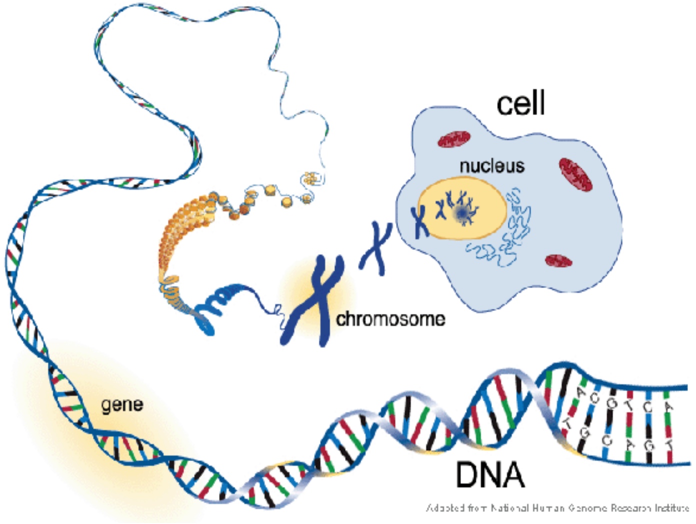


Adapted from National Human Genome Research Institute

Using Statistical and Computational Methods to Identify Genetic Variants in Large-scale Genomic Data

Xiaoyin Li, Ph.D

20/06/2023

ECODEP Seminary

St. Cloud State University

Outline

- Background
- How can we identify disease-related gene loci?
- Which loci in the genome govern the co-occurrence of disorders?
- how to understand the mechanism that genetic variants influence pairs of traits?

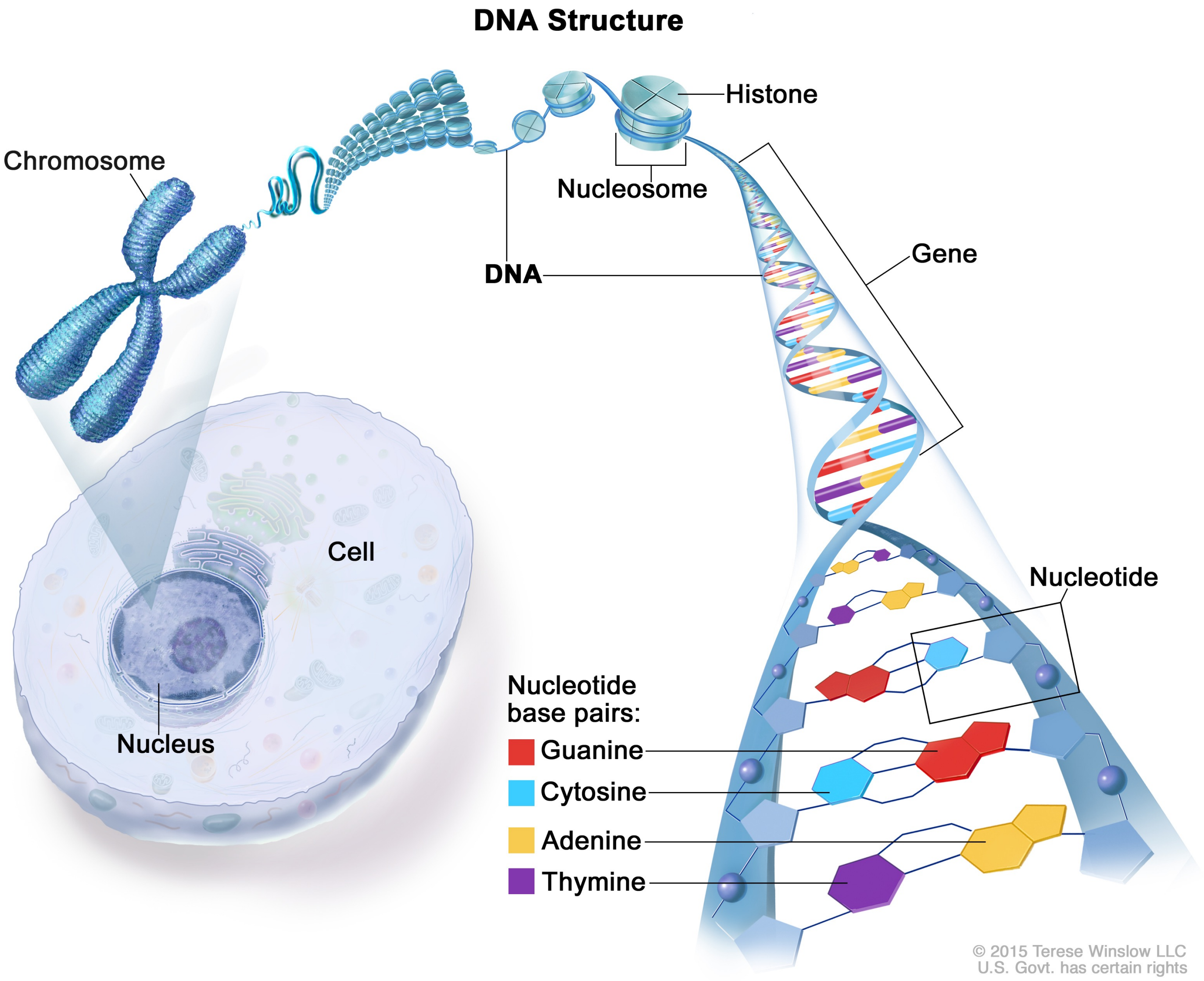
□ Most DNA is found inside the nucleus of a cell, where it forms the **chromosomes**.

□ Chromosomes have proteins called histones that bind to **DNA**.

□ DNA has two strands that twist into the shape of a spiral ladder called a **helix**.

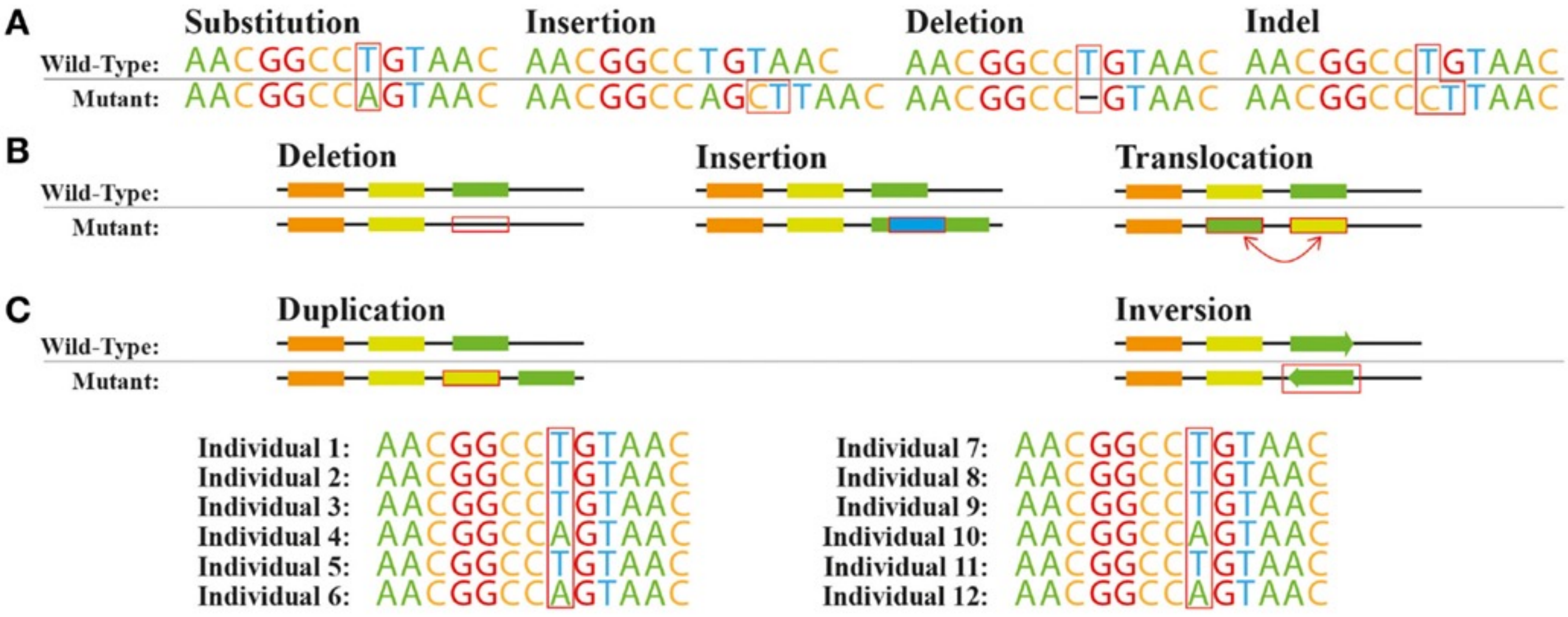
□ DNA is made up of four building blocks called **nucleotides**: adenine (A), thymine (T), guanine (G), and cytosine (C).

□ The nucleotides attach to each other (A with T, and G with C) to form chemical bonds called **base pairs**, which connect the two DNA strands. Genes are short pieces of DNA that carry information for creating proteins.



© 2015 Terese Winslow LLC
U.S. Govt. has certain rights

Type of variation


A
Substitution
Insertion
Deletion
Indel
Wild-Type:
AACGGCCTGTAAC
AACGGCCTGTAAC
AACGGCCTGTAAC
AACGGCCTGTAAC
Mutant:
AACGGCCAGTAAC
AACGGCCAGCTTAAC
AACGGCC-GTAAC
AACGGCCCTTAAC
B
Deletion
Insertion
Translocation
Wild-Type:
Mutant:
C
Duplication
Inversion
Wild-Type:
Mutant:
Individual 1: AACGGCCTGTAAC
Individual 2: AACGGCCTGTAAC
Individual 3: AACGGCCTGTAAC
Individual 4: AACGGCCAGTAAC
Individual 5: AACGGCCTGTAAC
Individual 6: AACGGCCAGTAAC
Individual 7: AACGGCCTGTAAC
Individual 8: AACGGCCTGTAAC
Individual 9: AACGGCCTGTAAC
Individual 10: AACGGCCAGTAAC
Individual 11: AACGGCCTGTAAC
Individual 12: AACGGCCAGTAAC

Nucleotide level variation

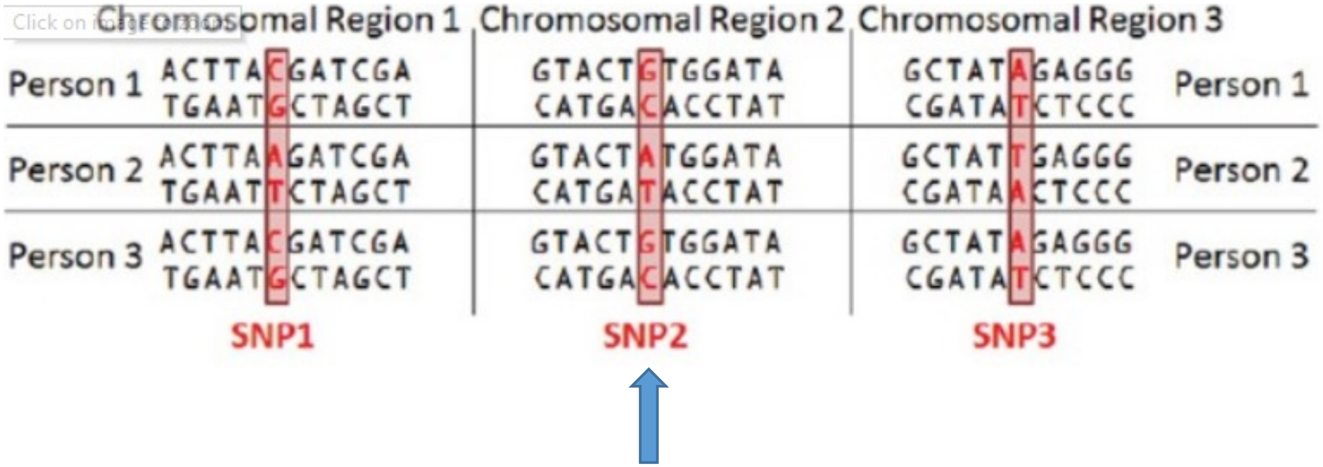
Structural variation

SNP A/T in population

Cardoso et al. 2015
Front. Bioeng. Biotechnol., 16 February 2015 | https://doi.org/10.3389/fbioe.2015.00013

Single Nucleotide Polymorphism (SNP)

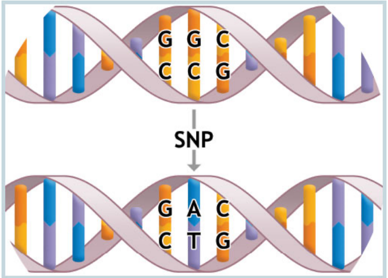
- SNPs occur normally throughout a person's DNA. They occur almost once in every 1,000 nucleotides on average, which means there are roughly 4 to 5 million SNPs in a person's genome.



