Biology, Chemistry, and Environmental Sciences Faculty Articles and Research

Science and Technology Faculty Articles and Research

2011

Efficient Replication of over 180 Genetic Associations with Self-Reported Medical Data

Joyce Y. Tung
23andMe

Chuong B. Do
23andMe

David A. Hinds
23andMe

Amy K. Kiefer
23andMe

J. Michael Macpherson
Chapman University, mmacpher@chapman.edu

See next page for additional authors

Recommended Citation

Efficient Replication of over 180 Genetic Associations with Self-Reported Medical Data

Creative Commons License

Authors

Joyce Y. Tung, Chuong B. Do, David A. Hinds, Amy K. Kiefer, J. Michael Macpherson, Arnab B. Chowdry, Uta Francke, Brian Naughton, Joanna Mountain, Anne Wojcicki, and Nicholas Eriksson

Efficient Replication of over 180 Genetic Associations with Self-Reported Medical Data

Joyce Y. Tung[1]*, Chuong B. Do[1], David A. Hinds[1], Amy K. Kiefer[1], J. Michael Macpherson[1], Arnab B. Chowdry[1], Uta Francke[1,2], Brian T. Naughton[1], Joanna L. Mountain[1], Anne Wojcicki[1], Nicholas Eriksson[1]

1 23andMe, Inc., Mountain View, California, United States of America, 2 Department of Genetics, Stanford University, Stanford, California, United States of America

Abstract

While the cost and speed of generating genomic data have come down dramatically in recent years, the slow pace of collecting medical data for large cohorts continues to hamper genetic research. Here we evaluate a novel online framework for obtaining large amounts of medical information from a recontactable cohort by assessing our ability to replicate genetic associations using these data. Using web-based questionnaires, we gathered self-reported data on 50 medical phenotypes from a generally unselected cohort of over 20,000 genotyped individuals. Of a list of genetic associations curated by NHGRI, we successfully replicated about 75% of the associations that we expected to (based on the number of cases in our cohort and reported odds ratios, and excluding a set of associations with contradictory published evidence). Altogether we replicated over 180 previously reported associations, including many for type 2 diabetes, prostate cancer, cholesterol levels, and multiple sclerosis. We found significant variation across categories of conditions in the percentage of expected associations that we were able to replicate, which may reflect systematic inflation of the effects in some initial reports, or differences across diseases in the likelihood of misdiagnosis or misreport. We also demonstrated that we could improve replication success by taking advantage of our recontactable cohort, offering more in-depth questions to refine self-reported diagnoses. Our data suggest that online collection of self-reported data from a recontactable cohort may be a viable method for both broad and deep phenotyping in large populations.

Competing Interests: JYT, CBD, DAH, AKK, JMM, ABC, UF, BTN, JLM, AW, AND NE are employed by 23andMe and own stock options in the company. There are no patents, products in development or marketed products to declare. This does not alter the authors' adherence to all the PLoS ONE policies on sharing data and materials, as detailed online in the guide for authors. However, the authors' obligations to protect their customers' privacy (as outlined in our Terms of Service and Privacy Statement) prevent them from making their customers' individual-level data publicly available. Aggregate-level data (for example, in the form of 2×3 tables that were used for their statistics) can be made available upon request.

* E-mail: joyce@23andme.com

Introduction

In the last few years, the cost of collecting genomic data has declined rapidly. However, advances in the collection of phenome data (the set of all phenotypic information from a single organism) have not kept pace [1,2]. New techniques are needed to complement the wealth of genomic data and build the large cohorts needed for highly-powered genome-wide association studies (GWAS).

The reliability of phenotyping is important for GWAS. Phenotyping error decreases power, which can be problematic as most GWAS are not sufficiently powered to explain a significant fraction of the underlying heritability. Although increasing sample sizes can counteract the problems caused by misclassification, it is the very issue of needing ever larger samples that necessitates more efficient methods for collecting data [3]. A number of associations reported in very large meta-analyses have not been replicated, and may never be, simply because of the difficulty of assembling such a sizeable cohort of patients. There is a need for more straightforward methods to quickly and reliably gather retrospective phenotype information from large cohorts of people, not only to validate existing associations, but to discover new ones.

Although self-report has frequently been used for physical traits, medical records have traditionally been the preferred source of retrospective information on medical conditions. Previous studies have reported good agreement between medical record data and self-reported medical conditions [4–8], and include a few common themes. First, there tends to be good concordance for well-defined and easily diagnosed diseases and for chronic conditions that require repeated medical follow-up (kappa ranging from 0.71–0.80 for diabetes, hypertension, myocardial infarction, and stroke) [4–9]. Second, a negative self-report is very likely to agree with a negative result from the medical record [5]. Third, self-reports of conditions that are milder, less specific (such as heart failure), or communicated in different ways by physicians (such as high cholesterol) tend to be less consistent with medical records, possibly because the original diagnosis was less certain or because of insufficient physician—patient communication [4,8,10]. Fourth, medical records, especially in countries without centralized healthcare, typically only address diseases from a limited portion of a patient's life; self-report can be more accurate for diseases outside this window [5].

To begin to address the phenomics problem, a structure that facilitates both broad and deep phenotyping and maximizes the

utility of information gathered while minimizing the burden on participants is needed. In this study, we evaluate a research model in which a large, recontactable cohort is surveyed online across a broad range of phenotypes. Subsets of this cohort with particular characteristics can then be contacted for further research with more in-depth phenotyping on specific topics as appropriate. We have demonstrated previously that this model can be used to discover and replicate associations with non-medical traits [11]. Here, by assessing our ability to replicate previously reported genetic associations across a wide range of conditions, we demonstrate that broad self-reported data collection online is useful for medically-related conditions as well. We show that some classes of conditions lend themselves particularly well to simple self-report, while others are more complex. We also show that the ability to recontact the cohort facilitates rapid refinement of phenotype characterization.

Results

We sought to replicate associations from the list curated by the National Human Genome Research Institute's Office of Population Genomics ("GWAS catalog") in a cohort of 20,182 participants of European ancestry who filled out surveys on the 23andMe website [12,13]. Members of the cohort, drawn from the 23andMe customer base, had been genotyped at approximately 600,000 SNPs, and had access to their raw genetic data as well as health- and ancestry-related interpretations of their data. The majority of the cohort was not selected based on disease status or other characteristics and is roughly a representative sample from the 23andMe customer base; approximately 4,000 members of the cohort had been recruited for a study on Parkinson's disease or a project with the National Senior Games Association (http://www.nsga.com/).

Phenotypes from the GWAS catalog were matched with available phenotypes from the online surveys (see Methods). In order to collect data on a wide range of phenotypes while keeping the time spent answering surveys low, we chose to assess most phenotypes using only single questions of the general form "Have you ever been diagnosed by a doctor with [Condition X]?". A total of 50 conditions from the GWAS list had direct analogues within the 23andMe database. For each condition, we used only one SNP from each linkage disequilibrium (LD) block (using a threshold of $r^2 > = 0.1$) and removed SNPs that were not on our platform or did not have a proxy SNP on our platform with LD of $r^2 > = 0.5$. Ultimately, we attempted to replicate a total of 392 different associations (315 case—control, 77 quantitative) for these 50 phenotypes.

