Evaluation of extending Proxy External Control Association Test (ProxECAT) to Poisson Regression

Makayla Cowles

May 3, 2022

Department of Mathematical & Statistical Sciences
UNIVERSITY OF COLORADO **DENVER**

Background

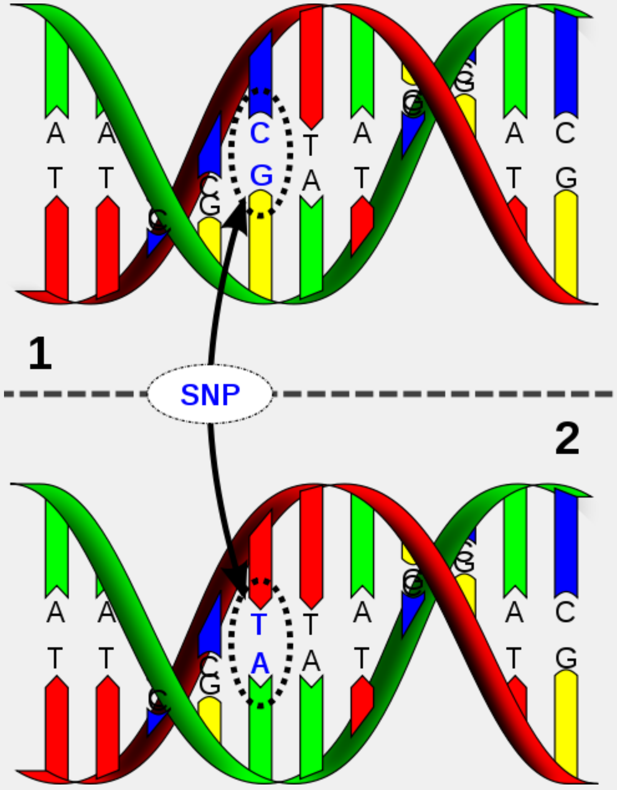
Objective

Methods

Results

Future Work

- Alleles: base pairs that differ
- Variants: location of the alleles
  - Common Variants: MAF > 0.05
  - Rare Variants: MAF < 0.01
- $$\text{Minor Allele Frequency (MAF)} = \frac{\textit{total \# alternate alleles in the observed variant}}{\textit{total \# of alternate alleles in a population}}$$

