Exploring the Heritable Component of Gene Expression Using Machine Learning Methods

Yan Li
Allen J Clark
Sumil Thakrar
Sakshi Arya

PUBH 7475/8475/Stat 8933 Final Project
Machine Learning and Data Mining
University of Minnesota

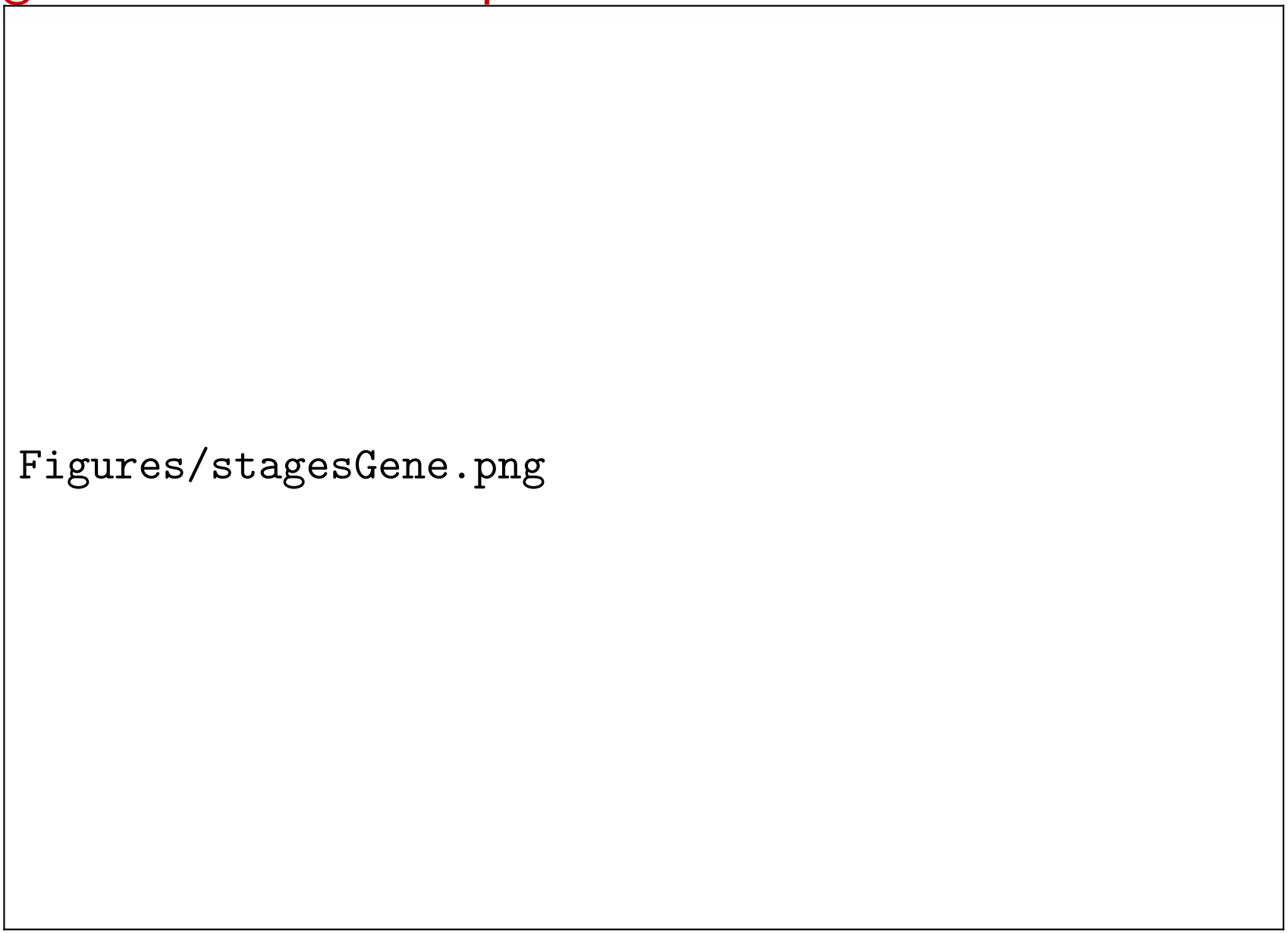
27th April, 2017

Need for Precision Medicine

'An emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment and lifestyle for each person.'

