

Genes for Good: engaging the public in genetics research using social media

Katharine Brieger,^{1,2,0} Gregory J.M. Zajac,^{2,0} Anita Pandit,^{2,0,*} Johanna R. Foerster,² Kevin W. Li,² Aubrey C. Annis,² Ellen M. Schmidt,^{2,3} Chris P. Clark,² Karly McMorro, ² Wei Zhou,⁴ Jingjing Yang,⁵ Alan M. Kwong,² Andrew P. Boughton,² Jinxi Wu,⁶ Chris Scheller,² Tanvi Parikh,⁷ Alejandro de la Vega,⁷ David M. Brazel,^{7,8} Maia Frieser,⁷ Gianna Rea-Sandin,⁹ Lars G. Fritsche,² Scott I. Vrieze,¹⁰ Gonçalo R. Abecasis^{2,*}

⁰The authors contributed equally as first authors.

* These authors are the corresponding authors.

¹Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI, 48109, United States of America

²Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI, 48109, United States of America

³Wellcome Sanger Institute, Hinxton, CB10 1SA, United Kingdom

⁴Department of Bioinformatics, University of Michigan School of Medicine, Ann Arbor, MI, 48109, United States of America

⁵Department of Human Genetics, Emory School of Medicine, Atlanta, GA, 30322, United States of America

⁶School of Information, University of Michigan, Ann Arbor, MI, 48109, United States of America

⁷Institute for Behavioral Genetics, University of Colorado Boulder, Boulder, CO, 80309, United States of America

⁸Department of Molecular, Cellular, and Developmental Biology, Boulder, CO, 80309, United States of America

⁹Department of Psychology, Arizona State University, Tempe, AZ, 85281, United States of America

¹⁰Department of Psychology, University of Minnesota, Minneapolis, MN, 55455, United States of America

1 **Abstract**

2 The Genes for Good study uses social media to engage a large, diverse participant pool in genetics
3 research and education. Health history and daily tracking surveys are administered through a Facebook
4 application, and participants who complete a minimum number of surveys are mailed a saliva kit to collect
5 DNA for genotyping. As of March 2019, we engaged >80,000 individuals, sent spit kits to >32,000 individuals
6 who met minimum participation requirements, and collected >27,000 spit kits. Participants come from all fifty
7 states and include a diversity of ancestral backgrounds. Rates of important chronic health indicators are
8 consistent with those estimated for the general U.S. population using more traditional study designs. However,
9 our sample is younger and contains a greater percentage of females than the general population. As one means
10 of verifying data quality, we have replicated genome-wide association studies (GWAS) for exemplar traits, such
11 as asthma, diabetes, body mass index (BMI), and pigmentation. The flexible framework of the web application
12 makes it relatively simple to add new questionnaires and for other researchers to collaborate. We anticipate that
13 the study sample will continue to grow and that future analyses may further capitalize on the strengths of the
14 longitudinal data in combination with genetic information.

15

16 **Introduction**

17 More than 10,000 genetic loci have been successfully linked to common and complex diseases¹. In
18 previous decades, the major challenge for human genetic studies was the cost and complexity of the genotyping
19 itself; however, researchers now face the bigger hurdle of obtaining large enough samples that also include
20 useful, linked medical and health data. The study designs typically used to collect such data are expensive and
21 often exclude individuals based on location or demographics. We reasoned that using social media platforms
22 would not only allow us to recruit a large population cohort, but also help us to reach populations that might not
23 typically participate in genetic studies due to the time commitment or distance to a research center. Potential
24 advantages of social media based study designs include the ability to reach diverse populations and the ability to
25 engage participants in research over time. Potential concerns include representativeness and the ability of this
26 approach to reproduce findings obtained using more traditional designs.

We present a new study design to take advantage of recent developments in health survey methods using social media and widespread interest in direct to consumer genetic testing^{2; 3}. Genes for Good is an ongoing, large-scale study of health, genetic, and behavioral information. We aim to engage tens of thousands of individuals in research through a Facebook application, reducing the expense of traditional epidemiologic designs and the exclusivity and high socioeconomic status associated with current direct to consumer efforts⁴.

Our model of using social media for genetic research invites participants to complete online health assessments at their convenience, as has been successfully applied in numerous studies of health, behavior⁵, and psychology⁶, including rare genetic diseases⁷, childbirth preferences⁸, and prediction of personality traits⁶. When a consenting participant has completed a minimum number of health history and health tracking surveys, they are mailed a spit kit to collect DNA for analysis. After genotyping, we test genetic variants for association with health, disease, and environmental information collected through online assessments.

In this paper, we demonstrate that the Genes for Good study model is a viable complement to more traditional research study designs. The phenotypic and genotypic data we have collected thus far appear valid and reliable. Further, the incentive structure of Genes for Good – namely, altruism combined with the return of survey response summaries and genetic data to participants – is effective, as demonstrated by exponential recruitment from all fifty U. S. states. Importantly, the recruitment happened organically, with participants publicizing the study through their own networks, without relying on paid advertising. We briefly explored the use of study recruitment websites (such as ResearchMatch⁹), but only several hundred participants were recruited this way. We also saw large influxes of participants after online articles in Reddit¹⁰ and BuzzFeed¹¹. While resources still go toward answering questions about the study and resolving technical issues, efficient participant recruitment and engagement allowed us to dedicate a larger fraction of resources to sample collection, processing and downstream analyses. The long-term goals of the study fall broadly into five main categories: (a) to identify novel genetic loci associated with a variety of phenotypes, (b) to longitudinally track an array of health and behavioral measures, (c) to enable genotype-first study designs (such as detailed phenotypic assessments of participants with naturally-occurring knockout variants), (d) to educate participants and make the data available to them, and (e) to encourage data sharing among researchers. Here, we present our

1 study design and methods, as well as initial findings about our sample demographics and important health
2 indicators.

3 One particular advantage of hosting our study on social media is that we can reach participants in an
4 environment that many already visit regularly as part of their daily routines. Social media use in the U.S. has
5 dramatically increased in the last decade – rising from 7% in 2005 to over 65% in 2015¹² – and so we have the
6 potential to reach a majority of the U.S. population through our application. In the last few years, several
7 research groups have recognized the major advantages social media offers: flexible timing, the possibility of
8 incentives and reminders, and the ability to reach non-urban communities. There has already been substantial
9 success in recruiting for studies via Facebook¹³ as well as in using it to prevent loss-to-follow-up¹⁴. Further, the
10 flexible framework of Genes for Good allows us and other research groups to continue adding new surveys and
11 activities to address future research questions. Our study takes advantage of the opportunity for repeated contact
12 that social media offers and represents the first large genetic study of tens of thousands of individuals conducted
13 via Facebook.

