Leveraging co-expression between genes to identify gene sets that are enriched for disease heritability

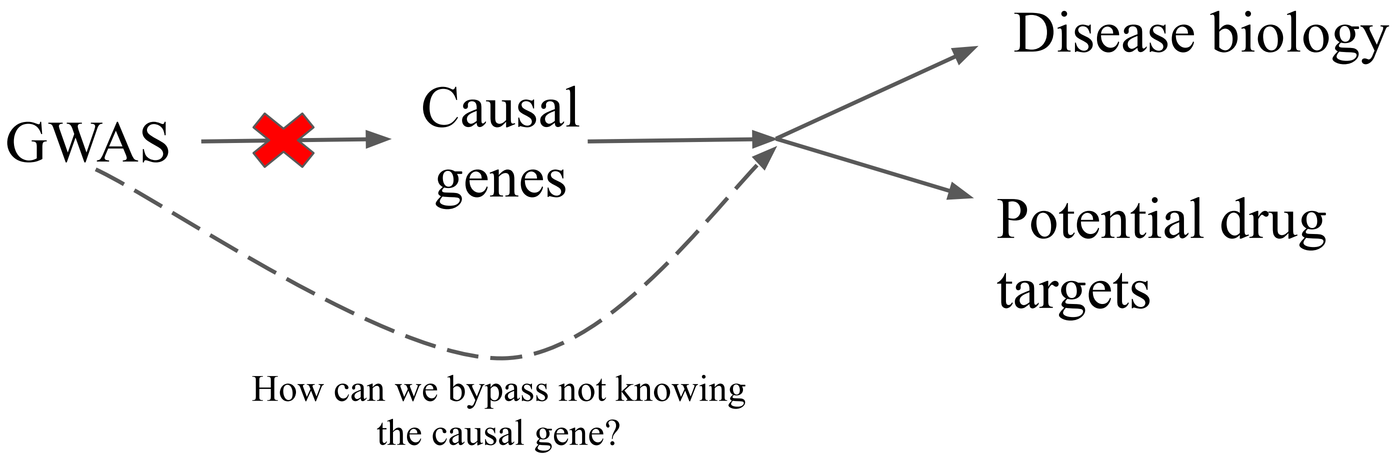
Katie Siewert
Post-doc, group of Alkes Price
Harvard T.H. Chan School of Public Health
5/13/2020

Learning from GWAS

GWAS → Causal genes → Disease biology

GWAS → Causal genes → Potential drug targets

Learning from GWAS

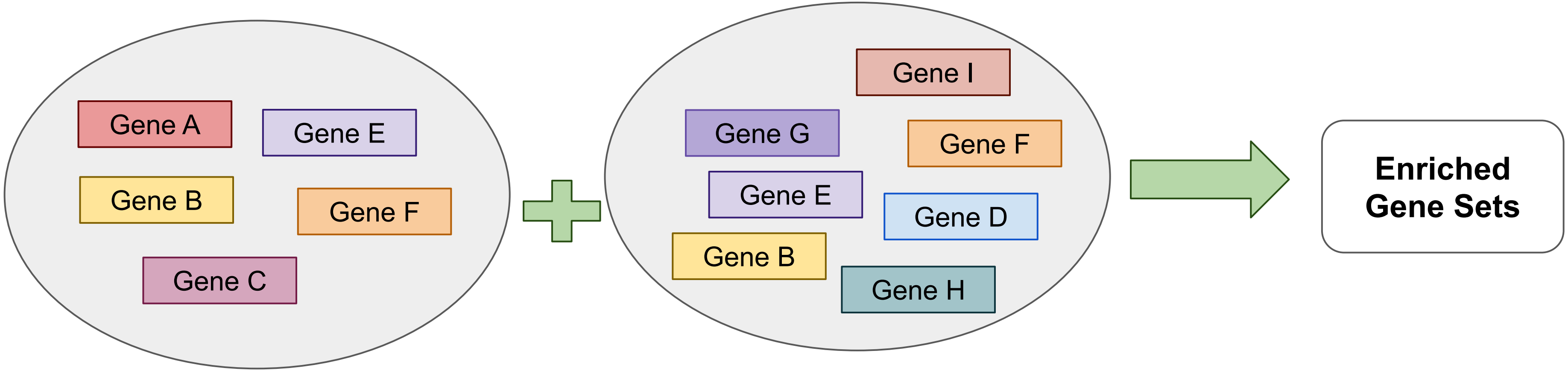


Causal genes are often unknown.

