Supplementary note for: Case-control association mapping without cases

Jimmy Z Liu [1,†], Yaniv Erlich [1,2], Joseph K Pickrell [1,3]

[1] New York Genome Center, New York, NY, USA

[2] Department of Computer Science, Columbia University, New York, NY, USA

[3] Department of Biological Sciences, Columbia University, New York, NY, USA

[†] Correspondence to: jliu@nygenome.org

June 15, 2016

Contents

1 Power calculations and simulations

1.1 Power of an association test using proxy-cases

Suppose a researcher has a fixed budget and is interested in whether to collect proxy-cases or true cases. If proxy-cases are easier to collect than true cases (for instance, if the disease is lethal), then how many proxy-cases will need to be collected instead of cases for power to detect association to be equivalent?

1.1.1 Method

Consider a study of N_A cases and N_U controls, and $N = N_A + N_U$ is the total sample size. For each SNP, let f_A be the allele frequency of allele A1 in cases and f_U be the allele frequency of A1 in controls. For all our simulations (except where noted otherwise), we assume perfect knowledge about individual phenotypes and family history. Controls do not include individuals with any affected relatives. This can be displayed in a 2×2 contingency table of expected allele counts under Hardy-Weinberg equilibrium:

	Cases	Controls
A1	$2f_A N_A$	$2f_U N_U$
A2	$2(1-f_A)N_A$	$2(1-f_U)N_U$

A test for association between genotype and disease status is then performed by calculating the Pearson χ^2 test statistic:

$$\chi^2 = \sum_{i=A1,A2} \sum_{j=A,U} \frac{(n_{ij} - E[n_{ij}])^2}{E[n_{ij}]} \tag{1}$$

where n_{ij} is the allele count of the cell for allele i and case/control status j and $E[n_{ij}] = \frac{n_{i.} n_{.j}}{n_{..}}$ is the expected cell count given the number of cases, controls and the allele frequency across all samples. Under the null hypothesis, χ^2 has a chi-squared (1-degree of freedom (df)) distribution.

For a given odds ratio (OR) and allele frequency in controls (f_U), we can calculate the expected allele frequency in cases: $f_A = \frac{f_U \times OR}{f_U \times OR + 1 - f_U}$ as well as the expected allele frequency in proxy-cases: $f_P = \frac{f_A + f_U}{2}$. We assume for now that all cases, proxy-cases and controls are perfectly classified, and that controls

do not include unaffected individuals with an affected first-degree relative. We consider the situation when some proxy-cases are misclassified as controls later on.

The power of the chi-squared test is dependent upon the noncentrality parameter (NCP), λ. Under the alternative hypothesis, χ^2 has a $\chi^2_1(\lambda)$ (noncentral chi-squared) distribution. The power of the test at significance level α is $P(\chi^2_1(\lambda) \leq \chi^2_{1;\alpha})$.

The NCP equals the test statistic minus the degrees of freedom, and is a function of the allele frequency and sample size [Evans and Purcell, 2012]. For a chi-squared (1-df) test using allele counts in true cases and controls, the NCP can be derived as:

$$\lambda_C = \frac{2N_A N_U (f_A - f_U)^2 (N_U + N_A)}{(N_A - N_A f_A + N_U - N_U f_U)(f_A N_A + f_U N_U)} \tag{2}$$

$$= \frac{2N_A r (f_A - f_U)^2 (1 + r)}{(f_A + r f_U)(1 + r - f_A - r f_U)} \tag{3}$$

and similarly for a chi-squared (1-df) test using proxy-cases and controls:

$$\lambda_P = \frac{2N_P N_U (f_P - f_U)^2 (N_U + N_P)}{(N_P - N_P f_P + N_U - N_U f_U)(f_P N_P + f_U N_U)} \tag{4}$$

$$= \frac{2N_P r (f_P - f_U)^2 (1 + r)}{(f_P + r f_U)(1 + r - f_P - r f_U)} \tag{5}$$

where N_P is the number of proxy-cases and r is the ratio of controls to cases (or controls to proxy-cases).

