Post-GWAS Analysis II

PEILIN JIA

BEIJING INSTITUTE OF GENOMICS

Outline

Association ←→ causal

Causal inference

- Colocalization analysis of GWAS data
- Mendelian randomization
- Fine-mapping
- Convolutional Neural Network in predicting functional impact of genetic variants

Transcriptome-wide association study (TWAS)

Epigenome-wide association study (EWAS)

PheWAS

Risk prediction: Polygenic Risk Score (PRS)

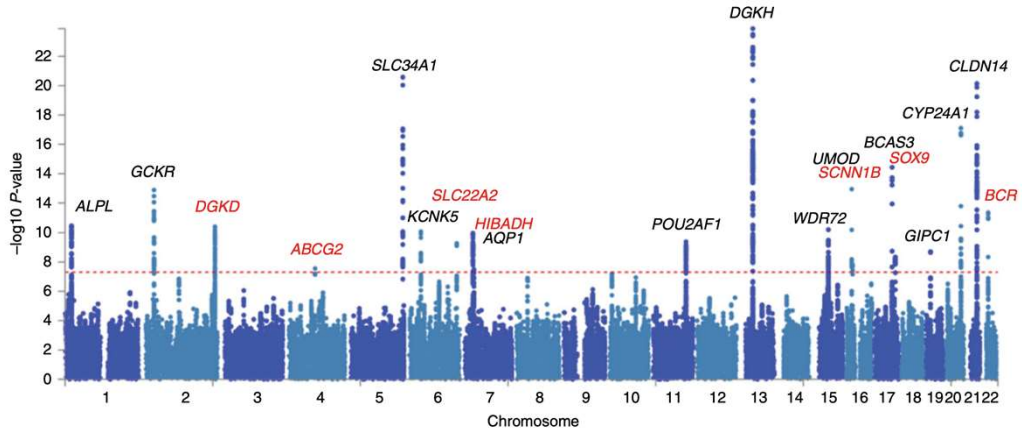
GWAS does not provide causal mechanisms

GWAS provides regions associated with disease risk

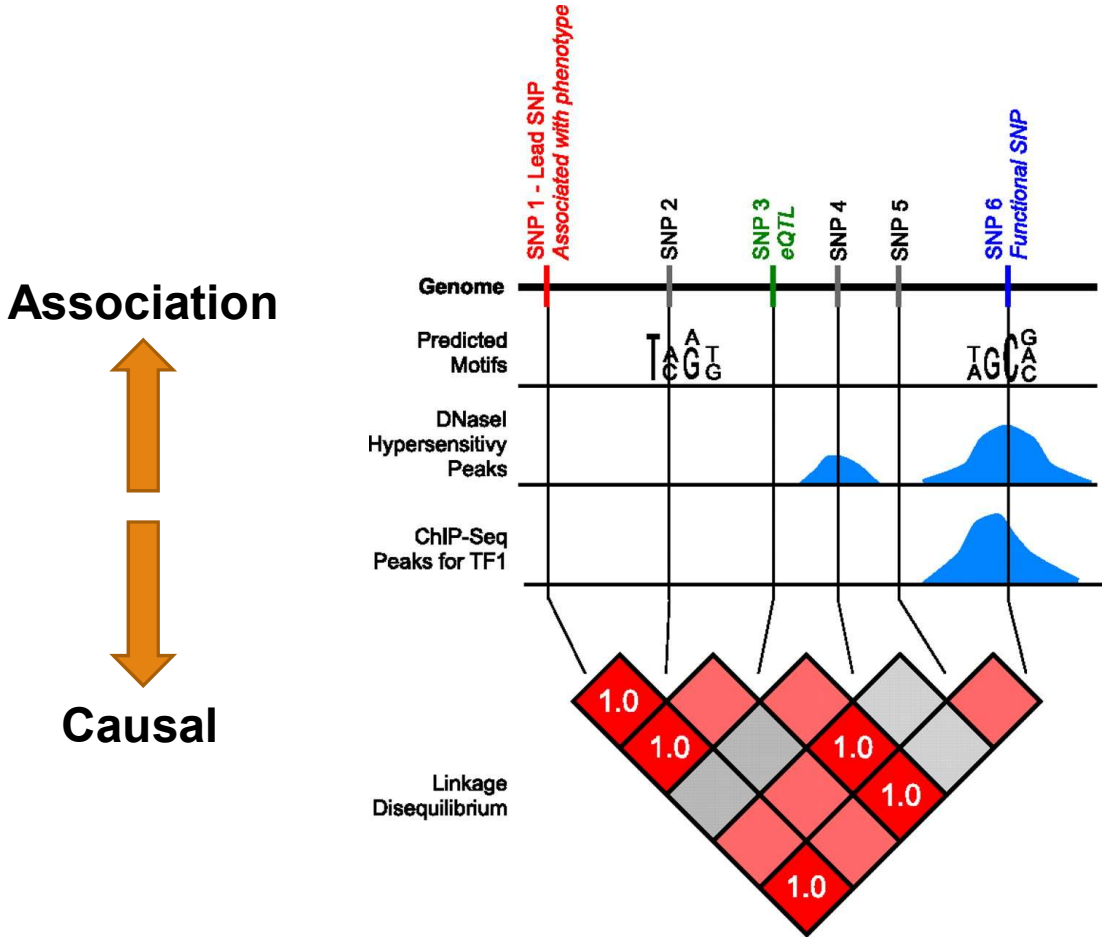
What are the mechanisms driving disease risk?

Most associations are non-coding

- SNPs with regulatory function are strong candidates

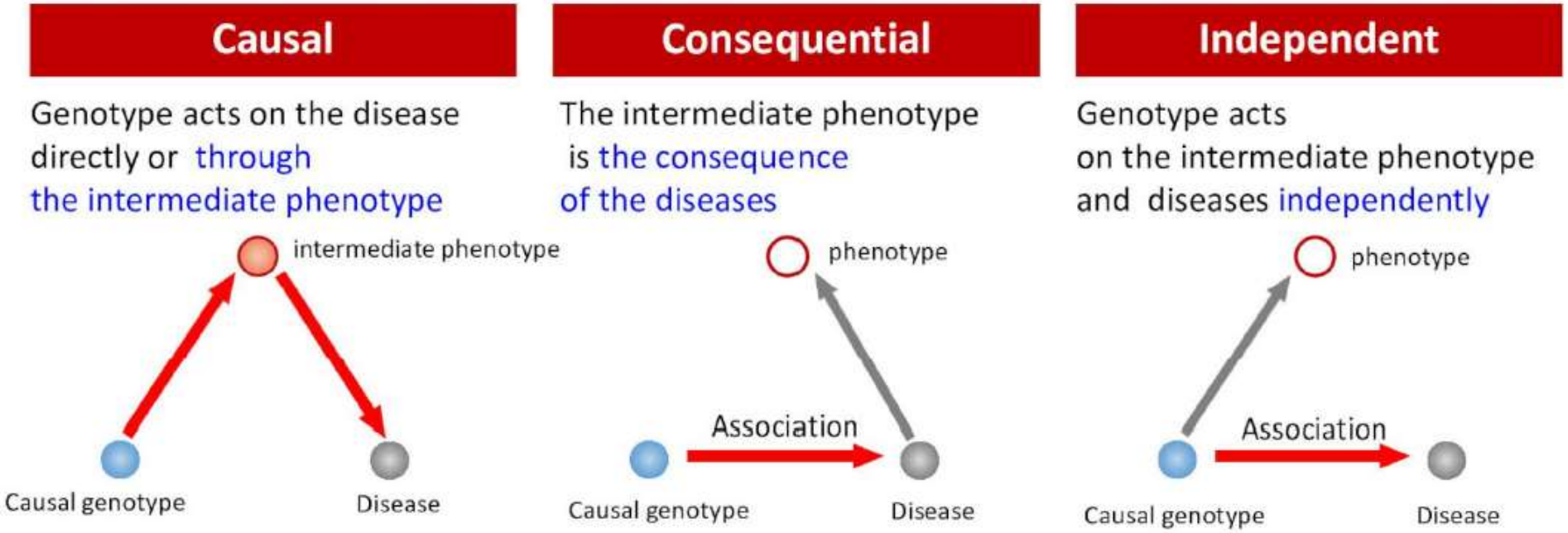


Schematic Overview of the Functional SNP Approach



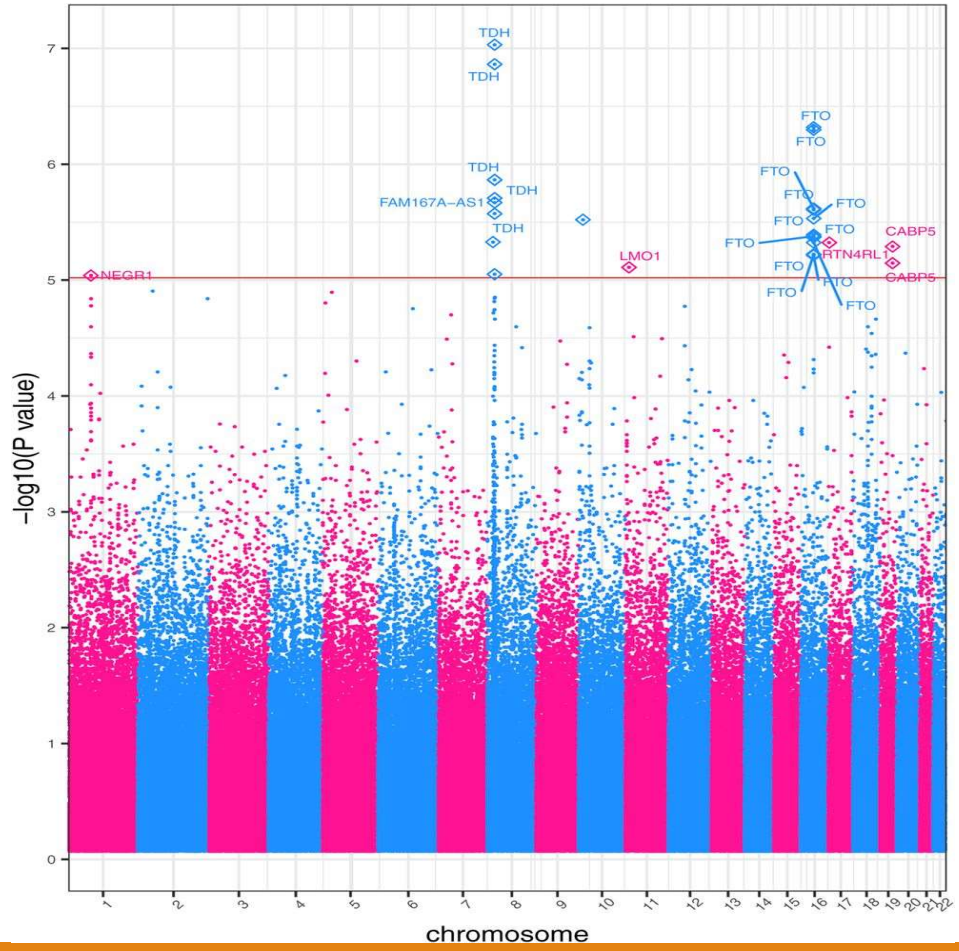
Marc A. Schaub et al. Genome Res. 2012;22:1748-1759

Possible models



Intermediate phenotypes involve gene expression, protein expression and epigenetic effects and others

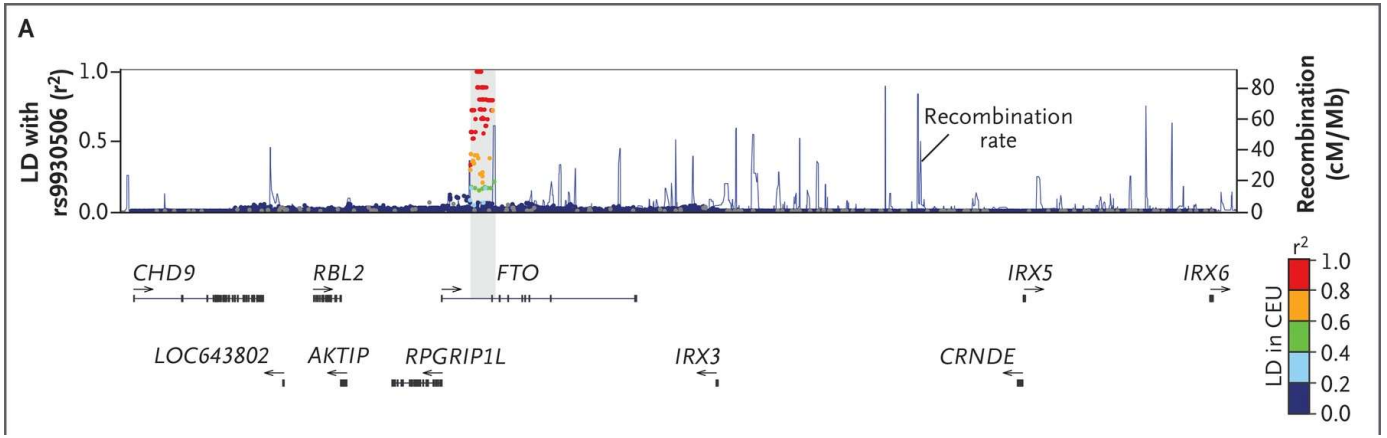
Example: *FTO* with obesity

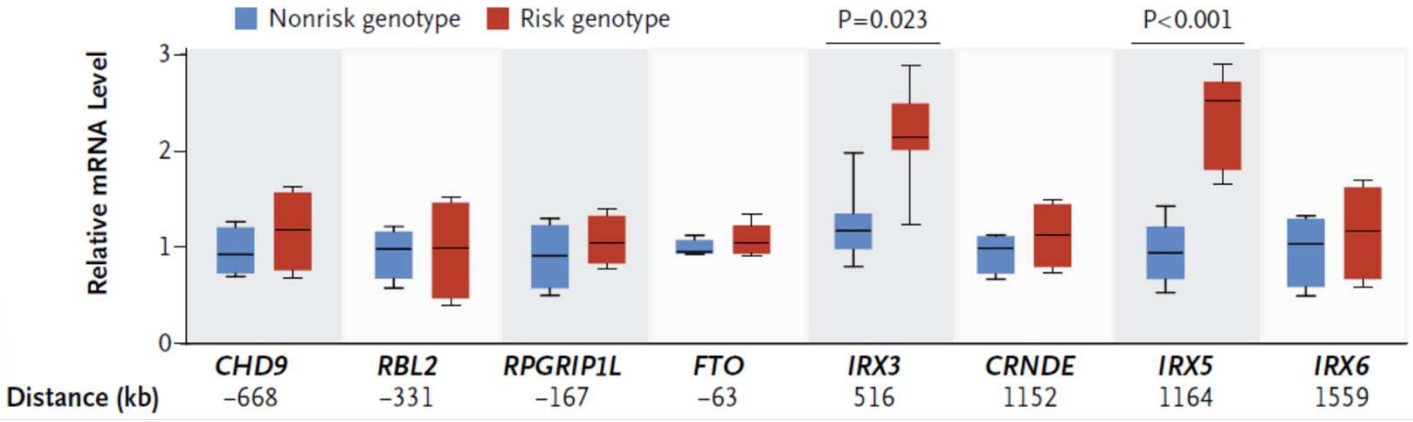


Example: *FTO* with obesity

Obesity-associated locus *FTO*

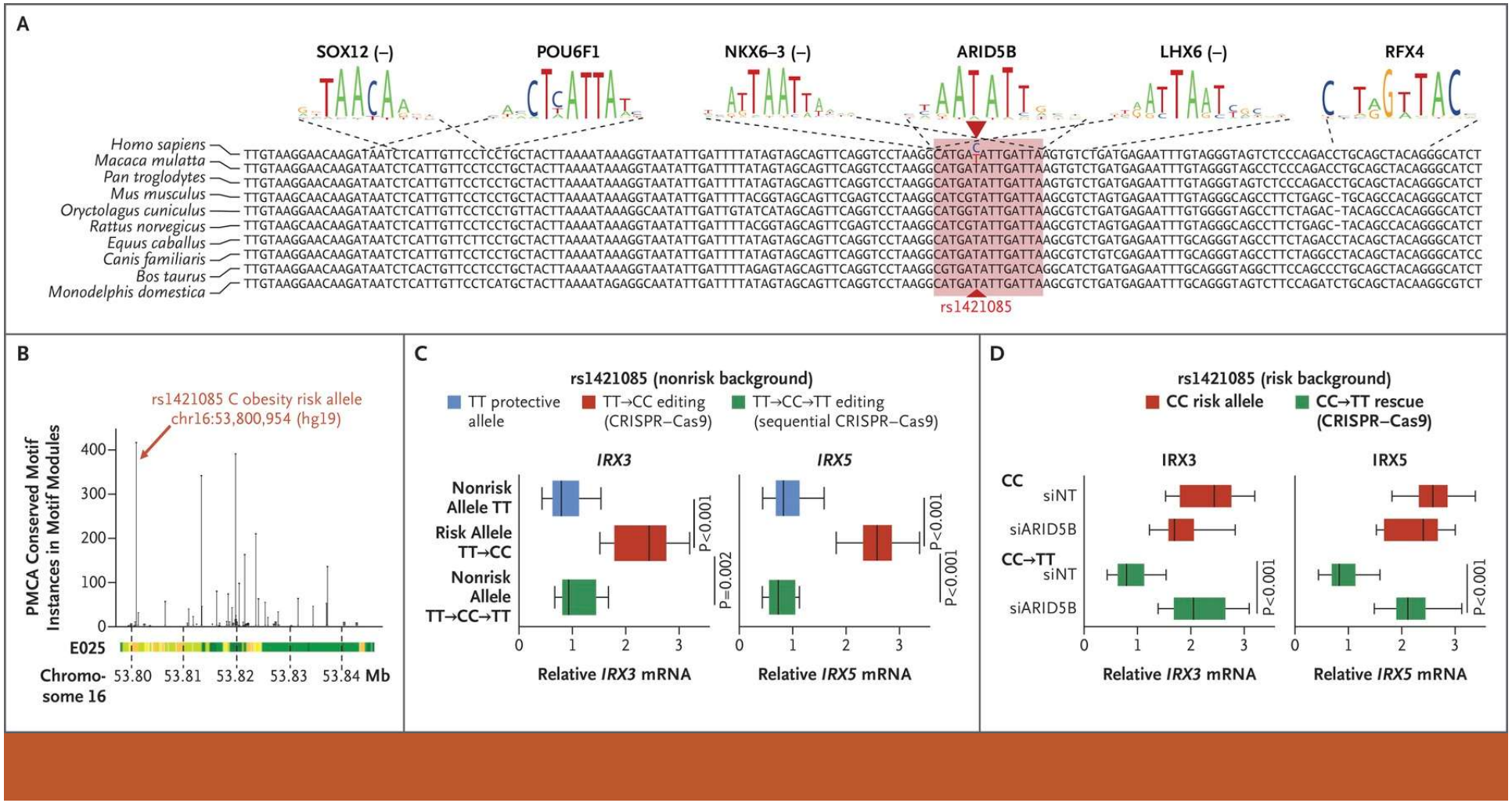
- Eight candidate genes





Analysis of the chromatin state of *FTO* across 127 human cell types revealed that it harbors an enhancer that is specific to pre-adipocyte cells.