Using a one-sided $P<0.05$ threshold for significance, we replicated 144 (93 case—control and 51 quantitative) of the 392 attempted associations in 36 of the 50 phenotypes (Figure 1, Table S1, Table S2). For some conditions, the size of our case group was quite small; however, the odds ratios for replicated SNPs were generally in good agreement with the published odds ratios. Of the case—control conditions, 84.6% of the replications had 95% confidence intervals containing the published odds ratio (replications for pigmentary phenotypes such as eye color, hair color, and freckling were not included as the assessment scales are not easy to match across different published reports). As one aspect of 23andMe's Personal Genome Service involves returning genetic data to our customers, we investigated the possibility that customers viewing a result of elevated risk for a certain disease before answering surveys may skew the results towards replication. To address this, we investigated the impact of seeing genetic risk results before versus after answering survey questions on self-reported disease status for a set of 20 conditions for which

participants were able to view a personal risk prediction. We observed that in general, the nature of the genetic risk result did not have a consistent or significant effect on the way questions were answered (Table S4).

As many of the known associations were discovered in large case—control studies, we expected to have low power to detect many of them using our mostly unselected cohort. Thus, to further assess our success in replication, we calculated our power to replicate each association for all case—control phenotypes (quantitative phenotypes were skipped in this calculation due to difficulties in matching scales in many of them). Power was calculated using the odds ratios reported in the GWAS catalog. To allow for phenotyping error, the calculations assumed that 5% of our reported cases are actually controls and that the minimum of the prevalence or 10% of our controls are (or will someday be) cases. For diseases with over 10% prevalence, controls were generally chosen to be of sufficient age so that at most 10% of people without the disease at that age would be expected to develop the disease. See Methods for full details.

We replicated 70% of the number of replications expected (93 replications against 132.7 expected), given our sample sizes (Figure 2). Some of the failed replications can be traced back to the possibility that the reported effect sizes for these associations are inflated or that the associations themselves are false positives. Table 1 shows all associations for which we had at least 80% power to replicate but failed to do so. Of these 19 associations, five have failed to replicate elsewhere despite high power to do so, two have shown significant heterogeneity of odds ratios in meta-analyses, and two exhibited significant signals when the two stages of a multi-stage study were combined but were not interpreted as significant by the authors. Removal of these nine SNPs for which our power may be substantially overestimated increased our replication rate to about 75% (93 out of 124.0). Furthermore, among the remaining 213 SNPs that we did not replicate, we observed the correct directionality of association for 126 out of an expected 172.0 SNPs (using a p-value threshold set to 0.5), yielding a rate of 73.2% relative to expected. Interestingly, an inflammatory bowel disease (IBD) association (rs7517847) that we had high power to replicate also failed to replicate in ulcerative colitis only cases in an Italian study of IBD, suggesting that the association may be specific to Crohn's disease, as opposed to all types of IBD [14]. Overall, our success rates differed vastly for different classes of diseases (Figure 2, Figure S1), suggesting that the difference between theoretical and actual power is to some extent explained by differences in phenotyping (discussed in more detail below).

Separate from these calculations, we also attempted to replicate 106 associations with phenotypes in our cohort that were only in weak correspondence with phenotypes in the published papers. Of the 106, we replicated 39 associations. For example, while we did not collect data on gallstones, we did collect data on gall bladder surgery which is often a consequence of gallstones, and were able to replicate one association with gallstones. Likewise, answers to the question "Have you ever been diagnosed by a doctor with high cholesterol (over 200 mg/dl) or hypercholesterolemia?" were sufficient to replicate 19 associations with cholesterol level. Data on receiving an abnormal result on a liver function test result were sufficient to replicate four associations with bilirubin levels. A summary list of these replications can be found in Table 2 (full list in Table S3).

Discussion

Advances in technology have driven down the price and difficulty of genotyping, but until recently, the same has not been

