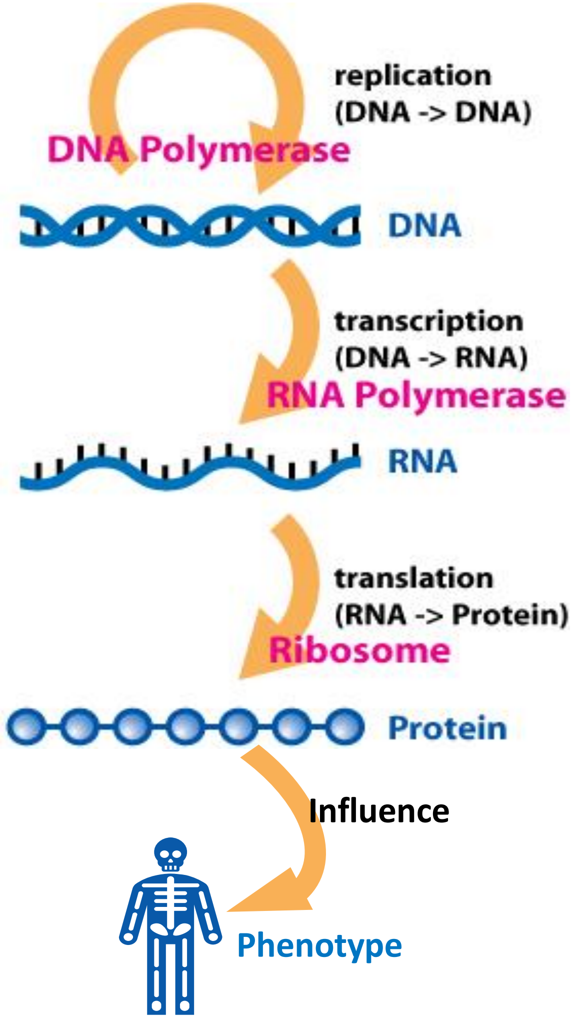
Identification of cell-type-specific genetic regulation of gene expression for transcriptome-wide association studies

Qiurui Ma; Duo Zhang; Brandon Jew; Sriram Sankararaman

The Central Dogma



File:Central Dogma of Molecular Biochemistry with Enzymes.jpg. (2019, September 4). Wikimedia Commons, the free media repository. Retrieved 00:50, September 5, 2019 from https://commons.wikimedia.org/w/index.php?title=File:Central_Dogma_of_Molecular_Biochemistry_with_Enzymes.jpg&oldid=364559532.

The Central Dogma

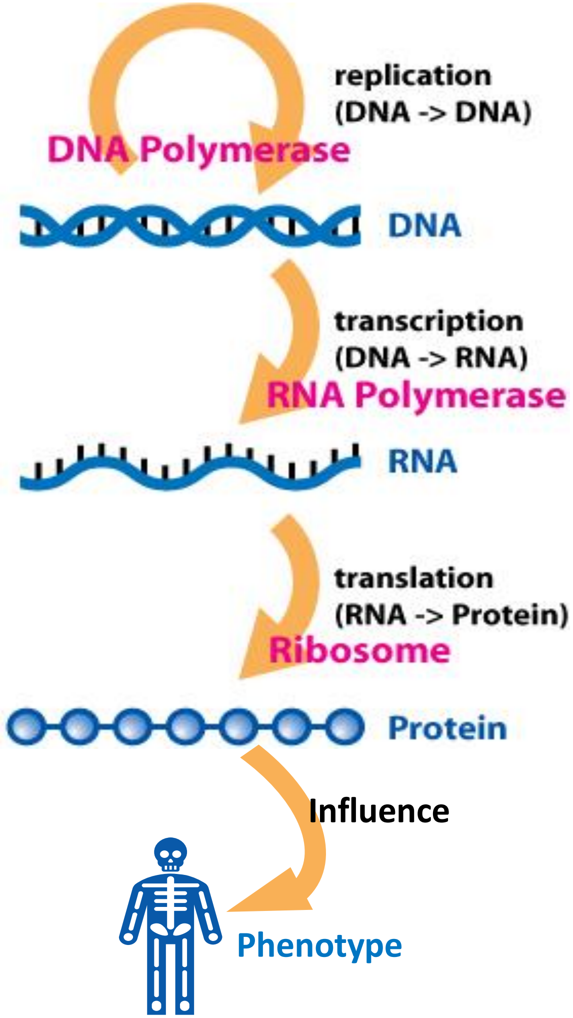
TCA-TWAS

Background
Model
Experiment
Conclusion

SNPs

Ind 1	Ind 2	Ind 3
AA	AG	AG
CT	CC	CC

Geno1	Geno2	Geno3
0	1	1
1	1	2



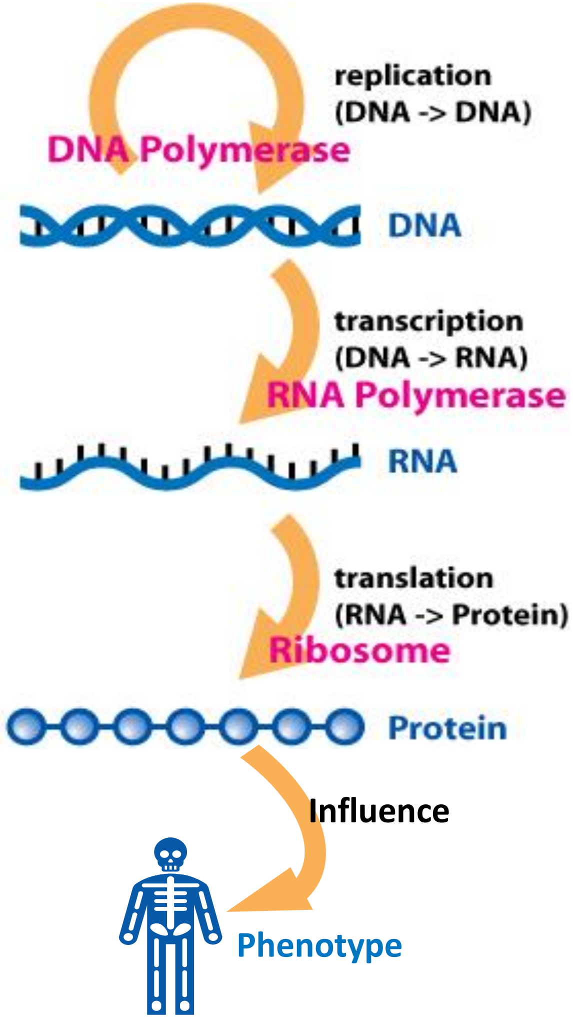
File:Central Dogma of Molecular Biochemistry with Enzymes.jpg. (2019, September 4). Wikimedia Commons, the free media repository. Retrieved 00:50, September 5, 2019 from https://commons.wikimedia.org/w/index.php?title=File:Central_Dogma_of_Molecular_Biochemistry_with_Enzymes.jpg&oldid=364559532.

The Central Dogma

SNPs

Ind 1	Ind 2	Ind 3
AA	AG	AG
CT	CC	CC

Geno1	Geno2	Geno3
0	1	1
1	1	2



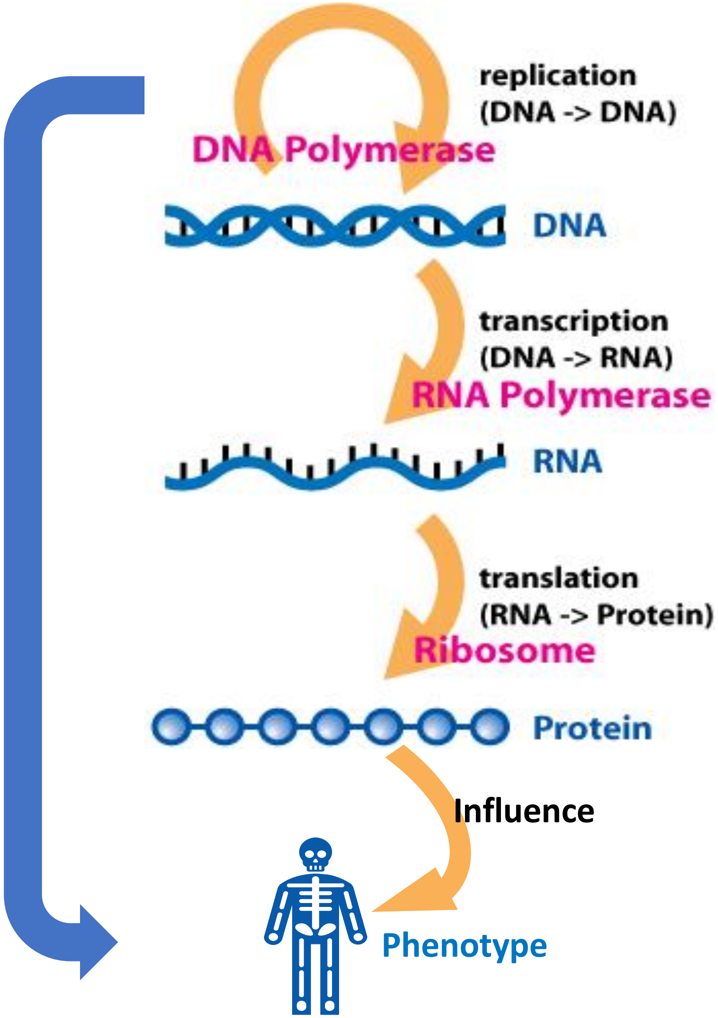
Gene Expression (GE):
cell 1, cell 2, cell 3, cell 4

Phenotype:
Height, Skin Color

File:Central Dogma of Molecular Biochemistry with Enzymes.jpg. (2019, September 4). Wikimedia Commons, the free media repository. Retrieved 00:50, September 5, 2019 from https://commons.wikimedia.org/w/index.php?title=File:Central_Dogma_of_Molecular_Biochemistry_with_Enzymes.jpg&oldid=364559532.