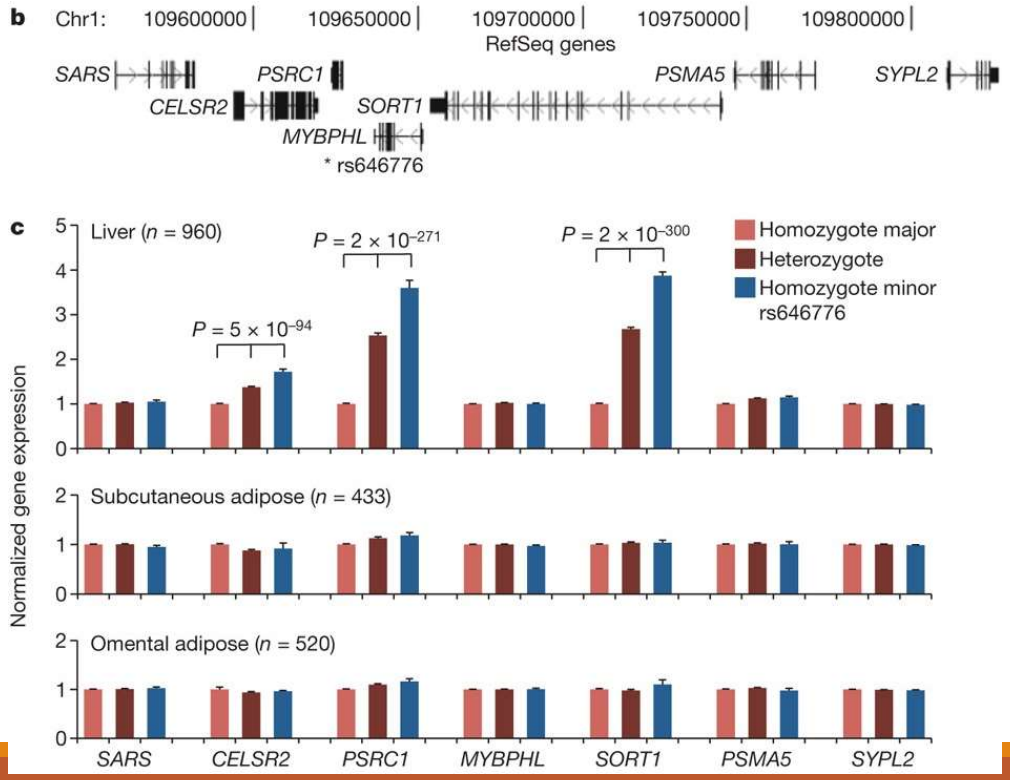
The rs1421085 T-to-C disrupts a conserved motif for the ARID5B repressor, which leads to derepression of a potent preadipocyte enhancer and a doubling of IRX3 and IRX5 expression during early adipocyte differentiation.



Example: *SORT1*

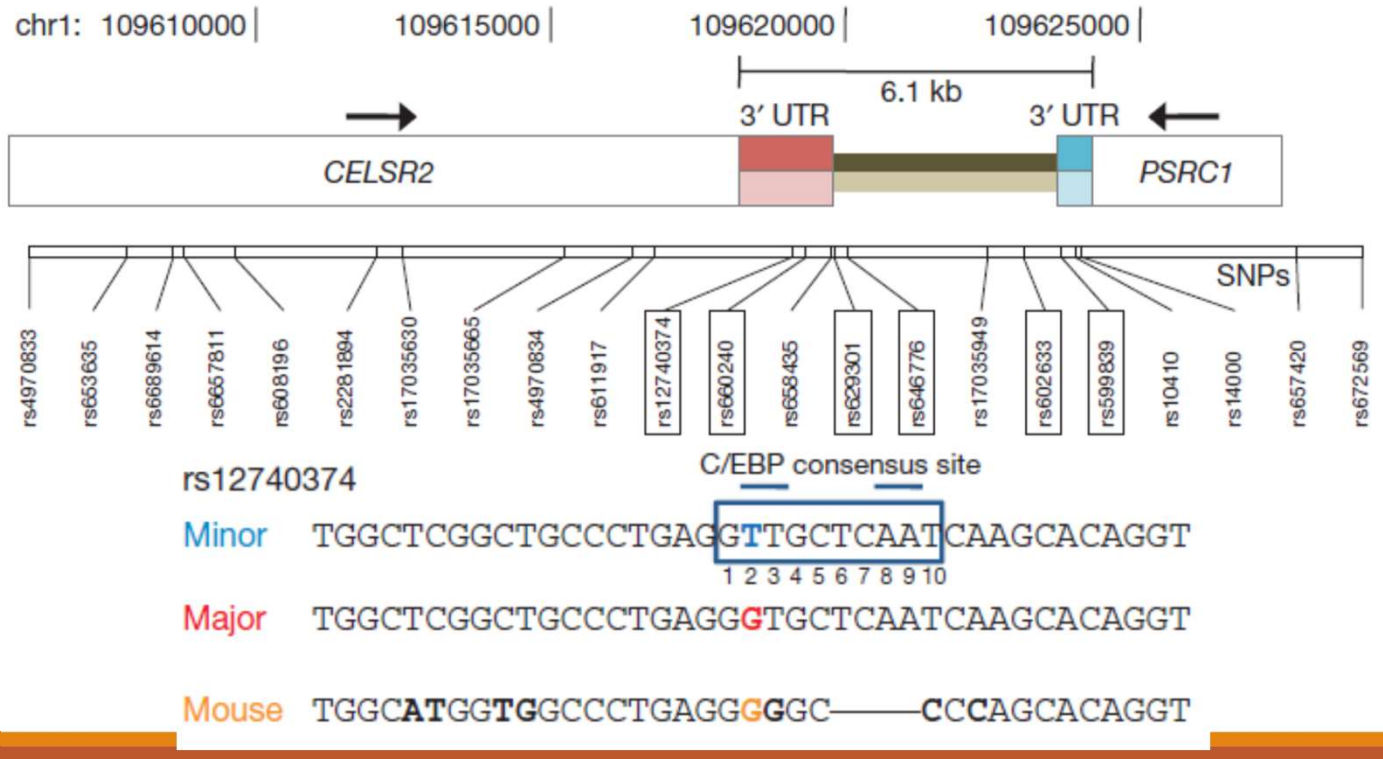
GWAS found a locus on chromosome 1p13 strongly associated with both plasma LDL-C and myocardial infarction.

Three genes showed tissue-specific expression changes among genotype groups.

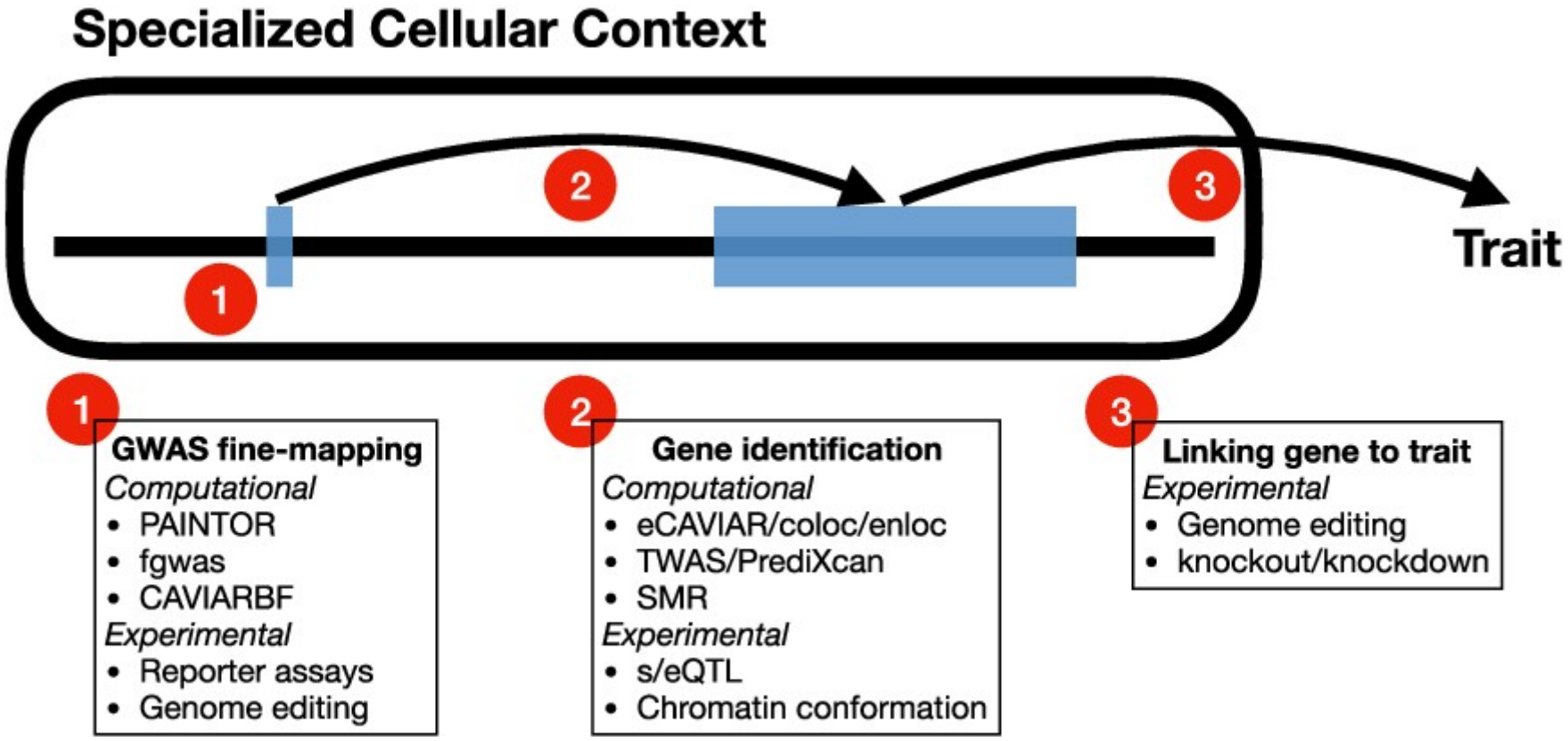


Example: *SORT1* (cont'd)

rs12740374, creates a C/EBP (CCAAT/enhancer binding protein) transcription factor binding site and alters the hepatic expression of the *SORT1* gene



Using Specialized Cell Types to Improve GWAS Follow-up Analysis



Outline

Association ←→ causal

Causal inference

- **Colocalization analysis of GWAS data**
- **Mendelian randomization**
- **Fine-mapping**
- **Convolutional Neural Network in predicting functional impact of genetic variants**

Transcriptome-wide association study (TWAS)

Epigenome-wide association study (EWAS)

PheWAS

Risk prediction: Polygenic Risk Score (PRS)

Integration of GWAS variants and eQTLs

Colocalization of pairs of association signals

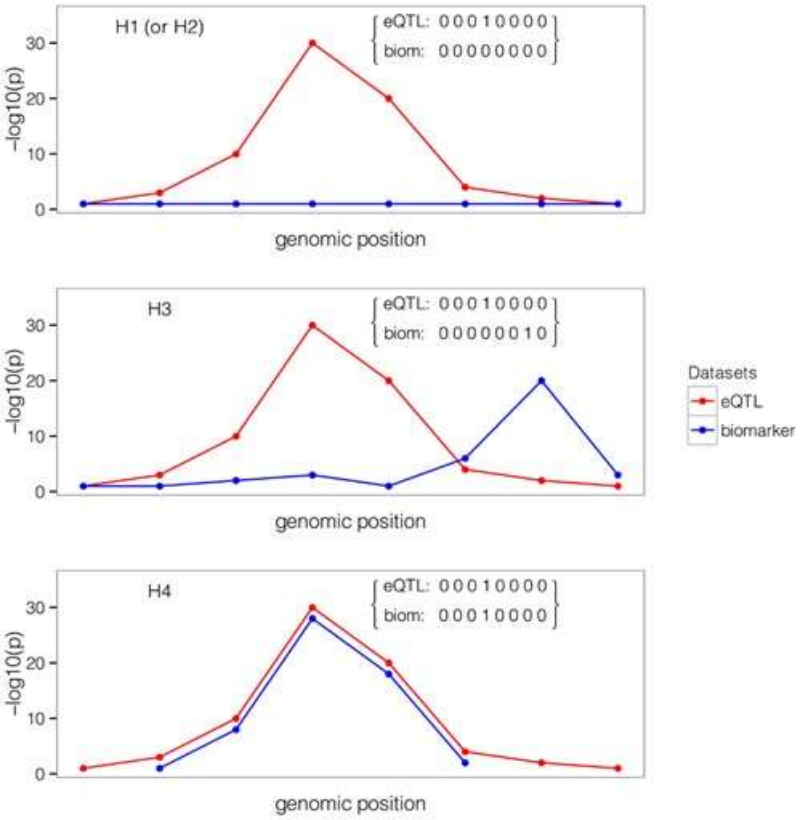
H1 is the hypothesis that there is only an eQTL signal at a locus

H2 is the hypothesis that there is only a GWAS signal at a locus.

H3 is the hypothesis that there are two independent eQTL and GWAS signals in linkage.

H4 is the strong hypothesis that the same SNP (not just the locus) is responsible for both the GWAS and eQTL.

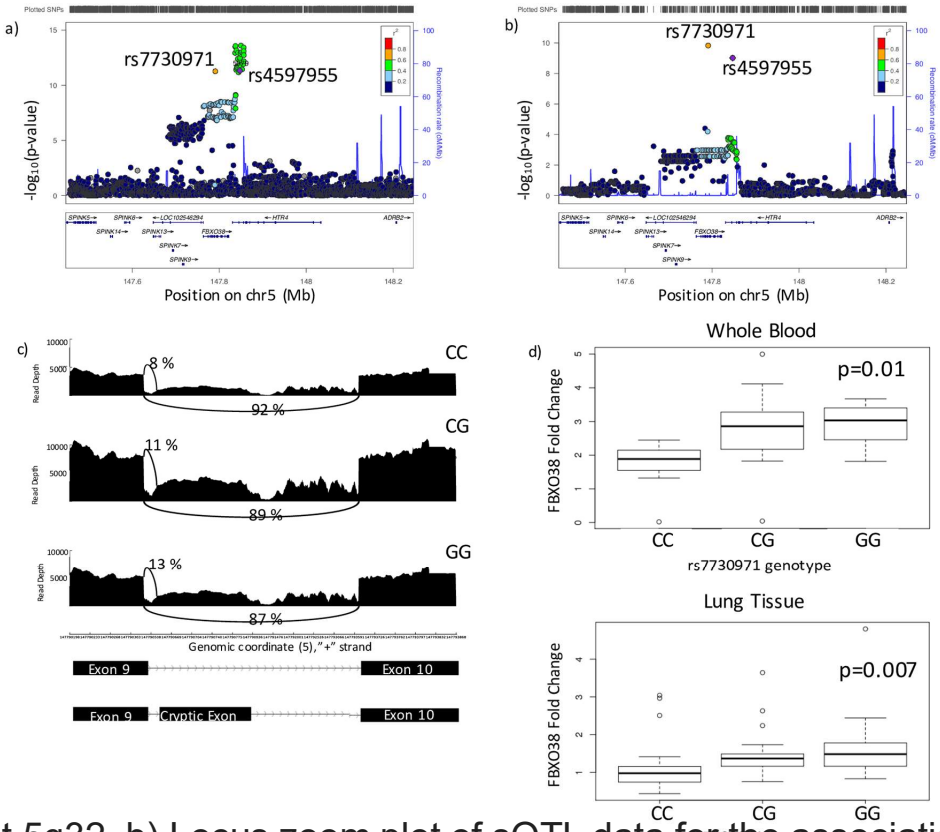
Bayesian analysis evaluate each H relative to the other four and generates a confidence level for the most likely one.



Giambartolomei et al (2014) *PLoS Genetics* **10(5)**: e1004383

Colocalization: GWAS + molQTL

Analysis of genetically driven alternative splicing identifies FBXO38 as a novel COPD susceptibility gene



a) Locus zoom plot of the GWAS association at 5q32. b) Locus zoom plot of sQTL data for the association between *FBXO38* splicing with genotype. c) Visualization of the FBXO38 splice site associated with rs7730971 genotype. d) Boxplot of qPCR results showing the fold change of the isoform congaing the cryptic exon compared to the CC genotype in whole blood (n = 30; selected based on expression levels) and lung tissue (n = 90, selected based on genotype).

