nature genetics

Case–control association mapping by proxy using family history of disease

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

Collecting cases for case–control genetic association studies can be time-consuming and expensive. In some situations (such as studies of late-onset or rapidly lethal diseases), it may be more practical to identify family members of cases. In randomly ascertained cohorts, replacing cases with their first-degree relatives enables studies of diseases that are absent (or nearly absent) in the cohort. We refer to this approach as genome-wide association study by proxy (GWAX) and apply it to 12 common diseases in 116,196 individuals from the UK Biobank. Meta-analysis with published genome-wide association study summary statistics replicated established risk loci and yielded four newly associated loci for Alzheimer's disease, eight for coronary artery disease and five for type 2 diabetes. In addition to informing disease biology, our results demonstrate the utility of association mapping without directly observing cases. We anticipate that GWAX will prove useful in future genetic studies of complex traits in large population cohorts.

The case–control association study is a powerful method for identifying genetic variants that influence disease risk. In a typical study, a researcher genotypes a set of individuals who have a disease (the 'cases') and a set of individuals who 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 1 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 given 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 years (at the time of recruitment)[1]. 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 (for example, see ref. 2 or our crowdsourcing website, https://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 rarer) and aspects of the sampling.

In this paper, we consider a study design where the researcher identifies family members of cases, rather than cases themselves (as the cases may be difficult or impossible to contact). This design is analogous to genomic selection in animal breeding, where the phenotypes of relatives are regularly incorporated into calculating breeding values[3–5]. In human studies, this design has been discussed in the context of family-based association studies[6,7] and is popular in studies of longevity (where 'cases' are long-lived individuals[8–11]) 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[12–14] 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 25 March 2016, there were 55 cases of Alzheimer's disease listed among the approximately 500,000 individuals in the UK Biobank. However, over 60,000 participants noted that one or both of their parents was/is affected with the disease. An individual with a single affected parent can be thought to have one chromosome sampled from a population of 'cases' and one sampled 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 $(f_U + f_A)/2$. 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).

[1]New York Genome Center, New York, New York, USA. [2]Department of Computer Science, Columbia University, New York, New York, USA. [3]Department of Biological Sciences, Columbia University, New York, New York, USA. Correspondence should be addressed to J.Z.L. (jliu@nygenome.org).

Received 25 March 2016; accepted 14 December 2016; published online 16 january 2017; doi:10.1038/ng.3766

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 (Online Methods). We initially considered the case where we have no phenotype information about genotyped individuals themselves, although we consider this case later on. For all our simulations (except where noted otherwise), controls are defined as unaffected individuals who have zero affected first-degree relatives and we assume perfect knowledge about individual phenotypes and family history of disease.

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 the 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 spectra, 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 (**Fig. 1a**). This ratio increases to ~4.9 if 10% of controls are in fact misclassified proxy cases (**Supplementary Fig. 1** and **Supplementary Note**). For late-onset diseases such as Alzheimer's disease (prevalence of 1.6% in the population versus 42% in those over the age of 84 years)[15] and Parkinson's disease (0.3% in the population versus 4% in those over 80 years)[16], 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 who have a parent with the disease) and controls. We considered approaches that treat all three of these groups separately (in a 3 × 2 χ^2 test) or that lump together the proxy cases and true cases and perform a standard 2 × 2 χ^2 test. When both true and proxy cases are available in a population cohort study, accounting for this fact increases power (**Fig. 1b** and **Supplementary Note**). For instance, for a disease with 5% prevalence and 50% heritability on the liability scale across 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 of 0.1 in controls and an odds ratio of 1.2, there is 60% power at $\alpha = 5 \times 10^{-8}$ to detect association using a standard 2 × 2 χ^2 test of true cases versus controls, 87% power using a 2 × 2 test where cases and proxy cases are lumped together and 90% 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× when compared to a case–control design. Overall, the boost in effective sample size ranges from 1.36× to 1.28× for disease prevalences from 1% to 20%, respectively. 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 because there are no further gains in effective sample size. Nevertheless, across simulated effect sizes, allele frequencies, heritabilities and disease prevalences, the 3 × 2 test is consistently more powerful than the case–control test (see the **Supplementary Note** for details).