National Institute of Health (NIH)

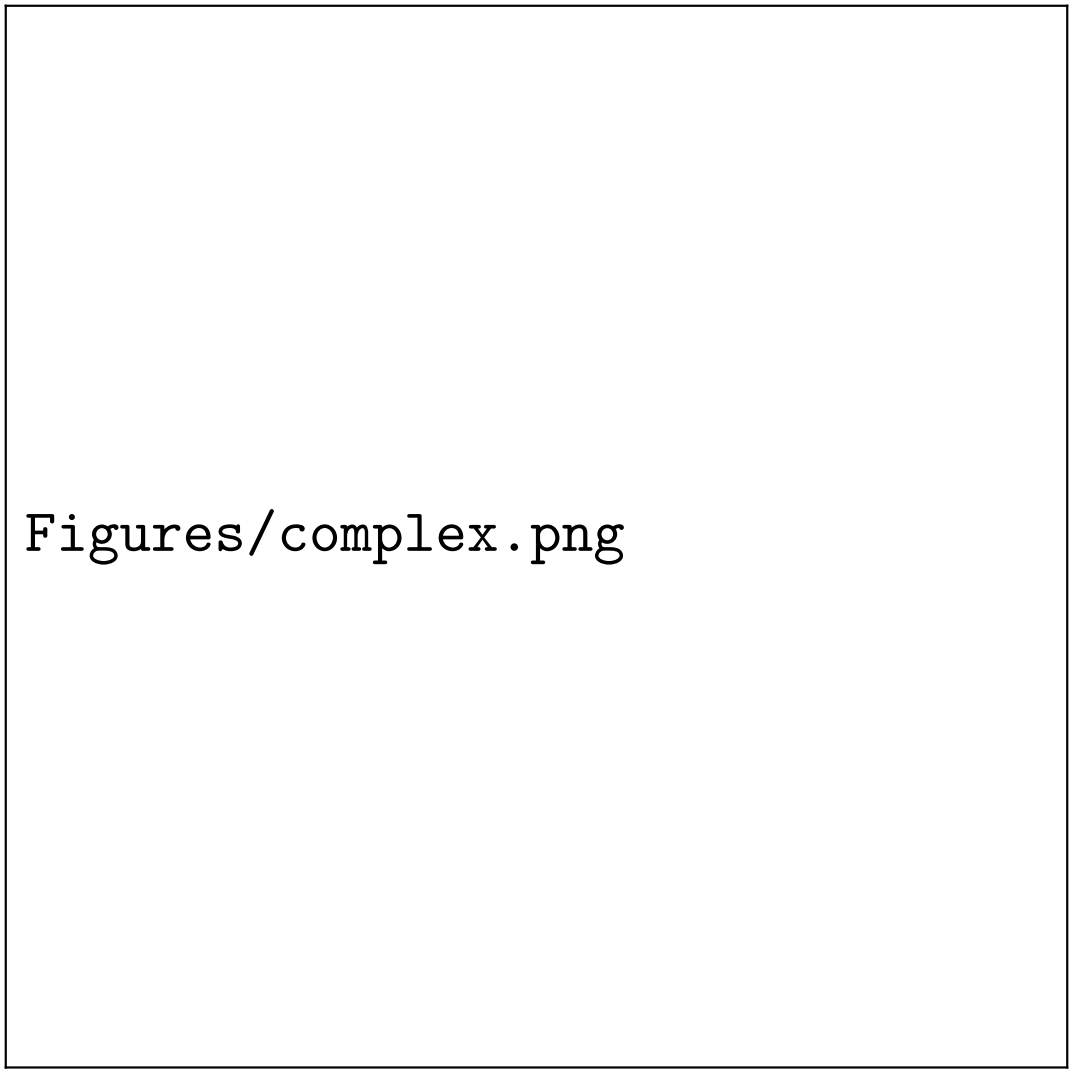
Stages of Genetic interpretation



Samuel J. Aronson, et al. Building the foundation for genomics in precision medicine, Nature, 2015