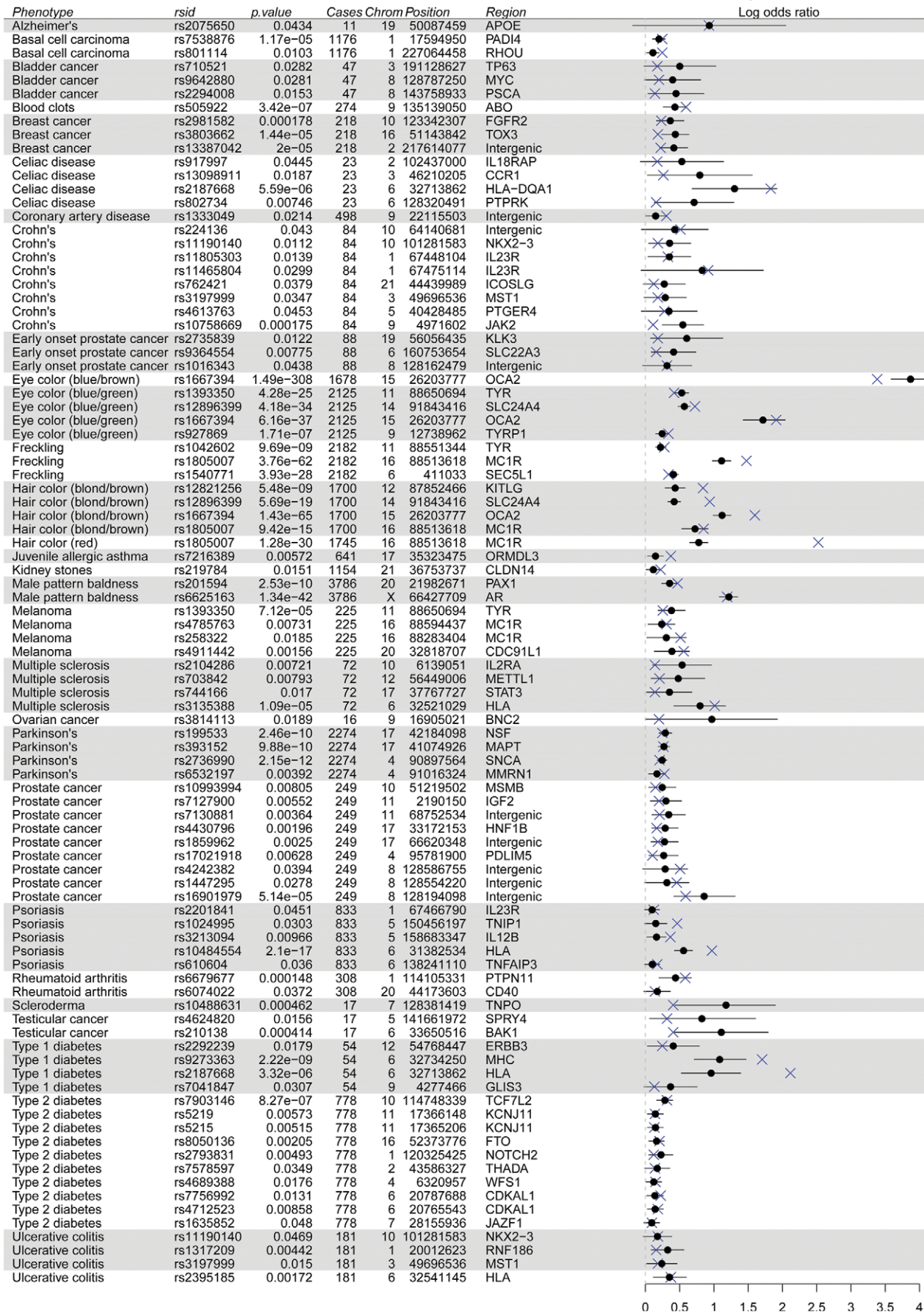| Phenotype	rsid	p.value	Cases	Chrom	Position	Region	Log odds ratio
Alzheimer's	rs2075650	0.0434	11	19	50087459	APOE	
Basal cell carcinoma	rs7538876	1.17e−05	1176	1	17594950	PADI4	
Basal cell carcinoma	rs801114	0.0103	1176	1	227064458	RHOU	
Bladder cancer	rs710521	0.0282	47	3	191128627	TP63	
Bladder cancer	rs9642880	0.0281	47	8	128787250	MYC	
Bladder cancer	rs2294008	0.0153	47	8	143758933	PSCA	
Blood clots	rs505922	3.42e−07	274	9	135139050	ABO	
Breast cancer	rs2981582	0.000178	218	10	123342307	FGFR2	
Breast cancer	rs3803662	1.44e−05	218	16	51143842	TOX3	
Breast cancer	rs13387042	2e−05	218	2	217614077	Intergenic	
Celiac disease	rs917997	0.0445	23	2	102437000	IL18RAP	
Celiac disease	rs13098911	0.0187	23	3	46210205	CCR1	
Celiac disease	rs2187668	5.59e−06	23	6	32713862	HLA−DQA1	
Celiac disease	rs802734	0.00746	23	6	128320491	PTPRK	
Coronary artery disease	rs1333049	0.0214	498	9	22115503	Intergenic	
Crohn's	rs224136	0.043	84	10	64140681	Intergenic	
Crohn's	rs11190140	0.0112	84	10	101281583	NKX2−3	
Crohn's	rs11805303	0.0139	84	1	67448104	IL23R	
Crohn's	rs11465804	0.0299	84	1	67475114	IL23R	
Crohn's	rs762421	0.0379	84	21	44439989	ICOSLG	
Crohn's	rs3197999	0.0347	84	3	49696536	MST1	
Crohn's	rs4613763	0.0453	84	5	40428485	PTGER4	
Crohn's	rs10758669	0.000175	84	9	4971602	JAK2	
Early onset prostate cancer	rs2735839	0.0122	88	19	56056435	KLK3	
Early onset prostate cancer	rs9364554	0.00775	88	6	160753654	SLC22A3	
Early onset prostate cancer	rs1016343	0.0438	88	8	128162479	Intergenic	
Eye color (blue/brown)	rs1667394	1.49e−308	1678	15	26203777	OCA2	
Eye color (blue/green)	rs1393350	4.28e−25	2125	11	88650694	TYR	
Eye color (blue/green)	rs12896399	4.18e−34	2125	14	91843416	SLC24A4	
Eye color (blue/green)	rs1667394	6.16e−37	2125	15	26203777	OCA2	
Eye color (blue/green)	rs927869	1.71e−07	2125	9	12738962	TYRP1	
Freckling	rs1042602	9.69e−09	2182	11	88551344	TYR	
Freckling	rs1805007	3.76e−62	2182	16	88513618	MC1R	
Freckling	rs1540771	3.93e−28	2182	6	411033	SEC5L1	
Hair color (blond/brown)	rs12821256	5.48e−09	1700	12	87852466	KITLG	
Hair color (blond/brown)	rs12896399	5.69e−19	1700	14	91843416	SLC24A4	
Hair color (blond/brown)	rs1667394	1.43e−65	1700	15	26203777	OCA2	
Hair color (blond/brown)	rs1805007	9.42e−15	1700	16	88513618	MC1R	
Hair color (red)	rs1805007	1.28e−30	1745	16	88513618	MC1R	
Juvenile allergic asthma	rs7216389	0.00572	641	17	35323475	ORMDL3	
Kidney stones	rs219784	0.0151	1154	21	36753737	CLDN14	
Male pattern baldness	rs201594	2.53e−10	3786	20	21982671	PAX1	
Male pattern baldness	rs6625163	1.34e−42	3786	X	66427709	AR	
Melanoma	rs1393350	7.12e−05	225	11	88650694	TYR	
Melanoma	rs4785763	0.00731	225	16	88594437	MC1R	
Melanoma	rs258322	0.0185	225	16	88283404	MC1R	
Melanoma	rs4911442	0.00156	225	20	32818707	CDC91L1	
Multiple sclerosis	rs2104286	0.00721	72	10	6139051	IL2RA	
Multiple sclerosis	rs703842	0.00793	72	12	56449006	METTL1	
Multiple sclerosis	rs744166	0.017	72	17	37767727	STAT3	
Multiple sclerosis	rs3135388	1.09e−05	72	6	32521029	HLA	
Ovarian cancer	rs3814113	0.0189	16	9	16905021	BNC2	
Parkinson's	rs199533	2.46e−10	2274	17	42184098	NSF	
Parkinson's	rs393152	9.88e−10	2274	17	41074926	MAPT	
Parkinson's	rs2736990	2.15e−12	2274	4	90897564	SNCA	
Parkinson's	rs6532197	0.00392	2274	4	91016324	MMRN1	
Prostate cancer	rs10993994	0.00805	249	10	51219502	MSMB	
Prostate cancer	rs7127900	0.00552	249	11	2190150	IGF2	
Prostate cancer	rs7130881	0.00364	249	11	68752534	Intergenic	
Prostate cancer	rs4430796	0.00196	249	17	33172153	HNF1B	
Prostate cancer	rs1859962	0.0025	249	17	66620348	Intergenic	
Prostate cancer	rs17021918	0.00628	249	4	95781900	PDLIM5	
Prostate cancer	rs4242382	0.0394	249	8	128586755	Intergenic	
Prostate cancer	rs1447295	0.0278	249	8	128554220	Intergenic	
Prostate cancer	rs16901979	5.14e−05	249	8	128194098	Intergenic	
Psoriasis	rs2201841	0.0451	833	1	67466790	IL23R	
Psoriasis	rs1024995	0.0303	833	5	150456197	TNIP1	
Psoriasis	rs3213094	0.00966	833	5	158683347	IL12B	
Psoriasis	rs10484554	2.1e−17	833	6	31382534	HLA	
Psoriasis	rs610604	0.036	833	6	138241110	TNFAIP3	
Rheumatoid arthritis	rs6679677	0.000148	308	1	114105331	PTPN11	
Rheumatoid arthritis	rs6074022	0.0372	308	20	44173603	CD40	
Scleroderma	rs10488631	0.000462	17	7	128381419	TNPO	
Testicular cancer	rs4624820	0.0156	17	5	141661972	SPRY4	
Testicular cancer	rs210138	0.000414	17	6	33650516	BAK1	
Type 1 diabetes	rs2292239	0.0179	54	12	54768447	ERBB3	
Type 1 diabetes	rs9273363	2.22e−09	54	6	32734250	MHC	
Type 1 diabetes	rs2187668	3.32e−06	54	6	32713862	HLA	
Type 1 diabetes	rs7041847	0.0307	54	9	4277466	GLIS3	
Type 2 diabetes	rs7903146	8.27e−07	778	10	114748339	TCF7L2	
Type 2 diabetes	rs5219	0.00573	778	11	17366148	KCNJ11	
Type 2 diabetes	rs5215	0.00515	778	11	17365206	KCNJ11	
Type 2 diabetes	rs8050136	0.00205	778	16	52373776	FTO	
Type 2 diabetes	rs2793831	0.00493	778	1	120325425	NOTCH2	
Type 2 diabetes	rs7578597	0.0349	778	2	43586327	THADA	
Type 2 diabetes	rs4689388	0.0176	778	4	6320957	WFS1	
Type 2 diabetes	rs7756992	0.0131	778	6	20787688	CDKAL1	
Type 2 diabetes	rs4712523	0.00858	778	6	20765543	CDKAL1	
Type 2 diabetes	rs1635852	0.048	778	7	28155936	JAZF1	
Ulcerative colitis	rs11190140	0.0469	181	10	101281583	NKX2−3	
Ulcerative colitis	rs1317209	0.00442	181	1	20012623	RNF186	
Ulcerative colitis	rs3197999	0.015	181	3	49696536	MST1	
Ulcerative colitis	rs2395185	0.00172	181	6	32541145	HLA	

Figure 1. Replicated SNPs for binary traits. Our log ORs and 95% confidence intervals are shown as black circles and lines. Published ORs are shown as blue Xs.

