# Professor Manichaikul Rotation Presentation

Yogindra Raghav
08/10/2022

UNIVERSITY of VIRGINIA
SCHOOL OF MEDICINE

# Overview

- Colocalization analysis
- Matching colocalized variants
- Some genes of interest
- Modeling COPD-related metrics with expression data and covariates
- Outputs from aforementioned model
- Future ideas
- GitHub Repository:

  https://github.com/YogiOnBioinformatics/Manichaikul-PhD-Rotation

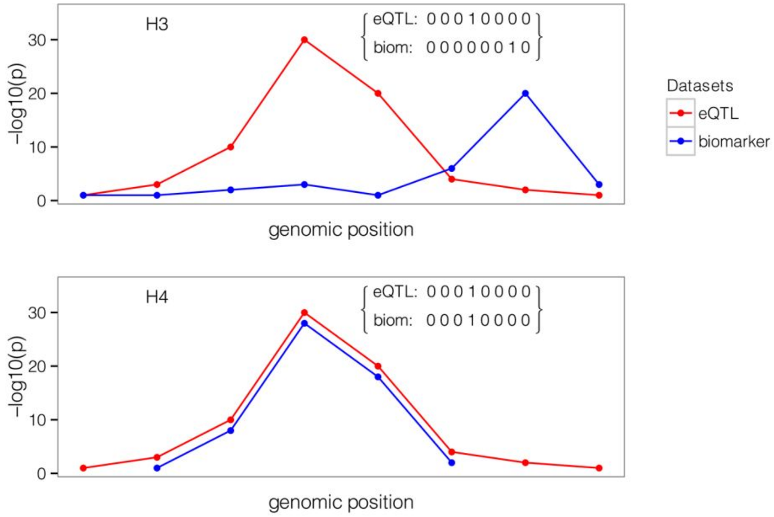
# Overview

- **Colocalization analysis**
- Matching colocalized variants
- Some genes of interest
- Modeling COPD-related metrics with expression data and covariates
- Outputs from aforementioned model
- Future ideas
- GitHub Repository:

  https://github.com/YogiOnBioinformatics/Manichaikul-PhD-Rotation

# Background: Colocalization Analysis

- Xiaowei ran colocalization analysis:
  - Understanding whether variants in linkage disequilibrium BOTH causally affect a COPD-related metric and some biological trait (e.g. expression of gene)

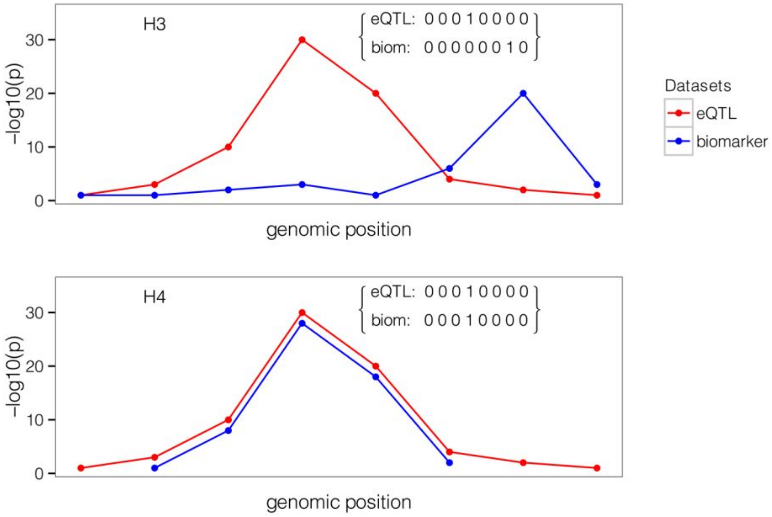


SOURCE: doi.org/10.1371/journal.pgen.1004383

# Background: Colocalization Analysis

- Xiaowei ran colocalization analysis:
  - Understanding whether variants in linkage disequilibrium BOTH causally affect a COPD-related metric and some biological trait (e.g. expression of gene)

- Analysis was ran for each of protein, expression, and methylation QTLs



SOURCE: doi.org/10.1371/journal.pgen.1004383

# Overview

- Colocalization analysis
- **Matching colocalized variants**
- Some genes of interest
- Modeling COPD-related metrics with expression data and covariates
- Outputs from aforementioned model
- Future ideas
- GitHub Repository:

  https://github.com/YogiOnBioinformatics/Manichaikul-PhD-Rotation

# Colocalized Variants Overlap Between Omics

- Colocalized variants that affected (ALL of the following):
  - gene’s expression
  - methylation site
  - protein expression