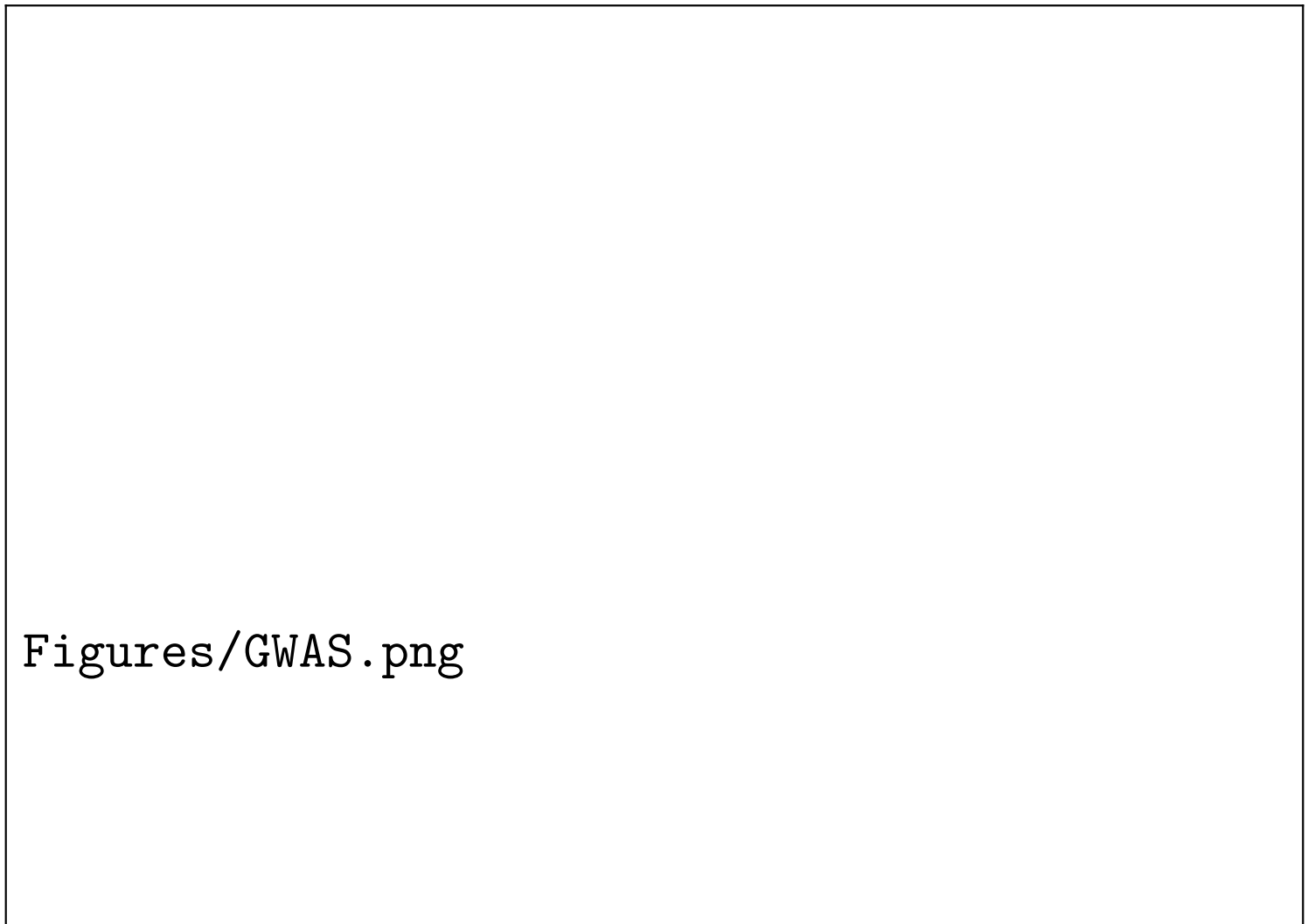
Complex Diseases: Combination of genetic and other factors

- Genetic factors partially associated, multiple genes or genetic variants involved



Genome-wide association study (GWAS)

Examining a genome-wide set of genetic variants (SNPs) in different individuals to see if any SNP is associated with a trait (complex disease)



Single-nucleotide polymorphism (SNPs)

- Variation in a single-nucleotide that occurs at a specific position in the genome (e.g.$>$ 1% of the population)
- Falls in coding and non-coding regions, or in the regions between genes
- Synonymous and nonsynonymous SNPs
- SNPs not in the protein coding regions could still effect gene expression

Challenges in GWAS

- Limited statistical power to identify all disease-associated loci
- Difficulty to interpret the biological relevance of susceptible loci and link with disease etiology.

  Understand how SNPs and diseases are associated biologically - studying eQTLs

eQTLs: Expression quantitative trait loci

- eQTLs are regions of the genome containing DNA sequence variants that influence the expression level of one or more genes
- Cis-eQTLs: SNPs have an effect on local genes
- Trans-eQTLs: SNPs have an effect on distant genes

Understanding eQTLs

?

?