	Chromosomal Region 1	Chromosomal Region 2	Chromosomal Region 3	
Person 1	ACTTACGATCGA TGAATGCTAGCT	GTACTGTGGATA CATGACACCTAT	GCTATAGAGGG CGATATCTCCC	Person 1
Person 2	ACTTAAGATCGA TGAATTCTAGCT	GTACTATGGATA CATGATACCTAT	GCTATTGAGGG CGATAACTCCC	Person 2
Person 3	ACTTACGATCGA TGAATGCTAGCT	GTACTGTGGATA CATGACACCTAT	GCTATAGAGGG CGATATCTCCC	Person 3
	SNP1	SNP2	SNP3	

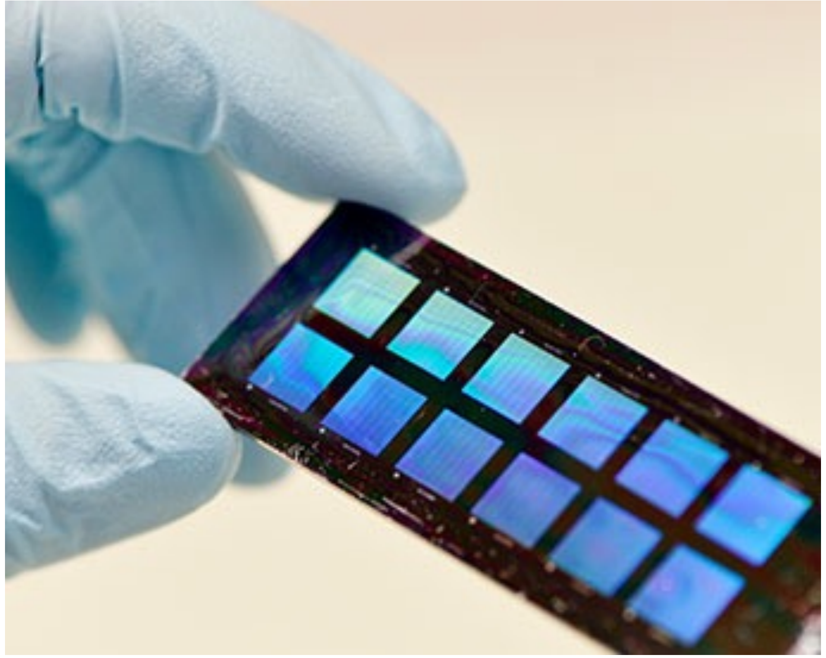
This is a SNP, with alleles: G / A,
minor allele frequency (MAF) = 4%

Genotypes at this SNP
in population
0: GG ~ 92.1%
1: GA ~ 7.7 %
2: AA~ 0.2 %

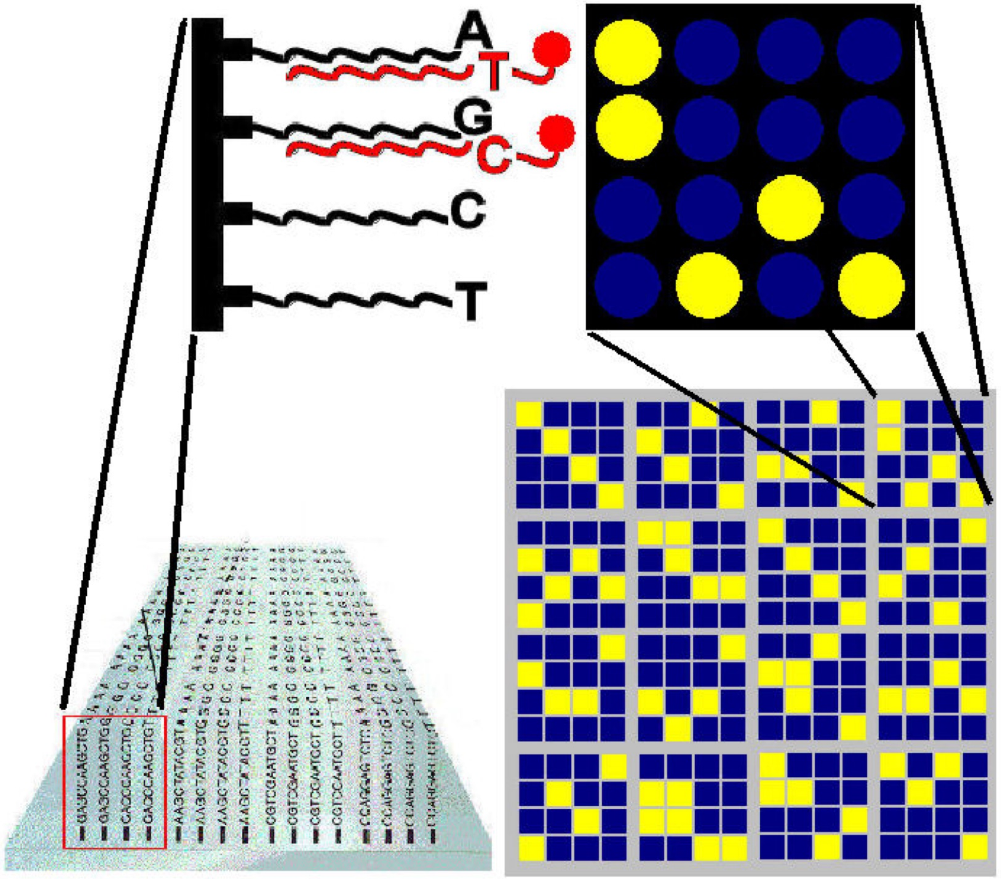


Reading SNPs

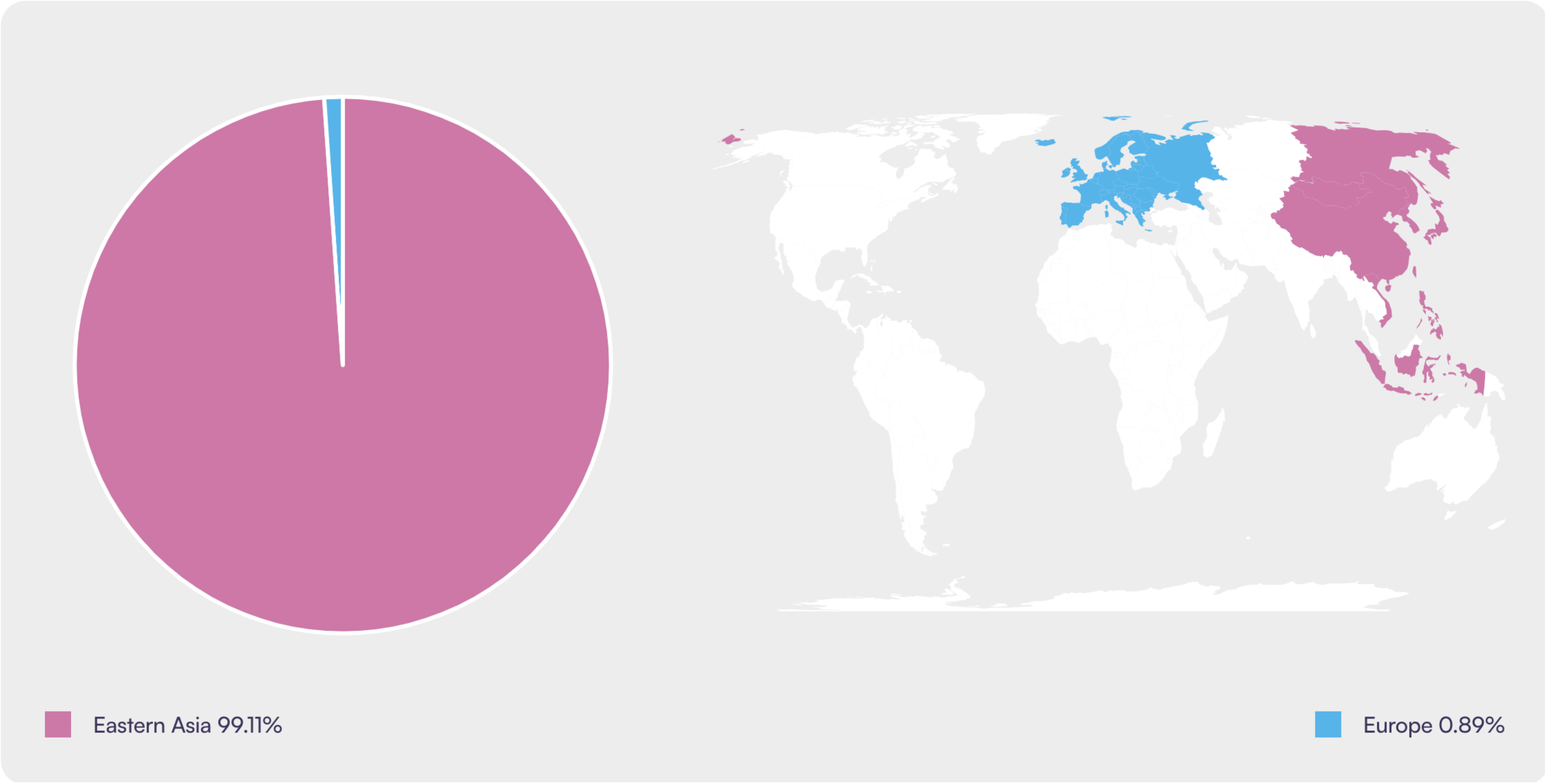
- Human SNP array can measure 10^6 SNPs
- Cost per individual ~100 dollars



This array can genotype 12 individuals at 10^6 SNPs

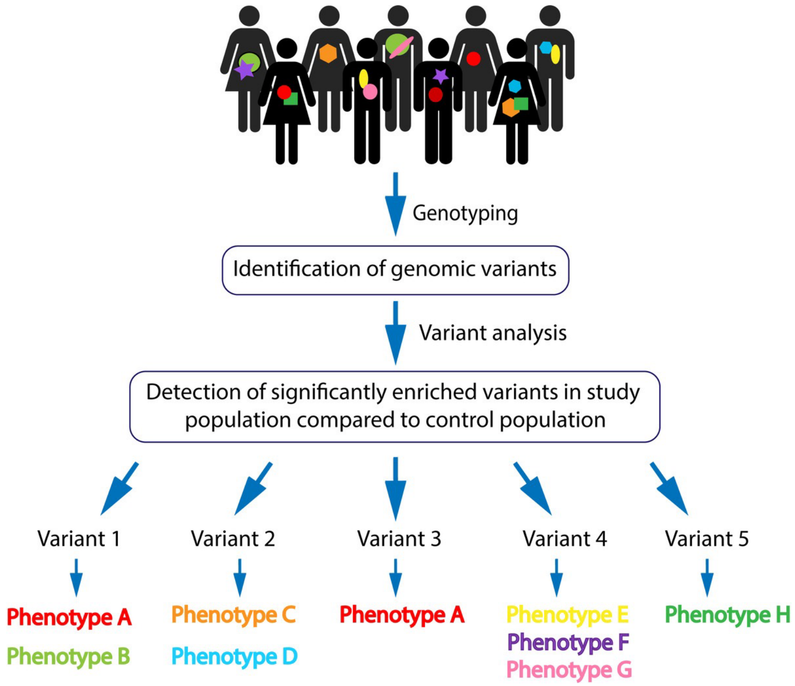


My ancestry analysis results



How can we identify disease-related gene loci?

