Integrative Methods

DCEG Statistical Genetics Workshop

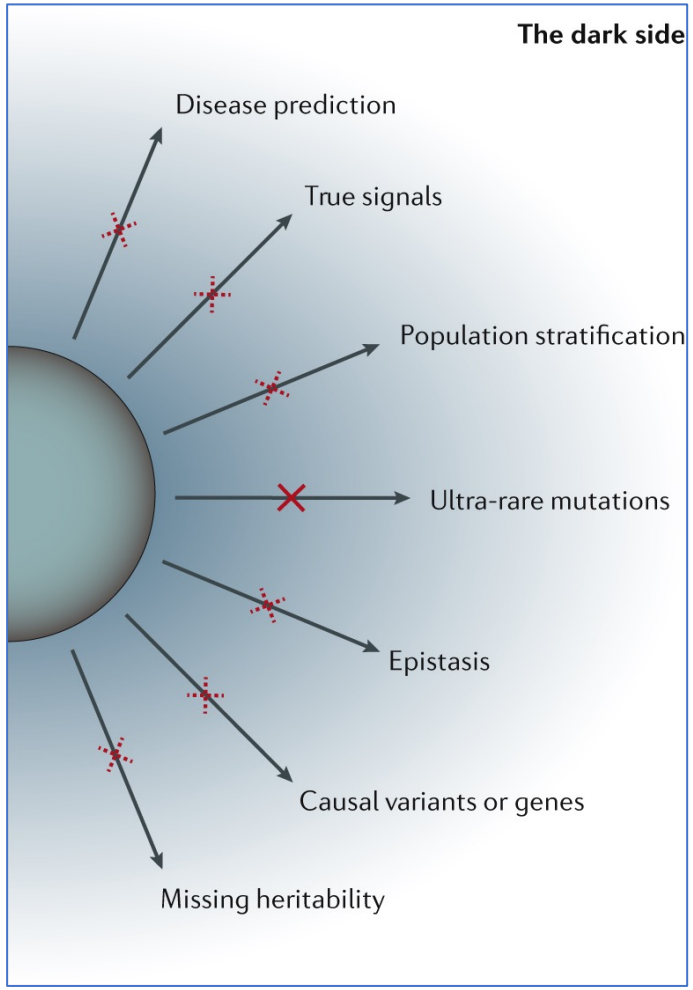
Lecture 6 (Part 1)

15 November 2023

Outline

- Transcriptome-Wide Association Studies (TWAS)
  - “General Steps”
  - Software
  - Why Perform a TWAS?
  - Limitations / Caveats
- Resources
- Other “–WAS” Approaches
  - pWAS
  - Metabo-WAS
  - Epigenome-WAS
  - Regulome-WAS
- Example Applications in Practice – Breast Cancer Risk

Background: Moving Beyond GWAS…



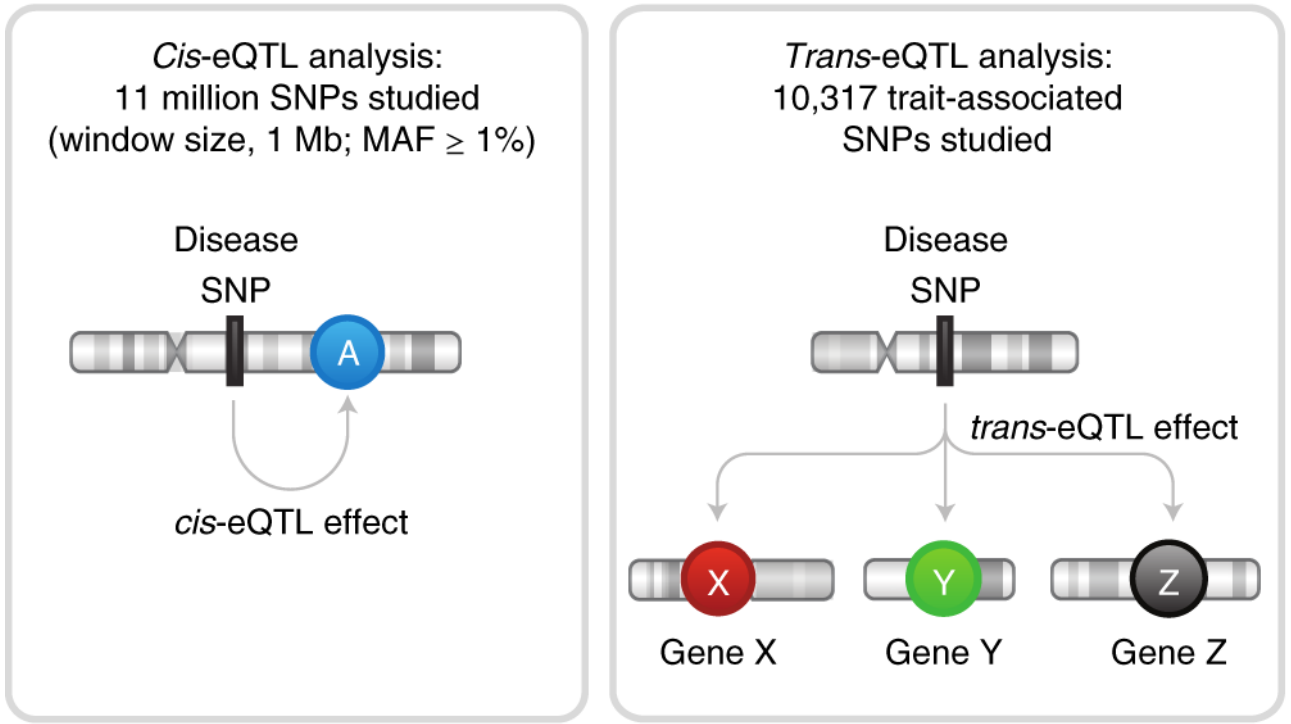
"The Dark Side" of GWAS:

=> Most associated variants not present in coding regions of the genome…

Tam, V., Patel, N., Turcotte, M. et al. Nat Rev Genet 20, 467–484 (2019).

Background: Expression Quantitative Trait Loci (eQTLs)

