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

Makayla Cowles

Abstract

Genome-wide association studies (GWAS) involve scanning complete sets of DNA, or genomes, in samples to identify genetic variations associated with particular diseases. Researchers have found that rare genetic variations can be indicators to having a pre-disposition for certain diseases and genetic conditions. Many statistical methods have been developed that can identify associations between genes and complex traits. In 2018, ProxECAT was developed for rare variant association studies, using case – control data from internal and external sources gathered from publicly available databases. While providing a robust approach to rare variant association studies using external controls, ProxECAT has opportunities for improvement. ProxECAT cannot control for covariates or incorporate internal and external control data sets in the same statistical test. We explore these areas for improvement using Poisson regression and compare the results to ProxECAT and ProxECAT using logistic regression. We find that the implemented Poisson regression models cannot account for the imbalance of cases and controls in the data. We explore other possible avenues for continuing the evaluation of ProxECAT to improve on its limitations.

1. Introduction

A genetics review is provided in the supplemental material.

During the mid-1990s, we saw the first genetic tests that were available to consumers for health, and in 2018 it was reported that the personal genomics usage had heightened (2). Understanding genetic factors is essential to promoting health and preventing disease. In genetic studies, thousands of common variants have been identified as being associated with various diseases. Yet, common genetic variants are often estimated to account for less than half of the heritability of particular diseases and genetic disorders, for instance for coronary heart disease and type 2 diabetes (10). It is hypothesized that the remaining heritability may be due, in part, to rare genetic variants (10). The identification of these genetic rare variants can lead to better disease diagnosis, treatment, and prevention.

Rare variant studies have increased over the years as the importance of testing the association between genes and health traits have become prominent. However, rare variant association tests are often underpowered due to the lack of large enough case and control sample sizes (6). To detect gene associations using disease-associated rare variants, case sizes of tens of thousands to hundreds of thousands are needed for sufficient power (10). In case-control studies, statistical power can be increased by increasing the control sample size. According to Lee et al., samples from other studies can be used as controls (external control samples) to substantially increase the power of rare variant tests (11). Methods like ProxECAT can utilize data from publicly available databases to increase sample size and power in genetic rare variant case-control association tests.

Researchers can use data found in public genetic databases in their rare variant association tests to increase sample size and potentially avoid underpowered studies. The data can either be individual-level data or genomic summary data, such as allele frequency. Since individual-level data can be difficult to access, these publicly available databases contain genomic summary data of allele frequency. The Exome Aggregation Consortium (ExAC) has more than 60,000 exomes, and a later version called the Genome Aggregation Database (GnomAD) has rare variant genetic data from approximately 140,000 individuals that could be utilized in rare variant association tests (6). Often, these databases are misused due to the difference in sequencing technology and processing between cases and controls. In a study by Stressman et al. (2017), it was reported that 91 genes were directly associated with autism spectrum disorders, intellectual disability, and development delay. One of the analyses the authors conducted was to assess a significance excess in private mutations, rare gene mutations that are usually found only in a single family or a small population. The researchers had 11,730 case samples that were analyzed with 45,375 control samples from ExAC. Unfortunately, the control samples from ExAC were not sequenced and analyzed in an identical way to the researchers’ 11,730 case samples (1). This systematic difference was not addressed in the authors analysis. There are multiple methods in which genetic data can be sequenced and analyzed. If controls and cases are not collected in the exact same manner, we cannot be confident in the results. Because of this misuse of comparing internal case samples with external control samples from different populations, it is suggested that there are far fewer significant genes than the authors observed (1).

Integrating External Controls into Association Test (iECAT) is a statistical association method that was developed to robustly incorporate controls from publicly available genetic databases of summary genetic data (6). iECAT does well with rare variant studies but is unstable for very rare variants that occur in a single individual (singletons) or two individuals (doubletons). In addition, iECAT requires both an external control sample and internal control sample (which not every study has) (11). Due to these limitations, a new statistical method, ProxECAT, was developed.

In 2018, Hendricks et al. developed ProxECAT, a statistical method to incorporate publicly available genomic summary data as external controls in case-control studies. In addressing limitations from previous methods, ProxECAT can use very rare variants such as singletons or doubletons and does not require an internal control sample. ProxECAT uses variants that are predicted to have both functional and non-functional effects on a gene's function (6). Often, to increase power, rare variant gene region association tests limit their analysis to only rare variants that have a functional effect (i.e., have an expected effect on a gene's function). ProxECAT uses the ratio of functional to synonymous rare variants between cases and controls to control for genetic bias by using the synonymous variants as a "proxy" for how well rare variants are sequenced within a gene region. ProxECAT assesses gene associations between cases and controls by testing if the ratio of functional to synonymous variants are equal among cases and controls (6).

