Genome-wide Association Study of Neutrophils in Mice

Adam Busis, Jen Kose, Tony Macula, Zach Pollock

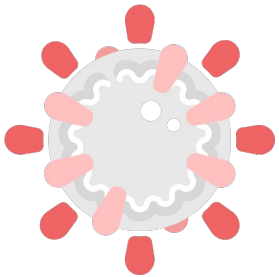
Features: Single-Nucleotide Polymorphisms (SNPs)

- Most DNA is the same between mice but there are some mutations
- SNPs are positions in DNA with two different possible nucleotides
- Each chromosome in a pair can have either option
- We assigned numbers (e.g. CC = 0, CT = 1, TT = 2)
- ~6000 SNPs, and only ~700 mice

SNP ID	JAXW202.2129	JAXW202.2130	JAXW202.2132	JAXW202.2133	JAXW202.2134
JAX00240603	TC	TT	CC	TC	TC
UNC010515443	GG	GG	AG	AG	GG
UNC010001943	NaN	AA	CC	AA	NaN
UNC010515539	AG	AG	AG	AG	GG
UNC010515556	GG	GG	AG	AA	AG

Target: Neutrophil count

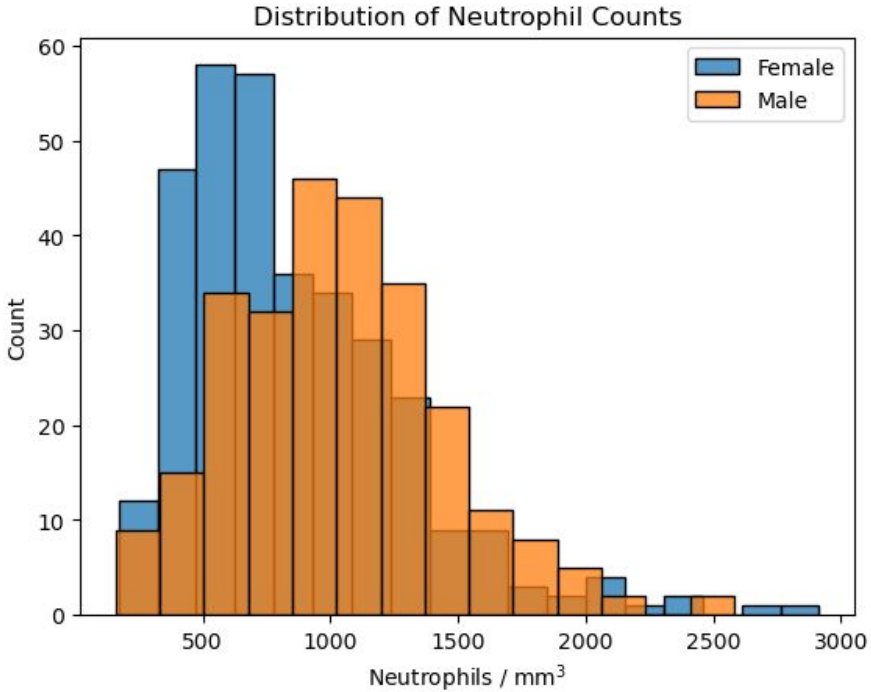
- Neutrophils are a type of white blood cell
- Measured in cells/mm^3
- Deficiency is known as neutropenia



- Data from The Jackson Laboratory Mouse Phenome Database
- Mouse data is more accessible than human!

Mouse Phenome Database
at The Jackson Laboratory

Baseline: Male vs Female Mice



- Female and male have different distributions
- For a baseline model, ignore genotype and take the average for each sex

Improved Model: *F*-tests

- For each SNP, try adding it to the linear regression model:

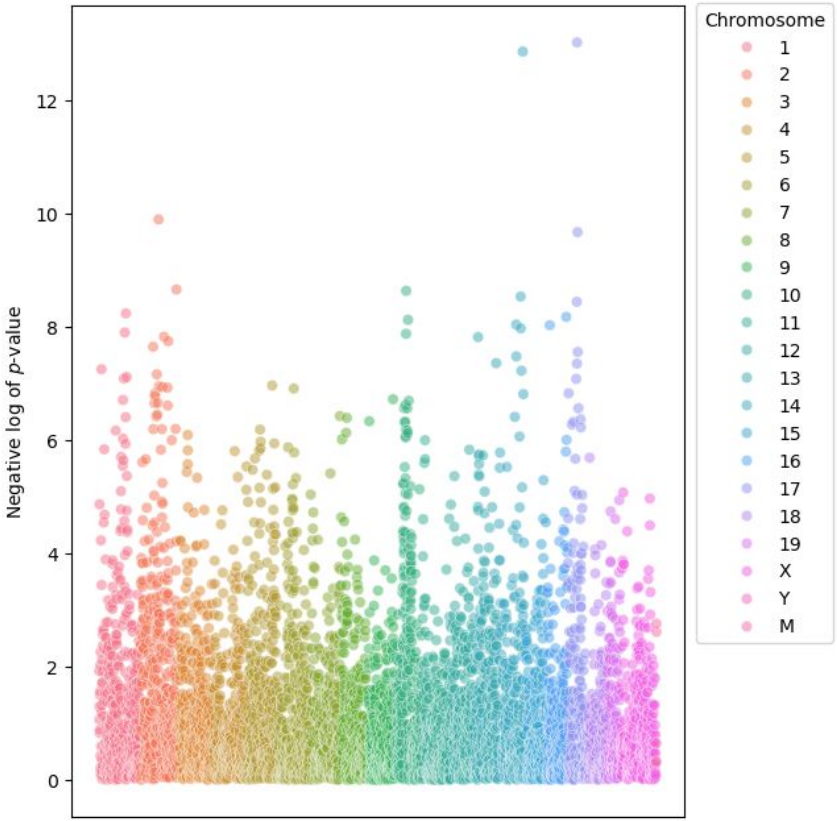
  $$y = \beta_0 + \beta_1 (\text{Sex}) + \beta_2 (\text{SNP})$$

- Compare to baseline:

  $$y = \beta_0 + \beta_1 (\text{Sex})$$

- Compare them with an *F*-test, look for significant *p*-values
- For significance level, use Bonferroni correction: $\alpha \rightarrow \alpha / n$
  - (In our case, $0.5 / 6000 = 8 * 10^{-6}$)

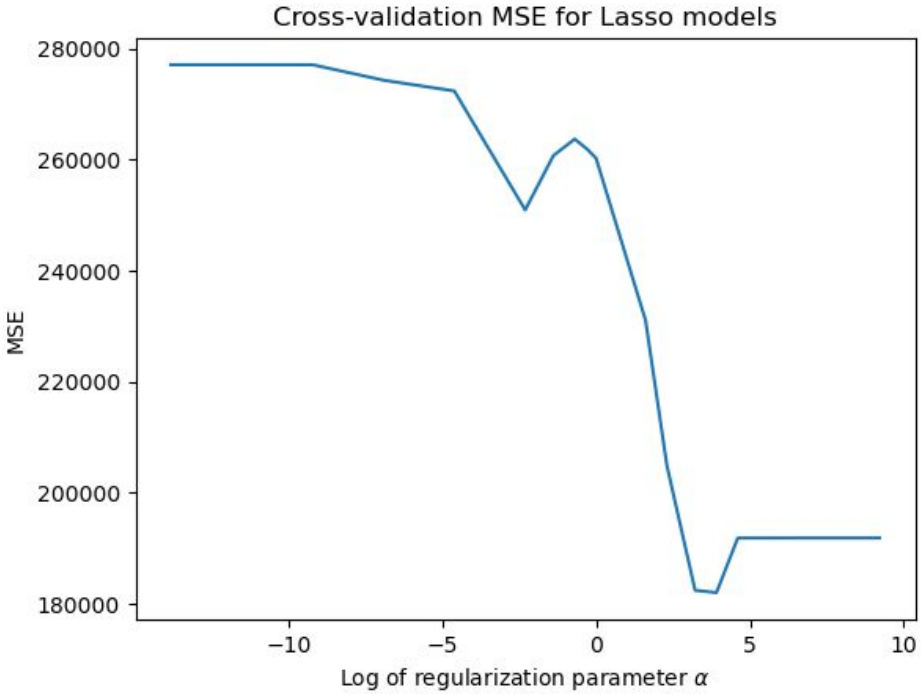
Improved Model: *F*-tests



- We found two statistically significant SNPs
- Improved model:

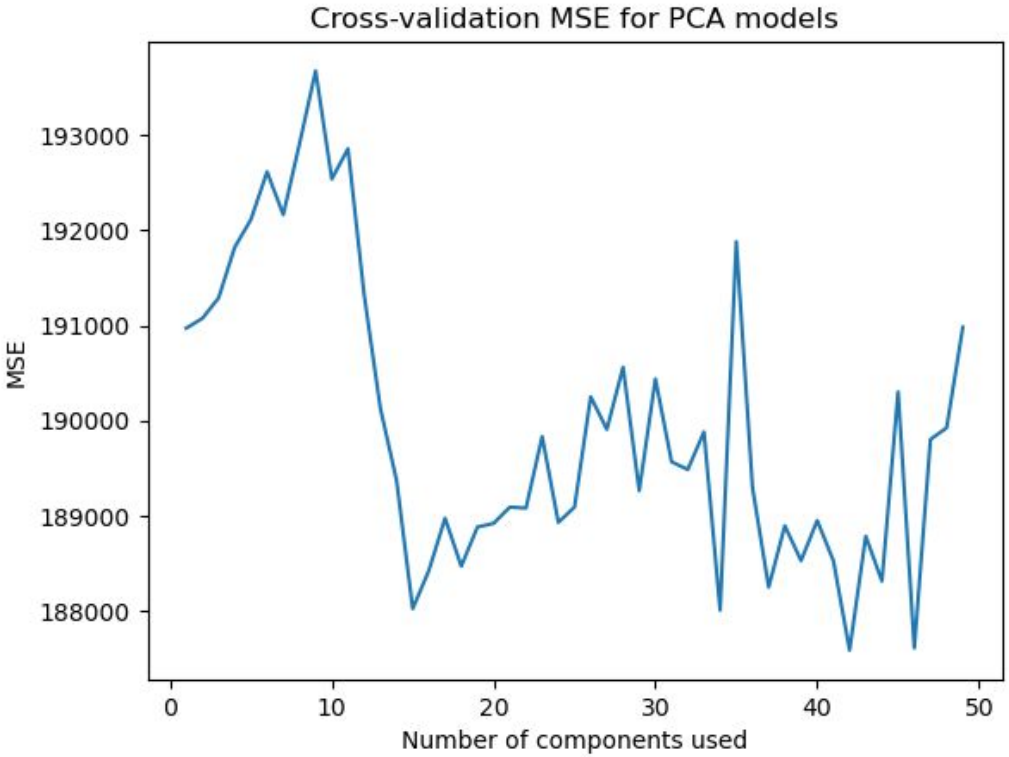
$$y = \beta_0 + \beta_1 (\text{Sex}) + \beta_2 (\text{SNP1}) + \beta_3 (\text{SNP2})$$

Further Models: Lasso



- Lasso regression has built-in feature selection by assigning most coefficients to 0
- We can do Lasso on all SNPs and let the regularization select which to use
- Minimum-MSE model had 40 nonzero coefficients

Further Models: PCA



- We can take a PCA of the genotype data and look at the most significant components
- Unsupervised dimension reduction reduces bias but might lose important information
- Ultimately this did not improve on the baseline

Cross-Validation MSE

Model	Description	CV MSE
0: Baseline	Linear regression using just sex	184250
1: Selected linear	Linear regression using sex and two significant SNPs	171906
2: Significant with interaction	Linear regression using the same SNPs as model 1, but with quadratic interaction terms	175384
3: Lasso	Lasso regression using sex and all SNPs with $\alpha = 50$	182051
4: PCA	Linear regression using the first 44 principal components	187587

- Linear regression with the two significant SNPs seems like the best model

Conclusion

- Unfortunately, the new model didn’t improve on the baseline in testing
- Were the two SNPs just false positives? Or did we get unlucky in the testing set?
- More data would be helpful! Most of the variation in neutrophil count isn’t explained by the genome
- It would be interesting to look at interactions between multiple SNPs

Model	Testing MSE
0: Baseline	163729
1: Selected linear	175798