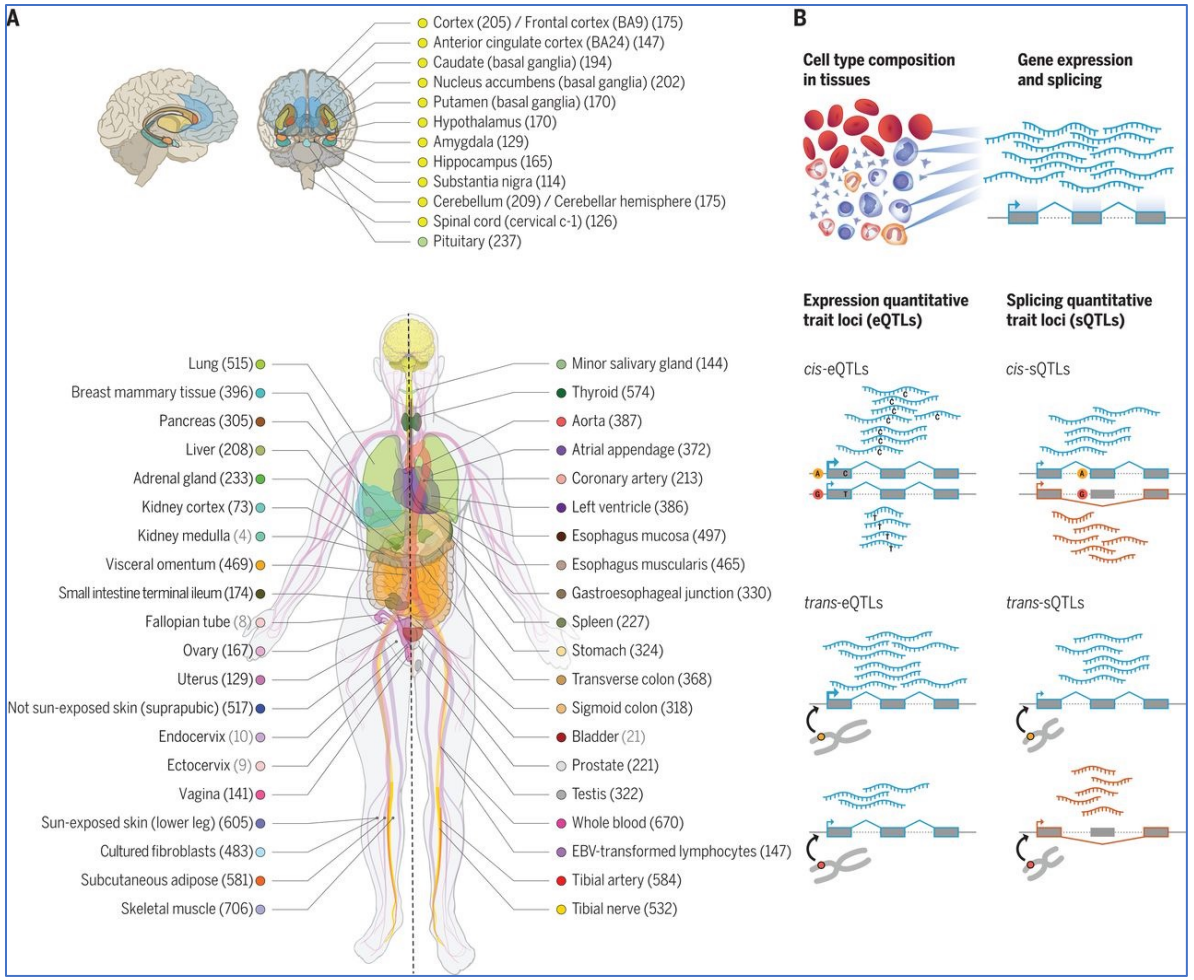
Locus (typically non-coding) that explains variation in mRNA / gene expression



Võsa, U., Claringbould, A., Westra, HJ. et al. Nat Genet 53, 1300–1310 (2021).

Background: Genotype-Tissue Expression Database (GTEx v8)

- Estimation of “normal” effects of genotype -> expression change per tissue
- Tissue-specific gene expression from rapid autopsy donors
- **54 tissues, 948 donors**

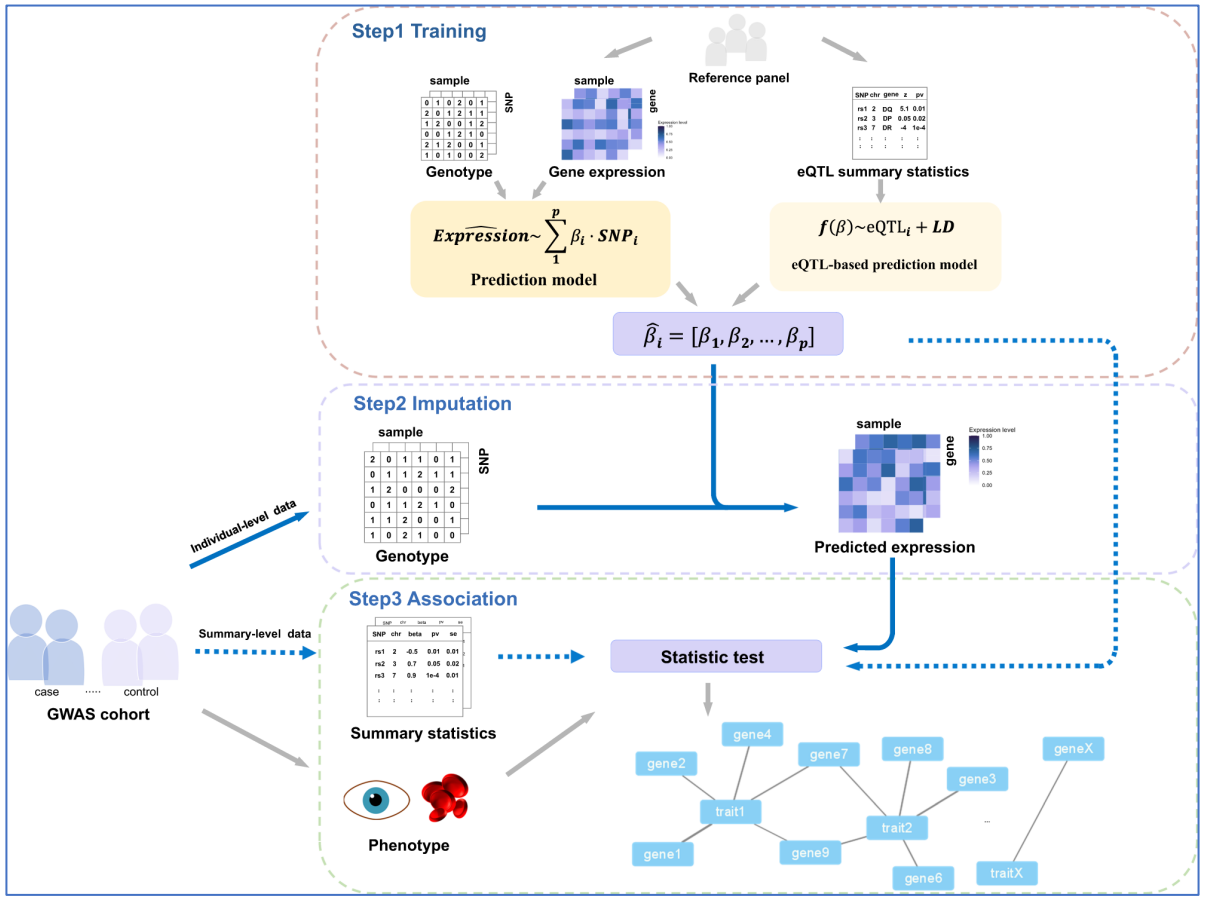


The GTEx Consortium , Science369,1318-1330(2020)

“General Steps” of a Transcriptome-Wide Association Study (TWAS)

- Define **genotype** and **phenotype**
- Step 1: Use **reference transcriptome / prediction models with regulatory weights** to train genotype -> expression
- Step 2: Impute genotype change or summary statistics -> predicted expression change
- Step 3: Associate imputed expression changes with phenotype of interest

RESULT: Gene-phenotype associated mediated through the transcriptome



Mai, J., Lu, M., Gao, Q. et al. Commun Biol 6, 899 (2023).

Prediction Model: Basic Design

Intercept term

Error term

$$E_g = \mu + X_{np} * \beta + \varepsilon$$

Expression level

X: n*p genotype matrix
p-vector SNPs
n individuals

β: p-vector of
SNP weights based on eQTL
effect sizes

Mai, J., Lu, M., Gao, Q. et al. Commun Biol 6, 899 (2023).