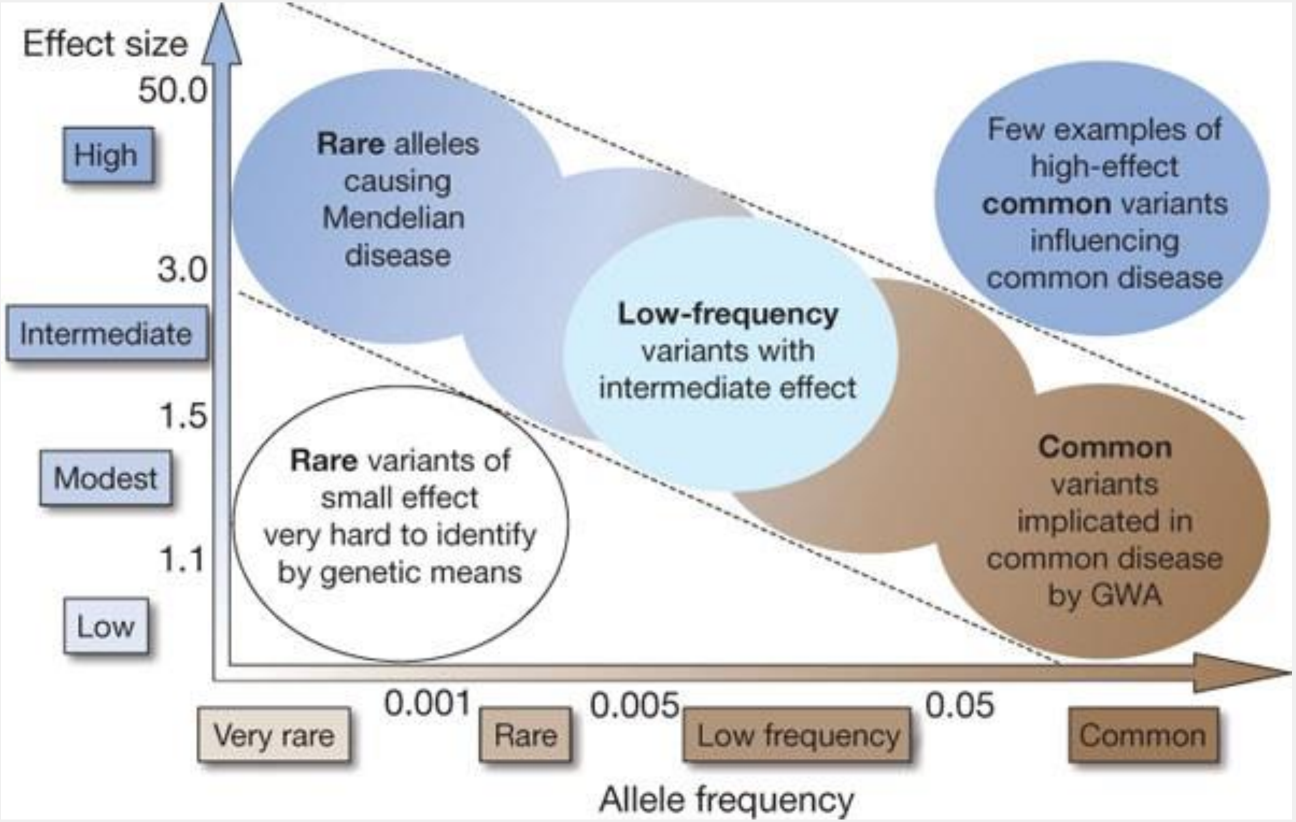


https://isogg.org/wiki/Single-nucleotide_polymorphism

Rare Variants

- Functional: alter a gene’s function
- Non-Functional (Synonymous): have no effect on the gene’s function

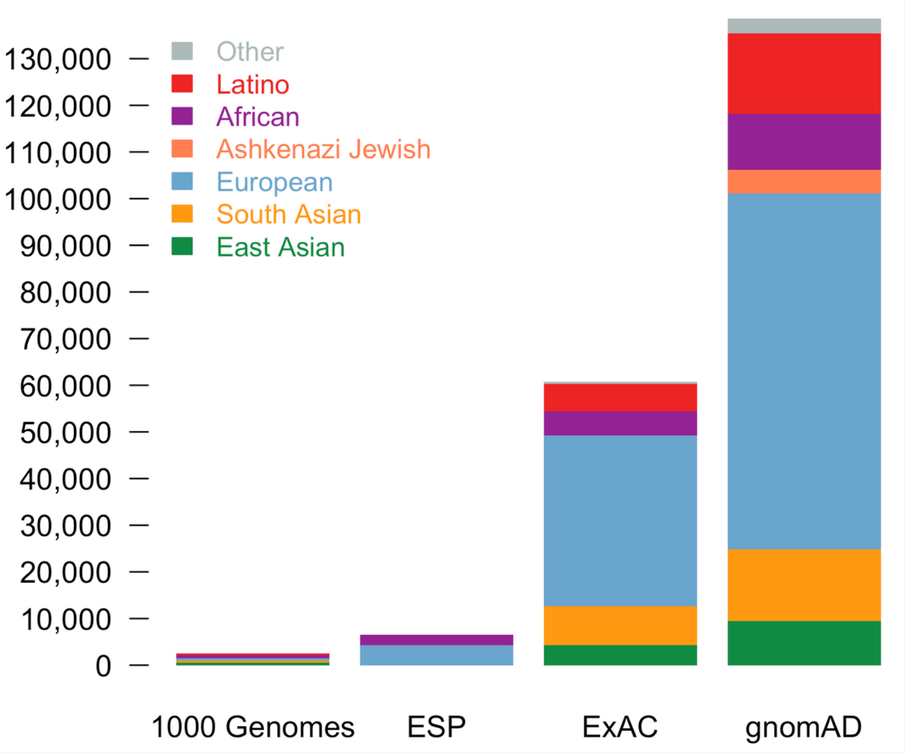
Genome Wide Association Study (GWAS)


Effect size
50.0
High
3.0
Intermediate
1.5
Modest
1.1
Low
Rare alleles causing Mendelian disease
Few examples of high-effect common variants influencing common disease
Low-frequency variants with intermediate effect
Rare variants of small effect very hard to identify by genetic means
Common variants implicated in common disease by GWA
Very rare
0.001
Rare
0.005
Low frequency
0.05
Common
Allele frequency