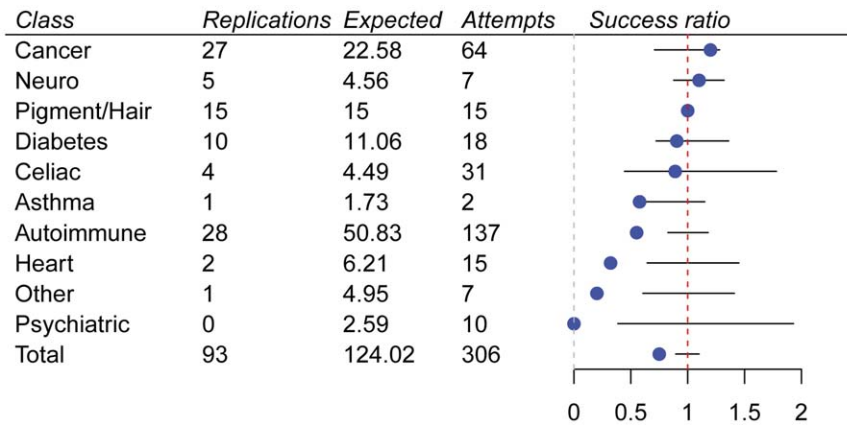doi:10.1371/journal.pone.0023473.g001

Class	Replications	Expected	Attempts	Success ratio
Cancer	27	22.58	64	
Neuro	5	4.56	7	
Pigment/Hair	15	15	15	
Diabetes	10	11.06	18	
Celiac	4	4.49	31	
Asthma	1	1.73	2	
Autoimmune	28	50.83	137	
Heart	2	6.21	15	
Other	1	4.95	7	
Psychiatric	0	2.59	10	
Total	93	124.02	306	

Figure 2. Success rate (versus total power) by disease class. Replications = number of associations we successfully replicated. Expected = number of associations we expected to replicate. Attempts = number of associations we attempted to replicate. The blue dot represents our success ratio (number of successful replications divided by number of expected replications). The black line represents the 95% prediction interval for the success ratio. The nine associations that we had high power to detect but had known conflicting data were not included in this figure (see text and Table 1). Conditions assigned to each class (also see Methods S1): Asthma: childhood asthma; Autoimmune: Crohn's disease, inflammatory bowel disease, lupus, multiple sclerosis, psoriasis, type 1 diabetes, ulcerative colitis; Cancer: basal cell carcinoma, bladder cancer, breast cancer, colorectal cancer, prostate cancer, lung cancer, melanoma, pancreatic cancer, scleroderma, testicular cancer, thyroid cancer; Celiac: celiac disease; Diabetes: type 2 diabetes; Heart: blood clots, coronary artery disease, heart attack; Pigment/Hair: eye color, freckling, hair color, red hair color, male pattern baldness; Neuro: Alzheimer's disease, autism, Parkinson's disease; Other: chronic obstructive pulmonary disease, kidney stones, stroke, osteoarthritis; Psychiatric: alcohol abuse, bipolar disorder, schizophrenia.
doi:10.1371/journal.pone.0023473.g002

Table 1. Associations with sufficient power for detection ($> = 80\%$) that failed to replicate.

Phenotype	SNP	Pub OR	Rep OR	P-value	Power	Cases	Controls	Replications in the Literature
Alcohol abuse	rs7590720	1.35	0.955	0.875	1	1811	8549	Failed to replicate [30]
Bipolar disorder	rs1012053	1.59	1	0.485	1	366	13030	Failed to replicate [31–34]
Bipolar disorder	rs420259	2.08	0.966	0.659	1	366	13030	Failed to replicate [32–34]
COPD[a]	rs13180	1.3	1.05	0.26	0.89	403	2306	Replicated [35]
COPD[a]	rs7671167	1.32	1.11	0.0968	0.93	403	2306	Replicated [35]
COPD[a]	rs1828591	1.38	1	0.489	0.97	403	2306	Replicated [36,37]
Crohn's disease	rs2066847	3.99	1.54	0.151	0.88	84	13288	Replicated [38]
IBD[b]	rs7517847	1.61	0.855	0.954	1	250	12808	Replicated [39][c]
Juvenile allergic asthma	rs2786098	1.43	1.07	0.181	1	641	6584	Failed to replicate [40]
Lupus	rs3131379	2.36	1.38	0.133	0.82	52	11675	Not yet replicated
Parkinson's disease	rs17115100	1.25	0.992	0.555	0.97	2274	5336	Not claimed [41][d]
Parkinson's disease	rs823128	1.52	1.17	0.0531	1	2274	5336	Not claimed [41][d]
Psoriasis	rs20541	1.27	1.09	0.0953	0.92	833	4291	Replicated [42,43]
Rheumatoid arthritis	rs10499194	1.33	0.927	0.797	0.9	308	12845	Failed to replicate [44]
Rheumatoid arthritis	rs3761847	1.32	1.01	0.437	0.93	308	12845	Between-study heterogeneity [45]
Type 2 diabetes	rs9300039	1.48	0.976	0.595	0.97	778	3273	Between-study heterogeneity [46]
Type 2 diabetes	rs2943641	1.19	1.03	0.328	0.81	778	3273	Not yet replicated
Thyroid cancer	rs965513	1.75	1.37	0.0559	0.83	52	11234	Replicated [47]
Ulcerative colitis	rs11209026	1.79	1.47	0.0577	0.85	181	13100	Replicated [48]

Pub OR = published odds ratio. Rep OR = 23andMe attempted replication odds ratio. Power = estimated power to detect association.
[a]COPD = Chronic Obstructive Pulmonary Disease. This analysis included smokers only.
[b]IBD = Inflammatory Bowel Disease.
[c]This SNP was initially associated with IBD, but replicated only for Crohn's disease [39], which is a subtype of IBD. Latiano et al. also replicates rs7517847 with Crohn's disease, but not with ulcerative colitis, which is the other major subtype of IBD [14].
[d]This association was curated into the GWAS catalog as significantly associated with Parkinson's disease but was not identified by the authors as significant.
doi:10.1371/journal.pone.0023473.t001

Table 2. Replications without strictly matching phenotypes.

23andMe Phenotype	Published Phenotype	# Replications	Genes
Liver test	Bilirubin levels	4	CHUK, GGT1, SAMM50, UGT1A1
High cholesterol	Cholesterol levels (quantitative)[a]	19	ABCG8, APOA1, APOB, CELSR2, CILP2, DNAH11, DOCK7, FADS1, GCKR, HNF1A, LDLR, LIPC (×2), MAFB, NCAN, PCSK9 (×2), TOMM40, TRIB1
Gall bladder removal	Gallstones	1	ABCG8
High blood pressure	Blood pressure (quantitative)[a]	8	ATP2B1, CYP17A1 (×2), CYP1A1, FGF5, SH2B3, ULK4, ZNF652
Osteoporosis	Bone mineral density (quantitative)[a]	5	MEF2C, MEPE, OSX, SOX6, SPTBN1
Macular degeneration	Advanced age-related macular degeneration	2	C2, C3
Nicotine abuse	Nicotine dependence	1	CHRNA3

[a]These phenotypes were measured quantitatively in the published reports, but the corresponding 23andMe phenotypes listed here were measured qualitatively (yes/no).

doi:10.1371/journal.pone.0023473.t002

true of phenotyping [15]. We propose that web-based collection of self-reported data on medical phenotypes is an efficient and effective method for phenotyping a large cohort of individuals, as evidenced by our ability to replicate a high percentage of associations across a wide range of conditions. Relative to medical record review, internet-based phenotyping is fast (we assessed more than 20,000 people for 50 phenotypes in approximately 12 months using only a small team of people). To our knowledge, this is the largest number of replications across a wide variety of diseases ever reported, demonstrating the value of gathering self-reported data on a large genotyped population.

