Lecture 5: GWAS (cont.)

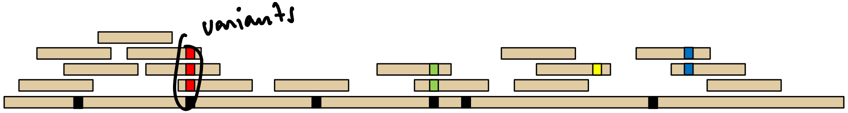
ECE 365 - Data Science and Genomics

Announcements:

- Lab 2 (Sequence alignment) due on Thursday
- Lab 3 (GWAS) released tomorrow

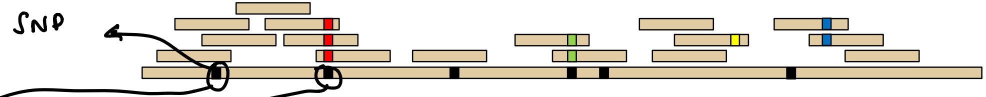
Genotype data

- Focus on a set of common variants in the genome



Genotype data

- □ Focus on a set of common variants in the genome



- □ Genotype data (VCF file) looks like this:

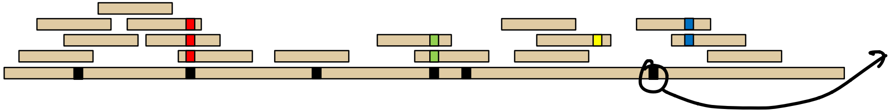
paternal

maternal

$$\begin{array}{l|cccccc|} & \text{person 1} & & & & & \text{person n} \\ \text{SNP 1} & 0|0 & 1|0 & 1|0 & 0|0 & 0|0 & 0|1 \\ \text{SNP 2} & 0|1 & 0|0 & 1|1 & 1|0 & 0|0 & 1|1 \\ & 0|1 & 0|1 & 0|1 & 0|0 & 0|0 & 0|0 \\ & 0|0 & 0|0 & 0|0 & 1|1 & 0|1 & 1|0 \\ & 1|0 & 0|1 & 1|1 & 0|0 & 0|0 & 0|0 \\ & 0|0 & 1|1 & 0|0 & 0|0 & 0|1 & 0|0 \\ \text{SNP m} & 0|1 & 0|1 & 0|0 & 1|0 & 0|0 & 1|1 \end{array}$$

Genotype data

- □ Focus on a set of common variants in the genome



REF = C (0)
ALT = T (1)

- □ Genotype data (VCF file) looks like this:

$$
\begin{array}{ccccccc}
 & \text{person 1} & & & & & \text{person n} \\
\text{SNP 1} & 0|0 & 1|0 & 1|0 & 0|0 & 0|0 & 0|1 \\
\text{SNP 2} & 0|1 & 0|0 & 1|1 & 1|0 & 0|0 & 1|1 \\
 & 0|1 & 0|1 & 0|1 & 0|0 & 0|0 & 0|0 \\
 & 0|0 & 0|0 & 0|0 & 1|1 & 0|1 & 1|0 \\
 & 1|0 & 0|1 & 1|1 & 0|0 & 0|0 & 0|0 \\
 & 0|0 & 1|1 & 0|0 & 0|0 & 0|1 & 0|0 \\
\text{SNP m} & 0|1 & 0|1 & 0|0 & 1|0 & 0|0 & 1|1
\end{array}
\xrightarrow{\text{count ALT occurrences}}
\begin{bmatrix}
0 & 1 & 1 & 0 & 0 & 1 \\
1 & 0 & 2 & 1 & 0 & 2 \\
1 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 2 & 1 & 1 \\
1 & 1 & 1 & 0 & 0 & 0 \\
0 & 2 & 0 & 0 & 1 & 0 \\
1 & 1 & 0 & 1 & 0 & 2
\end{bmatrix}
$$

Genome-Wide Association Studies (GWAS)

$$\begin{array}{cc}
 & \begin{array}{cccccc}\text{SNP 1} & & & & & \text{SNP m}\end{array} \\
\begin{array}{r}\text{person 1}\\ \\ \\ \\ \\ \\ \text{person n}\end{array} & \underbrace{\begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 2 & 1 & 0 & 2 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 2 \end{bmatrix}}_{\text{genome-wide}}
\end{array}
\qquad
\begin{array}{c}\text{phenotype} \\ \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ 1 \end{bmatrix}\end{array}
\nearrow$$

e.g. has Diabetes

(doesn't need to be binary. e.g., height)

Genome-Wide Association Studies (GWAS)

associated

$$\begin{array}{rcccccccc}
 & \text{SNP 1} & & \text{SNP}_{200} & & & \text{SNP m} & & \text{phenotype} \\
\text{person 1} & 0 & 1 & 1 & 0 & 0 & 1 & & 0 \\
 & 1 & 0 & 2 & 1 & 0 & 2 & & 0 \\
 & 1 & 1 & 1 & 0 & 0 & 0 & & 1 \\
 & 0 & 0 & 0 & 2 & 1 & 1 & & 1 \\
 & 1 & 1 & 1 & 0 & 0 & 0 & & 1 \\
 & 0 & 2 & 0 & 0 & 1 & 0 & & 0 \\
\text{person n} & 1 & 1 & 0 & 1 & 0 & 2 & & 1 \\
\text{new person} & 1 & 0 & 1 & 1 & 2 & 0 & & ? \quad (\text{predict})
\end{array}$$

- □ Which SNPs are associated with a given phenotype?
- □ Given a new individual's genotype, can you predict their phenotype?