TA Manolio *et al. Nature* **461**, 747-753 (2009) doi:10.1038/nature08494

External Controls for Rare Variant Association Test

- 10K to 100K sample sizes needed for adequate power
- Use external controls to increase power
- Public databases contain genetic summary data to be used as external controls
- Case control tests using external controls can be biased due to sequencing differences between cases and controls



https://gnomad.broadinstitute.org/news/2017-02-the-genome-aggregation-database/

Proxy External Controls Association Test (ProxECAT)

Can use very rare variants

- Singletons and Doubletons

Optimal when no or limited controls exist

- Only requires external controls

Utilizes both functional and synonymous variants

- synonymous variants are used as a “proxy” for how well rare variants are sequenced within a gene region

ProxECAT Limitations

- Does not enable internal controls to be analyzed with external controls
- Cannot adjust for covariates such as sex, ancestry, or proportion of alternate variant reads or depth of coverage

★ Extend ProxECAT to a Poisson Regression

Why?

- Regression can control for covariates
- Both internal and external controls can be evaluated together
- Rare allele counts are approximately distributed as a Poisson distribution

How?

- Compare results from Poisson regression to ProxECAT and logistic regression

Observational Units

Unit about which information is collected

In case control studies individuals is often the observational unit

Individual level data is hard to access

ProxECAT

- Alternate alleles are used as the observational unit
- Alternate allele counts are modeled as a random sample of four independent Poisson distributions

$$X_1^f \sim Poisson\left(\lambda_1^f\right),\ X_0^f \sim Poisson\left(\lambda_0^f\right), X_1^p \sim Poisson\left(\lambda_1^p\right), \text{and } X_0^p \sim Poisson\left(\lambda_1^p\right)$$

ProxECAT Data Notation

		Predicted Functional Impact		Total
		Functional	Not Functional (Proxy)	
Cases (Internal)	Y = 1	x_1^f	x_1^p	x_1
Controls (External)	Y = 0	x_0^f	x_0^p	x_0
Total		x^f	x^p	N