Current Research: GWAS

Linear Regression of SNPs and phenotype



Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W., ... & Sullivan, P. F. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics*, *48*(3), 245

Current Research: TWAS

TCA-TWAS

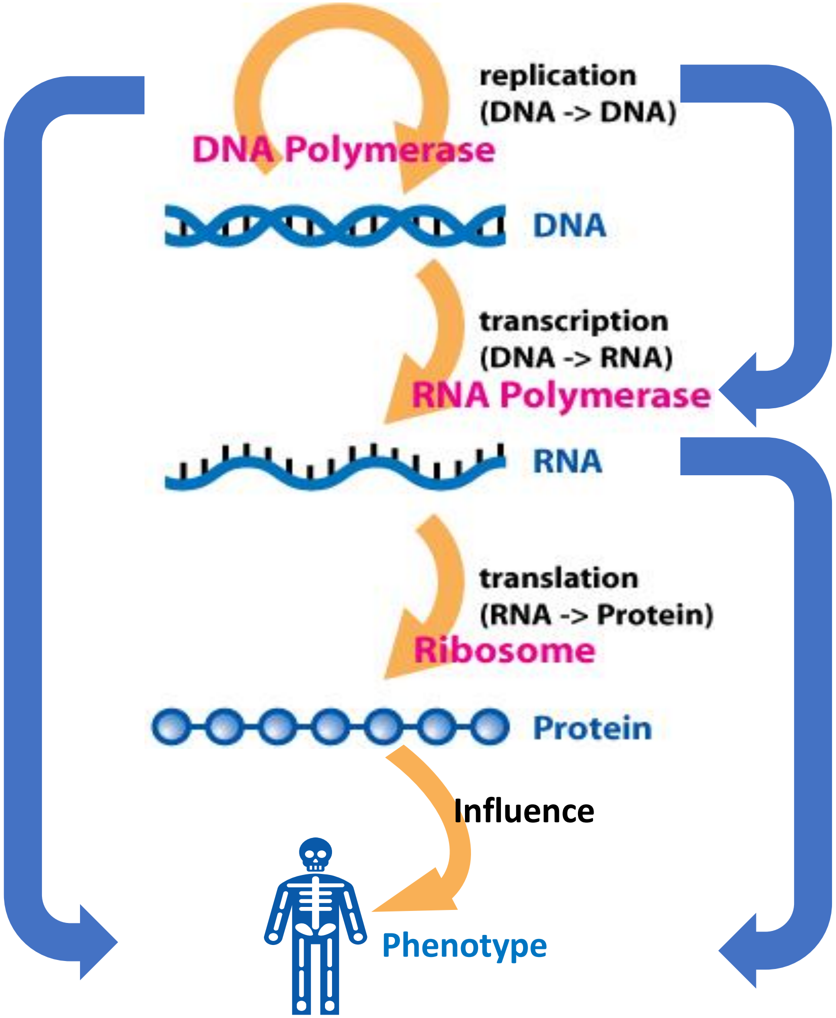
Background

Model

Experiment

Conclusion

Linear Regression of SNPs and phenotype

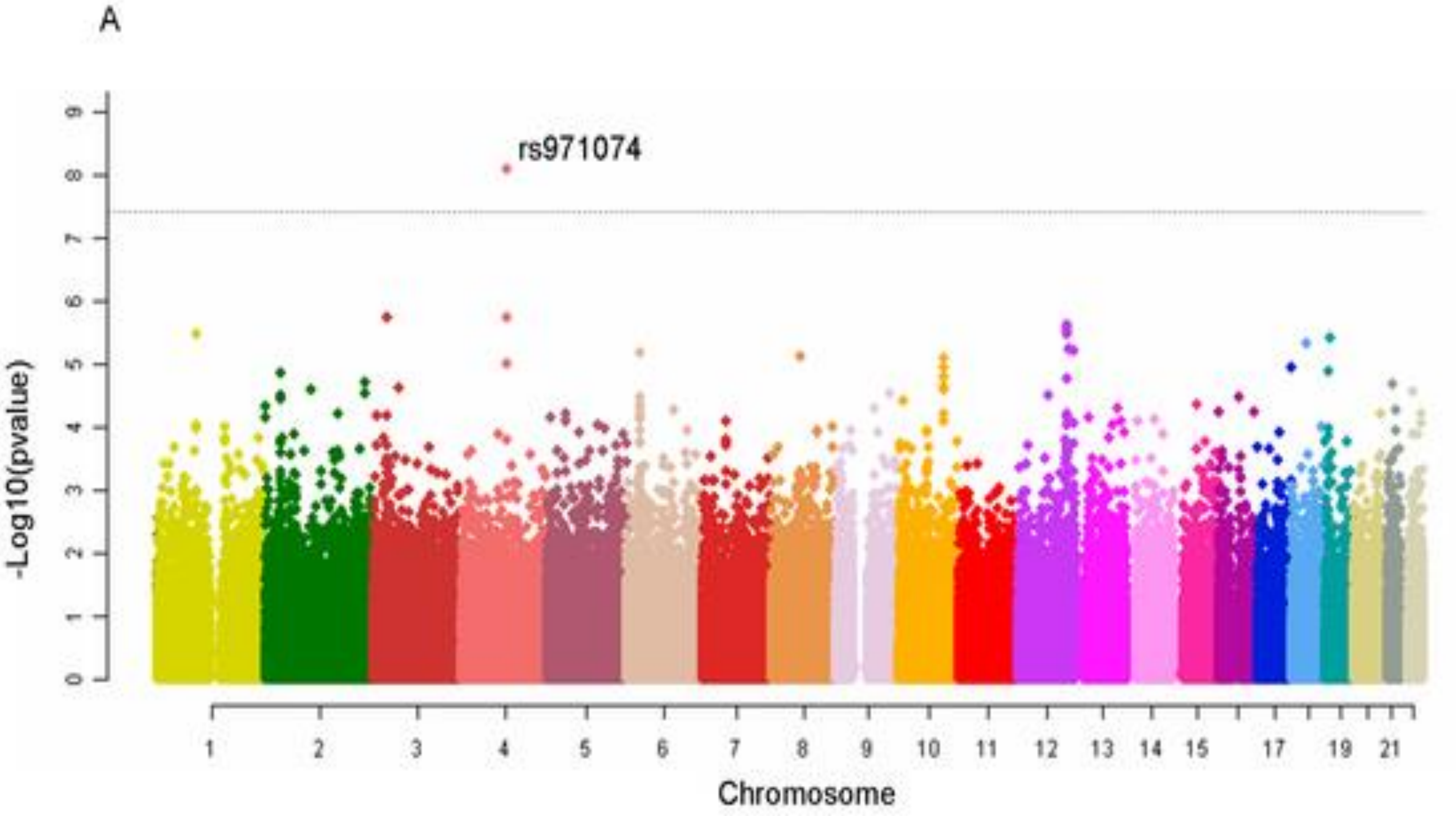


Linear Regression of SNPs and GE

Linear Regression of GE and Phenotype

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W., ... & Sullivan, P. F. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics*, *48*(3), 245

Current Research: Results



Manhattan Plot of UADT(Upper Aero-Digestive Tract) Cancer GWAS Discovery Phase

Mckay et, al. (2011). A Genome-Wide Association Study of Upper Aerodigestive Tract Cancers Conducted within the INHANCE Consortium. PLoS genetics. 7. e1001333. 10.1371/journal.pgen.1001333.

