RESEARCH ARTICLE

ProxECAT: Proxy External Controls Association Test. A new case-control gene region association test using allele frequencies from public controls

Audrey E. Hendricks[1,2,3]*, Stephen C. Billups[1], Hamish N. C. Pike[2], I. Sadaf Farooqi[4], Eleftheria Zeggini[5], Stephanie A. Santorico[1,2,3], Inês Barroso[4,5], Josée Dupuis[6]

1 Mathematical and Statistical Sciences Department, University of Colorado Denver, Denver, CO, United States of America, **2** Human Medical Genetics and Genomics Program, University of Colorado-Denver, Aurora, CO, United States of America, **3** Department of Biostatistics and Informatics, Colorado School of Public Health, Aurora, CO, United States of America, **4** University of Cambridge Metabolic Research Laboratories and NIHR Cambridge Biomedical Research Centre, Wellcome Trust-MRC Institute of Metabolic Science, Addenbrooke's Hospital, Cambridge, United Kingdom, **5** Human Genetics, Wellcome Sanger Institute, Cambridge, United Kingdom, **6** Department of Biostatistics, Boston University School of Public Health, Boston, MA, United States of America

* audrey.hendricks@ucdenver.edu

Citation: Hendricks AE, Billups SC, Pike HNC, Farooqi IS, Zeggini E, Santorico SA, et al. (2018) ProxECAT: Proxy External Controls Association Test. A new case-control gene region association test using allele frequencies from public controls. PLoS Genet 14(10): e1007591. https://doi.org/10.1371/journal.pgen.1007591

Editor: Ingrid Borecki, Regeneron Pharmaceuticals Inc, UNITED STATES

Received: March 7, 2018

Accepted: July 25, 2018

Published: October 16, 2018

Data Availability Statement: Data underlying this study are available from three publicly available datasets. The UK10K SCOOP data are available from the European Genome-Phenome Archive (EGA; https://ega-archive.org) (Study ID: EGAS00001000124; Dataset ID: EGAD00001000432). The UK10K Cohorts data are also available from the EGA (Study ID: EGAS00001000713; Dataset ID: EGAD00001000776). ExAC data are available from

Abstract

A primary goal of the recent investment in sequencing is to detect novel genetic associations in health and disease improving the development of treatments and playing a critical role in precision medicine. While this investment has resulted in an enormous total number of sequenced genomes, individual studies of complex traits and diseases are often smaller and underpowered to detect rare variant genetic associations. Existing genetic resources such as the Exome Aggregation Consortium (>60,000 exomes) and the Genome Aggregation Database (~140,000 sequenced samples) have the potential to be used as controls in these studies. Fully utilizing these and other existing sequencing resources may increase power and could be especially useful in studies where resources to sequence additional samples are limited. However, to date, these large, publicly available genetic resources remain underutilized, or even misused, in large part due to the lack of statistical methods that can appropriately use this summary level data. Here, we present a new method to incorporate external controls in case-control analysis called ProxECAT (Proxy External Controls Association Test). ProxECAT estimates enrichment of rare variants within a gene region using internally sequenced cases and external controls. We evaluated ProxECAT in simulations and empirical analyses of obesity cases using both low-depth of coverage (7x) whole-genome sequenced controls and ExAC as controls. We find that ProxECAT maintains the expected type I error rate with increased power as the number of external controls increases. With an accompanying R package, ProxECAT enables the use of publicly available allele frequencies as external controls in case-control analysis.

the ExAC website (http://exac.broadinstitute.org/downloads).

Funding: This work was supported in part by the Burroughs Wellcome Fund (AEH) (Award #1015182; https://www.bwfund.org) and National Institute of Health grant U01 DK078616 (JD). IB and EZ acknowledge funding from Wellcome (WT098051 and WT206194). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

Recent investments have produced sequence data on millions of people with the number of sequenced individuals continuing to grow. Although large sequencing studies exist, most sequencing data is gathered and processed in much smaller units of hundreds to thousands of samples. These silos of data result in underpowered studies for rare-variant association of complex diseases. Existing genetic resources such as the Exome Aggregation Consortium (>60,000 exomes) and the Genome Aggregation Database (~140,000 sequenced samples) have the potential to be used as controls in rare variant studies of complex diseases and traits. However, to date, these large, publicly available genetic resources remain underutilized, or even misused, in part due to the high potential for bias caused by differences in sequencing technology and processing. Here we present a new method, Proxy External Controls Association Test (ProxECAT), to integrate sequencing data from different, previously incompatible sources. ProxECAT provides a robust approach to using publicly available sequencing data enabling case-control analysis when no or limited internal controls exist. Further, ProxECAT's motivating insight, that readily available but often discarded information can be used as a proxy to adjust for differences in data generation, may motivate further method development in other big data technologies and platforms.

Introduction

Recent investments have produced sequence data on millions of people with the number of sequenced individuals continuing to grow. Although large sequencing studies, such as the Trans-Omics for Precision Medicine (TopMed) through the National Heart, Lung, and Blood Institute, exist, most sequencing data is gathered and processed in much smaller units of hundreds to thousands of samples. This is especially true in the study of diseases that are not very common but still likely to have a complex or oligogenic genetic architecture. These silos of data mean that most rare-variant association studies of uncommon, complex diseases are underpowered. Zuk et al. suggest that sample sizes in the tens, and perhaps hundreds of thousands are required for adequate power[1]. In addition to increasing the sample size of future studies, fully leveraging existing sequencing resources could increase power considerably and could be vital in scenarios where resources to sequence more samples are limited.