14 Considering their ubiquity and ease of use, social media and mobile devices as research tools are
15 important avenues to explore further¹⁵. However, we recognize some of the potential disadvantages we are
16 likely to face: (a) inaccurate data, (b) low response rate¹⁶, (c) high attrition, and (d) a sample limited to those
17 who have a Facebook account. In the first year of the study, we prioritized testing and combatting several of
18 these expected limitations. With the aforementioned challenges in mind, we implemented various methods to
19 assess the quality of our data. First, we looked at common diseases and phenotypes to validate our results – and
20 thus our approach to data collection – by comparing them to prior findings from traditional research and meta-
21 analysis designs. When expected phenotypic relationships hold true, such as that between BMI and Type 2
22 diabetes, we gain confidence in the quality of the survey responses we are collecting. Additionally, we assessed
23 the quality of the genetic data by replicating findings from genome wide association studies (GWAS) for a
24 variety of traits that are known to have genetic components, such as diabetes, asthma, BMI, hair color, and eye
25 color, confirming that our data yields the expected signals. We also examined rates of chronic health

1 conditions, such as hypertension and diabetes, to explore how our study participants compare to the overall U.S.
2 population.

3 **Material and Methods**

4 We have implemented a large, IRB-approved genetic study using social media. Participants must be at
5 least 18 years old, live in the U.S., and have a Facebook account. They are recruited via snowball sampling, i.e.
6 by finding our Genes for Good Facebook application through friends, family, and social media connections.
7 Once a person has consented, they are invited to complete online health history assessments at their
8 convenience. The surveys consist of health history questionnaires, daily tracking surveys, and an optional health
9 conditions module in which participants can list other conditions that they have. Once they have completed a
10 minimum number of required questionnaires, they are mailed a spit kit to collect DNA for analysis. The cost of
11 each participant is about \$80, which includes postage, DNA extraction, and genotyping; there is essentially no
12 cost associated with recruitment or data collection. Throughout the course of the study, we have typically
13 employed 2-3 full-time staff (study coordinator, developers), several graduate and undergraduate students, and a
14 part-time administrative assistant to assist with sending and receiving spit kits.

15

16 *Genetic Analysis*

17 DNA is genotyped at ~600,000 SNPs using either the Illumina Infinium CoreExome-24 v1.0 or v1.1
18 arrays for nonsynonymous exonic variants and a panel of common genome-wide markers⁴¹. The standard set of
19 markers on the array is augmented with missense, loss of function, and potential lipids and myocardial
20 infarction variants identified in the HUNT whole genome sequencing and whole exome sequencing projects⁴²;
21 height-associated variants from GIANT⁴³; potential stop-gain variants in 96 genes at loci potentially implicated
22 in type 2 diabetes, blood lipid levels, Alzheimer's disease, nicotine/alcohol metabolism, and several others with
23 mutations implicated in serious but treatable health conditions; complex trait associated variants in the
24 EBI/NHGRI GWAS catalog¹; a random subset of Neanderthal SNPs from the 1000 Genomes Project⁴⁴; ancestry
25 informative markers identified by Paschou et al. that were highly correlated with the principal components of
26 Human Genome Diversity Project samples⁴⁵; and pain related variants proposed by Dr. Chad Brummett of the

1 University of Michigan Division of Pain Research. Genotypes at an additional >30 million variants in the 1000
2 Genomes Phase 3 panel⁴⁶ are imputed using Minimac3⁴⁷. After quality control, local genetic ancestry is
3 estimated using RFMix⁴⁸, global ancestry with ADMIXTURE⁴⁹, and principal components analysis performed
4 with TRACE⁵⁰, using the Human Genome Diversity Project samples as a reference panel⁵¹ for all three
5 analyses. We provide each Genes for Good participant with a section in the app to view these estimates of
6 genetic ancestry on the sample they provided.

7 For the GWAS of Genes for Good participants' BMI, the BMI measurements were calculated from the
8 Height and Weight survey in the app, which was derived from height and weight questionnaires available from
9 PhenX Toolkit⁵². Weight measurements for the first several thousand genotyped participants were bottom-
10 coded at 80 lbs and top-coded at 251 lbs; then, the top-coded value was changed to 381 lbs partway through the
11 study to capture a greater range of variation. For participants that were pregnant at the time of answering the
12 survey, we used their pre-pregnancy weight obtained from the same survey. The BMI values were then
13 regressed on sex, age, array chip version, and the first five principal components; the residuals were inverse-
14 normal transformed in order to compare effect size estimates to the largest published meta-analysis of BMI⁵³
15 and to reduce the impact of extreme observations. We used the SAIGE software⁵⁴ to run a mixed model GWAS,
16 accounting for sample relatedness and population structure. Polygenic risk scores were calculated using
17 PLINK⁵⁵.

18
19 *Participant engagement*

20 We provide participants with several ways to interact with both their own data and the research study
21 as a whole. After each health history survey is completed, we provide charts summarizing the information, in
22 some cases comparing each participant's answers to the Genes for Good study population (example in Figure
23 7). Similarly, for daily tracking surveys, we generate summaries of each participant's health behavior over time
24 as well as summary statistics for the entire study (example in Figure 8). In addition to providing this ongoing
25 feedback and summary of the survey responses, we also offer participants who submit a sample a breakdown of
26 their genetic ancestry; the current version includes 7 continental human populations (Europe, Africa, East Asia,

Central/South Asia, West Asia/North Africa, Americas, and Oceania), and results are served in the form of a global ancestry estimate, local ancestry inference, and principal components analysis using the methods described previously (RFMIX, ADMIXTURE, TRACE). Before seeing their estimates of genetic ancestry, they are required to watch a short video on how to interpret their results. Participants can also download their array and imputed genotypes.

Privacy and data security

All Genes for Good data is divided into two classes: (a) personally identifiable information, such as email addresses, Facebook user IDs, and physical mailing addresses; and (b) research information, such as survey responses and genetic data. Each class of data is stored in a distinct relational database and served from a distinct server. Extracts for outside researchers include only research-specific data. We plan to ask participants to allow use of their mailing address to link to information such as geocode pollution, built environment (for instance, the number of fast food outlets or public parks within a certain radius of one's home), and census tract data. In these cases, the participants' physical address would still be withheld from external collaborators, but variables generated using addresses could be shared upon request.