Revisiting Logistic Regression

- □ Predict binary variable from real-valued features

$$\underset{\overset{\shortparallel}{p}}{p(1 \mid \underline{x})} = \frac{e^{\beta_0 + \underline{\beta}^T \underline{x}}}{1 + e^{\beta_0 + \underline{\beta}^T \underline{x}}} = \frac{1}{1 + e^{-(\beta_0 + \underline{\beta}^T \underline{x})}}$$

$$1 + e^{-(\beta_0 + \underline{\beta}^T \underline{x})} = \frac{1}{p} \Rightarrow e^{-(\beta_0 + \underline{\beta}^T \underline{x})} = \frac{1}{p} - 1 = \frac{1-p}{p}$$

$$\rightarrow \ln\left(\frac{p}{1-p}\right) = \underbrace{\beta_0 + \underline{\beta}^T \underline{x}}_{\text{linear model}}$$

$\frac{p}{1-p}$: odds ratio $\in (0, \infty)$

$\ln \frac{p}{1-p}$: log odds ratio $\in (-\infty, \infty)$

Revisiting Logistic Regression

- □ Can we use Logistic Regression for GWAS?

$$\begin{array}{r}\\ \\ \text{person 1}\\ \\ \underline{x}\\ \\ \\ \\ \text{person n}\end{array}
\begin{array}{c}
\overbrace{\text{SNP 1}\hspace{8em}\text{SNP m}}^{\text{millio}}\\
\begin{bmatrix}
0 & 1 & 1 & 0 & 0 & 1\\
1 & 0 & 2 & 1 & 0 & 2\\
1 & 1 & 1 & 0 & 0 & 0\\
0 & 0 & 0 & 2 & 1 & 1\\
1 & 1 & 1 & 0 & 0 & 0\\
0 & 2 & 0 & 0 & 1 & 0\\
1 & 1 & 0 & 1 & 0 & 2
\end{bmatrix}
\end{array}
\quad
\begin{array}{c}
\text{phenotype}\\
\begin{bmatrix}0\\0\\1\\1\\1\\0\\1\end{bmatrix}
\end{array}$$

$$p(1 \mid \underline{x}) = \frac{1}{1 + e^{-(\beta_0 + \underline{\beta}^T \underline{x})}}$$

risk of having a disease

Revisiting Logistic Regression

□ Can we use Logistic Regression for GWAS?

$$\text{1000s}\left\{\begin{array}{l} \text{person 1} \\ \\ \\ \\ \\ \\ \text{person n} \end{array}\right. \begin{array}{c} \overbrace{\text{SNP 1} \qquad\qquad\qquad\qquad \text{SNP m}}^{\text{millions}} \\ \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 2 & 1 & 0 & 2 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 2 \end{bmatrix} \end{array} \quad \begin{array}{c} \text{phenotype} \\ \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ 1 \end{bmatrix} \end{array}$$

□ Problem: Number of SNPs can be $\sim 10^6$

Revisiting Logistic Regression

- □ Can we use Logistic Regression for GWAS?

$$\begin{array}{ccc}
 & \text{SNP 1} \qquad\qquad\qquad\qquad \text{SNP m} & \text{phenotype} \\
\begin{array}{c}\text{person 1}\\ \\ \\ \\ \\ \\ \text{person n}\end{array} & \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 2 & 1 & 0 & 2 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 2 \end{bmatrix} & \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ 1 \end{bmatrix}
\end{array}$$

- □ Problem: Number of SNPs can be $\sim 10^6$

GWAS via *univariate* logistic regression

- □ Run a separate logistic regression for each SNP

$$
\begin{array}{c|cccccc|c|}
 & \text{SNP 1} & & & x_i & & \text{SNP m} & \text{phenotype} \\
\hline
\text{person 1} & 0 & 1 & 1 & \boxed{0} & 0 & 1 & 0 \\
 & 1 & 0 & 2 & \boxed{1} & 0 & 2 & 0 \\
 & 1 & 1 & 1 & \boxed{0} & 0 & 0 & 1 \\
 & 0 & 0 & 0 & \boxed{2} & 1 & 1 & 1 \\
 & 1 & 1 & 1 & \boxed{0} & 0 & 0 & 1 \\
 & 0 & 2 & 0 & \boxed{0} & 1 & 0 & 0 \\
\text{person n} & 1 & 1 & 0 & \boxed{1} & 0 & 2 & 1
\end{array}
$$

$$p(1 \mid x_i) = \frac{1}{1 + e^{-(\beta_0 + \boxed{\beta_i} x_i)}}$$

Captures association between i th SNP and phenotype

- □ Use this to identify small subset of SNPs associated with phenotype
- □ Let's look at some examples on a Jupyter notebook