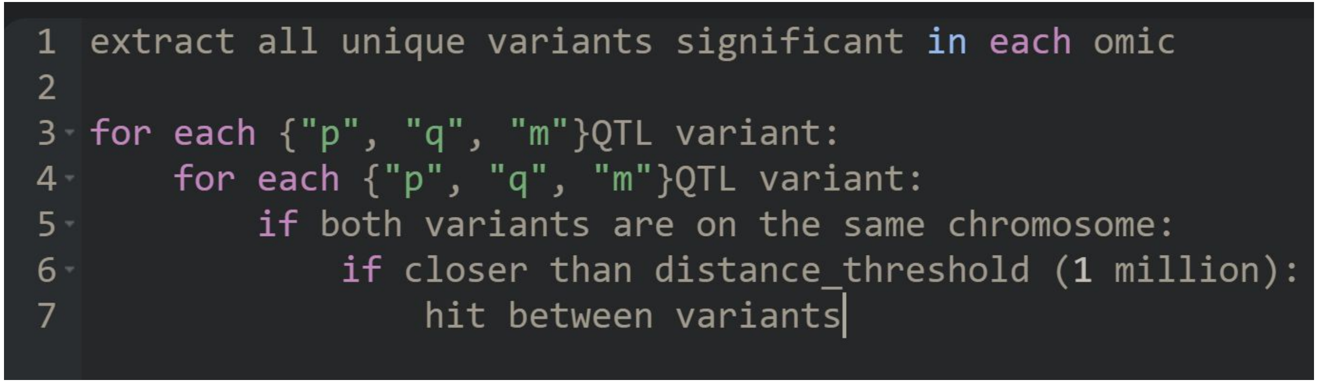
# Colocalized Variants Overlap Between Omics

- Colocalized variants that affected (ALL of the following):
  - gene’s expression
  - methylation site
  - protein expression

- chr6_31896897_T_C_b38 affects:
  - HLA-DQA2
  - SL003680
  - Multiple CpG sites
  - Ani mentioned this is under active investigation.

# Using Fuzzy Matching for Overlapping Between Omics

- Simple algorithm:



```
extract all unique variants significant in each omic

for each {"p", "q", "m"}QTL variant:
    for each {"p", "q", "m"}QTL variant:
        if both variants are on the same chromosome:
            if closer than distance_threshold (1 million):
                hit between variants
```

- Notebook source: https://github.com/YogiOnBioinformatics/Manichaikul-PhD-Rotation/blob/main/qtl-association/notebooks/matching.ipynb

# Overview

- Colocalization analysis
- Matching colocalized variants
- **Some genes of interest**
- Modeling COPD-related metrics with expression data and covariates
- Outputs from aforementioned model
- Future ideas
- GitHub Repository:

  https://github.com/YogiOnBioinformatics/Manichaikul-PhD-Rotation

# GSTCD and ADAM19 as Potentially Interesting Genes

- GSTCD affects multiple pathways in airway biology
- Validated in independent studies as important COPD gene
  - https://doi.org/10.1164/rccm.201102-0192OC
  - https://doi.org/10.1371/journal.pone.0074630
  - https://doi.org/10.1007/s12041-019-1119-9
  - https://doi.org/10.1186/s12931-019-1146-3

# GSTCD and ADAM19 as Potentially Interesting Genes

- ADAM19 is involved in early immune defense mechanisms in the lungs.
- Also validated as potentially interesting gene for COPD:
  - https://doi.org/10.3109/15412555.2016.1161017
  - https://doi.org/10.1183/13993003.congress-2021.PA2385
  - https://doi.org/10.1007/s12041-019-1119-9

# Overview

- Colocalization analysis
- Matching colocalized variants
- Some genes of interest
- **Modeling COPD-related metrics with expression data and covariates**
- Outputs from aforementioned model
- Future ideas
- GitHub Repository:
  https://github.com/YogiOnBioinformatics/Manichaikul-PhD-Rotation

# Modeling COPD Metrics ~ Expression Levels + Covariates

- Goal:
    - Understand if expression of a particular gene can significantly predict a key COPD-related metric while accounting for covariates.