While many of the associations tested here have been replicated before, there are a few that are, based on our literature review, the first independent replications of these associations in a population of European descent: basal cell carcinoma (*PADI4, RHOU*), plasma levels of liver enzymes (*PNPLA3*), and bone mineral density (*MEF2C, MEPE*—these have already been replicated in a population of Asian descent). Though our study has been performed in a population of European ancestry, a similar study would be feasible in other populations. Such a study could potentially improve risk prediction in non-European populations as well as further our understanding of disease architecture (e.g., understanding how effect size varies across populations could provide insight into how tightly linked associations are to the causal variants). Furthermore, while it is true that we are able to replicate previously identified associations using our research platform, the reverse is also true—novel discoveries using our method have been independently replicated using other modes of data collection for both traits and medical conditions [16–19].

Although most studies use medical records as the gold standard against which self-reported data are compared, there are some inherent challenges to the use of medical records [5]. As very few people have received all their health care from the same provider, the medical records from different stages of their lives are stored at different sites of care. Thus, a childhood diagnosis of asthma might be stored in a record at the pediatrician's office but not be reported in the record at the adult medical practice. In addition, extracting data from medical records often requires either manual curation, which is time-consuming and expensive, or reliance on ICD-9-CM or CPT codes which may have been miscoded. For example, a replication study was carried out using the BioVU DNA databank at Vanderbilt University by applying natural language processing techniques and billing-code queries to electronic medical records

[20]. Their algorithms achieved high positive predictive value (as measured by independent record review by two physicians) but required manual review and significant iterative work. Out of 21 SNPs in five phenotypes, they were able to replicate eight associations. In contrast, we were able to examine 50 phenotypes and replicate over 180 associations. For cases in which the information required may be difficult for individuals to report but can be extracted from electronic medical records (such as lab values), these two methods can provide complementary sources of data.

We replicated approximately 75% of the associations we expected to (excluding those for which our power may be substantially overestimated), based on power calculations. There are several possible reasons why we did not replicate all the associations we expected to (see Figure 2 and Figure S1 for instances in which our success ratio did not overlap the 95% prediction interval). One factor is systematic inflation of odds ratios in the initial reports due to the winner's curse—a bias in the effect size estimates from the first publication to report an association, generally occurring when the discovery sample is poorly powered to detect the association [21]. For example, if we were to assume a systematic inflation of 15% in the log-odds ratio, the replication rate would change from 70% to 77% (or 75% to 82% if we again exclude the nine associations that are not clearly true positives). This amount of inflation is entirely within the confidence intervals for most studies: it corresponds to an estimated odds ratio of 1.3 where the true odds ratio was 1.25 or an estimate of 1.5 where the true odds ratio was 1.41. There are more sophisticated methods to perform bias correction for odds ratios but these require an analysis of the original experimental design that is beyond the scope of this paper [22].

While winner's curse probably explains part of the deviation from expected, some classes of diseases were likely not well phenotyped in this study, through some combination of misdiagnosis and misreport. For example, autoimmune diseases are more challenging because they may be of low prevalence, have non-specific symptoms, and a high rate of misdiagnosis. In a study of rheumatoid arthritis diagnoses by non-rheumatologists, 23–82% were judged to be misdiagnoses [23], while another study showed that relative to assessment in a specialist setting, patients in a community setting who received a diagnosis of celiac disease were actually misdiagnosed more than 50% of the time [24]. Some of the underperformance of this approach for autoimmune diseases is

therefore likely due to patients reporting a mistaken diagnosis by a non-specialist.

Because we chose to keep the burden of answering surveys low for our participants, many of the conditions in this study were assessed with single questions such as "Have you ever been diagnosed by a doctor with schizophrenia?" This assessment likely led to reporting errors for some diseases. For example, psychiatric diseases or mental disorders such as Alzheimer's disease, for which diagnosis requires a somewhat subjective clinical evaluation of a patient's symptoms or an autopsy, were each assessed via a single question in this study. More questions are needed here to gather information about the clinical features that led to the diagnosis. In addition, in some cases it may make more sense to have a family member, friend, or caregiver provide information for an individual.

On occasion, the nature of people's answers to such single questions necessitated making judgment calls on how to define a phenotype. Because some people may have type 2 diabetes but are only aware of having high blood sugar, we included people who self-reported having hyperglycemia as type 2 diabetes cases. For chronic obstructive pulmonary disease (COPD), we included individuals who reported having emphysema or chronic bronchitis. However, there are likely to be individuals who repeatedly get bronchitis associated with a cold or flu and reported having "chronic bronchitis", not knowing that the clinical definition of this condition is developing bronchitis lasting at least three months in two consecutive years. This confusion may have reduced our power to replicate associations with COPD. In other cases, we were unable to come up with an acceptable match for a condition. For example, most GWAS of age-related macular degeneration (AMD) have focused on advanced AMD and generally only included cases with large drusen, geographic atrophy, and/or neovascularization. Our question asked only about AMD without assessing severity and thus our study may have included individuals with small or intermediate drusen and/or pigmentary abnormalities as cases. Such phenotypes from the GWAS catalog without direct analogs in our database were skipped for the main calculations in this paper. For all such conditions, more in-depth questions will be necessary to collect data more accurately.

These in-depth questions, which will be important when attempting to unravel the complex biological underpinnings of most phenotypes, can be asked up front for phenotypes that we suspect a priori may be challenging to assess. However, having a recontactable cohort makes the process of refinement possible when more information must be gathered. For celiac disease, starting with the question "Have you ever been diagnosed by a doctor with celiac disease?", we replicated only one association out of almost six expected. As the prevalence of celiac disease in our cohort appeared to be somewhat higher than the reported prevalence in the United States [25], we chose to return to our customer database with a refined question of "Have you ever been diagnosed with celiac disease, as confirmed by a biopsy of the small intestine? If your diagnosis was not confirmed by a biopsy, please select no." As a result, with a much smaller number of cases (which also reduced the number of associations we expected to replicate), we successfully replicated 4 out of 4.5 associations expected for celiac disease. This approach could also be used to examine endophenotypes or to divide broad phenotypes into subclasses with more defined characteristics.

The trend in GWAS research has been towards ever increasing sample sizes and reuse of previously genotyped cohorts whenever possible. Because it is relatively straightforward for our participants to provide information that is relevant for a variety of studies, any given individual can be a case or a control in multiple analyses at once. This could potentially reduce the total amount of work for the patient (sample collection needs to occur only once to participate in many studies) as well as potentially reducing the total number of people an investigator needs to genotype. In addition, for most conditions, this framework leads to a much larger number of controls than cases, which increases the study's power up to a certain point. Though self-report may lead to a slight increase in phenotyping error, in many cases, the lower phenotyping cost may lead to a more powerful study. For example, a study with 3,000 cases and 3,000 controls and a phenotyping error rate of 5% would have 77% power to detect a SNP at a minor allele frequency of 30% and an odds ratio of 1.3 with a p-value threshold of 10^{-7}. But a study with 5,000 cases and 5,000 controls with a phenotyping error rate of 10% would have 95% power to detect such an association. Even if the error rate were 15%, the 10,000 person study would have 77% power and would have many more people to follow up with. Although more data are needed to evaluate the true costs of this model relative to other models, we believe that this method has the potential to collect high-quality phenotype data in an efficient manner.