Existing genetic resources such as the Exome Aggregation Consortium (ExAC; >60,000 exomes)[2] and more recently, the Genome Aggregation Database (gnomAD; ~140,000 sequenced samples) have the potential to be used as controls in studies of complex diseases. However, to date, these large, publicly available genetic resources remain underutilized, or even misused[3], in large part due to the lack of statistical methods that can appropriately use this summary level data in complex disease studies. In particular, there is a large potential for bias caused by differences in sequencing technology, processing, and read depth[3].

Recently, Lee et al[4] developed iECAT, a method to incorporate publicly available allele frequencies from controls into an existing, unbiased, but underpowered case-control analysis. They found that iECAT controls for bias while increasing power to detect association to a genetic region and can be applied to both single variant analysis and gene region analysis using a SKAT-O framework[5]. iECAT cannot be applied to very rare variants such as singletons or doubletons and requires a set of controls that were sequenced and variant-called in parallel to the cases (i.e. internal controls). Additionally, the type I error rate for iECAT increases as the size of the internal control sample set decreases relative to the internal cases.

Thus, there is still the need for methods that can incorporate very rare variants and external controls without the explicit need for large internal control samples.

Here we present Proxy External Controls Association Test (ProxECAT), a method to estimate enrichment of rare variants within a gene region using internal cases and external controls. Our method addresses existing gaps such as using singleton and doubleton variants and requiring only external controls.

Rare-variant tests in a gene are often limited to variants predicted to have a functional effect on the protein, hence discarding non-functional variants. This can result in greater power[6, 7]. The development of ProxECAT was motivated by the observation that these discarded variants can be used as a proxy for how well variants within a genetic region are sequenced and called within a sample. ProxECAT is both simple and fast, requiring only allele frequency information, and is thus well suited to use publicly available resources such as ExAC and gnomAD.

We evaluate ProxECAT in simulations, and empirical analysis of high depth of coverage (80x) whole-exome sequenced childhood obesity cases (N = 927) using both low-depth of coverage (7x) whole-genome sequenced controls (N = 3,621), and ExAC (N = 33,370). Our method controls the type I error rate in simulations and yields the expected distribution of test statistics in real data settings. Given an accompanying R package, ProxECAT provides a robust and previously unavailable method to use publicly available allele frequencies as external controls in case-control analysis. This increases the utility of existing sequenced datasets to generate hypotheses and further research into the genetic basis of disease.

Results

Proxy external controls association test

For a gene region-based test, we consider the following. Let Y denote the disease status, with Y = 1 and Y = 0 for internal case and external control status, respectively. We split the variants into those that are predicted to have a functional genetic impact and those that are not predicted to have a functional impact. We use the latter as the proxy variants. Let, x_1^f and x_1^p denote the counts of the functional and proxy rare variant alleles respectively for internal cases and x_0^f and x_0^p denote the counts of functional and proxy rare variant alleles respectively for external controls (Table 1).

We model the observed variant minor allele counts in Table 1 as a random sample from four independent Poisson distributions, i.e., $X_1^f \sim Pois(\lambda_1^f), X_0^f \sim Pois(\lambda_0^f), X_1^p \sim Pois(\lambda_1^p)$, and $X_0^p \sim Pois(\lambda_0^p)$. The derivation of the ProxECAT test statistic follows from the null hypothesis in Eq (1):

$$H_0 : \frac{\lambda_1^f}{\lambda_1^p} = \frac{\lambda_0^f}{\lambda_0^p}. \qquad (1)$$

Table 1. Data notation for internal case and external control samples for ProxECAT.

		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	

https://doi.org/10.1371/journal.pgen.1007591.t001

Using the method of Lagrange Multipliers and the constraint as defined by the null hypothesis, we find the maximum likelihood estimates (MLEs) of our parameters: $\lambda_1^f, \lambda_1^p, \lambda_0^f, \lambda_0^p$. Details are in S1 Appendix.

Our MLEs under the null hypothesis are:

$$\hat{\lambda}_1^f = \frac{(x_1^f)^2 + x_1^f x_0^f + x_1^f x_1^p + x_0^f x_1^p}{x_1^f + x_0^f + x_1^p + x_0^p}$$

$$\hat{\lambda}_0^f = \frac{(x_0^f)^2 + x_1^f x_0^f + x_0^f x_0^p + x_1^f x_0^p}{x_1^f + x_0^f + x_1^p + x_0^p}$$

$$\hat{\lambda}_1^p = \frac{(x_1^p)^2 + x_1^f x_1^p + x_1^p x_0^p + x_1^f x_0^p}{x_1^f + x_0^f + x_1^p + x_0^p}$$

$$\hat{\lambda}_0^p = \frac{(x_0^p)^2 + x_0^f x_0^p + x_1^p x_0^p + x_0^f x_1^p}{x_1^f + x_0^f + x_1^p + x_0^p}. \quad (2)$$

We use the parameter estimates in the likelihood for the constrained null hypothesis. The MLEs for the unconstrained alternative hypothesis parameters are the variant allele counts for each group (i.e. $\tilde{\lambda}_1^f = x_1^f, \tilde{\lambda}_0^f = x_0^f, \tilde{\lambda}_1^p = x_1^p, \tilde{\lambda}_0^p = x_0^p$). We then complete a likelihood ratio test (LRT) as the ratio of the constrained (null hypothesis) and unconstrained (alternative hypothesis) likelihoods, which, by Wilk's theorem[8] can be transformed to have a chi-squared distribution with 1-df.