Mendelian Randomization

Strengthening causal inference within observational epidemiological data through the incorporation of the special properties of germline genetic variants

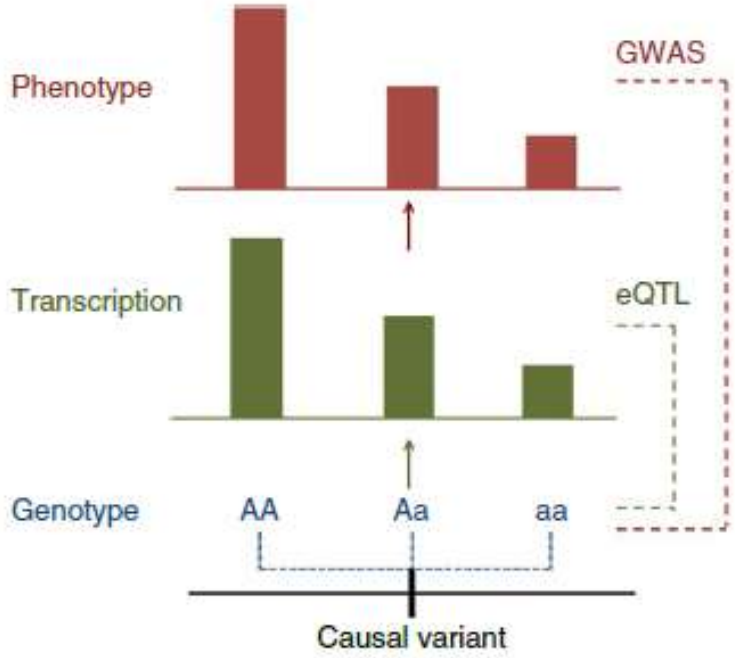
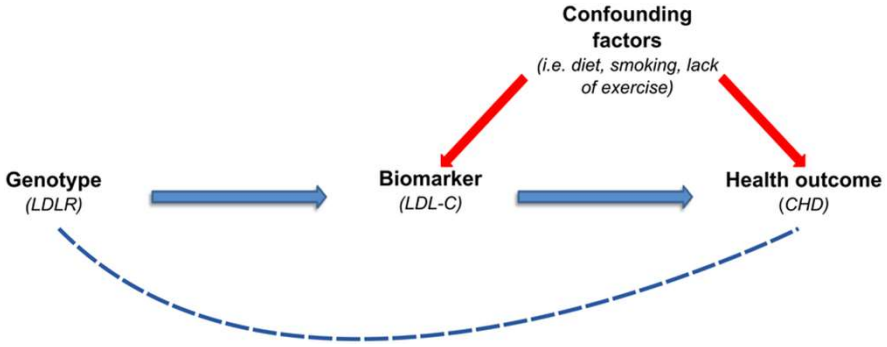


Fig 1. A model of causality where a difference in phenotype is caused by a difference in genotype mediated by gene expression (transcription)

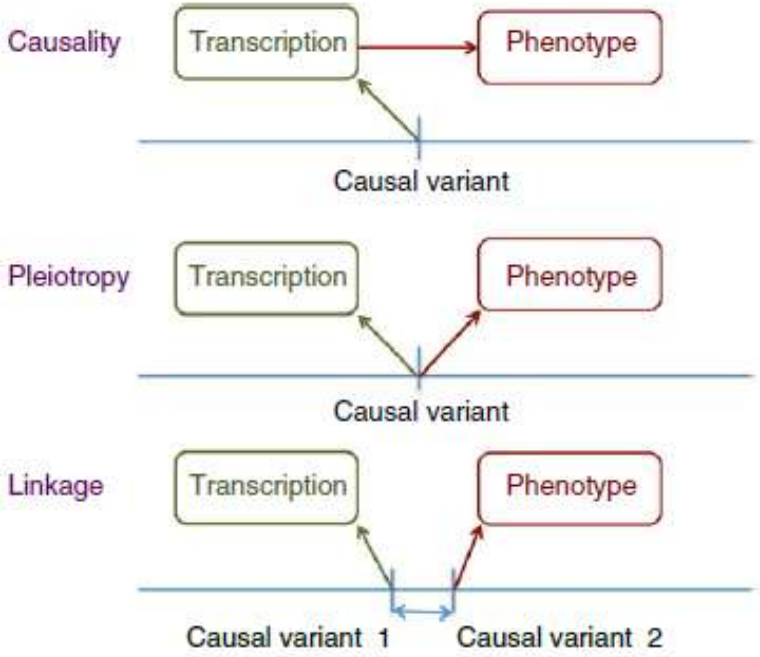


Mendelian randomization study design: if a biomarker is causal for a disease, then genetic variants which influence the levels of the biomarker should result in a higher risk of the disease.

[Zhu Z, Nature Genetics, 2016]

Mendelian Randomization

Strengthening causal inference within observational epidemiological data through the incorporation of the special properties of germline genetic variants



Three possible explanations for an observed association between a trait and gene expression through genotypes)

[Zhu Z, Nature Genetics, 2016]

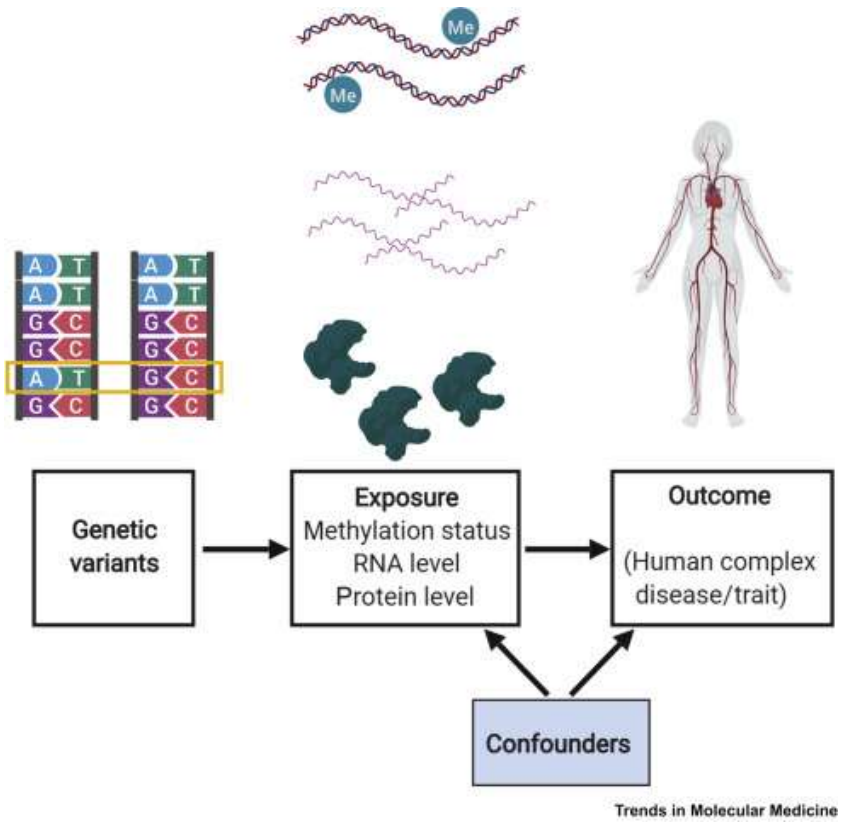
Mendelian Randomization

The laws of Mendelian inheritance assign alleles at conception to individuals independently of environmental risk factors and confounders.

Three assumptions:

- The genetic variants must be sufficiently associated with the exposure of interest;
- The genetic variants should not be associated with any confounder of the risk factor–outcome relationship
- There should not be any other pathway leading from genetic variants to outcome except through the exposure of interest. Except for the first assumption, which can be tested, the other two assumptions can only be addressed by sensitivity analyses

If genetic variants are sufficiently associated with the modifiable exposure of interest [methylation (Me), RNA expression levels, or protein levels], and are not associated with the outcome via a different pathway, they can then be used as instrumental variables for the exposure.

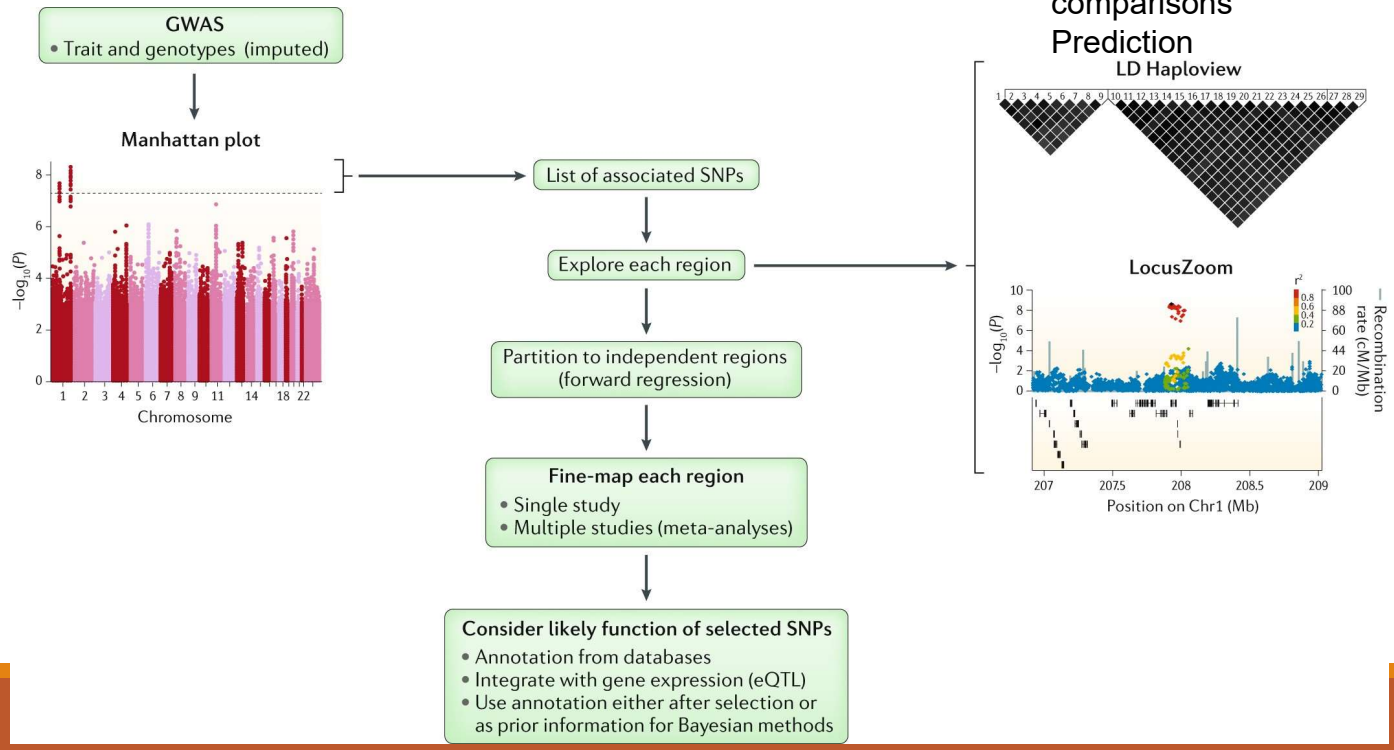


Fine-mapping

If genetic variants are sufficiently associated with the modifiable exposure of interest [in this case levels of methylation (Me), RNA expression levels, or protein levels], and are not associated with the outcome via a different pathway, they can then be used as instrumental variables for the exposure.

Fine-map

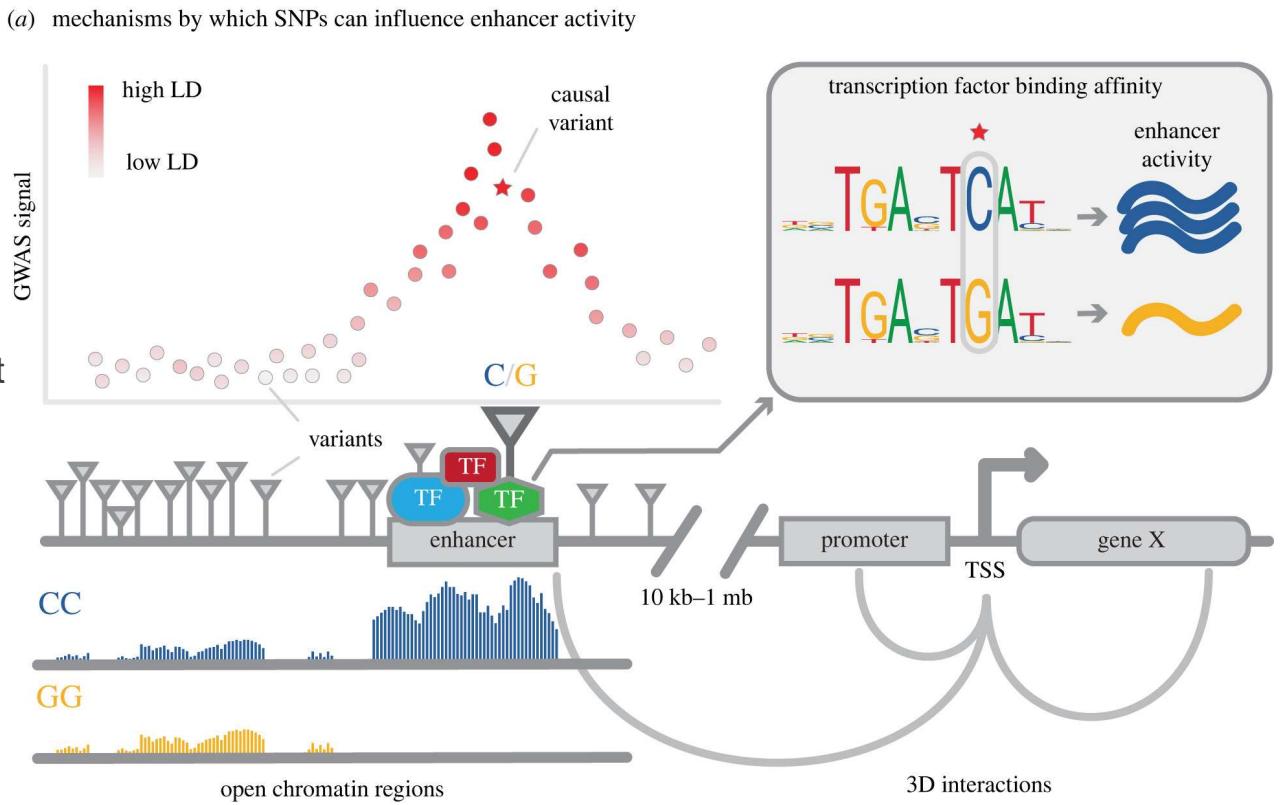
- To find causal genes
- To pinpoint variant → gene mechanisms
- To understand genetic architecture
  - Enrichment
  - Cross-population, cross-trait comparisons
  - Prediction



Fine-mapping: genetic variants

Fine-mapping from the variant perspective

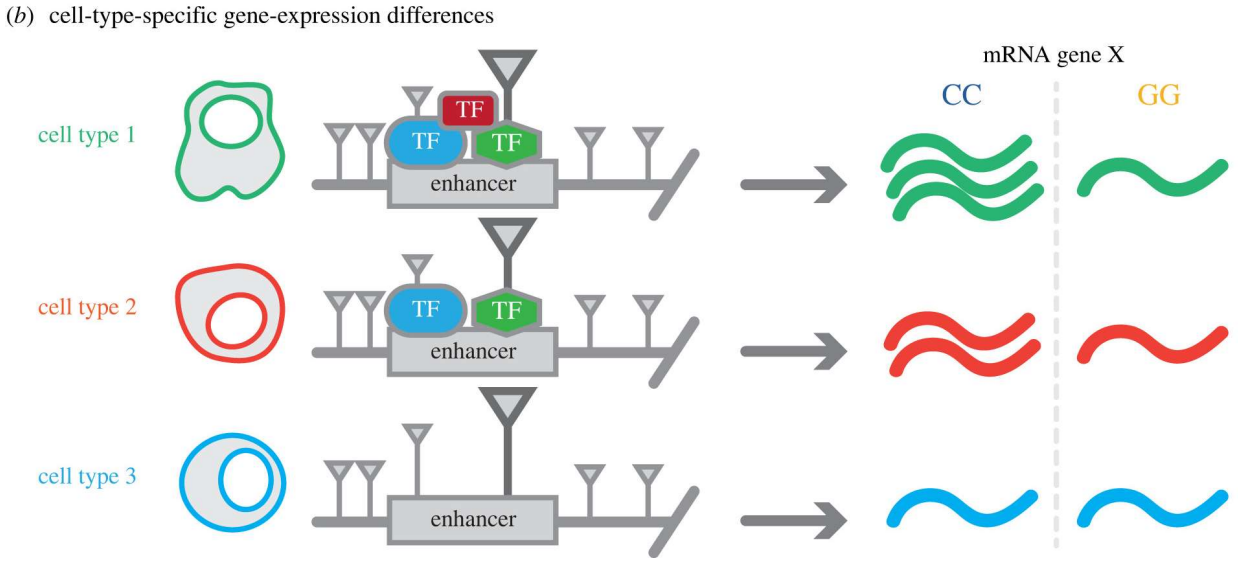
- Identifying overlap with functional elements
- Inferring allele-specific variant effects
- Identifying variants that disrupt underlying TF binding sites
- Fine-mapping by detection of regulatory region activity
- From causal variant to gene using the 3D interactome



Fine-mapping: genes

Gene prioritization using GWAS traits

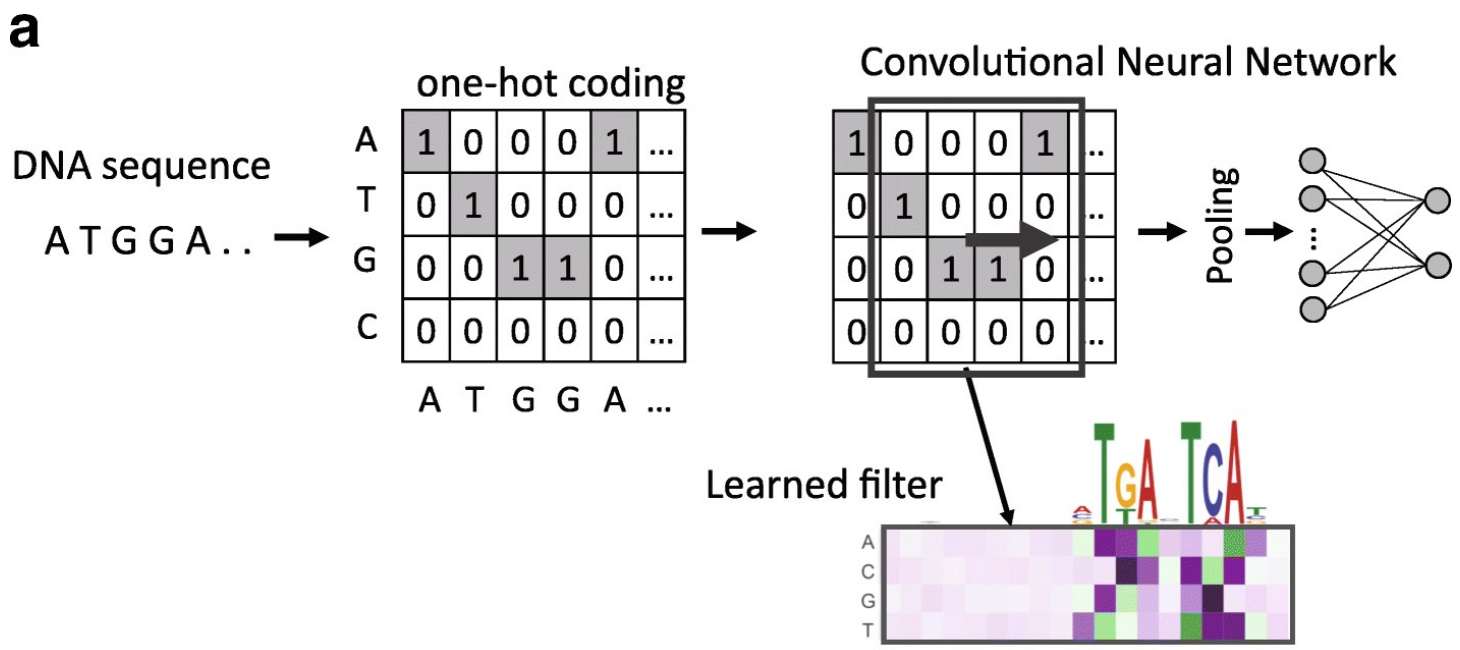
- Gene prioritization using expression quantitative trait loci
- Identifying downstream effects of GWAS loci using other QTLs
- Functional approaches to mapping genetic effects on expression
- Mapping gene–gene regulatory interactions using population data



Convolutional Neural Network (CNN) in predicting functional impact of genetic variants

One-hot coding to represent nucleotide → a DNA sequence can be represented using 4*N 'matrix'

Use convolutional neural networks to identify functional motifs



CNN in predicting functional impact of genetic variants

Illustration of the deep convolutional neural network.