Genome wide Association Study (GWAS)



Genotyping

Identification of genomic variants

Variant analysis

Detection of significantly enriched variants in study population compared to control population

Variant 1

Phenotype A

Phenotype B

Variant 2

Phenotype C

Phenotype D

Variant 3

Phenotype A

Variant 4

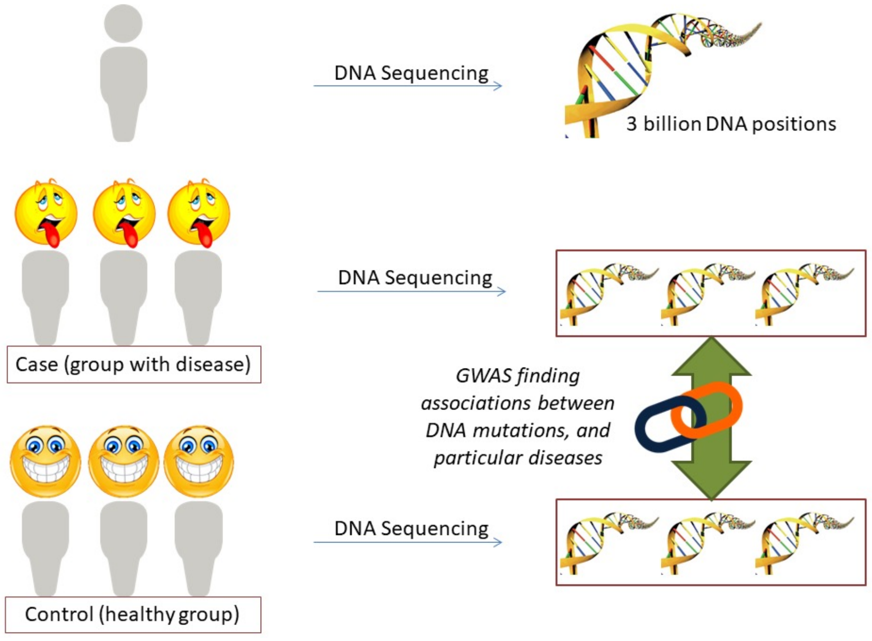
Phenotype E

Phenotype F

Phenotype G

Variant 5

Phenotype H



DNA Sequencing

3 billion DNA positions

DNA Sequencing

Case (group with disease)

GWAS finding associations between DNA mutations, and particular diseases

DNA Sequencing

Control (healthy group)

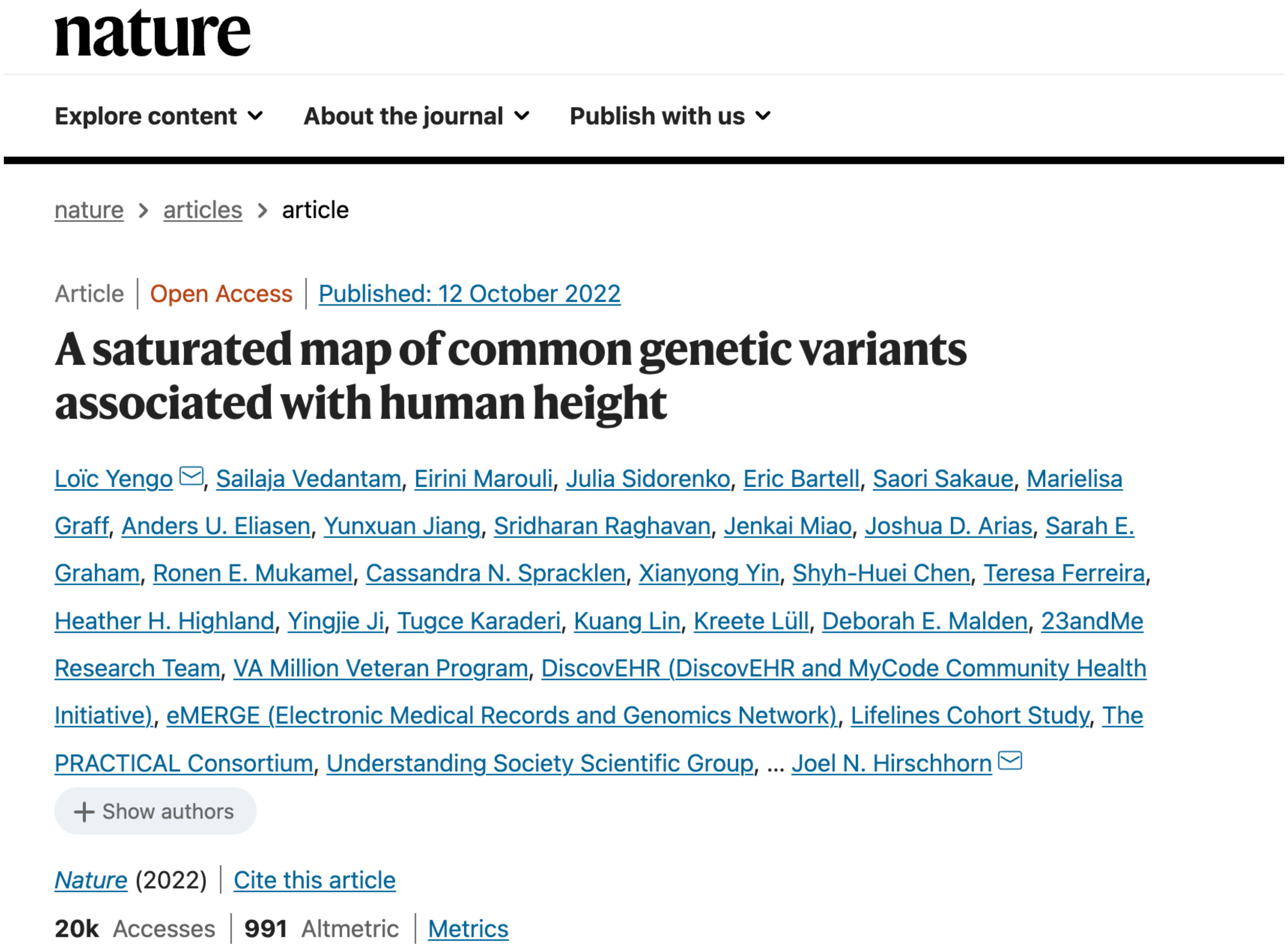
https://www.esat.kuleuven.be/cosic/privacy-preserving-gwas-practical/

St. Cloud State University

Genome wide Association Study (GWAS)

- Aim to identify which regions(or SNPs) in the genome are associated with disease or certain phenotype.
- Design:
  - Identify population structure
  - Select case subjects (those with disease)
  - Select control subjects (healthy)
  - Genotype a million SNPs for each subject
  - Determine which SNP is associated.

- 5.4 million individuals of diverse ancestries with genotype and height available
- 281 studies around the world participated
- 12,111 independent SNPs that are significantly associated with height
- Each locus is a hint to biology of height



Article | Open Access | Published: 12 October 2022

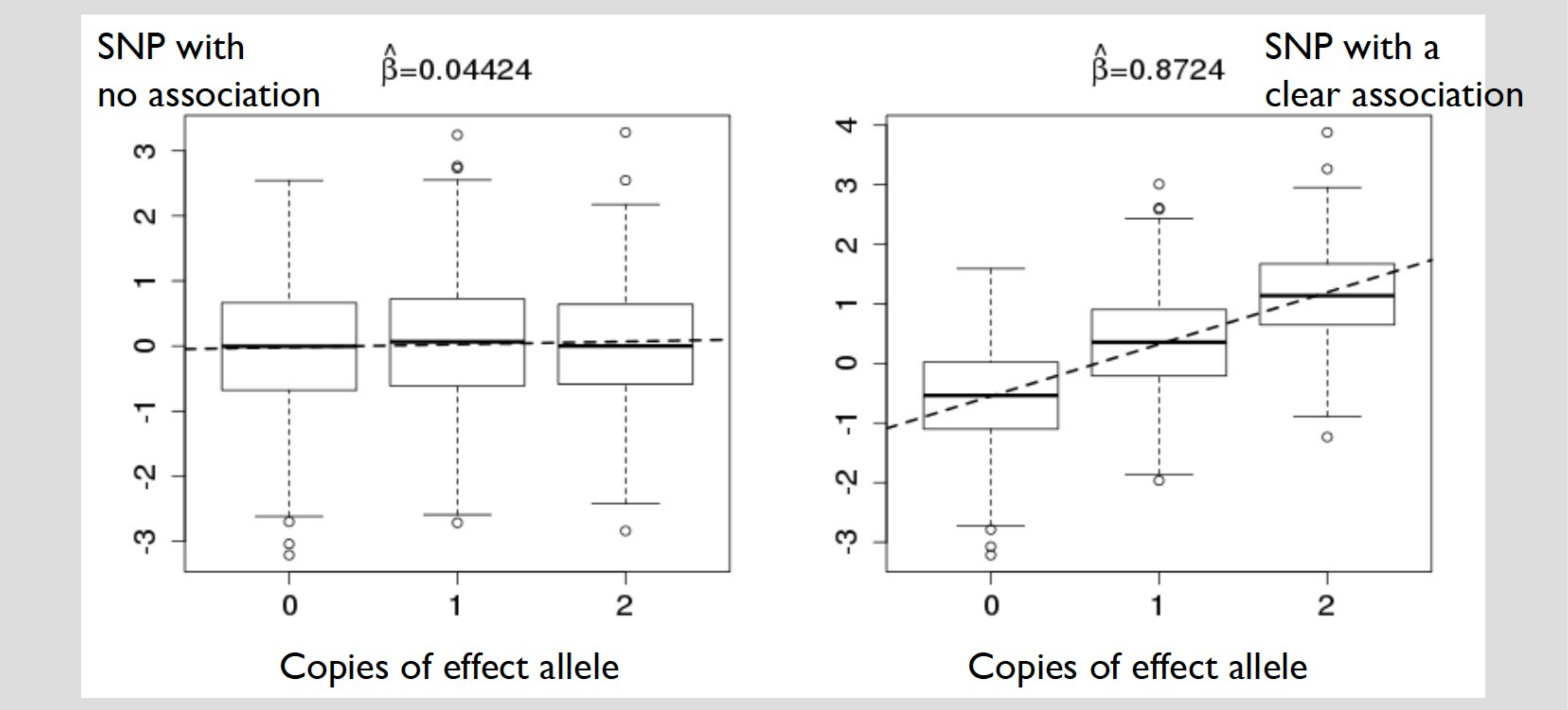
A saturated map of common genetic variants associated with human height

Loïc Yengo, Sailaja Vedantam, Eirini Marouli, Julia Sidorenko, Eric Bartell, Saori Sakaue, Marielisa Graff, Anders U. Eliasen, Yunxuan Jiang, Sridharan Raghavan, Jenkai Miao, Joshua D. Arias, Sarah E. Graham, Ronen E. Mukamel, Cassandra N. Spracklen, Xianyong Yin, Shyh-Huei Chen, Teresa Ferreira, Heather H. Highland, Yingjie Ji, Tugce Karaderi, Kuang Lin, Kreete Lüll, Deborah E. Malden, 23andMe Research Team, VA Million Veteran Program, DiscovEHR (DiscovEHR and MyCode Community Health Initiative), eMERGE (Electronic Medical Records and Genomics Network), Lifelines Cohort Study, The PRACTICAL Consortium, Understanding Society Scientific Group, … Joel N. Hirschhorn

Nature (2022) | Cite this article

20k Accesses | **991** Altmetric | Metrics

Association test: ”Does the mean height differ between genotype groups?”



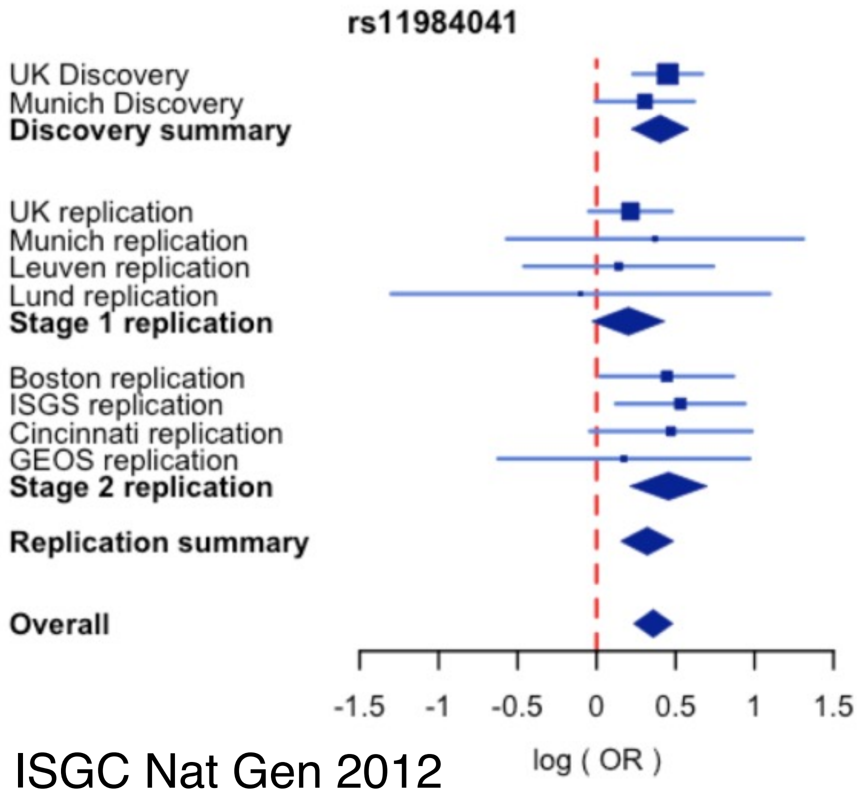
(output are linear regression slope $\hat{\beta}$, its standard error SE and P-value)