# Modeling COPD Metrics ~ Expression Levels + Covariates

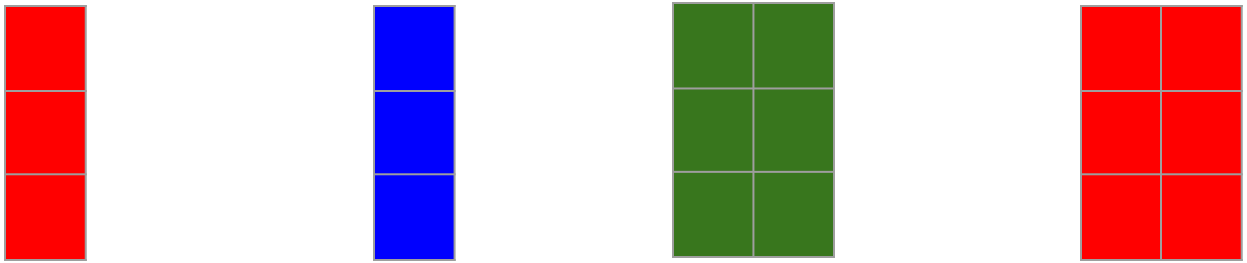
- Goal:
  - Understand if expression of a particular gene can significantly predict a key COPD-related metric while accounting for covariates.

- Model:
  - {insert COPD-related metric} ~ Gene “X” expression, sex, age, age^2, height, height^2, weight (FVC only), pack-years of smoking, current smoking, former smoking, first 10 principal components (PCs) of ancestry, race, PEERs factors

UNIVERSITY of VIRGINIA
SCHOOL OF MEDICINE

# Modeling COPD Metrics ~ Expression Levels + Covariates

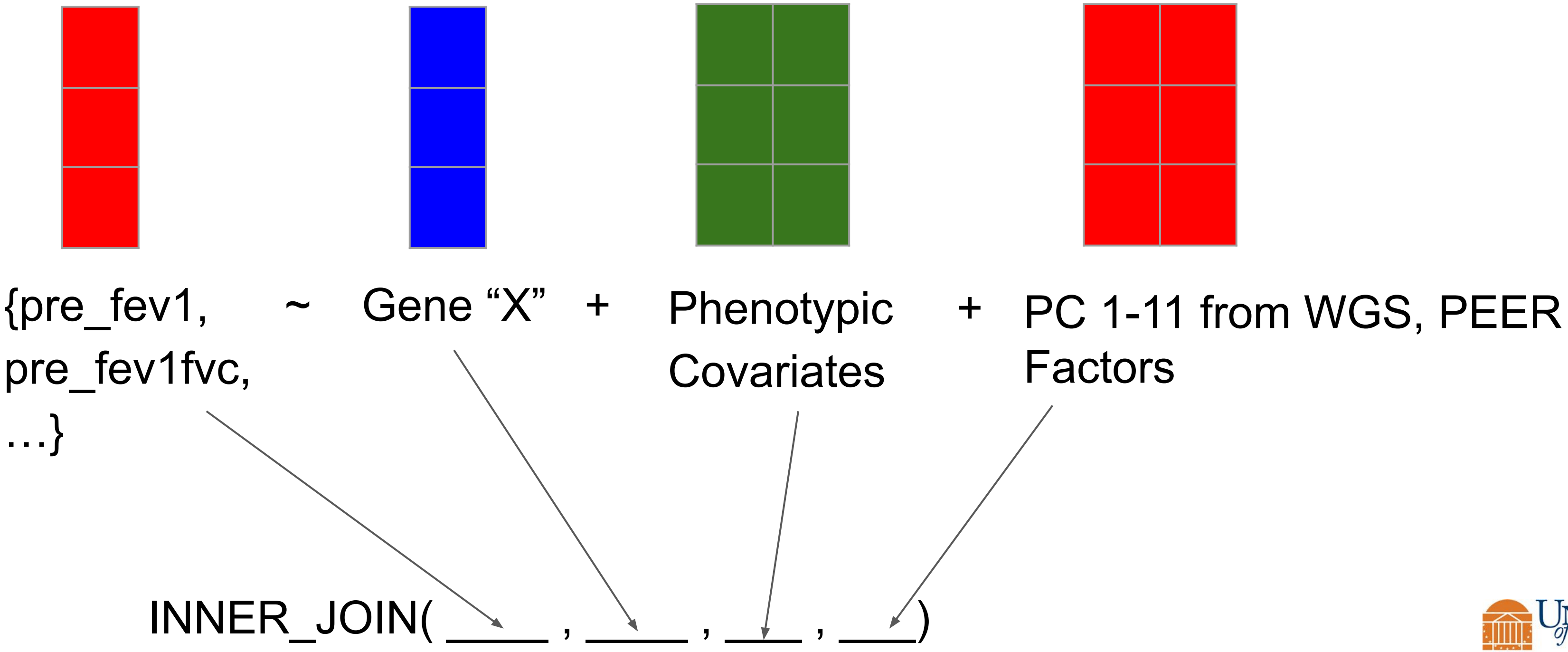
- Goal:
  - Understand if expression of a particular gene can significantly predict a key COPD-related metric while accounting for covariates.
- Model:



{pre_fev1, pre_fev1fvc, …} ~ Gene “X” + Phenotypic Covariates + PC 1-11 from WGS, PEER Factors

# Modeling COPD Metrics ~ Expression Levels + Covariates

- Model:



{pre_fev1, pre_fev1fvc, …} ~ Gene "X" + Phenotypic Covariates + PC 1-11 from WGS, PEER Factors

INNER_JOIN( _____ , _____ , ____ , ____)

University of Virginia School of Medicine

# Important Considerations

- Only using Exam 5 data.
    - No lung data prior.

# Important Considerations

- Only using Exam 5 data.
  - No lung data prior.
- Post-bronchodilator (“post_”) better than pre-bronchodilator (“pre_”).
  - Yet, few patients have “post_” vs “pre_” measurements.
  - Hence, using “pre_” for target variables.

# Important Considerations

- Only using Exam 5 data.
    - No lung data prior.
- Post-bronchodilator (“post_”) better than pre-bronchodilator (“pre_”).
    - Yet, few patients have “post_” vs “pre_” measurements.
    - Hence, using “pre_” for target variables.
- Do not impute any missing values.
    - Drop participants missing data for any covariate.

# Important Considerations

- Only using Exam 5 data.
  - No lung data prior.
- Post-bronchodilator (“post_”) better than pre-bronchodilator (“pre_”).
  - Yet, few patients have “post_” vs “pre_” measurements.
  - Hence, using “pre_” for target variables.
- Do not impute any missing values.
  - Drop participants missing data for any covariate.
- Convert ENSEMBL Gene IDs to Gene Symbols for interpretability.
  - e.g. ENSGXXXXXX —> HLA-DQA2

# Important Considerations

- Expression data already normalized using inverse normal transform.

$$Y_i^t = \Phi^{-1}\left(\frac{r_i - c}{N - 2c + 1}\right)$$

SOURCE:
https://doi.org/10.1007/s10519-009-9281-0

# Important Considerations

- Expression data already normalized using inverse normal transform.

$$Y_i^t = \Phi^{-1}\left(\frac{r_i - c}{N - 2c + 1}\right)$$

SOURCE:
https://doi.org/10.1007/s10519-009-9281-0

- Take square of “age” and “height” features
  - Accounts for potential non-linear effects.

# Important Considerations

- Expression data already normalized using inverse normal transform.

