bioRxiv preprint doi: https://doi.org/10.1101/045831; this version posted June 17, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 17, 2016

bioRxiv preprint doi: https://doi.org/10.1101/045831; this version posted June 17, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

The case-control association study is a powerful method for identifying genetic variants that influence disease risk. However, the collection of cases can be time-consuming and expensive; if a disease occurs late in life or is rapidly lethal, it may be more practical to identify family members of cases. Here, we show that replacing cases with their first-degree relatives enables genome-wide association studies by proxy (GWAX). In randomly-ascertained cohorts, this approach enables previously infeasible studies of diseases that are absent (or nearly absent) in the cohort. As an illustration, we performed GWAX of 12 common diseases in 116,196 individuals from the UK Biobank. By combining these results with published GWAS summary statistics in a meta-analysis, we replicated established risk loci and identified 17 newly associated risk loci: four in Alzheimer's disease, eight in coronary artery disease, and five in type 2 diabetes. In addition to informing disease biology, our results demonstrate the utility of association mapping using family history of disease as a phenotype to be mapped. We anticipate that this approach will prove useful in future genetic studies of complex traits in large population cohorts.

bioRxiv preprint doi: https://doi.org/10.1101/045831; this version posted June 17, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

In a typical case-control genetic association study, a researcher genotypes a set of individuals that have a disease (the "cases") and a set of individuals that do not have the disease (the "controls"). For each genetic variant, the difference in allele frequency between cases and controls can be used to estimate the causal effect of the genetic variant on the disease (assuming all potential confounders have been accounted for). While powerful, this study design requires an *a priori* decision about which disease is of interest, as well as substantial effort to identify matched cases and controls. An alternative approach is a cohort study, in which individuals are sampled from the general population and many phenotypes (along with genotypes) are collected on each individual. An advantage of a cohort study is that the cohort can be subdivided to create case-control studies of many different diseases.

However, cohort studies are limited by the fact that unbiased sampling may not yield sufficient numbers of cases to enable powerful case-control studies. For example, even in a perfectly representative sample of a million people one expects only 10,000 cases of a disease like schizophrenia with a population prevalence of 1%. Further, participants in a cohort study are rarely a fully representative sample of a population; a disease may also be rare in a cohort for the simple reason that the sampled population does not include the demographic group where the disease is most prevalent. For example, the UK Biobank (an ongoing and widely-available cohort study) sampled individuals in the age range of 40-69 (at the time of recruitment) [Sudlow et al., 2015]. By definition, this cohort does not include individuals with lethal childhood diseases, and at present there are only a handful of individuals with Alzheimer's disease or other late-onset diseases. Similarly, cohort studies that focus on individuals of a single sex (like the Nurses Health Study) have little power to study diseases that are more common in the other sex. Other sampling approaches, like cohorts made from customers of consumer genomics companies (e.g. Eriksson et al. [2010], `DNA.Land`), have analogous limitations. More generally, the number of cases of a given disease present in a cohort will be a function of aspects of the disease (with rarely-occurring or rapidly-lethal diseases being more rare) and aspects of the sampling.

In this paper, we consider a study design where the researcher genotypes family members of cases, rather than cases themselves (since the cases may be difficult or impossible to contact). This design is popular in studies of longevity (where "cases" are long-lived individuals, see e.g. Barzilai et al. [2003]; Joshi et al. [2016]; Pilling et al. [2016]; Tan et al. [2010]), but has not been widely used in other situations. The approach can be thought of as taking pedigree-based association methods that allow for missing genotype data (e.g. Gudbjartsson et al. [2008]; Kong et al. [2009]; Thornton and McPeek [2007]) to an extreme where no cases have been genotyped or phenotyped by the researcher.

As a motivating example for this type of design, consider Alzheimer's disease. As of March 25, 2016, there are 55 cases of Alzheimer's disease listed among the approximately 500,000 individuals in the UK Biobank. However, over 60,000 individuals note that one or both of their parents was/is affected with the

bioRxiv preprint doi: https://doi.org/10.1101/045831; this version posted June 17, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

disease. An individual with a single affected parent can be thought to have one chromosome sampled from a population of "cases" and one from a population of "controls". If the allele frequency (in the standard case-control setting) of some variant that increases risk of a disease is f_A in cases and f_U in controls, then the allele frequency in individuals with a single affected parent is $\frac{f_U+f_A}{2}$. This motivates a "proxy-case"-control association study where "proxy-cases" are the relatives of affected individuals and "controls" are the relatives of unaffected individuals. We refer to this approach as a Genome-Wide Association study by proXy (GWAX).

Results

Power of genome-wide association by proxy

We first explored the power of this approach with simulations and analytical calculations. Specifically, we focused on the situation where we have information about the diseases of the parents of an individual (Methods). We initially considered the case where we have no phenotype information about genotyped individuals themselves, though we consider this case later on.

The GWAX approach using proxy-cases who have one affected first-degree relative reduces the log odds ratios by a factor of around two when compared with a traditional case-control design (assuming an additive model for the impact of a genetic variant on a disease). This reduction in effect size reduces power to detect association. However, using proxy-cases may increase the effective sample size (in a cohort study) or be more logistically feasible than collecting standard cases, thus offsetting this loss in power. We calculated the number of proxy-cases and controls required such that the power to detect association is equivalent to using true cases and controls (Supplementary Note). Across the allele frequency and effect size spectrum, the proxy-case-control approach is more powerful when there are about four times (or more) as many proxy-cases and controls as there are true cases and controls, assuming the ratios of controls to cases and controls to proxy-cases are the same (Figure 1A). This ratio increases to ~ 4.9 if 10% of controls are in fact misclassified proxy-cases (Supplementary Note, Figure S8). For late onset diseases such as Alzheimer's disease (1.6% in the population vs. 42% in those over the age of 84 [Hebert et al., 2003]) and Parkinson's disease (0.3% in the population vs. 4% in those over 80 [de Lau and Breteler, 2006]), the proxy-case-control design gains substantial power if cohorts are sampled randomly from the population.

We next explored the situation where we have information about the phenotypes of the genotyped individuals as well. In this situation, we have "true cases" (genotyped individuals with a disease), "proxy-cases" (unaffected individuals with a parent with the disease), and "controls". We considered analysis approaches that treat all three of these groups separately (in a 3×2 chi-squared test) or lumping together the "proxy-cases" and "true cases" and performing a standard 2×2 chi-squared test. When both true and proxy-cases are available in a population cohort study, accounting for this fact increases power (Figure 1B, Supplementary Note). For instance, for a disease with 5% prevalence and 50% heritability on the liability scale across

bioRxiv preprint doi: https://doi.org/10.1101/045831; this version posted June 17, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