The framework described here, in which additional questions can be directed at participants at any time with relatively low marginal effort, facilitates follow-up on specific topics as shown in the celiac disease example. Thus, one possible model for large-scale phenotyping could start with broad but shallow phenotyping by self-report on a very large cohort of individuals, followed by targeted recontact of specific subsets of individuals for deeper phenotyping based on the initial information gathered. The additional phenotyping could involve more in-depth questions to the participants or a completely different type of data collection that may require an in-person visit. A platform like this one that maintains an ongoing relationship with the participants, including sharing data with them, may motivate individuals to participate and stay active in research (for example, more than 80% of our research participants have taken more than one research survey).

There are many benefits to having a large, recontactable cohort. Testing new hypotheses, following up on initial data, and assessing the accuracy of different risk prediction models are easier when the need to assemble a new cohort every time is obviated. This raises the question, how large of a cohort is needed? With 20,000 generally unselected people, we expected to replicate approximately 40% of the associations that we tested. Only a $10\times$ increase to 200,000 individuals would raise the expected proportion of replications to 80%, and with a million the expected replication rate would be more than 97%. A simple sum of the initial sample sizes in the papers reported in the GWAS catalog totals nearly 1,400,000. This is clearly an overestimate of the number of genotyped individuals as certain cohorts are reported in more than one study, but even if only 70% of these individuals are unique, this would constitute a resource of a million individuals with genome-wide genotype data who may be interested in participating in further research if given the opportunity. Unfortunately, because of the way research is currently done, these individuals come from dozens of different cohorts and it would be impractical if not impossible to recontact them all. As we move into studies that require ever larger sample sizes, such as those investigating gene—gene or gene—environment interactions, developing more efficient methods of conducting this type of research will become a necessity. We believe that this model in which investigators maintain long-term relationships with research participants and facilitate their participation through online tools is a significant step in that direction.

Methods

Cohort, Genotyping, and Phenotyping

Participants of European ancestry were drawn from the customer base of 23andMe. This group is almost 58% male, with an average age of 46 (approximately 95% of the group is between the ages of 20 and 80). Most of our participants are from the United States, with the next largest groups from Canada and Europe. Genotyping was performed on the Illumina Human-Hap550+ BeadChip and all SNPs tested had a call rate of at least 99%. All individuals provided informed consent and answered surveys online according to our human subjects protocol, which was reviewed and approved by Independent Review Consulting, now part of Ethical & Independent Review Services, a private institutional review board (http://www.eandireview.com). A number of the surveys were based on existing instruments in the literature; the remainder was developed by 23andMe scientists. In general, the new questionnaires were designed in collaboration with a medical professional and reviewed by an external scientific group. All surveys were accessible to customers who had logged in to their 23andMe accounts on a page labeled "Research Surveys". Each survey was labeled with a descriptive title and a link to a short explanation of the content and purpose of the survey. Surveys could be taken in any order and at any time. Because of this, the response rate (number of people who answered the survey divided by all genotyped users who had consented to participate in research, which is likely greater than the number of people who ever viewed the survey) varied from survey to survey but was generally in the range of 15–40%. The "Your Medical History" survey, from which the majority of the phenotype data used in the study were collected, had a response rate of 39%. We used data in our analysis that were collected prior to October 20, 2010.

Identification of SNPs for replication studies

The GWAS catalog is a list of genome-wide association studies curated by the National Human Genome Research Institute's Office of Population Genomics [12,13], and is a relatively complete catalog of SNPs found in published GWAS. We accessed the catalog on May 10, 2010 and sought to replicate as many SNP associations as possible from that list. We removed any SNPs from the list with reported p-values greater than 10^{-7} to limit the number of false positives we were attempting to replicate. In addition, we required that the entry had a reported odds ratio or regression coefficient and that the associations were to single SNPs, rather than haplotypes. For this analysis, as most GWAS are performed in populations of European descent, we restricted our attention to those associations reported in European populations to maximize the total number of associations we could test. We removed duplicate associations from the list, attempting to use the study with the largest total number of cases. To further avoid testing the same association twice, for SNPs that were in LD with each other (using a threshold of $r^2 >= 0.1$), we only picked one association, again attempting to use the study with the largest total number of cases. Where papers reported multiple SNPs in LD with each other, we chose the SNP with the smallest p-value. In cases where we did not have the reported SNP on our platform or where the SNP was not called in over 99% of our subjects, we used a proxy SNP if there was one with $r^2 >= 0.5$. We did not use results reported from papers that included the 23andMe database. Original data from the GWAS catalog can be found for all attempted replications with well-matching phenotypes in Table S7, and for all successful replications with less strictly matching phenotypes in Table S8.

Not all papers used the same stranding conventions, and some papers have misreported the risk allele. Therefore, we checked the stranding of the reported associations using a multi-step process. First, we confirmed that the CEU HapMap frequency information roughly matched the risk allele frequency reported in controls for all SNPs. Specifically, if the reported risk allele frequency and the HapMap frequency were both less than 0.35 or greater than 0.65, the frequency was judged to match. Unambiguous SNPs (i.e., SNPs whose two alleles are not reverse complements) with matching frequencies were judged to be correctly reported. All ambiguous SNPs and SNPs without HapMap data were checked manually in the original papers. This process turned up at least one SNP whose risk allele could not be determined from the original data (rs6457620 with rheumatoid arthritis, not replicated here).

Finally, we required that we could define cases in essentially the same manner as the original paper (using self-reported data for clinical data, where applicable). For example, we restricted our test to people in our database reporting disease onset before the age of 18 for associations to juvenile onset conditions. We also attempted to match smoking status and sex when applicable. In several cases the matching of diseases was a judgment call (for example, self-reports of hyperglycemia and type 2 diabetes were both coded as type 2 diabetes cases). For several common diseases, in an attempt to maximize power, we restricted our set of controls using incidence data for the disease. Specifically, we required controls to be at an age advanced enough that 90% of the controls would be expected never to develop the disease. See Methods S1 and Table S6 for how phenotypes were defined.

Power calculations

We calculated power only for binary traits, using the model from Freidlin et al. [26], modified to calculate power under a one-sided test and to allow for phenotyping error (specified as the percentage of cases incorrectly classified as controls, and vice versa). We set error rates at 5% for cases, based on general evidence that misdiagnosis rates are often over 5%. For example, misdiagnosis rates have been estimated to be 30–45% for celiac disease [24], 5% for multiple sclerosis [27], between 23% and 82% for rheumatoid arthritis diagnosed by a non-rheumatologist [23], and even for cancer with biopsy there are 1.4% discrepant diagnoses when comparing the original diagnosis with a second opinion [28]. We took the error rate for controls to be the minimum of the disease prevalence and 10%. For associations where we used a proxy SNP not in complete linkage disequilibrium with the original reported SNP, the total sample size was scaled by r^2 in the power calculation [29].

Statistical methods

We calculated the p-values for binary associations using the score test for a logistic regression (also known as the Armitage test). Odds ratio (OR) and effect sizes are specified for the risk allele reported in the GWAS catalog [12,13]. For non-binary traits, we used the Wald test for a linear regression. We used a threshold of 0.05 for significance of any individual test. All tests were one-sided in the direction of the published OR. Using different thresholds did not change the results substantially (Table S5). There is no substantial multiple testing burden in this study, as the vast majority of the associations are probably true signals.

For the prediction intervals in Figure 2, we used a model in which each attempted replication was considered to be an independent Bernoulli event with success probability equal to our estimated power for replicating that association. Using a dynamic programming recurrence, we explicitly computed the

probability distribution over the total number of successful replications based on this model. We then determined a 95% prediction interval [L, U] for the total number of successful replications by finding the largest L such that the probability of observing fewer than L replications (or analogously, the smallest U such that the probability of observing greater than U replications) was at most 2.5%. Finally, we determined the reported prediction intervals by dividing these lower and upper bounds by the expected total number of successful replications.