x – number of alternate alleles

N – total number of alternate alleles in gene region

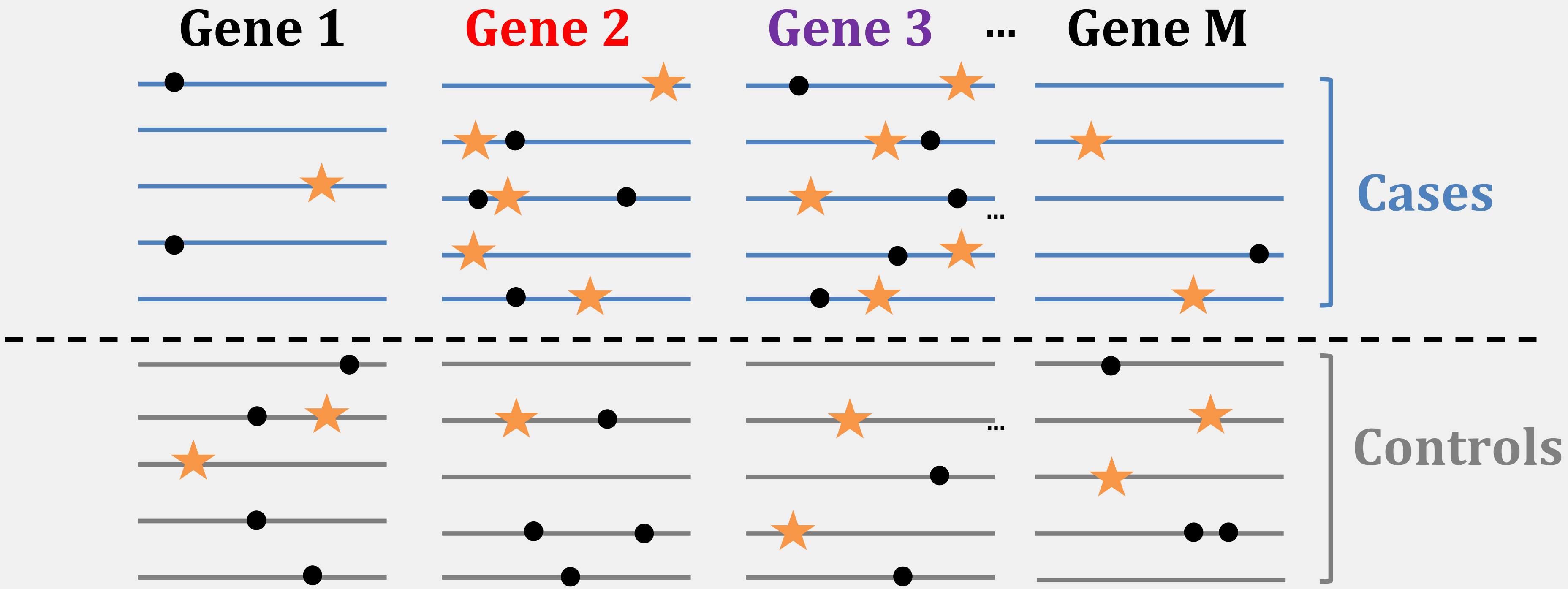
ProxECAT

$$H_0\colon \frac{\lambda_{FUN,g^*,case}}{\lambda_{SYN,g^*,case}} = \frac{\lambda_{FUN,g^*,control}}{\lambda_{SYN,g^*,control}}$$

$g^* - gene\ of\ interest$

$\lambda - rate\ of\ rare\ alternate\ alleles\ per\ N\ cases\ or\ controls$

- Controls for genetic bias in cases and controls



- Orange stars - rare, functional variants
- Dark circles - rare, synonymous variants

Generalized Linear Models (GLM)

- Assumptions

1. the observations are independent
2. the response variables follow a distribution from the exponential family
3. there exists a linear relationship between a transformation of the response variable and the predictor variables through a “link” function

$$g(\mu_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$$

μ_i: mean response for observation i
$g(\mu_i)$: link function
$x_{1i} \ldots x_{pi}$: p predictor variables for observation i
β_0: y-intercept
$\beta_1, \ldots, \beta_p$: regression estimates for each predictor variable p

Logistic Regression

- Data notation for ProxECAT can be observed as 2x2 Chi-square contingency table
- A 2x2 Chi-square contingency analysis is a specific case of logistic regression

ProxECAT Implementation of a Logistic Regression

- Alternate alleles are used as the observational unit

- Assume
  1. There does not exist a high correlation between the predictor variables
  2. The distribution of the alternate alleles is binomial
  3. Logit link function

Implementation of a Logistic Regression

$$\boxed{logit(y_i) = \beta_0 + \beta_1 x_i^{case} + \beta_2 x_i^{group}}$$

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

- y_i - functional status of alternate allele; 1 is a functional allele and 0 is synonymous alleles
- x_i^{case} - case status for the carrier of alternate allele
- x_i^{group} - group status for the carrier of the alternate allele; internal or external
- β_1 - regression estimate for the association between genetic region and case status
- β_2 - covariate regression estimate

Logistic Regression & Limitations of ProxECAT

- Can analyze both internal and external control data sets within the same test

- Can adjust for covariates e.g., internal/external, depth of coverage

Poisson Regression

- counts of rare alleles at a genetic variant are approximately distributed as a Poisson distribution

ProxECAT implementation of a Poisson Regression

- Genetic variants are used as the observational unit

- Assume
  1. Allele counts follow a Poisson distribution
  2. log link function

Implementation of Poisson Regression for functional (left) and non-functional (right) variants

Association without bias:

small p-value	high p-value
$\log(w_i) = \gamma_0 + \gamma_1 x_i^{case}$	$\log(z_i) = \alpha_0 + \alpha_1 x_i^{case}$
$H_0 \colon \gamma_1 = 0$	$H_0 \colon \alpha_1 = 0$
$\mathrm{H}_1 \colon \gamma_1 \neq 0$	$H_1 \colon \alpha_1 \neq 0$

- w_i - positive integer of functional allele counts per gene
- z_i - positive integer of non-functional allele counts per gene
- x_i^{case} - case status for the carrier of the rare variant
- γ_i, α_i - regression estimates between genetic region and case status

Implementation of Poisson Regression for Functional (left) and Non-Functional (right) Variants

Association with bias:

small p-value	low p-value
$\log(w_i) = \gamma_0 + \gamma_1 x_i^{case}$	$\log(z_i) = \alpha_0 + \alpha_1 x_i^{case}$
$H_0\colon \gamma_1 = 0$	$H_0\colon \alpha_1 = 0$
$\mathrm{H}_1\colon \gamma_1 \neq 0$	$H_1\colon \alpha_1 \neq 0$

- w_i - positive integer of functional allele counts per gene
- z_i - positive integer of non-functional allele counts per gene
- x_i^{case} - case status for the carrier of the rare variant
- γ_i, α_i - regression estimates between genetic region and case status