Given values N_A, r, f_U and OR, we wish to calculate N_P when $\lambda_C = \lambda_P$. That is, given a fixed number of true cases and controls, what is the number proxy-cases and controls required to have equivalent power to detect association as true cases? We can solve $\lambda_C = \lambda_P$ numerically by iterating through values of N_P in equations (2) and (4).

Alternatively, if it can be assumed that r is constant, the relationship between N_A and N_P when $\lambda_C = \lambda_P$ can be derived as following:

$$\frac{2N_A r (f_A - f_U)^2 (1 + r)}{(f_A + r f_U)(1 + r - f_A - r f_U)} = \frac{2N_P r (f_P - f_U)^2 (1 + r)}{(f_P + r f_U)(1 + r - f_P - r f_U)} \tag{6}$$

substituting $f_P = \frac{f_A + f_U}{2}$ and solving for N_P:

$$N_P = \frac{N_A(f_A^2 + 2f_A(2rf_U + f_U - r - 1) + f_U(2r+1)(2rf_U + f_U - 2(r+1)))}{f_A^2 + f_A((2f_U - 1)r - 1) + rf_U((f_U - 1)r - 1)} \tag{7}$$

$$= \frac{N_A(2r(f_U - 1) + f_U + f_A - 2)(f_U(2r+1) + f_A)}{(r(f_U - 1) + f_A - 1)(rf_U + f_A)} \tag{8}$$

and differentiating with respect to N_A:

$$\frac{\partial N_P}{\partial N_A} = \frac{(2r(f_U - 1) + f_U + f_A - 2)(f_A + f_U(2r+1))}{(f_A + r(f_U - 1) - 1)(f_A + rf_U)} \tag{9}$$

$$= \frac{(f_U - 2)(f_A - f_U)}{f_A + rf_U} - \frac{(f_U + 1)(f_A - f_U)}{f_A + r(f_U - 1) - 1} + 4 \tag{10}$$

Our analysis so far assume that cases, proxy-cases and controls are perfectly classified. In situations where a certain number of proxy-cases are misclassified as controls, we would expect a decline in effect size (and hence power to detect association) for a given sample size when compared with a perfectly classified proxy-case control design. We can model this by replacing f_U in equations (4) and (5) with $f_U^* = mf_P + (1-m)f_U$, where m is the proportion of controls that are actually proxy-cases.

1.1.2 Results

We estimated N_P for different values of N_A across different effect sizes and allele frequencies. When r is constant between cases and controls, there is a linear relationship between the number of true cases and number of proxy-cases to achieve the same power to detect association (Figure 1). In general, a study with proxy-cases will need four times as many samples as true cases to achieve equivalent power to detect association - in equation (10), $\frac{\partial N_P}{\partial N_A} \approx 4$ across the effect size and allele frequency spectrum. For example, a study of 1,000 true cases and 1,000 controls will have roughly have same power to detect association as one with 4,000 proxy-cases and 4,000 controls.

The above assumes that r remains constant between the case/control and its equivalent proxy-case/control study. In a situation where the number of controls is fixed and a researcher needs to decide whether to collect true cases