Extension to incorporate different depths of coverage

It has been shown that functional variants have a lower minor allele frequency (MAF) distribution compared to synonymous variants[9]. Further, high-depth of coverage sequencing will detect a higher amount of variation at lower MAFs compared to low-depth of coverage sequencing[9, 10]. This results in high-depth of coverage sequencing detecting more functional variation relative to synonymous variation compared to low-depth of coverage sequencing. To allow for scenarios where sequencing coverage varies considerably between cases and controls, we weight the observed functional variant minor allele counts. Specifically, we divide the number of minor alleles for functional variants by the median ratio of the number of minor alleles for functional to synonymous variants within cases (M_1) and within controls (M_0) separately:

$$x_{1,weighted}^f = \frac{x_1^f}{M_1}$$

$$x_{0,weighted}^f = \frac{x_0^f}{M_0}.$$

The weighted functional variant minor allele counts, $x_{1,weighted}^f$ and $x_{0,weighted}^f$, are used in place of the observed functional variant minor allele counts, x_1^f and x_0^f, respectively to estimate the parameters in (2). This new test statistic is called ProxECAT-weighted.

Extension to negative binomial

By assuming a Negative Binomial distribution for the number of minor alleles in a region instead of a Poisson distribution, we extend ProxECAT to incorporate possible overdispersion. We model the Negative Binomial distribution with the mean, λ, and over-dispersion, η, parameters where the distribution approaches Poisson as η becomes large (S1 Fig).

Type I error and power simulation results

We simulated a variety of confounding scenarios. Case-control confounding represents systematic, genome-wide differences in the number of rare minor alleles observed in cases and controls due to differences in sequencing technologies and pipelines. Gene confounding refers to a gene having a higher or lower number of rare minor alleles than expected based on gene length. Gene confounding can occur in both cases and controls for a variety of reasons including differences in mutation rates, ability to detect variants, and annotation quality. Confounding can also occur when a particular gene region has a different number of rare minor alleles in cases and in controls due to sequencing differences between cases and controls. This confounding is distinct from case-control confounding in that it is isolated to a particular gene region rather than genome-wide. Here, we refer to this confounding as gene confounding only in cases. The simulation scenarios and parameters are presented in Table 2 and Supplemental Table 1.

The case-control LRT (see *Software and Statistical Analysis* under Subjects and Methods) was robust to gene confounding scenarios maintaining the appropriate type I error rate but had an increased type I error rate in the presence of case-control confounding. The case-only LRT maintained appropriate type I error rate in the presence of case-control confounding but was inflated in the presence of gene-confounding. The inflation in the type I error for the case-control LRT and the case-only LRT increased further when both gene and case-control confounding were present. This was especially true for the case-control LRT (Fig 1).

Despite usually being within the 95% confidence interval for type I error, ProxECAT appeared to have a slight, but consistent inflation (Supplemental Table 2). This minor, but consistent inflation in the type I error rate can be addressed by using a more conservative significance threshold. We found that multiplying the significance level by 0.9 works well such that a 0.045 significance threshold maintains a 0.05 type I error rate, a 0.009 significance threshold maintains a 0.01 type I error rate, etc. Both the case-control LRT used here and ProxECAT assume a Poisson distribution and had inflated Type I Error rate in the presence of overdispersion (S3 Table). ProxECAT-over, which assumes a Negative Binomial distribution instead of a Poisson distribution, corrects for overdispersion in simulations when the overdispersion parameter is known and overdispersion is not too extreme (i.e. over-dispersion, $\eta \geq 5$) (S3 Table).

Case-control LRT had higher power than ProxECAT under scenarios of no case-control confounding and given the same sample size (S4 Table). However, the power of ProxECAT increased as the sample size of the external control set increased eventually reaching higher

Table 2. Simulation parameters.

Baseline variant minor allele rate	0.001 per subject per 1Kb
Association variant minor allele rate	0.001 * (1.2, 1.4, 1.6, 1.8, 2, 3)
Gene length	20, 40 Kb
Case set sample size	500, 1000
Control set sample size	500, 1000, 10000, 40000, 100000
Gene confounding	In cases and controls: 0.001 * (1, 1.2, 1.5, 2)
	Only in cases: 0.001 * (1, 1.2, 1.5, 2)
Case control confounding	In cases: 0.001 * (1, 1.1, 1.3, 1.5)

https://doi.org/10.1371/journal.pgen.1007591.t002

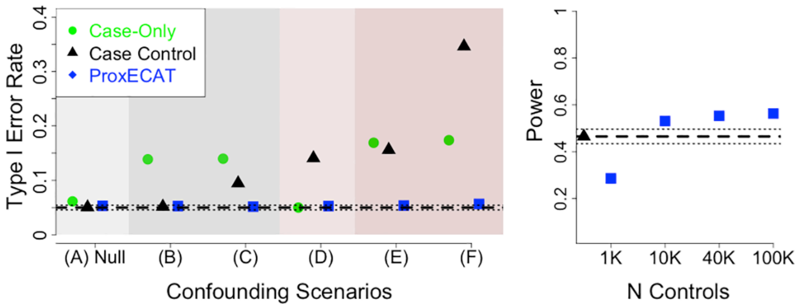

Fig 1. Type I error and power estimates for case-only LRT, case-control LRT, and ProxECAT. Estimates provided over various confounding simulation scenarios. General simulation parameters: gene-length = 20Kb, baseline mutation rate = 0.001 per person per 1Kb. **Left Plot**: type I error rate for $N_{cases} = N_{controls} = 1000$ and combinations of case-control confounding (mid level) and gene confounding (low level); dashed line represents expected type I error rate of 0.05 and dotted lines represent 95% confidence interval around the expected type I error rate. (A) Null simulation with no case-control or gene confounding bias; (B) gene-confounding; (C) gene-confounding only in cases; (D) case-control confounding; (E) case-control confounding and gene confounding; (F) case-control confounding and gene confounding only in cases. **Right Plot**: power for an effect size of 2 for case-control LRT (Ncases = 500; Ncontrols = 500) and ProxECAT (Ncases = 1000) and various external controls sample size. Dashed line is the case-control LRT power and dotted lines represent 95% confidence interval around the estimated power for case-control LRT.