Raw input sequences are first converted to a sequence matrix and screened by convolution filters, which mark the location and intensity of desired sequence motifs. These filtered signals are then collected if they reach above some threshold level, pooled and fed into a deep neural network, where simplified signals, such as the presence and absence of motifs, are synthesized to capture higher-level concepts

The motif discovery layers apply local sequence filters and extract relatively short motifs (convolution), and higher prediction layers synthesize local patterns in deep neural architecture (representational learning4).

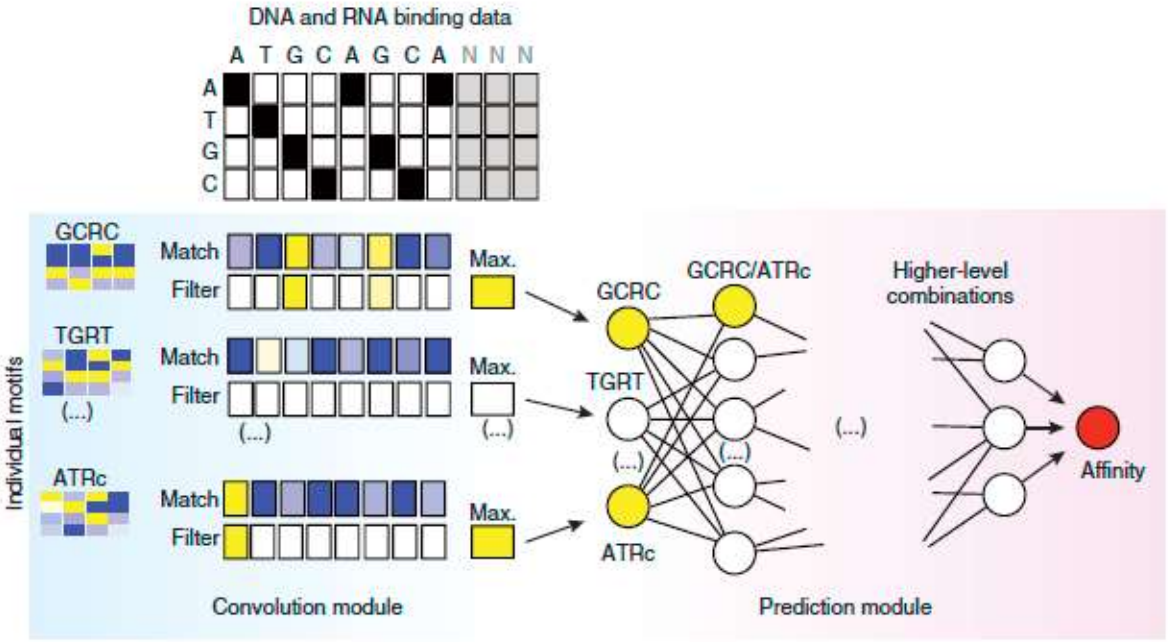


Table 2. Summary of the main variant annotation tools for non-coding DNA regions

Name	Uses	Main data sources	Advantages	Limitations
RegulomeDB	Prioritization of functional variants, using a score based on the number of elements with which the variant overlaps	ENCODE, Roadmap Epigenomics Project	Includes information from numerous functional annotation sources	The scoring system can be difficult to interpret
HaploReg	Annotation of variants in LD, located within or next to regulatory elements	ENCODE, GTEx, Roadmap Epigenomics Project	Allows the identification and mining of causal variants in LD that affect regulatory sites	Functional annotations are not updated periodically
FunciSNP	Identification and prioritization of putative regulatory SNPs	ENCODE, Roadmap Epigenomics Project	Large data queries are fast to perform	A minimum knowledge of R is needed for its use
rVarBase	Annotation of regulatory variants that are involved in transcriptional and post-transcriptional regulation	ENCODE, Roadmap Epigenomics Project	Uses annotations of numerous regulatory features, easy to use, intuitive website	Results summary can be initially confusing, i.e. a SNP can appear annotated with both strong and weak transcription
FunSeq2	Prioritization of cancer-associated SNVs in non-coding DNA	ENCODE	Can annotate and prioritize variants directly from BED or VCF files and the analysis can be customized	It is specifically designed to annotate cancer-associated variants but not for variants associated with other diseases
ENlight	Annotation of GWAS variants and analysing their putative effects through plot visualization	GWAS, ENCODE, GTEx	Plot system is useful to visually identify causal variants and the analysis can be customized	Functional annotations are not updated periodically
INFERNO	Characterization and prioritization of regulatory variants in different tissues	GTEx, FANTOM5, Roadmap Epigenomics Project	Prioritize variants by calculating an empirical *p*-value	Large Web queries take a long time to complete
Cepip	Prioritization of gene regulatory variants using tissue-expression data and predicted scores	GTEx, ENCODE, scores from different prediction tools	Integrates the effect of multiple chromatin states to identify and prioritize functional regulatory variants	A minimum knowledge of the command line is needed for its installation and use
GEMINI	Annotation of non-coding variants by integrating chromatin information for different cell types	ENCODE	Incorporates a workflow that automatically annotates variants from VCF or pedigree files	Requires command line use and lacks regulatory features in comparison with some other annotation tools

Outline

Association ←→ causal

Causal inference

- Colocalization analysis of GWAS data
- Mendelian randomization
- Fine-mapping
- Convolutional Neural Network in predicting functional impact of genetic variants

Transcriptome-wide association study (TWAS)

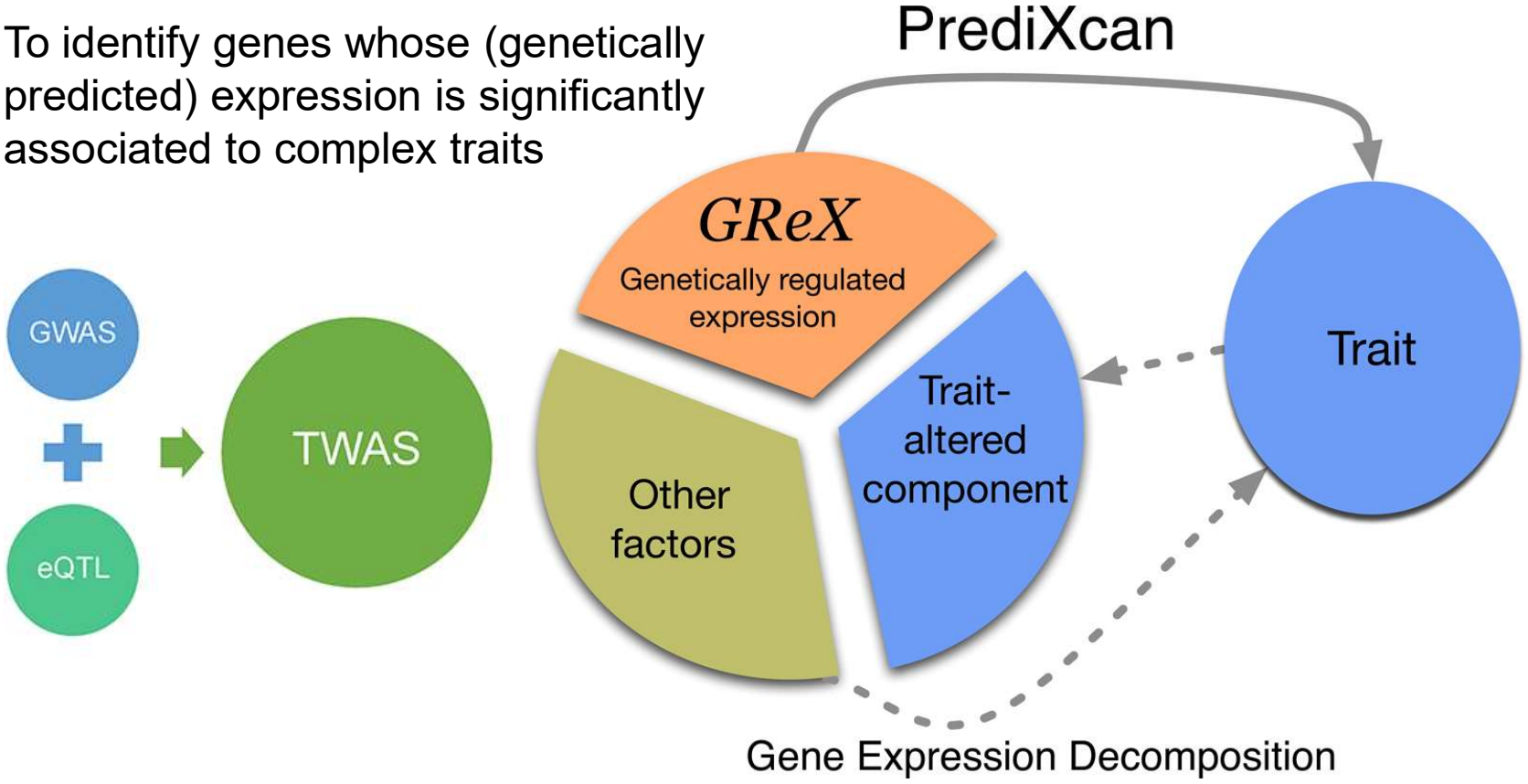
Epigenome-wide association study (EWAS)

PheWAS

Risk prediction: Polygenic Risk Score (PRS)

Transcriptome-wide association study (TWAS)

To identify genes whose (genetically predicted) expression is significantly associated to complex traits



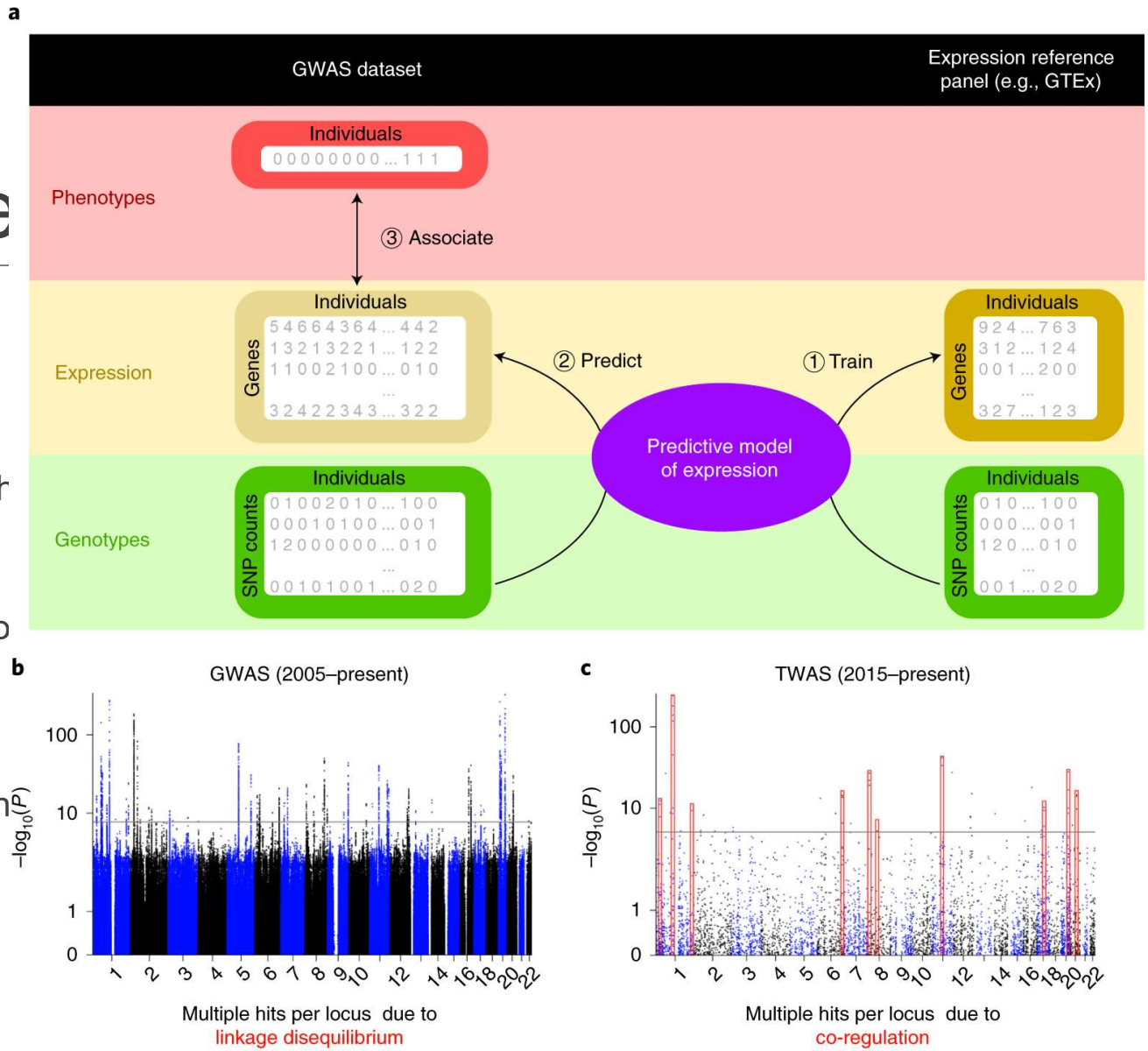
An overvie

TWAS involves:

(i) training a predictive model of expression from genotype on a reference panel such as GTEx;

(ii) using this model to predict expression fo individuals in the GWAS cohort; and

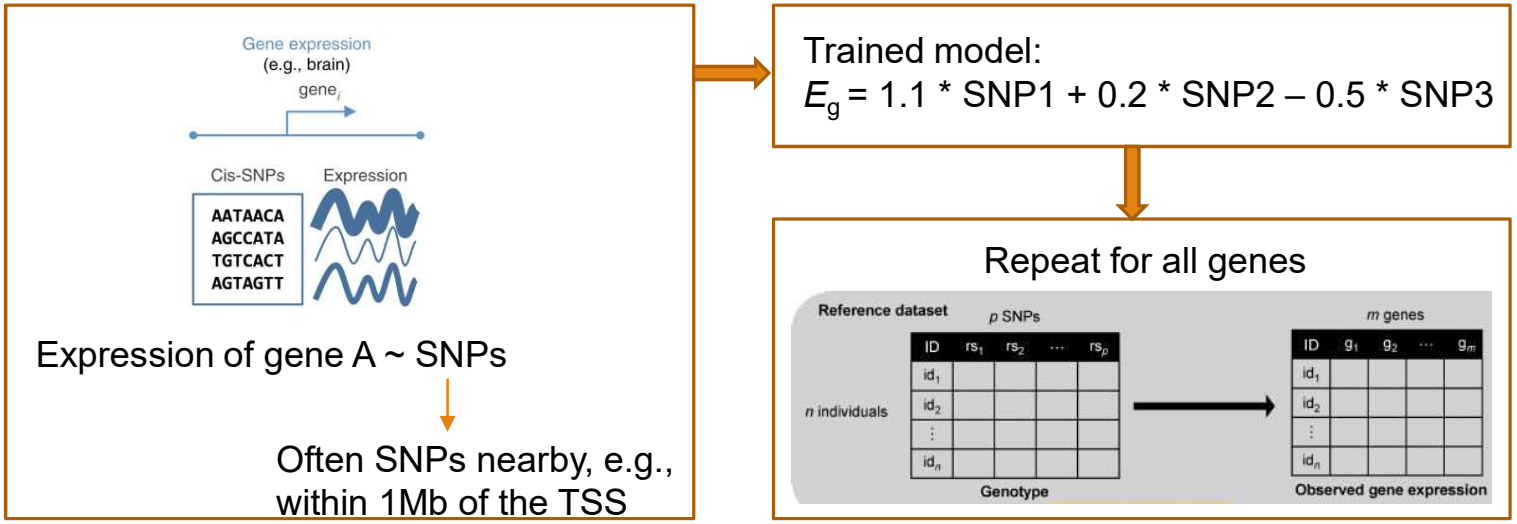
(iii) associating this predicted expression with the trait.