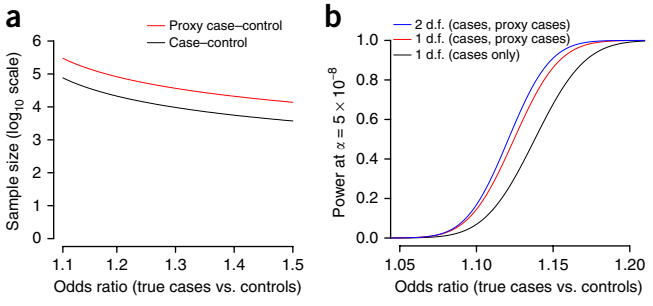

Figure 1 Power of proxy case–control association designs. (**a**) Total sample size required for 80% power to detect association at $\alpha = 5 \times 10^{-8}$ for case–control (black line) and proxy case–control (red line) designs at a SNP with 0.1 frequency in controls. (**b**) Power to detect association 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). Total sample size = 100,000, disease prevalence = 0.1, heritability of liability = 0.5 and allele frequency in controls = 0.1 (**Supplementary Note**). d.f., degrees of freedom.

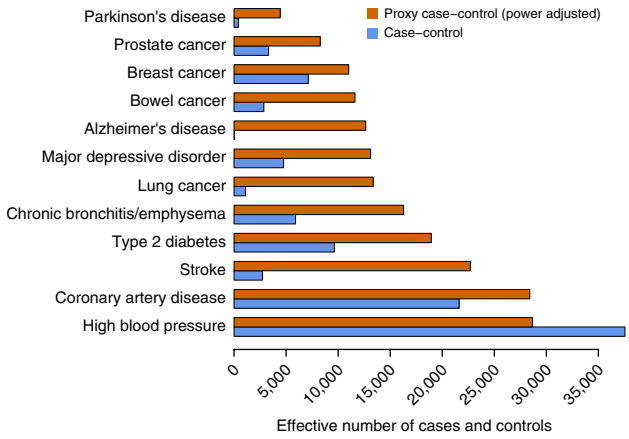

Figure 2 Effective sample sizes of case–control versus proxy case–control association designs in the UK Biobank. To adjust for power when considering proxy cases, we divided the effective sample size for proxy cases and controls by four. Case and control counts for breast cancer and prostate cancer only included females and males, respectively.

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 Project Phase 3 imputation (Online Methods), ~10.5 million low-frequency and common (minor allele frequency > 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 (although the medical records of the individuals themselves are available, we did not use them in this analysis 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 (**Supplementary Table 1**). On the basis of these sample sizes, we expect greater power to detect association using GWAX than a case–control GWAS for 11 of the 12 phenotypes (with high blood pressure being the exception) in the UK Biobank cohort (**Fig. 2**).

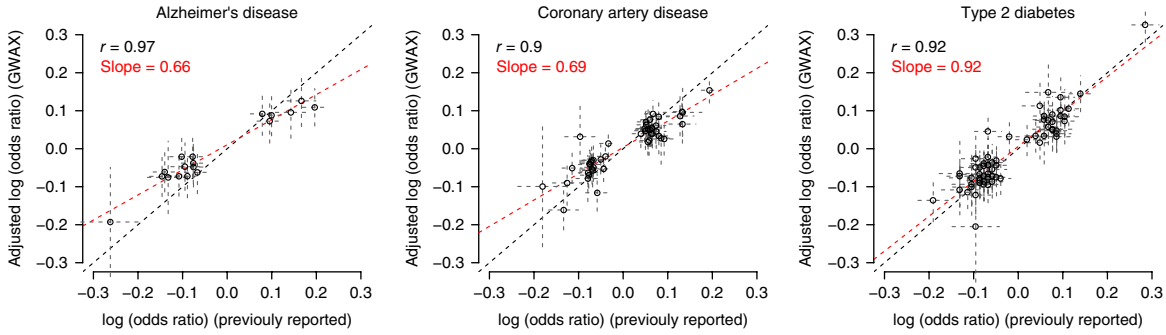

Figure 3 Comparison of adjusted odds ratios and previously reported case–control odds ratios at established risk loci for three diseases with publicly available summary statistics. Each point represents a previously reported risk variant and its corresponding effect size. The dashed gray lines are 95% confidence intervals. The dashed red line (and corresponding Pearson's *r* value and slope) is the fitted line from least-squares regression. The dashed black line corresponds to $y = x$. Reported effect sizes and the list of established risk loci were obtained from published GWAS for Alzheimer's disease[27], coronary artery disease[28] and type 2 diabetes[29].

We performed association testing using logistic regression with age, sex and the first four principal components as covariates. For each SNP, we calculated an adjusted odds ratio, which is directly comparable (under a standard additive model) with odds ratios estimated from traditional case–control study designs (Online Methods and **Supplementary Fig. 2**). The overall association results across the 12 phenotypes are shown in Manhattan plots, which have several clear peaks of association (**Supplementary Fig. 3**).