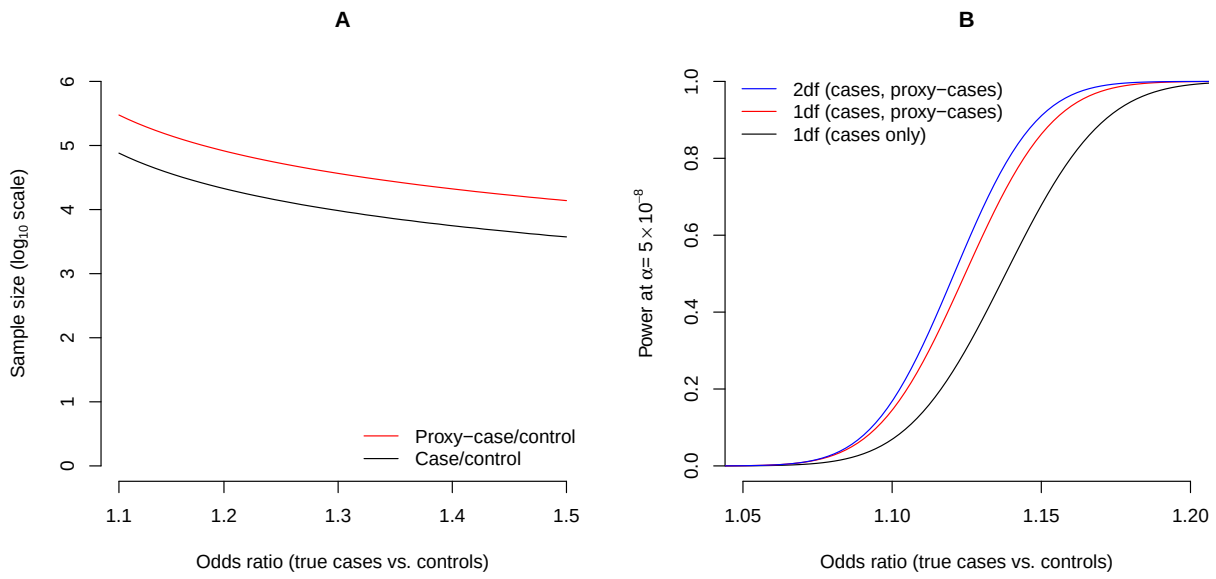

Figure 1. Power of proxy-case-control association designs. (A) Total sample size required for 80% power to detect association at $\alpha = 5 \times 10^{-8}$ for case-control (black line) and proxy-case-control (red line) designs at a SNP with 0.1 frequency in controls. (B) Power to detect association at $\alpha = 5 \times 10^{-8}$ using two designs that account for cases and proxy-cases (in red and blue) and a standard case-only/control design (in black). The total sample size = 100,000, disease prevalence = 0.1, heritability of liability = 0.5 and allele frequency in controls = 0.1 (See Supplementary Note).

all age groups, we expect to observe 5,000 cases and 8,597 proxy-cases in a randomly sampled cohort of 100,000. Here, for a SNP with allele frequency 0.1 in controls and an odds ratio of 1.2, there is 60.2% power at $\alpha = 5 \times 10^{-8}$ to detect association using a standard 2×2 chi-squared test of true cases vs. controls, 87.2% power using a 2×2 test where cases and proxy-cases are lumped together, and 89.8% power using a 3×2 test where true cases, proxy-cases and controls are treated separately (Supplementary Note). In this situation, treating cases, proxy-cases and controls separately boosts the effective sample size by $1.34\times$ when compared to a case-control design. Overall, the boost in effective sample size ranges from $1.36\times$ to $1.28\times$ for disease prevalences from 1% to 20%. When disease prevalence is greater than around 34%, the test where cases and proxy-cases are lumped together is less powerful than a standard case-control test since there are no further gains in effective sample size. Nevertheless, across simulated effect sizes, allele frequencies, heritability and disease prevalences, the 3×2 test is consistently more powerful than the case-control test (see Supplementary Note for details).

Application to the UK Biobank

We performed GWAX of 12 diseases in the UK Biobank (May 2015 Interim Release). After quality control and 1000 Genomes Phase 3 imputation (Methods), $\sim$ 10.5 million low-frequency and common ($MAF > 0.005$) SNPs from 116,196 individuals of European ancestry were available for analysis. All of these individuals answered questionnaires regarding the diseases of their family members (though the med-

bioRxiv preprint doi: https://doi.org/10.1101/045831; this version posted June 17, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

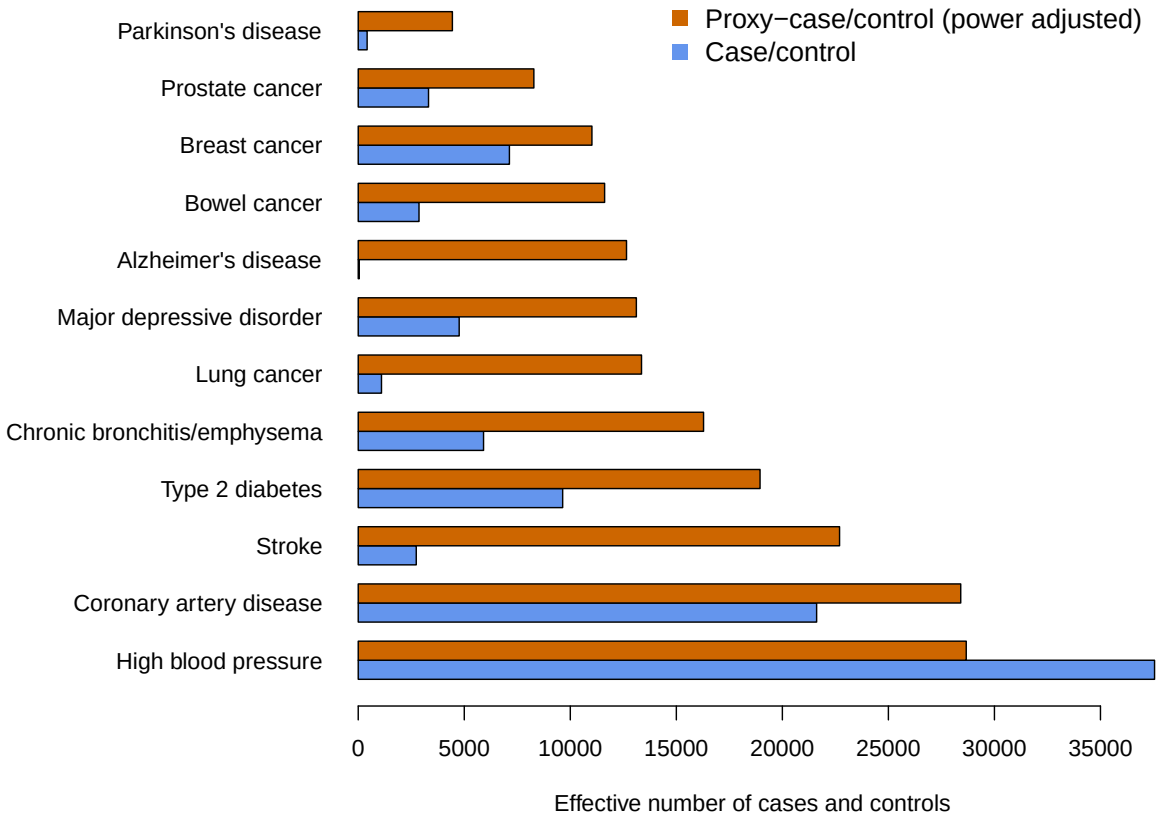

Figure 2. Effective sample sizes of case-control versus proxy-case-control association designs in the UK Biobank. The effective sample size for each design is: $N_{eff} = \frac{4}{1/N_A + 1/N_U}$ where N_A is the number of cases (or proxy-cases) and N_U is the number of controls. To account for power, we divided the effective sample size in proxy-cases/controls by four. Case/control counts for breast cancer and prostate cancer only include females and males, respectively.

ical records of the individuals themselves are available, we did not use them in this analysis in order to illustrate the approach without using cases). The number of proxy-cases per phenotype ranged from 4,627 for Parkinson's disease to 54,714 for high blood pressure (Table S1). Based on these sample sizes, we expect greater power to detect association using GWAX than a case-control GWAS for 11 of the 12 phenotypes (high blood pressure being the exception) in the UK Biobank cohort (Figure 2).