While providing a robust approach to rare variant association tests using external controls, ProxECAT does have opportunities for improvement. Currently, ProxECAT cannot adjust for covariates such as sex, ancestry, or proportion of alternate variant reads or depth of coverage. Additionally, ProxECAT does not enable internal controls to be analyzed with external controls; in other words, it cannot include multiple case or control sets within the same test (6). Instead of running two analyses in parallel, it would be most efficient if multiple case or control data sets could be included in the same test resulting in only one analysis.

In this paper, we evaluate extending ProxECAT to Poisson regression to address ProxECAT's current limitations. Three motives for why we chose to explore Poisson regression are 1) A major strength of regression analysis is its ability to control for covariates, 2) Both internal and external controls can be evaluated together using a regression model, and 3) Rare allele counts are approximately distributed as a Poisson distribution. An analysis of ProxECAT using logistic regression has previously been performed. Therefore, we compare our results from Poisson regression to ProxECAT and ProxECAT applied to logistic regression.

2. Methods

Here, we will review ProxECAT, discuss how ProxECAT was previously evaluated using logistic regression, explain how we evaluate ProxECAT using Poisson regression, and describe the data used in the analysis.

2.1 Observational Units

An observational unit is the unit about which information is collected. Most often, the observational units for rare variant case-control association tests are the individuals. However, when summarizing the data, information on the observational unit is lost. Instead, ProxECAT uses alternate alleles in a genetic region as the observational unit. If the observational unit is the alternate alleles, when summarizing the data over individuals, the observational units are retained. This information can then be used in ProxECAT to test for gene region associations between cases and controls.

When evaluating ProxECAT with logistic regression, the observational units are the alternate alleles, equivalently to ProxECAT. The information collected about the alleles are whether they are functional or synonymous (non-functional), if they are present within a case or control, and whether they come from an internal or external control sample. For logistic regression, we want to test if the functional status of the alleles depends on case status. Consequently, our outcome is either a functional allele or synonymous allele (two choices), which fits well within a binary logistic regression model.

Poisson regression requires the outcome of the model to be a count. Since the data collected about alternate alleles have binary responses (i.e., case vs control, function vs synonymous), it is not feasible to use the alternate allele as the observational unit. Alternatively, we set the observational unit to be the genetic variants. In doing so, the outcome is the count of alternate alleles in each genetic variant. For each rare variant, information is collected on its functionality, and the allele count for both case and control status. In our evaluation, we collect the allele count for functional variants and synonymous variants, then run a Poisson regression model for each gene.

2.2 ProxECAT "Proxy External Controls Association Test"

In genetic studies, a goal is to find genes that are associated with a disease. ProxECAT is a statistical method that tests for gene region associations between cases and controls. This method can use singletons and doubletons, incorporate both functional and synonymous variants, and properly use external controls. The main defining attribute of ProxECAT is that it is designed to be used with internal case samples and external controls samples of summary data sequenced and processed at different times and places. The observed alternate allele counts are modelled as a random sample of four independent Poisson distributions:

$$X_1^f \sim Poisson(\lambda_1^f),\ X_0^f \sim Poisson(\lambda_0^f),\ X_1^p \sim Poisson(\lambda_1^p), \text{ and } X_0^p \sim Poisson(\lambda_1^p) \qquad (6)$$

Table 1 is a contingency table that represents the data notation for ProxECAT. We denote x_1^f and x_0^f, with a 1 for case and 0 for control, as the alternate alleles that have a functional effect; and denote x_1^p and x_0^p , 1 for case and 0 for control, as the alternate alleles that have a nonfunctional effect. The null hypothesis for ProxECAT (Eq. 1) states whether the ratio of functional to proxy alleles is equal in cases and controls.

Table 1. Data notation for ProxECAT

				Total
		Functional	**Not Functional (Proxy)**	
Cases (Internal)	Y = 1	x_1^f	x_1^p	x_1
Controls (Externals)	Y = 0	x_0^f	x_0^p	x_0
Total		x_f	x_p	

https://doi.org/10.1371/journal.pgen.1007591

$$H_0: \frac{\lambda_1^f}{\lambda_1^p} = \frac{\lambda_0^f}{\lambda_0^p}$$

$$H_1: \frac{\lambda_1^f}{\lambda_1^p} \neq \frac{\lambda_0^f}{\lambda_0^p} \qquad \text{(Eq. 1)}$$

To test associations between a gene region and case status with ProxECAT, we collect data on the total observed allele count for functional case alleles, synonymous case alleles, functional control alleles, and synonymous control alleles by each gene. ProxECAT outputs a p-value for each gene, and it is then compared to a preset significance level. If the p-value < significance level, then we conclude that there is enough evidence to reject the null hypothesis that there exists no difference in the ratio of functional to proxy alternate alleles between cases and controls. Otherwise, if the p-value > significance level, we fail to reject the null hypothesis and conclude that there is not enough evidence to support an association between the gene and case status.