The privacy of Genes for Good data is monitored by the University of Michigan Institutional Review Board. All genetic and survey results are stored in a secure server on campus that is not directly connected to the public internet, and DNA samples are stored in physically secure spaces with restricted access. In addition, all archived data is de-identified to protect subject privacy including participants' demographic summary and genetic information. Even though Genes for Good uses Facebook to authenticate login, Facebook does not access information we collect through the App and we do not use participant's social media postings and connections in our research. We make efforts to communicate with participants about the extensive measures we take in ensuring the privacy of their data and to ease their worries about using social media as a platform for genetic research.

All communication to and from the application is encrypted. Participants are authenticated using a Facebook account and Facebook's OAuth implementation, ensuring that participants only have access to their

own data once inside the application. Communication with Facebook servers is limited to authentication only; although Facebook receives and retains information about which Facebook accounts have accessed the Genes for Good app, all other information provided by participants is provided directly to Genes for Good servers. Facebook cannot see any of the data entered by participants.

Once participants have their genetic data analyzed, they are notified that they may access results inside the app with a Results Access Code, a randomly generated alphanumeric code that must be requested by the participant and will be delivered to the email address on the participant's Genes for Good profile. Participant genotype data is processed internally on University of Michigan servers and is distributed to participants upon request via Box, a secure third-party file-sharing platform. Participants may request their raw genotypes as often as they like from within the genetic results section of the app. Each request compresses and uploads raw genotype data and supplementary information to a private, password-protected Box account directory. For security purposes, all requested genotypes automatically expire from Box servers three days after being uploaded.

Results

Since the launch of Genes for Good on January 19th, 2015 (Martin Luther King Jr. Day), we have seen steadily increasing participant recruitment and consistent use of the Facebook application. Genes for Good now has enough participants to begin conducting meaningful analyses with the data. As of March 2019, 117,652 participants had tried the app, with 81,110 who signed the electronic consent form. Consenting users have completed over 2.9 million surveys, answering >22 million questions. Genes for Good has mailed 33,427 spit kits to eligible participants, of which 27,470 have been returned (as of March 2019). The genetic data freeze used for this paper contains data from 20,232 participants whose genotypes passed quality control checks as of mid-2018.

Sample characteristics and phenotypes

Participants were recruited successfully from all fifty states, with areas of peak participant density roughly overlapping with major U.S. population centers (Figure 1). About 90% of users have residential addresses outside of Michigan. Compared to the U.S. population, our sample is younger (Genes for Good median age of 33, U.S. adult median age of 44) and enriched for females (74% of participants are women, compared to 51% for US adults, Table 1). Our sample also closely resembles the U.S. population on household income, although it is enriched for individuals from middle-income households with an annual income of \$35,000 - \$100,000; Table 3). In contrast, the majority of the participants in the research cohort from 23andMe are from households with an annual income over \$100,000¹⁷. To confirm the quality of the data collected from our sample, we also compared disease rates to those in the general U.S. population (Table 2). In looking at important risk factors for cardiovascular disease, we observed relatively similar rates of high cholesterol, hypertension, and smoking. However, our sample had lower rates of disease outcomes such as stroke and myocardial infarction. Our genotype data freeze contained 20,232 individuals, of which 76.3% were non-Hispanic white, 3.8% Asian, 2.7% African American, 8.8% multi-racial/other, and 8.3% Hispanic/Latino as determined by self-report through our Demographics survey.

In addition to the phenotype information collected from survey responses, 12,216 participants have reported 64,401 cases of 3,067 health conditions in an optional section of the app that allows participants to search for and report disorders using the Systematized Nomenclature of Medicine (SNOMED) dictionary¹⁸. These participant-entered data show that Genes for Good has attracted an unusually high proportion of individuals with certain rare diseases, like Ehlers-Danlos Syndrome (565 cases or 0.93% of GfG participants compared to ~0.02% prevalence worldwide)¹⁹. The 5 most commonly reported disorders were generalized anxiety disorder (1,803 cases), asthma (1,389), hypothyroidism (941), depressive disorder (920), and migraine (918). Higher BMI was associated with increased risk for all 5 conditions in logistic regression of each of the five traits on BMI, sex, and age (odds ratios of 1.02, 1.03, 1.04, 1.01, 1.03 per unit higher BMI, p-values of 7.6×10^{-9} , 2.1×10^{-20} , 3.9×10^{-24} , 1.5×10^{-4} , 6.3×10^{-14}).

To evaluate the quality of our data, we used our survey data to verify known phenotypic relationships. Taking diabetes as an example, we analyzed the association of the disease with BMI. Given the rapidly

1 increasing prevalence of diabetes in the U.S., this is a particularly important outcome to examine. Over the past
2 three decades, the number of diagnosed Americans has more than tripled, from 5.6 million in 1980 to 21 million
3 in 2012²⁰. And because about one-third of diabetics are undiagnosed, national survey statistics consistently
4 underestimate the true prevalence of diabetes²⁰. We compared rates of diabetes in our sample, within each BMI
5 bracket, to those reported from nationally representative samples²¹ and found a similar trend of increasing
6 diabetes prevalence as BMI increased (Figure 2). We further explored this relationship by calculating the
7 estimated effect of BMI on diabetes status, adjusting for age, sex, and race, using NHANES and Genes for
8 Good data separately. We found that the relationship between BMI and diabetes was comparable between
9 studies (95% CI for odds ratio per 1-unit increase in BMI, NHANES: 1.07-1.10; 95% CI, GFG: 1.08-1.10).
10 When comparing simple correlation coefficients between BMI and diabetes status across studies, we found no
11 notable difference between Genes for Good and NHANES ($r_{GFG}=0.18$, $r_{NHANES}=0.19$, $p = 0.83$). Though our
12 sample is quite different from NHANES in terms of wealth, age distribution, and ethnic diversity, we observe
13 similar trends in both cohorts when comparing diabetes cases and controls: diabetics typically have higher rates
14 of obesity, higher age, lower income, and lower education (Table S1 in supplemental data).

15
16 *Genetic associations*

17 To validate the quality of our self-reported phenotypes, we analyzed a data freeze of 20,232 genotypes
18 to see if we could replicate known genetic associations. We first analyzed traits related to pigmentation and
19 BMI, because these traits are known to have strong genetic factors. For example, most variation in eye color is
20 determined by 6 SNPs in *HERC2* and *OCA2*²². Figure 3 shows the number of participants with each
21 combination of eye color and genotype at one of the SNPs with the strongest association signal, rs12913832.
22 We observed strong evidence of association between eye color and genotype ($\chi^2 = 15,599$, $df = 8$, $p = 10^{-3376}$,
23 $N=19,974$), and the direction of effects is consistent with what was previously reported. Other pigmentation
24 traits like hair color, skin sun response, and hair texture are also consistent with prior studies. Table S2 shows
25 detailed GWAS results, and Table S3 compares our results to several larger studies. We show that Genes for

1 Good replicates the top pigmentation associations in prior studies at least nominally ($p < 0.05$), and frequently
2 does so at genome-wide significance ($p < 5 \times 10^{-8}$).