Bulk & Cell-Specific GE

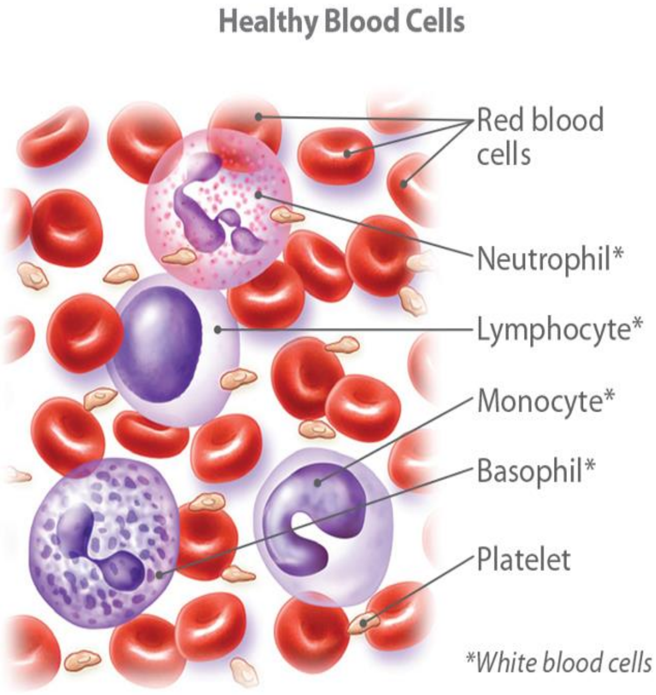
TCA-TWAS

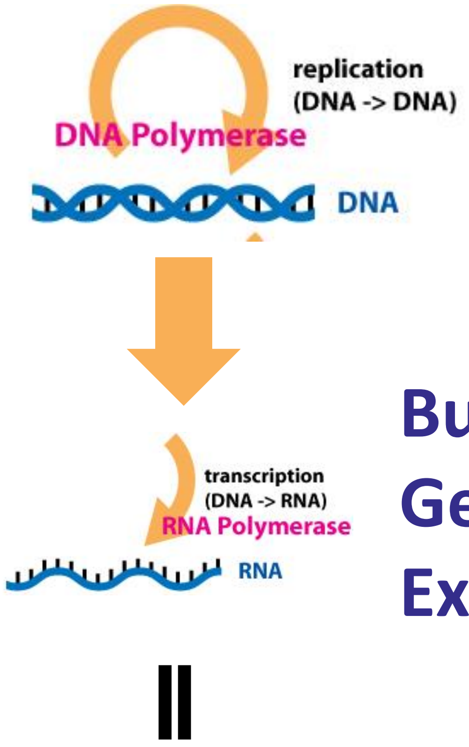
Background

Model

Experiment

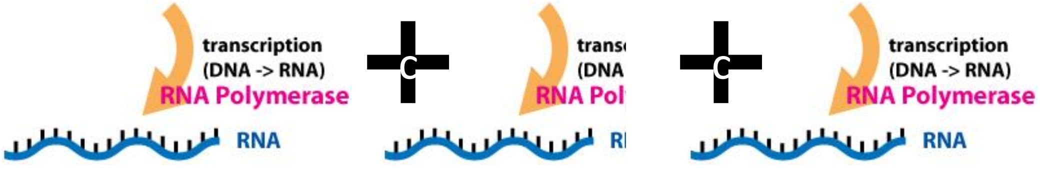
Conclusion



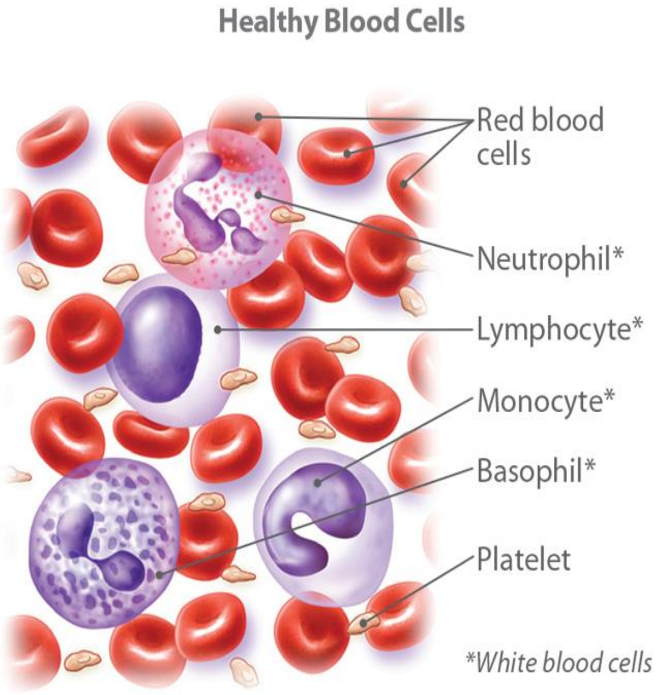


Bulk Level Gene Expression

Weighted Cell-Type-Specific Gene Expression



Challenges in Current Work



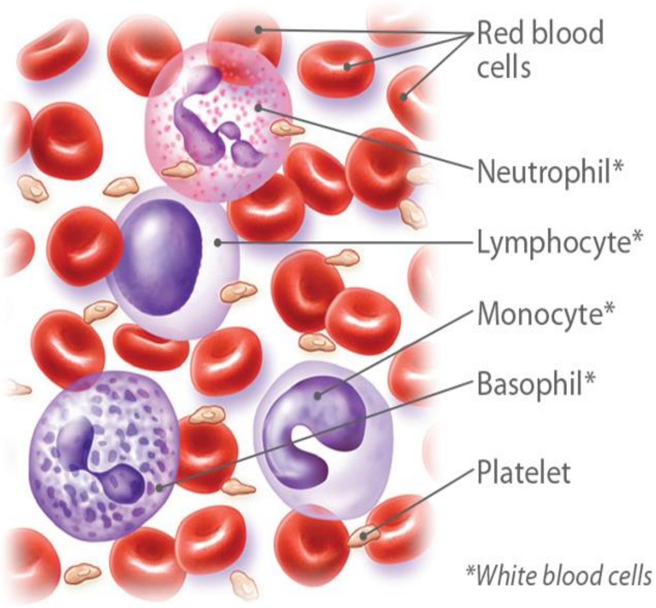
Methodological:
Unclear how SNPs affect phenotypes

- Missing cell type information
- Fail to tell causality from correlation

Practical:
Cell-Specific Biological Data being resource intensive, expensive to acquire

Challenges in Current Work

Healthy Blood Cells



*White blood cells

Methodological:

Unclear how SNPs affect phenotypes

- Missing cell type information
- Fail to tell causality from correlation

Practical:

Cell-Specific Biological Data being resource intensive, expensive to acquire

Goal:

Impute cell-type specific gene expressions from SNPs and bulk level gene expressions to perform downstream TWAS

Train & Target Dataset

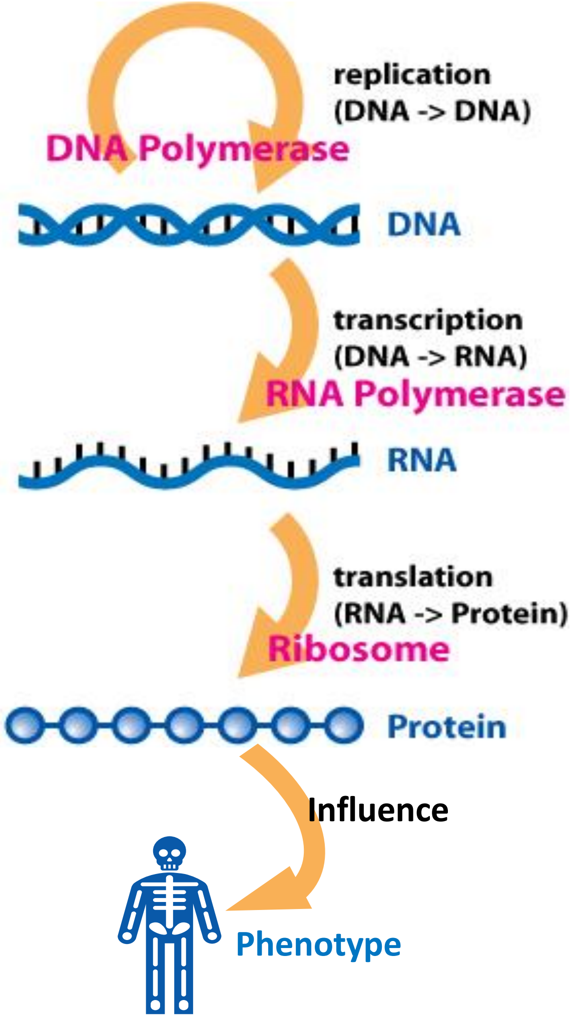
TCA-TWAS

Background

Model

Experiment

Conclusion



Train Dataset	Target Dataset
✓	✓
Bulk Level GE	✗
✓	✓

File:Central Dogma of Molecular Biochemistry with Enzymes.jpg. (2019, September 4). Wikimedia Commons, the free media repository. Retrieved 00:50, September 5, 2019 from https://commons.wikimedia.org/w/index.php?title=File:Central_Dogma_of_Molecular_Biochemistry_with_Enzymes.jpg&oldid=364559532.