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 type 2 diabetes, all of these represented replications of established associations (**Supplementary Table 2**). Among the most strongly associated loci were *APOE* (rs429358, $P = 9.72 \times 10^{-195}$) for Alzheimer's disease[17], *LPA* (rs10455872, $P = 2.55 \times 10^{-25}$) and *CDKN2A–CDNKN2B* (rs4007642, $P = 7.64 \times 10^{-21}$) for coronary artery disease[18,19], *FES–FURIN* (rs8027450, $P = 6.12 \times 10^{-13}$) for high blood pressure/hypertension[20], *FGFR2* (rs2981583, $P = 3.62 \times 10^{-12}$) for breast cancer[21], *TCF7L2* (rs34872471, $P = 7.76 \times 10^{-45}$) for type 2 diabetes[22], and *CHRNA5–CHRNA3* (rs5813926, $P = 1.67 \times 10^{-9}$) for lung cancer[23]. We identified two genome-wide significant loci for Parkinson's disease, one of which corresponds to the established *ASH1L* locus (rs35777901, $P = 2.25 \times 10^{-8}$)[24]. 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.*[24]. 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 odds ratios 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 (ref. 25). 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 $OR_{allelic} > OR_{het}$ (ref. 25). As such, the adjusted odds ratio 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 the children's parents). Indeed, across 11 of the 12 phenotypes, females were significantly more likely to report a first-degree relative with the disease than males ($P < 0.036$; **Supplementary Table 3**), 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 or 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 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.

Differences between the adjusted odds ratios and previously reported odds ratios 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[26] in the UK Biobank samples using previously reported odds ratios at established risk loci for Alzheimer's disease[27], coronary artery disease[28] and type 2 diabetes[29]. Dividing the UK Biobank individuals into those affected with disease and those unaffected but with two affected parents, we found significantly lower polygenic risk scores for individuals with two affected parents than cases in all three disorders ($P < 0.003$; **Supplementary Fig. 4**). 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 under 69 years of age) 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[27], coronary artery disease[28], major depressive disorder[30] and type 2 diabetes[29]. Across established loci for three of these diseases (no genome-wide significant loci were reported for