$$Y_i^t = \Phi^{-1}\left(\frac{r_i - c}{N - 2c + 1}\right)$$

SOURCE:
https://doi.org/10.1007/s10519-009-9281-0

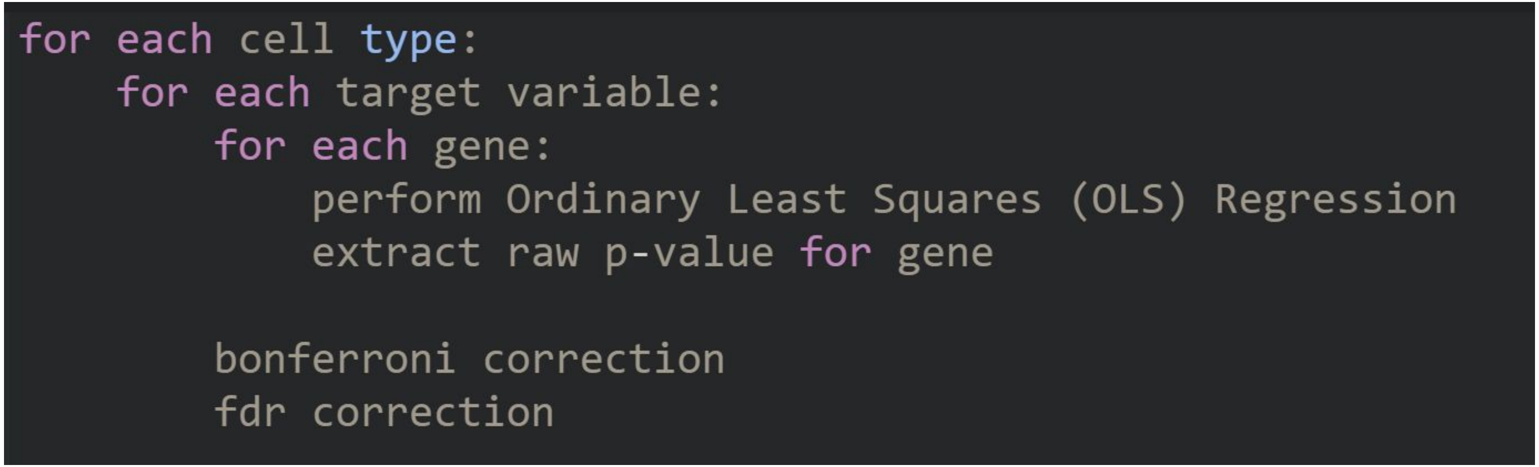
- Take square of “age” and “height” features
  - Accounts for potential non-linear effects.
- Use time-variant variables instead of baseline.

- Notebook source:

  https://github.com/YogiOnBioinformatics/Manichaikul-PhD-Rotation/blob/main/rna-seq-regression/notebooks/join_and_matrixize_data.ipynb

# Exhaustive Modeling

- Algorithm:



```
for each cell type:
    for each target variable:
        for each gene:
            perform Ordinary Least Squares (OLS) Regression
            extract raw p-value for gene

        bonferroni correction
        fdr correction
```

- Notebook source:

  https://github.com/YogiOnBioinformatics/Manichaikul-PhD-Rotation/blob/main/rna-seq-regression/notebooks/regression.ipynb

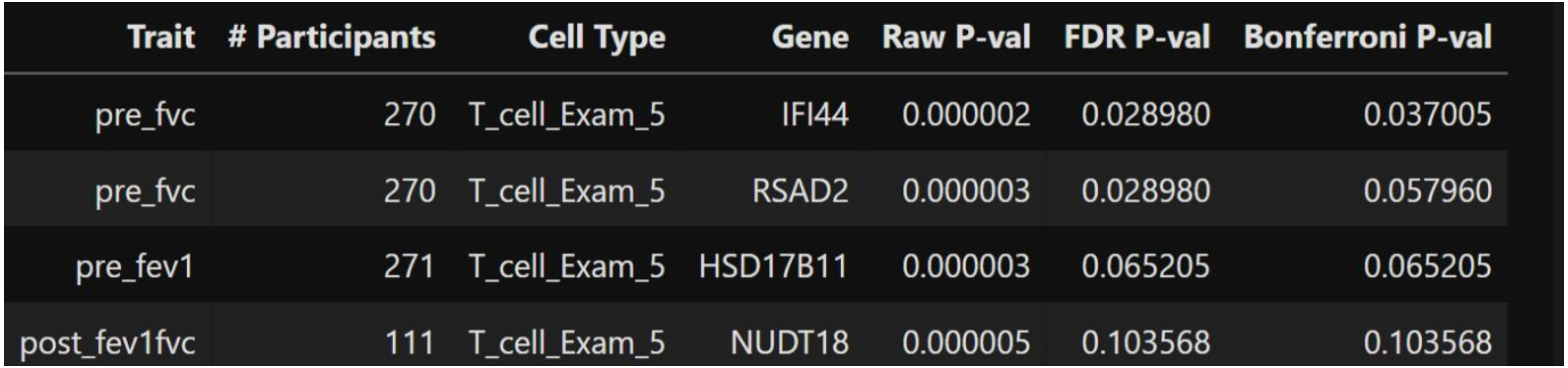
# Overview

- Colocalization analysis
- Matching colocalized variants
- Some genes of interest
- Modeling COPD-related metrics with expression data and covariates
- **Outputs from aforementioned model**
- Future ideas
- GitHub Repository:

  https://github.com/YogiOnBioinformatics/Manichaikul-PhD-Rotation

# Exhaustive Modeling Results

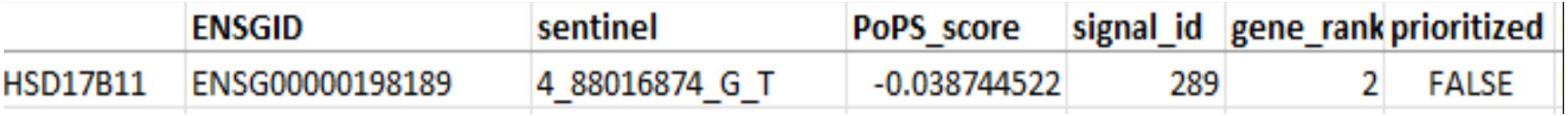
- When looking at corrected p-values for each gene per cell type per target variable:



| Trait | # Participants | Cell Type | Gene | Raw P-val | FDR P-val | Bonferroni P-val |
|---|---|---|---|---|---|---|
| pre_fvc | 270 | T_cell_Exam_5 | IFI44 | 0.000002 | 0.028980 | 0.037005 |
| pre_fvc | 270 | T_cell_Exam_5 | RSAD2 | 0.000003 | 0.028980 | 0.057960 |
| pre_fev1 | 271 | T_cell_Exam_5 | HSD17B11 | 0.000003 | 0.065205 | 0.065205 |
| post_fev1fvc | 111 | T_cell_Exam_5 | NUDT18 | 0.000005 | 0.103568 | 0.103568 |

# Exhaustive Modeling Results

- Only 1 of prior 4 genes overlapped with results from this pre-print.
  - https://www.medrxiv.org/content/10.1101/2022.05.11.22274314v2



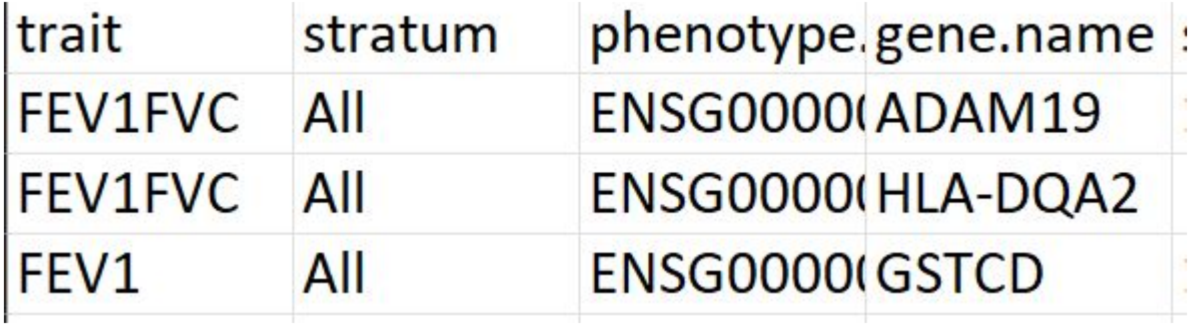
| | ENSGID | sentinel | PoPS_score | signal_id | gene_rank | prioritized |
|---|---|---|---|---|---|---|
| HSD17B11 | ENSG00000198189 | 4_88016874_G_T | -0.038744522 | 289 | 2 | FALSE |

- Seems to be a relatively unimportant gene in their analysis.