Pipeline Framework

TCA-TWAS

Background

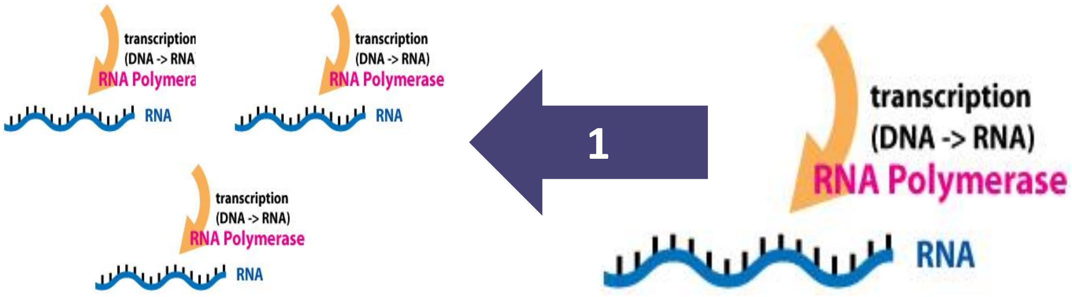
Model

Experiment

Conclusion

Train Dataset

TCA Deconvolution



Cell-Type-Specific Gene Expression

Bulk Level Gene Expression

Pipeline Framework

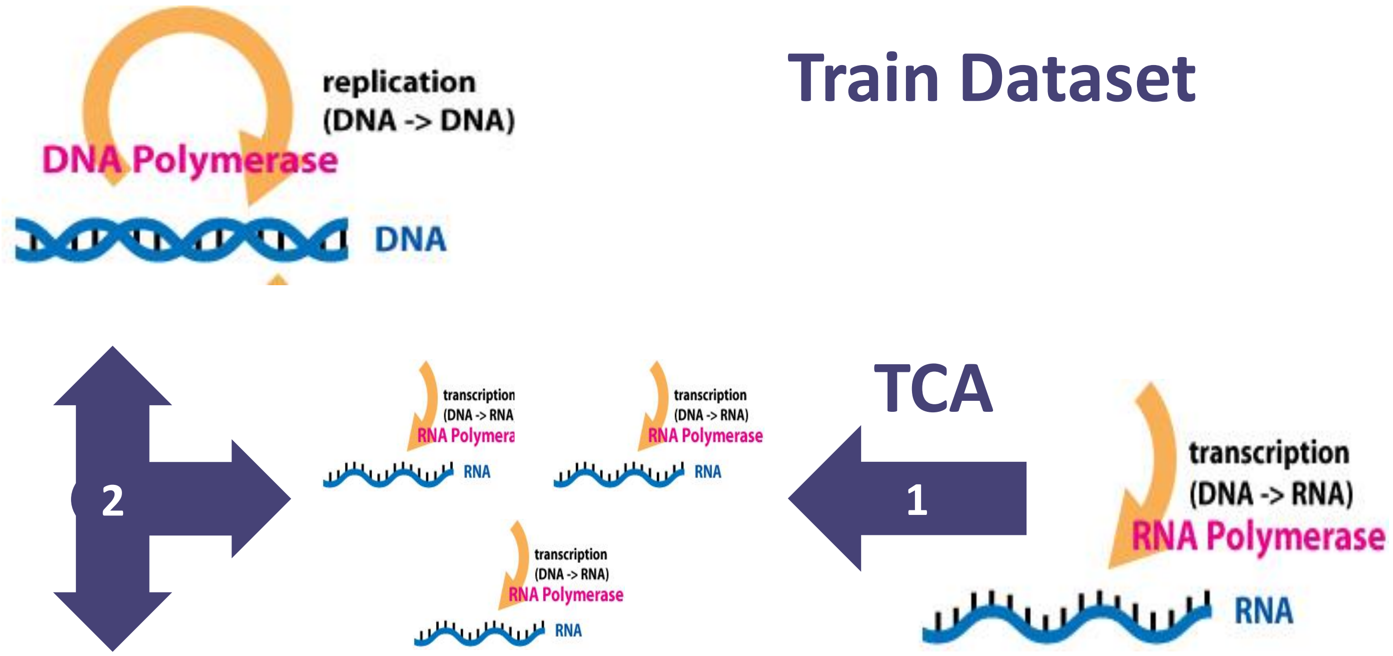
TCA-TWAS

Background

Model

Experiment

Conclusion



Train Dataset

TCA

LASSO

β

Estimated effect size of SNPs on GE

Pipeline Framework

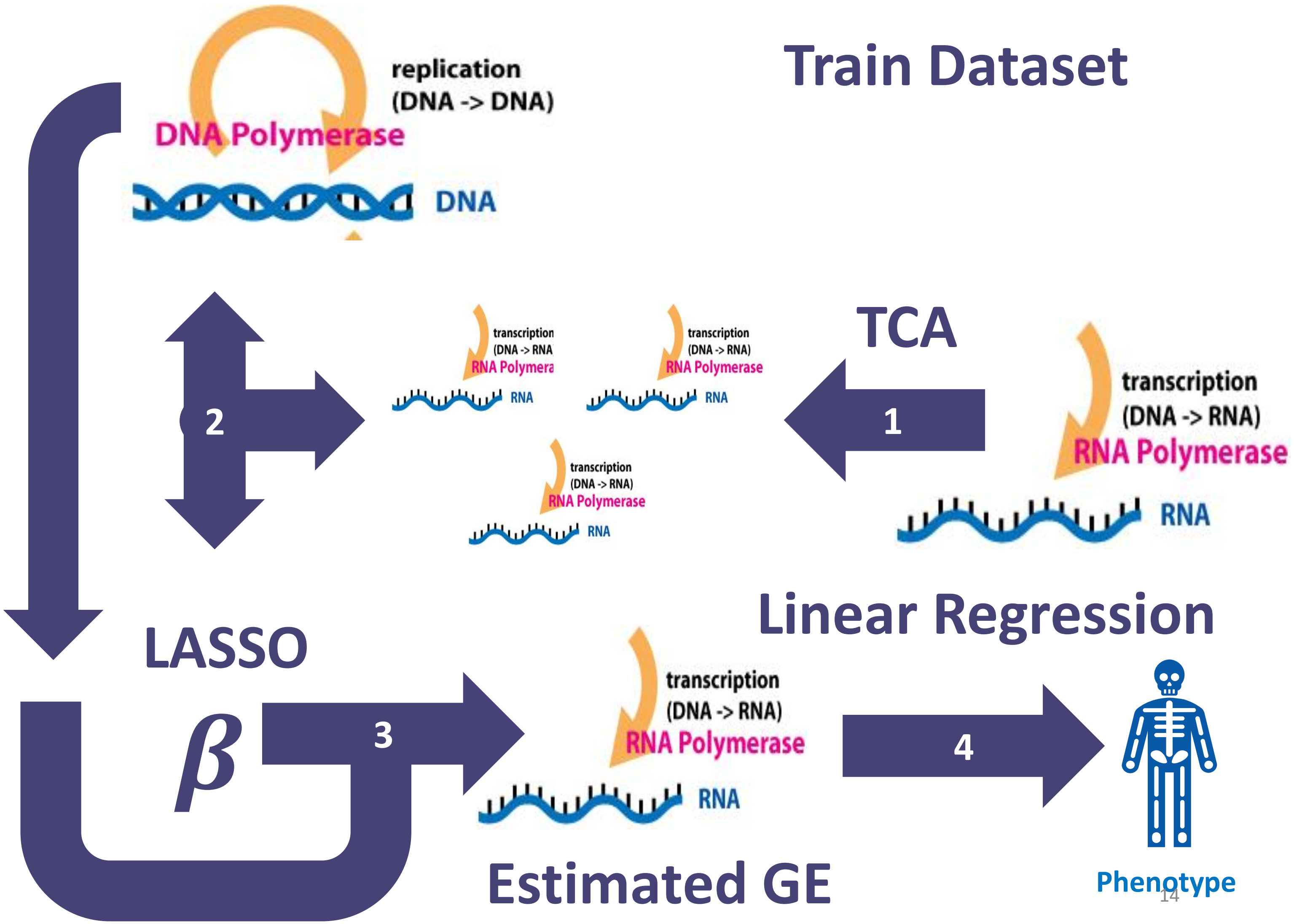
TCA-TWAS

Background

Model

Experiment

Conclusion



Train Dataset

TCA

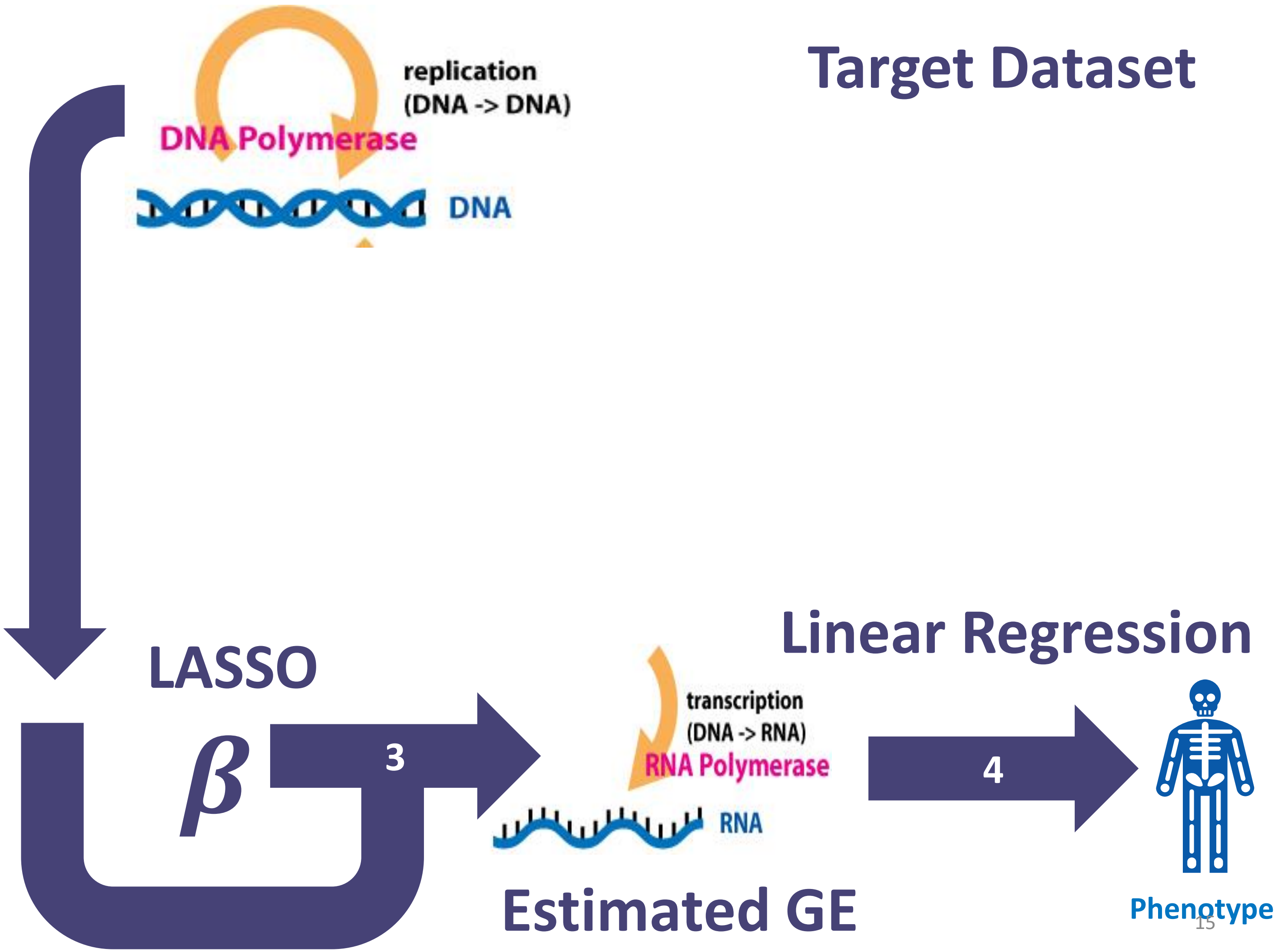
Linear Regression

LASSO

β

Estimated GE

Pipeline Framework



Target Dataset

LASSO

β

3

Linear Regression

4

Estimated GE

TCA model

TCA-TWAS

Background
Model
Experiment
Conclusion

$$Z_h^i = \overbrace{\overbrace{\epsilon_z + \mu_h + \begin{bmatrix}1\\27\\0\end{bmatrix}^T \gamma_h}^{\text{TCA Model: mdl2}} + \begin{bmatrix}0\\2\\1\\\vdots\\0\\1\end{bmatrix}^T}^{\text{added SNPs effect: mdl1}} \beta_h$$

Cell-Specific GE

Covariate:
(Gender, Age, Smoking)

SNPs

Rahmani, E., Schweiger, R., Rhead, B., Criswell, L. A., Barcellos, L. F., Eskin, E., ... & Halperin, E. (2019). Cell-type-specific resolution epigenetics without the need for cell sorting or single-cell biology. *BioRxiv*, 437368.