Gene set analysis

Potentially causal genes
(often determined via distance from GWAS peak)

Gene set(s) or pathway(s)
(externally sourced)



Example Methods: DEPICT (Pers 2015) , MAGMA (de Leeuw 2015), S-LDSC (Kim 2019)

de Leeuw NRG 2016, Feng 2019 NG

Gene set analysis: Loss of power

- Nearest gene is causal in only ~50% of cases (Gamazon 2018)
  - Causes noise in gene set analysis ⇒ Reduces power

Gene set analysis: Loss of power

- Nearest gene is causal in only ~50% of cases (Gamazon 2018)
    - Causes noise in gene set analysis ⇒ Reduces power

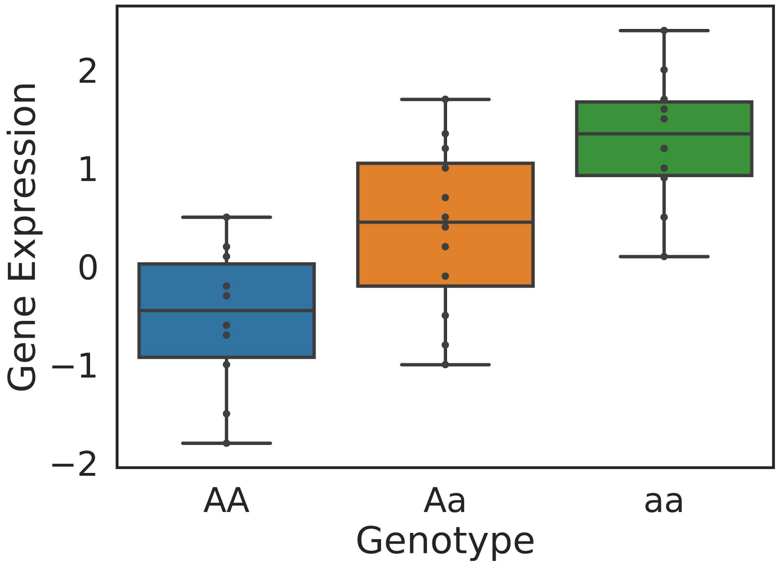
Can we do better than nearest gene approaches?

Outline

- Background
  - eQTLs, TWAS and gene co-expression
- Our approach for gene set enrichment: Gene Co-expression Score Regression (GCSC)
- Results
  - Simulations
  - Enriched gene sets

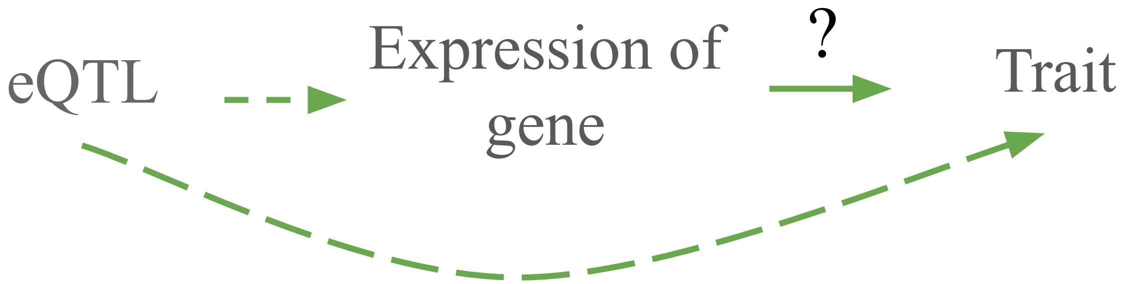
eQTLs: SNPs associated with a gene's expression

- eQTLs identified by measuring gene expression levels in genotyped individuals
    - e.g. the GTEx project