The setup:

Data description

- Measurements from $n = 471$ subjects
- High dimensional data: p ranges from 7000 to 32000 SNPs
- 88 datasets corresponding to 88 genes
- Response: gene expression residuals in normal prostate tissue after regressing out all relevant covariates

index	phenotype	rs111417370_T	rs2003816_T	rs116897926_A	$\cdots$	rs114857962_A
1	0.004	0.0	1	0.0	$\cdots$	0.002
2	0.328	0.0	1	0.0	$\cdots$	0.004
3	0.164	0.4	0	0.0	$\cdots$	0.000
4	-0.107	0.0	1	0.0	$\cdots$	0.989
5	-0.109	0.0	0	0.9	$\cdots$	1.000
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Goals

Predict gene expression in a normal prostrate tissue as a function of SNPs

1. Only focus on using neighborhood SNPs

2. Identify genes that were well predicted

3. For the genes that were not well predicted, extend to SNPs from the same chromosome to see if the predicting effects are better.

Statistical methods

- **Intention:** To compare the existing method, PrediXan (used in the paper), which uses elastic net, with other methods.
- Penalization methods will be good because of the $p > n$ situation

Methods used:

- LASSO
- SCAD
- TLP
- Random Forest
- Elastic Net

Parallel Processing

- Parallel Processing speeds up the code.
- Code run in parallel with one core for each CV fold
- Reduced running time by 56% (450 sec $\rightarrow$ 200 sec)

```
library ( parallel )

no_cores <- 5 # Number of Cores

cl <- makeCluster(no_cores) # Initiate Cluster

# Export Necessary Functions
toexport = c("tlp_fold", ... )
clusterExport(cl, toexport)

parApply(cl = cl, X = CVFolds ,MARGIN = 2, FUN = tlp_fold, data = data)
```

Model evaluation

- Divided data into 80% training and 20% test data sets.
- 5-Fold Cross validation to evaluate the performance of the models.
- Performance measures:
  - MSE
  - Relative Error (MSE/variance(y))

Methods	Parameters tuned
LASSO	λ
SCAD	λ
TLP	λ
Elastic Net	α, λ
Random Forest	*mtry*, *ntree*

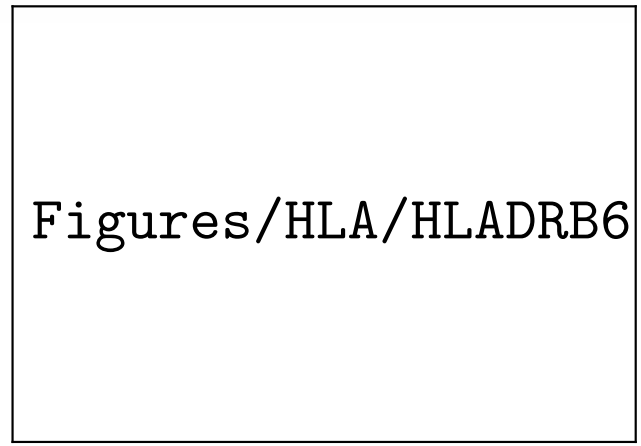
- Repeat the procedure for each method for 88 data sets

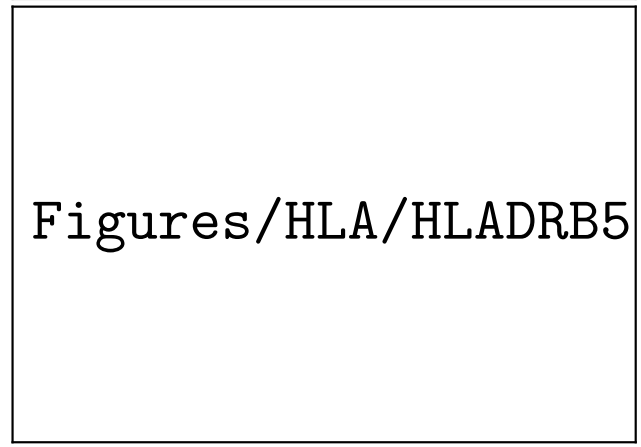
Method comparison

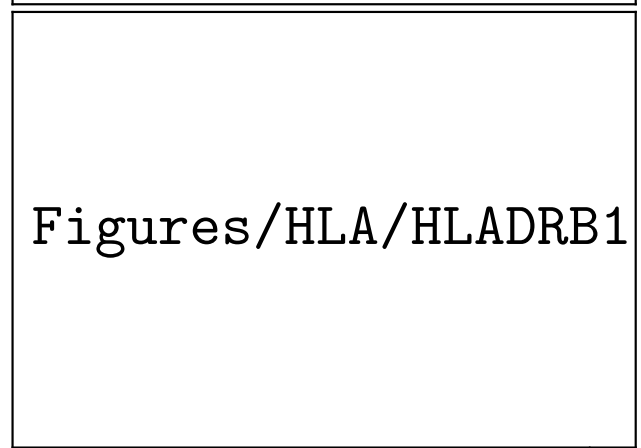
- Multiple pairwise comparisons for comparing average relative errors of different methods