3 We next compared results for a mixed model GWAS of BMI, using measurements obtained from the
4 Height and Weight health history survey, to results from the GIANT consortium²³. We obtained effect sizes
5 consistent with those published for the top ten GIANT loci. We also obtained nominally significant ($p < 0.05$)
6 association results at all 10 loci. Figure 4 summarizes the comparison of our results with published GIANT
7 results, showing consistency of direction of effect, magnitude, and relative significance (Figure 5 shows
8 regional association in our top signal, at FTO). Given the relatively small sample size of our data, our effect
9 estimates necessarily have wider confidence limits compared to the meta-analysis. However, the meta-analysis
10 point estimates are contained within these limits for nearly every SNP, which provide evidence that self-
11 reported phenotypes collected within our cohort are reliable.

12 We next expanded our comparison of GWAS results obtained with Genes for Good data to include the
13 traits of type 1 diabetes, type 2 diabetes, and asthma. For all traits except asthma, our association signals are
14 consistent with reports from published large GWAS and show some significant hits (Tables S2, S3, and S4;
15 Figure S1). Our asthma analysis did not give any genome-wide significant results, but when we examined the
16 eighteen SNPs associated with asthma in the study of Demenais et al.²⁴ we found that all had a consistent
17 direction of effect in Genes for Good data but with smaller effect sizes (Table S4). Our asthma cases and
18 controls were defined based on answers to “Was your asthma ever confirmed by a doctor?” with 4,378 cases
19 and 11,715 controls reported. Given the large proportion of cases (27.2%), we believe that some individuals
20 who answered “Yes” did not meet the standard for an asthma diagnosis used in Demenais et al.²⁴ A similar
21 observation has been made in other studies of self-reported phenotypes — for example, in a study of psoriasis
22 including data from 23andMe customers, it was estimated that only ~36% of individuals who self-reported
23 having psoriasis met the criteria used in clinical studies, diluting association signals and effect size estimates²⁵.
24 We did an adjustment proposed by Duffy et al. (2004) to account for the apparent over-reporting of cases²⁶. We
25 also did a power calculation at the 0.05 significance level to determine our ability to replicate the findings in
26 Demenais et al. and estimated that we should replicate approximately 7 of 18 SNPs (summing estimated power

across eighteen variants gives expected number of 6.8 replicated signals). After the Duffy adjustment over half of our odds ratios were closer to the effect sizes reported in Demenais et al., though some odds ratios were overcorrected to have effect sizes larger than those reported in Demenais et al. As our power calculation suggested, we were able to replicate 7 of the 18 SNPs at the 0.05 significance level (Table S4)^{24; 25}. Reassuringly, we also found that, when we calculated polygenic risk scores (PRS) for type 1 and type 2 diabetes using publicly available GWAS summary statistics^{27; 28}, PRS for type 2 diabetes was strongly associated with self-reported type 2 diabetes status (OR increase per PRS quintile=1.47; $p=7.63 \times 10^{-37}$) and that PRS for type 1 diabetes PRS was strongly associated with self-reported type 1 diabetes status (OR increase per PRS quintile=1.66; $p=5.13 \times 10^{-9}$) (Figure 6). We found similar support for an association between asthma PRS and self-reported asthma (OR increase per PRS quintile=1.16; $p=3.17 \times 10^{-26}$) (Figure 6).

Somewhat unexpectedly, we observed that in our type 2 diabetes results the signal at *CDKAL1* was stronger than at *TCF7L2*, which is typically the top signal reported for type 2 diabetes GWAS. Hypothesizing that this might be due to the younger age of Genes for Good participants, we split the Genes for Good data at the median age to test for changes in diabetes risk between the below-median age and above-median age groups for the *TCF7L2* and *CDKAL1* variants (median age = 32; cases_{Below-Median} = 65, controls_{Below-Median} = 8,385; cases_{Above-Median} = 722, controls_{Above-Median} = 7,728). Although we saw a trend to a larger diabetes risk for carriers of the *TCF7L2* variant rs7903146 in the above-median group (OR_{Below-Median} = 1.21, OR_{Above-Median} = 1.34), we saw the same trend for carriers of the *CDKAL1* variant rs7756992 (OR_{Below-Median} = 1.04, OR_{Above-Median} = 1.37). Regardless, the differences between the below-median and above-median age groups for both SNPs were not significant ($p > 0.05$).

Discussion

We set out to recruit a large, diverse sample of engaged volunteers that might provide information about the diverse U.S. population. For each volunteer, we used surveys to collect health and behavioral data that might inform a variety of genomic research studies. With rapid and inexpensive recruitment, we have quickly developed a participant pool with which to validate the quality of the data. We are optimistic about our ability to obtain the large sample size required for valid genetic association studies of complex diseases and behaviors.

1 With our current analysis of 20,232 individuals, we have successfully validated several known genotype-
2 phenotype relationships and contributed to several consortium meta-analyses²⁹⁻³².

3 We have good representation with respect to geography, age, and gender, though our sample does have
4 some noticeable differences from a sample of random U.S. adults. One characteristic that presents both an
5 opportunity and a challenge is the younger age of Genes for Good participants compared to the U.S. adult
6 population. While a younger demographic may be more interesting for some measures (behavioral data, activity
7 levels), it will be less useful for others (age-associated cancers and development of other late-onset chronic
8 disease). We do see slightly lower rates of the chronic conditions examined here compared to the general U.S.
9 population, which we attribute to the lower average age of our participants; even if participants have the
10 relevant risk factors, they may not have had the time to develop those long-term outcomes. For instance, we see
11 much lower rates of heart attack in our participants despite comparable hypertension rates, and we see lower
12 rates of type 2 diabetes despite comparable BMI (Figure 2). At the same time, Genes for Good's recruitment
13 strategy may have led to an enrichment of individuals with certain rare diseases like Ehlers-Danlos Syndrome,
14 perhaps because of network effects within these communities.

15 Most participants completed the minimum number of health history surveys required to receive a spit kit
16 (15 surveys), with many going well above that number. Completion of daily tracking surveys was modest, with
17 most genotyped participants completing only minimum number required to obtain a spit kit. None of our
18 surveys are mandatory and it is certainly possible that participants will avoid surveys that are more onerous or
19 which they are not comfortable with, introducing ascertainment biases (for example, individuals who are not
20 skilled at reasoning puzzles might choose to skip the reasoning). The most completed surveys were generally
21 those that appear higher in the list of available surveys within our app (Figure S2; Figure S3 provides additional
22 details of survey completion rates).