Association Testing: Basic Design

- Gene-trait (per tissue) associations
  - Linear regression
  - Logistic regression
  - Cox regression
  - Spearman model

“1st Generation” TWAS Software

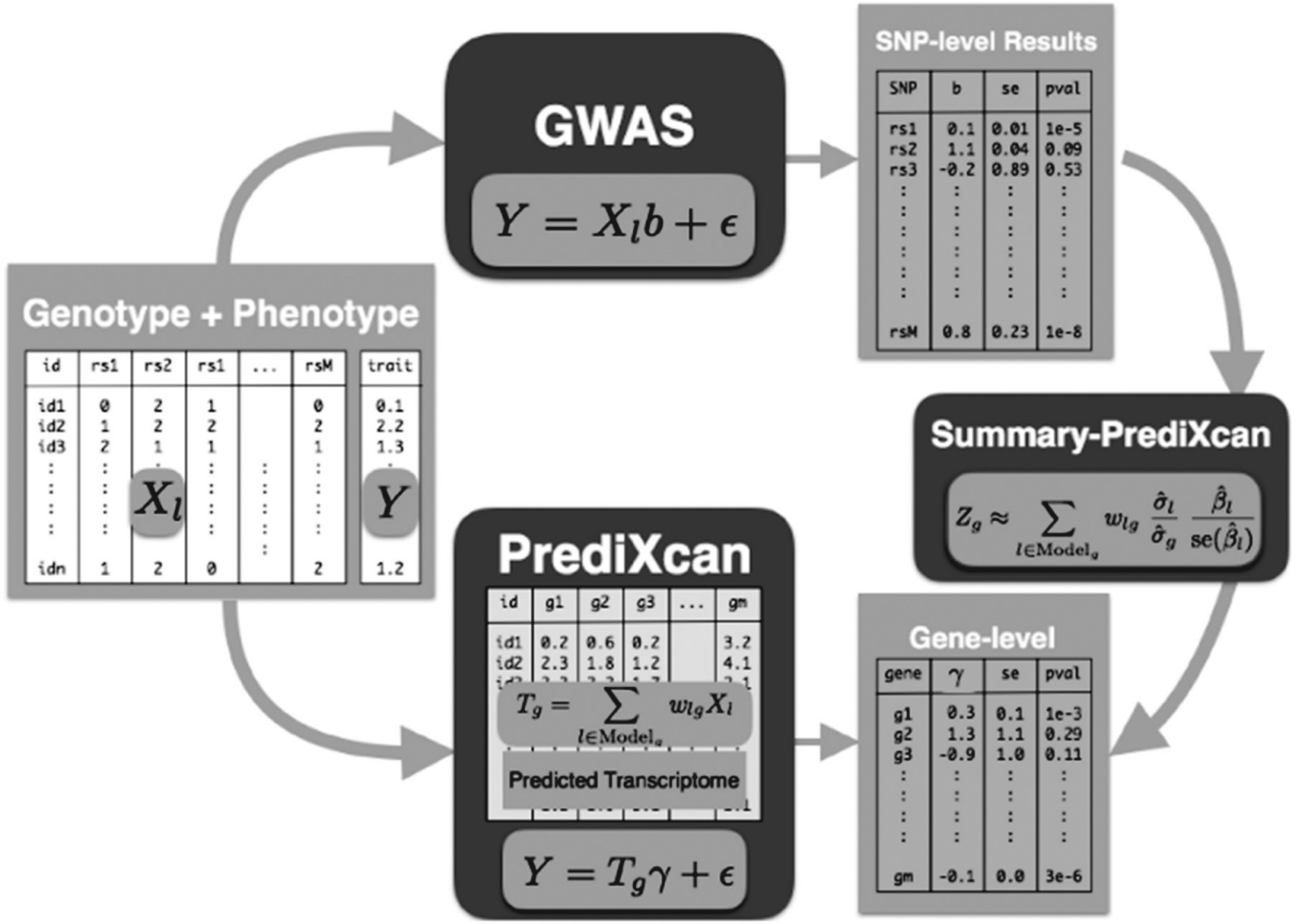
PrediXcan -> S-PrediXcan

- PredictDB for prediction models (leverages lasso/elastic net)
- *Cis*-eQTLs
- Individual tissues
- Imputation / Summary statistics

FUSION

- Bayesian sparse linear mixed model (BSLMM) for prediction models
- *Cis*-eQTLs
- Individual tissues
- Summary statistics

PrediXcan/S-PrediXcan Framework



Barbeira, A.N., Dickinson, S.P., Bonazzola, R. et al. Nat Commun 9, 1825 (2018).

“2nd Generation” TWAS Software

MultiXcan / S-MultiXcan

- PredictDB
- *Cis*-eQTLs
- Across tissues with an F test
- Summary statistics

UTMOST

- Across tissue prediction model with Lasso, Berk-Jones test
- *Cis*-eQTLs
- Summary statistics

TIGAR

- Non-parametric Dirichlet process regression for prediction models
- *Cis*-eQTLs
- Individual tissues
- Imputation

“3rd Generation” TWAS Software

MOSTWAS: Multi-Omics

- Chromatin status, transcription factor binding sites, microRNA effects, and CpG methylation also included
- *Cis*- and *trans*-eQTLs
- Individual tissues
- Summary statistics

Multi-Ancestry

- MATS
- TESLA

Why Perform a TWAS?

- Identifies genes regulated by disease-associated variants
- Gene-level association to phenotype of interest =>
  - Statistical power (improves multiple testing burden of variants)
  - Easier experimental validation (gene changes vs. specific variant changes)
- Lower computing complexity vs. GWAS
- Tissue effect specificity

Limitations / Caveats

- Accuracy of TWAS Depends On:
  - Prediction Models!!
  - Underlying GWAS assumptions

- Correlated prediction: Contamination from linkage disequilibrium
  - Fine mapping and colocalization => **Return back to lecture 3!**

- Correlated expression: True co-regulation

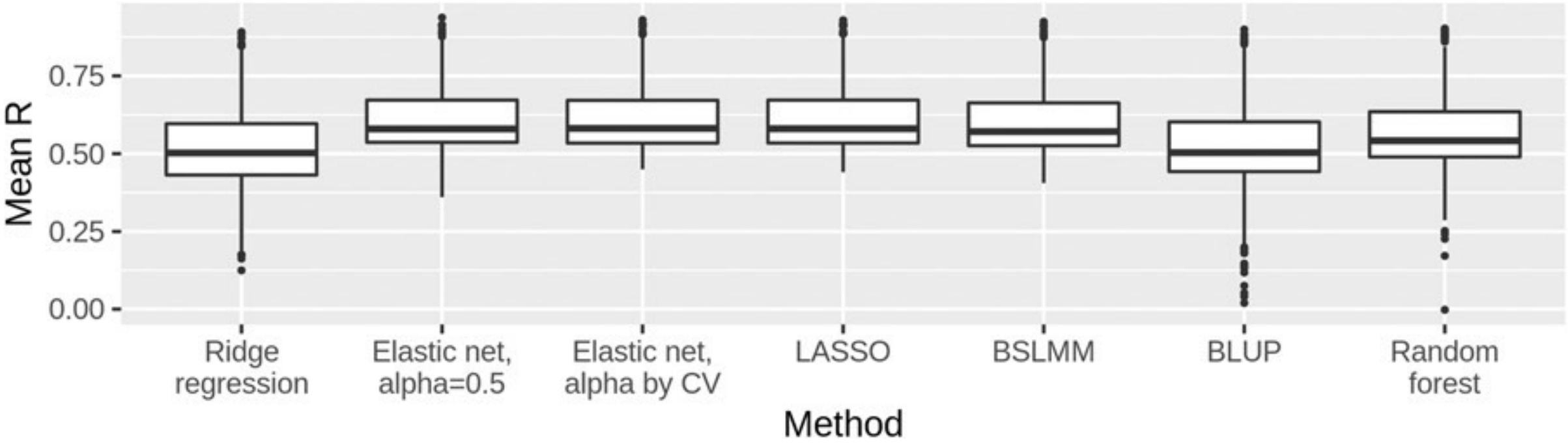
Accuracy of TWAS Approaches: Prediction Models

- Prediction model ancestry needs to correspond to TWAS population ancestry or be multi-ancestry

- Focus on *cis*-eQTLs biases against *trans*-eQTLs

- SUMMIT: Estimates cis-eQTL effect size (weights for prediction models) with eQTL summary statistics, LD information w/ shrinkage estimator
- Multi-ancestry modeling

Accuracy of TWAS Approaches: Prediction Models – Statistical Design



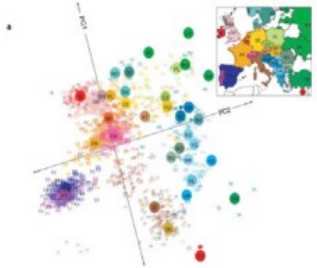
Fryett JJ, Morris AP, Cordell HJ.. Genet Epidemiol. 2020 Jul;44(5):425-441.