Average relative error vs gene by method (Nbhd SNPs)

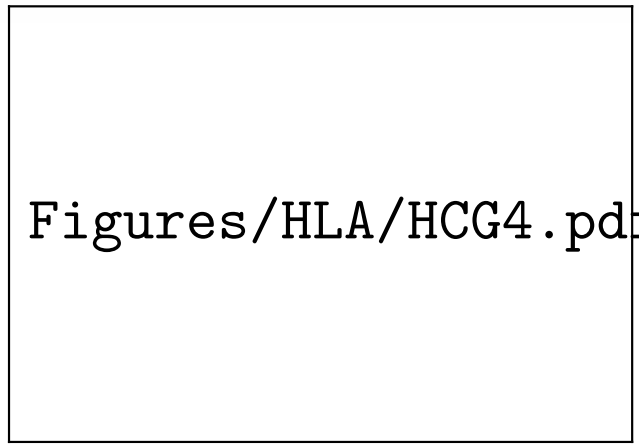
HLA

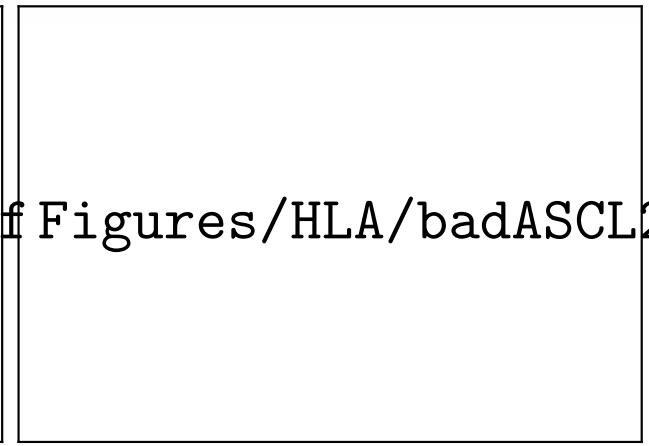

Figures/HLA/HLADRB6.

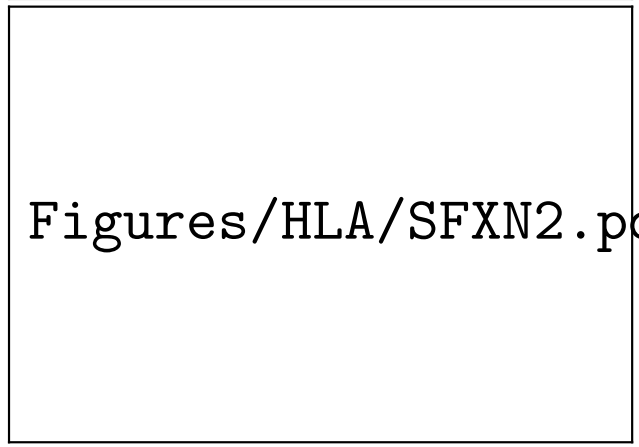

Figures/HLA/HLADRB5.

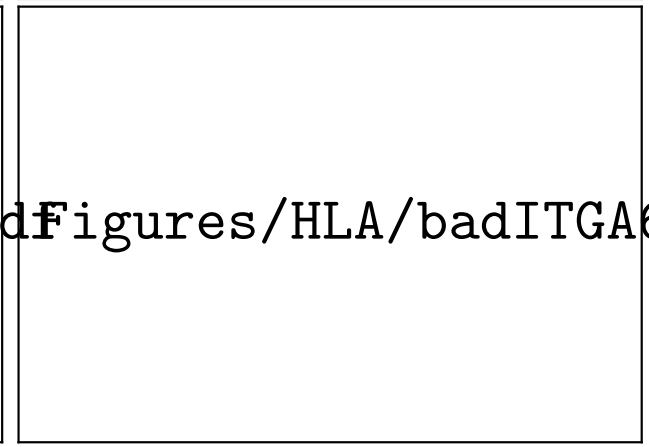

Figures/HLA/HLADRB1.