TCA model

TCA-TWAS

Background

Model

Experiment

Conclusion

$$z_h^i = \overbrace{\overbrace{\epsilon_z + \mu_h + \begin{bmatrix} 1 \\ 27 \\ 0 \end{bmatrix}^T \gamma_h}^{\text{TCA Model: mdl2}} + \begin{bmatrix} 0 \\ 2 \\ 1 \\ \vdots \\ 0 \\ 1 \end{bmatrix}^T \beta_h}^{\text{added SNPs effect: mdl1}}$$

Cell-Specific GE

Covariate:
(Gender, Age, Smoking)

SNPs

$$G_i = c_i^2 \delta + \sum_{h=1}^{k} w_{hi} z_{hi} + \epsilon_g$$

Cell Type Weight
Extra Information

Rahmani, E., Schweiger, R., Rhead, B., Criswell, L. A., Barcellos, L. F., Eskin, E., ... & Halperin, E. (2019). Cell-type-specific resolution epigenetics without the need for cell sorting or single-cell biology. *BioRxiv*, 437368.

TCA model

TCA-TWAS

Background

Model

Experiment

Conclusion

$$Z_h^i = \overbrace{\overbrace{\epsilon_z + \mu_h + \begin{bmatrix}1\\27\\0\end{bmatrix}^T \gamma_h}^{\text{TCA Model: mdl2}} + \begin{bmatrix}0\\2\\1\\\vdots\\0\\1\end{bmatrix}^T}^{\text{added SNPs effect: mdl1}} \beta_h$$

Cell-Specific GE

Covariate:
(Gender, Age, Smoking)

SNPs

$$G_i = c_i^2\delta + \sum_{h=1}^{k} w_{hi} z_{hi} + \epsilon_g$$

Cell Type Weight
Extra Information

$$Pr(Z_h^i | G_i, w_i, \mu_h, \sigma_z, \sigma_g, \sigma_\delta)$$

EM to infer parameters

Rahmani, E., Schweiger, R., Rhead, B., Criswell, L. A., Barcellos, L. F., Eskin, E., ... & Halperin, E. (2019). Cell-type-specific resolution epigenetics without the need for cell sorting or single-cell biology. *BioRxiv*, 437368.

Simulated Data

TCA-TWAS

Background

Model

Experiment

Conclusion

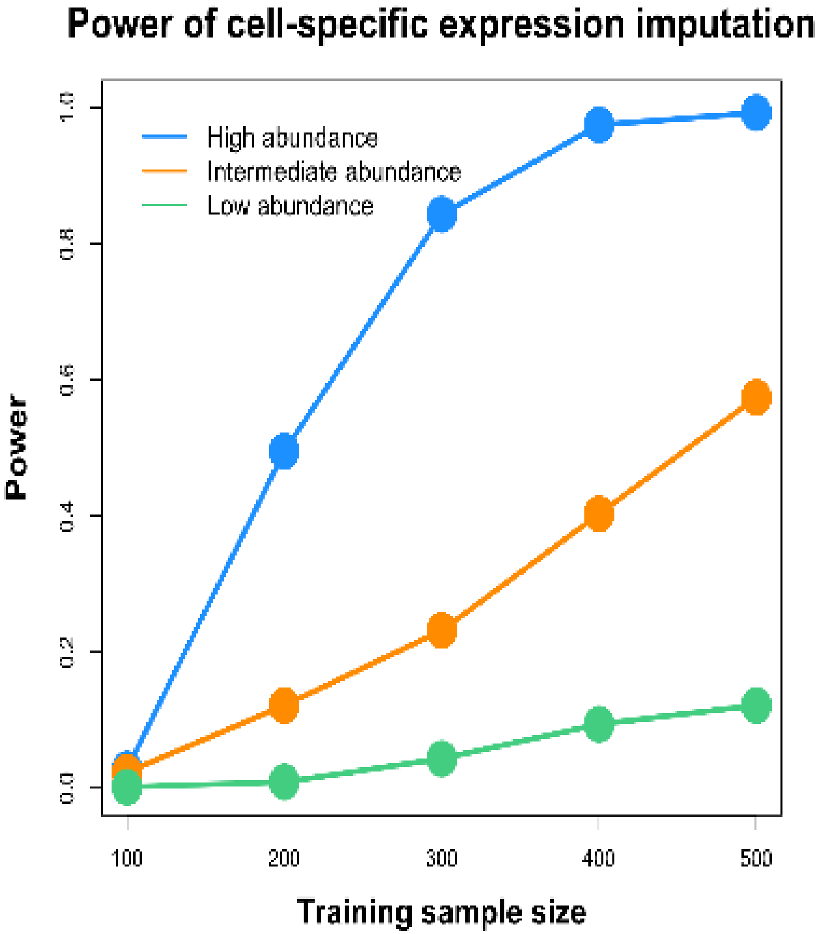
Is the prediction result of proposed model significant?

transcription
(DNA -> RNA)
RNA Polymerase
RNA

Estimated GE

4

Phenotype


Power of cell-specific expression imputation

Power calculated as percentage of p value less than threshold
Heritability 0.3, 1000 genes, 100000 sample size

Simulated Data

TCA-TWAS

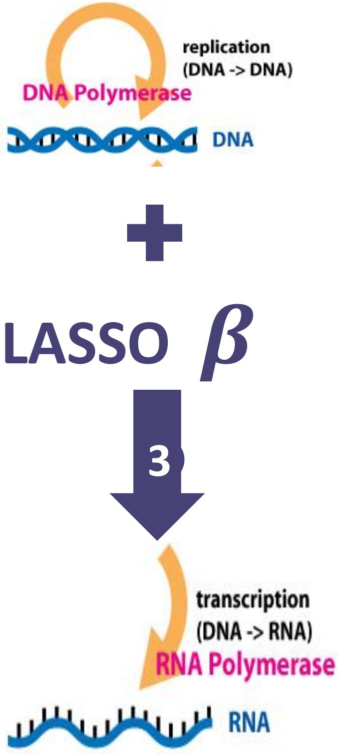
Background

Model

Experiment

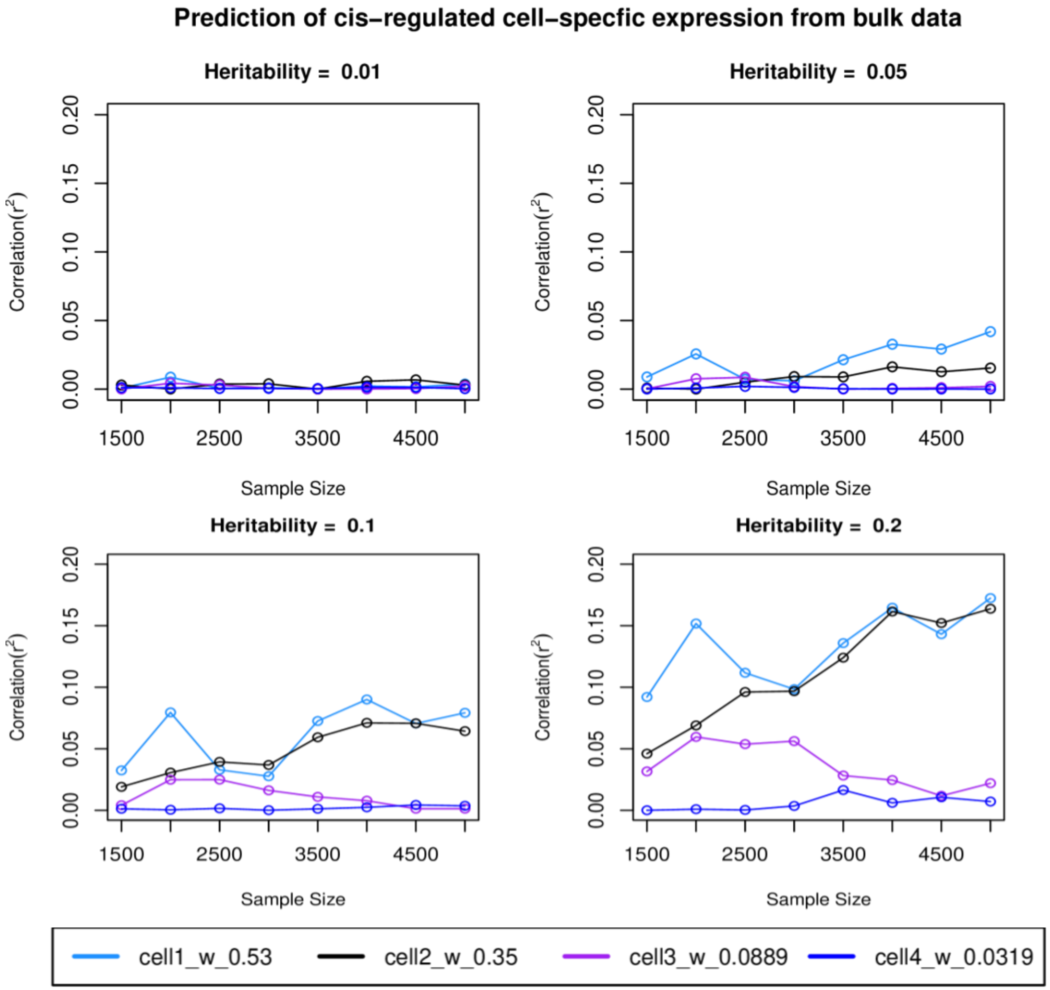
Conclusion

Is the model underfitting or overfitting?