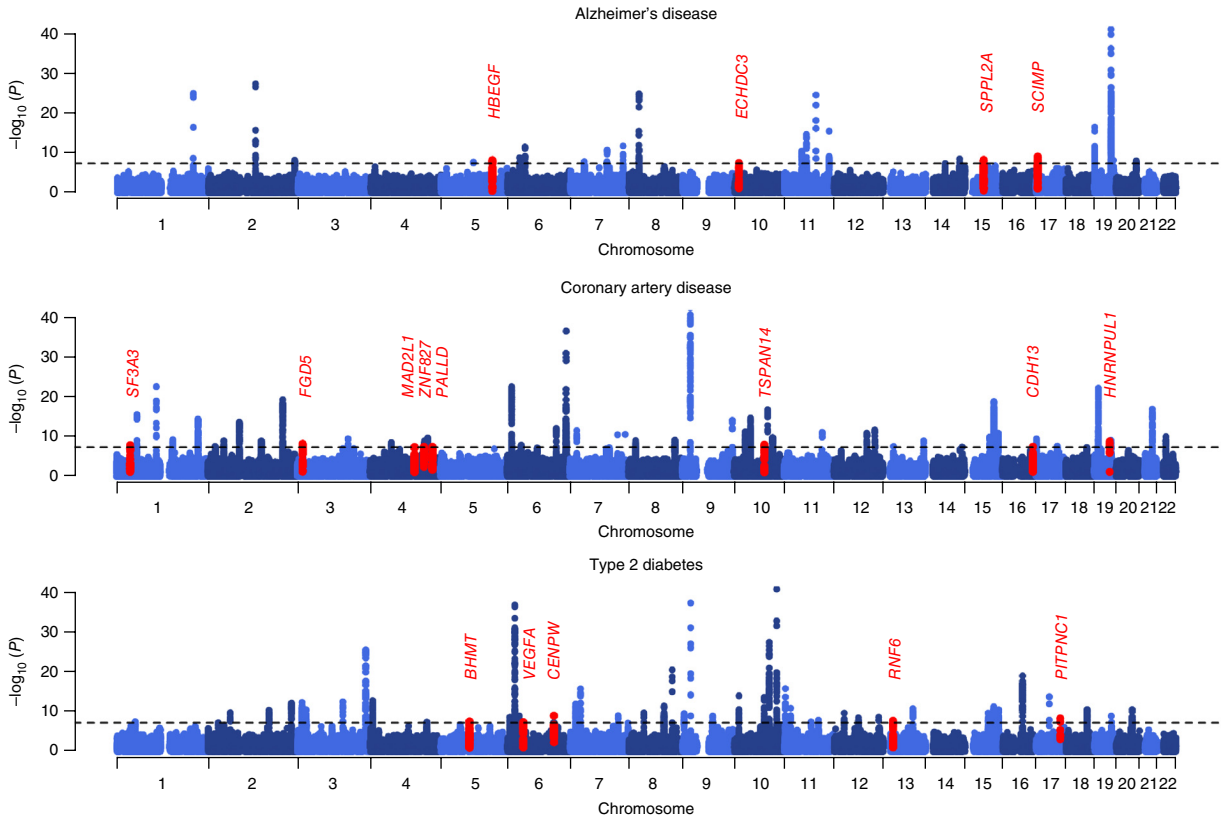

Figure 4 Manhattan plots of fixed-effects meta-analysis results for Alzheimer's disease, coronary artery disease and type 2 diabetes. Chromosome position is plotted on the *x* axis. Strength of association is plotted on the *y* axis. New risk loci are highlighted in red. The dashed horizontal line corresponds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

major depressive disorder), the direction and relative size of effects were consistent between our adjusted odds ratios and those reported previously (0.92 < Pearson's r < 0.97), although the adjusted odds ratios were slightly underestimated, with regression slopes between 0.66 and 0.92 (**Fig. 3**). We observed significant ($P < 0.01$) genetic correlations (r_g; the proportion of variance in disease liability that is shared by 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 new risk loci at genome-wide significance associated with Alzheimer's disease, coronary artery disease and type 2 diabetes (**Fig. 4**, **Table 1** and **Supplementary Figs. 5**–**7**).

Among the newly identified loci for Alzheimer's disease were genes involved in immune surveillance (*SPPL2A*; encoding signal peptide peptidase like 2A) and major histocompatibility complex class II signal transduction (*SCIMP*; encoding SLP-adaptor- and SCK-interacting membrane protein)[31], further highlighting the role of the innate immune system in Alzheimer's disease etiology[32,33]. For coronary artery disease, one newly identified 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)[34], a key cytokine in the formation of new vessels and a potential therapeutic target for heart disease[35]. For type 2 diabetes, we identified a new risk locus in *PITPNC1* (phosphatidylinositol transfer protein, cytoplasmic 1), a member of the phosphatidylinositol transfer protein family that has been shown to be involved in lipid transport between membrane compartments[36].

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 and 109,347 controls) and type 2 diabetes (2,463 cases and 112,273 controls), the results of which were combined with the previously published summary GWAS statistics described above. Of the eight new 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 new type 2 diabetes loci exceeded genome-wide significance in the GWAS meta-analysis (**Supplementary Table 4**). These results further demonstrate that using proxy cases is more powerful than using 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 approach 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

