide association analysis of OA



30 significant loci (12 novel variants)
SNP-based heritability: 8.5% (s.e. = 0.4%)

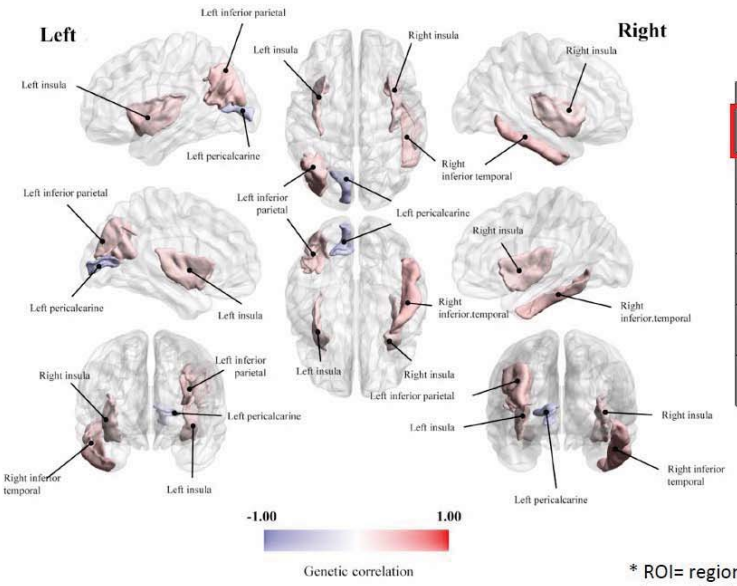
Partitioned heritability was enriched in the central nervous system and brain tissues



Name	Coefficient	Coefficient_std_error	Coefficient_P_value	Coefficient_Pval_FDR
Neuron	5.12E-09	1.60E-09	0.000664049	0.001992147
Oligodendrocyte	9.08E-10	1.53E-09	0.276438945	0.414658417
Astrocyte	-2.79E-09	1.29E-09	0.984970189	0.984970189

Abbreviation: FDR, false discovery rate

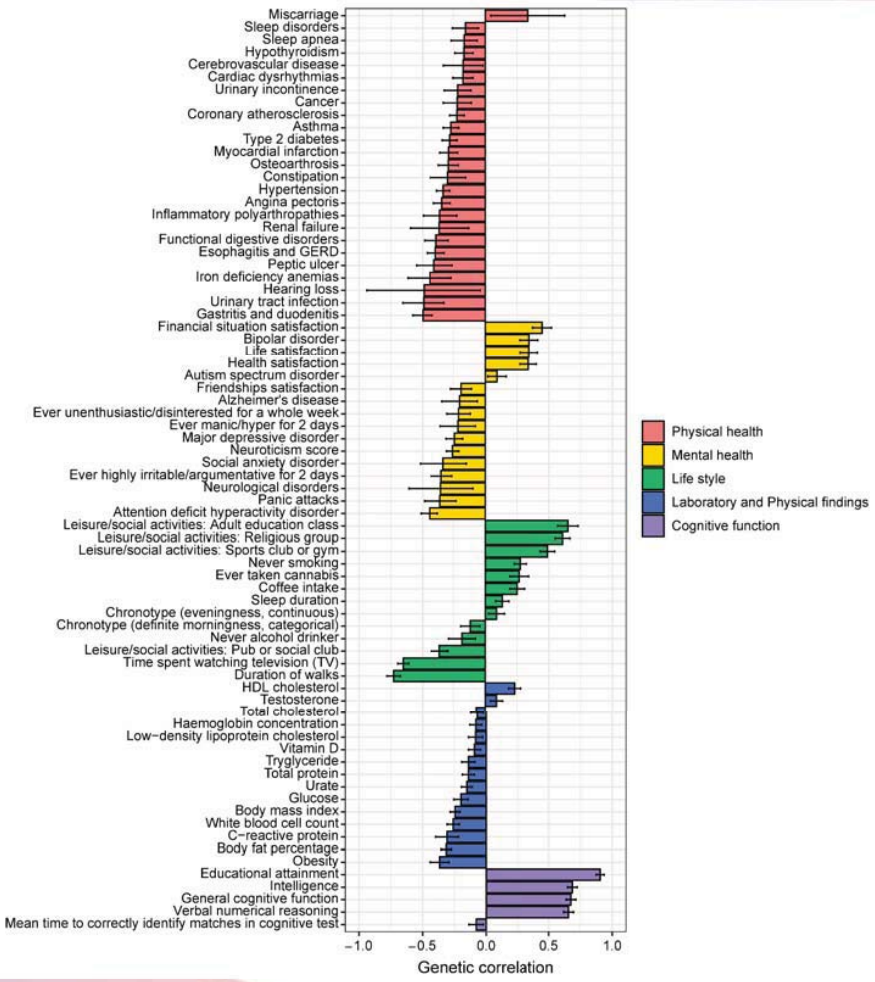
Total brain volume was genetically correlated with occupational attainment