TWAS method

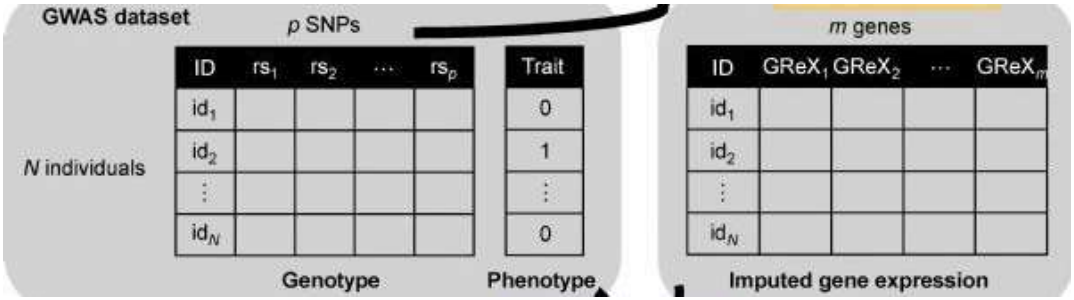
Step 1. Model training: the genetic predictor of gene expression (E_g) is learned in a reference panel

Data needed: a reference panel with both genotype data and transcriptome data

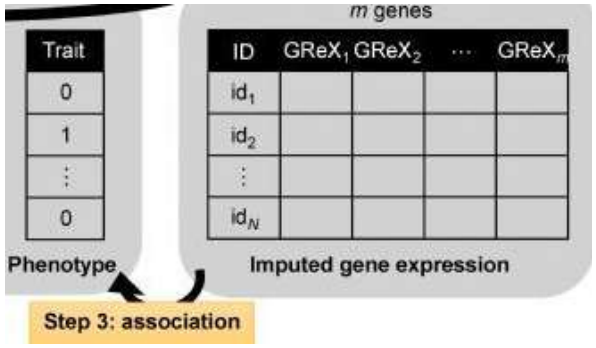


TWAS method

Step 2. Imputation of gene expression in GWAS datasets



Step 3. Association



TWAS Example: Blood Cell Traits

Blood cell traits (BCTs)

- Clinically important
- Complete blood cell count
- Acute and chronic disease risk
- Insights into inflammation, oxygen transport, blood clotting

Genome-wide association studies (GWAS)

- Identified >2, 700 variants for BCTs
- Limited in identifying causal genes and pathways

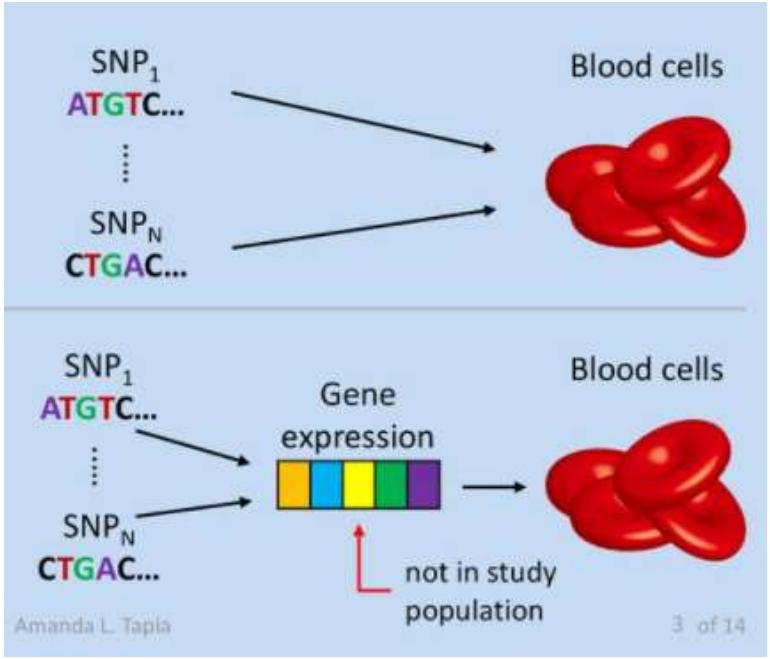
Transcriptome-wide association study (TWAS)

- Transcriptome = gene expression
- Advantages beyond GWAS
  - Better understanding biological mechanisms of association
  - Reduced multiple test correction
  - Potentially identify novel loci missed by GWAS

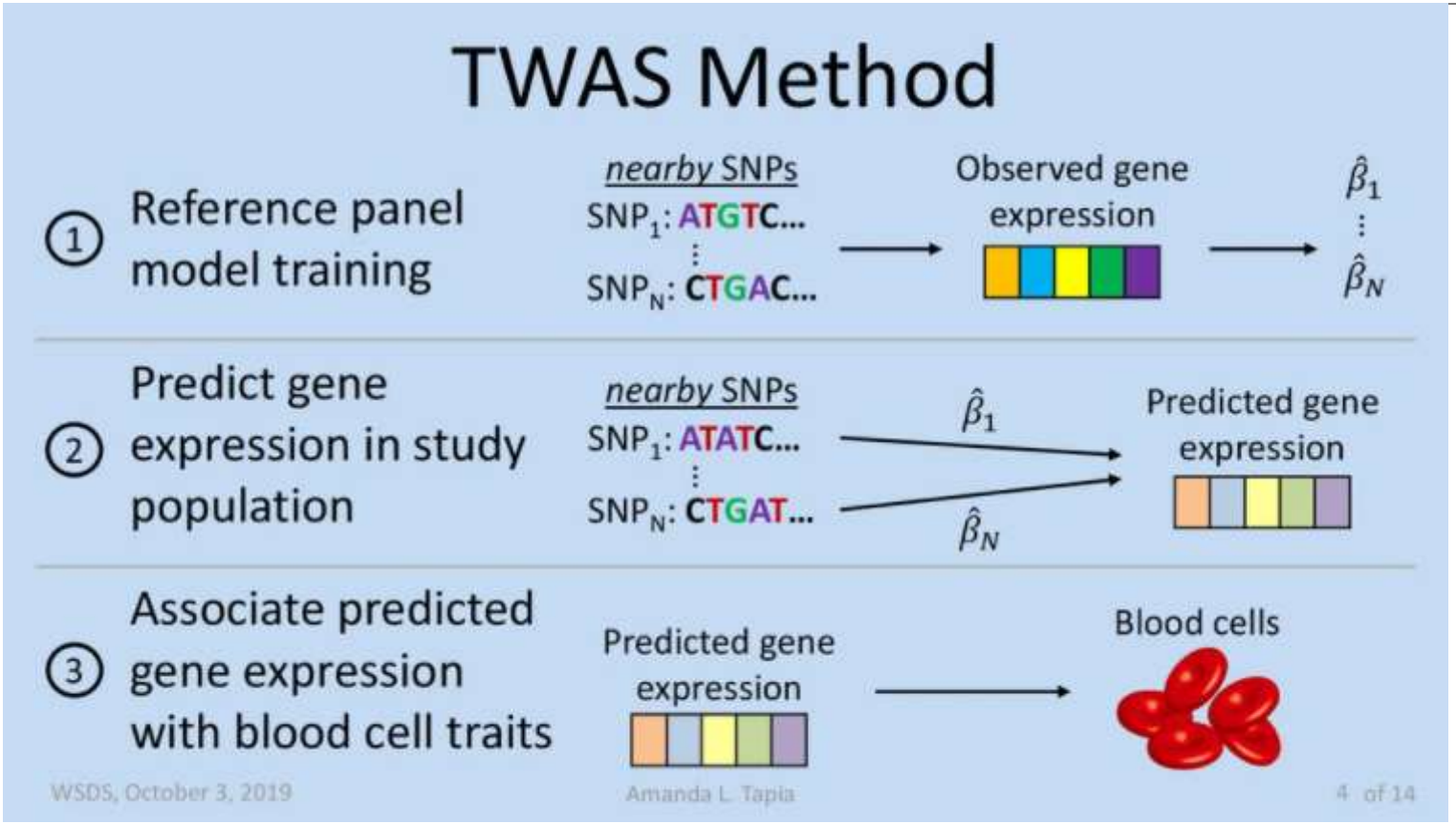
TWAS Example: Blood Cell Traits

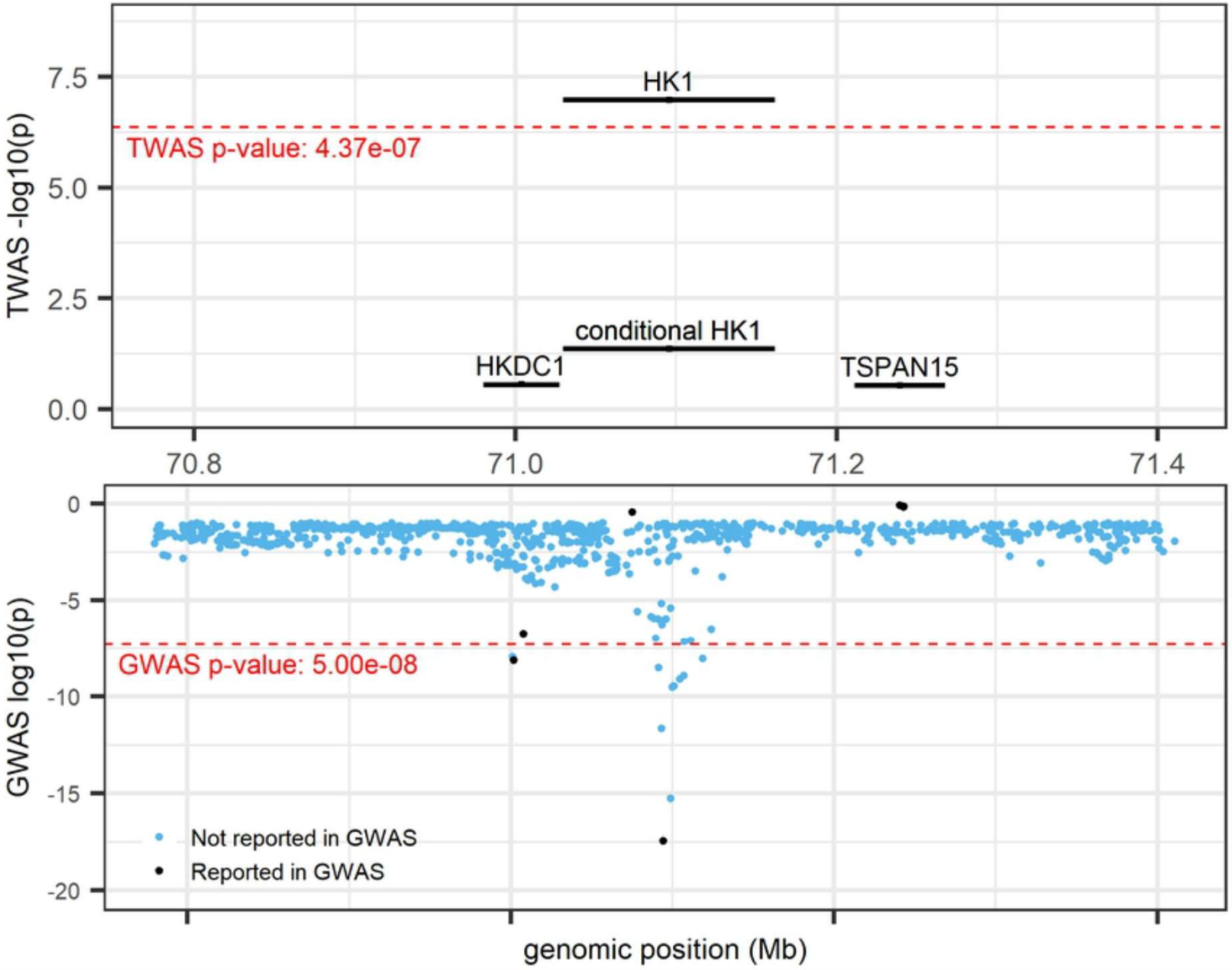
GWAS:
Relates genetic variation directly to blood cells (~ millions of SNPs)

TWAS:
Relates genetically regulated gene expression to blood cell (~ 10,000 genes)

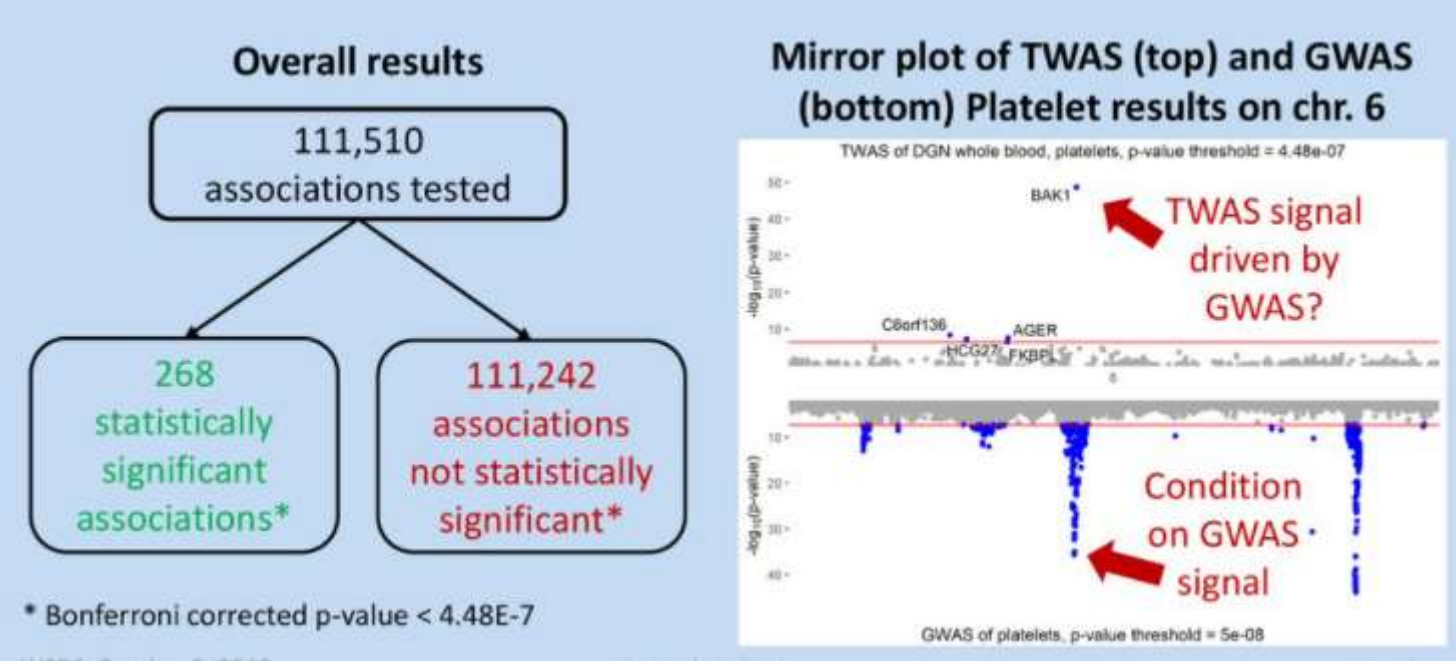


TWAS Example: Blood Cell Traits




TWAS -log10(p)
7.5
5.0
2.5
0.0
HK1
TWAS p-value: 4.37e-07
conditional HK1
HKDC1
TSPAN15
70.8
71.0
71.2
71.4
GWAS log10(p)
0
-5
-10
-15
-20
GWAS p-value: 5.00e-08
Not reported in GWAS
Reported in GWAS
genomic position (Mb)

Results


Overall results
111,510
associations tested
268
statistically
significant
associations*
111,242
associations
not statistically
significant*
* Bonferroni corrected p-value < 4.48E-7
Mirror plot of TWAS (top) and GWAS (bottom) Platelet results on chr. 6
TWAS of DGN whole blood, platelets, p-value threshold = 4.48e-07
BAK1
TWAS signal driven by GWAS?
C6orf136
AGER
HCG27
FKBPL
Condition on GWAS signal
GWAS of platelets, p-value threshold = 5e-08

Outline

Association ←→ causal

Causal inference

- Colocalization analysis of GWAS data
- Mendelian randomization
- Fine-mapping
- Convolutional Neural Network in predicting functional impact of genetic variants

Transcriptome-wide association study (TWAS)

Epigenome-wide association study (EWAS)

PheWAS

Risk prediction: Polygenic Risk Score (PRS)

Epigenome-Wide Association Study (EWAS)

1) identification of differentially methylated regions/loci (e.g., HPG-DHunter, DMRcaller),

2) analysis of the association between epigenetic variation and disease/phenotype (e.g., EWAS2.0, EWAS1.0),

3) comprehensive analysis of DNA methylation data (e.g., GLINT, TABSAT),

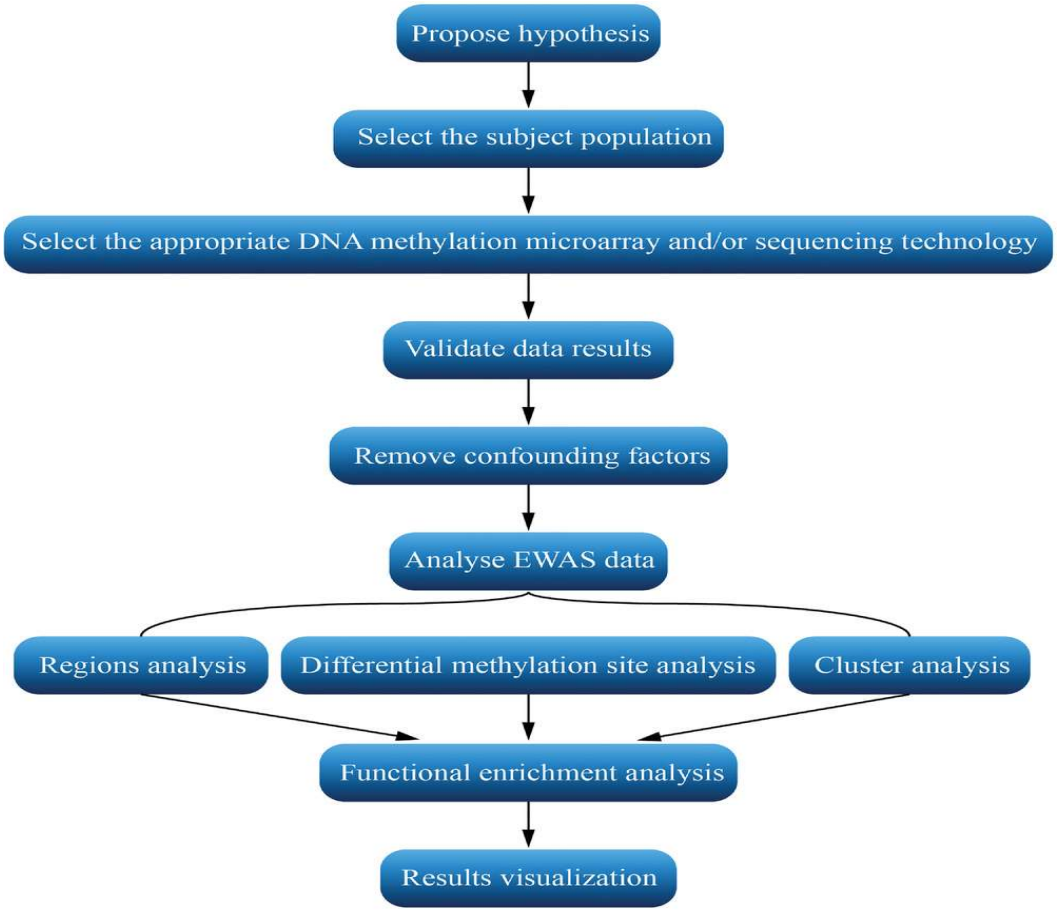
4) prediction of histone modifications and DNA methylation level (e.g., Pancancer DNA Methylation Trackhub, Epigram),

5) prediction of complex traits based on methylation (e.g., TANDEM, OmicKriging),

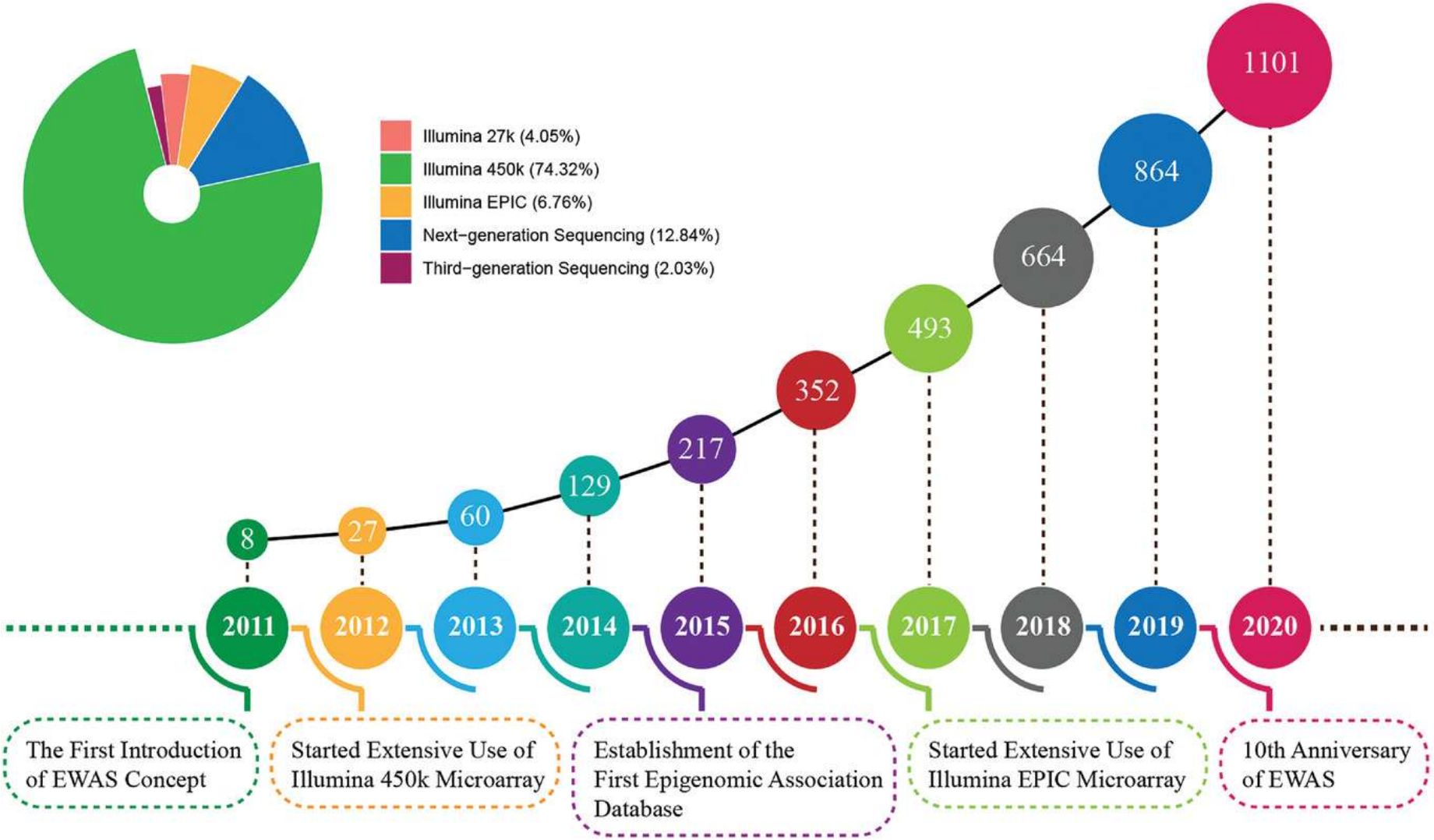
6) identification of differential cell types based on methylation (e.g., CellDMC, BPRMeth), and

7) methylation data processing and normalization (e.g., omicsPrint, FuntooNorm).