GWAS via *univariate* logistic regression

- We will get a β for each SNP

$$\begin{array}{rcc}
 & \begin{array}{cccccc}\text{SNP 1} & & & & & \text{SNP m}\end{array} & \text{phenotype} \\
\begin{array}{r}\text{person 1}\\ \\ \\ \\ \\ \\ \text{person n}\end{array} & \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 2 & 1 & 0 & 2 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 2 \end{bmatrix} & \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ 1 \end{bmatrix} \\
 & \begin{array}{cccccc}\beta_1 & \beta_2 & \cdots & & & \beta_m\end{array} &
\end{array}$$

model for ith SNP :

$$\ln\left(\frac{p(1 \mid x_i)}{1 - p(1 \mid x_i)}\right) = \underset{\downarrow}{\beta_0} + \beta_i x_i$$

LOR for reference genome (no ALT SNPs)

GWAS via *univariate* logistic regression

- Idea: combine all beta coefficients into a single model:

$$\ln\left(\frac{p(1|x)}{1-p(1|x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$$

GWAS via *univariate* logistic regression

- Idea: combine all beta coefficients into a single model:

$$\ln\left(\frac{p(1|x)}{1-p(1|x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$$

- Problems with this approach:

GWAS via *univariate* logistic regression

- Idea: combine all beta coefficients into a single model:

$$\ln\left(\frac{p(1|x)}{1-p(1|x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$$

- Problems with this approach:
  - Since m is very large, some β_is will be large **by chance**

GWAS via *univariate* logistic regression

- Idea: combine all beta coefficients into a single model:

$$\ln\left(\frac{p(1|x)}{1-p(1|x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$$

- Problems with this approach:
  - Since m is very large, some β_is will be large **by chance**
  - Some x_is are correlated

(e.g., anyone with $x_3 = 1$ has $x_4 = 1$)

Identifying statistically significant SNPs

- To measure the significance of the association, we use the p-value

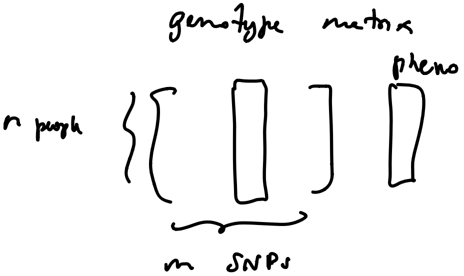
Identifying statistically significant SNPs

- To measure the significance of the association, we use the p-value
  - Probability that the coefficient β_i would be obtained by chance if there was **no** association

$$p\left(|\hat{\beta}_i| > |\beta_i| \text{ found} \;\middle|\; \text{no association}\right)$$

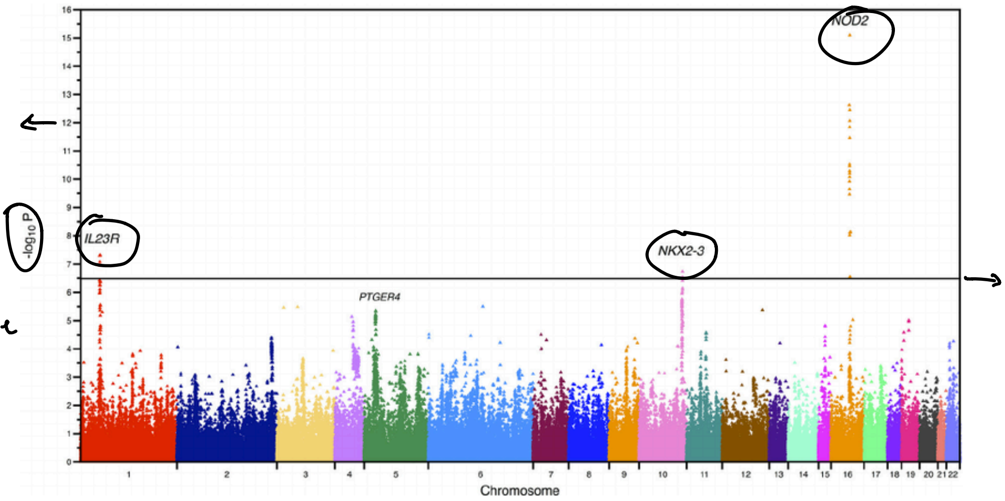
- We can use the statsmodels Python package to perform the logistic regressions and compute the p-values
- Let's return to the jupyter notebook

Manhattan plots



- □ Allow us to see the significance of all SNPs in the genome
- □ We plot $-\log_{10}(p\text{–value})$

Manhattan plot of SNP association with Irritable Bowel Disease



$10^{-12} \leftarrow$

$-\log_{10}$ p-value

$\rightarrow$ threshold