https://doi.org/10.1371/journal.pgen.1007591.g001

power than the case-control LRT for the same number of internal sequences (Fig 1). This increase in power for ProxECAT is due, in part, to being able to sequence more cases with ProxECAT (N = 1000) than with a case-control LRT where sequencing resources need to be split between cases and controls (here Ncases = 500 and Ncontrols = 500). ProxECAT's power increased while the type I error stayed the same under confounding scenarios where the number of functional variants in the cases increases (S4 Table).

Assessing fit of the Poisson distribution

To assess the fit of the Poisson distribution and specifically look for over dispersion, we simulated rare minor alleles assuming a Binomial distribution for each variant and compared these

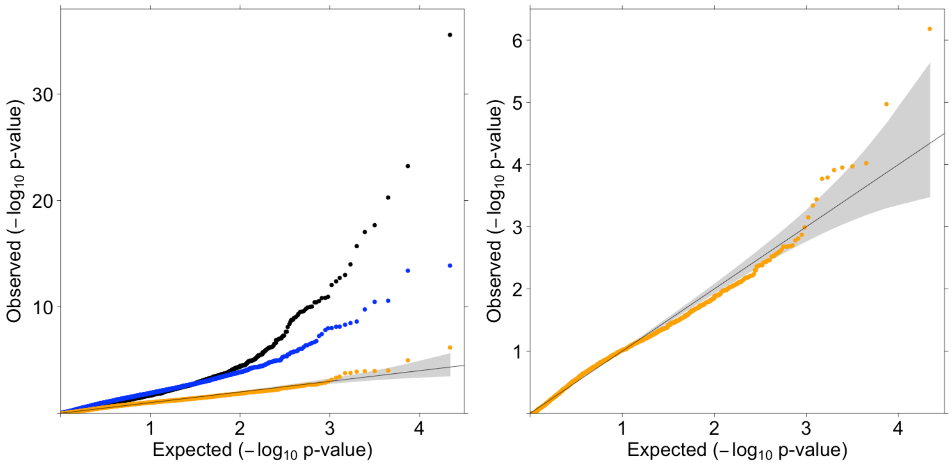

Fig 2. Quantile-Quantile plots for SCOOP cases vs. UK10K Cohort controls. Internal MAF < 0.01 in both cases and controls and number of variant minor alleles per gene ≥ 5. N genes = 11,051. 95% confidence interval of expected results in gray. ProxECAT (blue, lambda = 3.151), ProxECAT-weighted (orange, lambda = 1.026), case-control (black, lambda = 1.971). A) all tests, B) ProxECAT-weighted only.

https://doi.org/10.1371/journal.pgen.1007591.g002

results to the theoretical Poisson distribution for the number of rare minor alleles in a genetic region. No over dispersion was apparent as the sampling mean and variance of the simulated scenarios were similar across different sample sizes, MAFs, and number of minor alleles per gene (S2 and S3 Figs). When the expected number of minor alleles per gene was greater than 20, the Poisson approximation for the number of minor alleles started to look more continuous. In other words, as the expected number of variants per gene decreased, the Poisson approximation became more discrete and multimodal (S2 and S3 Figs). The theoretical distribution for the number of minor alleles per gene created from simulating genotypes for individual, independent variants from a Binomial distribution was more robust to discretization maintaining a mostly continuous distribution until the expected number of minor alleles per gene was equal to or less than four.

SCOOP data analysis

We evaluated ProxECAT using 926 cases from the Severe Childhood Onset Obesity Project (SCOOP) sample as cases and either 3,621 UK10K Cohort or 33,370 ExAC non-Finnish Europeans as controls. High-depth of coverage WES SCOOP cases vs. low-depth of coverage WGS UK10K Cohort controls had an inflated distribution of test statistics for the case-control LRT both at the center (lambda = 1.971) and in the tail of the distribution. While we did not observe inflation in the tail of the distribution for ProxECAT (Fig 2), there was a large inflation in the overall distribution of test statistics (lambda = 3.151). We observed a much higher ratio of the number of minor alleles in functional to synonymous variants per gene for the high-depth of coverage cases, median = 3.00, versus the low-depth of coverage controls, median = 1.89 (Table 3). ProxECAT-weighted, which adjusts for this systematic difference in sequencing coverage, resulted in a distribution of observed test statistics that more closely matches the expected distribution (lambda = 1.026, Fig 2).

A large strength of this method is the ability to use allele frequency data directly, rather than individual level allele calls. To assess the ability of this method to use publicly available allele frequency data, we used ExAC allele frequencies as controls for the SCOOP cases. The standard case-control LRT was inflated at both the median, lambda = 1.713, and tail (Fig 3) while our method maintained the expected distribution of test statistics. Because the depth of sequencing coverage is comparable and high for both SCOOP cases and ExAC controls, ProxECAT-weighted produced similar results to the standard, un-weighted test.

For both analyses, filtering to very rare variants was essential to avoid inflation in the distribution of observed test-statistics. This can be accomplished using moderate internal frequency filters and an external dataset such as 1000Genomes (MAF < 1%) as in the SCOOP vs UK Cohort analysis or using more stringent internal frequency filters (MAF < 0.1%) and no external dataset as in the SCOOP vs ExAC analysis.