Epigenome-Wide Association Study (EWAS)



Advanced Science, Volume: 8, Issue: 20, First published: 11 August 2021, DOI: (10.1002/advs.202100727)


Illumina 27k (4.05%)
Illumina 450k (74.32%)
Illumina EPIC (6.76%)
Next-generation Sequencing (12.84%)
Third-generation Sequencing (2.03%)
8
27
60
129
217
352
493
664
864
1101
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
The First Introduction of EWAS Concept
Started Extensive Use of Illumina 450k Microarray
Establishment of the First Epigenomic Association Database
Started Extensive Use of Illumina EPIC Microarray
10th Anniversary of EWAS

EWAS Achievements

Prediction of Disease Risk

- Predict specific disease risk by identifying specific DNA methylation loci as biomarkers
- A study developed a methylation risk score (MRS) based on levels of methylation change. Researchers used this score together with information on 187 CpG loci associated with obesity to predict the risk of developing type 2 diabetes (T2D) in the future.

Early Diagnosis of Disease

- Indicating biomarkers early in the disease process can help alter the disease process or even stop its progression, e.g., autism spectrum disorders
- An EWAS has identified three CpG loci that can be used as biomarkers for the early diagnosis of colorectal cancer (CRC).

EWAS Achievements

Identifying Drug Targets

- One effective way to fight cancer is to inhibit methylation, and epigenetic drugs can have an impact on DNA methylation patterns
- Several epigenetic drugs targeting histone methyltransferases and DNA methyltransferases are currently available for the treatment of many types of cancer.[71] For instance, Zebularine, Azacitidine, and Chaetocin have been broadly used in the clinical practice

Measuring Drug Response by Monitoring Drug-Induced Epigenetic Changes

- Examining drug-induced epigenetic changes is a novel way to measure drug response and evaluate prognostic ability

Word cloud of traits in EWASs. Top 100 traits in EWAS, including phenotypes, behaviors, environmental factors, cancer and noncancer diseases.

Outline

Association ←→ causal

Causal inference

- Colocalization analysis of GWAS data
- Mendelian randomization
- Fine-mapping
- Convolutional Neural Network in predicting functional impact of genetic variants

Transcriptome-wide association study (TWAS)

Epigenome-wide association study (EWAS)

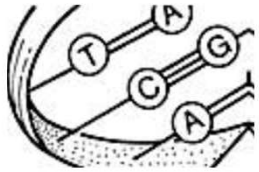
Phenome-wide association studies (PheWAS)

Risk prediction: Polygenic Risk Score (PRS)

Phenome-wide association studies (PheWAS)

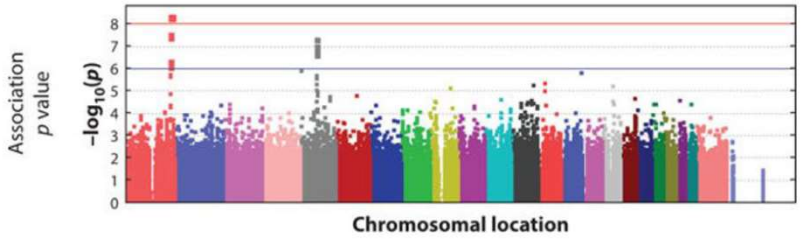
Genome-wide association studies (**GWAS**) –focused on a single disease or a small set of diseases at a time in order to find specific genotype/phenotype associations

Genetic variants



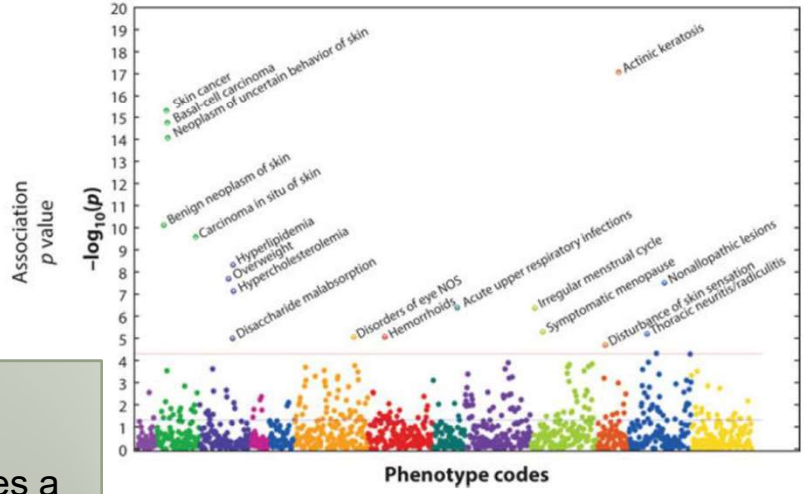
Phenotype

GWAS: Target phenotype →



PheWAS: Target genotype →

(or other input variable, e.g., a specific disease, trait, or exposure)



- Phenome-wide association studies (PheWAS): Combining patient genotype data and electronic health record (EHR)
  - Begins with a genotype and then systematically queries a large number of phenotypes
  - Allows to study multiple phenotypes

Electronic Health Records (EHR)

The Electronic Health Record (EHR) is an electronic compilation of longitudinal data related to the complete healthcare of an individual.

EHRs proved a valuable resource for analyzing pharmacogenetic traits and developing reverse genetics approaches such as phenome-wide association studies

1000s of phenotypes constructed usin

- ICD codes
- PheCodes (Denny et al. 2010)
- Manually curated phenotypes

TENTH REVISION

ICD-10

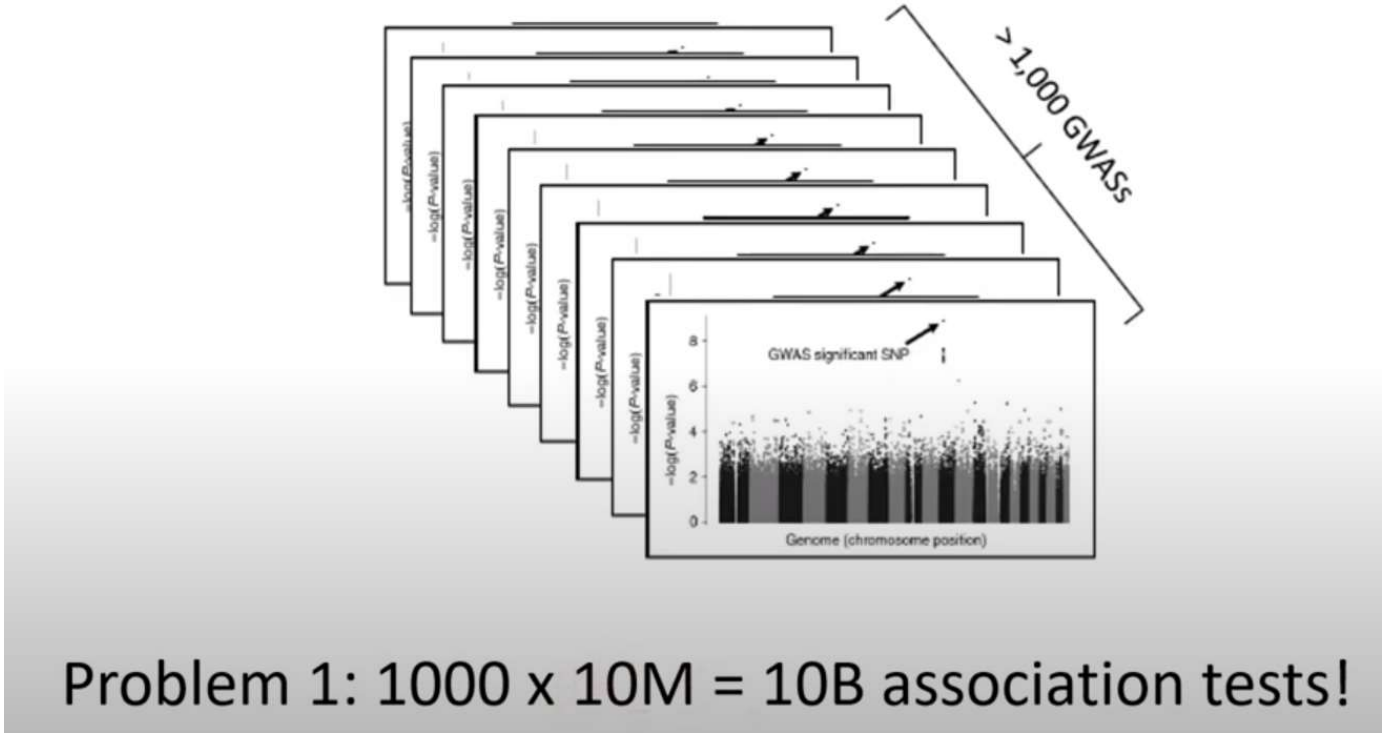
INTERNATIONAL CLASSIFICATION OF DISEASES

ICD-10 is a new code set for reporting medical diagnoses & inpatient procedures.

International Statistical Classification of Diseases and Related Health Problems (ICD)

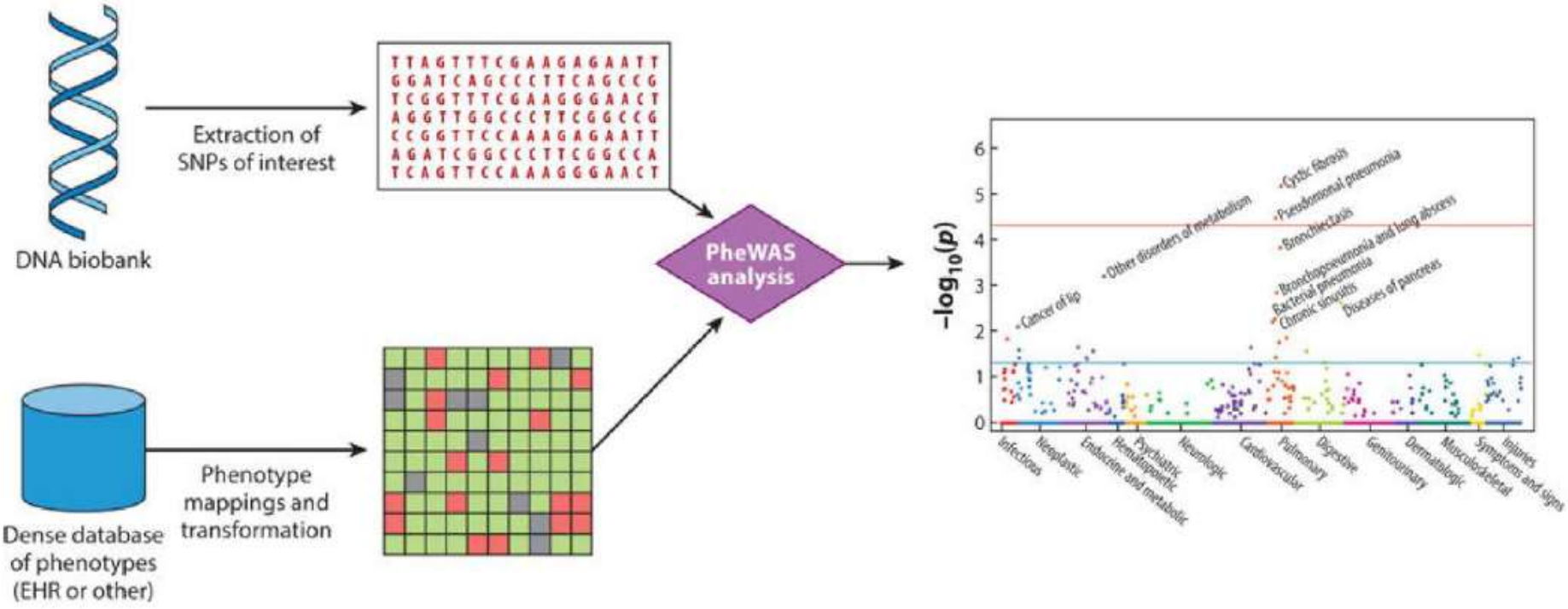
Übersicht über die Klassifikation psychischer und Verhaltensstörungen nach ICD-10 Kap. V (F)

EHR + GWAS = PheWAS



Methodology

- PheWASs are designed to survey which of many phenotypes may be associated with a given genetic variant
- The essential process – **identify a large list of phenotypes, ideally collected systematically (and not restricted to phenotypes of predefined interest)**.



A typical transformation would take ~14,000 diagnostic billing codes and identify ~1,600 distinct case phenotypes, each matched to a control group.

First PheWAS study (2010)

1) Identify individual cases and controls for 776 diseases
2) Tested of 5 single-nucleotide polymorphisms (SNPs) already known to be associated with 7 of these diseases

SNP	Gene/region	Disease	Cases	Previous OR	PheWAS *P*-value	PheWAS OR
rs3135388	DRB1*1501	MS	89	1.99[a]	2.77×10^{-6}	2.24 (1.56–3.16)
		SLE	141	2.06[b]	0.51	1.13 (0.79–1.58)
rs17234657	Chr. 5	CD	200	1.54[c]	0.00080	1.57 (1.19–2.04)
rs2200733	Chr. 4q25	AF and flutter	606	1.75[d]	0.14	1.15 (0.95–1.39)
rs1333049	Chr. 9p21	CAD	1181	1.20–1.47[e]	0.011	1.13 (1.03–1.23)
		Carotid atherosclerosis	333	1.46[f]	0.82	0.98 (0.84–1.15)
rs6457620	Chr. 6	RA[g]	392	2.36[c]	0.0002	1.35 (1.15–1.58)

Results

- Replicated **4 of 7 associations**
- Established 19 new SNP-disease association

Bioinformatics. 2010 May 1;26(9):1205-10. doi: 10.1093/bioinformatics/btq126. Epub 2010 Mar 24.

PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations.

Denny JC[1], Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, Wang D, Masys DR, Roden DM, Crawford DC.

Author information

1 Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, USA. josh.denny@vanderbilt.edu

PheWAS limitations

Multiple hypotheses testing – number of phenotypes (on the order of 10^3) ×genotypes (on the order of 10^6). Many phenotypes are highly correlated (~LD of genetic variants).

- Accuracy of the phenotypes derived via PheWASs, especially for EHR data.
- Pseudopleiotropy and true pleiotropy. Pseudopleiotropy – differences along a causal pathway.

Phenogram

Plots phenotypes that have been associated with SNPs or other locations along the genome.

Outline

Association ←→ causal

Causal inference

- Colocalization analysis of GWAS data
- Mendelian randomization
- Fine-mapping
- Convolutional Neural Network in predicting functional impact of genetic variants

Transcriptome-wide association study (TWAS)

Epigenome-wide association study (EWAS)

PheWAS

Risk prediction: Polygenic Risk Score (PRS)

Complex phenotype prediction

Risk prediction:

Clinical decision-making, early disease detection and prevention of common adult-onset conditions.

Current practice:

- Basic demographic characteristics (such as age, gender and ethnicity)
- Basic health parameters and lifestyle factors (such as body mass index, smoking status, alcohol consumption and physical exercise habits)
- Measurement of clinical risk factors proximal to overt disease onset (such as blood pressure levels, blood chemistries or biomarkers indicative of ongoing disease processes)
- Ascertainment of environmental exposures (such as air pollution, heavy metals and other environmental toxins)
- Family history.