In its current model structure, ProxECAT cannot adjust for possible covariates. Looking at the null hypothesis, the only data ProxECAT can utilize are allele counts for cases and controls and whether the alleles are located in a functional or synonymous variant. In addition, ProxECAT cannot incorporate internal controls with external controls. The objective of extending ProxECAT to a logistic regression and a Poisson regression is to address both limitations, while maintaining the strengths of ProxECAT.

2.3 Generalized Linear Models (GLMs)

We will next explain the implementation of ProxECAT in two GLMs, logistic regression and Poisson regression. First, giving a brief review of generalized linear models.

Generalized linear models are generalized extensions of ordinary linear regression. When generalizing linear regression, we consider the following assumptions: 1) the observational units are independent 2) the response variables follow a distribution from the exponential family, such as the normal, Poisson, negative binomial, binomial, and gamma distributions and 3) there exists a linear relationship between a transformation of the response variable and the predictor variables through a "link" function (4). The equation for GLMs are as follows

$$g(\mu_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} \qquad \text{(Eq. 2)}$$

where μ_i is the mean response for observation i; $g(\mu_i)$ is the link function; $x_{1i} \dots x_{pi}$ are p predictor variables for observation i; β_0 is the y-intercept; and $\beta_1, \dots, \beta_p$ are the regression estimates for each predictor variable p.

2.3.1 Logistic Regression

In improving on the limitations that exist with ProxECAT, an analysis extending ProxECAT to a logistic regression was previously performed. A connection between ProxECAT and logistic regression can be made, making it reasonable to use logistic regression. The structure of ProxECAT has similarities when compared with a Chi-square test for independence. Table 1 shows the data notation for ProxECAT which can be observed as $2x2$ Chi-square contingency table of functional and nonfunctional allele counts for case status. Logistic regression tests for relationships between a dichotomous dependent variable (consisting of two values) and either dichotomous or continuous predictor variables. Chi-squared test for independence tests for a relationship between two dichotomous variables. Hence a $2x2$ Chi-square contingency analysis is a specific case of logistic regression, where the dependent and independent variables are both dichotomous (9). Observing these connections makes it appropriate to implement a logistic regression.

Before determining the logistic regression model, we chose the observational unit as the alternate alleles. Information can be gathered on alternate alleles including functionality (functional or not functional), case status (case or control), and group status (internal control or external control). Logistic regression requires a response variable with a binary outcome. Hence, we use the allele functionality as the response variable, leaving case status and group status as the predictor variables. To run a logistic regression, we assume that there does not exist a high correlation between the predictor variables and the distribution of the alternate alleles is binomial. The applied logistic regression model is

$$logit(y_i) = \beta_0 + \beta_1 x_i^{case} + \beta_2 x_i^{group} \qquad \text{(Eq. 3)}$$

There is a total of N alternate alleles in each gene, indexed by i. The response variable y_i, is binary, where 1 is a functional allele and 0 is a nonfunctional allele per gene i. Our predictor variables are denoted by x_i^{case} and x_i^{group}, for each allele observed in a case or control and whether the allele comes from an internal or external control per gene i, respectfully. However, x_i^{group} is a covariate because our goal is testing for gene associations between case and control, not necessarily between internal and external controls. The coefficients β_1 and β_2 are the regression estimates for the association between the response variable and predictor variables. The estimate of interest is β_1, therefore, the null hypothesis and alternative hypothesis are as follows

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0 \qquad \text{(Eq. 4)}$$

If the null hypothesis is rejected, then we conclude there exists evidence supporting a relationship between allele functionality and allele case status. Alternatively, if the null hypothesis is not rejected, we conclude that there does not exist evidence to support a relationship between the functionality and case status of the observed alleles.

This logistic regression model addresses both disadvantages of ProxECAT. We can analyze both internal and external control data sets in one test by adding a predictor variable for these data sets. In Eq. 3, this predictor variable is denoted as x_i^{group}. This variable also serves as a covariate, indicating that logistic regression can adjust for covariates.