LASSO β

Estimated GE


Prediction of cis–regulated cell–specfic expression from bulk data

Heritability = 0.01

Heritability = 0.05

Heritability = 0.1

Heritability = 0.2

Heritability being theoretical upper bound for r^2

Simulated Data

TCA-TWAS

Background

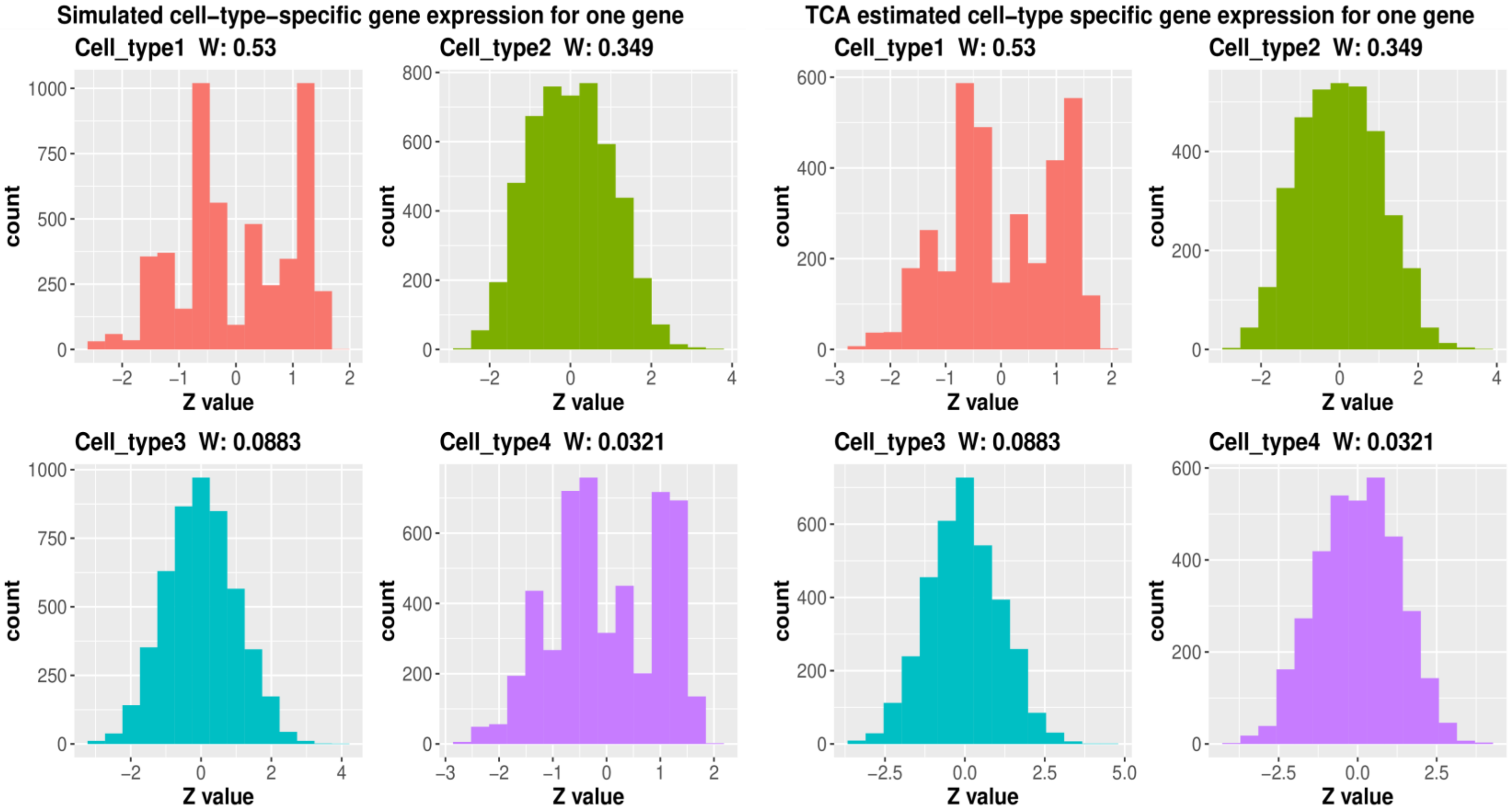
Model

Experiment

Conclusion

Could TCA recover the ground truth data distribution?

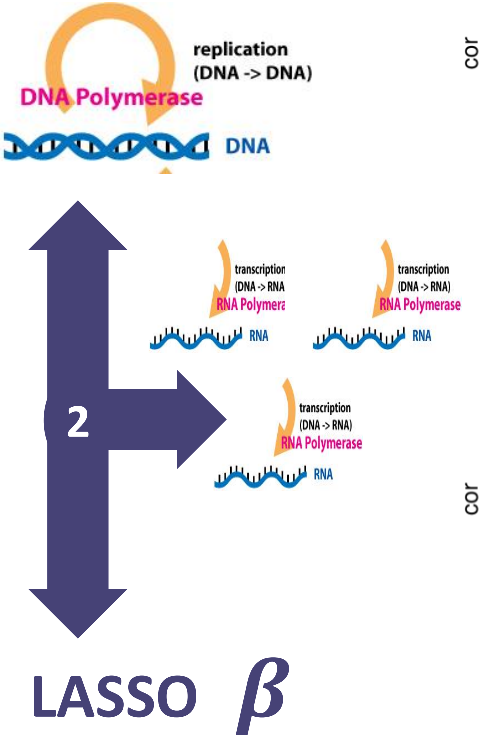
Step 1

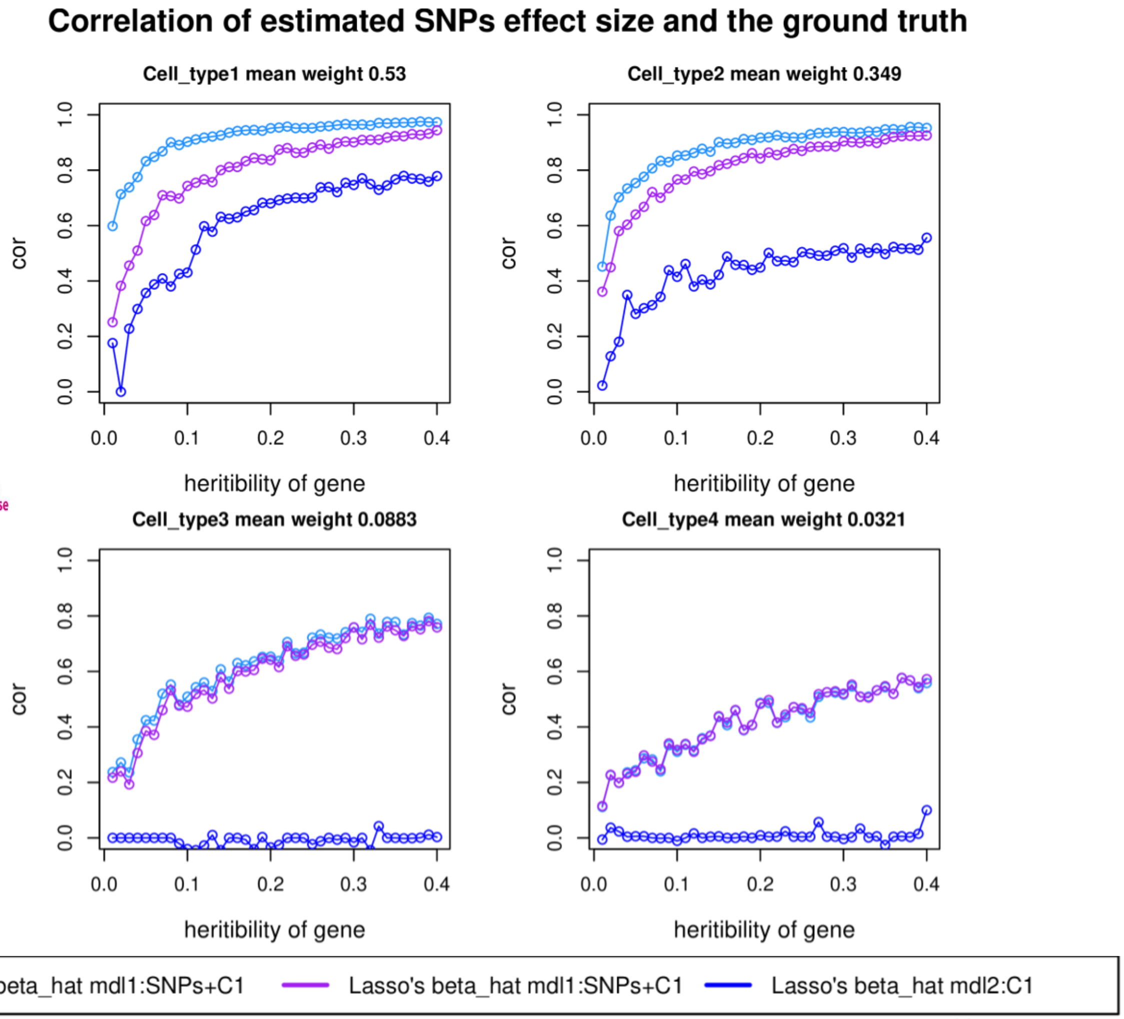


TCA recover abundant cell type distribution, but not the lesser ones

Simulated Data

Compare our modified TCA and the original TCA, which is better?