23 Another challenge we face is that our sample is heavily skewed female. While targeted recruitment in
24 the future may bring the gender distribution into balance, we also recognize the immediate potential to conduct
25 a large-scale study of women's health and have implemented relevant survey measures regarding polycystic
26 ovarian syndrome and pregnancy outcomes.

Genetic information, privacy, and ethics

There are a number of incentives for participation in Genes for Good besides the altruistic contribution and potential positive impact of genetics research on society. Firstly, we provide interactive graphs and visualizations by which users can compare their survey responses to those of other participants (examples in Figures 7 and 8). Secondly, Genes for Good allows participants to view estimates of their genetic ancestry and download their raw genetic data, which some have argued should be the fundamental right of participants who contribute DNA to research²⁸. When downloading genetic data, we require participants to review a short slide show that explains the data we generate is suitable for a research study but does not meet the standards used for clinical genetic tests. We emphasize that, compared to the data used in clinical tests, research data might be more susceptible to error. Around 70% of participants with genotypes available have requested a download link for their raw genetic data, which we provide in 23andMe format, a format known to be widely accepted at third-party interpretation sites. Many participants have told us they upload their data to third-party sites to obtain more detailed ancestry estimates, find DNA relatives, and even seek health interpretation. A recent review paper³³ investigating reactions to a clinical genetic risk assessment concluded that in general, patients do not engage in risk-reducing behavior after receiving information about genetic predisposition. We expect that Genes for Good participants are unlikely to base major health or life decisions on the research-grade data we have returned. In addition, we will continue to develop Genes for Good web-based software applications to promote literacy of individuals about their genetic information.

Along with raw genetic data, we also return to participants their genetic ancestry information based on DNA analysis. The primary anticipated risk of the return of ancestry information is the discovery or suspicion of non-paternity and/or secret adoption by participants, i.e. discovering one's ancestry is inconsistent with what the participant knows about the ancestry of their supposedly biological parents. This has the potential to cause emotional or psychological stress on participants and their families, and we provide education about this risk during the informed consent.

1 *Significance and future directions*

2 The online platform implemented in Genes for Good is a viable study design for population-based
3 genetic research. Now in the study's fourth year, we have already had great success in recruitment, health
4 history survey analysis, and genetic analysis. We are currently exploring the more than 300 phenotypes
5 collected so far and continue to participate in ongoing collaborations. As the sample size grows, our power to
6 detect novel associations and our ability to contribute more meaningful data to researchers will increase.

7 The flexibility of the study design and our ongoing relationship with participants also makes it possible
8 to implement new methods of data collection with relative ease. Additional data collection techniques are being
9 developed and validated in a wide array of studies, including wireless sensors for continuous collection of data
10 related to physical activity^{34; 35}, heart rate³⁶, body temperature, sleep³⁷, and GPS location logging to infer habits
11 and environmental exposures³⁸. These measures and more are currently available through a combination of
12 smartphone and wrist sensors (e.g. FitBit), and many more wireless sensors exist for more specialized tasks
13 (e.g. breathalyzers, insulin levels, QT interval). These and other novel data collection methods are developing
14 rapidly, holding great promise in the near future for the efficient collection of large quantities of precise
15 longitudinal data with minimal participant burden. The implementation of such devices would facilitate the
16 collection of tracking data within Genes for Good.

17 Having verified the quality of our data and several known associations with particular loci, we are now
18 poised to begin exploring new genotypic-phenotypic relationships, such as those with behavioral and health
19 tracking information. Research in other settings with Genes for Good data show that our results are consistent
20 with those of prior studies. Liu et al.³⁹ show that a PRS calculated from SSGAC's educational attainment data is
21 effective in predicting 4% of the trait variance, which is consistent with previously reported out-of-sample
22 predictive power for educational attainment⁴⁰. We are also working to streamline data sharing methods to
23 facilitate collaborations with other researchers. Finally, we are actively developing new tools to provide
24 participants with meaningful data summaries at the personal and study level. We believe these steps will keep
25 participants engaged and invested in the genetic research and will also help encourage longitudinal survey
26 completions.

As we seek opportunities for long term funding of the study, we are currently not collecting spit kits from new participants. Although enrollment has decreased since we stopped offering spit kits (we currently collect only health survey responses), interest remains high, as evidenced by the email inquiries we receive on a weekly basis. We plan to collect and genotype additional samples when future funding becomes available; when doing so, we expect to implement several changes to study protocol that will solve issues observed throughout the course of the study. For example, we noticed that survey completion correlates with the order that the survey appears on the app homepage (Figure S2); something as simple as randomizing the order upon refresh may remedy this.

Supplemental Data

Supplemental data contain four tables and three figures.

Conflicts of Interest

Gonçalo R Abecasis is currently an employee of Regeneron Pharmaceuticals and the beneficiary of stock options and grants in Regeneron. Previously, he served on scientific advisory boards for 23andMe, Regeneron Pharmaceuticals and Helix.

Acknowledgements

This research has been conducted using the UK Biobank Resource under application number 24460 (specifically, calculation of PRS for type 1 diabetes and asthma was conducted using GWAS results from UK Biobank).

Administrative Support: Irene Felicetti, Stephanie Bachoura, Samantha Bachoura, Laura Baker

IT Support: Sean Caron

UM Sequencing Core: Robert Lyons, Susan Dagenais, Christopher Krebs, David Erdody

Web Resources

Genes for Good Facebook application: <https://apps.facebook.com/genesforgood>

Genes for Good informational website: <http://www.genesforgood.org>

Full text of all Genes for Good survey: http://genesforgood.sph.umich.edu/for_researchers

Information on Box compliance with HIPAA guidelines: <https://www.box.com/industries/healthcare>