2.3.2 Poisson Regression

The counts of rare alleles are approximately distributed as a Poisson distribution (7). Hendricks et al. (2018), assessed the fit of a Poisson distribution on the number of rare alleles in a genetic region. To assess the fit, rare minor alleles were simulated assuming a binomial distribution for each variant and compared with the theoretical Poisson distribution (6). We believe it appropriate to evaluate ProxECAT using a Poisson regression model due to the fit of a Poisson distribution on the rare alleles.

To evaluate ProxECAT using a Poisson regression model, the observational unit must include count data. Considering rare variants, information can be collected on their functionality, case status, and the total number of alternate alleles present in the corresponding variant. Unlike logistic regression, the response variable is not a binary outcome of whether the alternate alleles have a functional or nonfunctional effect, but an outcome of how many total alternate alleles are observed in each rare variant. As a result of the Poisson regression structure, we will describe three different Poisson regression models that are used in this analysis. Each of the three models assumes the allele counts follow a Poisson distribution and a linear relationship exists between the response variable and predictor variables through the logarithm link function.

The first two Poisson regression models are similar to each other, because we want to incorporate both functional and nonfunctional rare variants. The predictor variable is rare variant case status, and the response variable is the allele count for functional and nonfunctional rare variants. The two applied models are

$$\log(w_i) = \gamma_0 + \gamma_1 x_i^{case} \qquad \text{(Eq. 5)}$$

$$\log(z_i) = \alpha_0 + \alpha_1 x_i^{case} \qquad \text{(Eq. 6)}$$

In Eq. 5, w_i is a positive integer of functional allele counts per gene, indexed by i. While in Eq. 6, z_i is a positive integer of nonfunctional allele counts per gene, i. In both Eq. 5 and Eq. 6, x_i^{case} denotes whether the observational rare variant is a case or control per gene, i, for functional alleles and nonfunctional alleles, respectfully. Our regression estimates for the association between our predictor and response variables are γ_1 and α_1. The following are the null and alternative hypothesis for Eq. 5

$$H_0: \gamma_1 = 0$$

$$H_1: \gamma_1 \neq 0 \qquad \text{(Eq. 7)}$$

and Eq. 6

$$H_0\colon \alpha_1 = 0$$

$$H_1\colon \alpha_1 \neq 0 \qquad \text{(Eq. 8)}$$

Recall, ProxECAT uses the non-functional variants as proxies for how well the gene region was captured by the sequencing pipeline. Consequentially, we would want to see an association in the functional model (Eq. 5) but not in the non-functional model (Eq. 6). When running these two models in parallel, it would be ideal if there exists enough evidence to conclude that an association occurs between functional rare variants and case status; yet not enough evidence to reject the null hypothesis (Eq. 8) stating that there is no association between nonfunctional rare variants and case status. Eq. 5 and Eq. 6 retain a similar structure compared with ProxECAT and logistic regression. Each has a null hypothesis of no significant relationships between case status and the functionality of alleles.

The third Poisson regression model is different from previously applied regression models because the response variable is not defined in terms of functional or synonymous alternate alleles, but in total allele count for all rare variants. The third model we evaluate is

$$\log(p_i) = \omega_0 + \omega_1 x_i^{fun} + \omega_2 x_i^{case} + \omega_3 x_i^{int} \qquad \text{(Eq. 11)}$$

where p_i denotes the allele count for each variant per gene, indexed by i, x_i^{fun} represents the functionality for the observed rare variant, x_i^{case} represents the case status for the observed rare variant, and x_i^{int} is the interaction variable for x_i^{case} and x_i^{fun}. When evaluating ProxECAT using this Poisson model, we run two models. The first does not incorporate the interaction variable x_i^{int}, and only uses x_i^{fun} and x_i^{case} as predictor variables. The second model is Eq. 11, incorporating the interaction variable, x_i^{int}. The estimate of interest is ω_3; the null and alternative hypotheses are as follows

$$H_0\colon \omega_3 = 0$$

$$H_1\colon \omega_3 \neq 0 \qquad \text{(Eq. 12)}$$

Implementing this model, we test for relationships between the interaction of these two variables with the total allele count for each rare variant.

In the three Poisson models, we do not incorporate internal and external controls, nor do we directly add covariates. If warranted, these could be added into a Poisson regression. However, our aim is to explore if Poisson regression could appropriately be used for rare variant association studies. We avoid adding any extra variables as to reduce the probability of overfitting the model before getting a simpler model functioning correctly.