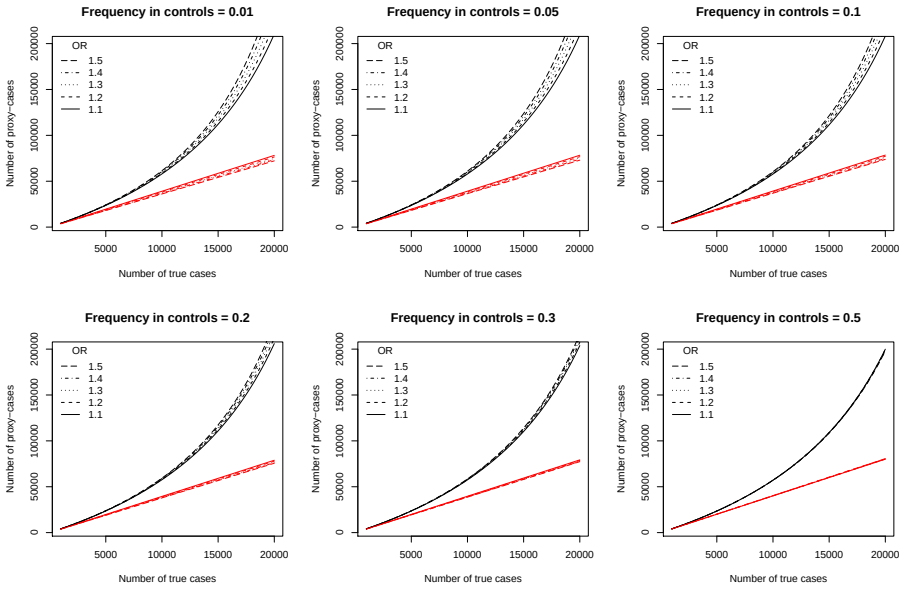

Figure 1: Number of proxy-cases and true cases when power to detect association is equivalent. Red lines indicate a study design where the ratio of cases to controls r is constant between true case and proxy-case study. Black lines indicates a study design where number of controls is fixed at 100,000.

or proxy-cases, r will decrease as the number of true cases increases. In practice, this means that the ratio of proxy-cases to true cases when power is equivalent will increase as the number of true cases increases (Figure 1). For instance, with the number of controls fixed at 100,000, a study of 1,000 true cases is equivalent to one with $\sim$4,125 proxy-cases; while one with 5,000 true cases is equivalent to $\sim$23,700 proxy-cases.

1.2 Power of an association test when true cases and proxy-cases are considered jointly

We consider a cohort study where both true cases and proxy-cases are available for analysis, and compare power to detect association under three different association models: 1) a 2×2 chi-square (1-df) test using true cases and controls, 2) a 2×2 chi-square (1-df) test when cases and proxy-cases are lumped together vs. controls, and 3) a 3×2 chi-square (2-df) test where cases, proxy-cases and controls considered separately. The expected number of true cases and proxy-cases in a randomly sampled cohort is based on the disease prevalence and heritability of disease liability.

1.2.1 Method

Let there be a cohort composed of N individuals. For a typical case/control association study, the cohort is split into approximately $N_A = N \times K$ cases and $N_U = N \times (1 - K)$ controls where K is the disease prevalence. For each SNP, the allele counts can be denoted as before:

	Cases	Controls
A1	$2f_A N_A$	$2f_U N_U$
A2	$2(1 - f_A)N_A$	$2(1 - f_U)N_U$

where f_A is the frequency of A1 in cases and f_U is the frequency of A1 in controls. The odds ratio is:

$$OR = \frac{f_A(1 - f_U)}{f_U(1 - f_A)} \tag{11}$$

with NCP:

$$\lambda_C = \frac{2N_A N_U (f_A - f_U)^2 (N_U + N_A)}{(N_A - N_A f_A + N_U - N_U f_U)(f_A N_A + f_U N_U)} \tag{12}$$

We wish to calculate the NCP when cases are composed of both true cases and proxy-cases. To do this, we need to know how the allele counts are expected to vary when a proportion of controls are reassigned to be proxy-cases. The number of expected proxy-cases in a randomly sampled cohort can be estimated using the liability threshold model of disease.