meta-analysis

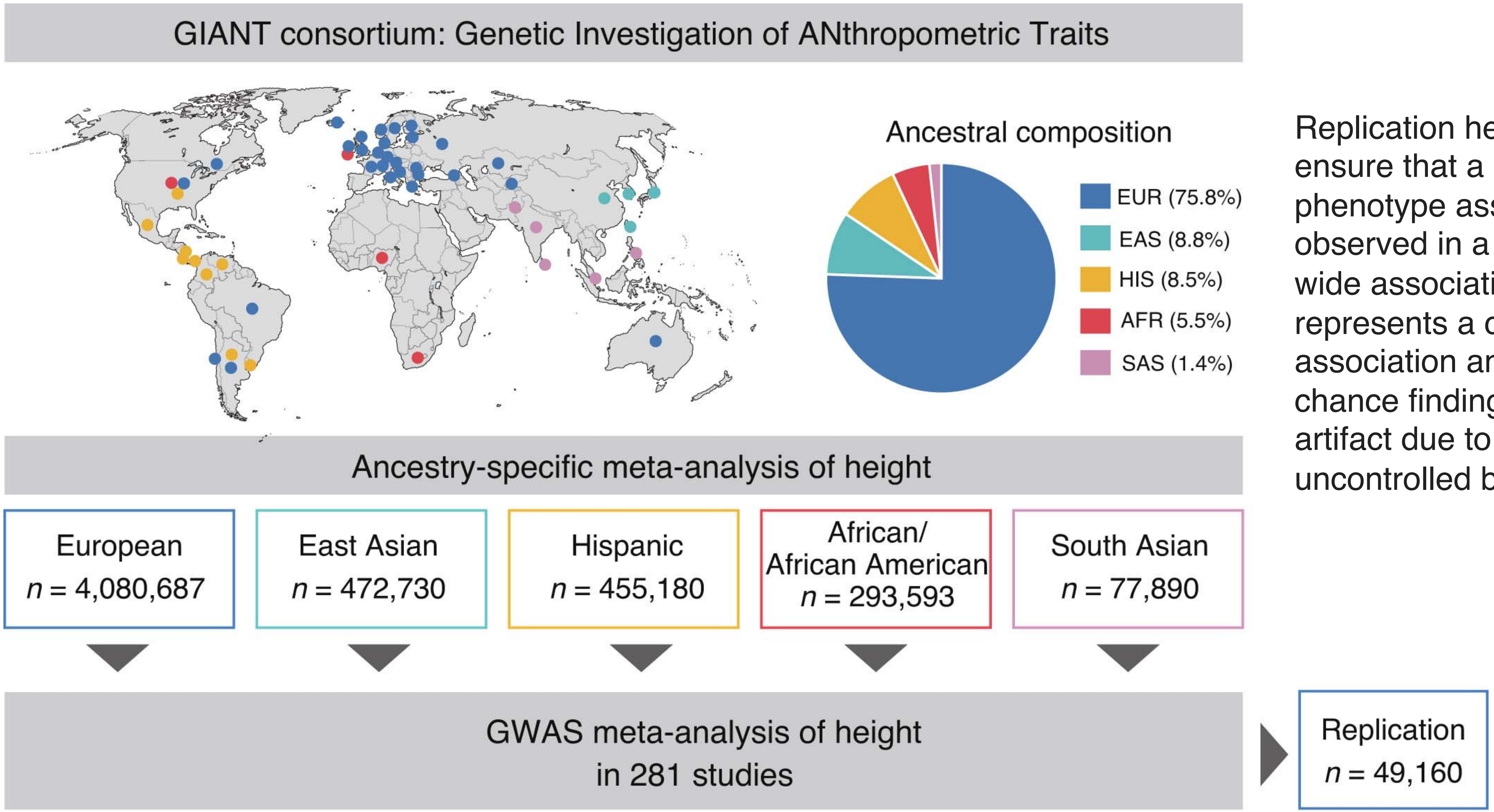
- 5.4 million individuals of diverse ancestries
- 281 studies around the world participated

A **meta-analysis** is a statistical analysis that combines the results of multiple scientific studies on the same question.

Here it works on GWAS results, not requiring original genotype-phenotype data.



ISGC Nat Gen 2012



GIANT consortium: Genetic Investigation of ANthropometric Traits

Ancestry-specific meta-analysis of height

European $n = 4{,}080{,}687$

East Asian $n = 472{,}730$

Hispanic $n = 455{,}180$

African/ African American $n = 293{,}593$

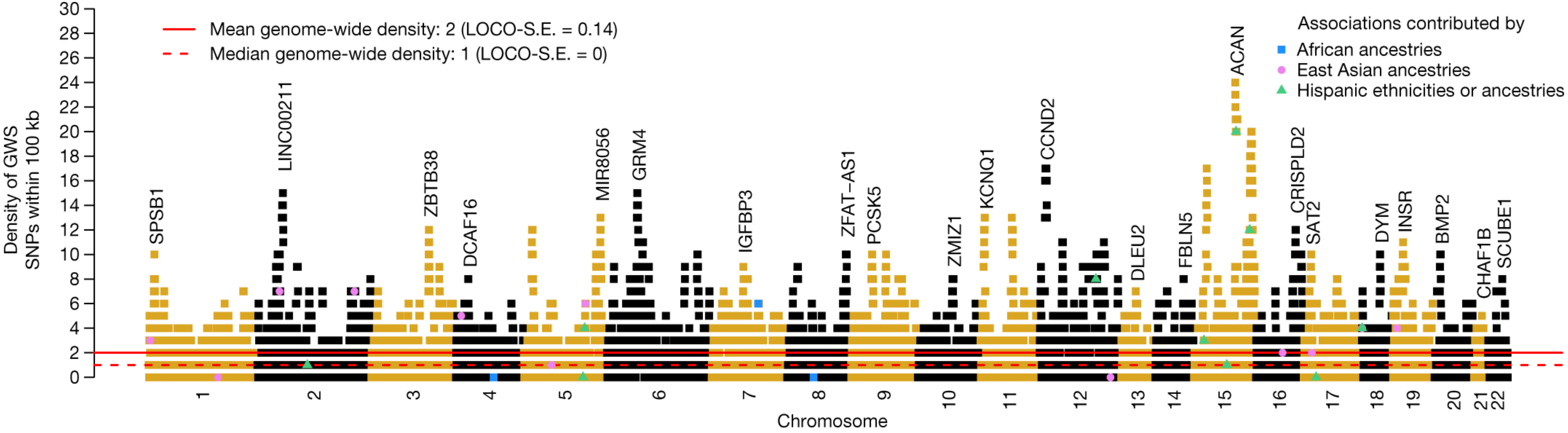
South Asian $n = 77{,}890$

GWAS meta-analysis of height in 281 studies

Replication $n = 49{,}160$

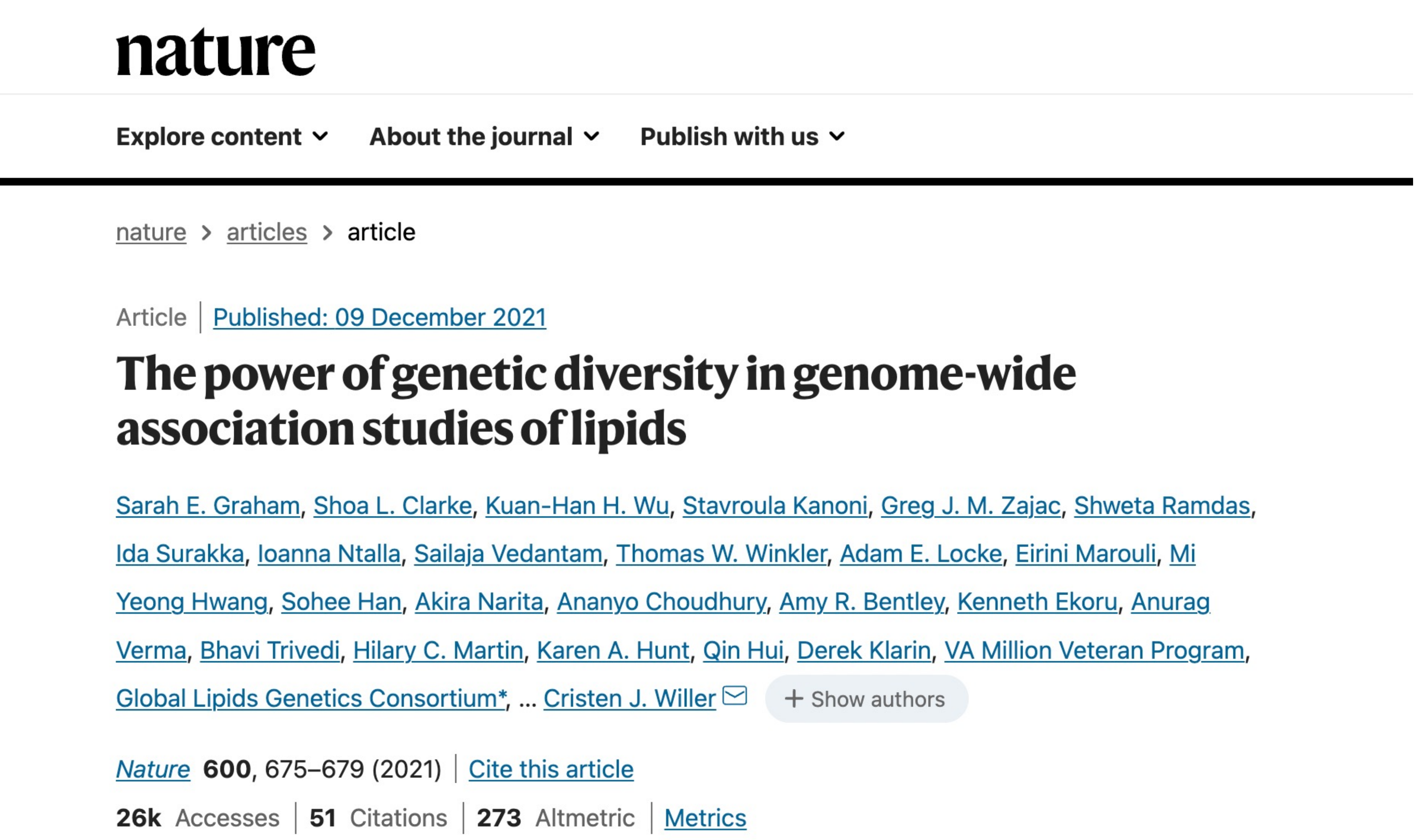
Replication helps ensure that a genotype-phenotype association observed in a genome-wide association study represents a credible association and is not a chance finding or an artifact due to uncontrolled biases.

Brisbane plot



- Each dot represents one of the 12,111 quasi-independent GWS ($P<5\times10^{-8}$) height-associated SNPs identified using our cross-ancestry GWAS meta-analysis.
- GWS SNPs with the largest density on each chromosome were annotated with the closest gene.
- Signal density was calculated for each associated SNP as the number of other independent associations within 100 kb.