Table 1 New 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

SNP	Chr.[a]	Position (GRCh37)	EA/OA[b]	Freq.[c]	Previous GWAS			UK Biobank GWAX			Combined				Nearby gene(s)
					OR	95% CI[d]	*P*	OR	95% CI[d]	*P*	OR	95% CI[d]	*P*	P_{het}/I^2[e]	
Alzheimer's disease															
rs2074612	5	139,714,690	T/C	0.438	1.09	1.05–1.12	1.63×10^{-7}	1.07	1.01–1.13	1.25×10^{-2}	1.08	1.05–1.11	8.00×10^{-9}	0.64/0	*HBEGF*
rs7920721	10	11,720,308	G/A	0.38	1.07	1.04–1.10	3.08×10^{-7}	1.06	1.00–1.11	4.16×10^{-2}	1.07	1.04–1.10	4.27×10^{-8}	0.65/0	*ECHDC3*
rs59685680	15	51,001,534	G/T	0.198	0.92	0.89–0.95	4.30×10^{-7}	0.91	0.85–0.97	4.61×10^{-3}	0.92	0.89–0.95	7.32×10^{-9}	0.71/0	*SPPL2A*, *TRPM7*, *USP50*
rs77493189	17	5,118,951	G/T	0.123	1.10	1.06–1.14	5.01×10^{-7}	1.16	1.07–1.25	2.62×10^{-4}	1.11	1.07–1.15	9.60×10^{-10}	0.26/0.21	*SCIMP*, *ZNF594*, *RABEP1*, *USP6*
Coronary artery disease															
rs61776719	1	38,461,319	C/A	0.446	0.95	0.93–0.97	2.57×10^{-6}	0.95	0.92–0.98	1.74×10^{-3}	0.95	0.93–0.97	1.63×10^{-8}	0.92/0	*SF3A3*, *FHL3*
rs4685219	3	14,894,188	A/G	0.376	1.05	1.03–1.07	1.26×10^{-6}	1.06	1.02–1.10	1.30×10^{-3}	1.05	1.03–1.07	7.33×10^{-9}	0.57/0	*FGD5*, *NR2C2*
rs11723436	4	120,901,336	G/A	0.313	1.06	1.03–1.08	1.68×10^{-7}	1.03	1.00–1.07	6.76×10^{-2}	1.05	1.03–1.07	4.77×10^{-8}	0.34/0	*MAD2L1*
rs13109172	4	146,759,483	C/A	0.355	0.95	0.94–0.97	9.54×10^{-6}	0.94	0.91–0.98	9.70×10^{-4}	0.95	0.93–0.97	4.16×10^{-8}	0.53/0	*ZNF827*
rs869396	4	169,688,000	A/C	0.467	0.96	0.94–0.98	3.75×10^{-5}	0.93	0.90–0.97	8.18×10^{-5}	0.96	0.95–0.97	4.09×10^{-8}	0.12/0.58	*PALLD*
rs17680741	10	82,251,514	C/T	0.288	0.96	0.94–0.98	1.52×10^{-5}	0.93	0.89–0.96	7.21×10^{-5}	0.95	0.93–0.97	1.22×10^{-8}	0.15/0.51	*TSPAN14*, *SH2D4B*, *FAM213A*
rs7500448	16	83,045,790	G/A	0.252	0.95	0.92–0.97	2.11×10^{-6}	0.95	0.91–0.98	5.74×10^{-3}	0.95	0.93–0.97	4.09×10^{-8}	1.00/0	*CDH13*
rs15052	19	41,813,375	C/T	0.176	1.08	1.05–1.11	2.21×10^{-7}	1.07	1.03–1.12	2.11×10^{-3}	1.08	1.05–1.11	1.82×10^{-9}	0.77/0	*HNRNPUL1*, *TGFB1*, *CCDC97*, *AXL*
Type 2 diabetes															
rs1291041	5	78,443,735	T/G	0.352	0.95	0.93–0.98	5.70×10^{-5}	0.91	0.87–0.95	1.48×10^{-5}	0.94	0.92–0.96	2.26×10^{-8}	0.05/0.73	*BHMT*
rs744103	6	43,805,362	T/A	0.312	1.07	1.04–1.10	1.90×10^{-5}	1.09	1.04–1.14	4.32×10^{-4}	1.07	1.04–1.10	3.32×10^{-8}	0.60/0	*VEGFA*
rs4273712	6	126,964,510	G/A	0.266	1.07	1.04–1.10	1.20×10^{-6}	1.10	1.05–1.15	1.07×10^{-4}	1.08	1.05–1.11	8.38×10^{-10}	0.35/0	*CENPW*
rs301047	13	26,788,114	G/A	0.186	1.07	1.04–1.10	1.30×10^{-5}	1.11	1.05–1.17	1.61×10^{-4}	1.08	1.05–1.11	1.56×10^{-8}	0.26/0.22	*RNF6*
rs11658220	17	65,646,092	A/G	0.103	1.13	1.07–1.20	4.30×10^{-5}	1.16	1.08–1.24	2.03×10^{-5}	1.14	1.09–1.19	4.13×10^{-9}	0.57/0	*PITPNC1*

[a]Chromosome. [b]EA, effect allele; OA, other allele. [c]Effect allele frequency in controls. [d]Confidence interval of OR. [e]*P* value for Cochrane's *Q* statistic.

known risk loci and identified 17 new risk loci at genome-wide significance associated with Alzheimer's disease, coronary artery disease and type 2 diabetes.

Two future applications of this type of principle seem promising. First, for genetic studies of diseases with specific properties, it is likely advantageous to design a GWAX rather than a GWAS. These diseases include those that are rapidly lethal, like sudden infant death syndrome, glioblastoma or aortic rupture. Second, incorporating family history questionnaires into cohort studies may be immediately feasible. Large population cohorts such as the UK Biobank and NIH Precision Medicine Initiative along with participant-driven projects[2,37] are valuable resources in biomedical research. By performing association mapping using the family members of affected individuals, the ascertainment limitations inherent in these cohorts can be overcome.

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[38] 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[39]. Using liability as a phenotype can also account for ascertainment biases of case–control studies[40,41] and allow for much greater power to identify disease susceptibility variants.

URLs. UK Biobank, http://www.ukbiobank.ac.uk/; genotyping and quality control of UK Biobank, 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.*[24], http://pdgene.org/view?study=1.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank T. Hayeck and J. Jakobsdottir for comments on a draft of this manuscript. J.Z.L. and J.K.P. are partially supported by the National Institute of Mental Health (NIH grant R01MH106842). This research has been conducted using the UK Biobank Resource.