Association testing was performed using logistic regression with age, sex and the first four principal components as covariates. For each SNP, we calculated an adjusted odds ratio (OR), which is directly comparable (under a standard additive model) with ORs estimated from traditional case-control study designs (Methods, Figure S1). The overall association results across the 12 phenotypes are shown in Manhattan plots, which show several clear peaks of association (Figure S2).

In the GWAX of these 12 diseases, 24 loci reached "genome-wide significance" ($P < 5 \times 10^{-8}$). For Alzheimer's disease, breast cancer, heart disease, high blood pressure, lung cancer, prostate cancer and

bioRxiv preprint doi: https://doi.org/10.1101/045831; this version posted June 17, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

type 2 diabetes, all of these represented replications of established associations (Table S3). Among the most strongly associated loci include *APOE* (rs429358, $P = 9.72 \times 10^{-195}$) for Alzheimer's disease [Corder et al., 1993], *LPA* (rs10455872, $P = 2.55 \times 10^{-25}$) and *CDKN2A/CDNKN2B* (rs4007642, $P = 7.64 \times 10^{-21}$) for coronary artery disease [Danesh et al., 2000; The Wellcome Trust Case Control Consortium et al., 2007], *FES/FURIN* (rs8027450, $P = 6.12 \times 10^{-13}$) for high blood pressure/hypertension [The International Consortium for Blood Pressure Genome-Wide Association Studies, 2011], *FGFR2* (rs2981583, $P = 3.62 \times 10^{-12}$) for breast cancer [Hunter et al., 2007], *TCF7L2* (rs34872471, $P = 7.76 \times 10^{-45}$) for type 2 diabetes [Grant et al., 2006], and *CHRNA5/CHRNA3* (rs5813926, $P = 1.67 \times 10^{-9}$ for lung cancer [Hung et al., 2008]. We identified two genome-wide significant loci for Parkinson's disease, one of which corresponds to the established *ASH1L* locus (rs35777901, $P = 2.25 \times 10^{-8}$) [Nalls et al., 2014]. The second locus at *SLIT3* (rs1806840, $P = 6.39 \times 10^{-9}$) is implicated in Parkinson's disease risk at genome-wide significance for the first time, although this SNP is reported as "non-significant" ($P > 0.05$, see URLs) in Nalls et al. [2014]. The locus remains genome-wide significant (rs1806840, $P = 5.90 \times 10^{-9}$) when running a linear mixed model association, suggesting that the signal is unlikely to be driven by cryptic population structure (Supplementary Note). Future genetic studies of Parkinson's disease will be needed to determine whether *SLIT3* is a true risk locus.

Effect size comparisons

In principle, the adjusted odds ratios obtained from a proxy-case-control design might differ from those obtained from a standard case-control design for a number of reasons. First, non-additive effects will distort these ORs in different ways in the two study designs. For example, under an additive model and Hardy-Weinberg equilibrium, the allelic odds ratio ($OR_{allelic}$, estimated from allele counts) is equivalent to the heterozygote odds ratio (OR_{het}, estimated from genotype counts), and the homozygote odds ratio (OR_{hom}) is simply OR_{het}^2 [Sasieni, 1997]. When the risk-increasing allele is partially recessive, then $OR_{hom} > OR_{het}^2$. In this case, if additivity is assumed, then the observed $OR_{allelic}$ will be inflated by recessive effects, such that $OR_{allelic} > OR_{het}$ [Sasieni, 1997]. As such, the adjusted OR from GWAX (which is equivalent to OR_{het} under additivity) will underestimate the observed $OR_{allelic}$ from a case-control design.

Similarly, errors made by offspring in recalling the diseases of their parents would bias our estimates, as would direct causal effects of an offspring's genotype on a parental phenotype (if, for example, a partially-heritable childhood behavior influences the diseases of their parents). Indeed, across 11 of the 12 phenotypes, females were significantly more likely to report a first-degree relative with the disease than males (Table S2), indicating at least some recall bias. Phenotype misclassification will also bias the effect size estimates. For instance, UK Biobank participants were asked whether their parents/siblings were diagnosed with "diabetes", without any distinction between type 1 and type 2 diabetes. Given the population prevalences of type 1 and type 2 diabetes, we would expect over 90% of the proxy-cases to be type 2 diabetes. As such, we refer to this group as type 2 diabetes throughout this study.

bioRxiv preprint doi: https://doi.org/10.1101/045831; this version posted June 17, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Differences between the adjusted ORs and those previously reported may also reflect inherent differences in the samples that are collected as part of a case-control study versus a population cohort. Our additive model assumes that the frequency of a risk allele in individuals with two affected parents is the same as that in the population of cases generally. To test whether this is the case, we constructed polygenic risk scores [The International Schizophrenia Consortium, 2009] in the UK Biobank samples using previously reported ORs at established risk loci for Alzheimer's disease [Lambert et al., 2013], coronary artery disease [The CARDIoGRAMplusC4D Consortium, 2015] and type 2 diabetes [DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium et al., 2014]. Dividing the UK Biobank individuals into those affected with disease and those unaffected with two affected parents, we found significant differences in the mean polygenic risk scores between the two groups across all three disorders ($P < 0.003$, Figure S3). These results may reflect non-additive effects, or alternatively may represent a true difference in polygenic risk between the two groups. That is, given that these disorders generally occur later in life, cases ascertained as part of a case-control study (or UK Biobank participants who are aged under 69) may represent a more extreme version of the disease, harboring a greater burden of risk variants than cases that are truly sampled randomly from the general population.

To test the extent of these biases, we obtained summary association statistics from previously published GWAS for four phenotypes: Alzheimer's disease [Lambert et al., 2013], coronary artery disease [The CARDIoGRAMplusC4D Consortium, 2015], major depressive disorder [Ripke et al., 2013] and type 2 diabetes [DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium et al., 2014]. Across established loci for three of these diseases (no genome-wide significant loci were reported for major depressive disorder), the direction and relative size of effects were consistent between our adjusted ORs and those reported previously ($0.92 <$ Pearson's $r < 0.97$), though the adjusted ORs were slightly underestimated, with regression slopes between 0.66 and 0.92 (Figure 3). We observed significant ($P < 0.01$) genetic correlations (r_g, the proportion of variance in disease liability that is shared between two phenotypes) between our GWAX results and the published GWAS summary statistics for coronary artery disease ($r_g = 0.93$), major depressive disorder ($r_g = 0.67$), type 2 diabetes ($r_g = 0.91$) and Alzheimer's disease ($r_g = 0.44$).