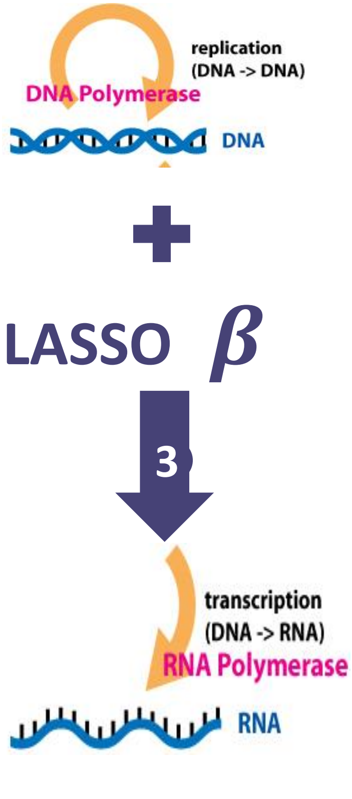

Correlation of estimated SNPs effect size and the ground truth

Simulated Data

TCA-TWAS

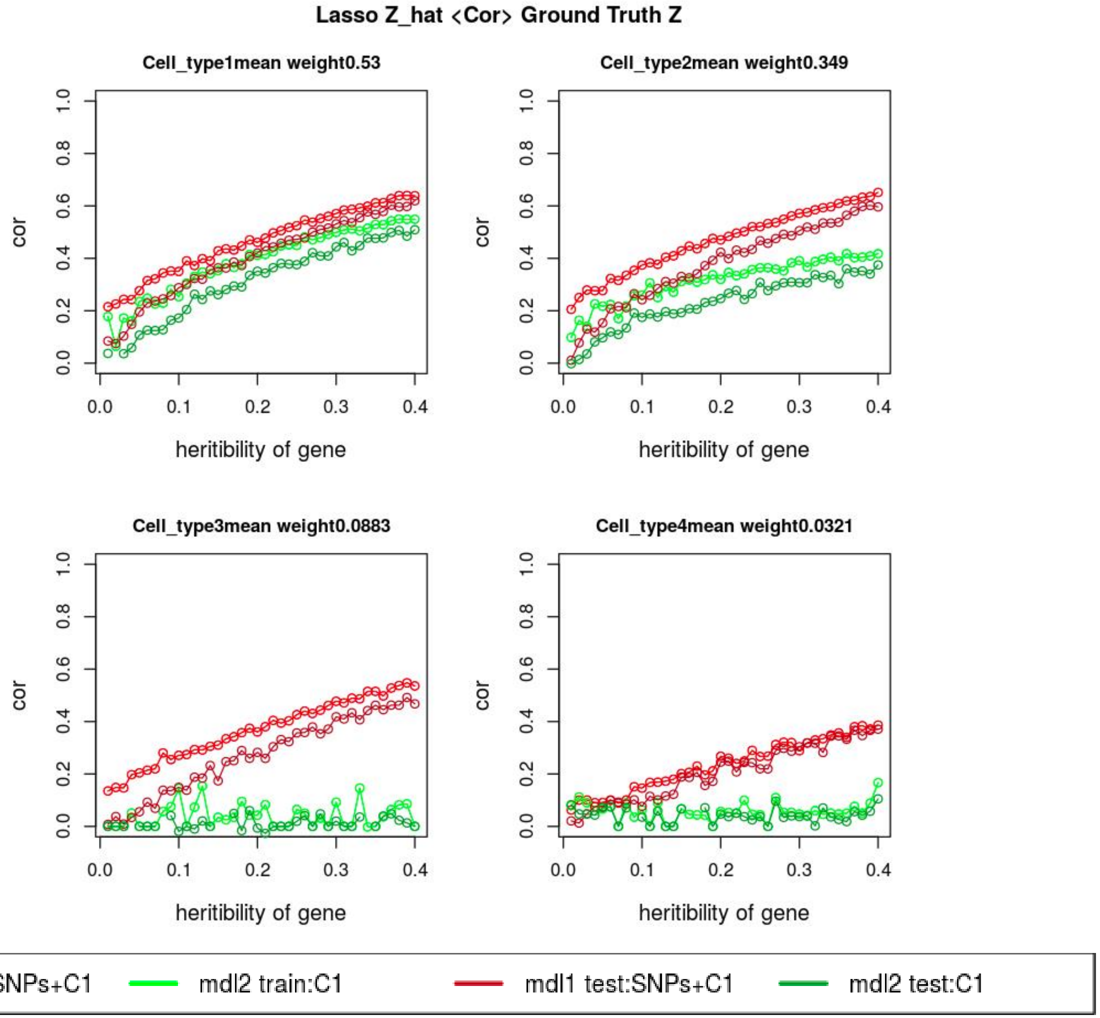
Background
Model
Experiment
Conclusion

Compare our modified TCA and the original TCA, which is better?



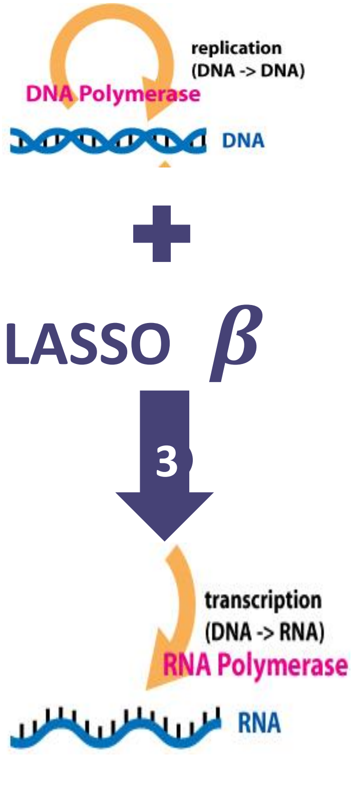
LASSO β

Estimated GE



Simulated Data

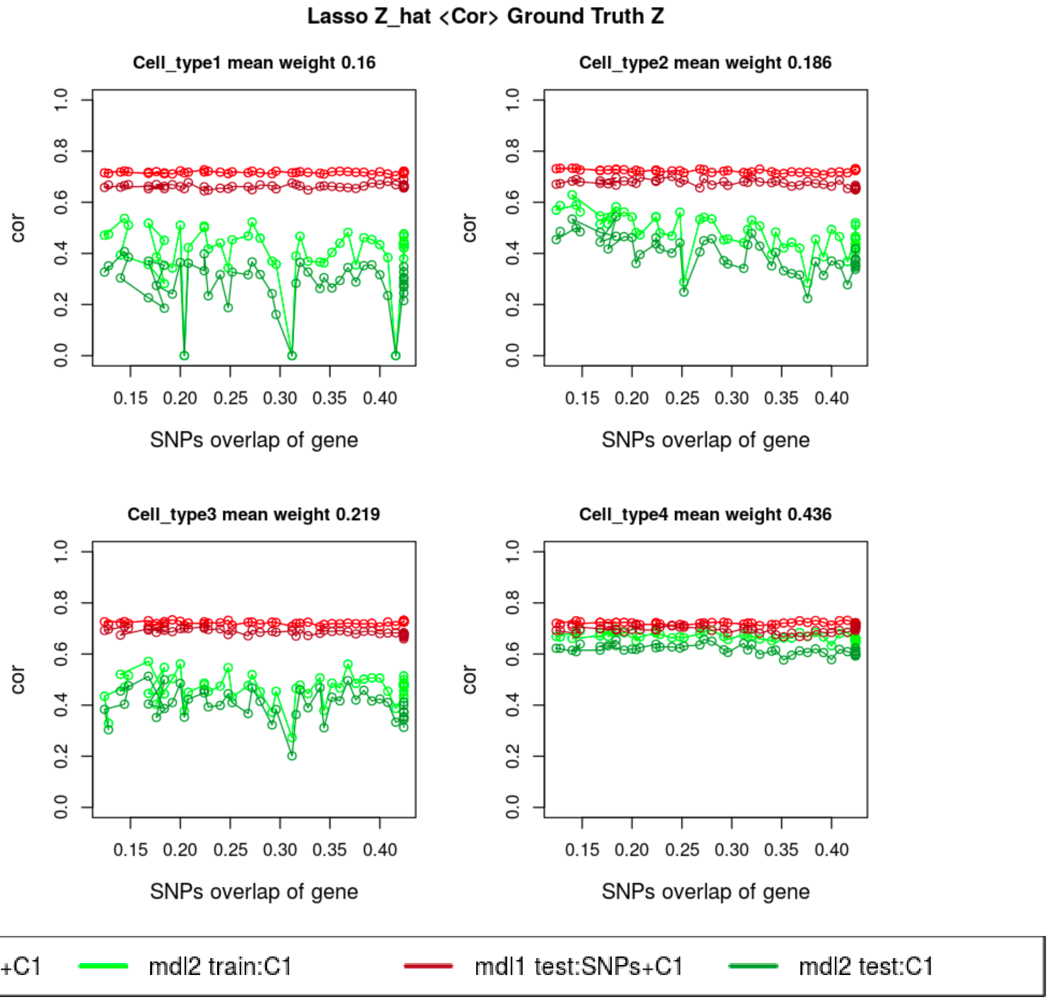
Can the model leverage other structural information inside data?



+

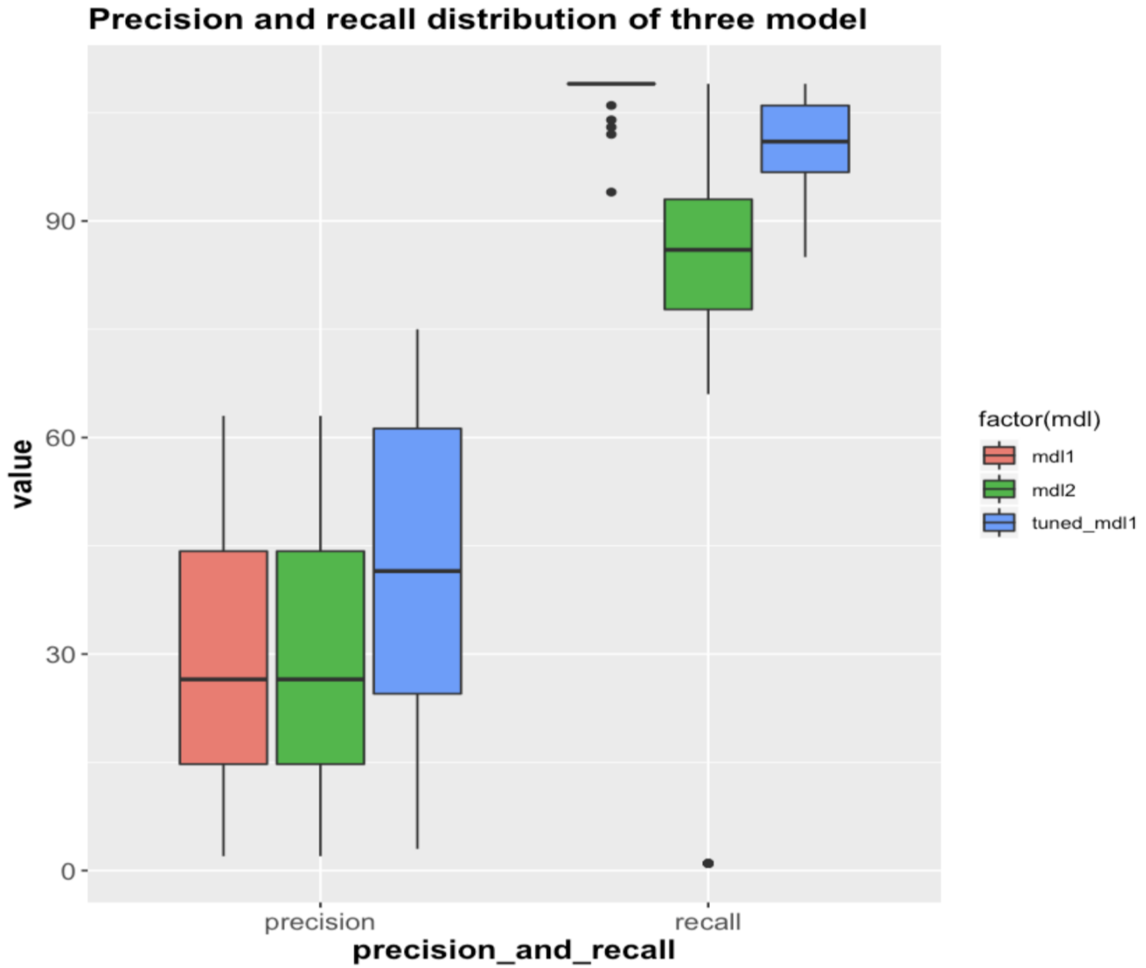
LASSO β

Estimated GE



Dutch Dataset

Is the model's correctly identifying causal effects?