Complex phenotype prediction

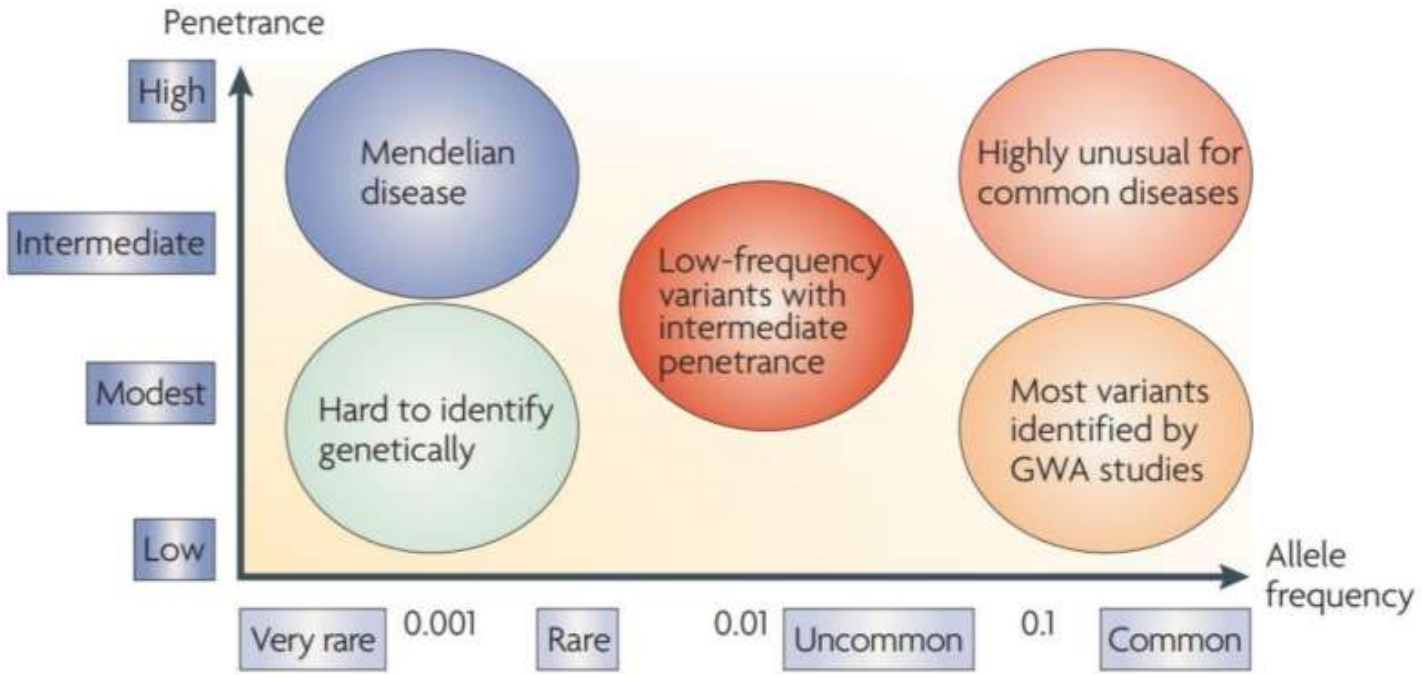
For complex phenotypes, rare mutations *may* exist to confer several-fold increased risk in heterozygous carriers.

- Familial Hypercholesterolemia (FH, 家族性高胆固醇症) mutation (LDLR, APOB, PCSK9): 0.4% population, conferring ~3-fold increased risk for coronary artery disease
- p.Glu508Lys (HNF1A) 0.1% of the general population, conferring ~5-fold increased risk for type 2 diabetes

The 'Angelina Effect'

Although the ascertainment of monogenic mutations can be highly relevant for carriers and their families, the vast majority of disease occurs in those without such mutations.

Complex phenotype prediction



McCarthy, M., Abecasis, G., Cardon, L. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* **9,** 356–369 (2008).

CDCV/RAME/infinitesimal/Broad-sense-heritability

Common Disease Common Variant

- Complex disease is largely attributable to a moderate number of common variants, each of which explains several per cent of the risk in a population.

The rare alleles of major effect (RAME) model

- a large number of large-effect rare variants

The infinitesimal model

- A large number of small-effect common variants across the entire allele frequency spectrum

Broad sense heritability model

- Non-additive G×G and G×E interactions and epigenetic effects

Complex phenotypic prediction

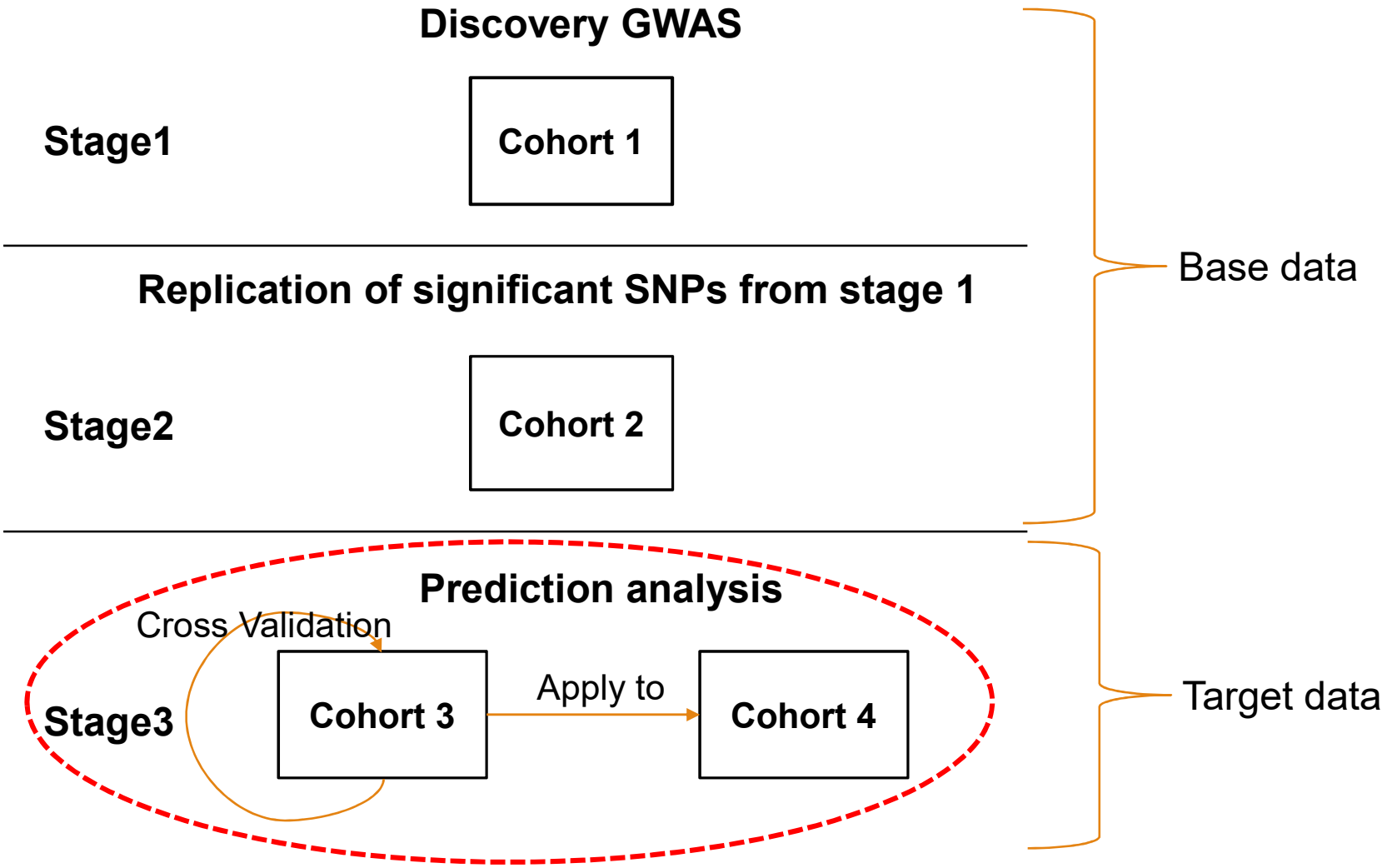
GWAS have identified thousands of common susceptibility variants for a wide spectrum of complex traits.

Genetic information remains largely unchanged through life

Combination of identified SNPs explain a significant portion of the trait's variation

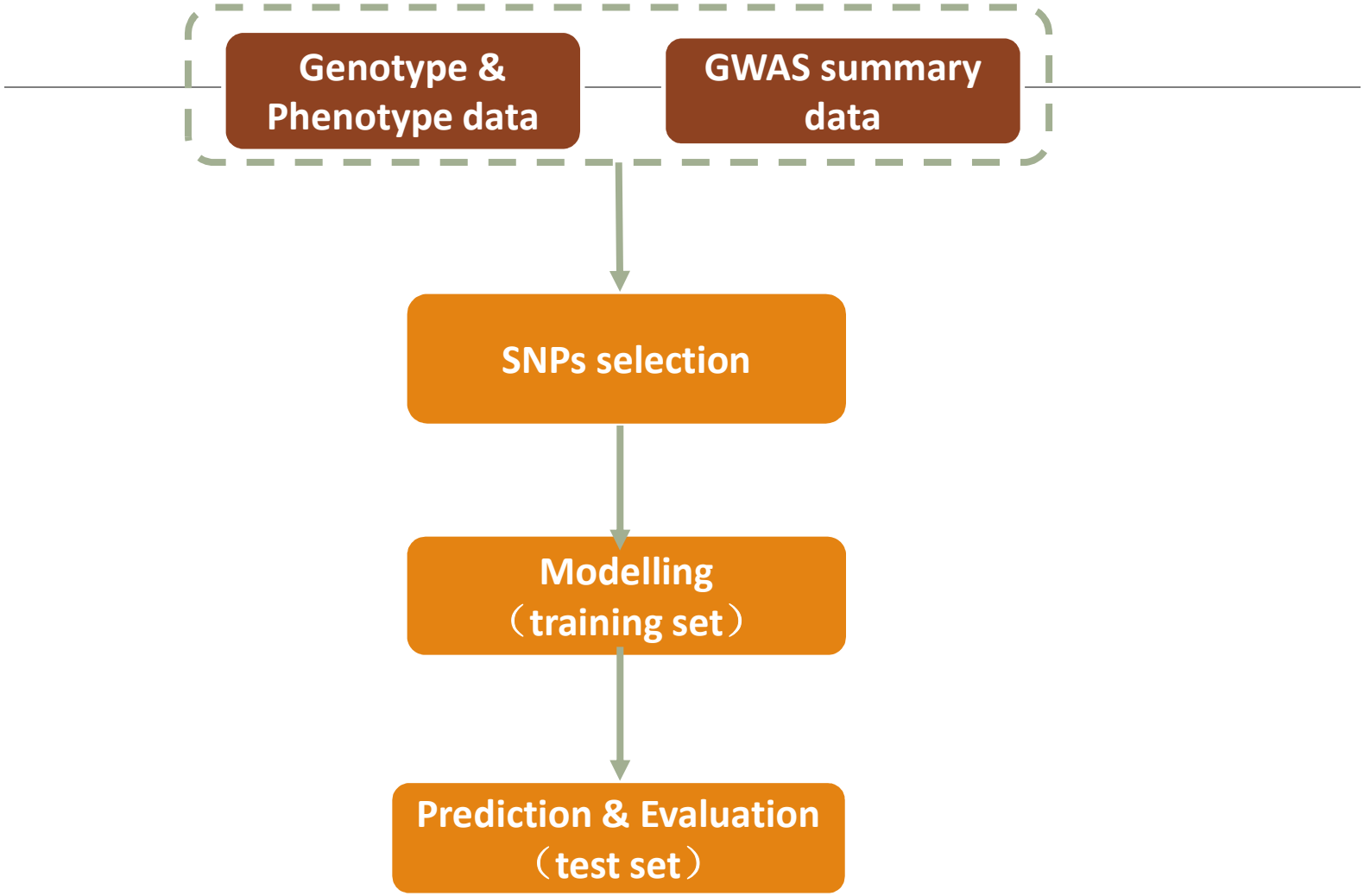
SNPs identified from GWAS can be used for the assessment of polygenic disease risks and prediction of other complex traits.

Pipeline of Complex phenotypic prediction



Cohort 1-4 are independent cohorts.

Pipeline of Complex phenotypic prediction



PRS (polygenic risk score)
PGS (polygenic score)

A polygenic risk score (PRS) is a sum of trait-associated alleles across many genetic loci, typically weighted by effect sizes estimated from a genome-wide association study.

$$PGS_i = \sum_{j=1}^{M} a_{ij} w_j$$

i 表示第i个个体, j 为第j个SNP, wj为该SNP的权重, a则为第i个个体第j个SNP的等位基因拷贝数

Method to select SNPs

Selecting appropriate SNPs from GWAS results for phenotypic prediction

Many SNPs affecting the phenotype

Too many SNPs included model, weakening the stability and generalization ability of the prediction

Too few SNPs may lead to too much deviation from the actual model

Shrinkage of the effect estimates of all SNPs via standard or tailored statistical techniques

- LASSO, ridge regression, Bayesian approaches…

Clumping and p-value selection thresholds

- Only those SNPs with a GWAS association P value below a certain threshold (e.g., $P < 1 \times 10-5$) are included in the calculation of the PRS

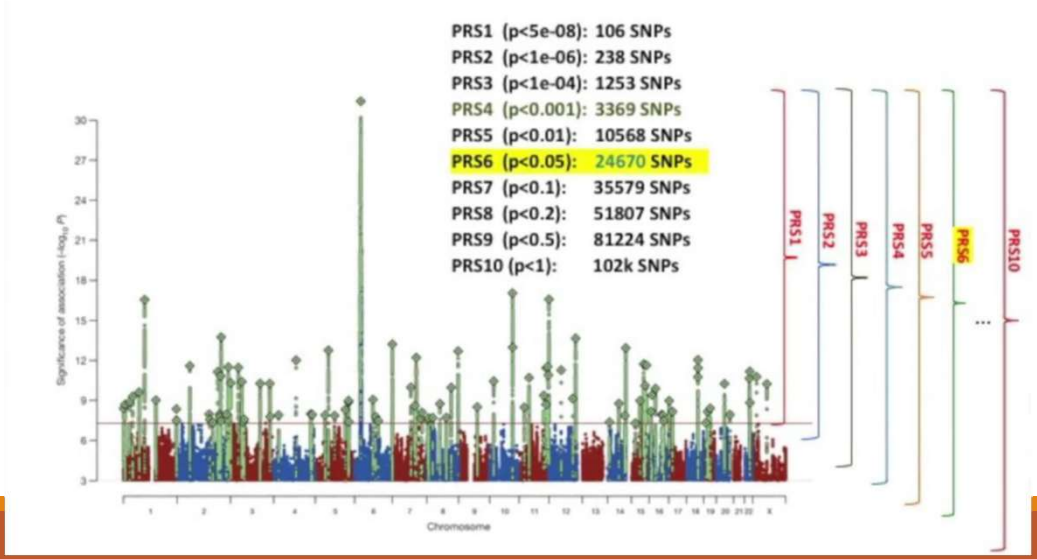
C+T method (clumping + thresholding)

To add up a list of SNPs, the SNPs need to be independent, or near independent.

SNPs are clumped (i.e., thinned, prioritizing SNPs at the locus with the smallest GWAS P value) so that the retained SNPs are largely independent of each other, and, thus, their effects can be summed, assuming additivity.

PRS can be calculated using the clumped SNPs at different p-value threshold.

- 0.001?
- 0.05?
- 0.1?

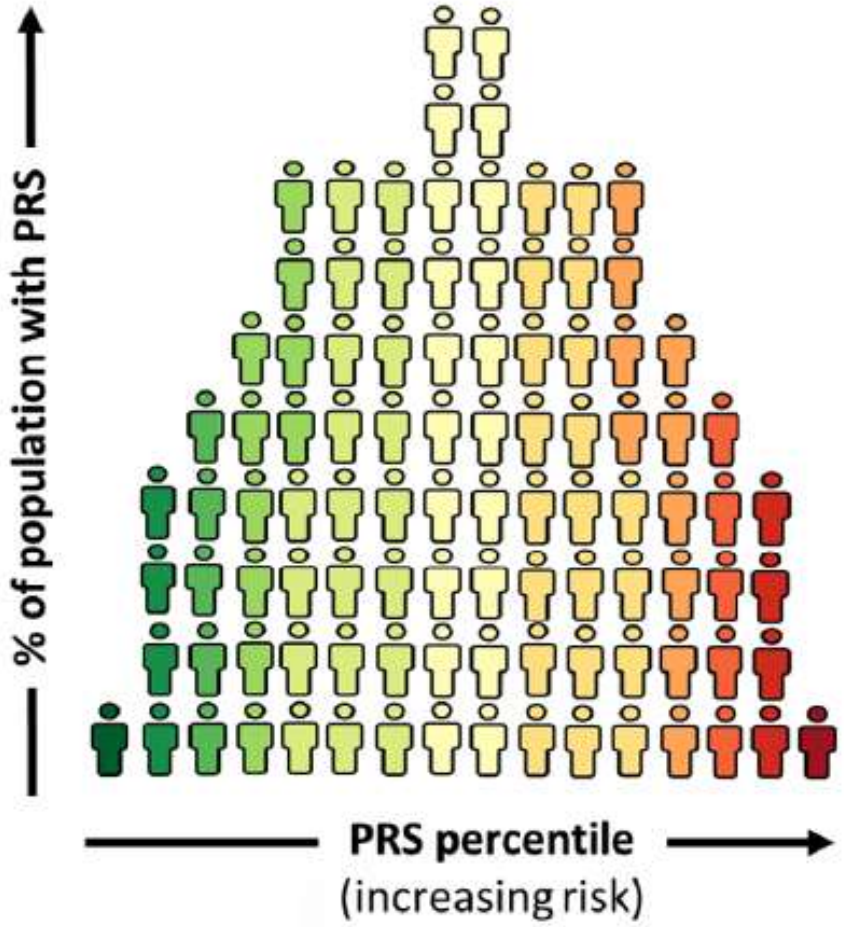


Polygenic risk score

$$PGS_i = \sum_{j=1}^{M} a_{ij} w_j$$

PRS Distribution

PRS percentile	Risk of disease vs. reference group
0-1	Lowest
1-5	
5-10	
10-20	
20-40	
40-60 (reference)	**1**
60-80	
80-90	
90-95	
95-99	
99-100	Highest



Source: RGA

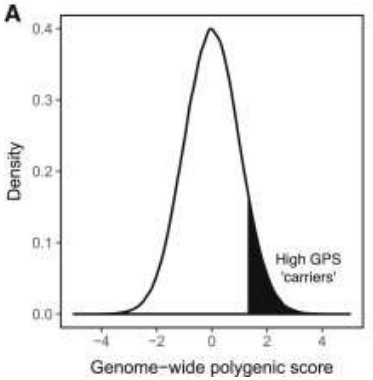
Successful applications of PRS

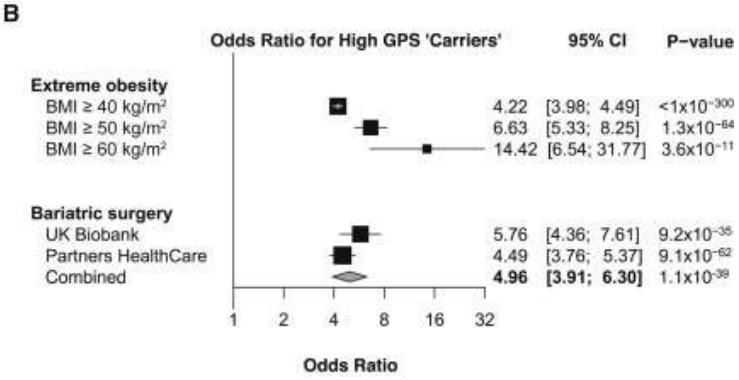
Disorder	No. of Genetic Variants	Relative risk, comparing top 20% to bottom 20% PRS	Reference
Coronary artery disease	**50**	**2.0**	Khera AV. *et al*. (2016), N Engl J Med.
Coronary artery disease	**49,310**	**1.8 to 4.5**	Abraham G. *et al*. (2016), Eur Heart J.
Type 2 diabetes	**1000**	**3.5**	Läll K. *et al*. (2017), Genet Med.
Ischemic stroke	**10**	**1.2 to 2.0**	Hachiya T. *et al*. (2017), Stroke
Breast cancer	**77**	**3.0**	Mavaddat N. *et al*. (2015), J Natl Cancer Inst.
Breast cancer (East Asian ancestry)	**44**	**2.9**	Wen W. *et al*. (2016), Breast Cancer Res.
Prostate cancer	**25**	**3.7 (25%)**	Amin Al Olama A. *et al*. (2015), Cancer Epidemiol Biomarkers Prev.
Lung cancer	**38**	**4.6 (25%)**	Cheng Y. *et al*. (2016), Oncotarget