The liability distribution of individuals with an affected first-degree relative is:

$$L_1 \sim N\left(\frac{h_L^2 i}{2}, 1 - \frac{h_L^4 i(i-T)}{4}\right) \tag{13}$$

where h_L^2 is the narrow-sense heritability of liability, T is the truncation point of a standard normal distribution at disease prevalence K, z is the height of the standard normal distribution at T and $i = z/K$ [Falconer and Mackay, 1996]. The probability that an affected individual's first degree relative is also affected is then $K_1 = 1 - \Phi(T_1)$ where $\Phi(.)$ is the standard normal cumulative distribution function and

$$T_1 = \frac{T - h_L^2 i/2}{\sqrt{1 - h_L^4 i(i-T)/4}} \tag{14}$$

In a cohort study, the total number of individuals with at least one affected parent is approximately $(1-(1-K)^2) \times N$. Of these, $K_1 \times N_A$ individuals are also affected themselves. Subtracting one from the other, the total number of proxy-cases (unaffected individuals with an affected parent) is:

$$N_P = (1-(1-K)^2) \times N - (1-(1-K_1)^2) \times N_A \tag{15}$$

Let $f_P = \frac{f_A + f_U^*}{2}$ be the allele frequency of A1 in proxy-cases, where f_U^* is the allele frequency of A1 in controls that do not include proxy-cases ("true controls"). In an association study where cases are composed of both truly affected individuals and proxy-cases, the allele count of A1 in cases increases by $2f_P N_P$ while the number in controls decreases by the same amount. Hence, we can estimate the allele frequency in true controls:

$$f_U^* = \frac{f_U N_U - f_P N_P}{N_U - N_P} \tag{16}$$

Substituting $f_P = \frac{f_A + f_U^*}{2}$:

$$f_U^* = \frac{2f_U N_U - f_A N_P}{2N_U - N_P} \tag{17}$$

Similarly, the allele frequency in cases that now include proxy-cases:

$$f_A^* = \frac{f_A N_A + f_P N_P}{N_A + N_P} \tag{18}$$

Hence the allele counts are now:

	Cases	Controls
A1	$2f_A^*(N_A + N_P)$	$2f_U^*(N_U - N_P)$
A2	$2(1 - f_A^*)(N_A + N_P)$	$2(1 - f_U^*)(N_U - N_P)$

and NCP:

$$\lambda^* = \frac{2(N_A + N_P)(f_A^* - f_U^*)^2(N_U + N_A)}{(N_A - f_A^*(N_A + N_P) + N_U - f_U^*(N_U - N_P))(f_A^*(N_A + N_P) + f_U^*(N_U - N_P))} \tag{19}$$

The power of the chi-square (1-df) test at significance level α using just true cases is $P(\chi_1^2(\lambda_C) \leq \chi_{1;\alpha}^2)$ and when using both true cases and proxy-cases is $P(\chi_1^2(\lambda^*) \leq \chi_{1;\alpha}^2)$.

We next consider power calculations for a test where cases, proxy-cases and controls are considered separately. The allele counts are now:

	Cases	proxy-cases	Controls
A1	$2f_A N_A$	$2f_P N_P$	$2f_U^*(N_U - N_P)$
A2	$2(1 - f_A)N_A$	$2(1 - f_P)N_P$	$2(1 - f_U^*)(N_U - N_P)$

with NCP for a 3×2 chi-square (2-df) test:

$$\lambda^\dagger = 2(N_A + N_U) \sum_{i=1}^{6} \frac{(p_{1i} - p_{0i})^2}{p_{0i}} \tag{20}$$

where p_{0i} and p_{1i} are the proportions in cell i under the null and alterate hypotheses respectively Cohen [1977]. The power at significance level α of the chi-square (2-df) test is then given by $P(\chi_2^2(\lambda^\dagger) \leq \chi_{2;\alpha}^2)$.

1.2.2 Results

We estimated the power to detect association at $\alpha = 5 \times 10^{-8}$ in a cohort of 100,000 individuals across different values of K, h^2, f_U, and OR using three models: 1) true cases vs. controls (1-df test), 2) true cases + proxy-cases vs. controls (1-df test) and 3) true cases vs. proxy-cases vs. controls (2-df test) (Figure 2).