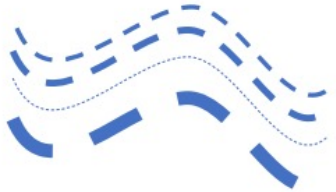
Accuracy of TWAS Approaches: Relation to GWAS

SNPs → Genetic component of expression ⇢ Trait

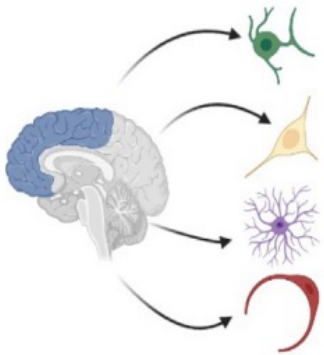
TGCTAGA
GAGACAT
CTAGAGT
TCATTGT

CHALLENGES:
- Genetic ancestry
- Population structure
- Linkage disequilibrium



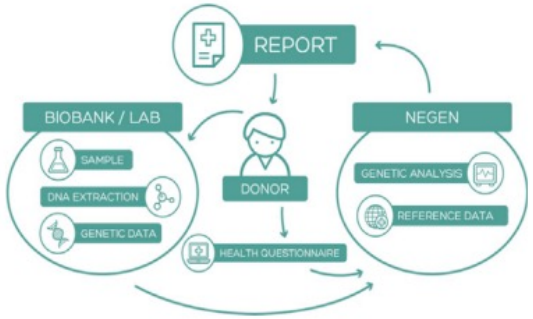


CHALLENGES:
- Tissue specificity
- Cell-type- and state specificity
- Dynamic trajectories



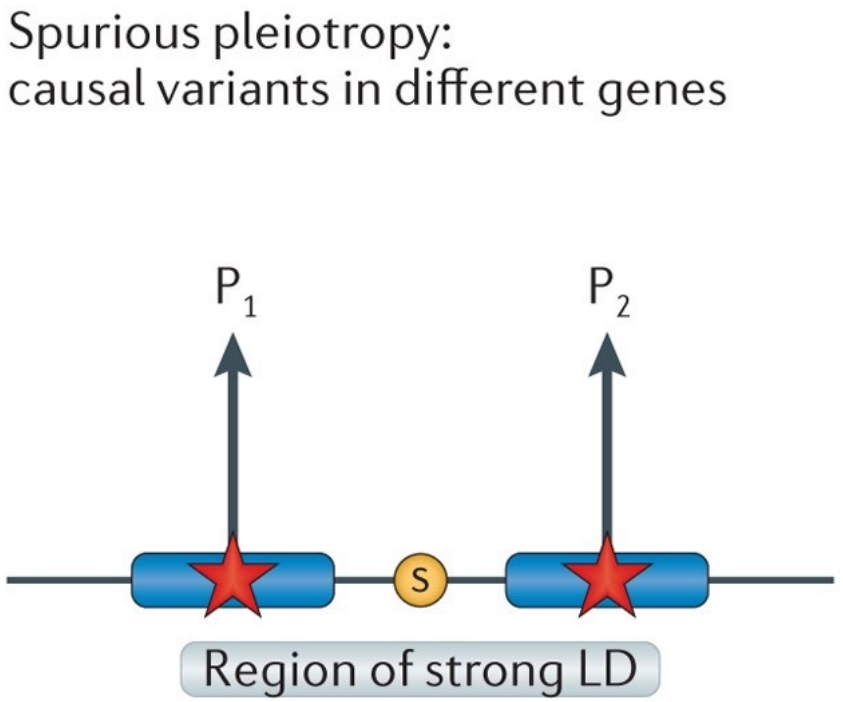
Case
Control
Control
Case

CHALLENGES:
- Phenotype acquisition
- Case-control definition
- Ascertainment bias



Bhattacharya A, et al. Cell Genom. 2022 Oct 12;2(10):100180.

Limitations: LD Contamination



MUST USE: Fine mapping and colocalization to resolve

Solovieff, N., Cotsapas, C., Lee, P. et al.. Nat Rev Genet 14, 483–495 (2013).

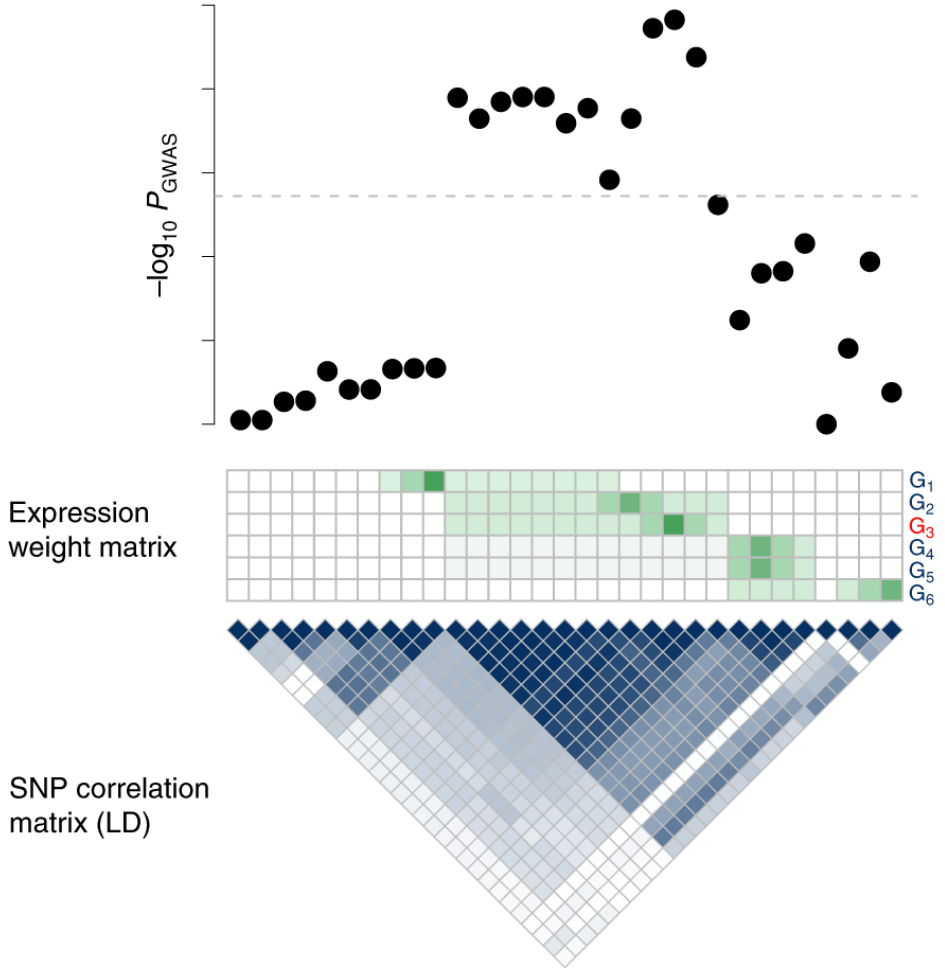
Revisiting: Fine Mapping and Colocalization (in TWAS)

Fine Mapping

- Goal: To refine more likely causal genes within a given LD block
- Prior fine mapping tools could be applied (e.g. eCAVIAR, DAP-G, SuSIE-R)