Polygenic Prediction of Weight and Obesity Trajectories

Table 1. Genome-wide Polygenic Score for Obesity, Assessed in F

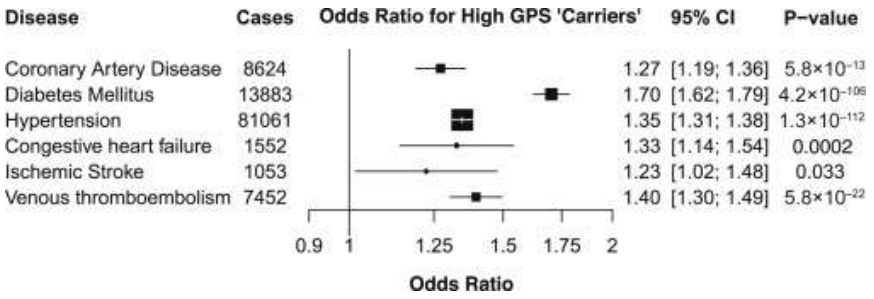
	UK Biobank	Partners HealthCare
n participants	288,016	6,536
Study design	cross-sectional	case-control
Age range	40–69 years	≥18 years
Female sex	55%	61%
Outcomes	weight, severe obesity, bariatric surgery, cardiometabolic diseases, mortality	bariatric surgery





Association of High GPS with Extreme Obesity

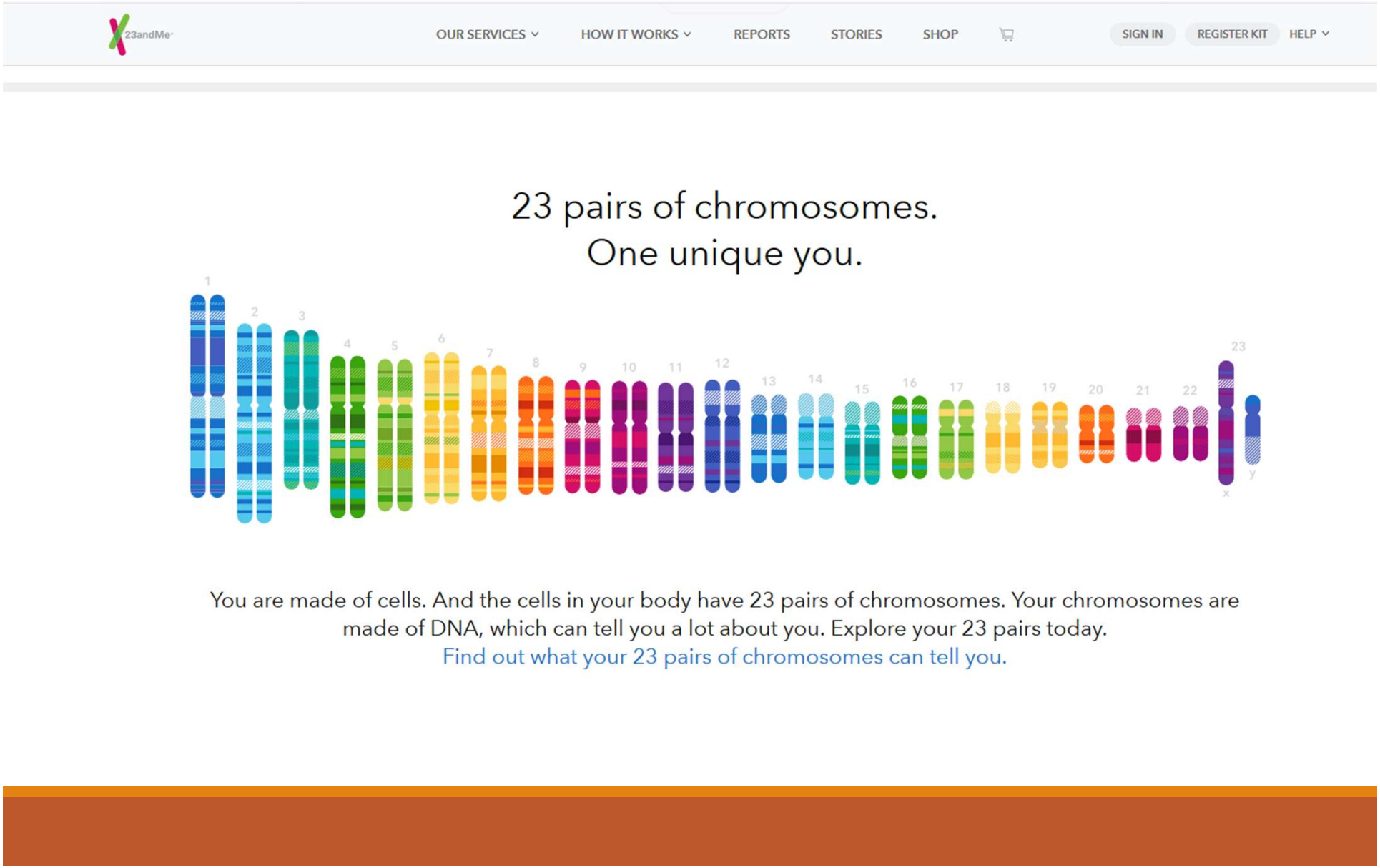
- A genome-wide polygenic score can quantify inherited susceptibility to obesity
- Polygenic score effect on weight emerges early in life and increases into adulthood
- Effect of polygenic score can be similar to a rare, monogenic obesity mutation
- High polygenic score is a strong risk factor for severe obesity and associated diseases



Association of High GPS with Cardiometabolic Diseases

Khera et al. Cell, 177(3)587-596, 2019

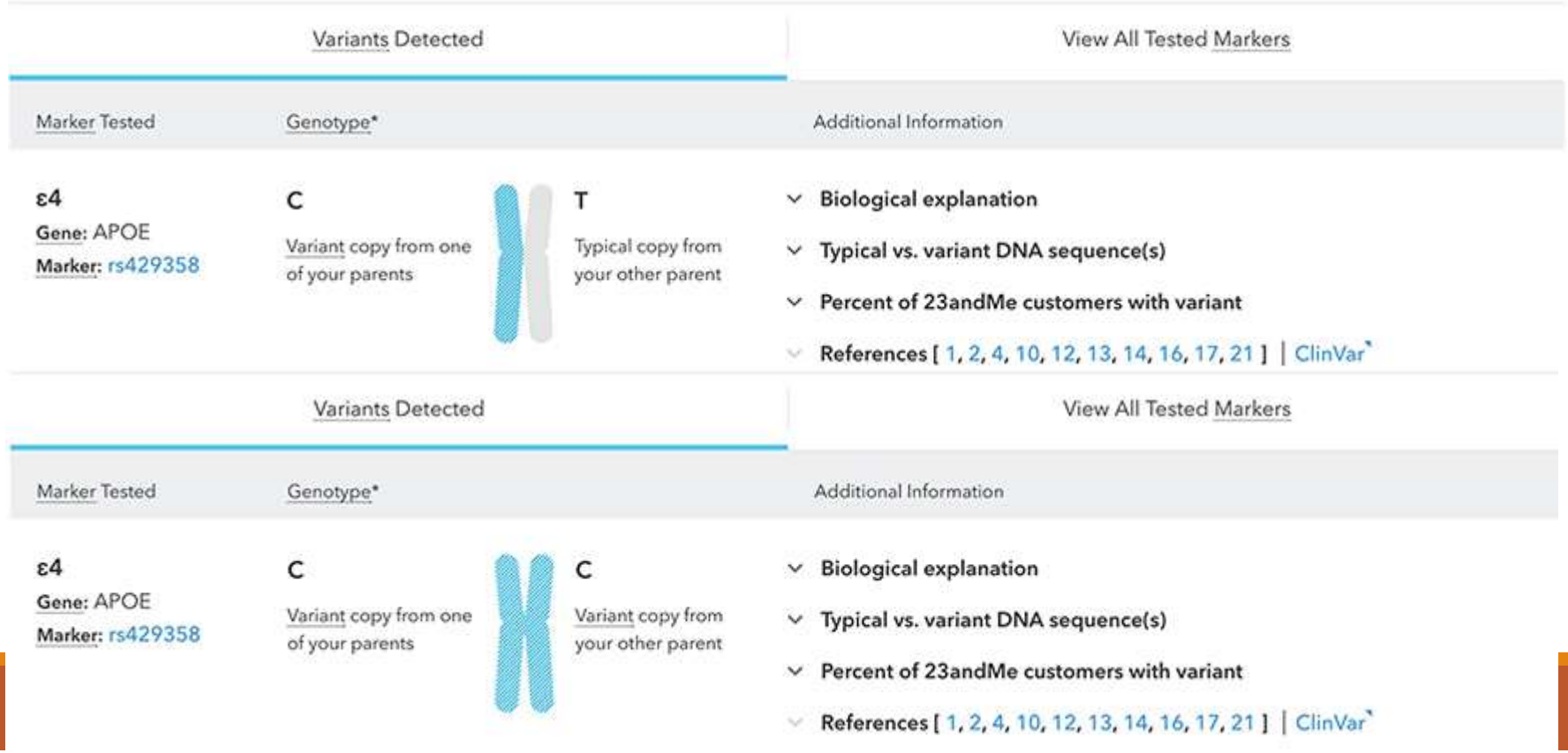
23andme


23andMe
OUR SERVICES
HOW IT WORKS
REPORTS
STORIES
SHOP
SIGN IN
REGISTER KIT
HELP
23 pairs of chromosomes.
One unique you.
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
x y
You are made of cells. And the cells in your body have 23 pairs of chromosomes. Your chromosomes are made of DNA, which can tell you a lot about you. Explore your 23 pairs today.
Find out what your 23 pairs of chromosomes can tell you.

Report APOE

Late-Onset Alzheimer's Disease

- 0 variants – female
- 0 variants – male
- 1 variant, 1 copy – female
- 1 variant, 1 copy – male
- Variant detected, 2 copies – female
- Variant detected, 2 copies – male
- Variant not determined



Ten years of GWAS (Summary)

Complex traits are highly polygenic

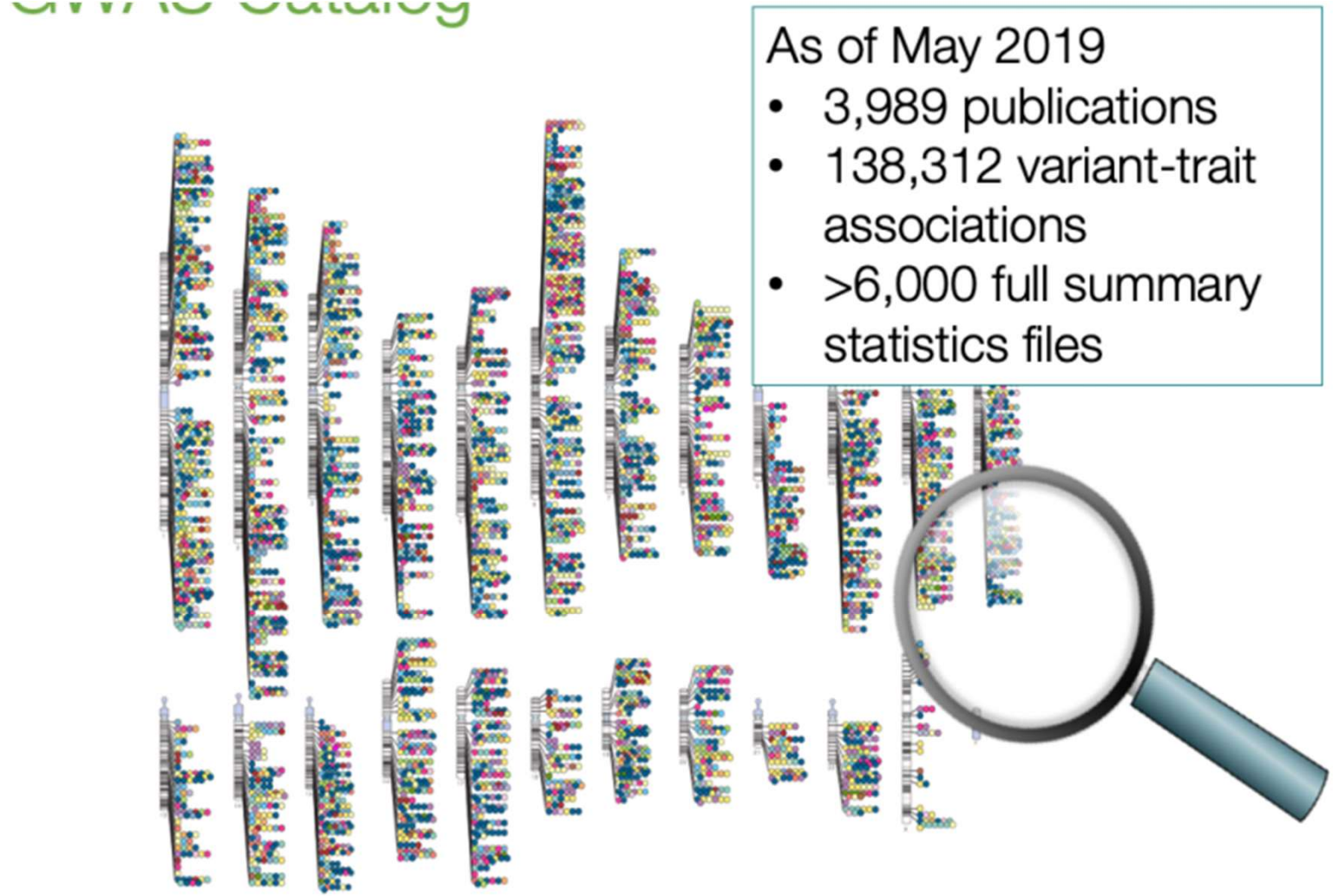
Pleiotropy is pervasive

The missing heritability problem

New Analysis Methodology Underpinning New Discovery

The Utility of GWAS-Derived Genetic Predictors

GWAS Catalog



Complex Traits Are Highly Polygenic

For almost any complex trait that has been studied, many loci contribute to standing genetic variation.

- On average, the proportion of variance explained at the individual variants is small

Larger experimental sample sizes will lead to new discoveries.

The term polygenic describes the genetic architecture underpinning variation in a trait between individuals in a population.

- Each individual will carry a number of alleles that increase (+) and a number of alleles that decrease (-) the trait or disease risk.

Pleiotropy Is Pervasive

Multiple lines of evidence are consistent with widespread pleiotropy for complex traits.

“One gene, one function, one trait” – not standing

Mendelian mutations that cause specific syndromes or diseases are frequently associated with multiple phenotypes in an affected individual.

Pedigree studies have reported genetic correlations between traits, implying that a number of the same variants affect two or more traits in a consistent direction.

GWASs have shown that the same genetic variants can be significantly associated with multiple diseases and traits when the phenotypes are measured on different individuals.

Analytical methods that estimate genetic correlations from GWAS data have provided evidence for widespread pleiotropy

The Missing Heritability Problem

The heritability for height explained by significantly associated SNPs is only 10%, while that explained by all measured SNPs is 45% -- still much smaller than a frequently quoted h^2 of 80% from family or twin studies.

New Analysis Methodology Underpinning New Discovery

methods of better modeling population structure and relatedness between individuals in a sample during association analyses

methods of detecting novel variants and gene loci on the basis of GWAS summary statistics

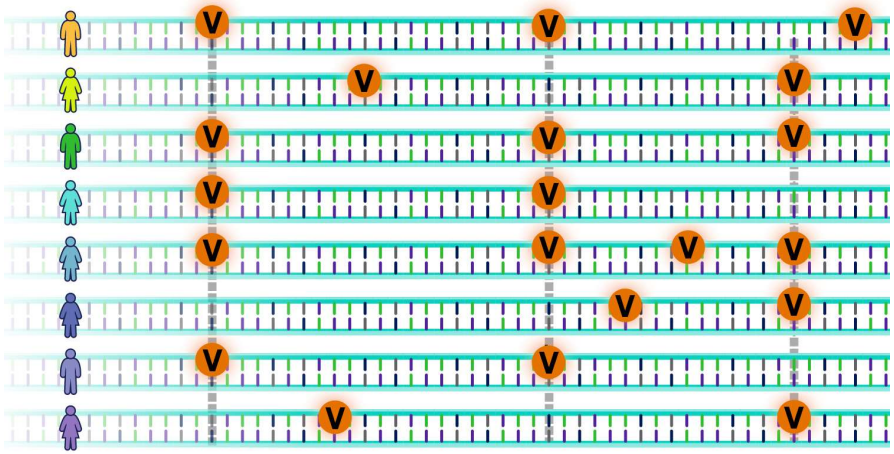
methods of estimating and partitioning genetic (co)variance

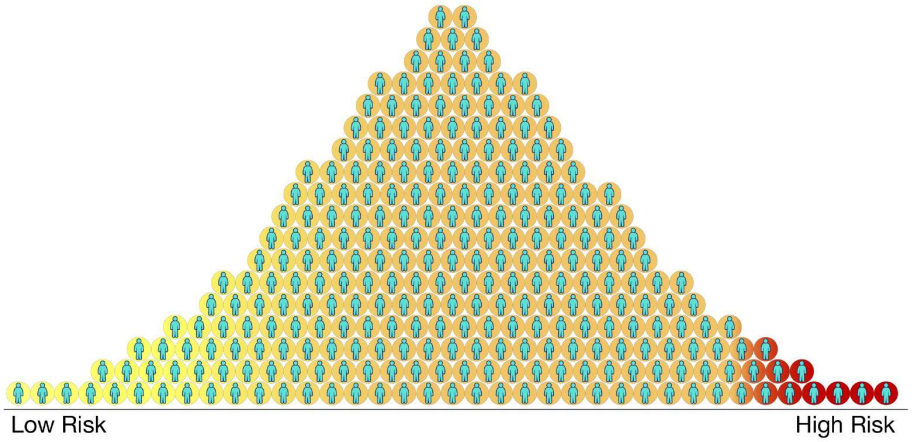
methods of inferring causality

The Utility of GWAS-Derived Genetic Predictors

Generate a polygenic risk score (PRS) per individual

Some variants increase the risk of developing diseases, while others may reduce such risk; others have no effect on disease risk.





Thank you for your attention!