Implementation of Poisson Regression with Functional Status and Case Status

$$\log(p_i) = \omega_0 + \omega_1 x_i^{fun} + \omega_2 x_i^{case}$$

$$H_0 : \omega_1 = 0, \omega_2 = 0$$

$$H_1 : \omega_1 = 0, \omega_2 \neq 0$$

- p_i - total allele count for each rare variant per gene
- x_i^{fun} - functional status of the rare variant
- x_i^{case} - case status for the carrier of the rare variant
- $\omega_1, \omega_2, \omega_3$ - regression estimates

Implementation of Poisson Regression with Interaction

$$\log(p_i) = \omega_0 + \omega_1 x_i^{fun} + \omega_2 x_i^{case} + \omega_3 x_i^{int}$$

$$H_0\colon\ \omega_3 = 0$$

$$H_1\colon\ \omega_3 \neq 0$$

- p_i - total allele count for each rare variant per gene
- x_i^{fun} - functional status of the rare variant
- x_i^{case} - case status for the carrier of the rare variant
- x_i^{int} - interaction between functional status and case status
- $\omega_1, \omega_2, \omega_3$ - regression estimates

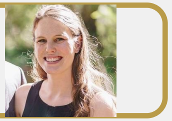
Poisson Regression & Limitations of ProxECAT

- Can analyze both internal and external control data sets within the same test

- Can adjust for covariates e.g., internal/external

Data

Simulations by Jessica Murphy



RARESim (Null, 2022) & Hapgen2

Non-Finnish European haplotypes from 1000 Genome Phase 3

Chromosome 19

12 genes

Simulated under the null hypothesis of no association

Dataset

Simulation Replicates	100
Genes	12
Total Sample Size	22,000
Internal Case Sample Size	1,000
Internal Control Sample Size	1,000
External Control Sample Size	10,000 ; 10,000
Rare Variants	MAF < 0.01

Type I Error Rate Comparisons

ProxECAT, Logistic , Poisson

3 main implementations of Poisson

- Functional variant model and non-functional variant model
- Model with functional status and case status
- Interaction model

Sample sizes

- Balanced (1,000 cases vs 1,000 controls)
- Unbalanced (1,000 cases vs 11,000 controls)

Type I Error rates for Poisson and Logistic (1,000 cases vs 11,000 controls)

	ProxECAT	Logistic
ADGRE 2	0.02	0.06
ADGRE 3	0.10	0.06
ADGRE 5	0.03	0.04
CLEC17A	0.03	0.01
DDX39A	0.09	0.02
DNAJB1	0.03	0.02
GIPC1	0.02	0.02
NDUFB7	0.04	0.01
PKN1	0.06	0.08
PTGER1	0.07	0.01
TECR	0.05	0.00
ZNF333	0.06	0.03
Average	**0.050**	**0.030**

Type I Error Rates for Poisson for Functional and Non-Functional Rare Variants (1,000 cases vs 11,000 controls)

	Functional Rare Variant Model	Non-Functional Rare Variant Model
ADGRE 2	1.00	1.00
ADGRE 3	1.00	1.00
ADGRE 5	1.00	1.00
CLEC17A	0.99	0.72
DDX39A	0.96	0.97
DNAJB1	1.00	1.00
GIPC1	1.00	0.99
NDUFB7	1.00	0.84
PKN1	1.00	1.00
PTGER1	0.98	0.92
TECR	0.90	0.99
ZNF333	1.00	0.99
Average	**0.986**	**0.952**

Type I Error Rates for Poisson for Functional and Non-Functional Rare Variants (1,000 cases vs 1,000 controls)

	Functional Rare Variant Model	Non-Functional Rare Variant Model
ADGRE 2	0.04	0.02
ADGRE 3	0.03	0.04
ADGRE 5	0.02	0.00
CLEC17A	0.03	0.00
DDX39A	0.01	0.00
DNAJB1	0.07	0.03
GIPC1	0.01	0.04
NDUFB7	0.04	0.03
PKN1	0.04	0.03
PTGER1	0.03	0.00
TECR	0.00	0.03
ZNF333	0.04	0.03
Average	**0.030**	**0.021**

Type I Error Rates for Poisson with Functional and Case Status (1,000 cases vs 11,000 controls)

	Functional Status	Case Status
ADGRE 2	0.72	1.00
ADGRE 3	0.66	1.00
ADGRE 5	0.75	1.00
CLEC17A	0.19	1.00
DDX39A	0.73	1.00
DNAJB1	0.84	1.00
GIPC1	0.67	1.00
NDUFB7	0.69	1.00
PKN1	0.87	1.00
PTGER1	0.55	1.00
TECR	0.84	1.00
ZNF333	0.68	1.00
Average	**0.682**	**1.00**