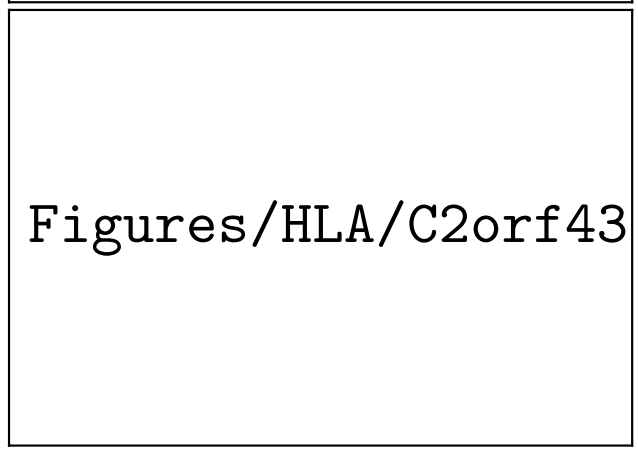
Other well and not so well predicted genes by methods

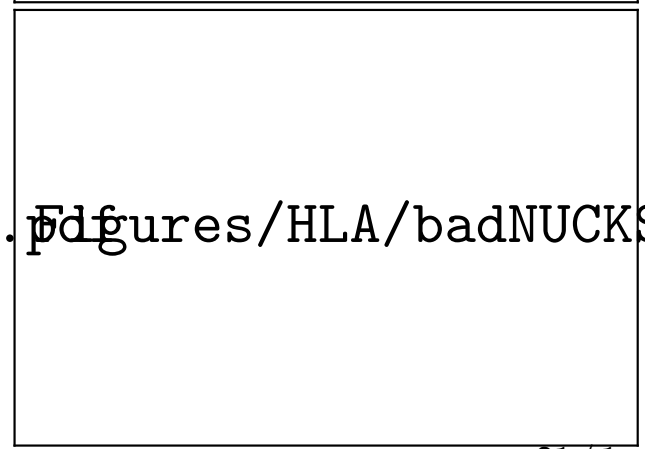






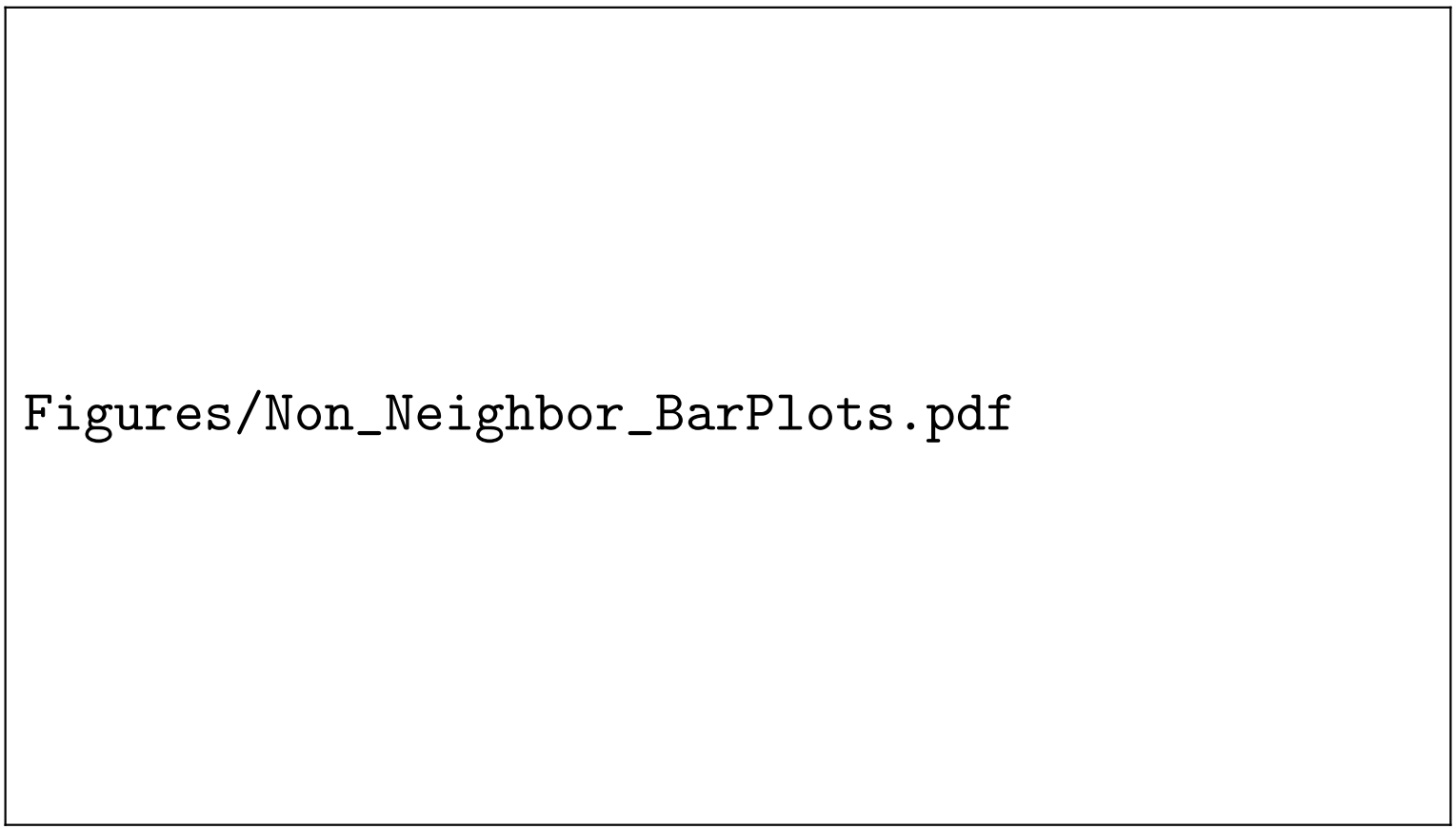






Genes within the same chromosome

- Idea: Use all SNPs from the same chromosome
- Analysis: No improvement for Chromosomes 12 and 22



Badly predicted genes

	Genes	ave	sd
1	ASCL2	0.997	0.089
2	COL2A1	0.991	0.043
3	GDF7	0.951	0.031
4	GGCX	0.974	0.030
5	ITGA6	0.997	0.119
6	KRT18	0.978	0.015
7	KRT8	0.963	0.040
8	LIME1	0.982	0.029
9	LMTK2	0.960	0.036
10	NUCKS1	1.009	0.054
11	TBX5	0.976	0.022
12	ZGPAT	0.958	0.024

Table: Badly predicted genes

Moderately predicted genes

	Genes	Average relative error	sd
7	C10orf32	0.527	0.085
20	FAM57A	0.567	0.099
27	HCG27	0.561	0.076
28	HCG4	0.597	0.157
56	NCOA4	0.545	0.108
71	SFXN2	0.543	0.154
79	TBX1	0.533	0.066