Region	rg	s.e.	P
Total brain volume	0.239	0.045	1.30×10^{-7}
Left insula	0.145	0.038	1.00×10^{-4}
Right inferior temporal	0.210	0.062	6.00×10^{-4}
Right insula	0.133	0.041	1.10×10^{-3}
Left inferior parietal	0.174	0.055	1.50×10^{-3}
Left pericalcarine	-0.182	0.058	1.60×10^{-3}

* ROI= regions of interest (n=33,292) from Zhao et al. Science (2021)
https://github.com/BIG-S2/GWAS

Genetic correlation



MR between occupational attainment and Alzheimer's disease

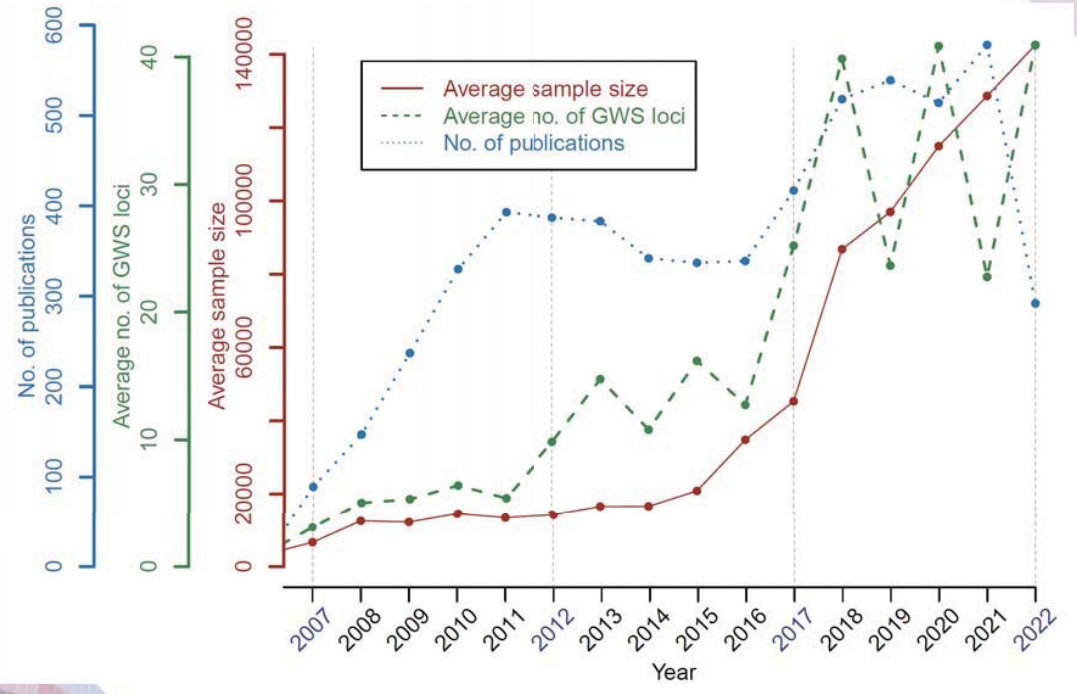
Method	*n* SNPs	OR (95% CI)	*P*
Primary MR for occupational attainment on risk of Alzheimers disease			
Inverse variance weighted	18	0.78 (0.65 to 0.92)	4.26×10^{-3}
Weighted median		0.73 (0.57 to 0.92)	9.10×10^{-3}
MR-Egger (P for pleiotropy = 0.90)		0.73 (0.27 to 1.95)	0.54
Sensitivity analysis for occupational attainment on risk of Alzheimer's disease after the exclusion of pleiotropic SNPs			
Inverse variance weighted	11	0.72 (0.57 to 0.91)	5.54×10^{-3}
Weighted median		0.72 (0.53 to 0.97)	0.03
MR-Egger (P for pleiotropy = 0.97)		0.70 (0.12 to 4.00)	0.70
Sensitivity analysis for independent effect of occupational attainment on risk of Alzheimer's disease by multivariate MR controlling for educational attainment			
Exposure: Occupational attainment			
Inverse variance weighted	69	0.72 (0.54 to 0.95)	0.02
Median based		0.68 (0.48 to 0.97)	0.04
MR-Egger (P for pleiotropy[a] = 0.21)		0.63 (0.45 to 0.89)	8.27×10^{-3}
Exposure: Educational attainment			
Inverse variance weighted	69	1.08 (0.62 to 1.91)	0.78
Median based		1.10 (0.53 to 2.30)	0.79
MR-Egger (P for pleiotropy[a] = 0.21)		0.63 (0.23 to 1.74)	0.38