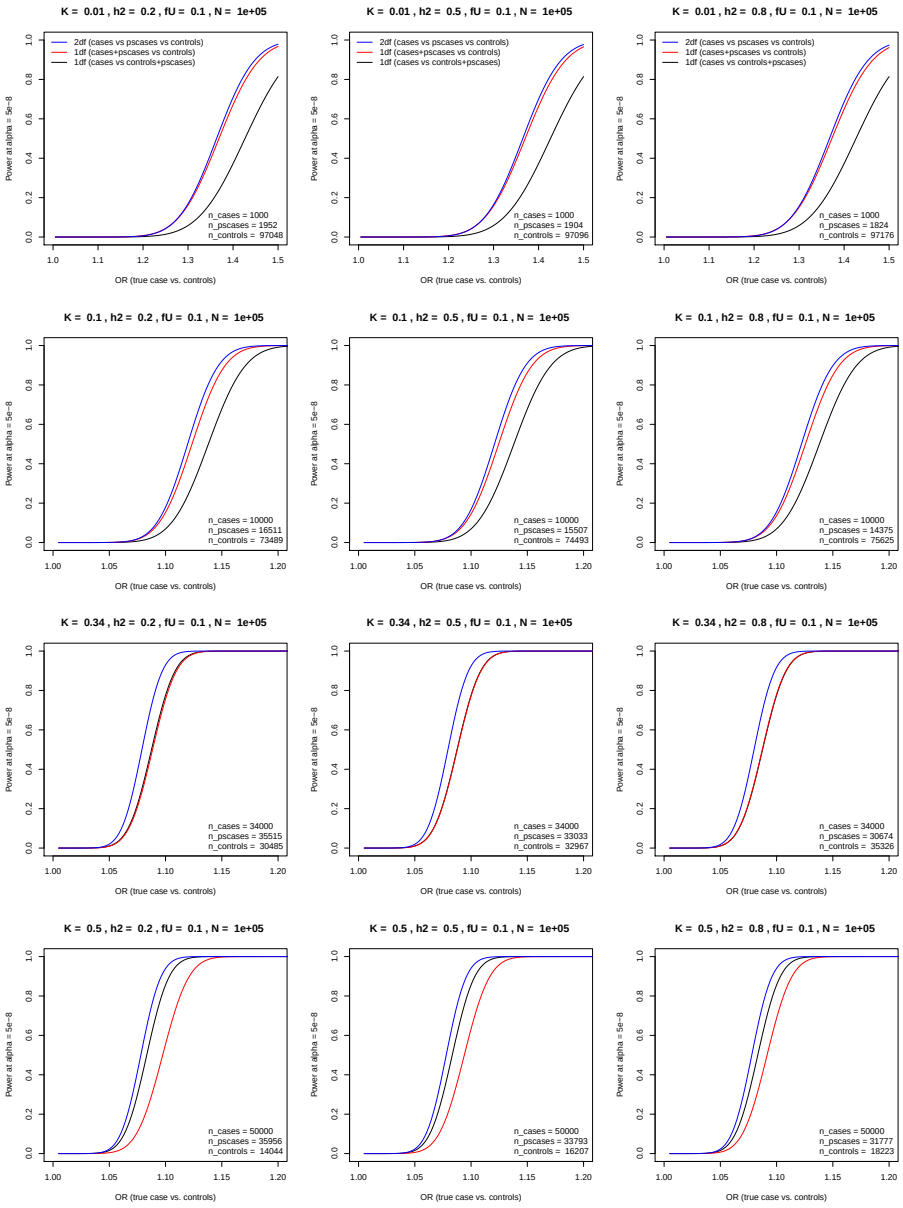

Figure 2: Power to detect association at $\alpha = 5 \times 10^{-8}$ using three study designs. The power of the 2-df test comparing cases against proxycases and against controls is shown in blue, the 1-df test comparing cases and proxycases against controls is in red, and the 1-df test comparing cases against controls and proxycases is shown in black. We set the total sample size (N) to 100,000, the allele frequency in controls (fU) to 0.1, and then varied the disease prevalence (K), heritability of disease liability (h2) and odds ratio (OR). The expected number of cases and proxy-cases is shown in the bottom right corner of each panel.