AUTHOR CONTRIBUTIONS

All authors contributed to the study design and writing, and all approved this manuscript. J.Z.L. performed the statistical analysis.

COMPETING FINANCIAL INTERESTS

The authors declare no completing financial interests.

1. Sudlow, C. *et al.* 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**, e1001779 (2015).
2. Eriksson, N. *et al.* Web-based, participant-driven studies yield novel genetic associations for common traits. *PLoS Genet.* **6**, e1000993 (2010).
3. Hayes, B.J., Bowman, P.J., Chamberlain, A.J. & Goddard, M.E. Genomic selection in dairy cattle: progress and challenges. *J. Dairy Sci.* **92**, 433–443 (2009).
4. Garrick, D.J., Taylor, J.F. & Fernando, R.L. Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genet. Sel. Evol.* **41**, 55 (2009).
5. Cole, J.B. *et al.* Distribution and location of genetic effects for dairy traits. *J. Dairy Sci.* **92**, 2931–2946 (2009).
6. Visscher, P.M. & Duffy, D.L. The value of relatives with phenotypes but missing genotypes in association studies for quantitative traits. *Genet. Epidemiol.* **30**, 30–36 (2006).
7. Chen, W.-M. & Abecasis, G.R. Family-based association tests for genomewide association scans. *Am. J. Hum. Genet.* **81**, 913–926 (2007).
8. Barzilai, N. *et al.* Unique lipoprotein phenotype and genotype associated with exceptional longevity. *J. Am. Med. Assoc.* **290**, 2030–2040 (2003).
9. Joshi, P.K. *et al.* Variants near *CHRNA3/5* and *APOE* have age- and sex-related effects on human lifespan. *Nat. Commun.* **7**, 11174 (2016).
10. Pilling, L.C. *et al.* Human longevity is influenced by many genetic variants: evidence from 75,000 UK Biobank participants. *Aging* **8**, 547–560 (2016).
11. Tan, Q., Zhao, J.H., Li, S., Kruse, T.A. & Christensen, K. Power assessment for genetic association study of human longevity using offspring of long-lived subjects. *Eur. J. Epidemiol.* **25**, 501–506 (2010).
12. Gudbjartsson, D.F. *et al.* Many sequence variants affecting diversity of adult human height. *Nat. Genet.* **40**, 609–615 (2008).
13. Kong, A. *et al.* Parental origin of sequence variants associated with complex diseases. *Nature* **462**, 868–874 (2009).
14. Thornton, T. & McPeek, M.S. Case–control association testing with related individuals: a more powerful quasi-likelihood score test. *Am. J. Hum. Genet.* **81**, 321–337 (2007).
15. Hebert, L.E., Scherr, P.A., Bienias, J.L., Bennett, D.A. & Evans, D.A. Alzheimer disease in the US population: prevalence estimates using the 2000 census. *Arch. Neurol.* **60**, 1119–1122 (2003).
16. de Lau, L.M. & Breteler, M.M. Epidemiology of Parkinson's disease. *Lancet Neurol.* **5**, 525–535 (2006).
17. Corder, E.H. *et al.* Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* **261**, 921–923 (1993).
18. Danesh, J., Collins, R. & Peto, R. Lipoprotein(a) and coronary heart disease. Meta-analysis of prospective studies. *Circulation* **102**, 1082–1085 (2000).
19. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
20. International Consortium for Blood Pressure Genome-Wide Association Studies. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* **478**, 103–109 (2011).
21. Hunter, D.J. *et al.* A genome-wide association study identifies alleles in *FGFR2* associated with risk of sporadic postmenopausal breast cancer. *Nat. Genet.* **39**, 870–874 (2007).
22. Grant, S.F.A. *et al.* Variant of transcription factor 7–like 2 (*TCF7L2*) gene confers risk of type 2 diabetes. *Nat. Genet.* **38**, 320–323 (2006).
23. Hung, R.J. *et al.* A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature* **452**, 633–637 (2008).
24. Nalls, M.A. *et al.* Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat. Genet.* **46**, 989–993 (2014).
25. Sasieni, P.D. From genotypes to genes: doubling the sample size. *Biometrics* **53**, 1253–1261 (1997).
26. International Schizophrenia Consortium. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
27. Lambert, J.-C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45**, 1452–1458 (2013).
28. CARDIoGRAMplusC4D Consortium. A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* **47**, 1121–1130 (2015).
29. DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium. *et al.* Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat. Genet.* **46**, 234–244 (2014).
30. Ripke, S. *et al.* A mega-analysis of genome-wide association studies for major depressive disorder. *Mol. Psychiatry* **18**, 497–511 (2013).
31. Friedmann, E. *et al.* Consensus analysis of signal peptide peptidase and homologous human aspartic proteases reveals opposite topology of catalytic domains compared with presenilins. *J. Biol. Chem.* **279**, 50790–50798 (2004).
32. Chan, G. *et al.* CD33 modulates TREM2: convergence of Alzheimer loci. *Nat. Neurosci.* **18**, 1556–1558 (2015).
33. Gjoneska, E. *et al.* Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. *Nature* **518**, 365–369 (2015).
34. Kurogane, Y. *et al.* FGD5 mediates proangiogenic action of vascular endothelial growth factor in human vascular endothelial cells. *Arterioscler. Thromb. Vasc. Biol.* **32**, 988–996 (2012).