Colocalization

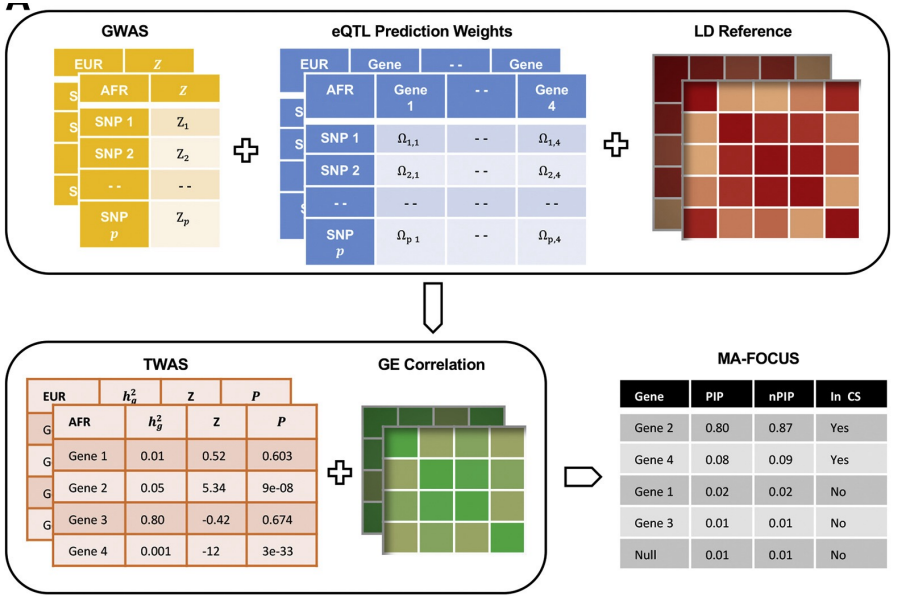
- Goal: To limit effect of LD contamination by prioritizing most likely signal using posterior probability estimates based on regional structure
- Prior colocalization tools can be applied (coloc, enloc, eCAVIAR)

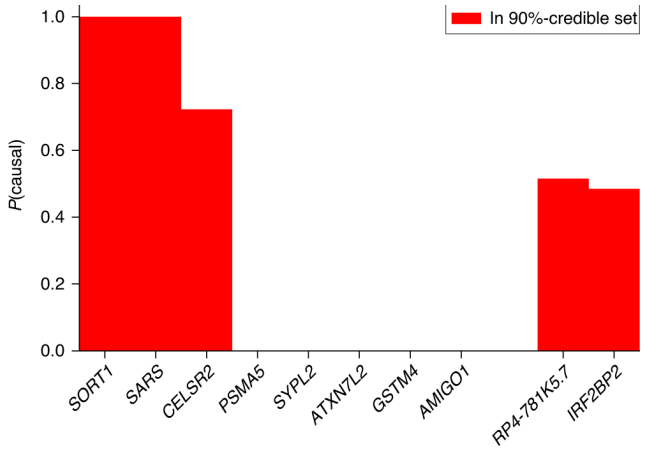


Mancuso N, et al. Nat Genet. 2019 Apr;51(4):675-682.

FOCUS/MA-FOCUS: TWAS-Specific Estimates of Causal Probabilities

- Models predicted expression correlations and uses them to assign genes posterior probabilities of causality (MA = Multi-Ancestry)
- **Input:** GWAS summary data, expression prediction weights, LD among SNPs in region
- **Output:** Probability for each gene in region to explain TWAS signal (PIP and CS from fine-mapping)

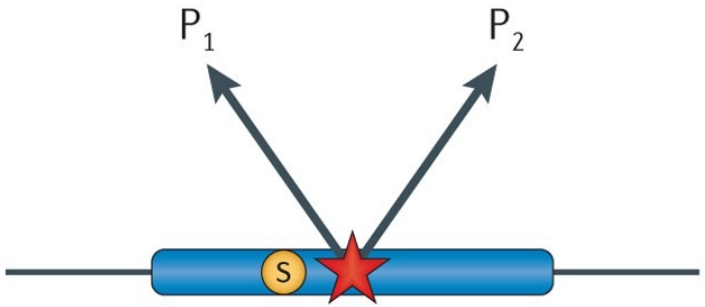




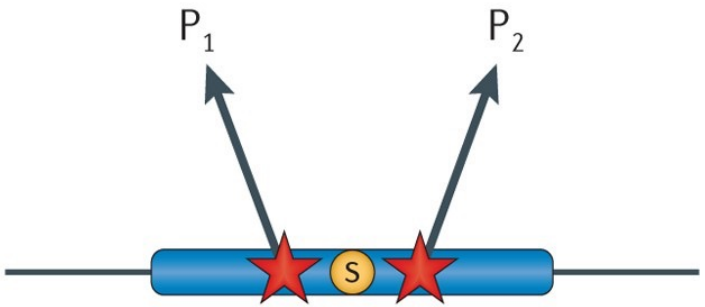
Mancuso N, et al. Nat Genet. 2019 Apr;51(4):675-682. AND Lu Z, et al. Am J Hum Genet. 2022 AND Wainberg M, et al. *Nat Genet*. 2019

Limitations: True Co-Regulation

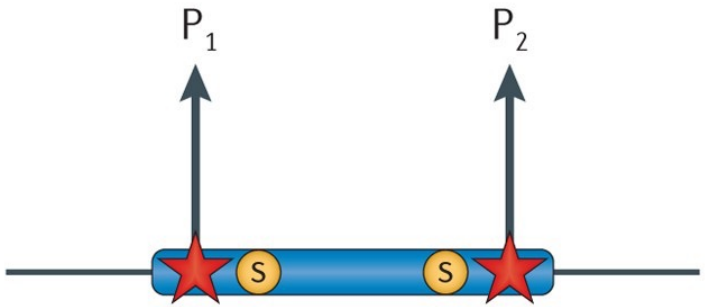
Biological pleiotropy: single causal variant



Biological pleiotropy: different causal variants colocalizing in same gene and tagged by the same genetic variant



Biological pleiotropy: different causal variants colocalizing the same gene



Solovieff, N., Cotsapas, C., Lee, P. et al.. Nat Rev Genet 14, 483–495 (2013).