In all situations, accounting for proxy-cases using the 2-df test improved power to detect association. For diseases with prevalence less than $\sim$ 10%, the 2-df test was marginally more powerful that the 1-df test of lumping true and proxy-cases together. For instance, for a disease with 5% prevalence and 50% heritability, we expect to observe 5,000 cases and 8,597 proxy-cases in a randomly sampled cohort of 100,000 individuals. Here, for a SNP with allele frequency 0.1 in controls, an OR of 1.2 and $\alpha = 5 \times 10^{-8}$, there is 60.2% power to detect association using a 2×2 test of true cases vs. controls, 83.7% using a 2×2 test of true cases + proxy-cases vs. controls and 87.7% using a 3×2 test of true cases vs. proxy-cases vs. controls.

The 1-df test of lumping cases and proxy-cases becomes less powerful than a standard case/control test when disease prevalence becomes higher than ~ 0.34. This is because as K increases, so does the number of proxy-cases, such that reassigning them from controls to cases leads to a much smaller proportional increase (and sometimes a decrease) in the total effective sample size than if prevalence was lower (while at the same time losing power due to the decrease in the effect size). For instance, when $K = 0.34$ and $h^2 = 0.5$, we expect 34,000 cases and 33,033 proxy-cases in a cohort of 100,000. The effective sample size ($N_{eff} = \frac{4}{\frac{1}{N_A}+\frac{1}{N_U}}$) of a standard case/control design is 89,760, while lumping cases with proxy-cases gives an effective sample size of 88,395. In contrast, using the same parameters except $K = 0.05$, lumping proxy-cases with cases increases the effective sample size 19,000 to 46,993.

The overall gain in effective sample size from the 2-df design is shown in Figure 3. The improvement in effective sample size for the 2-df test when compared with the 1df case/control test is greatest when disease prevalence is low. For instance, when $K = 0.005$, the 2-df design represents a 1.36$\times$ increase in sample size compared with the 1-df case/control design. This ratio decreases to 1.2 when $K = 0.2$.

1.3 Discussion

We compared power to detect association with proxy-cases under two broad scenarios. In the first, a choice needs to be made whether to collect only cases or