References

1. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 42, D1001-1006.
2. Stoekle, H.C., Mamzer-Bruneel, M.F., Vogt, G., and Herve, C. (2016). 23andMe: a new two-sided data-banking market model. *BMC medical ethics* 17, 9.
3. Royal, C.D., Novembre, J., Fullerton, S.M., Goldstein, D.B., Long, J.C., Bamshad, M.J., and Clark, A.G. (2010). Inferring genetic ancestry: opportunities, challenges, and implications. *American Journal of Human Genetics* 86, 661-673.
4. Agurs-Collins, T., Ferrer, R., Ottenbacher, A., Waters, E.A., O'Connell, M.E., and Hamilton, J.G. (2015). Public Awareness of Direct-to-Consumer Genetic Tests: Findings from the 2013 U.S. Health Information National Trends Survey. *Journal of cancer education : the official journal of the American Association for Cancer Education* 30, 799-807.
5. Pedersen, E.R., and Kurz, J. (2016). Using Facebook for Health-related Research Study Recruitment and Program Delivery. *Current opinion in psychology* 9, 38-43.
6. Kosinski, M., Matz, S.C., Gosling, S.D., Popov, V., and Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *The American psychologist* 70, 543-556.
7. Abiad, J.E., Robbins, S., Morris, C., and Sobreira, M. (2018). Survey of Patients with Ollier Disease and Maffucci Syndrome Over Facebook Compared to Review of Clinical Literature (Abstract #9). Platform talk presented at the 2018 ACMG Annual Clinical Genetics Meeting, April 10-14, 2018, Charlotte, NC.
8. Arcia, A. (2014). Facebook Advertisements for Inexpensive Participant Recruitment Among Women in Early Pregnancy. *Health education & behavior : the official publication of the Society for Public Health Education* 41, 237-241.
9. Harris, P.A., Scott, K.W., Lebo, L., Hassan, N., Lightner, C., and Pulley, J. (2012). ResearchMatch: a national registry to recruit volunteers for clinical research. *Acad Med* 87, 66-73.
10. (2017). Free DNA Test from the University of Michigan. Reddit r/freebies, https://www.reddit.com/r/freebies/comments/67v69c65/free_dna_test_from_the_university_of_michigan/.
11. Hughes, V. (2015). A New Facebook App Wants To Test Your DNA. BuzzFeed News, <https://www.buzzfeed.com/virginiahughes/a-new-facebook-app-wants-to-test-your-dna>.
12. Perrin, A. (2015). Social Networking Usage: 2005-2015. Pew Research Center, <http://www.pewinternet.org/2015/2010/2008/social-networking-usage-2005-2015>.
13. Fenner, Y., Garland, S.M., Moore, E.E., Jayasinghe, Y., Fletcher, A., Tabrizi, S.N., Gunasekaran, B., and Wark, J.D. (2012). Web-based recruiting for health research using a social networking site: an exploratory study. *J Med Internet Res* 14, e20.
14. Mychasiuk, R., and Benzies, K. (2012). Facebook: an effective tool for participant retention in longitudinal research. *Child Care Health Dev* 38, 753-756.
15. Steinhubl, S.R., Muse, E.D., and Topol, E.J. (2015). The emerging field of mobile health. *Science translational medicine* 7, 283rv283.
16. Kapp, J.M., Peters, C., and Oliver, D.P. (2013). Research recruitment using Facebook advertising: big potential, big challenges. *J Cancer Educ* 28, 134-137.
17. Tung, J.Y., Eriksson, N., Kiefer, A.K., Macpherson, J.M., Naughton, B.T., Chowdry, A.B., Do, C.B., Hinds, D.A., Wojcicki, A., and Mountain, J.L. (2011). Characteristics of an Online Consumer Genetic Research Cohort (Abstract #914T). Poster presented at the 61st Annual Meeting of The American Society of Human Genetics, October 11-15, 2011, Montreal, Canada.
18. Lee, D., Cornet, R., Lau, F., and de Keizer, N. (2013). A survey of SNOMED CT implementations. *J Biomed Inform* 46, 87-96.
19. Levy, H.P. (2018). Hypermobility Ehlers-Danlos Syndrome. In *GeneReviews®*, M.P. Adam, H.H. Ardinger, R.A. Pagon, and S.E. Wallace, eds. (Seattle, WA, University of Washington), p <https://www.ncbi.nlm.nih.gov/books/NBK1279/>.

20. CDC. (2014). National Diabetes Statistics Report: Estimates of Diabetes and Its Burden in the United States, 2014. Department of Health and Human Services, <https://www.cdc.gov/diabetes/pubs/statsreport14/national-diabetes-report-web.pdf>.
21. Bays, H.E., Chapman, R.H., Grundy, S., and Group, S.I. (2007). The relationship of body mass index to diabetes mellitus, hypertension and dyslipidaemia: comparison of data from two national surveys. *Int J Clin Pract* 61, 737-747.
22. Liu, F., van Duijn, K., Vingerling, J.R., Hofman, A., Uitterlinden, A.G., Janssens, A.C., and Kayser, M. (2009). Eye color and the prediction of complex phenotypes from genotypes. *Curr Biol* 19, 192.
23. Locke, A.E., Kahali, B., Berndt, S.I., Justice, A.E., Pers, T.H., Day, F.R., Powell, C., Vedantam, S., Buchkovich, M.L., Yang, J., et al. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature* 518, 197-206.
24. Demenais, F., Margaritte-Jeannin, P., Barnes, K.C., Cookson, W.O.C., Altmüller, J., Ang, W., Barr, R.G., Beaty, T.H., Becker, A.B., Beilby, J., et al. (2018). Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. *Nat Genet* 50, 42-53.
25. Tsoi, L.C., Stuart, P.E., Tian, C., Gudjonsson, J.E., Das, S., Zawistowski, M., Ellinghaus, E., Barker, J.N., Chandran, V., Dand, N., et al. (2017). Large scale meta-analysis characterizes genetic architecture for common psoriasis associated variants. *Nat Commun* 8, 15382.
26. Duffy, S.W., Warwick, J., Williams, A.R.W., Keshavarz, H., Kaffashian, F., Rohan, T.E., Nili, F., and Sadeghi-Hassanabadi, A. (2004). A simple model for potential use with a misclassified binary outcome in epidemiology. *Journal of Epidemiology and Community Health* 58, 712-717.
27. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562, 203-209.
28. Lunshof, J.E., Church, G.M., and Prainsack, B. (2014). Information access. Raw personal data: providing access. *Science (New York, NY)* 343, 373.
29. Jiang, Y., Chen, S., McGuire, D., Chen, F., Liu, M., Iacono, W.G., Hewitt, J.K., Hokanson, J.E., Krauter, K., Laakso, M., et al. (2018). Proper conditional analysis in the presence of missing data: Application to large scale meta-analysis of tobacco use phenotypes. *PLoS Genet* 14, e1007452.
30. Zhan, X., Chen, S., Jiang, Y., Liu, M., Iacono, W.G., Hewitt, J.K., Hokanson, J.E., Krauter, K., Laakso, M., Li, K.W., et al. (2017). Association Analysis and Meta-Analysis of Multi-allelic Variants for Large Scale Sequence Data. *bioRxiv*, 197913.
31. Liu, M., Jiang, Y., Wedow, R., Li, Y., Brazel, D.M., Chen, F., Datta, G., Davila-Velderrain, J., McGuire, D., Tian, C., et al. (2019). Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat Genet* 51, 237-244.
32. Sanchez-Roige, S., Fontanillas, P., Elson, S.L., the 23andMe Research, T., Pandit, A., Schmidt, E.M., Foerster, J.R., Abecasis, G.R., Gray, J.C., de Wit, H., et al. (2017). Genome-wide association study of delay discounting in 23,217 adult research participants of European ancestry. *Nature Neuroscience*.
33. Hollands, G.J., French, D.P., Griffin, S.J., Prevost, A.T., Sutton, S., King, S., and Marteau, T.M. (2016). The impact of communicating genetic risks of disease on risk-reducing health behaviour: systematic review with meta-analysis. *BMJ* 352, i1102.
34. Dobkin, B.H., and Dorsch, A. (2011). The promise of mHealth: daily activity monitoring and outcome assessments by wearable sensors. *Neurorehabil Neural Repair* 25, 788-798.
35. Appelboom, G., Camacho, E., Abraham, M.E., Bruce, S.S., Dumont, E.L., Zacharia, B.E., D'Amico, R., Slomian, J., Reginster, J.Y., Bruyere, O., et al. (2014). Smart wearable body sensors for patient self-assessment and monitoring. *Arch Public Health* 72, 28.
36. El-Amrawy, F., and Nounou, M.I. (2015). Are Currently Available Wearable Devices for Activity Tracking and Heart Rate Monitoring Accurate, Precise, and Medically Beneficial? *Healthc Inform Res* 21, 315-320.
37. Montgomery-Downs, H., Insana, S.P., and Bond, J.A. (2012). Movement toward a novel activity monitoring device. *Sleep Breath* 16, 913-917.
38. Glasgow, M.L., Rudra, C.B., Yoo, E.H., Demirbas, M., Merriman, J., Nayak, P., Crabtree-Ide, C., Szpiro, A.A., Rudra, A., Wactawski-Wende, J., et al. (2016). Using smartphones to collect time-activity data for long-term personal-level air pollution exposure assessment. *J Expo Sci Environ Epidemiol* 26, 356-364.