2.4 Data

Prior to this analysis simulated data was created to confirm the validity of updating ProxECAT with GLMs. The data was simulated with simulation methods Hapgen2 and RARESim (8). Hapgen2 is a simulation method that simulates case and control datasets for genetic variations, and RARESim is a simulation method that can be used for very rare genetic variants. The input datasets are haplotypes, legend files (an accompanying variant list) and a recombination map from 1000 Genomes Phase 3 (8). A Non-Finnish European population was used for the simulation haplotypes. GnomAD exome data was then merged with the 1000 Genome legend file. The variant functional annotation was performed using refGene database in AANOVAR, an efficient software tool to functionally annotate genetic variants. Using the input dataset, Hapgen2 outputs a haplotype file and legend file with an abundance of rare variants. The allele counts outputted from Hapgen2, and data from GnomAD is used as input data for RARESim. RARESim then outputs a list of variants to remove that do not follow the distribution seen in the real world (8). Our four final simulated data sets are a haplotype file, legend file, allele count file, and a sample file. The data was simulated under the null hypothesis of no association between gene region and case status. Consequently, the result for any statistical method that uses this simulated data for analyses should follow the null hypothesis.

A total of 100 simulation replicates were produced for 12 genes. There are a total of 22,000 individuals in the sample file, however the number of variants and the allele count differs for each simulation replicate. Using the 22,000 individuals, case status is randomly assigned into four groups: 1,000 internal cases, 1,000 internal controls, and two sets of 10,000 external controls. One set of external controls is removed, leaving 12,000 individuals that are used for the association tests. Additionally, all common variants are filtered out using the remaining 12,000 cases and controls and the minor allele formula (MAF). If the MAF > 0.01, the variants are removed, leaving only the rare variants for association tests.

ProxECAT requires data collected on the alternate alleles. To run ProxECAT, we create a data frame of 12 rows for each 12 genes, and four columns for the total count of functional case alleles, functional control alleles, synonymous case alleles, and synonymous control alleles. We use a total of 1,000 cases and 11,000 controls, combining both external and internal controls to create one control set for evaluating ProxECAT. To evaluate ProxECAT using logistic regression, we organize our data by alternate allele. For each 12 genes, we create a new data frame where the rows are each alternate allele observed in an individual, and the columns are which gene region the allele is located, allele functionality, allele case status, and allele group status. For logistic regression, we use 1,000 cases, 1,000 internal controls, and 10,000 external controls. To apply Poisson regression, we organize our data by rare variants. Using the haplotype file, the allele count is found by calculating the rare variant row sums for cases and controls, separately. The rows for the new data frame are the rare variants, and the columns are the allele count for functional variants and synonymous variants and the case status. We run Poisson regression with two different sample sizes. The first run uses 1,000 cases and 11,000 controls, combining external and internal controls. The second run uses 1,000 cases, but only uses 1,000 controls. We randomly assign the 1,000 controls from the larger set of 11,000. Our output for each method is a 12 x 100 table of p-values: 12 rows for each gene region and 100 columns for each simulation replicate. Lastly, we output one table of type I error results by gene per method.

3. Results

A total of nine models were run with two different case and control samples. Table 2 shows each model, its corresponding formula, and the number of case and control samples that were used. Each model has been labeled with a reference letter which will be used in the type I error rates table (Table 3).

Table 2. Reference Table for Models

	Model	Formula	Number of Cases	Number of Controls
A	ProxECAT	$proxecat(x_1, x_2, x_3, x_4)$	1,000	11,000
B	Logistic Regression	$logit(y_i) = \beta_0 + \beta_1 x_i^{case} + \beta_2 x_i^{group}$ (Eq. 3)	1,000	11,000
C	Functional Poisson Regression	$\log(w_i) = \gamma_0 + \gamma_1 x_i^{case}$ (Eq. 5)	1,000	11,000
D	Non-Functional Poisson Regression	$\log(z_i) = \alpha_0 + \alpha_1 x_i^{case}$ (Eq. 6)	1,000	11,000
E	Functional Poisson Regression	$\log(w_i) = \gamma_0 + \gamma_1 x_i^{case}$ (Eq. 5)	1,000	1,000
F	Non-Functional Poisson Regression	$\log(z_i) = \alpha_0 + \alpha_1 x_i^{case}$ (Eq. 6)	1,000	1,000
G	Poisson Regression w/out Interaction	$\log(p_i) = \omega_0 + \omega_1 x_i^{fun} + x_i^{case}$	1,000	11,000
H	Poisson Regression w/out Interaction	$\log(p_i) = \omega_0 + \omega_1 x_i^{fun} + x_i^{case}$	1,000	1,000
I	Poisson Regression w/ Interaction	$\log(p_i) = \omega_0 + \omega_1 x_i^{fun} + x_i^{case} + \omega_3 x_i^{int}$ (Eq. 11)	1,000	11,000

To determine if our models follow the null hypothesis of no association, we output a table of type I error rates (Table 3). Table 3 states the type I error rate for each gene and the average type I error rate for each method. The first two columns in Table 3 (A and B) are the output for type I error rates of ProxECAT and logistic regression. The type I error rates for logistic regression come from the p-values corresponding to case status, x_i^{case}, from Eq. 3. To determine model efficiency for Poisson regression, we will compare the type I error rates of each Poisson model in Table 2.