St. Cloud State University



Article | Published: 09 December 2021

The power of genetic diversity in genome-wide association studies of lipids

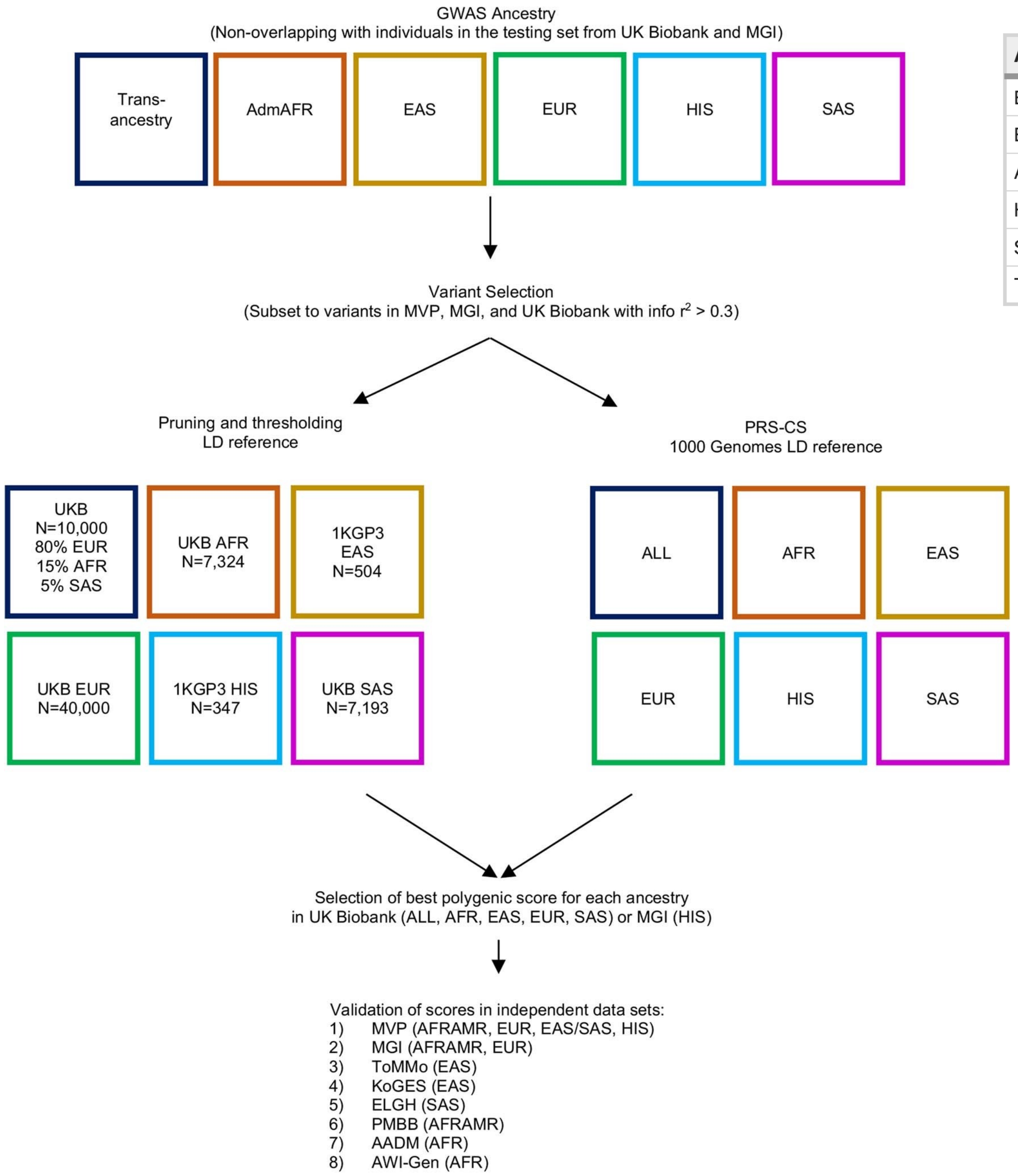
Sarah E. Graham, Shoa L. Clarke, Kuan-Han H. Wu, Stavroula Kanoni, Greg J. M. Zajac, Shweta Ramdas, Ida Surakka, Ioanna Ntalla, Sailaja Vedantam, Thomas W. Winkler, Adam E. Locke, Eirini Marouli, Mi Yeong Hwang, Sohee Han, Akira Narita, Ananyo Choudhury, Amy R. Bentley, Kenneth Ekoru, Anurag Verma, Bhavi Trivedi, Hilary C. Martin, Karen A. Hunt, Qin Hui, Derek Klarin, VA Million Veteran Program, Global Lipids Genetics Consortium*, … Cristen J. Willer ✉ + Show authors

Nature **600**, 675–679 (2021) | Cite this article

26k Accesses | **51** Citations | **273** Altmetric | Metrics

- A multi-ancestry, GWAS meta-analysis of lipid levels
- Approximately 1.65 million individuals, (350,000 of non-European).
- 91 million variants

St. Cloud State University



GWAS Ancestry
(Non-overlapping with individuals in the testing set from UK Biobank and MGI)

Trans-ancestry | AdmAFR | EAS | EUR | HIS | SAS

Variant Selection
(Subset to variants in MVP, MGI, and UK Biobank with info $r^2 > 0.3$)

Pruning and thresholding
LD reference

UKB
N=10,000
80% EUR
15% AFR
5% SAS

UKB AFR
N=7,324

1KGP3
EAS
N=504

UKB EUR
N=40,000

1KGP3 HIS
N=347

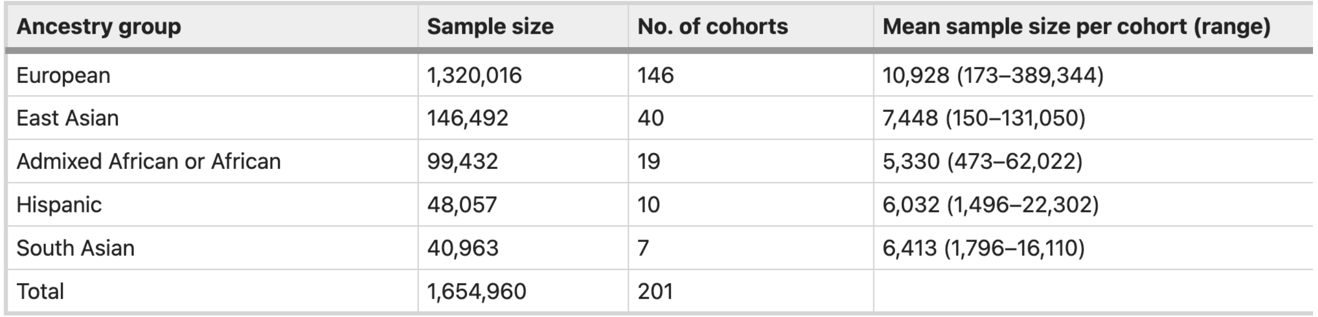
UKB SAS
N=7,193

PRS-CS
1000 Genomes LD reference

ALL | AFR | EAS | EUR | HIS | SAS

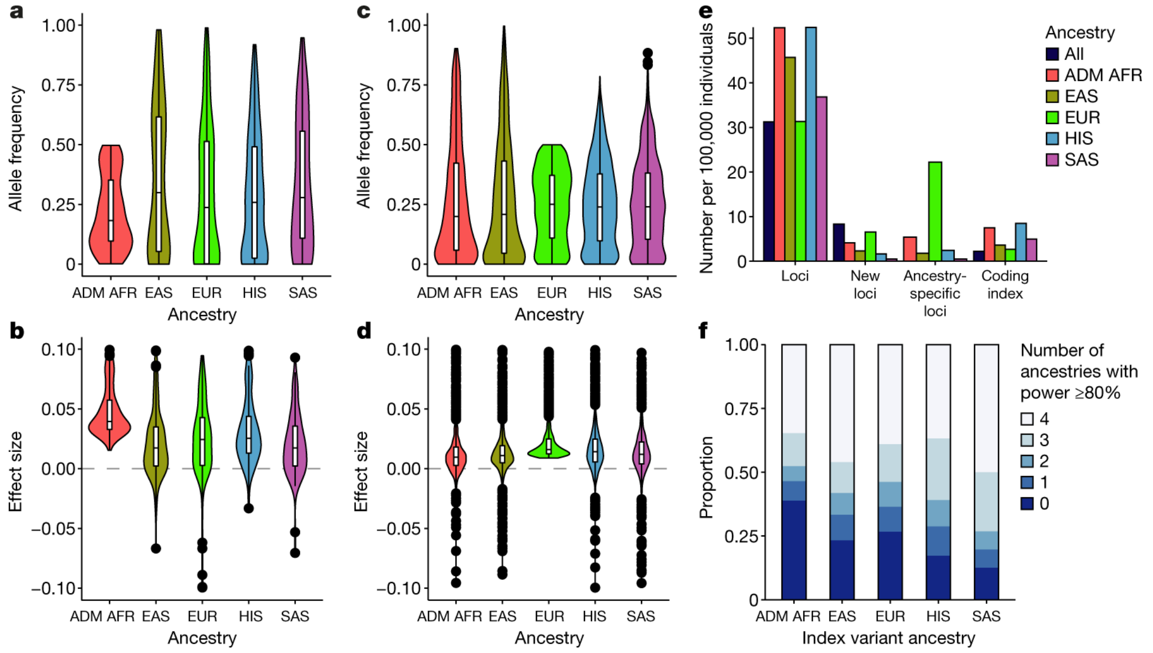
Selection of best polygenic score for each ancestry
in UK Biobank (ALL, AFR, EAS, EUR, SAS) or MGI (HIS)

Validation of scores in independent data sets:
1) MVP (AFRAMR, EUR, EAS/SAS, HIS)
2) MGI (AFRAMR, EUR)
3) ToMMo (EAS)
4) KoGES (EAS)
5) ELGH (SAS)
6) PMBB (AFRAMR)
7) AADM (AFR)
8) AWI-Gen (AFR)

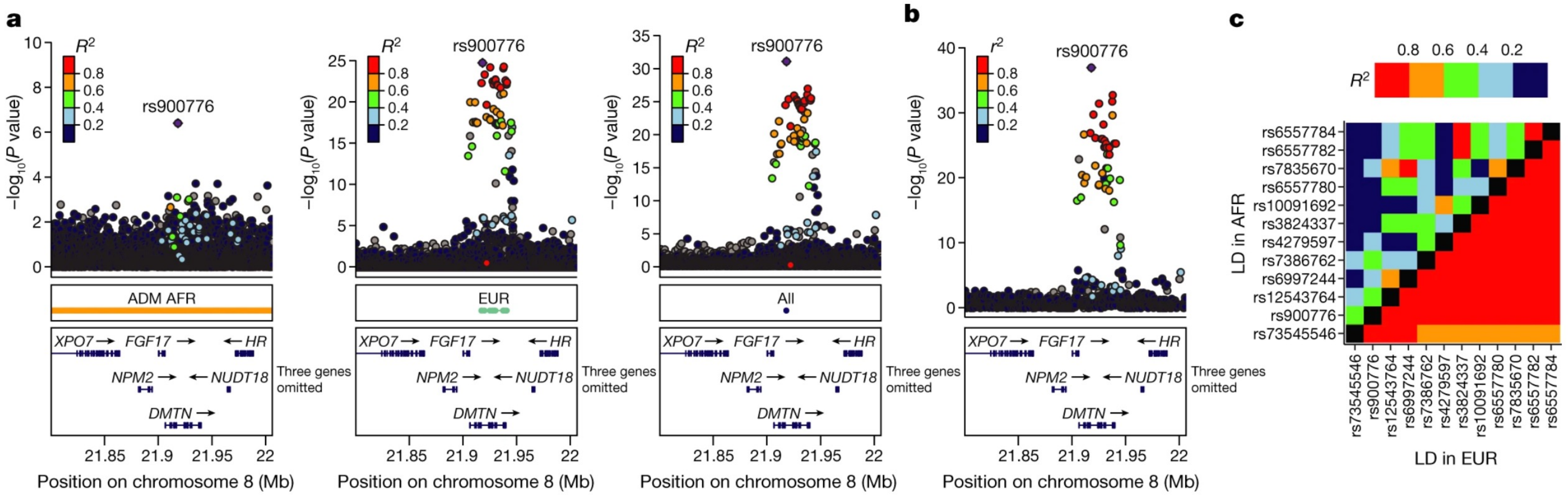


Ancestry group	Sample size	No. of cohorts	Mean sample size per cohort (range)
European	1,320,016	146	10,928 (173–389,344)
East Asian	146,492	40	7,448 (150–131,050)
Admixed African or African	99,432	19	5,330 (473–62,022)
Hispanic	48,057	10	6,032 (1,496–22,302)
South Asian	40,963	7	6,413 (1,796–16,110)
Total	1,654,960	201	

We found 773 lipid-associated genomic regions that contained 1,765 distinct index variants that reached genome-wide significance.



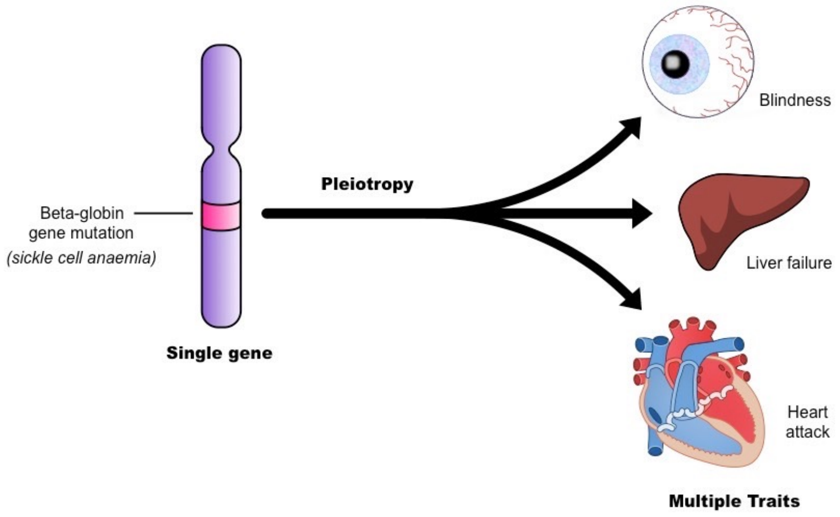
Fine-mapping of rs900776



a, **b**, Association of the *DMTN* intron variant rs900776 with LDL-C in the admixed African, European, or multi-ancestry meta-analysis (**a**) or *DMTN* expression quantitative trait loci (**b**). **c**, The LD patterns for variants in the European ancestry 99% credible set differ greatly between African (AFR) and European ancestry individuals in 1000 Genomes.