GTEx Consortium 2017 NG

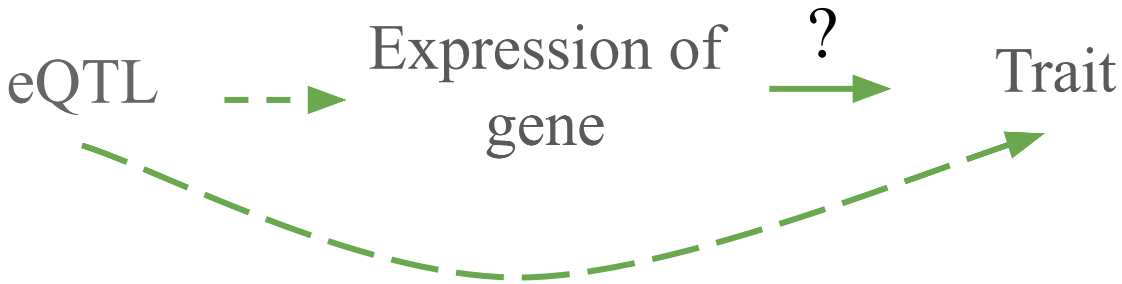
eQTLs: SNPs associated with a gene's expression



- eQTLs can inform causal genes at GWAS loci

Lappalainen 2013 Nature, Wainberg 2019 NG

eQTLs: SNPs associated with a gene's expression



- eQTLs can inform causal genes at GWAS loci
- However, not proof of causality
  - eQTL could also be regulating other genes, maybe in other tissues or conditions

Lappalainen 2013 Nature, Wainberg 2019 NG

TWAS: Transcriptome-wide association study

Tests for association between genetically predicted gene expression & disease

Step 1) Make gene model

- Input: Assayed gene expression & genotypes in same individuals
- Learn: Predictive model of gene expression (weighted combo of SNPs)

TWAS: Transcriptome-wide association study

Tests for association between genetically predicted gene expression & disease

Step 1) Make gene model

- Input: Assayed gene expression & genotypes in same individuals
- Learn: Predictive model of gene expression (weighted combo of SNPs)

Step 2) Test for association

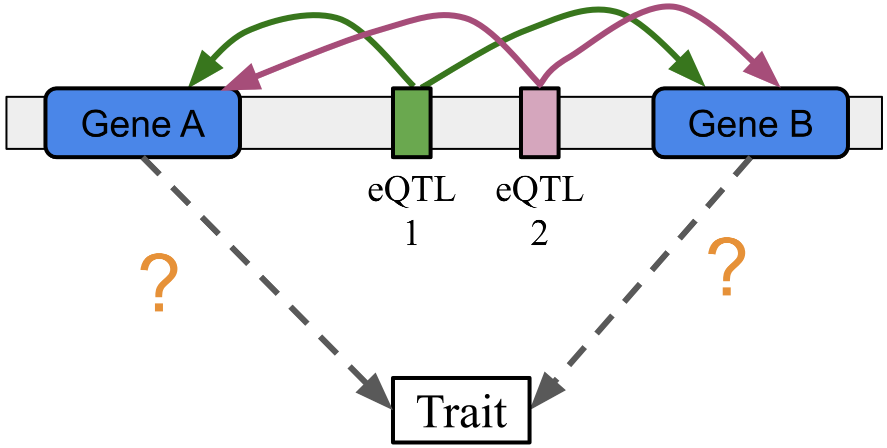
- Input: eQTL weights in gene model & GWAS z-scores
- Tests: Correlation between eQTL weights for a gene and Z_{GWAS}

TWAS pools information across eQTLs

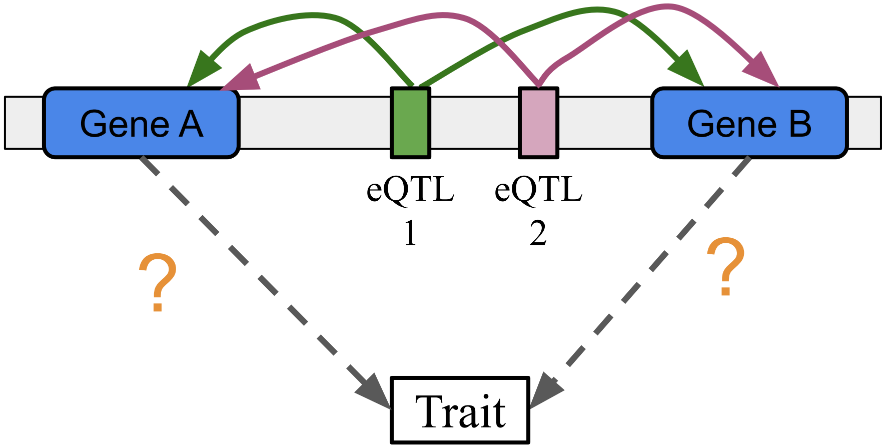
- By pooling info from all eQTLs in a gene model, TWAS looks for consistency in magnitude & direction of effect.