35. Taimeh, Z., Loughran, J., Birks, E.J. & Bolli, R. Vascular endothelial growth factor in heart failure. *Nat. Rev. Cardiol.* **10**, 519–530 (2013).
36. Garner, K. *et al.* Phosphatidylinositol transfer protein, cytoplasmic 1 (PITPNC1) binds and transfers phosphatidic acid. *J. Biol. Chem.* **287**, 32263–32276 (2012).
37. Dolgin, E. Personalized investigation. *Nat. Med.* **16**, 953–955 (2010).
38. Ledford, H. Genome hacker uncovers largest-ever family tree. *Nature* http://dx.doi.org/10.1038/nature.2013.14037 (2013).
39. Campbell, D.D., Sham, P.C., Knight, J., Wickham, H. & Landau, S. Software for generating liability distributions for pedigrees conditional on their observed disease states and covariates. *Genet. Epidemiol.* **34**, 159–170 (2010).
40. Hayeck, T.J. *et al.* Mixed model with correction for case–control ascertainment increases association power. *Am. J. Hum. Genet.* **96**, 720–730 (2015).
41. Weissbrod, O., Lippert, C., Geiger, D. & Heckerman, D. Accurate liability estimation improves power in ascertained case–control studies. *Nat. Methods* **12**, 332–334 (2015).

ONLINE 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 was equivalent between the two designs. 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 these power calculations are described in the **Supplementary Note**.

UK Biobank data collection. The UK Biobank[1] is a large population-based study of over 500,000 subjects aged 40–69 years recruited from 2006–2010. Participants entered information about their family history of disease by answering three questions: (1) "Has/did your father ever suffer from any of the following diseases?"; (2) "Has/did your mother ever suffer from any of the following diseases?"; 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 answer" or "None of the above." Throughout this manuscript, we use heart disease, severe depression and diabetes to refer specifically to coronary artery disease, major depressive disorder and type 2 diabetes, respectively. Case/control status for the participants themselves was available via health records (ICD-10 diagnoses; **Supplementary Table 5**). The UK Biobank received ethics approval from the National Health Service National Research Ethics Service (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 equivalent power to detect association as the observed sample size with an unequal case/control balance. The test statistic for a standard 2 × 2 1-degree-of-freedom χ^2 test when the numbers of cases and controls differ is

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

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

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

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

$$N_{\text{eff}} = \frac{4}{1/N_A + 1/N_U}$$

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 includes directly genotyped and imputed data for 152,529 individuals. Around 90% of individuals were genotyped on the Affymetrix UK Biobank Axiom array, while the remaining individuals were genotyped on the Affymetrix UK BiLEVE array. The two platforms are similar, with >95% marker content in common (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 frequencies >1%. Genotyped individuals were phased using SHAPEIT2 (ref. 42) and then imputed with the IMPUTE2 (ref. 43) algorithm using a reference panel consisting of 12,570 haplotypes from a combined UK10K[44] and 1000 Genomes Project Phase 3 data set[45]. In total, 73,355,667 polymorphic variants were successfully imputed. Additional information on the genotyping arrays, sample preparation and quality control can be found in the documents referenced in the URLs section. After quality control, we took forward 116,196 unrelated individuals of European descent for analysis.

Genome-wide association by proxy. The GWAX study design is an extreme version of approaches that try to impute unknown genotypes in phenotyped individuals on the basis of the genotypes of close relatives, although these approaches require accurate pedigree information[12,46] and/or sparse genotypes (for example, microsatellites) on which to impute[47]. Our approach is also similar to that of MQLS[14], 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 (J. Jakobsdottir (Icelandic Heart Association), 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, standard logistic regression scales easily to large data sets 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 had 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 is listed in **Supplementary Table 1**.

Analysis of 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 (ref. 48). For all analyses, we included the subjects' reported sex, age at recruitment and the first four principal components (estimated directly from the post–quality control set of UK Biobank individuals) as covariates.