What is pleiotropy?

- Pleiotropy occurs when a single genetic variant (gene) influences multiple traits.
- A recent study analyzed publicly available GWAS on 558 unique traits, discovered that 90% of those loci are associated with multiple trait domains. (*Watanabe et al. 2019 nature genetics*)
- Dissecting the association pathways from a variant to multiple traits is extremely important but has not been well studied.



http://ib.bioninja.com.au/standard-level/topic-3-genetics/34-inheritance/pleiotropy.html

Which loci in the genome govern the co-occurrence of disorders?

How to detect pleiotropy?

Cross Phenotype association implies potential pleiotropy where a variant is associated with multiple traits regardless of underlying causes. Detecting cross-phenotype effects using GWAS data can help us to identify pleiotropy variants.

- ➢ **Multivariate regression**: the response variable will be a matrix, where each row represents an individual and each column represents one phenotype.
- ➢ **Univariate regression**: the response variable (i.e. the phenotype) will be a vector, with one data point for each individual in the study

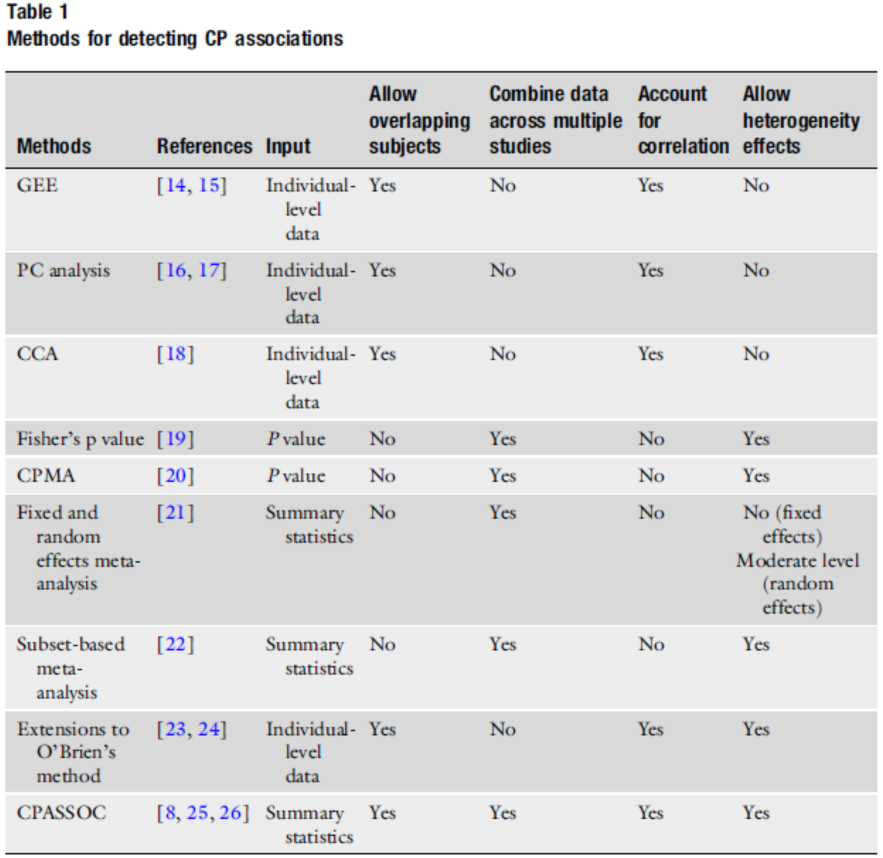


Table 1
Methods for detecting CP associations

Methods	References	Input	Allow overlapping subjects	Combine data across multiple studies	Account for correlation	Allow heterogeneity effects
GEE	[14, 15]	Individual-level data	Yes	No	Yes	No
PC analysis	[16, 17]	Individual-level data	Yes	No	Yes	No
CCA	[18]	Individual-level data	Yes	No	Yes	No
Fisher's p value	[19]	*P* value	No	Yes	No	Yes
CPMA	[20]	*P* value	No	Yes	No	Yes
Fixed and random effects meta-analysis	[21]	Summary statistics	No	Yes	No	No (fixed effects) Moderate level (random effects)
Subset-based meta-analysis	[22]	Summary statistics	No	Yes	No	Yes
Extensions to O'Brien's method	[23, 24]	Individual-level data	Yes	No	Yes	Yes
CPASSOC	[8, 25, 26]	Summary statistics	Yes	Yes	Yes	Yes

X Li, X Zhu - Statistical Human Genetics, 2017

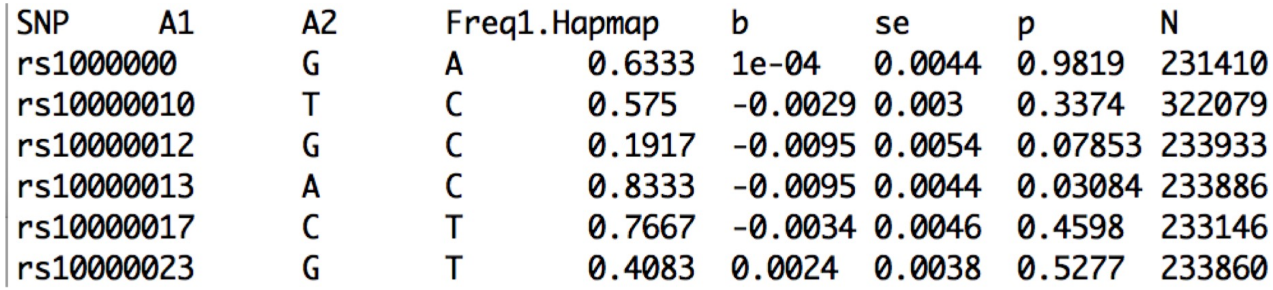
Cross Phenotype Association Analysis (CPASSOC)

Cross Phenotype Association Analysis can integrate association evidence from multiple correlated continuous and binary traits via summary statistics.

There are advantages to using summary statistics instead of individual-level data.

➢First, there is no asymptotic efficiency gain by analyzing individual-level data.

➢Second, in practice it is easier and more feasible to obtain summary statistics than individual-level data.

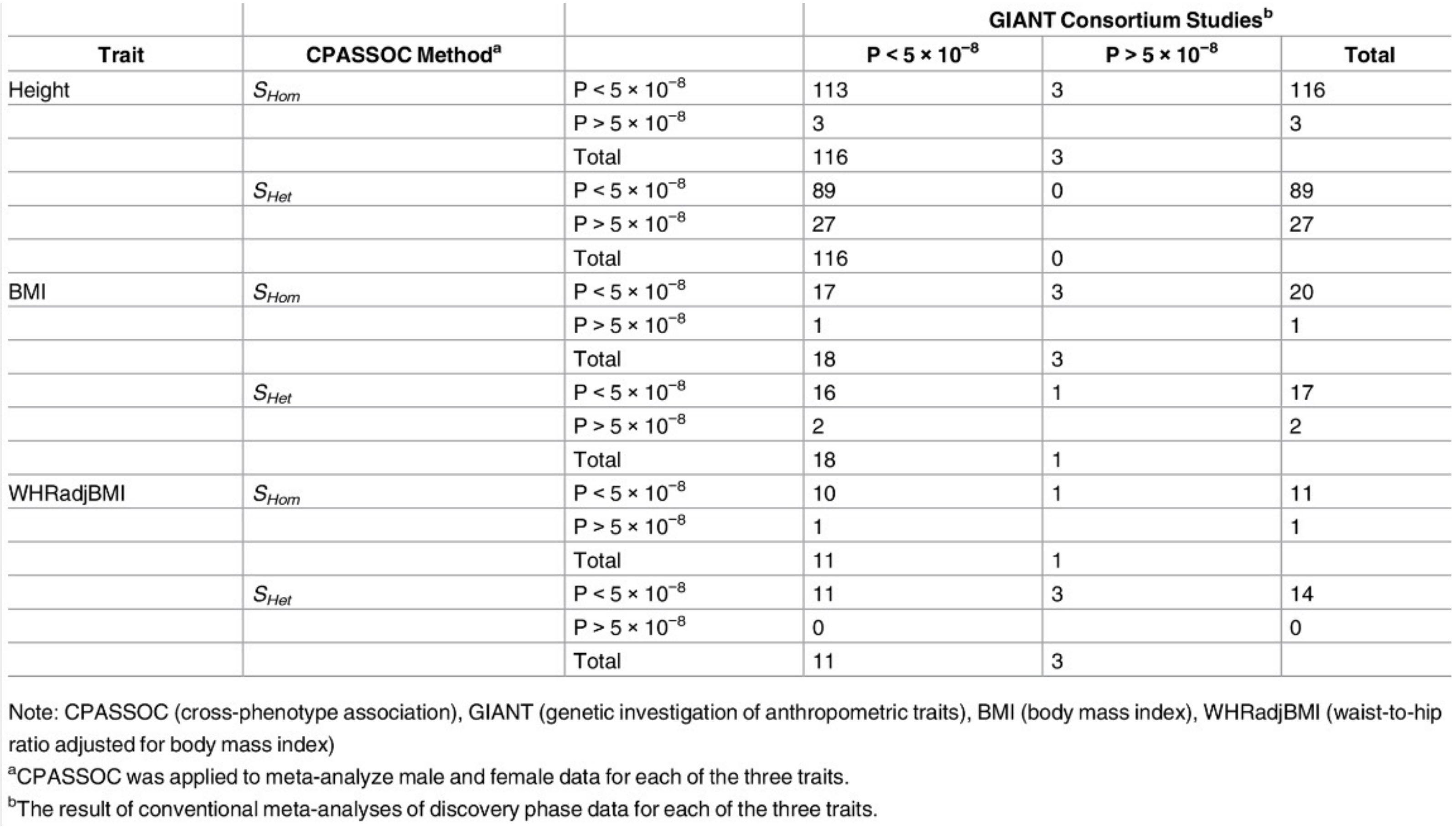


```
SNP      A1       A2       Freq1.Hapmap     b        se       p        N
rs1000000         G        A        0.6333  1e-04    0.0044   0.9819   231410
rs10000010        T        C        0.575   -0.0029  0.003    0.3374   322079
rs10000012        G        C        0.1917  -0.0095  0.0054   0.07853  233933
rs10000013        A        C        0.8333  -0.0095  0.0044   0.03084  233886
rs10000017        C        T        0.7667  -0.0034  0.0046   0.4598   233146
rs10000023        G        T        0.4083  0.0024   0.0038   0.5277   233860
```

While no-one has access to all original genotype-phenotype data, everyone can access the meta-analyzed GWAS results as they are (often) publicly available.