- Result: stronger evidence for gene⇒trait association than looking at a single eQTL

Co-expression can cause multiple association in TWAS



Co-expression can cause multiple association in TWAS



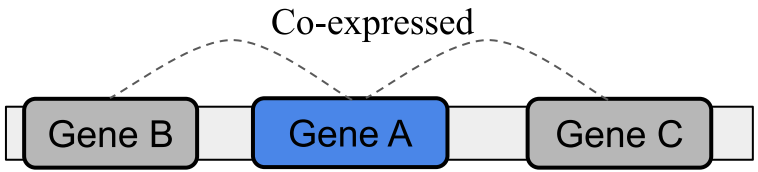
Correlation in predicted gene expression can be caused by:

- Shared causal eQTL(s)
- Causal eQTL(s) in LD
- Errors in gene model

Wainberg 2019 NG

Co-expression increases TWAS associations

Significant χ^2_{TWAS} only if gene A is causal



Significant χ^2_{TWAS} if A, B or C is causal

Co-expression increases TWAS associations

Significant χ^2_{TWAS} only if gene A is causal

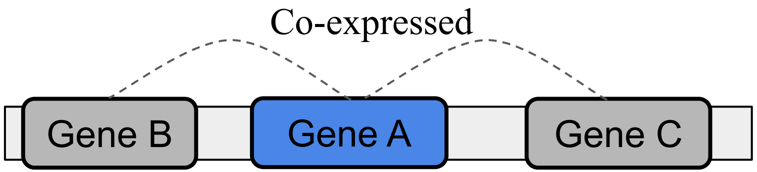
Significant χ^2_{TWAS} if A, B or C is causal

The more genes that a gene is co-expressed with, the more likely its expression is correlated with a causal gene's expression

➢ Increases $E[\chi^2_{TWAS}]$

Co-expression increases TWAS associations

Significant χ^2_{TWAS} only if gene A is causal



Significant χ^2_{TWAS} if A, B or C is causal

Can't simply compare the TWAS χ^2 of genes in a set to other genes

- Co-expression adds noise and confounds

Goal: Develop method to quantify heritability enrichment in gene sets using gene expression

1. eQTLs allow us to more accurately map SNPs to genes
2. TWAS allows us to pool information across eQTLs, looking for directional consistency
3. Is not confounded by co-expression

Co-expression can be calculated using models for gene expression and a reference panel

Gene expression model
weights obtained using feature selection

Reference panel
e.g. 1000 Genomes

Predict expression in reference panel individuals → Compute correlation squared between genes (co-expression) → Co-expression scores: sum of co-expression of gene with all other genes

Gene Co-expression Score Regression regresses on co-expression to estimate heritability

GCSC Regression equation

$$\underbrace{E[\chi^2_{TWAS_j}]}_{\text{Gene expression->trait association}} = N\left(\tau_b c_{j,b} + \tau_s c_{j,s}\right) + 1$$

Terms we calculate ahead of time:

Gene expression->trait association

Co-expression score with all genes

Co-expression score with genes in gene set

Gene Co-expression Score Regression regresses on co-expression to estimate heritability

<u>GCSC Regression equation</u>

Terms we estimate using GCSC:

Heritability explained by predicted gene expression

Additional heritability explained by genes in the gene set

$$\underbrace{E[\chi^2_{TWASj}]}_{\text{Gene expression->trait association}} = N(\tau_b c_{j,b} + \tau_s c_{j,s}) + 1$$

Terms we calculate ahead of time:

GWAS Sample Size

Co-expression score with all genes

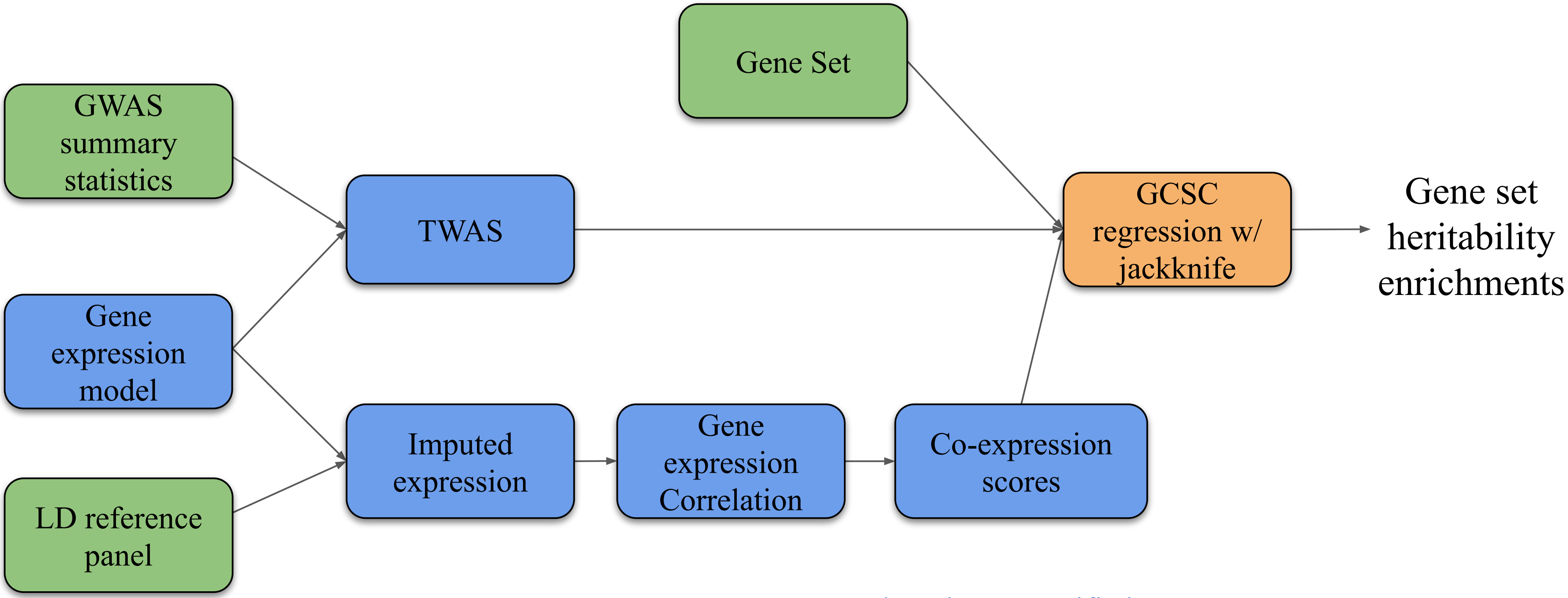
Co-expression score with genes in gene set

Gene Co-expression Score Regression is analogous to stratified LD score regression

- Stratified LD score regression: GWAS χ^2 are regressed against LD scores for an annotation to estimate heritability explained by the annotation

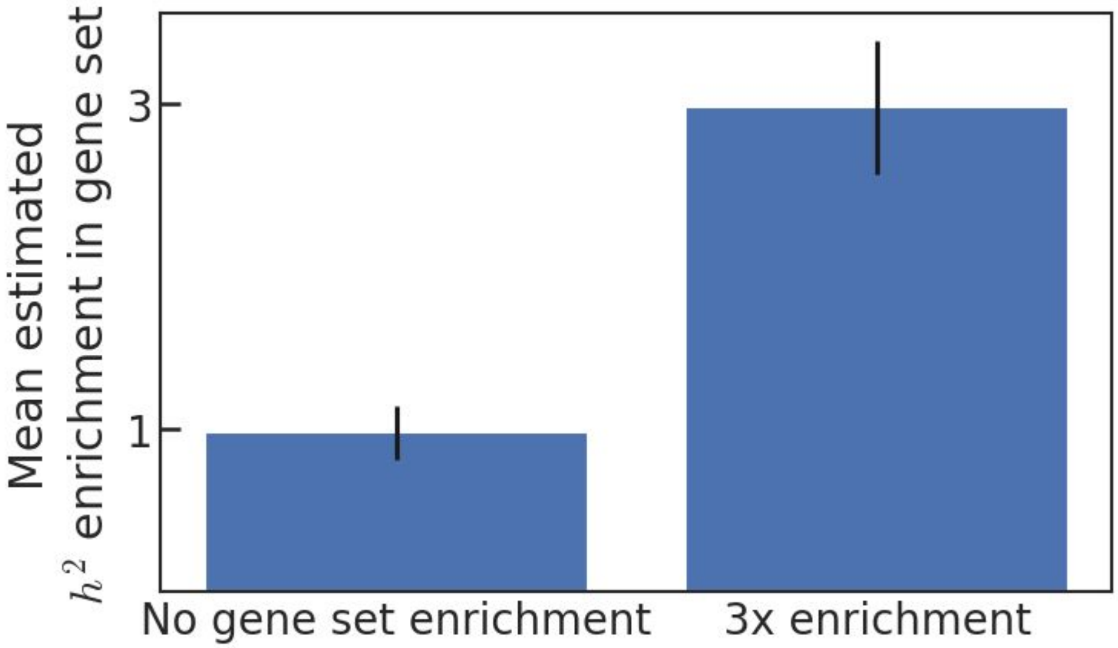
- GCSC regression: TWAS χ^2 are regressed against co-expression scores for a gene set to estimate heritability explained by predicted expression of genes in a set