Meta analysis

Motivated by these consistent odds ratios, and in an effort to identify additional risk loci, we performed fixed-effects meta-analysis combining our proxy-case-control association summary statistics with those from the previously published GWAS. This approach implicated 17 novel risk loci at genome-wide significance associated with Alzheimer's disease, coronary artery disease and type 2 diabetes (Table 1, Figure 4, Figure S4-S6).

Among the novel loci for Alzheimer's disease include genes invovled in immune surveillance (*SPPL2A*, signal peptide peptidase like 2A) and major histocompatability complex class II signal transduction (*SCIMP*, SLP adaptor and SCK interacting membrane protein) [Friedmann et al., 2004], further highlighting the role

bioRxiv preprint doi: https://doi.org/10.1101/045831; this version posted June 17, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Phenotype	SNP	Chr	Position (GRCh37)	Eff/Alt allele	Freq	OR (adj.)	P-value	Nearby genes
Alzheimer's disease	rs2074612	5	139,714,690	T/C	0.438	1.08	8.00×10^{-9}	*HBEGF*
	rs7920721	10	11,720,308	G/A	0.380	1.07	4.27×10^{-8}	*ECHDC3*
	rs59685680	15	51,001,534	G/T	0.198	0.92	7.32×10^{-9}	*SPPL2A, TRPM7, USP50*
	rs77493189	17	5,118,951	G/T	0.123	1.11	9.60×10^{-10}	*SCIMP, ZNF594, RABEP1, USP6*
Coronary artery disease	rs61776719	1	38,461,319	C/A	0.446	0.95	1.63×10^{-8}	*SF3A3, FHL3*
	rs4685219	3	14,894,188	A/G	0.376	1.05	7.33×10^{-9}	*FGD5, NR2C2*
	rs11723436	4	120,901,336	G/A	0.313	1.05	4.77×10^{-8}	*MAD2L1*
	rs13109172	4	146,759,483	C/A	0.355	0.95	4.16×10^{-8}	*ZNF827*
	rs869396	4	169,688,000	A/C	0.467	0.96	4.09×10^{-8}	*PALLD*
	rs17680741	10	82,251,514	C/T	0.288	0.95	1.22×10^{-8}	*TSPAN14, SH2D4B, FAM213A*
	rs7500448	16	83,045,790	G/A	0.252	0.95	4.09×10^{-8}	*CDH13*
	rs15052	19	41,813,375	C/T	0.176	1.08	1.82×10^{-9}	*HNRNPUL1, TGFB1, CCDC97, AXL*
Type 2 diabetes	rs1291041	5	78,443,735	T/G	0.352	0.94	2.26×10^{-8}	*BHMT*
	rs744103	6	43,805,362	T/A	0.312	1.07	3.32×10^{-8}	*VEGFA*
	rs4273712	6	126,964,510	G/A	0.266	1.08	8.38×10^{-10}	-
	rs301047	13	26,788,114	G/A	0.186	1.08	1.56×10^{-8}	*RNF6*
	rs11658220	17	65,646,092	A/G	0.103	1.14	4.13×10^{-9}	*PITPNC1*

Table 1. Novel genome-wide significant risk loci identified through proxy-case-control analysis and meta-analysis with published genome-wide association studies of Alzheimer's disease, coronary artery disease and type 2 diabetes.

bioRxiv preprint doi: https://doi.org/10.1101/045831; this version posted June 17, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

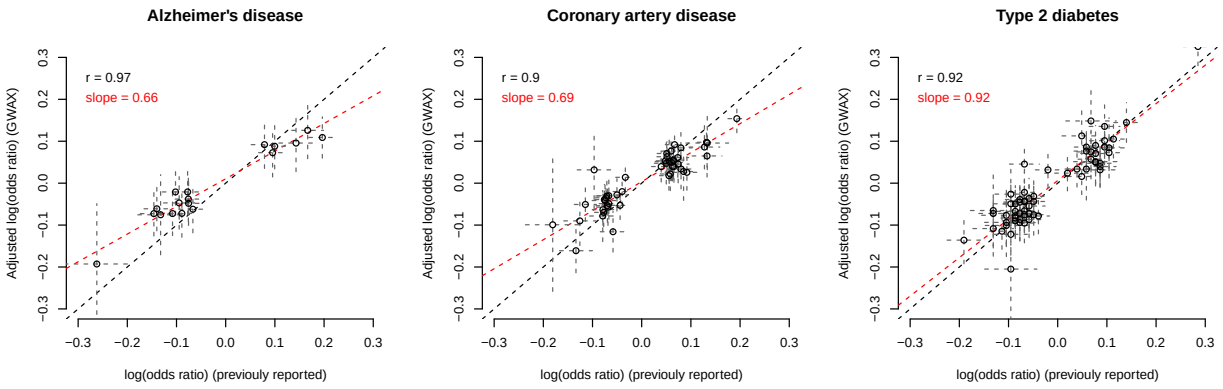

Figure 3. Comparison of adjusted ORs and previously reported case-control ORs at established risk loci for three diseases with publicly available summary statistics. Each point represents a previously reported risk variant and its corresponding effect size. The dashed gray lines are 95% confidence intervals. The dashed red line (and corresponding slope) is the fitted line from least squares regression. The dashed black line is is $y = x$. Reported effect sizes and list of established risk loci were obtained from - Alzheimer's disease: Lambert et al. [2013]; coronary artery disease: The CARDIoGRAMplusC4D Consortium [2015]; type 2 diabetes: DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium et al. [2014].

of the innate immune system in Alzheimer's disease etiology [Chan et al., 2015; Gjoneska et al., 2015]. For coronary artery disease, one novel locus resides in an intron of *FGD5* (FYVE, RohGEF and PH domain containing 5), a member of the FGD family of guanine nucleotide exchange factors. *FGD5* has been shown to regulate *VEGFA* (vascular endothelial growth factor) [Kurogane et al., 2012], a key cytokine in the formation of new vessels and potential therapeutic target for heart disease [Taimeh et al., 2013]. For type 2 diabetes, we identified a novel locus in *PITPNC1* (phosphatidylinositol transfer protein, cytoplasmic 1), a member of the phosphatidylinositol transfer protein family and has been show to be involved in lipid transport between membrane compartments [Garner et al., 2012].

To further illustrate the utility of the GWAX approach, we also performed case-control GWAS in the UK Biobank (taking case status from medical records) for coronary artery disease (5,685 cases, 109,347 controls) and type 2 diabetes (2,463 cases, 112,273 controls), the results of which were combined with the previously published summary GWAS statistics described above. Of the eight novel coronary artery disease risk loci identified in the GWAX meta-analysis, only two were genome-wide significant in the GWAS meta-analysis. Similarly, none of the five novel type 2 diabetes loci exceeded genome-wide significance in the GWAS meta-analysis (Table S4). These results further demonstrate that using proxy-cases is more powerful than cases for identifying risk loci in population cohorts.