Type I Error Rates for Poisson with Functional and Case Status (1,000 cases vs 1,000 controls)

	Functional Status	Case Status
ADGRE 2	0.40	0.03
ADGRE 3	0.37	0.02
ADGRE 5	0.42	0.04
CLEC17A	0.02	0.01
DDX39A	0.26	0.03
DNAJB1	0.58	0.05
GIPC1	0.34	0.04
NDUFB7	0.53	0.04
PKN1	0.71	0.03
PTGER1	0.22	0.03
TECR	0.47	0.03
ZNF333	0.39	0.07
Average	**0.392**	**0.035**

Type I Error Rates for Poisson with Interaction (1,000 cases vs 11,000 controls)

	Interaction between Functional and Case Status
ADGRE 2	0.02
ADGRE 3	0.06
ADGRE 5	0.02
CLEC17A	0.00
DDX39A	0.02
DNAJB1	0.03
GIPC1	0.01
NDUFB7	0.00
PKN1	0.05
PTGER1	0.03
TECR	0.01
ZNF333	0.02
Average	**0.022**

- ProxECAT and logistic regression both perform appropriately
- Poisson without interaction cannot account for the imbalance in cases and controls
- Interaction between functional allele status and case status shows the last Poisson model could be explored more

Discussion

Moving Forward

- Standardize population to 1,000 or 10,000
- Different ratios of case and control samples
- Incorporate ratio of functional to synonymous alleles
- Additional distributions
- Logistic regression

Questions

Acknowledgments

Audrey Hendricks, PhD

Megan Null, PhD

Jessica Murphy

Erin Austin, PhD

References

- *Deoxyribonucleic Acid (DNA) Fact Sheet*. (2020, August 24). Retrieved from National Human Genome Research Institution. https://www.genome.gov/about-genomics/fact-sheets/Deoxyribonucleic-Acid-Fact-Sheet.
- Gonzalez, M. W., & Kann, M. G. (2012). Chapter 4: Protein interactions and disease. *PLoS computational biology*, *8*(12), e1002819. https://doi.org/10.1371/journal.pcbi.1002819.
- "Single-Nucleotide Polymorphism." Single-Nucleotide Polymorphism - ISOGG Wiki, 31 Jan. 2020, https://isogg.org/wiki/Single-nucleotide_polymorphism.
- Manolio, T., Collins, F., Cox, N. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–753 (2009). https://doi.org/10.1038/nature08494
- Karczewski, Konrad, and Laurent Francioli. "The Genome Aggregation Database (GnomAD)." *The Genome Aggregation Database (GnomAD) | GnomAD News*, 27 Feb. 2017, https://gnomad.broadinstitute.org/news/2017-02-the-genome-aggregation-database/.
- Zuk O, Schaffner SF, Samocha K, Do R, Hechter E, Kathiresan S, et al. Searching for missing heritability: designing rare variant association studies. Proc Natl Acad Sci U S A. 2014;111(4):E455–64. Epub 2014/01/17. pmid:24443550; PubMed Central PMCID: PMCPMC3910587. https://doi.org/10.1073/pnas.1322563111.
- Barrett, J.C., Buxbaum, J.D., Cutler, D.J., Daly, M.J., Devlin, B., Gratten, J., Hurles, M.E., Kosmicki, J.A., Lander, E.S., MacArthur, D.G., Neale, B.M., Roeder, K., Visscher, P.M., & Wray, N.R. (2017). New mutations, old statistical challenges. bioRxiv.2. https://www.biorxiv.org/content/10.1101/115964v3.
- Dunn, P. K., & Smyth, G. K. (2018). *Generalized Linear Models With Examples in R*. New York: Springer Science+Business Media, LLC.
- Newsom, J. T. (1999-2007). *Lecture 21: Logistic Regression*. Retrieved from newsomj. http://web.pdx.edu/~newsomj/pa551/lectur21.htm.
- Null, M., Dupuis, J., Sheinidashtegol, P., Layer, R. M., Gignoux, C. R., & Hendricks, A. E. (2022). RAREsim: A simulation method for very rare genetic variants. *American journal of human genetics*, *109*(4), 680–691. https://doi.org/10.1016/j.ajhg.2022.02.009
- Murphy, J. (2021, June 3). Common Controls Simulations.