GCSC pipeline



Blue: tissue-specific inputs
Green: not related to tissue
Can perform tissue-specific, or combined tissue GCSC

GCSC simulation results



Error bars denote +- 2 s.e.

Appropriate standard errors also verified

Gene sets enriched for heritability using all tissue GCSC

- Tested 59 gene sets, results are meta-analyzed across 44 independent traits
- Significant (after Bonferroni) enrichments include:
  - LoF constraint genes: ExAC pLI genes (1.2x enrichment P:7.5e-34)
  - Olfactory receptors (0.13x, 1.3e-33)
  - Top decile of genes with most LD-independent SNPs (0.68x, 7.8e-22)
  - Essential genes (1.14x , 6.5e-19)
  - High Enhancer domain score genes (1.13x, 8.4e-14)
  - High protein-protein closeness centrality genes (1.1x , 1.3e-11)
  - Haploinsufficient genes (1.4x, 1.5e-8)
  - eQTL deficient genes (0.82x, 1.1e-7)
  - Educational and developmental disorder genes (1.1x, 5.2e-4)

Gene sets enriched for heritability using all tissue GCSC

- Tested 59 gene sets, results are meta-analyzed across 44 independent traits
- Significant (after Bonferroni) enrichments include:
  - LoF constraint genes: ExAC pLI genes (1.2x enrichment P:7.5e-34)
  - Olfactory receptors (0.13x, 1.3e-33)
  - Top decile of genes with most L
  - Essential genes (1.14x , 6.5e-19)
  - High Enhancer domain score ge
  - High protein-protein closeness c
  - Haploinsufficient genes (1.4x, 1.5e-8)
  - eQTL deficient genes (0.82x, 1.1e-7)
  - Educational and developmental disorder genes (1.1x, 5.2e-4)

<u>Compare to s-LDSC approaches using:</u>

- fine-mapped eQTLs (Hormozdiari 2018 NG): p-value 4.9e-17
- eQTL effect sizes (Yao et al 2020 bioRXiv): p-value 2.3e-25

GCSC finds enrichment of gene specifically expressed in blood and immune cell types in Alzheimer's

Tested for enrichment of heritability explained in 44 traits for specifically expressed genes in 53 tissues