CI = confidence interval; IV = instrumental variable; OR = odds ratio; SNP = single nucleotide polymorphism.

Summary of post-GWAS analysis and tools

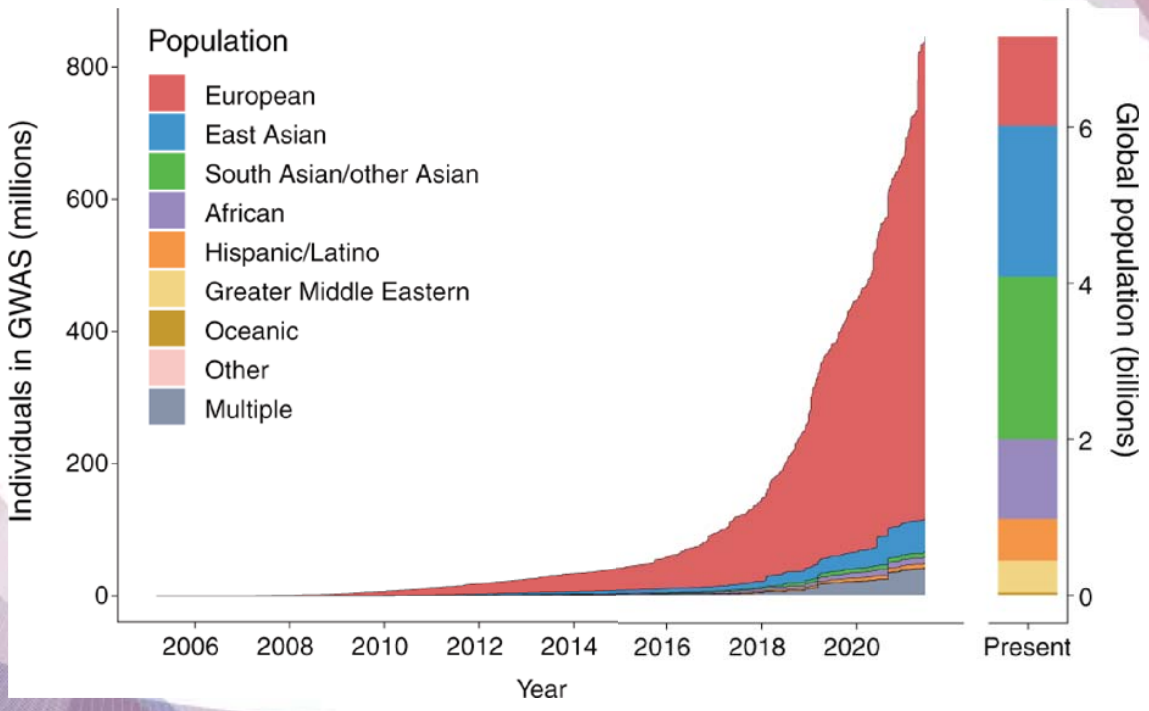
- Understanding genetic architecture
  - SNP-based heritability: LDSC, *GCTA (if genotype available)*
  - Genetic correlation: LDSC (same ancestry), POPCORN or S-LDXR (transancestry)
  - SNP heritability in specific tissues or cells: LDSC-SEG
- Finding causal variants, genes, and pathways
  - Fine-mapping (causal variants): CAVIAR, FINEMAP, PAINTOR, SUSIE
  - eQTL and colocalization analysis (genes): COLOC2
  - Pathway enrichment analysis (pathways or gene sets): MAGMA
- *Identifying individuals at high genetic risk (genotype required)*
  - Polygenic risk score: PRSICE-2, LDPRED, PRS-CS
- Inferring causality between traits
  - Mendelian randomization: MR-BASE, TwoSampleMR (R package)