Four genes, passing a 0.01 level of significance in both the SCOOP vs UK10K Cohort analysis and in the SCOOP vs ExAC analysis, are shown in Table 4. These results are putative novel obesity candidates meriting further replication. *MIB2* may be of particular interest as it is

Table 3. Genome-wide descriptive statistics for the ratio of the number of functional and synonymous variant minor alleles per gene in cases and controls.

		min	Q1	Median	Q3	Max
SCOOP vs UK10K Cohort	**SCOOP cases**	0.01	2.00	3.00	6.00	124
	UK10K Cohort controls	0.02	1.02	1.89	3.33	120
SCOOP vs ExAC	**SCOOP cases**	0.07	1.00	1.40	3.00	29
	ExAC	0.02	1.00	1.65	2.55	109

https://doi.org/10.1371/journal.pgen.1007591.t003

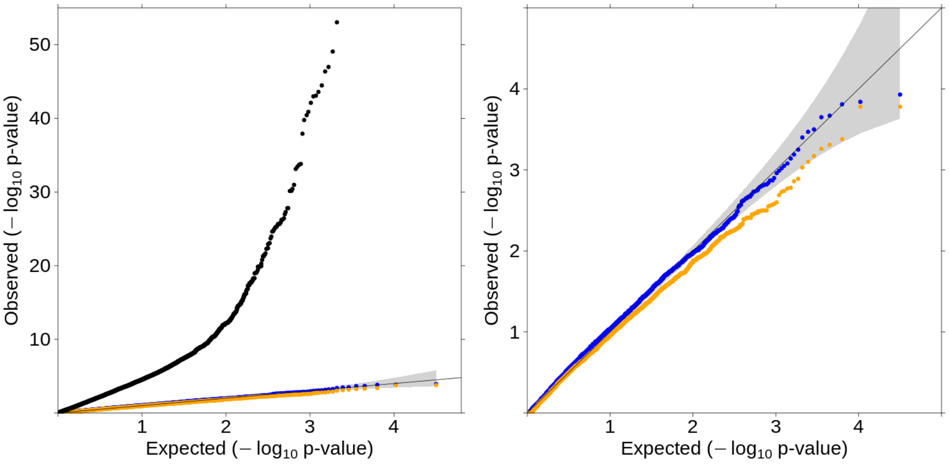

Fig 3. Quantile-Quantile plots for SCOOP cases vs. ExAC controls. Internal MAF < 0.001 in both cases and controls and number of variant minor alleles per gene ≥ 5. N genes = 15,863. 95% confidence interval of expected results in gray. ProxECAT (blue, lambda = 1.163), ProxECAT-weighted (orange, lambda = 1.069), case-control (black, lambda = 1.713) A) all tests, B) ProxECAT and ProxECAT-weighted only.

https://doi.org/10.1371/journal.pgen.1007591.g003

associated with decreased body weight in mice in the International Mouse Phenotyping Consortium (p-value = $7.49^{*}10^{-10}$, http://www.mousephenotype.org/data/genes/MGI:2679684). Additional genes with the smallest p-values are found in S5–S7 Tables.

Sensitivity of proxy selection

Within the SCOOP vs. ExAC analysis, we completed a sensitivity analysis using three increasingly broad proxy selection strategies of Sequence Ontology terms: (1) synonymous (SYN); (2) predicted low impact rating from Ensembl [11] (LOW); and (3) not in our functional category (NOT FUNC). These strategies are nested with LOW Sequence Ontology terms included in NOT FUNC, and SYN Sequence Ontology terms included in both LOW and NOT FUNC. We assessed consistency across the number of alternate alleles and in the distribution of test statistics across the three proxy selection strategies.

As expected given the nested nature of the proxy selection strategies, SYN had a smaller number of alternate alleles than either LOW or NOT FUNC and LOW had a smaller number of alternate alleles than NOT FUNC. SYN and LOW proxy selection strategies produced similar numbers of alternate alleles per gene while the correlation was lower for NOT FUNC with either SYN or LOW (S4 Fig). We found similar consistency in the distributions of test statistics between the proxy selection strategies (S5 Fig).

Table 4. Gene-based results for genes with p–value < 0.01 in SCOOP vs. Cohort and SCOOP vs ExAC.

	SCOOP vs Cohort					SCOOP vs ExAC				
	SCOOP	Cohort	p-values			SCOOP	ExAC	p-values		
			ProxECAT	ProxECAT	case			ProxECAT	ProxECAT	case
Gene	x_1^f/x_1^p	x_0^f/x_0^p		weighted	control	x_1^f/x_1^p	x_0^f/x_0^p		weighted	control
CD22	15/0	13/18	1.1E-05	2.1E-03	1.1E-04	16/1	380/247	1.5E-03	1.4E-03	1.3E-01
MIB2	0/8	62/16	1.9E-06	1.2E-04	1.1E-07	0/4	600/361	5.2E-03	1.8E-02	9.8E-09
NDEL1	13/0	18/25	1.7E-05	2.0E-03	6.6E-03	11/1	357/268	8.1E-03	5.7E-03	7.4E-01
PRDM13	9/0	13/33	1.1E-05	8.0E-03	2.9E-02	7/0	173/116	7.8E-03	6.8E-03	3.6E-01

https://doi.org/10.1371/journal.pgen.1007591.t004

Discussion

We propose a new method, ProxECAT, to test for enrichment of an accumulation of very rare variant alleles in a gene-region using publicly available external allele frequencies. ProxECAT only requires allele frequencies and uses exclusively external controls enabling the use of large, publicly available datasets such as ExAC and gnomAD. Analyses in simulations and using UK10K Cohort and ExAC as control sets for childhood obesity cases show that ProxECAT keeps the type I error rate and expected distribution of test statistics under control despite differences in sequencing technology and processing. Because ProxECAT uses external controls, additional resources can be devoted to sequencing cases. This results in greater power for ProxECAT compared to the case-control LRT test for the same number of internally sequenced individuals.