Evaluating ProxECAT using a Poisson regression model involves multiple runs using different sample sizes for four different Poisson models. Models C and D use the initial sample sizes of 1,000 cases and 11,000 controls, and their type I error rates can be seen in Table 3. The type I error rates for both models come from the p-values for the corresponding to case status, x_i^{case}, from Eq. 5 And Eq. 6. Models E and F use the exact same formula as models C and D; however, we reduce the control sample size to 1,000. There exists an imbalance of cases and controls when looking at the type I error rate results for models C and D. Whereas the results from models E and F come from a reduced data set to get an equal sample size of cases and controls.

The last Poisson model was run thrice. For the first run, we only use x_i^{fun} and x_i^{case} as predictor variables, not incorporating the interaction variable, and use the initial 1,000 case samples and 11,000 control samples. For the second run, we only use x_i^{fun} and x_i^{case} as predictor variables but reduce our controls sample size to equal our case sample size (1,000 cases/1,000 controls). For the last run, we incorporate all three predictor variables shown in Eq. 11 and use the whole sample size of 1,000 cases and 11,000 controls. For both models G and H, the first listed type I error rate is found from the p-values of the functionality, x_i^{fun}, and the listed second type I error rate is found from the p-values of case status, x_i^{case}. Model I outputs the type I error rate for the interaction, x_i^{int}, with 1,000 cases and 11,000 controls.

Table 3. Type I error rates for each model per gene

	A	B	C	D	E	F	G		H		I
ADGRE 2	0.02	0.06	1.00	1.00	0.04	0.02	0.72	1.00	0.40	0.03	0.02
ADGRE 3	0.10	0.06	1.00	1.00	0.03	0.04	0.66	1.00	0.37	0.02	0.06
ADGRE 5	0.03	0.04	1.00	1.00	0.02	0.00	0.75	1.00	0.42	0.04	0.02
CLEC17A	0.03	0.01	0.99	0.72	0.03	0.00	0.19	1.00	0.02	0.01	0.00
DDX39A	0.09	0.02	0.96	0.97	0.01	0.00	0.73	1.00	0.26	0.03	0.02
DNAJB1	0.03	0.02	1.00	1.00	0.07	0.03	0.84	1.00	0.58	0.05	0.03
GIPC1	0.02	0.02	1.00	0.99	0.01	0.04	0.67	1.00	0.34	0.04	0.01
NDUFB7	0.04	0.01	1.00	0.84	0.04	0.03	0.69	1.00	0.53	0.04	0.00
PKN1	0.06	0.08	1.00	1.00	0.04	0.03	0.87	1.00	0.71	0.03	0.05
PTGER1	0.07	0.01	0.98	0.92	0.03	0.00	0.55	1.00	0.22	0.03	0.03
TECR	0.05	0.00	0.90	0.99	0.00	0.03	0.84	1.00	0.47	0.03	0.01
ZNF333	0.06	0.03	1.00	0.99	0.04	0.03	0.68	1.00	0.39	0.07	0.02
Average	**0.050**	**0.030**	**0.986**	**0.952**	**0.030**	**0.021**	**0.682**	**1.00**	**0.392**	**0.035**	**0.022**

We observe high type I error rates for models C and D using 1,000 cases and 11,000 controls. However, for models E and F the type I error rates are significantly reduced when using 1,000 cases and only 1,000 controls. For Model G the type I error rate for case status is at a maximum value of 1.00, furthermore, the type I error rate for functionality is lower than case status, but it is still high. After reducing the control size to 1,000, (model H) the type I error rate for case status significantly decreases. The type I error rate for functionality also decreases, however not as considerably as case status. Lastly, for model I, that the type I error rate for the interaction starts at a much lower rate than previous methods.

4. Discussion

Our overall goal is to extend ProxECAT to Poisson regression. Three Poisson regression models were generated with the idea to test for gene region associations between cases and controls. The models were run with data simulated under the null hypothesis of no association. For each Poisson model, type I error rates were calculated per gene and compared with type I error rates for ProxECAT and ProxECAT extended to logistic regression.