Discussion

This study demonstrates proof of principle that complex disease risk loci can be identified using the genotypes of unaffected individuals and the phenotypes of their affected relatives. We applied the GWAX ap-

bioRxiv preprint doi: https://doi.org/10.1101/045831; this version posted June 17, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

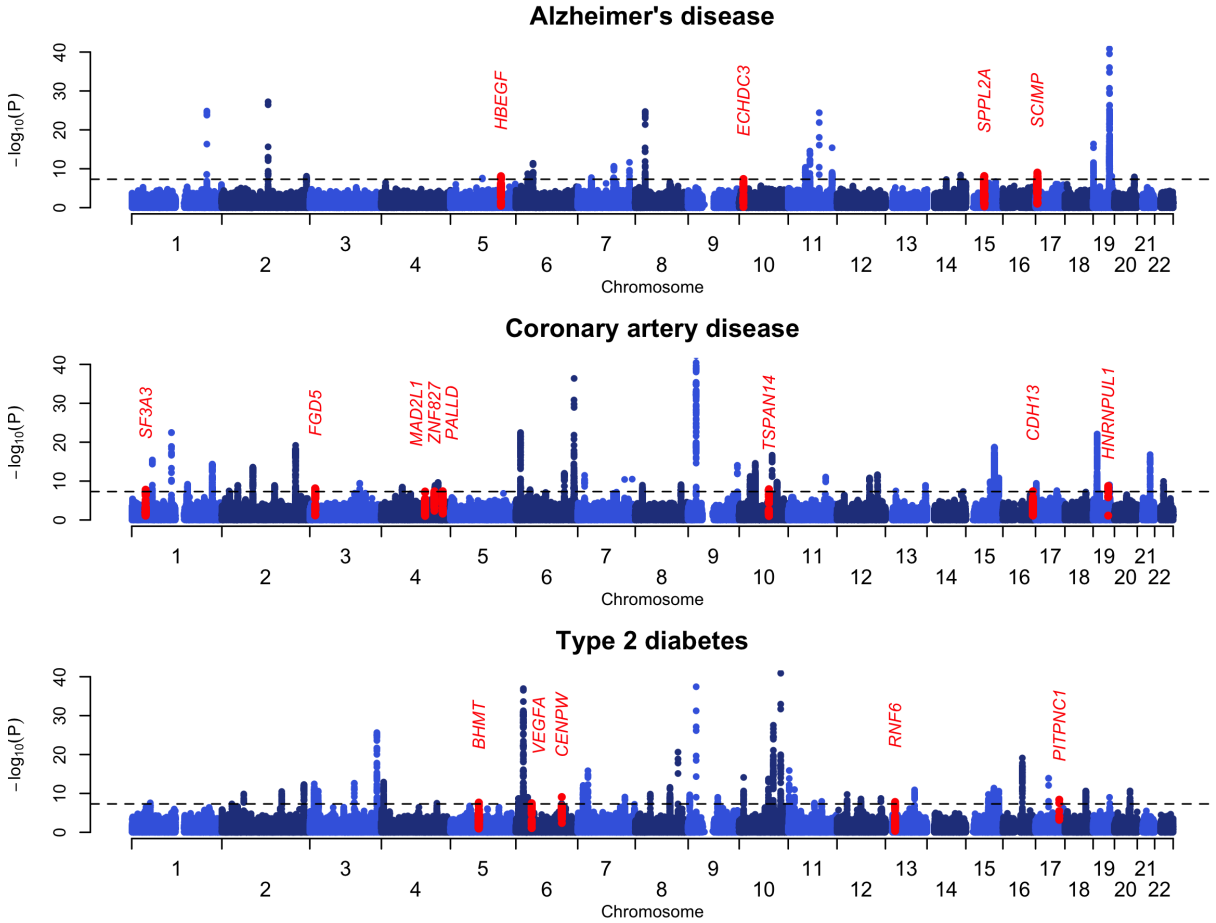

Figure 4. Manhattan plots of fixed-effects meta-analysis results for Alzheimer's disease, coronary artery disease and type 2 diabetes. Chromosome and positions are plotted on the x-axis. Strength of association is plotted on the y-axis. Novel risk loci are indicated in red. The dashed horizontal line indicates the genome-wide significant threshold of $P < 5 \times 10^{-8}$. $-\log_{10}$ P-values are truncated at 40 for illustrative purposes.

bioRxiv preprint doi: https://doi.org/10.1101/045831; this version posted June 17, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

proach to 12 common diseases in 116,196 individuals from the UK Biobank and combined our results with publicly available GWAS summary statistics for four of these diseases. We replicated known risk loci and identified 17 novel risk loci at genome-wide significance associated with Alzheimer's disease, coronary artery disease and type 2 diabetes.

Large population cohorts such as the UK Biobank and NIH Precision Medicine Initiative along with participant-driven projects [Dolgin, 2010; Eriksson et al., 2010] are emerging as valuable resources in biomedical research. By performing association mapping using the family members of affected individuals, we partly overcome the ascertainment limitations inherent in these studies.

Future expansions of these approaches may take into account more distant relatives in a formal way, allowing for the phenotypes of all known relatives to be accounted for and analyzed in conjunction with directly genotyped individuals. Genetic studies of complex disorders may progress beyond simple "case" and "control" phenotypes, and instead leverage multiple layers of information into a direct estimate of disease liability. Large crowd-sourced family trees [Ledford, 2013] along with reported phenotypes, demographics, lifestyle surveys, medical records and epidemiological information can be combined to provide robust estimates of both the genetic and environmental components of disease liability [Campbell et al., 2010]. Using liability as a phenotype can also account for ascertainment biases of case-control studies [Hayeck et al., 2015; Weissbrod et al., 2015], and allow for much greater power to identify disease susceptibility variants.

Methods

Power calculations

We performed power calculations comparing a study design using true cases and controls to one with proxy-cases and controls, and estimated the sample sizes of each such that power to detect association is equivalent. We also considered the situation where both cases and proxy-cases are available in the context of a population cohort study, where the expected number of cases and proxy-cases depends on disease prevalence and heritability on the liability scale. Details of the power calculations are described in the Supplementary Note.

UK Biobank data collection

The UK (United Kingdom) Biobank is a large population-based study of over 500,000 subjects aged 40-69 years recruited from 2006-2010 [Sudlow et al., 2015]. Participants entered information about their family history of disease by answering three questions: 1) "Has/did your father ever suffer from?", 2) "Has/did your mother ever suffer from?", and 3) "Have any of your brothers or sisters suffered from any of the following diseases?". Participants were asked to choose among 12 conditions (heart disease, stroke, high blood pressure, chronic bronchitis/emphysema, Alzheimer's disease/dementia, diabetes, Parkinson's disease, severe depression, lung cancer, bowel cancer, prostate cancer and breast cancer) and were allowed to select more than one condition. Participants were also given the choice of entering "Do not know", "Prefer not to

bioRxiv preprint doi: https://doi.org/10.1101/045831; this version posted June 17, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

answer" or "None of the above". Throughout this manuscript, we denote heart disease, severe depression and diabetes to refer specifically to coronary artery disease, major depressive disorder and type 2 diabetes, respectively. Case/control statuses for the participants themselves were available via health records (ICD10 diagnoses, Table S5). The UK Biobank received ethics approval from the National Health Service National Research Ethics Service (Ref 11/NW/0382).