There are several limitations to the method proposed here. First, ProxECAT has a minor, but consistent inflation in the type I error rate. This limitation is easily addressed by using a more conservative significance threshold. Second, ProxECAT cannot currently include covariates such as sex, and ancestry. Thus, internal cases and external controls should be closely matched by ancestry and, as with any association study, findings will need independent replication preferably using a study where cases and controls are sequenced and processed in parallel. Third, the current approach does not enable internal controls to be analyzed along with external controls. While two analyses can be done in parallel and compared, it would be ideal to incorporate internal and external controls into the same statistical test. We are actively working on extensions to address these limitations.

It is important to highlight that research utilizing solely external controls is more susceptible to confounding due to known or unknown factors. Thus, any genes identified using ProxECAT or any method that uses only external controls should be carefully followed up in further validation, replication, and functional studies.

ProxECAT provides a robust approach to using allele frequencies from existing, publicly available sequencing data enabling case-control analysis when no or limited internal controls exist. ProxECAT uses the insight that readily available genomic information often discarded from analyses (here synonymous variation) can adjust for sizeable confounding due to differences in data generation. In the era of big data, we hope that both this insight and the ProxECAT method will enable additional genetic discoveries and will also motivate future methodological advancements in analyzing data across technologies and platforms.

Materials and methods

Software and statistical analysis

All tests were implemented using functions from our accompanying R package ProxECAT (https://github.com/hendriau/ProxECAT). Our primary test, which can model both ProxECAT and ProxECAT-weighted, was implemented with the *proxecat* function and our secondary test modeling over-dispersion was implemented using the *proxecat.over* function. We also implemented a **case-control LRT** to test for enrichment of rare, functional variant alleles in cases vs. controls and a **case-only LRT** similar to that performed by Zhi and Chen in 2012 [12]. The case-only LRT tests for enrichment of rare alleles for functional variants in each gene of interest compared to the genome-wide average number of minor alleles per gene in cases only adjusting for the length of each gene. Unless otherwise specified, we assumed the data follow a Poisson distribution for all LRTs.

Type I error and power simulations

Within each case-control confounding simulation, we simulated 20,000 independent genes under four gene-disease association and gene confounding states. The four distinct gene states are: (1) association with case status and no gene confounding, (2) association with case status and gene confounding, (3) no association with case status and gene confounding, (4) no association with case status and no gene confounding. The number of rare minor alleles per gene was simulated under a Poisson distribution or an over-dispersed Poisson modeled using a Negative Binomial parameterization using the R functions *rpois* and *rnbinom*, respectively. The mu and size parameters in *rnbinom* represent the mean and over-dispersion, respectively.

Assessing fit of the Poisson distribution

To assess the fit of the Poisson distribution, we simulated the number of each genotype group for each variant assuming Hardy-Weinberg Equilibrium and a Binomial distribution where p was the MAF. We varied the MAF (0.0001, 0.0005, 0.001, 0.005), the sample size (1000; 10,000), and the maximum number of variable variants within the gene region (5, 10, 20). We then assessed how closely the simulated distributions of the number of minor alleles observed per gene region matched a theoretical Poisson distribution where λ was the mean from each simulation scenario.

UK10K SCOOP

Whole-exome sequenced (WES) cases are from the Severe Childhood Onset Obesity Project (SCOOP) cohort[6, 13], which is a self-reported UK European subset of the Genetics of Obesity Study (GOOS). GOOS includes individuals with severe early-onset obesity body mass index (BMI) standard deviation score (SDS) $>$ 3 and age at onset of obesity $<$ 10 years. Leptin deficient individuals (identified by biochemical measurement) and those with mutations in the *MC4R* gene were excluded.

We used VerifyBamID (v1.0)[14] and a threshold of $\geq$3% to identify contaminated samples. We computed principal components with the 1000Genomes Phase I integrated call set[9] using EIGENSTRAT v4.2[15] to identify non-Europeans, and pairwise identity by descent estimates from PLINK v1.07[16] with a threshold of $\geq$0.125 to identify related individuals. Contaminated, non-European, and related samples were removed resulting in 927 SCOOP cases for analysis. Details about sequencing and variant calling for the SCOOP cases, as part of the UK10K exomes can be found elsewhere[17]. All participants gave written informed consent and all methods were performed in accordance with the relevant laboratory/clinical guidelines and regulations.

UK10K cohort

The whole-genome sequenced (WGS) controls consist of the UK10K Cohort sample, comprised of two population cohorts: the Avon Longitudinal Study of Parents and Children (ALSPAC) and the TwinsUK study from the Department of Twin Research and Genetic Epidemiology at King's College London (TwinsUK). We used allele frequency data for 3,621 individuals that passed sample QC as described elsewhere[17].

Exome aggregate consortium

We used allele frequency values for the N = 33,370 non-Finnish European (NFE) group from the ExAC variant site dataset version 1.0 (http://exac.broadinstitute.org/downloads)[2].