To test whether having seen personal risk estimates for a disease had an effect on self-report of that disease, we looked for an interaction between reported disease risk and whether the individual had possibly seen their report before answering the question (Table S4). This was possible as many people filled out surveys before their results became available. More precisely, we regressed reported phenotype on predicted risk, a "results available" indicator variable, the interaction of these two variables, and age, sex and five principal components of ancestry, and tested the interaction term for significance.

Supporting Information

Figure S1 Success rate (versus total power) by disease. Replications = number of associations we successfully replicated. Expected = number of associations we expected to replicate. Attempts = number of associations we attempted to replicate. The blue dot represents our success ratio (number of successful replications divided by number of expected replications). The black line represents the 95% prediction interval for the success ratio. For ovarian cancer, the success ratio is 12.8 (not within the scale of the graph). (DOCX)

Methods S1 Survey text. Unless otherwise noted, if multiple questions were asked, any subject who answered positively (bolded answer) to at least one was included as a case. Controls were those who answered negatively to these questions. Individuals who gave neither an affirmative nor a negative reply (such as "I'm not sure" or "Decline to state") were not included in the analysis. A subject who answered questions inconsistently (for example, yes to one breast cancer question, no to another) was removed from that analysis. Questions marked with "RS" were asked as Research Snippets, which are questions that are asked singly, as opposed to being part of a larger survey. For additional parameters, see Table S6. (DOCX)

Table S1 All binary replications attempted. Risk = risk allele for original SNP. Chr = chromosome. log(OR) = 23andMe log odds ratio. Pub log(OR) = Published log odds ratio. Rep = replicated. (DOCX)

Table S2 All quantitative replications attempted. Risk = risk allele for original SNP. Size = number of participants. Chr = chromosome. Beta = 23andMe beta (effect size). Rep = replicated. (DOCX)

Table S3 All successful replications without strictly matching phenotypes. Phenotype = 23andMe phenotype. Published = published phenotype. OR = 23andMe odds ratio. [a] Note that this replication is in the opposite direction from that reported in the GWAS catalog. However, a close reading of the original report shows that the direction of the effect was misreported in the GWAS catalog. (DOCX)

Table S4 Effect of viewing genetic risk data on reported disease status. Because one aspect of 23andMe's Personal Genome Service involves returning genetic data to our customers, it is possible that viewing a result of elevated risk for a certain disease may make it more likely for an individual to recall a previous diagnosis of that disease, thus potentially skewing the results towards replication. To address this, we investigated the impact of seeing genetic risk results before versus after answering survey questions on self-reported disease status for a set of 20 conditions for which participants were able to view a personal risk prediction. Only for psoriasis was there a statistically significant impact of seeing one's results on self-report of disease status, and this impact was no longer observed once the direction of the estimated risk (increased or decreased) was taken into account, suggesting that in general, the nature of the genetic risk result did not have a consistent or significant effect on the way questions were answered. Estimated Risk = p-value for association of estimated risk with reported disease status. We expect to see an association with any risk model that is reasonably predictive. Saw Data First = p-value for association of viewing genetic risk results before answering survey questions with reported disease status. Estimated Risk * Saw Data First = p-value for association of interaction between estimated risk and viewing this risk before answering survey questions with reported disease status. (DOCX)

Table S5 Success rate by replication p-value threshold. Alpha = p-value threshold for replication. Replications = number of associations successfully replicated. Expected = number of associations we expected to replicate. Ratio = expected / replications. These calculations do not include the nine associations for which our power may be substantially overestimated. (DOCX)

Table S6 Additional parameters for phenotype classification. (DOCX)

Table S7 Data from GWAS catalog for associations with strictly matching phenotypes. PMID: PubMed ID. RAF = risk allele frequency. OR/Beta = odds ratio or beta (effect size). CI = confidence interval. (DOCX)

Table S8 Data from GWAS catalog for successful associations without strictly matching phenotypes. PMID: PubMed ID. RAF = risk allele frequency. OR/Beta = odds ratio or beta (effect size). CI = confidence interval. (DOCX)

Author Contributions

Conceived and designed the experiments: JYT CBD DAH AKK JMM ABC UF BTN JLM AW NE. Performed the experiments: JYT CBD NE. Analyzed the data: JYT CBD NE. Contributed reagents/materials/analysis tools: JYT CBD AKK JMM ABC BTN NE. Wrote the paper: JYT CBD NE.

References

1. Bilder RM, Sabb FW, Cannon TD, London ED, Jentsch JD, et al. (2009) Phenomics: the systematic study of phenotypes on a genome-wide scale. Neuroscience 164: 30–42. doi:10.1016/j.neuroscience.2009.01.027.

2. Houle D, Govindaraju DR, Omholt S (2010) Phenomics: the next challenge. Nat Rev Genet 11: 855–866. doi:10.1038/nrg2897.

3. Lee K, Sawcer S (2010) Detecting genes in complex disease: does phenotype accuracy limit the horizon? Trends Genet 26: 241–242; author reply 242–243. doi:10.1016/j.tig.2010.03.003.

4. Okura Y, Urban LH, Mahoney DW, Jacobsen SJ, Rodeheffer RJ (2004) Agreement between self-report questionnaires and medical record data was substantial for diabetes, hypertension, myocardial infarction and stroke but not for heart failure. J Clin Epidemiol 57: 1096–1103. doi:10.1016/j.jclinepi.2004.04.005.

5. Smith B, Chu LK, Smith TC, Amoroso PJ, Boyko EJ, et al. (2008) Challenges of self-reported medical conditions and electronic medical records among members of a large military cohort. BMC Med Res Methodol 8: 37. doi:10.1186/1471-2288-8-37.

6. O'Mahony PG, Dobson R, Rodgers H, James OF, Thomson RG (1995) Validation of a population screening questionnaire to assess prevalence of stroke. Stroke 26: 1334–1337.

7. Walker MK, Whincup PH, Shaper AG, Lennon LT, Thomson AG (1998) Validation of patient recall of doctor-diagnosed heart attack and stroke: a postal questionnaire and record review comparison. Am J Epidemiol 148: 355–361.

8. Martin LM, Leff M, Calonge N, Garrett C, Nelson DE (2000) Validation of self-reported chronic conditions and health services in a managed care population. Am J Prev Med 18: 215–218.

9. Colditz GA, Martin P, Stampfer MJ, Willett WC, Sampson L, et al. (1986) Validation of questionnaire information on risk factors and disease outcomes in a prospective cohort study of women. Am J Epidemiol 123: 894–900.

10. Burgess AM, Martel MU, Wyman DK (1971) Validation of interview-based disease classifications: a mail survey of physicians. J Chronic Dis 24: 45–59.

11. Eriksson N, Macpherson JM, Tung JY, Hon LS, Naughton B, et al. (2010) Web-based, participant-driven studies yield novel genetic associations for common traits. PLoS Genet 6: e1000993. doi:10.1371/journal.pgen.1000993.

12. Genome.gov | A Catalog of Published Genome-Wide Association Studies (n.d.) Available: http://www.genome.gov/gwastudies/. Accessed 17 Nov 2010.

13. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106: 9362–9367. doi:10.1073/pnas.0903103106.

14. Latiano A, Palmieri O, Valvano MR, D'Incà R, Cucchiara S, et al. (2008) Replication of interleukin 23 receptor and autophagy-related 16-like 1 association in adult- and pediatric-onset inflammatory bowel disease in Italy. World J Gastroenterol 14: 4643–4651.