Effective sample size comparisons

For each of the 12 phenotypes, we converted the observed number of cases (or proxy-cases) and controls into an effective sample size (N_{eff}). The effective sample size is the total sample size where there is an equal number of cases (or proxy-cases) and controls that gives the equivalent power to detect association as the observed unequal sample size. The test statistic for a standard 2×2 1-df chi-square test when the number of cases and controls differ is:

$$\chi^2_{unbalanced} = \frac{(f_A - f_U)^2}{(1/N_A + 1/N_U)(f(1-f))}, \tag{1}$$

where N_A is the number of cases (or proxy-cases), N_U the number of controls, f_A is the allele frequency in cases (or proxy-cases), f_U the allele frequency in controls and f is the overall allele frequency. Under a balanced design where $N_A = N_U = \frac{N_{eff}}{2}$, the test statistic becomes:

$$\chi^2_{balanced} = \frac{(f_A - f_U)^2}{(2/N_{eff} + 2/N_{eff})f(1-f)}. \tag{2}$$

Setting $\chi^2_{unblanaced} = \chi^2_{balanced}$ and solving for N_{eff}, we have the effective sample size as a function of observed number of cases (or proxy-cases) and controls:

$$N_{eff} = \frac{4}{1/N_A + 1/N_U}. \tag{3}$$

When we report the effective sample size in proxy-cases and controls, we divide N_{eff} by four to account for power, enabling a direct comparison with the effective sample size when using cases and controls (Supplementary Note).

Genotyping, imputation and quality control

The UK Biobank May 2015 Interim Data Release included directly genotyped and imputed data for 152,529 individuals. Around 90% of individuals were genotyped on the Affymetrix UK Biobank Axiom array, while the remainder were genotyped on the Affymetrix UK BiLEVE array. The two platforms are similar with $> 95\%$ common marker content (847,441 markers in total). Markers were selected on the basis of known associations with phenotypes, coding variants across a range of minor allele frequencies, and content to provide good genome-wide imputation coverage in European populations for variants with minor allele

bioRxiv preprint doi: https://doi.org/10.1101/045831; this version posted June 17, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

frequencies $> 1\%$. Genotyped individuals were phased using SHAPEIT2 [Delaneau et al., 2014] and then imputed with the IMPUTE2 [Howie et al., 2009] algorithm using a reference panel consisting of 12,570 haplotypes from a combined UK10K [The UK10K Consortium, 2015] and 1000 Genomes Phase 3 dataset [The 1000 Genomes Project Consortium, 2015]. In total, 73,355,667 polymorphic variants were successfully imputed. Additional information on the genotyping array, sample preparation and quality control can be found in the documents the URLs section. After QC, We took forward 116,196 unrelated individuals of European descent for analysis.

Genome-wide association by proxy

The genome-wide association by proxy study design is an extreme version of approaches that try to impute unknown genotypes in phenotyped individuals based on the genotypes of close relatives, though these approaches either require accurate pedigree information [Gudbjartsson et al., 2008] and/or sparse genotypes (e.g. microsatellites) on which to impute [Burdick et al., 2006]. Our approach is also similar to that of MQLS [Thornton and McPeek, 2007], a method for association testing in related individuals that allows for combinations of known and unknown phenotypes and genotypes. Indeed, when the genotyped individuals are all of unknown phenotype but with the phenotype of one parent available, MQLS and our approach (using Pearson's χ^2 test) are mathematically equivalent (Johanna Jakobsdottir, personal communication). However, the current implementation of MQLS does not allow for situations where all genotyped individuals have unknown phenotypes, does not scale to large cohorts and genome-wide data, and cannot handle covariates like principal components to account for population structure. By contrast, using standard logistic regression scales easily to large datasets and can handle covariates without issues.

To perform GWAX in the UK Biobank, for each of the 12 common diseases, subjects were considered proxy-cases if they have at least one affected mother, father or sibling. Subjects who answered "Do not know", or "Prefer not to answer" were removed from the analysis. All other subjects were considered controls. The total number of proxy-cases and controls for each phenotype are listed in Table S1.

Association between genotype and phenotype was performed on best-guess imputed genotypes (allelic likelihood > 0.9, missingness $< 10\%$, minor allele frequency > 0.005) using logistic regression in PLINK2 [Chang et al., 2015]. For all analyses, we included the subjects' reported sex, age at recruitment and the first four principal components (estimated directly from the post-QC set of UK Biobank individuals) as covariates. We observed modest genomic inflation across the 12 diseases ($1.05 < \lambda < 1.07$; Figure S4).

To test whether these inflation is due to population stratification or reflect a true polygenic signal, we performed LD score regression on the summary association statistics using a set of 1.2 million common SNPs from HapMap3 [Bulik-Sullivan et al., 2015b]. The LD Score regression intercepts were between 0.99 and 1.02 (Figure S7), suggesting that the inflation is due to a true polygenic signal. For the 24 lead SNPs identified with $P < 5 \times 10^{-8}$, 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

bioRxiv preprint doi: https://doi.org/10.1101/045831; this version posted June 17, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

623,852 directly genotyped SNPs. P-values were very similar with those from the logistic regression using 4PCs (Table S3). Together, these results suggest that the effects of population stratification were minimal. As such, we did not adjust association statistics using genomic control.

To enable direct comparison of our effect sizes to those from traditional case-control designs (as well as enabling fixed-effect meta-analysis), we calculate odds ratios using the following approximation. For each SNP, let f_A and f_U be the allele frequencies in true-cases and controls respectively, and

$$OR = \frac{f_A(1-f_U)}{f_U(1-f_A)} \tag{4}$$

be the true case-control odds ratio. If f_P is the allele frequency in proxy-cases (the vast majority of whom have only one first-degree relative affected with disease), then

$$f_P = \frac{f_U + f_A}{2}. \tag{5}$$

In order to estimate the adjusted odds ratio as a function of the observed allele frequencies in proxy-cases and controls, we substitute f_A into (1):

$$\hat{OR} = \frac{(2f_P - f_U)(1-f_U)}{f_U(1-2f_P+f_U)}. \tag{6}$$

For the range of ORs (< 1.4) typically reported in a GWAS, the log of the adjusted odds ratio derived here is approximately double that of the log odds ratio directly estimated from logistic regression using proxy-cases and controls (Figure S1). As the odds ratios and standard errors from logistic regression take into account covariates, we report adjusted log odds ratios using this doubling approximation rather than directly estimating them from allele frequencies using equation (6). The corresponding adjusted standard error is also double the standard error of the log odds ratio from logistic regression, since $se^2 = Var(2\beta) = 2^2 Var(\beta)$.

Polygenic risk scores

Publicly available GWAS summary association statistics were obtained for Alzheimer's disease (17,008 cases and 37,154 controls for stage 1 SNPs; plus 8,572 cases and 11,312 controls for 11,632 stage 2 SNPs) [Lambert et al., 2013], coronary artery disease (60,801 cases and 123,504 controls) [The CARDIoGRAMplusC4D Consortium, 2015], major depressive disorder (9,249 cases and 9,519 controls) [Ripke et al., 2013] and type 2 diabetes (26,488 cases and 83,964 controls) [DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium et al., 2014].