As with standard GWAS, spurious associations may be the result of population structure, and associations at loci such as the human leukocyte antigen (HLA) region may be difficult to interpret owing to complex patterns of linkage disequilibrium (LD). We observed modest genomic inflation across the 12 diseases ($1.05 < \lambda < 1.07$; **Supplementary Fig. 8**). To test whether this inflation is due to population stratification or reflects a true polygenic signal, we performed LD score regression on the summary association statistics using a set of 1.2 million common SNPs from HapMap 3 (ref. 49). The LD score regression intercepts were between 0.99 and 1.02 (**Supplementary Fig. 8**), 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[50], where genetic relatedness within the UK Biobank was estimated using 623,852 directly genotyped SNPs. *P* values were very similar to those from the logistic regression using four principal components (**Supplementary Table 2**). 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 to enable fixed-effects meta-analysis), we calculated 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 let

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

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

$$f_P = \frac{f_U + f_A}{2}$$

doi:10.1038/ng.3766

To estimate the adjusted odds ratio as a function of the observed allele frequency in proxy cases and controls, we substitute f_A into f_P

$$\widehat{\mathrm{OR}} = \frac{(2f_P - f_U)(1 - f_U)}{f_U(1 - 2f_P + f_U)}$$

For the range of odds ratios (<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 (**Supplementary Fig. 2**). 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 the $\widehat{\mathrm{OR}}$ derivation above. The corresponding adjusted standard error is also double the standard error of the log odds ratio from logistic regression, as $\mathrm{se}^2 = \mathrm{var}(2\beta) = 2^2\mathrm{var}(\beta)$, where se is the standard error and β is the effect size.

Polygenic risk scores. Publically availably 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)[27], coronary artery disease (60,801 cases and 123,504 controls)[28], major depressive disorder (9,249 cases and 9,519 controls)[30] and type 2 diabetes (26,488 cases and 83,964 controls)[29].

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[26] for each individual in the UK Biobank. No genome-wide significant loci were reported for major depressive disorder. For a disease with m associated SNPs, the polygenic risk score for individual i is

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

where $\hat{\beta}_j$ is the reported effect size (log odds ratio) 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 in the UK Biobank who were (i) affected with disease, (ii) unaffected but had two affected parents, and (iii) unaffected but had one affected parent. For the last group, the mean risk score was doubled so that it was equivalent (assuming additivity) 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 published GWAS summary statistics using LD score regression with a set of ~1.2 million common SNPs from HapMap 3 (ref. 49).

Meta-analysis. Fixed-effects meta-analysis was performed for Alzheimer's disease, coronary artery disease, major depressive disorder and type 2 diabetes using the inverse-variance-weighted method for all SNPs that overlapped in 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 $\widehat{\mathrm{se}}_i$, respectively, where i = 1 or 2 corresponds to the GWAX (adjusted log odds ratio) and GWAS results, the combined effect size is

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

with corresponding standard error and P value

$$\widehat{\mathrm{se}}_{\mathrm{meta}} = \sqrt{1/\sum_i w_i}$$

and

$$P_{\mathrm{meta}} = 2\Phi(-|\hat{\beta}_{\mathrm{meta}}|/\widehat{\mathrm{se}}_{\mathrm{meta}})$$

where $w_i = 1/\mathrm{se}_i^2$ and Φ is the cumulative standard normal distribution. Genomic inflation factors ranged from 1.06 to 1.17, while LD score regression intercepts ranged from 0.99 to 1.05 (**Supplementary Fig. 9**).

Identification of independent risk loci. A locus was considered to be 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 loci were identified using the approximate conditional and joint association method implemented in GCTA (GCTA-COJO). The method performs approximate stepwise 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 used a reference panel consisting of 2,500 randomly selected individuals from the UK Biobank cohort[51].

Data availability. Summary association statistics from the UK Biobank GWAX and GWAX + GWAS meta-analyses can be found at http://gwas-browser.nygenome.org/downloads/gwas-browser/.

42. Delaneau, O., Marchini, J. & 1000 Genomes Project Consortium. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nat. Commun.* **5**, 3934 (2014).
43. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
44. UK10K Consortium. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).
45. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
46. Rafnar, T. *et al.* Mutations in *BRIP1* confer high risk of ovarian cancer. *Nat. Genet.* **43**, 1104–1107 (2011).
47. Burdick, J.T., Chen, W.-M., Abecasis, G.R. & Cheung, V.G. *In silico* method for inferring genotypes in pedigrees. *Nat. Genet.* **38**, 1002–1004 (2006).
48. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
49. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
50. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
51. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375, S1–S3 (2012).