ST. CLOUD STATE UNIVERSITY

CPASSOC VS conventional GWAS



			GIANT Consortium Studies[b]		
Trait	**CPASSOC Method[a]**		**$P < 5 \times 10^{-8}$**	**$P > 5 \times 10^{-8}$**	**Total**
Height	S_{Hom}	$P < 5 \times 10^{-8}$	113	3	116
		$P > 5 \times 10^{-8}$	3		3
		Total	116	3	
	S_{Het}	$P < 5 \times 10^{-8}$	89	0	89
		$P > 5 \times 10^{-8}$	27		27
		Total	116	0	
BMI	S_{Hom}	$P < 5 \times 10^{-8}$	17	3	20
		$P > 5 \times 10^{-8}$	1		1
		Total	18	3	
	S_{Het}	$P < 5 \times 10^{-8}$	16	1	17
		$P > 5 \times 10^{-8}$	2		2
		Total	18	1	
WHRadjBMI	S_{Hom}	$P < 5 \times 10^{-8}$	10	1	11
		$P > 5 \times 10^{-8}$	1		1
		Total	11	1	
	S_{Het}	$P < 5 \times 10^{-8}$	11	3	14
		$P > 5 \times 10^{-8}$	0		0
		Total	11	3	

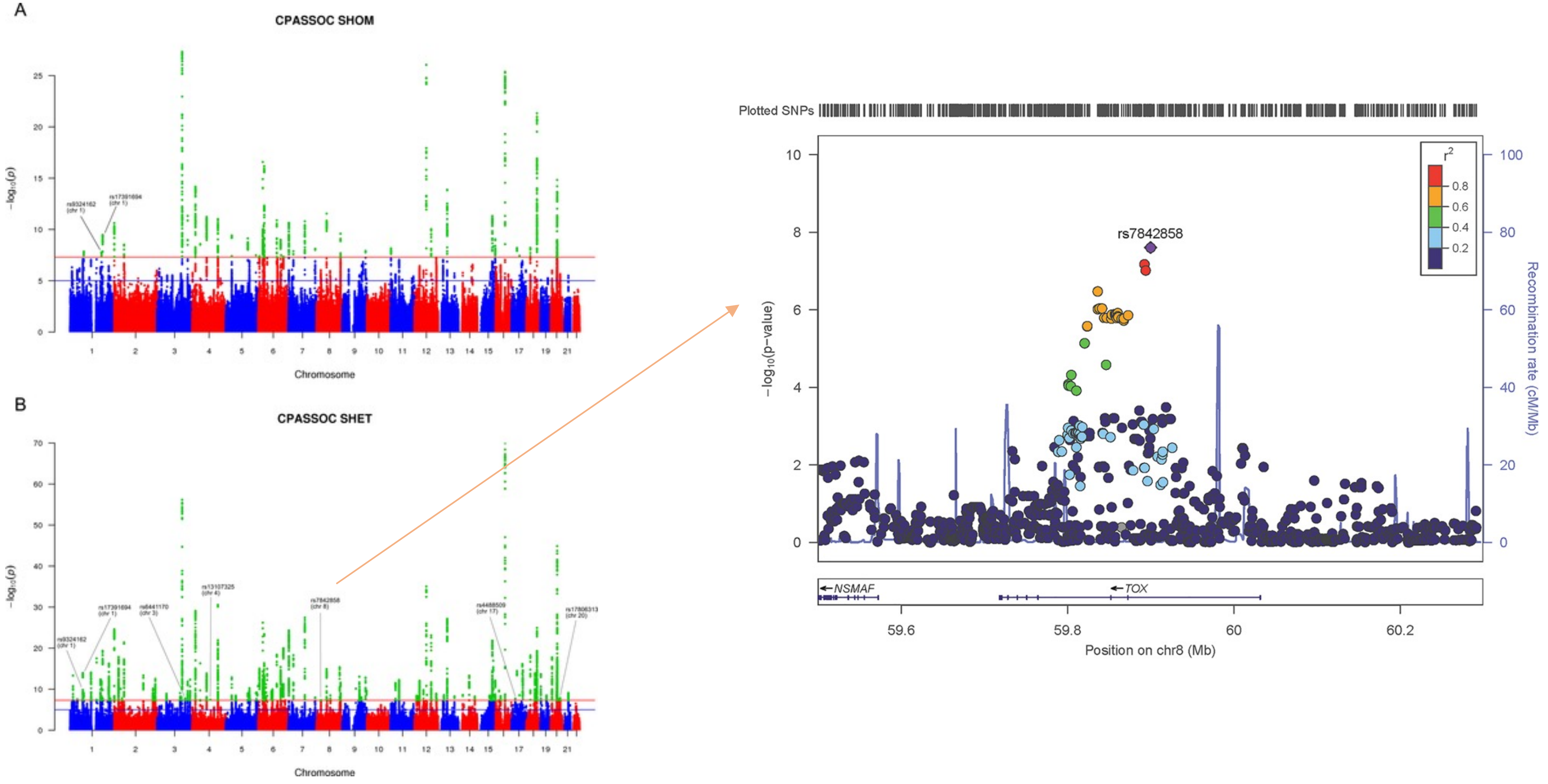
Note: CPASSOC (cross-phenotype association), GIANT (genetic investigation of anthropometric traits), BMI (body mass index), WHRadjBMI (waist-to-hip ratio adjusted for body mass index)

[a] CPASSOC was applied to meta-analyze male and female data for each of the three traits.

[b] The result of conventional meta-analyses of discovery phase data for each of the three traits.

Park H, Li X, Song YE, He KY, Zhu X (2016) Multivariate Analysis of Anthropometric Traits Using Summary Statistics of Genome-Wide Association Studies from GIANT Consortium. PLOS ONE 11(10): e0163912. https://doi.org/10.1371/journal.pone.0163912

Manhattan plots of CPASSOC for combining three gender specific traits

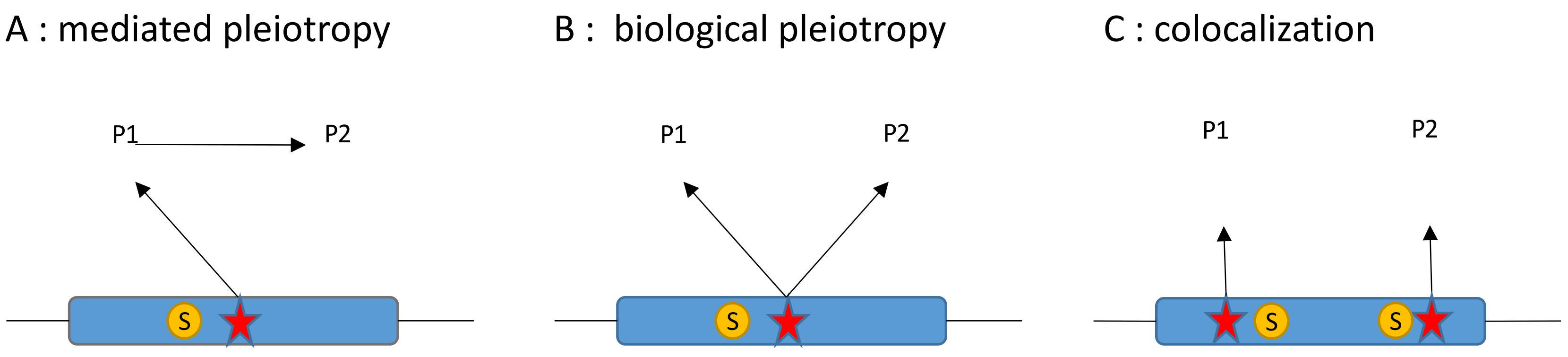


Park H, Li X, Song YE, He KY, Zhu X (2016) Multivariate Analysis of Anthropometric Traits Using Summary Statistics of Genome-Wide Association Studies from GIANT Consortium. PLOS ONE 11(10): e0163912. https://doi.org/10.1371/journal.pone.0163912

St. Cloud State University

how to understand the mechanism that genetic variants influence pairs of traits?

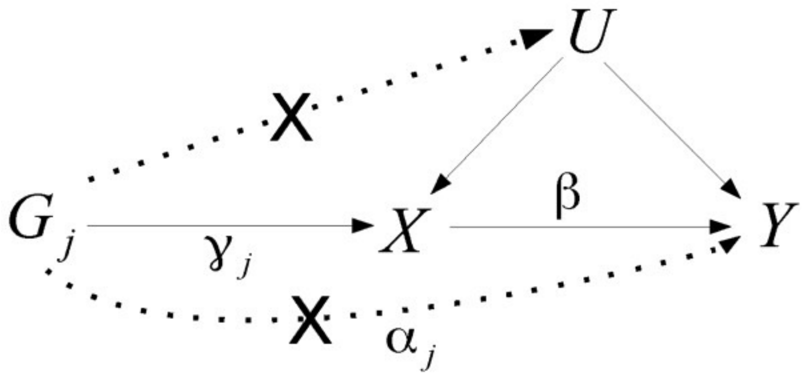
Different types of pleiotropy can underlie a CP association



- **A** l Mediated pleiotropy: the causal variant affects P_1, which lies on the causal path to P_2
- **B** l Biological (Horizontal) pleiotropy: the causal variant affects both phenotypes.
- **C**l Colocalization: two causal variants in strong LD that affect different phenotypes.

Mendelian randomization

- MR is an approach to infer causality of an exposure for a complex disease outcome
- MR uses genetic variants as instrumental variables (IVs) that are robustly associated with the exposure and tests whether the exposure has a causal role in the etiology of a disease
- If the genetic variants have pleiotropic effects on the outcome, these causal estimates will be biased

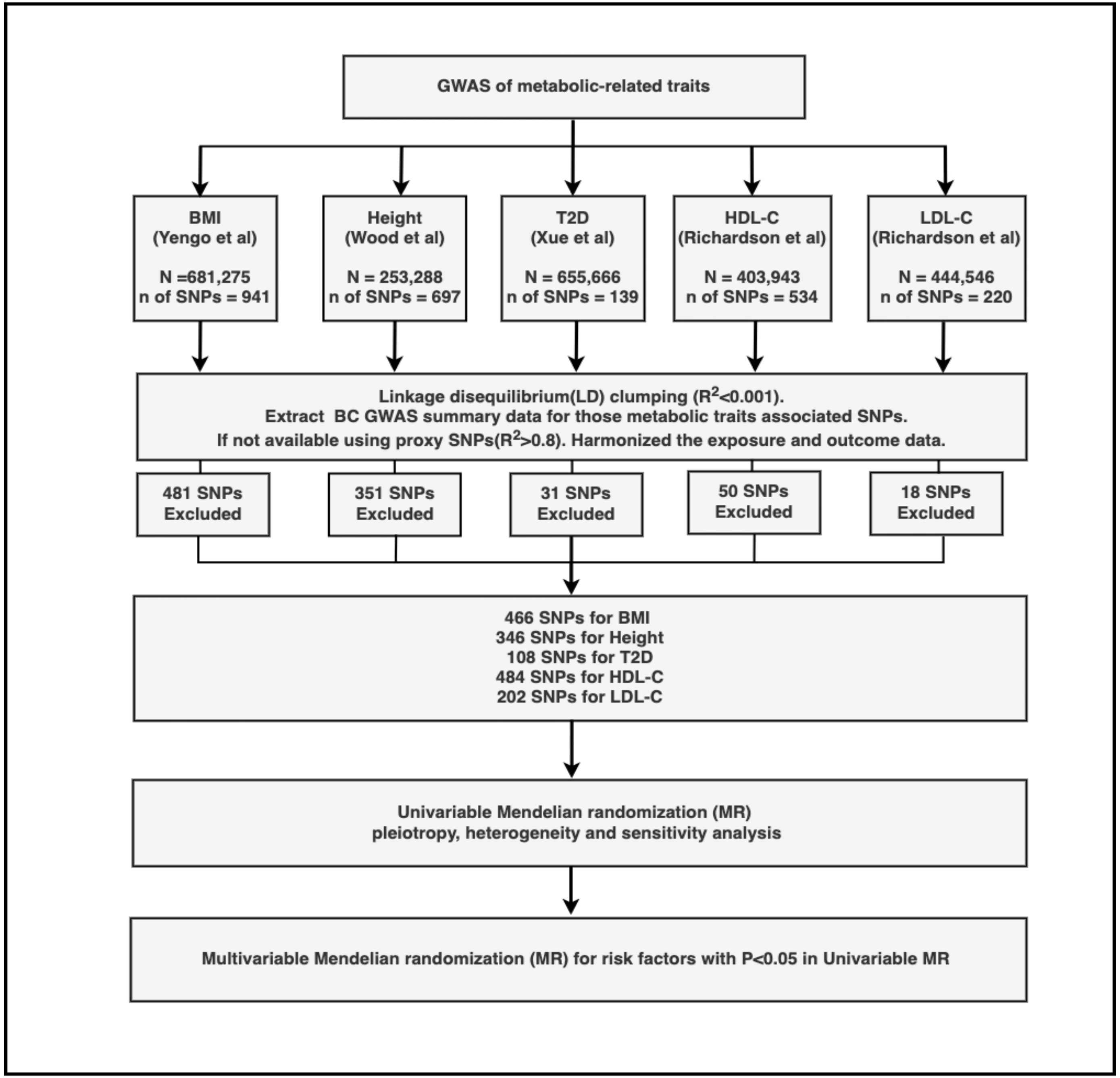


IV1: The genetic variant is independent of confounders U;
IV2: The genetic variant is associated with the exposure X;
IV3: The genetic variant is independent of the outcome Y conditional on the exposure X and confounders U.

ONLY genetic variants that manifest mediated pleiotropy for both exposure and outcome are valid IVs in MR analysis

Mendelian randomization analysis revealed potential metabolic causal factors for breast cancer

- Breast cancer (BC) is the most common invasive cancer and the second leading cause of cancer death in women.
- In this study, we sought to use human genetics to disentangle which of the five established metabolic risk factors account for a causal relationship with BC risk.