From these summary statistics, we extracted the reported effect sizes at established loci for Alzheimer's disease (20 SNPs), coronary artery disease (55 SNPs) and type 2 diabetes (71 SNPs) and constructed polygenic risk scores for each individual in the UK Biobank. No genome-wide significant risk loci were reported for major depression. For a disease with m associated SNPs, the polygenic risk score for individual i is:

bioRxiv preprint doi: https://doi.org/10.1101/045831; this version posted June 17, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$S_i = \sum_{j=1}^{m} \hat{\beta}_j g_{ij} \tag{7}$$

where $\hat{\beta}_j$ is the reported effect size (log(OR)) of the reference allele of SNP j from the previous GWAS, and g_{ij} is the allele count of the reference allele for individual i at SNP j. Scores were normalized to mean = 0 and variance = 1. The means of the normalized polygenic risk scores were calculated for groups of individuals who were 1) affected with the disease, 2) unaffected with two affected parents and 3) unaffected with one affected parent. For this last group, the mean risk score was doubled so that it is (in theory) equivalent to the risk score for unaffected individuals with two affected parents. We tested for a significant difference in the mean risk scores for each pair of groups using Welch's t-test.

Genetic correlation

For each of the four phenotypes, we estimated the genetic correlation between our GWAX summary statistics and the published GWAS summary statistics using LD score regression with a set of ~ 1.2 million common SNPs from HapMap3 [Bulik-Sullivan et al., 2015a].

Meta-analysis

Fixed-effects meta-analysis was performed for Alzheimer's disease, coronary artery disease, major depressive disorder and type 2 diabetes using inverse variance-weighted method for all SNPs that overlap between the publicly available summary statistics and our adjusted odds ratio GWAX results. That is, for each SNP with estimated log odds ratios and standard errors, $\hat{\beta}_i$ and $\hat{se}_i$ respectively, where $i = 1$ or 2 correponding to the GWAX (adjusted log odds ratio) or GWAS results, the combined effect size is:

$$\hat{\beta}_{meta} = \frac{\sum_i \hat{\beta}_i w_i}{\sum_i w_i} \tag{8}$$

with corresponding standard error and P-value:

$$\hat{se}_{meta} = \sqrt{1/\sum_i w_i} \tag{9}$$

$$P_{meta} = 2\Phi(|\hat{\beta}_{meta}|/\hat{se}_{meta}) \tag{10}$$

where $w_i = 1/\hat{se}_i^2$ and Φ is the cumulative standard normal distribution.

Identification of independent risk loci

A locus was considered genome-wide significant if it includes a SNP with association $P < 5 \times 10^{-8}$. For both the primary proxy-case-control analysis in UK Biobank individuals and meta analyses, independent risk

bioRxiv preprint doi: https://doi.org/10.1101/045831; this version posted June 17, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

loci were identified using the approximate conditional and joint association method implemented in GCTA (GCTA-COJO). The method performs approximate step-wise conditional association testing using summary association statistics and LD structure from a set of reference genotypes. As such, the SNPs selected from this procedure can be thought to represent the strongest independent signals associated with the phenotype. We ran GCTA-COJO with settings $r^2 > 0.9$ and $P < 5 \times 10^{-8}$, and a reference panel consisting of 2,500 randomly selected individuals from the UK Biobank cohort. [Yang et al., 2012].

URLs

UK Biobank - `http://www.ukbiobank.ac.uk`

Genotyping and quality control of UK Biobank, a large-scale, extensively phenotyped prospective resource - `http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/UKBiobank_genotyping_QC_documentation-web.pdf`

Genotype imputation and genetic association studies of UK Biobank - `http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/imputation_documentation_May2015.pdf`

Parkinson's disease GWAS summary statistics from Nalls et al. [2014] - `http://pdgene.org/view?study=1`

References

Barzilai, N., Atzmon, G., Schechter, C., Schaefer, E. J., Cupples, A. L., Lipton, R., Cheng, S., and Shuldiner, A. R., 2003. Unique lipoprotein phenotype and genotype associated with exceptional longevity. *JAMA*, **290**(15):2030–40.

Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., Consortium, R., Consortium, P. G., for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3, G. C., Duncan, L., *et al.*, 2015a. An atlas of genetic correlations across human diseases and traits. *Nat Genet*, **47**(11):1236–1241.

Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., of the Psychiatric Genomics Consortium, S. W. G., Patterson, N., Daly, M. J., Price, A. L., and Neale, B. M., *et al.*, 2015b. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet*, **47**(3):291–295.

Burdick, J. T., Chen, W.-M., Abecasis, G. R., and Cheung, V. G., 2006. In silico method for inferring genotypes in pedigrees. *Nat Genet*, **38**(9):1002–1004.

Campbell, D. D., Sham, P. C., Knight, J., Wickham, H., and Landau, S., 2010. Software for generating

bioRxiv preprint doi: https://doi.org/10.1101/045831; this version posted June 17, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

liability distributions for pedigrees conditional on their observed disease states and covariates. *Genetic Epidemiology*, **34**(2):159–170.

Chan, G., White, C. C., Winn, P. A., Cimpean, M., Replogle, J. M., Glick, L. R., Cuerdon, N. E., Ryan, K. J., Johnson, K. A., Schneider, J. A., *et al.*, 2015. CD33 modulates TREM2: convergence of Alzheimer loci. *Nature Neuroscience*, **18**(11):1556–1558.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J., 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, **4**(1):1–16.

Corder, E. H., Saunders, A. M., Strittmatter, W. J., Schmechel, D. E., Gaskell, P. C., Small, G. W., Roses, A. D., Haines, J. L., and Pericak-Vance, M. A., 1993. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science*, **261**(5123):921–923.

Danesh, J., Collins, R., and Peto, R., 2000. Lipoprotein(a) and Coronary Heart Disease Meta-Analysis of Prospective Studies. *Circulation*, **102**(10):1082–1085.

de Lau, L. M. and Breteler, M. M., 2006. Epidemiology of Parkinson's disease. *The Lancet Neurology*, **5**(6):525 – 535.

Delaneau, O., Marchini, J., and Consortium, T. . G. P., 2014. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nat Commun*, **5**.

DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Asian Genetic Epidemiology Network Type 2 Diabetes (AGEN-T2D) Consortium, South Asian Type 2 Diabetes (SAT2D) Consortium, Mexican American Type 2 Diabetes (MAT2D) Consortium, Type 2 Diabetes Genetic Exploration by Next-generation sequencing in multi-Ethnic Samples (T2D-GENES) Consortium, Mahajan, A., Go, M. J., Zhang, W., Below, J. E., Gaulton, K. J., *et al.*, 2014. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature Genetics*, **46**(3):234–244.

Dolgin, E., 2010. Personalized investigation. *Nat Med*, **16**(9):953–955.

Eriksson, N., Macpherson, J. M., Tung, J. Y., Hon, L. S., Naughton, B., Saxonov, S., Avey, L., Wojcicki, A., Pe'er, I., and Mountain, J., *et al.*, 2010. Web-Based, Participant-Driven Studies Yield Novel Genetic Associations for Common Traits. *PLoS Genet*, **6**(6):e1000993.

Friedmann, E., Lemberg, M. K., Weihofen, A., Dev, K. K., Dengler, U., Rovelli, G., and Martoglio, B., 2004. Consensus Analysis of Signal Peptide Peptidase and Homologous Human Aspartic Proteases Reveals Opposite Topology of Catalytic Domains Compared with Presenilins. *Journal of Biological Chemistry*, **279**(49):50790–50798.

bioRxiv preprint doi: https://doi.org/10.1101/045831; this version posted June 17, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Garner, K., Hunt, A. N., Koster, G., Somerharju, P., Groves, E., Li, M., Raghu, P., Holic, R., and Cockcroft, S., 2012. Phosphatidylinositol Transfer Protein, Cytoplasmic 1 (PITPNC1) Binds and Transfers Phosphatidic Acid. *Journal of Biological Chemistry*, **287**(38):32263–32276.