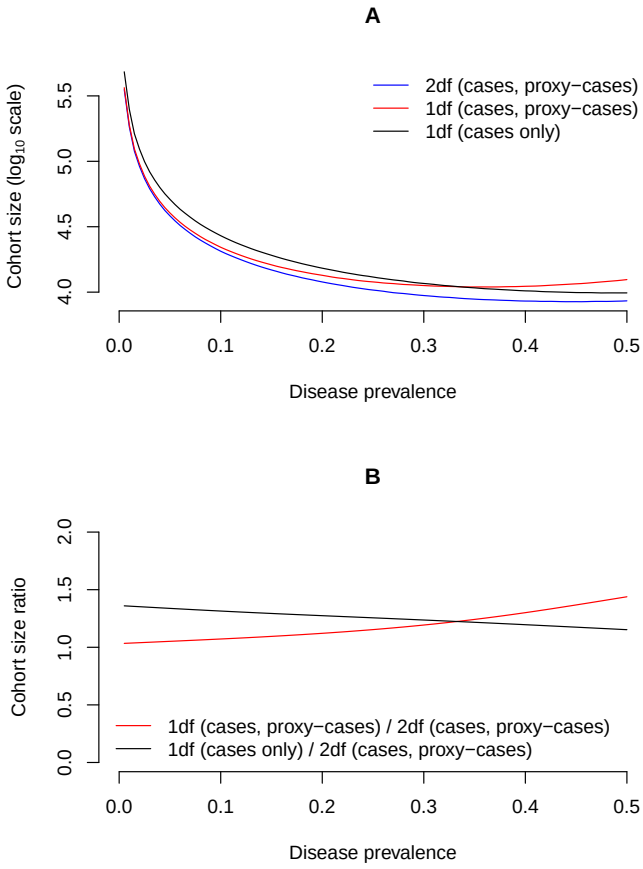

Figure 3: Equivalent sample sizes of case/proxy-case/control designs in a population cohort. Disease prevalence is shown in the x-axis. (A) The y-axis shows the total cohort sizes of the 2-df, 1-df (cases, proxy-cases) and 1-df (cases only) study designs required for 80% power to detect association at $\alpha = 5 \times 10^{-8}$ for a SNP with odds ratio = 1.2, allele frequency = 0.4 and heritability = 0.5. (B) The y-axis shows the relative size of the 1-df (cases, proxy-cases) and 1-df (cases only) study designs compared with the 2-df design that is required to achieve the same power to detect association. These ratios remain constant across effect sizes and allele frequencies.

only proxy-cases, and the second, where both true and proxy-cases are collected. The first situation may prove useful if the collection of true cases is somehow prohibitive compared to proxy-cases - for instance, the disease may be very rare, lethal or very late onset.

The second situation is perhaps more likely to be encountered in the context of large population cohort studies where phenotypes of genotyped individuals as well as their relatives are collected. Here, we show that it is always more powerful to explicitly account for proxy-cases using the 3×2 chi-squared test than performing a standard case/control test. The approach of lumping cases and proxy-cases together is marginally less powerful than the 3×2 test for diseases with prevalence $< 10\%$, though this approach loses power compared to the others as prevalence increases. In practice, using the latter approach and performing association using logistic regression or mixed models so that covariates and population structure can be accounted for may prove to be effective for diseases with low to moderate prevalence.

The expected number of proxy-cases in a population cohort is estimated assuming that the disease prevalence is the same between genotyped individuals and their first degree relatives. For late onset diseases where the prevalence is higher in the parents of genotyped individuals, accounting for proxy-cases will become even more attractive than just using cases.

2 A novel Parkinson's disease association at *SLIT3*(?)

In the primary GWAX in the UK Biobank individuals, we identified a previously unreported Parkinson's disease risk locus near the *SLIT3* gene (rs1806840, $P = 6.39 \times 10^{-9}$, Figure 4). Given the sample size of 4051 proxy-cases and 110,402 controls (effectively equivalent to a case/control study with N = $\sim$ 4440) is an order of magnitude smaller than in Nalls et al. [2014] (13,708 cases and 95,282 controls), we would not have expected to identify any genome-wide significant locus that had not already been identified this earlier study. Indeed, all other 23 genome-wide significant loci were at established risk loci for their respective diseases. The lead SNP at this locus, rs1806840, is also listed with association $P > 0.05$ in Nalls et al. [2014] at `pdgene.org`.