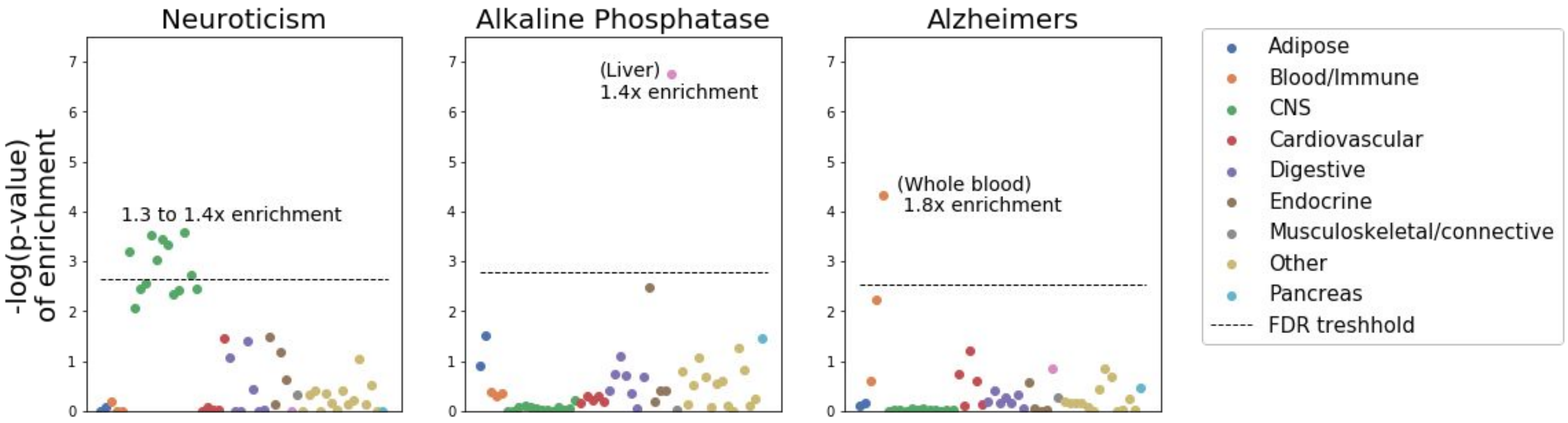
-Found 118 significantly enriched trait/tissue combinations

Gene sets from Finucane 2018 NG

GCSC finds enrichment of gene specifically expressed in blood and immune cell types in Alzheimer's

Tested for enrichment of heritability explained in 44 traits for specifically expressed genes in 53 tissues

-Found 118 significantly enriched trait/tissue combinations

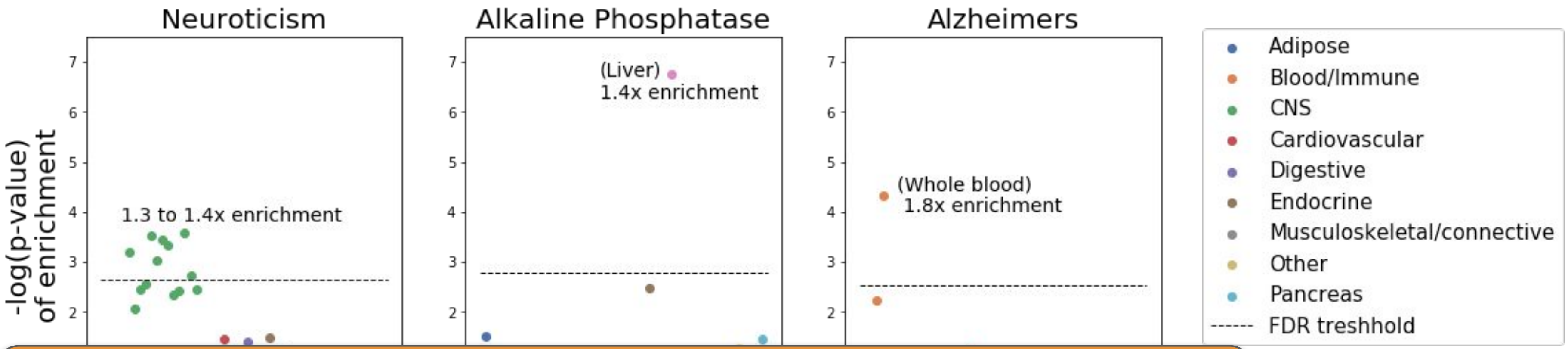


Gene sets from Finucane 2018 NG

GCSC finds enrichment of gene specifically expressed in blood and immune cell types in Alzheimer's

Tested for enrichment of heritability explained in 44 traits for specifically expressed genes in 53 tissues

-Found 118 significantly enriched trait/tissue combinations



Corroborates findings that expression of immune and blood genes play role in Alzheimer's (Gjoneska 2015 Nature, Sims 2017 NG)

Gene sets from Finucane 2018 NG

Conclusions

GCSC (Gene Co-expression Score Regression) for gene set enrichment

- Uses TWAS to detect sets of genes whose expression is enriched for trait heritability

- Found large number of heritability enrichments, including specifically expressed genes

Acknowledgements

Alkes Price

Huwenbo Shi

Doug Yao

Omer Weissbrod

Sam Kim

Sasha Gusev

Price group

Funding: NIH NIDKK T32

Thank you!