Variant and gene filtering

To focus on rare or very rare variants, we limited to variants below a pre-specified MAF threshold in both cases and controls. We used MAF $\leq$ 1% in the SCOOP cases vs. UK10K cohort controls analysis and MAF $\leq$ 0.1% in the SCOOP vs. ExAC analysis. For the SCOOP cases vs. UK10K controls analysis, we also applied external filtering excluding variants with a MAF > 1% in at least one of the 1000Genomes five primary ancestry groups. Exclusion by 1000Genomes MAF was not possible when using ExAC as 1000Genomes sample are included in the ExAC genotype frequencies. We explored the distribution of test statistics over several thresholds for the minimum number of functional (x^f) and proxy (x^P) variants within each gene (5, 10, and 20).

Analysis regions were limited to the intersection of respective target regions for SCOOP vs. UK10K Cohort and for SCOOP vs. ExAC. All variant annotation was applied using the GRCh37 human reference. The Ensembl Variant Effect Predictor (VEP, http://www.ensembl.org/info/docs/tools/vep/index.html [11] v79 and v90.1) from Ensembl was used to add variant consequence annotations for SCOOP vs. UK10K Cohort and SCOOP vs. ExAC respectively. We defined functional variation using the following Sequence Ontology terms[18] variant consequences: splice_donor_variant, splice_acceptor_variant, stop_gained, frameshift_variant, stop_lost, initiator_codon_variant, inframe_insertion, inframe_deletion, missense_variant, and protein_altering_variant. Variants were considered synonymous if they had the "synonymous_variant" flag. We defined the LOW proxy group as having a predicted low impact rating from Ensembl, SO terms: splice_region_variant, incomplete_terminal_codon_variant, stop_retained_variant, synonymous_variant.

Assessing results from real data analysis

We used quantile-quantile plots (QQ-plots) to assess the resulting distribution of test statistics from the real data applications. Specifically, we looked at the middle of the distribution of test statistics as assessed by the lambda value (i.e. the median of the observed test statistic divided by the median of the expected test statistic) and the tail of the distribution of test statistics, which we assessed visually.

R Package. ProxECAT R package and functions are available on github: https://github.com/hendriau/ProxECAT.

Supporting information

S1 Fig. Comparison of Poisson and Negative Binomial distributions for μ = 20.
(PNG)

S2 Fig. Comparison of the number of rare alleles in a gene region from a simulated variant level binomial distribution (black) and a theoretical Poisson distribution (red) for a sample size of 10,000. MAF = 0.0001, 0.0005, 0.001, 0.005; number of minor variant alleles within the gene region = 5, 10, 20.
(PDF)

S3 Fig. Comparison of the number of rare alleles in a gene region from a simulated variant level binomial distribution (black) and a theoretical Poisson distribution (red) for a sample size of 1,000. MAF = 0.0001, 0.0005, 0.001, 0.005; number of variants within the gene region = 5, 10, 20.
(PDF)

S4 Fig. Sensitivity analysis of proxy selection strategies in SCOOP vs. ExAC; scatter plots. Comparison of the natural log of the number of alternate alleles observed in each gene region for functional variants (FUNC) and three proxy selection strategies: synonymous (SYN), low impact (LOW), not functional (NOT FUNC). Top right panels: scatter plots with y = x line.

Bottom left panels: correlation coefficient.
(PDF)

S5 Fig. QQplots for the test results of proxy selection strategies for SCOOP vs. ExAC: Synonymous (SYN), low impact (LOW), not functional (NOT FUNC). Internal MAF < 0.001 and number of alleles per gene ≥ 5 for functional and proxy. ProxECAT (blue), ProxECAT-weighted (orange), 95% confidence interval of expected results in gray. Left: SYN, Ngenes = 15,779 (ProxECAT lambda = 1.233, ProxECAT-weighted = 1.081). Middle: LOW, Ngenes = 15,874, (ProxECAT lambda = 1.215, ProxECAT-weighted lambda = 1.119). Right: NOT FUNC, Ngenes = 16,011 (ProxECAT lambda = 1.18, ProxECAT-weighted = 1.18). For the NOT FUNC proxy group, the weights for ProxECAT-weighted are one for both cases and controls resulting in identical distributions of test statistics for ProxECAT and ProxECAT-weighted.
(PNG)

S1 Table. Gene confounding and case-control confounding simulation design. Darker shading indicates a higher level of gene confounding. Solid shading indicates gene confounding in both cases and controls. Stripped shading indicates gene confounding in only cases.
(XLSX)

S2 Table. Type I Error over all simulation scenarios.
(XLSX)

S3 Table. Type I Error for over-dispersed simulations. Gene length = 20Kb, Ncases = 1000, Ncontrols = 1000, no confounding.
(XLSX)

S4 Table. Power over all simulation scenarios.
(XLSX)

S5 Table. Top 100 results for SCOOP vs. Cohort ordered by ProxECAT-weighted p-value.
(XLSX)

S6 Table. Top 100 results for SCOOP vs. ExAC ordered by ProxECAT p-value.
(XLSX)

S7 Table. Results with p-value < 0.05 for both SCOOP vs. Cohort and SCOOP vs. ExAC.
(XLSX)

S1 Appendix. Derivation of ProxECAT.
(PDF)

S1 Results. Full results for SCOOP vs Cohort analysis.
(ZIP)

S2 Results. Full results for SCOOP vs ExAC analysis.
(ZIP)

S3 Results. Read me file for S1 and S2 Results.
(TXT)

Acknowledgments

We are indebted to the patients and their families for their participation and to the physicians involved in the Genetics of Obesity Study (GOOS).

Author Contributions

Conceptualization: Audrey E. Hendricks.