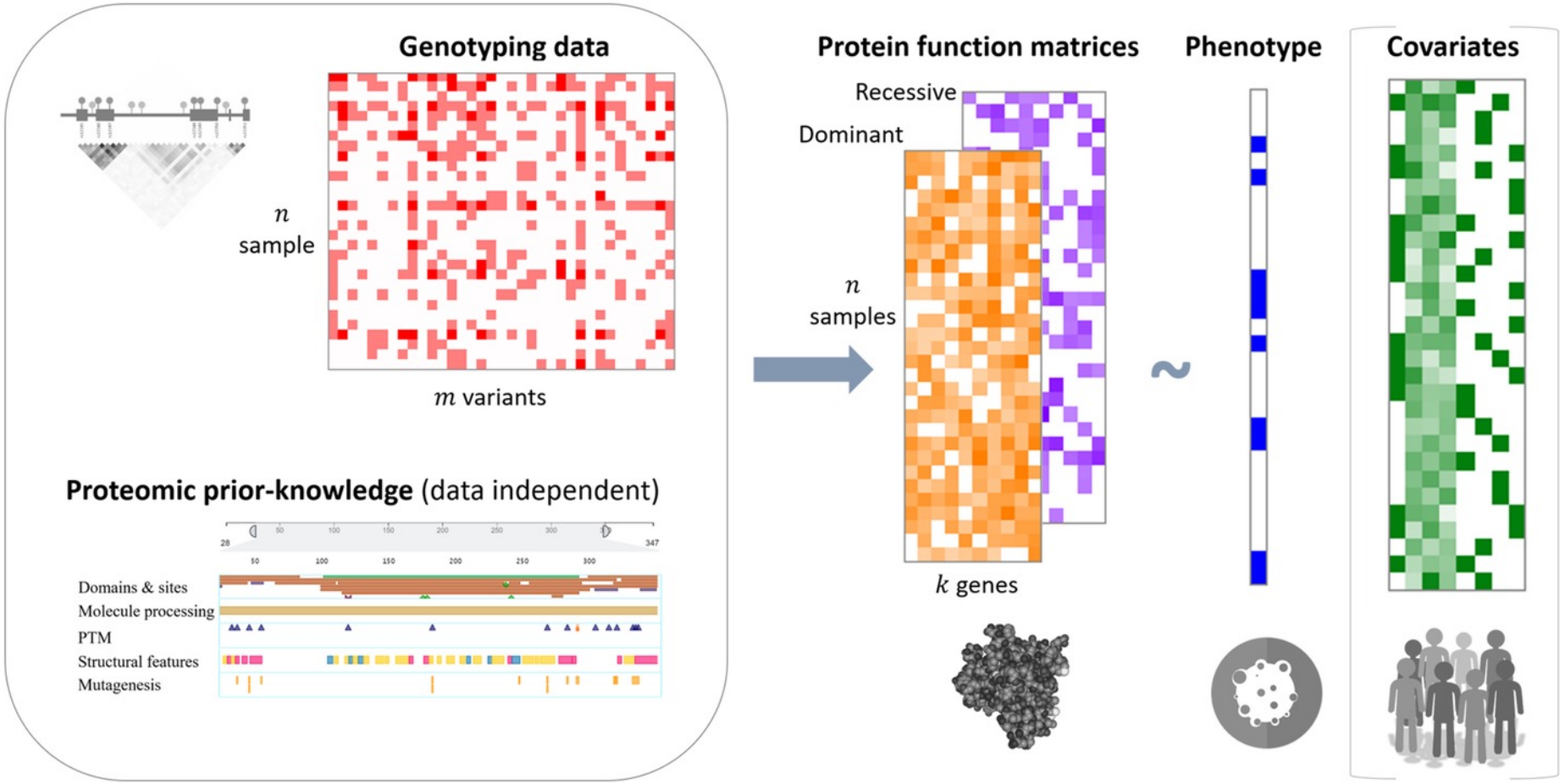
Resources: Reference Prediction Models / Datasets

- GTEx: https://gtexportal.org/home/
- eQTLGen Consortium: https://www.eqtlgen.org/
- Predict DB: https://predictdb.org/
- FUSION Prediction Models: http://gusevlab.org/projects/fusion/#gtex-v8-multi-tissue-expression

Resources: TWAS Databases

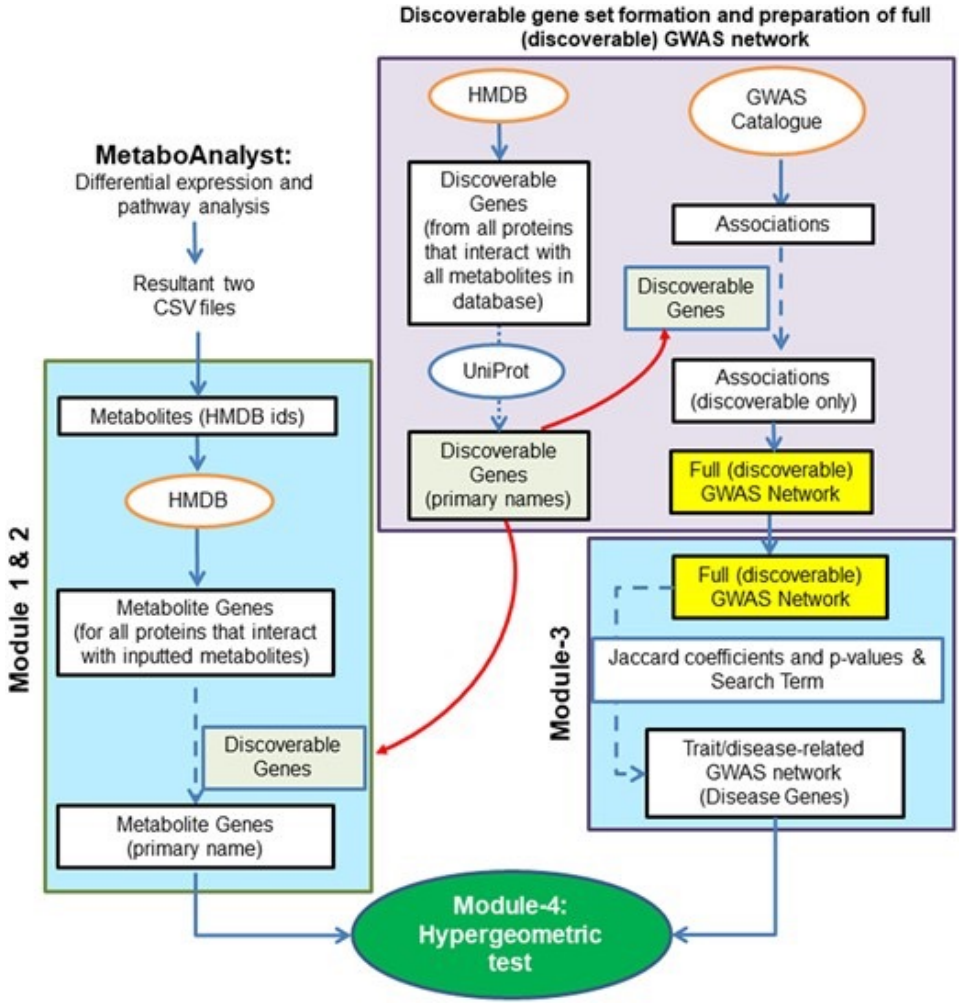
- http://twas-hub.org/
- http://www.webtwas.net/
- https://ngdc.cncb.ac.cn/twas/