Real Data

TCA-TWAS

Background

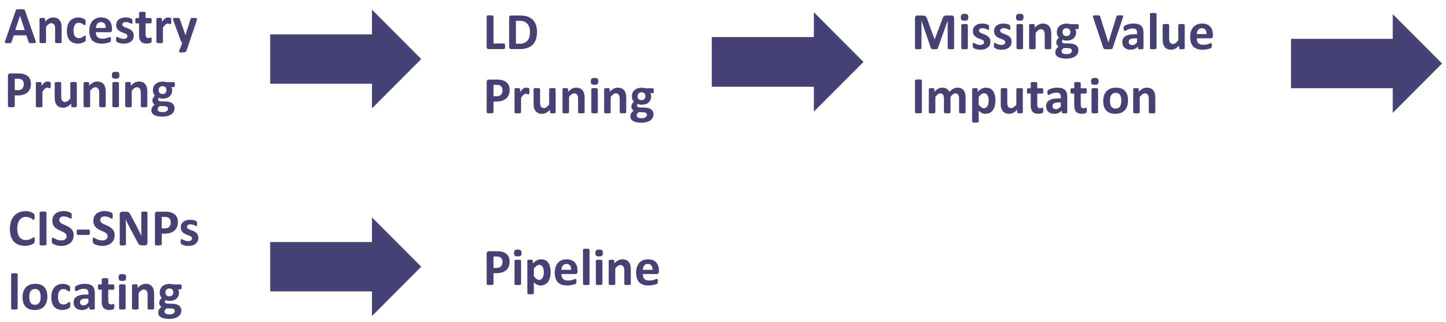
Model

Experiment

Conclusion

Train Dataset: Dutch Dataset with GE

- 5000 individuals, blood, 4 cell types
- 10201 Genes, 801501 SNPs, 10201 GE



Ancestry Pruning → LD Pruning → Missing Value Imputation → CIS-SNPs locating → Pipeline

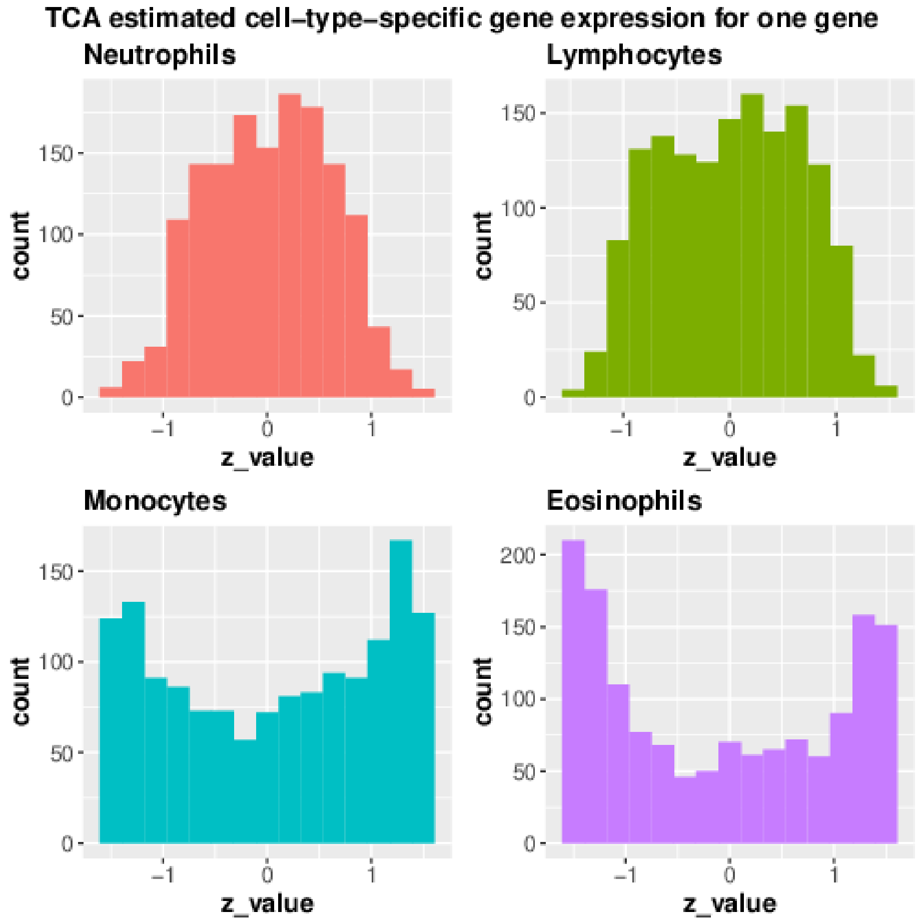
Target Dataset: UKBiobank Dataset without GE

- 500,000 individuals
- SNPs, Phenotypes

File:Central Dogma of Molecular Biochemistry with Enzymes.jpg. (2019, September 4). Wikimedia Commons, the free media repository. Retrieved 00:50, September 5, 2019 from https://commons.wikimedia.org/w/index.php?title=File:Central_Dogma_of_Molecular_Biochemistry_with_Enzymes.jpg&oldid=364559532.

Dutch Dataset

Is the model's performance consistent between cell types?



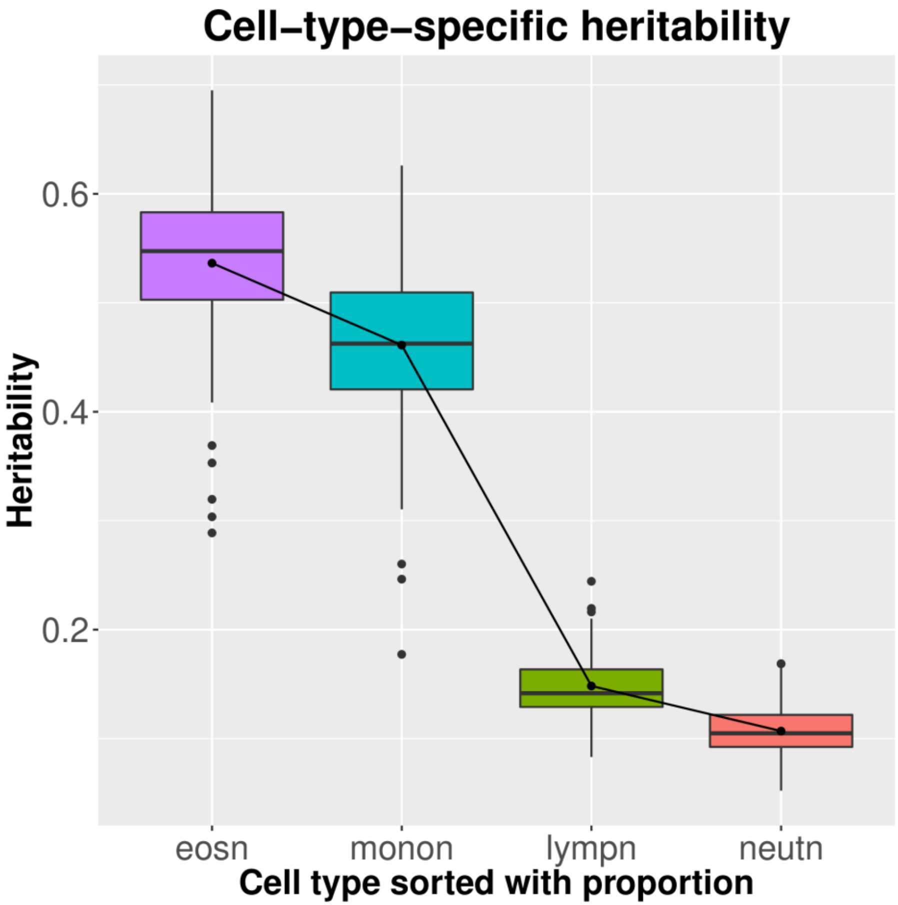
TCA's estimation result fall short of normal distribution, which is the original assumption

Dutch Dataset

TCA-TWAS

Background
Model
Experiment
Conclusion

Is the model's performance consistent between cell types?



Heritability calculated from LASSO's prediction for gene expression. Overfitting occurs for low cell types

Summary

- Cell-type-specific GE can be imputed from SNPs in a relative low cost to conduct the downstream phenotype association study

Key Contribution

- Incorporated genetic effects into TCA to deconvolute bulk level GE into cell-type specific ones
- Produced effect size estimators on Dutch Dataset to impute cell-type specific gene expressions on UKBiobank Dataset
- Functional R package on Cran for standard usage

Future Work

- Add sparsity constraints to TCA parameter estimates
- Utilize extra correlation structure to enhance TCA performance
- Consider batch effects when transferring effect size from train dataset to target dataset

Q&A