Gjoneska, E., Pfenning, A. R., Mathys, H., Quon, G., Kundaje, A., Tsai, L.-H., and Kellis, M., 2015. Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer/'s disease. *Nature*, **518**(7539):365–369.

Grant, S. F. A., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Manolescu, A., Sainz, J., Helgason, A., Stefansson, H., Emilsson, V., Helgadottir, A., *et al.*, 2006. Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat Genet*, **38**(3):320–323.

Gudbjartsson, D. F., Walters, G. B., Thorleifsson, G., Stefansson, H., Halldorsson, B. V., Zusmanovich, P., Sulem, P., Thorlacius, S., Gylfason, A., Steinberg, S., *et al.*, 2008. Many sequence variants affecting diversity of adult human height. *Nat Genet*, **40**(5):609–615.

Hayeck, T. J., Zaitlen, N. A., Loh, P.-R., Vilhjalmsson, B., Pollack, S., Gusev, A., Yang, J., Chen, G.-B., Goddard, M. E., Visscher, P. M., *et al.*, 2015. Mixed Model with Correction for Case-Control Ascertainment Increases Association Power. *The American Journal of Human Genetics*, **96**(5):720 – 730.

Hebert, L., Scherr, P., Bienias, J., Bennett, D., and Evans, D., 2003. Alzheimer disease in the us population: Prevalence estimates using the 2000 census. *Archives of Neurology*, **60**(8):1119–1122.

Howie, B. N., Donnelly, P., and Marchini, J., 2009. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genet*, **5**(6):1–15.

Hung, R. J., McKay, J. D., Gaborieau, V., Boffetta, P., Hashibe, M., Zaridze, D., Mukeria, A., Szeszenia-Dabrowska, N., Lissowska, J., Rudnai, P., *et al.*, 2008. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature*, **452**(7187):633–637.

Hunter, D. J., Kraft, P., Jacobs, K. B., Cox, D. G., Yeager, M., Hankinson, S. E., Wacholder, S., Wang, Z., Welch, R., Hutchinson, A., *et al.*, 2007. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet*, **39**(7):870–874.

Joshi, P. K., Fischer, K., Schraut, K. E., Campbell, H., Esko, T., and Wilson, J. F., 2016. Variants near CHRNA3/5 and APOE have age- and sex-related effects on human lifespan. *Nat Commun*, **7**:11174.

Kong, A., Steinthorsdottir, V., Masson, G., Thorleifsson, G., Sulem, P., Besenbacher, S., Jonasdottir, A., Sigurdsson, A., Kristinsson, K. T., Jonasdottir, A., *et al.*, 2009. Parental origin of sequence variants associated with complex diseases. *Nature*, **462**(7275):868–874.

bioRxiv preprint doi: https://doi.org/10.1101/045831; this version posted June 17, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Kurogane, Y., Miyata, M., Kubo, Y., Nagamatsu, Y., Kundu, R. K., Uemura, A., Ishida, T., Quertermous, T., Hirata, K.-i., and Rikitake, Y., *et al.*, 2012. FGD5 Mediates Proangiogenic Action of Vascular Endothelial Growth Factor in Human Vascular Endothelial Cells. *Arteriosclerosis, Thrombosis, and Vascular Biology*, **32**(4):988–996.

Lambert, J.-C., Ibrahim-Verbaas, C. A., Harold, D., Naj, A. C., Sims, R., Bellenguez, C., Jun, G., DeStefano, A. L., Bis, J. C., Beecham, G. W., *et al.*, 2013. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics*, **45**(12):1452–1458.

Ledford, H., 2013. Genome hacker uncovers largest-ever family tree.

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.

Pilling, L. C., Atkins, J. L., Bowman, K., Jones, S. E., Tyrrell, J., Beaumont, R. N., Ruth, K. S., Tuke, M. A., Yaghootkar, H., Wood, A. R., *et al.*, 2016. Human longevity is influenced by many genetic variants: evidence from 75,000 UK Biobank participants. *Aging (Albany NY)*, **8**(3):547–63.

Ripke, S., Wray, N. R., Lewis, C. M., Hamilton, S. P., Weissman, M. M., Breen, G., Byrne, E. M., Blackwood, D. H. R., Boomsma, D. I., Cichon, S., *et al.*, 2013. A mega-analysis of genome-wide association studies for major depressive disorder. *Molecular Psychiatry*, **18**(4):497–511.

Sasieni, P. D., 1997. From Genotypes to Genes: Doubling the Sample Size. *Biometrics*, **53**(4):1253–1261.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., *et al.*, 2015. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med*, **12**(3):e1001779.

Taimeh, Z., Loughran, J., Birks, E. J., and Bolli, R., 2013. Vascular endothelial growth factor in heart failure. *Nat Rev Cardiol*, **10**(9):519–530.

Tan, Q., Zhao, J. H., Li, S., Kruse, T. A., and Christensen, K., 2010. Power assessment for genetic association study of human longevity using offspring of long-lived subjects. *Eur J Epidemiol*, **25**(7):501–6.

The 1000 Genomes Project Consortium, 2015. A global reference for human genetic variation. *Nature*, **526**(7571):68–74.

bioRxiv preprint doi: https://doi.org/10.1101/045831; this version posted June 17, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The CARDIoGRAMplusC4D Consortium, 2015. A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics*, **47**(10):1121–1130.

The International Consortium for Blood Pressure Genome-Wide Association Studies, 2011. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*, **478**(7367):103–109.

The International Schizophrenia Consortium, 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, **460**(7256):748–752.

The UK10K Consortium, 2015. The UK10K project identifies rare variants in health and disease. *Nature*, **526**(7571):82–90.

The Wellcome Trust Case Control Consortium, Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D. P., McCarthy, M. I., Ouwehand, W. H., *et al.*, 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**(7145):661–678.

Thornton, T. and McPeek, M. S., 2007. Case-Control Association Testing with Related Individuals: A More Powerful Quasi-Likelihood Score Test. *The American Journal of Human Genetics*, **81**(2):321 – 337.

Weissbrod, O., Lippert, C., Geiger, D., and Heckerman, D., 2015. Accurate liability estimation improves power in ascertained case-control studies. *Nature Methods*, **12**(4):332–334.

Yang, J., Ferreira, T., Morris, A. P., Medland, S. E., Madden, P. A. F., Heath, A. C., Martin, N. G., Montgomery, G. W., Weedon, M. N., Loos, R. J., *et al.*, 2012. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet*, **44**(4):369–375.

Acknowledgements

JZL and JKP are partially supported by the National Institute of Mental Health (NIH grant R01MH106842). We thank Tristan Hayeck and Johanna Jakobsdottir for comments on a draft of this manuscript. This research has been conducted using the UK Biobank Resource. We would like to thank all UK Biobank participants and those involved in sample collection, genotyping, quality control and imputation.