Other “-WAS” Approaches: Proteome-WAS



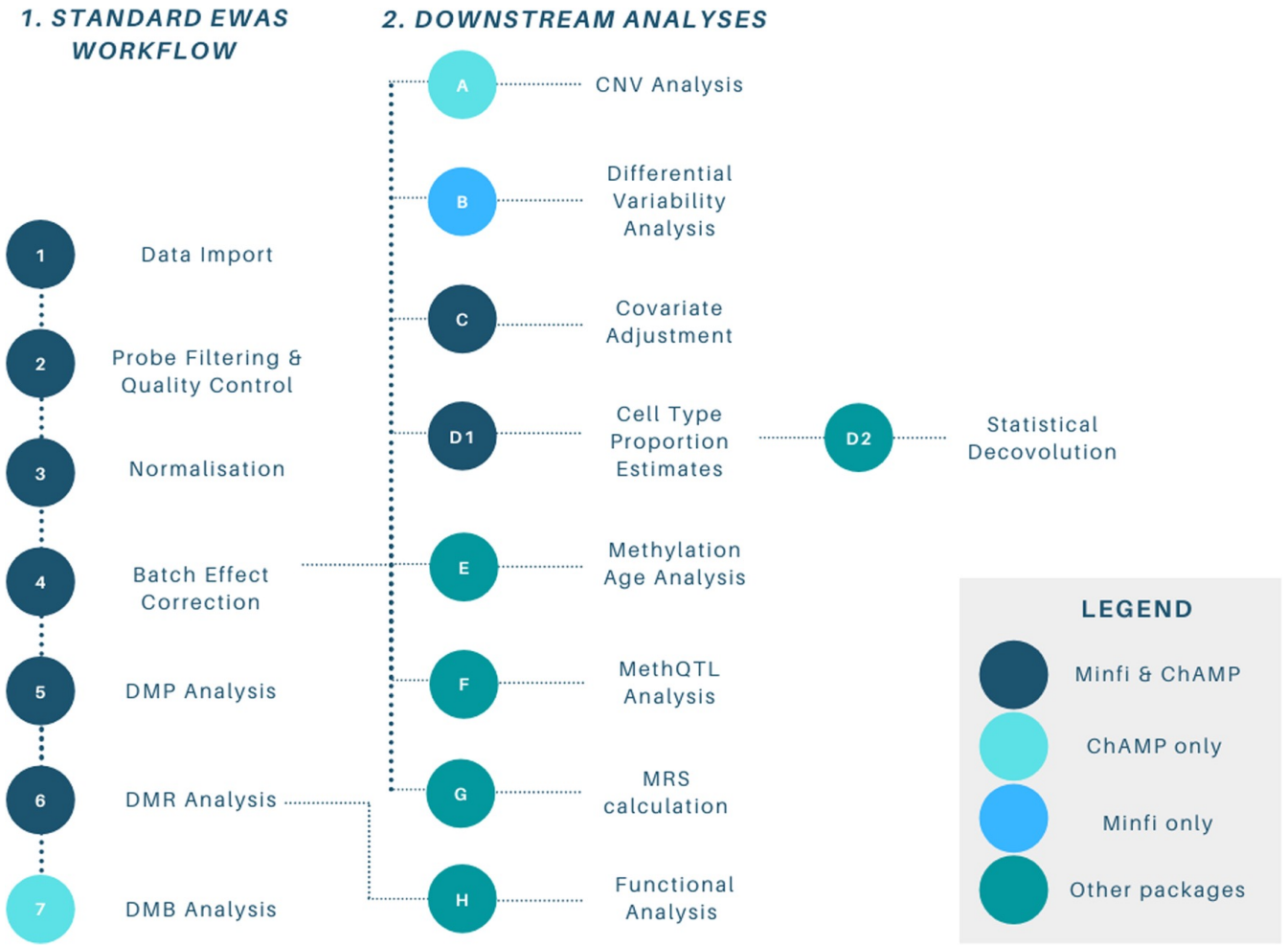
https://github.com/nadavbra/pwas; http://nilanjanchatterjeelab.org/pwas/

Other “-WAS” Approaches: Metabo-WAS



- https://github.com/saifurbd28/metGWAS-1.0

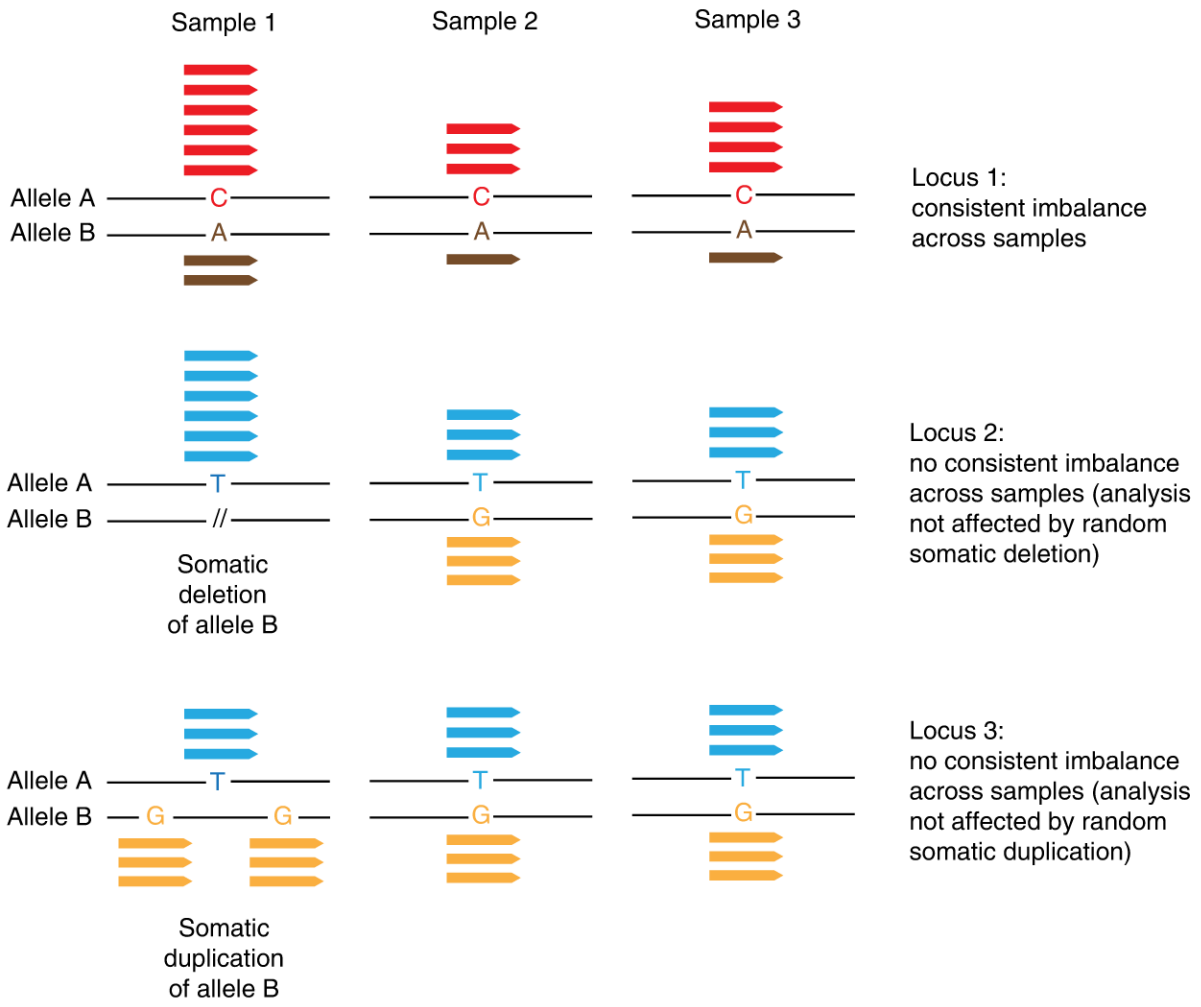
Other “-WAS” Approaches: Epigenome-WAS



Campagna, M.P., Xavier, A., Lechner-Scott, J. et al. Clin Epigenet 13, 214 (2021).