39. Liu, M., Rea-Sandin, G., Foerster, J., Fritsche, L., Brieger, K., Clark, C., Li, K., Pandit, A., Zajac, G., Abecasis, G.R., et al. (2017). Validating Online Measures of Cognitive Ability in Genes for Good, a Genetic Study of Health and Behavior. *Assessment*, 1073191117744048.
40. Branigan, A.R., McCallum, K.J., and Freese, J. (2013). Variation in the Heritability of Educational Attainment: An International Meta-Analysis. *Social Forces* 92, 109-140.
41. Illumina. (2017). Infinium® CoreExome- 24 v1.2 BeadChip. https://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet_human_core_exome_beadchip.pdf.
42. Krokstad, S., Langhammer, A., Hveem, K., Holmen, T.L., Midthjell, K., Stene, T.R., Bratberg, G., Heggland, J., and Holmen, J. (2013). Cohort Profile: the HUNT Study, Norway. *International journal of epidemiology* 42, 968-977.
43. Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan, J., Kutalik, Z., et al. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* 46, 1173-1186.
44. Sankararaman, S., Mallick, S., Dannemann, M., Prufer, K., Kelso, J., Paabo, S., Patterson, N., and Reich, D. (2014). The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* 507, 354-357.
45. Paschou, P., Lewis, J., Javed, A., and Drineas, P. (2010). Ancestry informative markers for fine-scale individual assignment to worldwide populations. *Journal of medical genetics* 47, 835-847.
46. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. *Nature* 526, 68-74.
47. Das, S., Forer, L., Schonherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and methods. *Nature genetics* 48, 1284-1287.
48. Maples, B.K., Gravel, S., Kenny, E.E., and Bustamante, C.D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am J Hum Genet* 93, 278-288.
49. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 19, 1655-1664.
50. Wang, C., Zhan, X., Liang, L., Abecasis, G.R., and Lin, X. (2015). Improved ancestry estimation for both genotyping and sequencing data using projection procrustes analysis and genotype imputation. *Am J Hum Genet* 96, 926-937.
51. Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L., et al. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319, 1100-1104.
52. Hamilton, C.M., Strader, L.C., Pratt, J.G., Maiese, D., Hendershot, T., Kwok, R.K., Hammond, J.A., Huggins, W., Jackman, D., Pan, H., et al. (2011). The PhenX Toolkit: get the most from your measures. *American Journal of Epidemiology* 174, 253-260.
53. Locke, A.E., Kahali, B., Berndt, S.I., Justice, A.E., Pers, T.H., Day, F.R., Powell, C., Vedantam, S., Buchkovich, M.L., Yang, J., et al. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature* 518, 197-206.
54. Zhou, W., Nielsen, J.B., Fritsche, L.G., Dey, R., Gabrielsen, M.E., Wolford, B.N., LeFaive, J., VandeHaar, P., Gagliano, S.A., Gifford, A., et al. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet* 50, 1335-1341.
55. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4, 7.
56. Locke, A.E., Kahali, B., Berndt, S.I., Justice, A.E., Pers, T.H., Day, F.R., Powell, C., Vedantam, S., Buchkovich, M.L., Yang, J., et al. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature* 518, 197-206.
57. Howden, L.M., and Meyer, J.A. (2011). Age and Sex Composition: 2010. In *Census Briefs*, C2010BR-03. (Washington, D.C., U.S. Census Bureau), pp 1-16 <https://www.census.gov/prod/cen2010/briefs/c2010br-2003.pdf>.
58. (2017). Distribution of Facebook users in the United States as of January 2017, by age group and gender. We Are Social, <https://www.statista.com/statistics/187041/us-user-age-distribution-on-facebook/>.
59. eMarketer, and Squarespace. (2017). Number of Facebook users in the United States as of January 2017, by age group (in millions). Statista, <https://www.statista.com/statistics/398136/us-facebook-user-age-groups/>.
60. CDC, and NCHS. (2017). National Health and Nutrition Examination Survey Data 2015-2016. Hyattsville, MD: US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, <https://wwwn.cdc.gov/Nchs/Nhanes/Search/DataPage.aspx?Component=Examination&CycleBeginYear=2015>.