In Table 3 for ProxECAT (model A), we observe that most of the type I error rates are close to 0.05, and the average is exactly our significance level of 0.05. Hence, we can conclude that ProxECAT performs appropriately for rare variant association tests. In the second column of Table 3 (Model B), we analyze the type I error rates for logistic regression. Even though, the average is not exactly equal to 0.05, the calculated average of 0.03 allows us to conclude that logistic regression performs sufficiently well albeit slightly conservative for rare variant association tests. Now that we have the type I error rates for both methods, we will compare the method performance for each Poisson model with both ProxECAT and logistic regression.

The results of the initial two Poisson models that we evaluate do not follow the null hypothesis. Most type I error rates for models C and D are 1.00 resulting in a very high type I error rate average. To explore possible reasons for the high error rates, we apply the same two models to a reduced sample size of 1,000 cases and 1,000 controls (models E and F). These results produce much lower type I error rates compared to the two previous columns. Implying that Eq. 5 and Eq. 6 cannot currently control for the imbalance of case status in the data. Though, when the control sample size is reduced, Eq. 5 is just as efficient as the logistic regression model, both having a type I error rate of 0.03. We observe similar results for the third Poisson regression model, models G and H). To address the problem that our applied Poisson model cannot sufficiently run when there is an imbalance in case status, we run a Poisson model with an added interaction variable between the two initial predictors, case status and functionality (Eq. 11). Comparing the results for model G and I, we observe that the interaction variable has type I error rates far closer to 0.05 compared to the case status and functionality. We see the interaction variable does not seem to be affected by the data imbalance in case status, and it accounts for the interaction between functionality and case status. Though, this third Poisson model might not be appropriate for testing gene region associations between case status. The associations we are testing with Eq. 11 are between case status and functionality with the total number of alleles per variant in a genetic region. This association does not directly give us information we are after for finding gene region associations.

Moving forward, it might be feasible to use a Poisson regression to address the limitations in ProxECAT. However, more analyses should be completed before a Poisson regression model can be confidently applied. To address the problem with the imbalance in case and control size, we could standardize the population to 1,000 or 10,000. Although, a disadvantage of this approach is the possibility of missing rare variants in our data, since rare variants are not observed frequently in a population. It would also be interesting to look at different ratios of case samples vs control samples. We know the models run well with an equal number of cases and controls, but could we analyze how the type I error rate increases with an increase in case and control sample sizes. We could explore the third Poisson regression model (Eq. 11) further since the interaction variable deems appropriate to use. Optimally, we would want to incorporate the ratio of functional to synonymous alleles as ProxECAT does. It could be possible to evaluate extending ProxECAT to a gamma regression to incorporate ratios. There is a possibility that extending ProxECAT to a Poisson regression will not work. In such a case, we could adjust our focus back to logistic regression. At a current average type I error rate of 0.03, we could explore ways to improve the model to match or get closer to our significance level of 0.05.

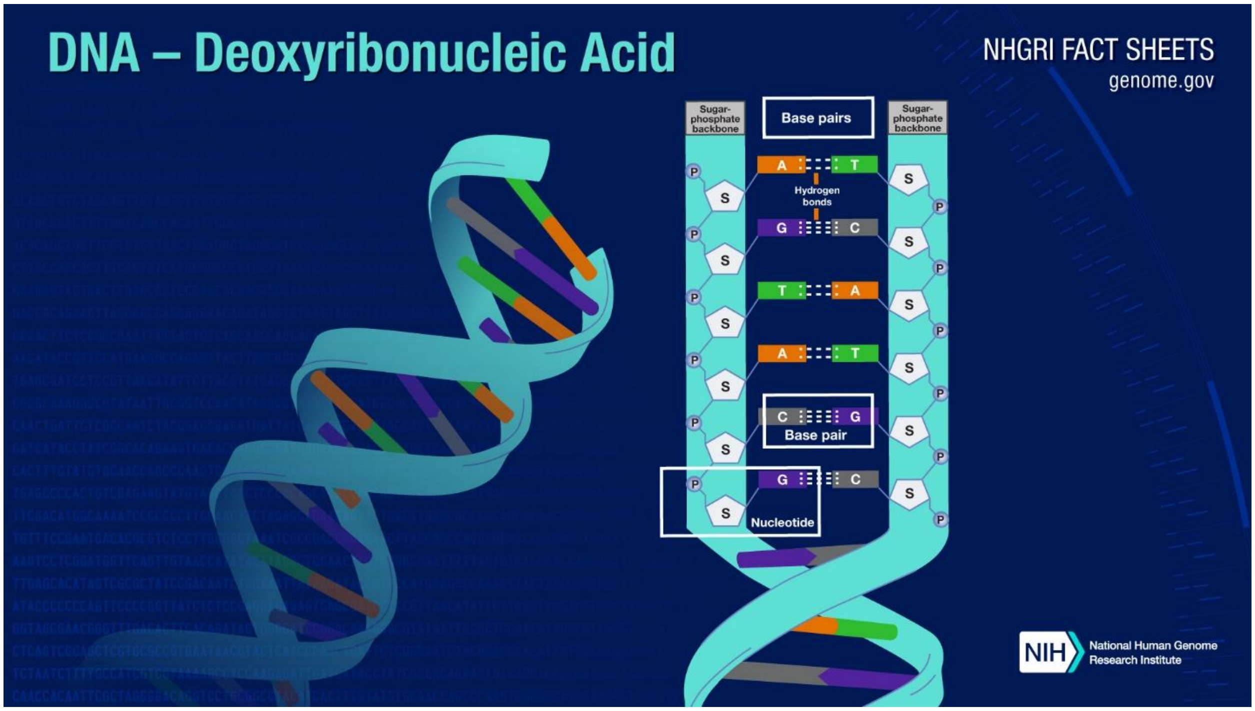
Our applied Poisson regression models are not currently appropriate for rare variation association tests. We observe that our models perform better when decreasing the sample size of controls to equal our case sample size. For rare variant association studies, it is ideal if we have multiple controls per case. Thus, it is important that we have a model that can control for difference in sample sizes regarding case status. Further exploratory analysis can be performed on more a complex Poisson model, differently distributed regression models, or the logistic regression model.