Other “-WAS” Approaches: Regulome-WAS

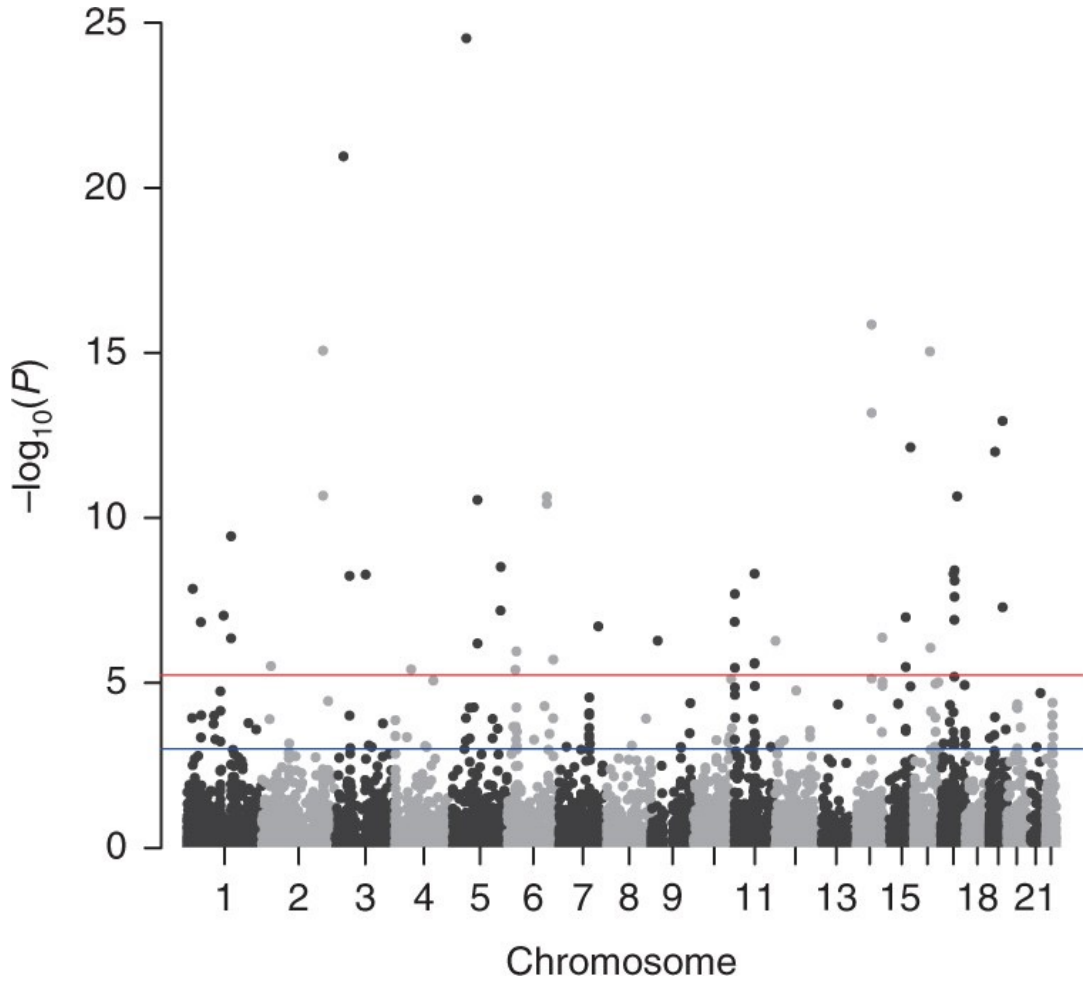
- Leverages ATAC-Seq samples for allele-specific accessible eQTLs



Grishin D, Gusev A. Nat Genet. 2022 Jun;54(6):837-849.

Applications in Practice: Breast Cancer Risk

Region	Gene
1p34.1	ZSWIM5
3p24.1	LRRC3B
4q12	SPATA18
6p22.1	UBD
7p32.2	KLHDC10
9p21.3	MIR31HG
11p15.5	RIC8A
11q13.2	B3GNT1
11q13.2	RP11-867G23.10
12q13.33	RP11-218M22.1
14q24.1	GALNT16
14q24.1	PLEKHD1
15q24.2	MAN2C1
15q24.2	CTD-2323K18.1



Wu, L., Shi, W., Long, J. et al. Nat Genet 50, 968–978 (2018).

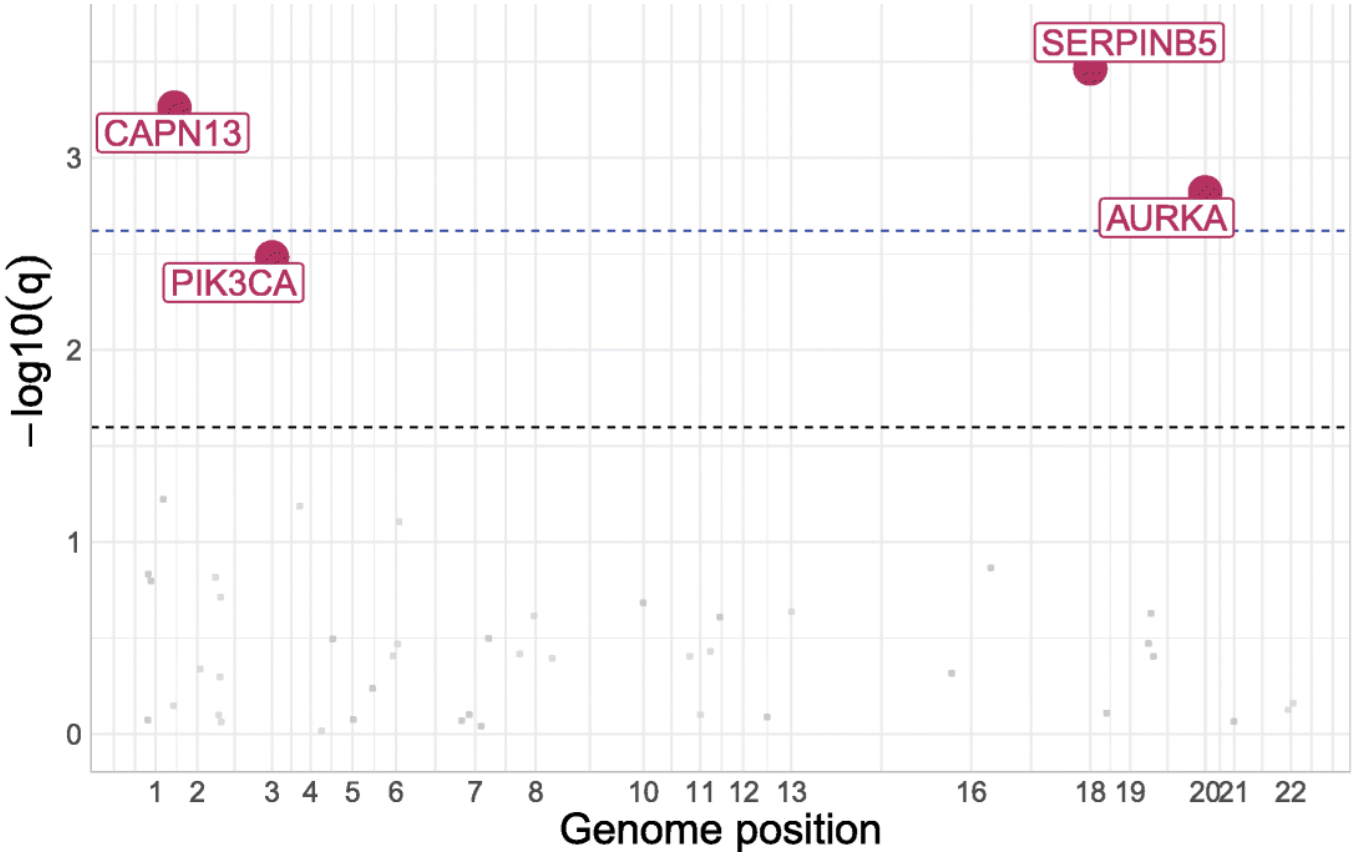
Applications in Practice: Breast Cancer Risk - Splicing-Associated Variants

- Conditional joint TWAS across multiple tissues
- At least 1MB from prior GWAS sites and noncoding -> potential splicing related variants = Greater likelihood of association with splicing

Bhattacharya, A., García-Closas, M., Olshan, A.F. et al. Genome Biol 21, 42 (2020).

Applications in Practice: Breast Cancer Risk - Ancestry-Specific

- Race-stratified predictive models of germline genotype -> tumor expression in 406 genes related to breast cancer
- Models not applicable across race / subtype
- **Identified associations in Black women near AURKA, CAPN13, PIK3CA, and SERPINB5 via TWAS that are underpowered in GWAS**



Bhattacharya, A., García-Closas, M., Olshan, A.F. et al. Genome Biol 21, 42 (2020).

Overview

- Transcriptome-Wide Association Studies (TWAS)
  - “General Steps”
  - Software
  - Why Perform a TWAS?
  - Limitations / Caveats
- Resources
- Other “–WAS” Approaches
  - pWAS
  - Metabo-WAS
  - Epigenome-WAS
  - Regulome-WAS
- Applications in Practice – Breast Cancer Risk