61. Nwankwo, T., Yoon, S.S., Burt, V., and Gu, Q. (2013). Hypertension among adults in the United States: National Health and Nutrition Examination Survey, 2011-2012. NCHS Data Brief, 1-8.
62. Mozaffarian, D., Benjamin, E.J., Go, A.S., Arnett, D.K., Blaha, M.J., Cushman, M., de Ferranti, S., Despres, J.P., Fullerton, H.J., Howard, V.J., et al. (2015). Heart disease and stroke statistics--2015 update: a report from the American Heart Association. *Circulation* 131, e29-322.
63. Ward, B.W., Clarke, T.C., Nugent, C.N., and Schiller, J.S. (2016). Early Release of Selected Estimates Based on Data From the 2015 National Health Interview Survey. National Center for Health Statistics, May 2016 <https://www.cdc.gov/nchs/data/nhis/earlyrelease/earlyrelease201605.pdf>.
64. Semega, J.L., Fontenot, K.R., and Kollar, M.A. (2017). Households by Total Money Income, Race, and Hispanic Origin of Householder: 1967 to 2016. In US Census Bureau, Current Population Reports, P60-259, Income and Poverty in the United States: 2016. (Washington, DC, U.S. Government Printing Office), pp 23-29.

Figure Titles and Legends

Figure 1. Geographic Distribution. The geographic distribution of Genes for Good participants as of October 2017. The colors indicate the number of participants who have logged into the app from that county, with darker colors representing higher density.

Figure 2. Relationship between BMI and diabetes rates in participants is consistent with that seen in the general U.S. population. Type 2 diabetes is a phenotype of particular interest because of its increasing prevalence, impact on cardiovascular health, and relatively well-characterized genetics. Here, we have compared the rates of diabetes in Genes for Good participants to the rates found in the nationally representative studies SHIELD and NHANES²¹.

*Imputed variant

Figure 4. Effect size estimates of a GWAS for BMI in our study sample compared to findings from a meta-analysis. We compare effect estimates from Genes for Good to published findings from the Locke et al 2015 meta-analysis of BMI GWAS⁵⁶. Specifically, we looked at the top 10 reported signals and were able to replicate all of these effects in direction and nominal significance ($p < 0.05$). The forest plot on the right compares effect size estimates across studies; the dashed lines represent the confidence intervals around the Genes for Good estimates, while the solid lines represent results from Locke et al. Given the relatively small sample size available in this data freeze, our estimates have fairly wide confidence limits. However, Locke's estimates are completely contained within our limits for 8 of 10 SNPs.

Figure 3. Distribution of eye color among participants with different genotypes at rs12913832 (the top signal when performing GWAS using blue eye color in Genes for Good participants), a marker in the *HERC2* gene known to play a role in eye color determination.

Figure 5. LocusZoom plot showing single-variant association results for BMI in the gene *FTO*. This result is consistent with other studies that reported their strongest evidence for association in this gene. The effect size at the nearby SNP rs1558902 (0.081) was consistent with the effect size (0.081) reported previously in Locke et al.⁵³.

Figure 6. Prevalence for self-reported Type 1 and Type 2 diabetes across polygenic risk score quintiles (five bins of equal sample size). An increase in the genetic risk score is associated with increasing prevalence of disease. We also evaluated associations between polygenic risk score quintile and Type 1 diabetes, Type 2 diabetes, and asthma status, adjusted for age and sex. We found that all three self-reported traits were significantly associated with calculated PRS quintile ($p_{T1D}=5.13 \times 10^{-9}$, $p_{T2D}=7.63 \times 10^{-37}$, $p_{asthma}=3.17 \times 10^{-26}$).

Figure 7. Example Health History result. An example of how participants' results to the Personality survey are displayed within the Genes for Good app. The bars show this participant's percentile scores on the five personality attributes measured by the survey.

Figure 8. Example daily tracking result. An example of how participants’ answers to the daily sleep tracking survey are displayed, showing (A) average hours of sleep for this participant, compared to other participants of the same age range and sex, and to all other Genes for Good participants, (B) average hours of sleep reported for different days of the week when this participant has taken the survey, (C) average hours of sleep over the past 7 days, past 30 days, and over all responses from this participant, and (D) average hours of sleep reported for different days of the week for all Genes for Good participants stratified by sex.

Tables

Table 1: Demographics

	Genes for Good ^a	U.S. Population ^b	Facebook-using population ^c
Age			
Median, years	33	44 ^d	
18-24	17.0%	13.2%	19.5%
25-34	37.1%	17.1%	27.0%
35-44	21.6%	16.4%	19.6%
45-54	11.9%	18.3%	16.5%
55+	12.4%	35.5%	17.4%
Sex			
Male	25.9%	49.2%	49%
Female	74.1%	50.8%	51%

^a Data source for our study data is based on all valid responses as of August 9th, 2017

^b Data for U.S. population from the 2010 U.S. Census⁵⁷

^c Data for Facebook population from Statistica^{58; 59}

^d Median age of U.S. persons over age 18 reported in the U.S. 2010 Census

Table 2: Chronic health indicators in study sample compared to overall United States population

	Genes for Good ^a	U.S. Population ^b
BMI, mean, kg/m ²	29.80	29.38
Underweight (BMI < 18.5)	1.9%	1.6%
Normal weight (BMI 18.5 - 24.9)	31.6%	27.2%
Overweight (BMI 25 - 29.9)	26.0%	31.6%
Obese (BMI ≥ 30)	40.4%	39.7%
Chronic Health Indicators		
High cholesterol	26.1%	29.3%
Hypertension	24.9%	29%
Previous stroke	1.3%	2.9%
Previous MI	1.5%	4.5%
Diabetes (Type 1 or 2)	6.5%	9.3%
Current smoker	17.0%	15.1%

^a Data source for our study data is based on all valid responses as of August 9th, 2017

^b Data from nationally representative samples to determine U.S. rates of obesity⁶⁰, high cholesterol, hypertension⁶¹, stroke⁶², MI, diabetes, and smoking⁶³

1 **Table 3: Income distribution**

2
3 Income table
4

Income Category	Genes for Good (%)	US Population ^a (%)	23andMe ^b (%)
Less than \$35,000	28.0	30.2	10.2
\$35,000 to \$50,000	18.9	12.9	7.2
\$50,000 to \$75,000	19.8	17.0	13.9
\$75,000 to \$100,000	14.5	12.3	14.7
More than \$100,000	18.9	27.7	54.0

5
6 Distribution of household income among Genes for Good participants based on answers to the Demographics
7 survey as of August 9, 2017 compared to the general U.S. population.

8
9 ^a Data from U.S. Census Table H-17⁶⁴

10 ^b Data describing 23andMe research cohort approximated from 2011 ASHG poster¹⁷

11
12
13