Supplemental Materials

Genetics Review

DNA is made up of chemical building blocks called nucleotides. There are four nucleotides, or bases: adenine (A), thymine (T), guanine (G), and cytosine (C). These bases bond together to form base pairs; an A always pairs with a T, and a C always pairs with a G. The base pairs hold the two DNA strands together, creating a double helix structure (3).

Figure 1. Structure of DNA



"Courtesy: National Human Genome Research Institute" genome.gov

Between any two people, there exists approximately a 0.1% difference in the order of the base pairs. Yet, this small difference explains our unique features such as eye color and height. The base pairs that differ among individuals are known as alleles, and the location of the alleles in the genetic sequence are variants. In general, variants are often broken into two types: common variants and rare variants. To differentiate between common and rare variants, we use information provided by the Minor Allele Frequency (MAF). The MAF is the frequency at which the least most common allele occurs in a given population. It is calculated by dividing the total number of alternate alleles in a population by the total number alleles in the observed variant. Rare variants are often defined as having a MAF < 0.01. Rare variants can be classified into two subgroups: functional rare variants and synonymous rare variants. Functional rare variants alter a gene's function by preventing proteins from working properly. Undesired protein interactions are a cause of many diseases (5). Alternatively, synonymous rare variants do not alter a gene's function.

Code

https://github.com/mcowles33/ProxECAT.git

Acknowledgments

I would like to express my gratitude to my advisor Audrey Hendricks who guided me in this project, and Erin Austin who pushed me along the way. I also wish to acknowledge the help and previous work provided by Jessica Murphy.

References

1. 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.

2. Bowen, S., & Muin, K. J. (2018, June 12). *Consumer Genetic Testing Is Booming: But What are the Benefits and Harms to Individuals and Populations?* Retrieved from Centers for Disease Control and Prevention. https://blogs.cdc.gov/genomics/2018/06/12/consumer-genetic-testing/.

3. *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.

4. Dunn, P. K., & Smyth, G. K. (2018). *Generalized Linear Models With Examples in R.* New York: Springer Science+Business Media, LLC.

5. 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.

6. Hendricks, A. E., Billups, S. C., Pike, H., Farooqi, I. S., Zeggini, E., Santorico, S. A., Barroso, I., & Dupuis, J. (2018). ProxECAT: Proxy External Controls Association Test. A new case-control gene region association test using allele frequencies from public controls. *PLoS genetics*, *14*(10), e1007591. https://doi.org/10.1371/journal.pgen.1007591.

7. Joyce, P., Tavaré, S. The distribution of rare alleles. *J. Math. Biology* **33,** 602–618 (1995). https://doi.org/10.1007/BF00298645.

8. Murphy, J. (2021, June 3). Common Controls Simulations.

9. Newsom, J. T. (1999-2007). *Lecture 21: Logistic Regression.* Retrieved from newsomj. http://web.pdx.edu/~newsomj/pa551/lectur21.htm.

10. 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.

11. Lee, S., Kim, S., & Fuchsberger, C. (2017). Improving power for rare-variant tests by integrating external controls. Genetic Epidemiology, 41(7), 610-619. https://doi.org/10.1002/gepi.22057.