Data curation: Audrey E. Hendricks, I. Sadaf Farooqi, Eleftheria Zeggini.

Formal analysis: Audrey E. Hendricks, Hamish N. C. Pike, Inês Barroso, Josée Dupuis.

Investigation: Audrey E. Hendricks, Inês Barroso.

Methodology: Audrey E. Hendricks, Stephen C. Billups, Stephanie A. Santorico, Josée Dupuis.

Resources: Audrey E. Hendricks, I. Sadaf Farooqi, Inês Barroso.

Software: Audrey E. Hendricks.

Supervision: Audrey E. Hendricks.

Visualization: Audrey E. Hendricks.

Writing – original draft: Audrey E. Hendricks.

Writing – review & editing: Audrey E. Hendricks, Stephen C. Billups, Hamish N. C. Pike, Eleftheria Zeggini, Stephanie A. Santorico, Inês Barroso, Josée Dupuis.

References

1. 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. https://doi.org/10.1073/pnas.1322563111 PMID: 24443550; PubMed Central PMCID: PMCPMC3910587.
2. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016; 536(7616):285–91. Epub 2016/08/19. https://doi.org/10.1038/nature19057 PMID: 27535533; PubMed Central PMCID: PMCPMC5018207.
3. Barrett JC, Buxbaum J, Cutler D, Daly M, Devlin B, Gratten J, et al. New mutations, old statistical challenges. bioRxiv. 2017. https://doi.org/10.1101/115964.
4. Lee S, Kim S, Fuchsberger C. Improving power for rare-variant tests by integrating external controls. Genet Epidemiol. 2017; 41(7):610–9. Epub 2017/06/28. https://doi.org/10.1002/gepi.22057 PMID: 28657150.
5. Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. Sequence kernel association tests for the combined effect of rare and common variants. Am J Hum Genet. 2013; 92(6):841–53. Epub 2013/05/21. https://doi.org/10.1016/j.ajhg.2013.04.015 PMID: 23684009; PubMed Central PMCID: PMCPMC3675243.
6. Hendricks AE, Bochukova EG, Marenne G, Keogh JM, Atanassova N, Bounds R, et al. Rare Variant Analysis of Human and Rodent Obesity Genes in Individuals with Severe Childhood Obesity. Sci Rep. 2017; 7(1):4394. Epub 2017/07/01. https://doi.org/10.1038/s41598-017-03054-8 PMID: 28663568; PubMed Central PMCID: PMCPMC5491520.
7. Purcell SM, Moran JL, Fromer M, Ruderfer D, Solovieff N, Roussos P, et al. A polygenic burden of rare disruptive mutations in schizophrenia. Nature. 2014; 506(7487):185–90. Epub 2014/01/22. https://doi.org/10.1038/nature12975 PMID: 24463508; PubMed Central PMCID: PMCPMC4136494.
8. Wilks S. The large-sample distribution of the likelihood ratio for testing composite hypotheses The Annals of Mathematical Statistics. 1938; 9:60–2.
9. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, et al. A map of human genome variation from population-scale sequencing. Nature. 2010; 467(7319):1061–73. https://doi.org/10.1038/nature09534 PMID: 20981092; PubMed Central PMCID: PMCPMC3042601.
10. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. A global reference for human genetic variation. Nature. 2015; 526(7571):68–74. https://doi.org/10.1038/nature15393 PMID: 26432245; PubMed Central PMCID: PMCPMC4750478.
11. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016; 17(1):122. Epub 2016/06/06. https://doi.org/10.1186/s13059-016-0974-4 PMID: 27268795; PubMed Central PMCID: PMCPMC4893825.

12. Zhi D, Chen R. Statistical guidance for experimental design and data analysis of mutation detection in rare monogenic mendelian diseases by exome sequencing. PLoS One. 2012; 7(2):e31358. Epub 2012/02/22. https://doi.org/10.1371/journal.pone.0031358 PMID: 22348076; PubMed Central PMCID: PMC3277495.

13. Wheeler E, Huang N, Bochukova EG, Keogh JM, Lindsay S, Garg S, et al. Genome-wide SNP and CNV analysis identifies common and low-frequency variants associated with severe early-onset obesity. Nature genetics. 2013; 45(5):513–7. Epub 2013/04/07. https://doi.org/10.1038/ng.2607 PMID: 23563609; PubMed Central PMCID: PMCPMC4106235.

14. Jun G, Flickinger M, Hetrick KN, Romm JM, Doheny KF, Abecasis GR, et al. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Am J Hum Genet. 2012; 91(5):839–48. Epub 2012/10/25. https://doi.org/10.1016/j.ajhg.2012.09.004 PMID: 23103226; PubMed Central PMCID: PMCPMC3487130.

15. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009; 19(9):1655–64. Epub 2009/07/31. https://doi.org/10.1101/gr.094052.109 PMID: 19648217; PubMed Central PMCID: PMCPMC2752134.

16. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81(3):559–75. https://doi.org/10.1086/519795 PMID: 17701901; PubMed Central PMCID: PMCPMC1950838.

17. Consortium UK, Walter K, Min JL, Huang J, Crooks L, Memari Y, et al. The UK10K project identifies rare variants in health and disease. Nature. 2015; 526(7571):82–90. https://doi.org/10.1038/nature14962 PMID: 26367797.

18. Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, et al. The Sequence Ontology: a tool for the unification of genome annotations. Genome Biol. 2005; 6(5):R44. Epub 2005/04/29. https://doi.org/10.1186/gb-2005-6-5-r44 PMID: 15892872; PubMed Central PMCID: PMCPMC1175956.