Table: Moderately predicted genes

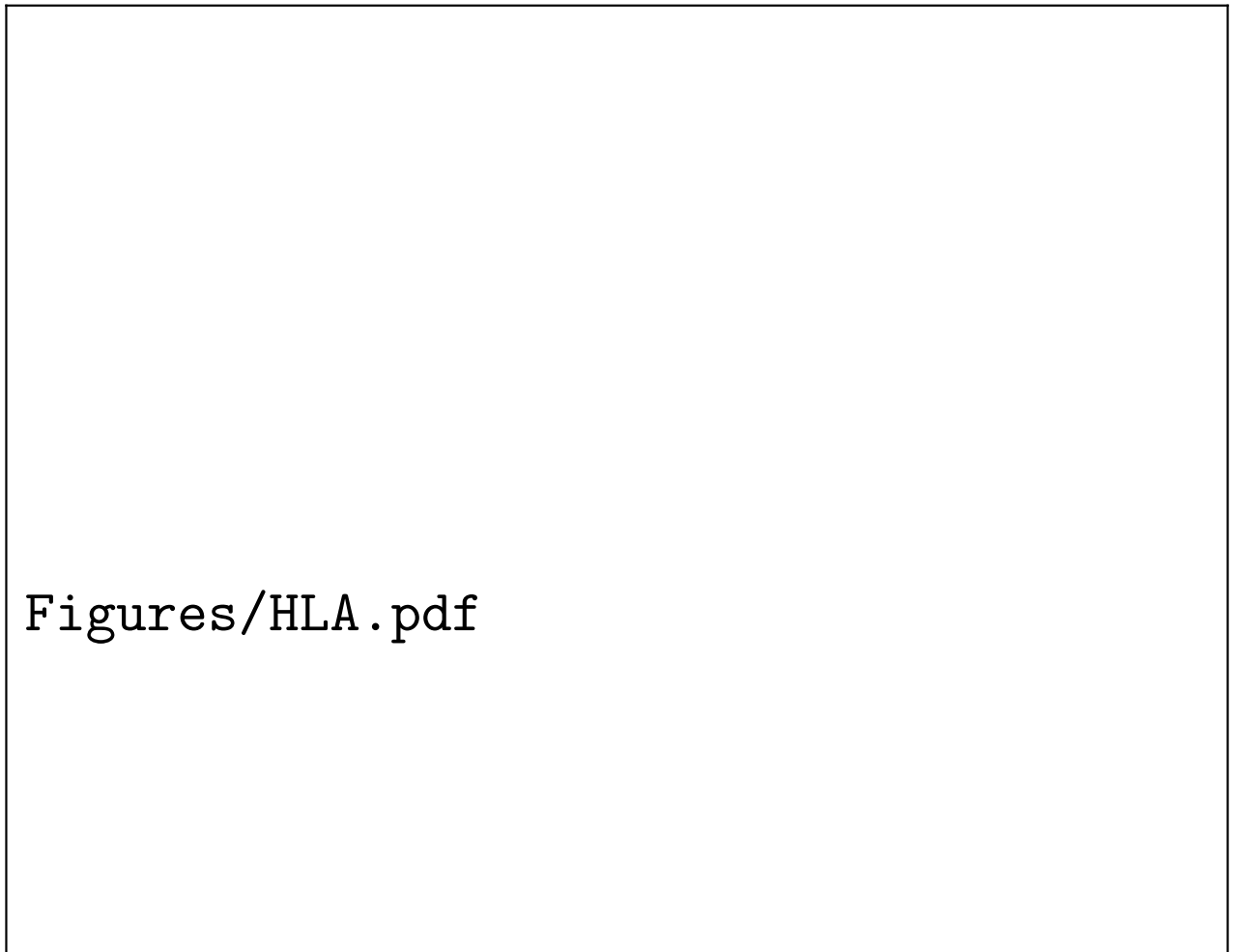
Well predicted genes

	Genes	Average Relative Errors	sd
1	HLA-DQA1	0.393	0.049
2	HLA-DQA2	0.289	0.061
3	HLA-DQB1	0.204	0.023
4	HLA-DQB2	0.461	0.093
5	HLA-DRB1	0.314	0.060
6	HLA-DRB5	0.220	0.025
7	HLA-DRB6	0.372	0.012
8	HLA-L	0.361	0.045
9	IRX4	0.404	0.045
10	LOC284581	0.231	0.068
11	MICA	0.406	0.088
12	NUDT11	0.249	0.029
13	PPP1R14A	0.422	0.050
14	PSORS1C1	0.430	0.063
15	PSORS1C3	0.343	0.047
16	RGS17	0.395	0.035
17	SEPT2	0.425	0.068
18	TMEM180	0.425	0.086
19	VARS2	0.301	0.045

Table: Well predicted genes

HLA

- HLA: the human leukocyte antigen;
  an extensively studied molecule involved in immunity
  has been associated with more than 100 different diseases,
  primarily autoimmune diseases



NUDT11

- NUDT11 locates on chromosome X, and belongs to a subgroup of phosphohydrolases that preferentially attack diphosphoinositol polyphosphates .
- Previously been found to be strongly associated with prostate cancer risk variants (NUDT11, MSMB, NCOA4, SLC22A3, and HNF1B).
- Suppressing the expression of NUDT11 influences cellular phenotypes associated with tumor-related properties in prostate cancer cells.
- rs4907792, which show significantly altered androgen sensitivity in LNCaP cells, is highly associated with eQTL for NUDT11 in prostate tissue.

Variable selection differences from penalized regression methods

Gene: HLA-DQB2.chr6

- 24743 SNPs
- Shrunk to $\sim 100 - 300$

s/venndiagram.pdf

Figures/HLA/HLADQB2edit.pdf

- TLP - simple and interpretable
- SCAD - strange
- To do - check for correlations

Conclusion

- Overall, elastic net performs statistically better than other four methods, but the effect sizes are small.
- Prediction performance varies among different genes for cis-eQTL.
- Different genes may require different method for good prediction effect, due to non-linear association. Random forest performed much better for 3 HLA genes.
- Some well predicted genes have interesting biological relevance.

Future Directions

- Evaluate other machine learning methods, such as SVM.
- Explore how trans-eQTLs predict gene expression: 1) extend to 5Mbp or 10Mbp up and down stream of genes; 2) use SNPs from different chromosome for gene prediction(need to consider correlation between different genes).
- Identify causal SNPs from eQTLs for good predicted genes.

References