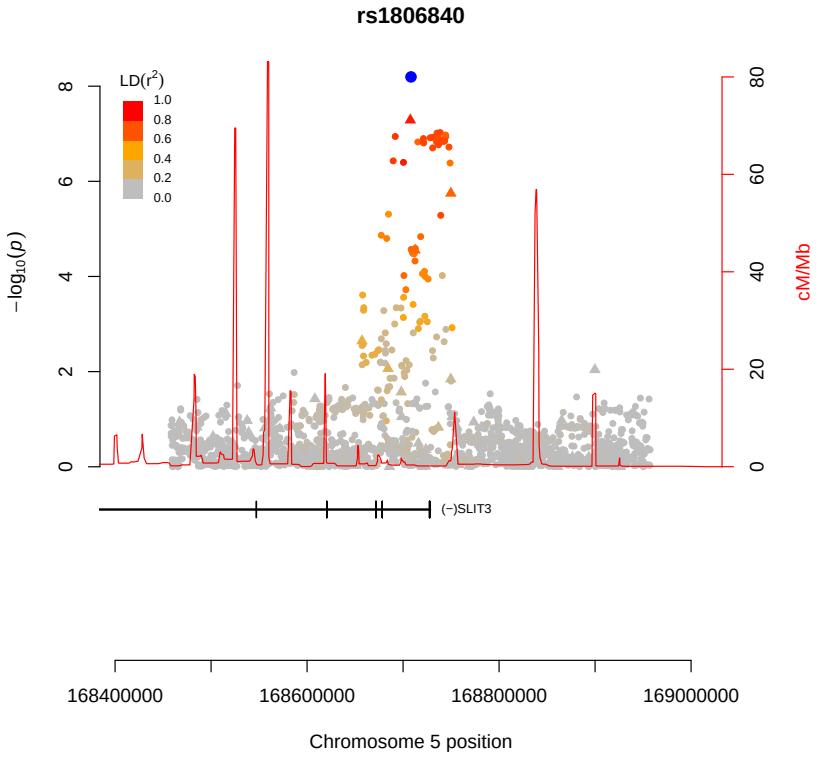

Figure 4: Regional association plot of the novel Parkinson's disease association at *SLIT3*. The coordinates on the x-axis are GRCh37. SNPs are plotted according to their base-pair position and strength of association. The color of each point indicates the degree of linkage disequilibrium with rs1806840 (in blue). Circle points represent imputed SNPs; triangle points represent directly genotyped SNPs.

The lead SNP does not appear to have any obvious QC issues, with Hardy-Weinberg $P = 0.79$, missingness = 0.018, genotyping batch effects on missingness $P = 0.87$, and batch effects on allele frequency $P = 0.73$. Imputation quality is also high, with IMPUTE2 INFO = 0.99. The strongest association signal at a directly genotyped SNP, rs11134568, is $P = 5.18 \times 10^{-8}$. In EUR individuals from 1000 Genomes Project Phase 3, LD between rs1806840 and rs11134568 is $r^2 = 0.87$. The strength of association in the region appears consistent with LD patterns (Figure 4).

We were also unable to rule out population stratification driving the association at this locus. The signal persisted when the first 20PCs were used as covariates in the logistic regression ($P = 2.15 \times 10^{-10}$). We also performed association testing using a linear mixed model implemented in BOLT-LMM [Loh et al., 2015], where genetic relatedness between the UK Biobank was estimated using 623,852 directly genotyped SNPs. Using this method, which tries to account for cryptic population stratification, rs1806840 remained genome-wide significant (Table S2, $P = 5.9 \times 10^{-9}$).

While we urge caution in interpreting this signal as a bona fide Parkinson's disease risk locus, there is no obvious evidence from our data to suggest that it is a false positive. We hope that future large genetic studies of Parkinson's disease, especially those that can compare UK vs other populations, will shed further light on this locus.

References

Cohen, J., 1977. Chapter 7 - Chi-Square Tests for Goodness of Fit and Contingency Tables. In Cohen, J., editor, *Statistical Power Analysis for the Behavioral Sciences (Revised Edition)*, pages 215 – 271. Academic Press, revised edition.

Evans, D. M. and Purcell, S., 2012. Power Calculations in Genetic Studies. *Cold Spring Harbor Protocols*, **2012**(6):pdb.top069559.

Falconer, D. and Mackay, T., 1996. Chapter 18 - Threshold Characters. In *Introduction to Quantitative Genetics*, pages 299–311. Benjamin Cummings, fourth edition.

Loh, P.-R., Tucker, G., Bulik-Sullivan, B. K., Vilhjalmsson, B. J., Finucane, H. K., Salem, R. M., Chasman, D. I., Ridker, P. M., Neale, B. M., Berger, B., *et al.*, 2015. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet*, **47**(3):284–290.

Nalls, M. A., Pankratz, N., Lill, C. M., Do, C. B., Hernandez, D. G., Saad, M., DeStefano, A. L., Kara, E., Bras, J., Sharma, M., *et al.*, 2014. Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nature Genetics*, **46**(9):989–993.