GWAS of metabolic-related traits

BMI
(Yengo et al)
N =681,275
n of SNPs = 941

Height
(Wood et al)
N = 253,288
n of SNPs = 697

T2D
(Xue et al)
N = 655,666
n of SNPs = 139

HDL-C
(Richardson et al)
N = 403,943
n of SNPs = 534

LDL-C
(Richardson et al)
N = 444,546
n of SNPs = 220

Linkage disequilibrium(LD) clumping (R^2<0.001).
Extract BC GWAS summary data for those metabolic traits associated SNPs.
If not available using proxy SNPs(R^2>0.8). Harmonized the exposure and outcome data.

481 SNPs Excluded

351 SNPs Excluded

31 SNPs Excluded

50 SNPs Excluded

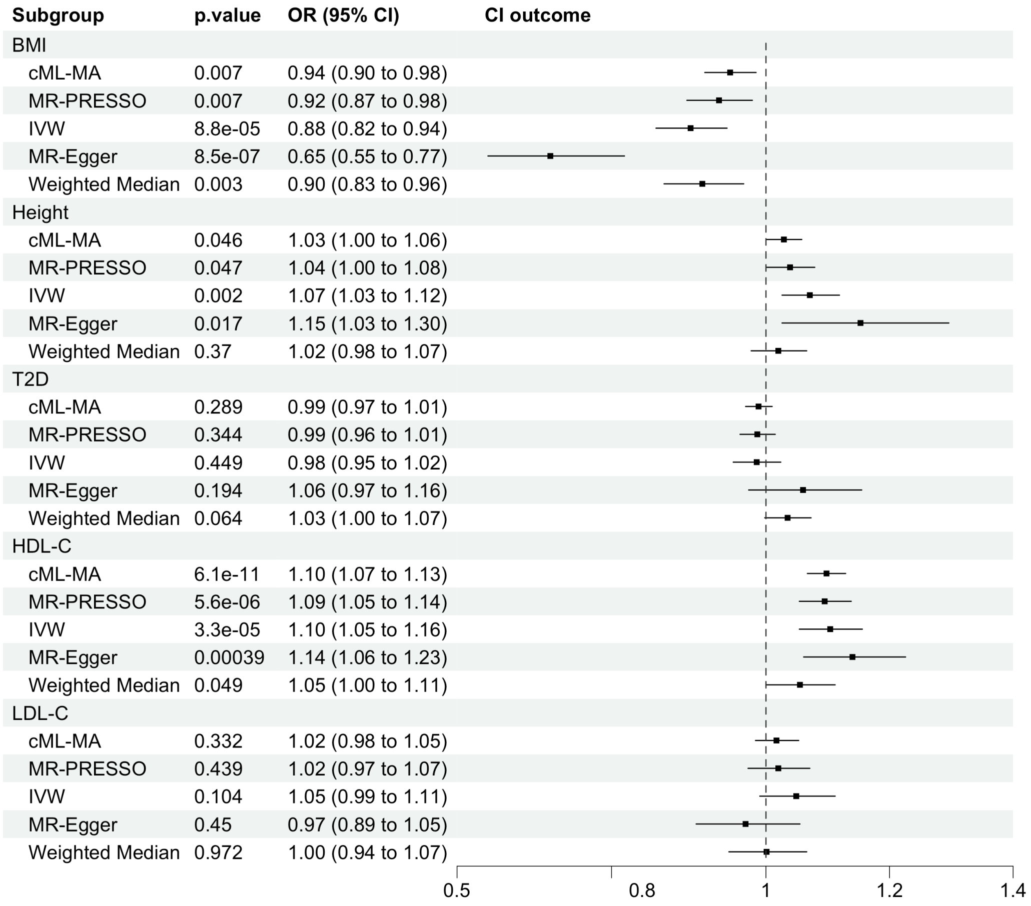
18 SNPs Excluded

466 SNPs for BMI
346 SNPs for Height
108 SNPs for T2D
484 SNPs for HDL-C
202 SNPs for LDL-C

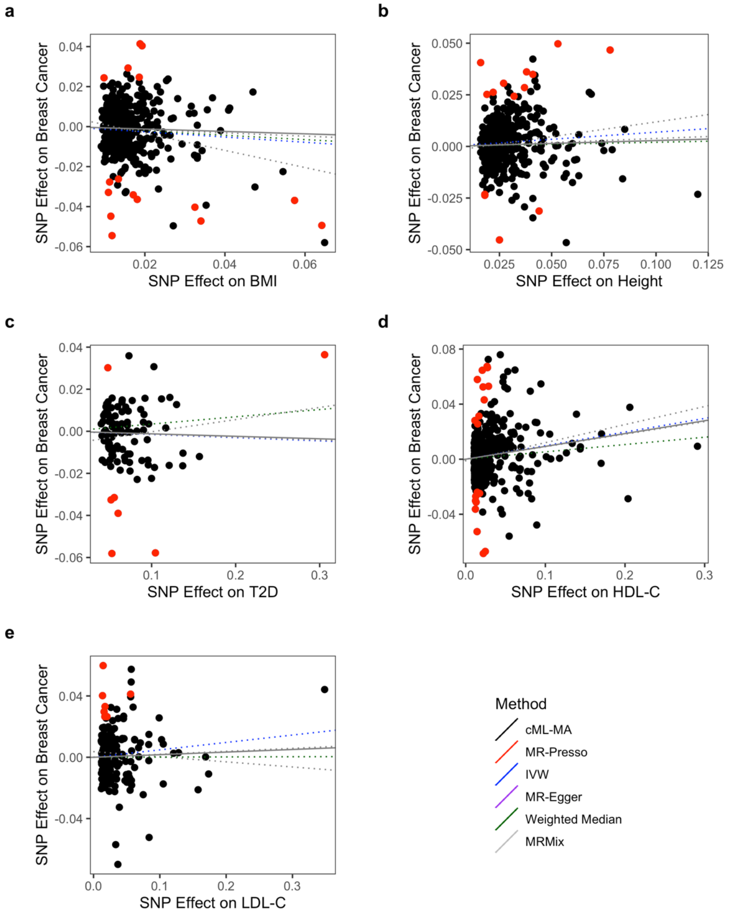
Univariable Mendelian randomization (MR)
pleiotropy, heterogeneity and sensitivity analysis

Multivariable Mendelian randomization (MR) for risk factors with P<0.05 in Univariable MR

Univariable MR analysis

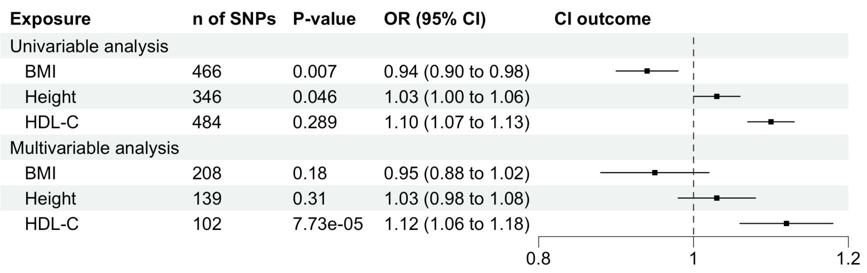


Subgroup	p.value	OR (95% CI)	CI outcome
BMI			
cML-MA	0.007	0.94 (0.90 to 0.98)	
MR-PRESSO	0.007	0.92 (0.87 to 0.98)	
IVW	8.8e-05	0.88 (0.82 to 0.94)	
MR-Egger	8.5e-07	0.65 (0.55 to 0.77)	
Weighted Median	0.003	0.90 (0.83 to 0.96)	
Height			
cML-MA	0.046	1.03 (1.00 to 1.06)	
MR-PRESSO	0.047	1.04 (1.00 to 1.08)	
IVW	0.002	1.07 (1.03 to 1.12)	
MR-Egger	0.017	1.15 (1.03 to 1.30)	
Weighted Median	0.37	1.02 (0.98 to 1.07)	
T2D			
cML-MA	0.289	0.99 (0.97 to 1.01)	
MR-PRESSO	0.344	0.99 (0.96 to 1.01)	
IVW	0.449	0.98 (0.95 to 1.02)	
MR-Egger	0.194	1.06 (0.97 to 1.16)	
Weighted Median	0.064	1.03 (1.00 to 1.07)	
HDL-C			
cML-MA	6.1e-11	1.10 (1.07 to 1.13)	
MR-PRESSO	5.6e-06	1.09 (1.05 to 1.14)	
IVW	3.3e-05	1.10 (1.05 to 1.16)	
MR-Egger	0.00039	1.14 (1.06 to 1.23)	
Weighted Median	0.049	1.05 (1.00 to 1.11)	
LDL-C			
cML-MA	0.332	1.02 (0.98 to 1.05)	
MR-PRESSO	0.439	1.02 (0.97 to 1.07)	
IVW	0.104	1.05 (0.99 to 1.11)	
MR-Egger	0.45	0.97 (0.89 to 1.05)	
Weighted Median	0.972	1.00 (0.94 to 1.07)	



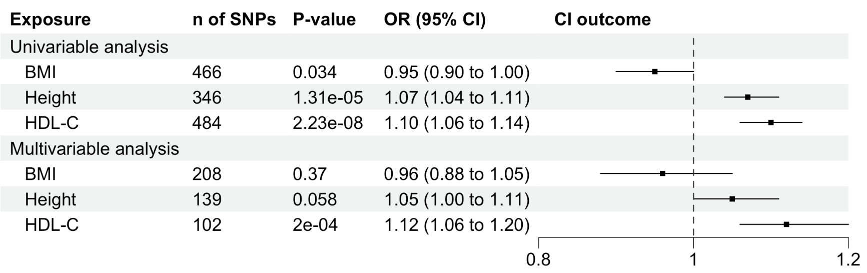
Multivariable MR analysis

a



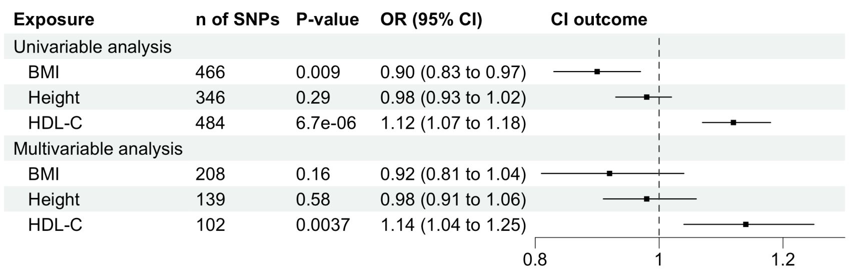
Exposure	n of SNPs	P-value	OR (95% CI)
Univariable analysis			
BMI	466	0.007	0.94 (0.90 to 0.98)
Height	346	0.046	1.03 (1.00 to 1.06)
HDL-C	484	0.289	1.10 (1.07 to 1.13)
Multivariable analysis			
BMI	208	0.18	0.95 (0.88 to 1.02)
Height	139	0.31	1.03 (0.98 to 1.08)
HDL-C	102	7.73e-05	1.12 (1.06 to 1.18)

b



Exposure	n of SNPs	P-value	OR (95% CI)
Univariable analysis			
BMI	466	0.034	0.95 (0.90 to 1.00)
Height	346	1.31e-05	1.07 (1.04 to 1.11)
HDL-C	484	2.23e-08	1.10 (1.06 to 1.14)
Multivariable analysis			
BMI	208	0.37	0.96 (0.88 to 1.05)
Height	139	0.058	1.05 (1.00 to 1.11)
HDL-C	102	2e-04	1.12 (1.06 to 1.20)

c



Exposure	n of SNPs	P-value	OR (95% CI)
Univariable analysis			
BMI	466	0.009	0.90 (0.83 to 0.97)
Height	346	0.29	0.98 (0.93 to 1.02)
HDL-C	484	6.7e-06	1.12 (1.07 to 1.18)
Multivariable analysis			
BMI	208	0.16	0.92 (0.81 to 1.04)
Height	139	0.58	0.98 (0.91 to 1.06)
HDL-C	102	0.0037	1.14 (1.04 to 1.25)

- Using univariable MR analysis, we found BMI and HDL-C causally linked to the BC risk.
- When BMI, height, and HDL-C were taken into account in multivariable MR, the relationship between BMI and height and BC risk was attenuated. Only HDL-C retained a robust effect with BC risk, indicating that HDL-C was responsible for the the genetic association between BMI, and height with BC risk in the univariable study.

St. Cloud State University

Summary

- GWAS study is a powerful tool to detect associations between genetic variants and traits in samples from populations.
- Cross phenotype association will increase statistical power when analyzing traits share common variants or common genetic pathways, which may reflect the relevance of pleiotropy.
- MR analysis techniques can be employed to determine the causal relationship between risk factors and trait.
- Our findings demonstrated that HDL-C was critical in facilitating the causal effects of breast cancer risk.

Questions?