15. Illumina .::. Investor Relations 2011 Press Releases (n.d.) Available: http://investor.illumina.com/phoenix.zhtml?c = 121127&p = irol-news&nyo = 0. Accessed 28 Jun 2011.

16. Pelchat ML, Bykowski C, Duke FF, Reed DR (2011) Excretion and perception of a characteristic odor in urine after asparagus ingestion: a psychophysical and genetic study. Chem Senses 36: 9–17. doi:10.1093/chemse/bjq081.

17. Medland SE, Nyholt DR, Painter JN, McEvoy BP, McRae AF, et al. (2009) Common variants in the trichohyalin gene are associated with straight hair in Europeans. Am J Hum Genet 85: 750–755. doi:10.1016/j.ajhg.2009.10.009.

18. Do CB, Tung JY, Dorfman E, Kiefer AK, Drabant EM, et al. (2011) Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease. PLoS Genetics;In press.

19. The International Parkinson's Disease Genetics Consortium (2011) A two-stage meta-analysis identifies several new loci for Parkinson's disease. PLoS Genetics;In press.

20. Ritchie MD, Denny JC, Crawford DC, Ramirez AH, Weiner JB, et al. (2010) Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. Am J Hum Genet 86: 560–572. doi:10.1016/j.ajhg.2010.03.003.

21. Kraft P, Zeggini E, Ioannidis JPA (2009) Replication in genome-wide association studies. Stat Sci 24: 561–573. doi:10.1214/09-STS290.

22. Xiao R, Boehnke M (2009) Quantifying and correcting for the winner's curse in genetic association studies. Genet Epidemiol 33: 453–462. doi:10.1002/gepi.20398.

23. Feldman DE, Bernatsky S, Haggerty J, Leffondré K, Tousignant P, et al. (2007) Delay in consultation with specialists for persons with suspected new-onset rheumatoid arthritis: a population-based study. Arthritis Rheum 57: 1419–1425. doi:10.1002/art.23086.

24. Pinto Sánchez MI, Smecuol E, Vázquez H, Mazure R, Mauriño E, et al. (2009) Very high rate of misdiagnosis of celiac disease in clinical practice. Acta Gastroenterol Latinoam 39: 250–253.

25. Fasano A, Berti I, Gerarduzzi T, Not T, Colletti RB, et al. (2003) Prevalence of celiac disease in at-risk and not-at-risk groups in the United States: a large multicenter study. Arch Intern Med 163: 286–292.

26. Freidlin B, Zheng G, Li Z, Gastwirth JL (2002) Trend tests for case-control studies of genetic markers: power, sample size and robustness. Hum Hered 53: 146–152.

27. Gasperini C (2001) Differential diagnosis in multiple sclerosis. Neurol Sci 22 Suppl 2: S93–97.

28. Kronz JD, Westra WH, Epstein JI (1999) Mandatory second opinion surgical pathology at a large referral hospital. Cancer 86: 2426–2435.

29. Pritchard JK, Przeworski M (2001) Linkage disequilibrium in humans: models and data. Am J Hum Genet 69: 1–14. doi:10.1086/321275.

30. Bierut LJ, Agrawal A, Bucholz KK, Doheny KF, Laurie C, et al. (2010) A genome-wide association study of alcohol dependence. Proc Natl Acad Sci U S A 107: 5082–5087. doi:10.1073/pnas.0911109107.

31. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls (2007) Nature 447: 661–678. doi:10.1038/nature05911.

32. Sklar P, Smoller JW, Fan J, Ferreira MAR, Perlis RH, et al. (2008) Whole-genome association study of bipolar disorder. Mol Psychiatry 13: 558–569. doi:10.1038/sj.mp.4002151.

33. Ferreira MAR, O'Donovan MC, Meng YA, Jones IR, Ruderfer DM, et al. (2008) Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. Nat Genet 40: 1056–1058. doi:10.1038/ng.209.

34. Scott LJ, Muglia P, Kong XQ, Guan W, Flickinger M, et al. (2009) Genome-wide association and meta-analysis of bipolar disorder in individuals of European ancestry. Proc Natl Acad Sci U S A 106: 7501–7506. doi:10.1073/pnas.0813386106.

35. Pillai SG, Ge D, Zhu G, Kong X, Shianna KV, et al. (2009) A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. PLoS Genet 5: e1000421. doi:10.1371/journal.pgen.1000421.

36. Van Durme YMTA, Eijgelsheim M, Joos GF, Hofman A, Uitterlinden AG, et al. (2010) Hedgehog-interacting protein is a COPD susceptibility gene: the Rotterdam Study. Eur Respir J 36: 89–95. doi:10.1183/09031936.00129509.

37. Pillai SG, Kong X, Edwards LD, Cho M, Anderson WH, et al. (2010) Loci Identified by Genome-wide Association Studies Influence Different Disease-related Phenotypes in COPD. Am J Respir Crit Care Med;Available: http://www.ncbi.nlm.nih.gov/pubmed/20656943. Accessed 17 Nov 2010.

38. McGovern DPB, Jones MR, Taylor KD, Marciante K, Yan X, et al. (2010) Fucosyltransferase 2 (FUT2) non-secretor status is associated with Crohn's disease. Hum Mol Genet 19: 3468–3476. doi:10.1093/hmg/ddq248.

39. Rioux JD, Xavier RJ, Taylor KD, Silverberg MS, Goyette P, et al. (2007) Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. Nat Genet 39: 596–604. doi:10.1038/ng2032.

40. Moffatt MF, Gut IG, Demenais F, Strachan DP, Bouzigon E, et al. (2010) A large-scale, consortium-based genomewide association study of asthma. N Engl J Med 363: 1211–1221. doi:10.1056/NEJMoa0906312.

41. Simón-Sánchez J, Schulte C, Bras JM, Sharma M, Gibbs JR, et al. (2009) Genome-wide association study reveals genetic risk underlying Parkinson's disease. Nat Genet 41: 1308–1312. doi:10.1038/ng.487.

42. Chang M, Li Y, Yan C, Callis-Duffin KP, Matsunami N, et al. (2008) Variants in the 5q31 cytokine gene cluster are associated with psoriasis. Genes Immun 9: 176–181. doi:10.1038/sj.gene.6364451.

43. Nair RP, Duffin KC, Helms C, Ding J, Stuart PE, et al. (2009) Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. Nat Genet 41: 199–204. doi:10.1038/ng.311.

44. Perdigones N, Lamas JR, Vigo AG, de la Concha EG, Jover JA, et al. (2009) 6q23 polymorphisms in rheumatoid arthritis Spanish patients. Rheumatology (Oxford) 48: 618–621. doi:10.1093/rheumatology/kep053.

45. Patsopoulos NA, Ioannidis JPA (2010) Susceptibility variants for rheumatoid arthritis in the TRAF1-C5 and 6q23 loci: a meta-analysis. Ann Rheum Dis 69: 561–566. doi:10.1136/ard.2009.109447.

46. Ioannidis JPA, Patsopoulos NA, Evangelou E (2007) Heterogeneity in meta-analyses of genome-wide association investigations. PLoS ONE 2: e841. doi:10.1371/journal.pone.0000841.

47. Takahashi M, Saenko VA, Rogounovitch TI, Kawaguchi T, Drozd VM, et al. (2010) The FOXE1 locus is a major genetic determinant for radiation-related thyroid carcinoma in Chernobyl. Hum Mol Genet 19: 2516–2523. doi:10.1093/hmg/ddq123.

48. Barrett JC, Lee JC, Lees CW, Prescott NJ, Anderson CA, et al. (2009) Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. Nat Genet 41: 1330–1334. doi:10.1038/ng.483.