# Modeling “Hit Genes”

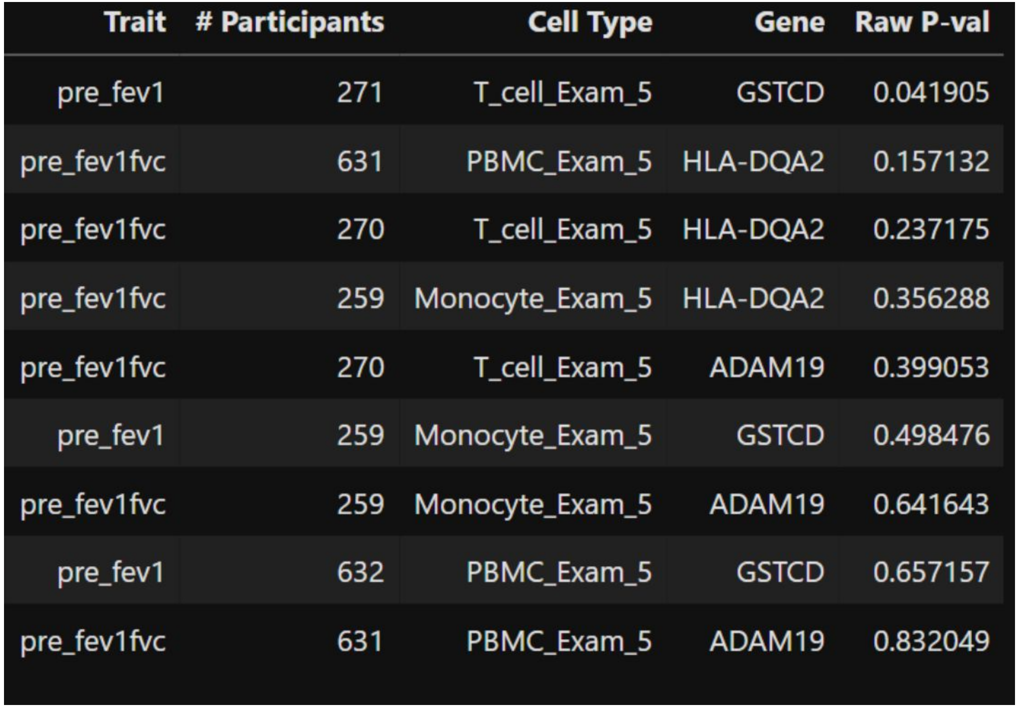
- For a given trait and given gene that has 2 variants that causally affect both the trait and gene…
  - Can the trait be modelled successfully by the gene’s expression level, accounting for covariates?

- eQTL analysis from Xiaowei:



| trait | stratum | phenotype. | gene.name |
|---|---|---|---|
| FEV1FVC | All | ENSG0000 | ADAM19 |
| FEV1FVC | All | ENSG0000 | HLA-DQA2 |
| FEV1 | All | ENSG0000 | GSTCD |

# Modeling “Hit Genes” Yields Mostly Non-Significant Results



| Trait | # Participants | Cell Type | Gene | Raw P-val |
|---:|---:|---:|---:|---:|
| pre_fev1 | 271 | T_cell_Exam_5 | GSTCD | 0.041905 |
| pre_fev1fvc | 631 | PBMC_Exam_5 | HLA-DQA2 | 0.157132 |
| pre_fev1fvc | 270 | T_cell_Exam_5 | HLA-DQA2 | 0.237175 |
| pre_fev1fvc | 259 | Monocyte_Exam_5 | HLA-DQA2 | 0.356288 |
| pre_fev1fvc | 270 | T_cell_Exam_5 | ADAM19 | 0.399053 |
| pre_fev1 | 259 | Monocyte_Exam_5 | GSTCD | 0.498476 |
| pre_fev1fvc | 259 | Monocyte_Exam_5 | ADAM19 | 0.641643 |
| pre_fev1 | 632 | PBMC_Exam_5 | GSTCD | 0.657157 |
| pre_fev1fvc | 631 | PBMC_Exam_5 | ADAM19 | 0.832049 |

# Overview

- Colocalization analysis
- Matching colocalized variants
- Some genes of interest
- Modeling COPD-related metrics with expression data and covariates
- Outputs from aforementioned model
- **Future ideas**
- GitHub Repository:

  https://github.com/YogiOnBioinformatics/Manichaikul-PhD-Rotation

# Future Directions & Acknowledgements

- Run same type of modeling but…
  - Replace gene expression vector with methylation vector
    - Ani has covariates for the methylation data.
  - If available, could do the same with protein expression vector

# Future Directions & Acknowledgements

- Run same type of modeling but…
  - Replace gene expression vector with methylation vector
    - Ani has covariates for the methylation data.
  - If available, could do the same with protein expression vector
- If many proteins or methylation sites near genes turn up significant…
  - Gene Ontology enrichment analysis is worth a shot.

# Future Directions & Acknowledgements

- Run same type of modeling but…
  - Replace gene expression vector with methylation vector
    - Ani has covariates for the methylation data.
  - If available, could do the same with protein expression vector
- If many proteins or methylation sites near genes turn up significant…
  - Gene Ontology enrichment analysis is worth a shot.
- Thanks to:
  - Ani
  - Jisu
  - Xiaowei
  - Catherine