15 years of GWAS discovery



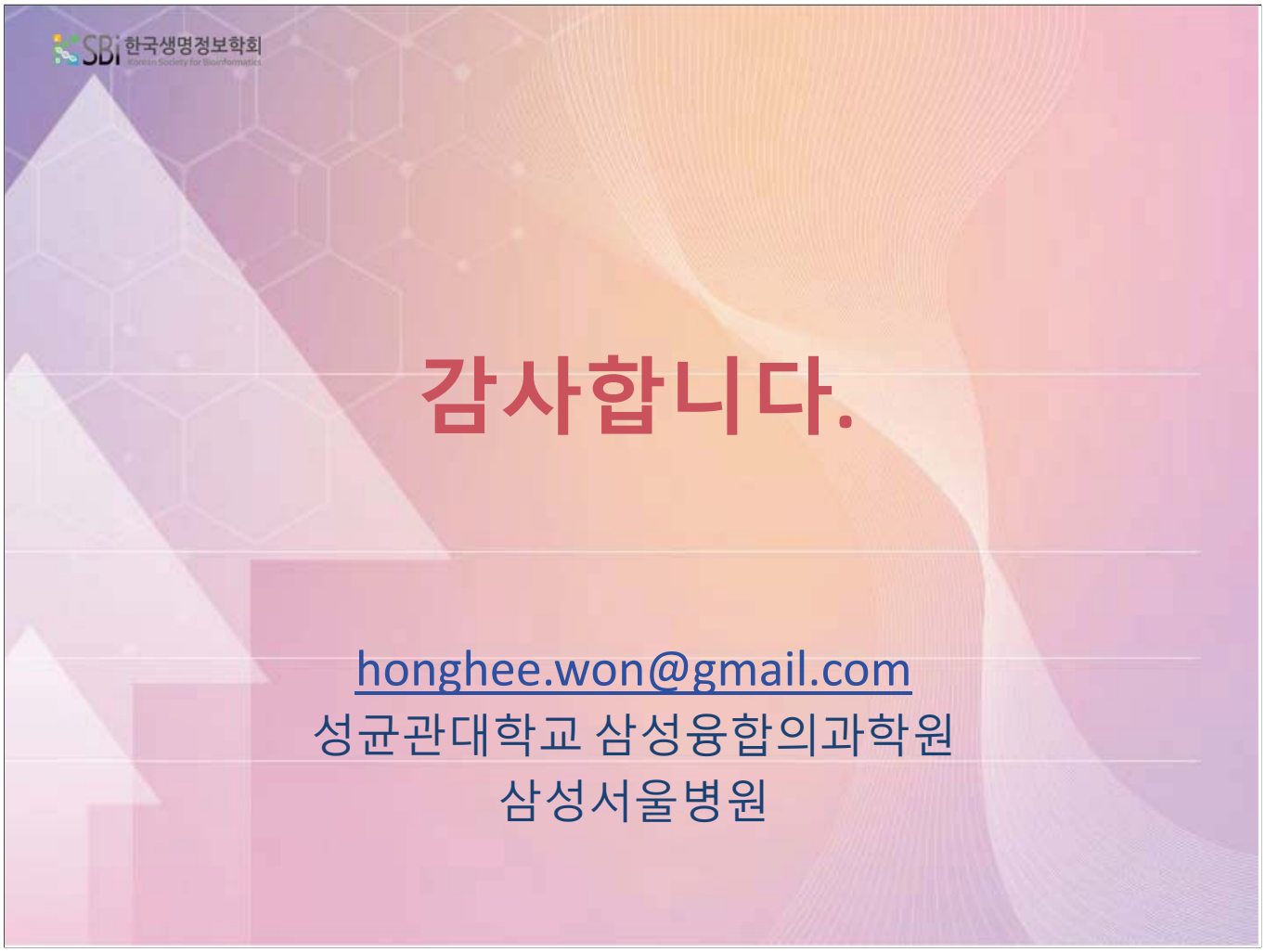
AJHG 110, 1–16, February 2, 2023

As of June 2021, the vast majority (86%) of genomics studies have been conducted in individuals of European descent



Summary

- Common variants account for a large portion of heritability
- Post-GWAS analyses use GWAS summary statistics that are publicly available
- Post-GWAS analyses reveal the genetic architecture of human traits
- Omics data with GWAS are helpful in identifying target genes
- However, the current imbalance between ancestries may limit the clinical utility of genomics in non-European populations


SBi 한국생명정보학회
Korean Society for Bioinformatics
감사합니다.
honghee.won@gmail.com
성균관대학교 삼성융합의과학원
삼성서울병원