A public resource facilitating clinical use of genomes

Madeleine P. Ball[a,1], Joseph V. Thakuria[a,b,c,1], Alexander Wait Zaranek[a,c,1], Tom Clegg[c], Abraham M. Rosenbaum[a,d], Xiaodi Wu[a,e], Misha Angrist[f], Jong Bhak[g,h], Jason Bobe[i], Matthew J. Callow[j], Carlos Cano[k], Michael F. Chou[a], Wendy K. Chung[l], Shawn M. Douglas[a], Preston W. Estep[i,m], Athurva Gore[n], Peter Hulick[o], Alberto Labarga[k], Je-Hyuk Lee[a], Jeantine E. Lunshof[p,q], Byung Chul Kim[h], Jong-Il Kim[r,s], Zhe Li[n], Michael F. Murray[t], Geoffrey B. Nilsen[j], Brock A. Peters[j], Anugraha M. Raman[a], Hugh Y. Rienhoff[u], Kimberly Robasky[a,v], Matthew T. Wheeler[w], Ward Vandewege[c], Daniel B. Vorhaus[x], Joyce L. Yang[a], Luhan Yang[a], John Aach[a], Euan A. Ashley[w,y], Radoje Drmanac[j], Seong-Jin Kim[z], Jin Billy Li[a,aa], Leonid Peshkin[bb], Christine E. Seidman[cc], Jeong-Sun Seo[r,dd], Kun Zhang[n], Heidi L. Rehm[ee], and George M. Church[a,2]

[a]Department of Genetics, Harvard Medical School, Boston, MA 02115; [b]Division of Medical Genetics, Massachusetts General Hospital, Boston, MA 02114; [c]Clinical Future Inc., Cambridge, MA 02142; [d]Ion Torrent by Life Technologies, Guilford, CT 06437; [e]Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO 63110; [f]Duke University Institute for Genome Sciences and Policy, Durham, NC 27708-0141; [g]Theragen BiO Institute, TheragenEtex Inc., Suwon, 443-270, Korea; [h]Genomics Department, Personal Genomics Institute, Suwon 443-766, Korea; [i]PersonalGenomes.org, Boston, MA 02215; [j]Complete Genomics, Inc., Mountain View, CA 94043; [k]Department of Computer Science and A.I., University of Granada, 18071 Granada, Spain; [l]Departments of Pediatrics and Medicine, Columbia University, New York, NY 10032; [m]TeloMe, Inc., Waltham, MA 02451; [n]Department of Bioengineering, University of California at San Diego, La Jolla, CA 92093; [o]Division of Genetics, NorthShore University HealthSystem, Evanston, IL 60201; [p]Faculty of Earth and Life Sciences, Department of Molecular Cell Physiology, VU University Amsterdam, 1081 HV Amsterdam, The Netherlands; [q]Faculty of Health, Medicine and Life Sciences, Maastricht University, 6200 MD Maastricht, The Netherlands; [r]Genomic Medicine Institute, Medical Research Center, College of Medicine, Seoul National University, Seoul, Korea; [s]Psoma Therapeutics Inc., Gasan-dong, Kumchun-gu, Seoul 153-781, Korea; [t]Division of Genetics, Brigham and Women's Hospital, Boston, MA 02115; [u]www.MyDaughtersDNA.org, San Carlos, California 94070; [v]Bioinformatics Program, Boston University, Boston, MA 02215; [w]Stanford Center for Inherited Cardiovascular Disease, Stanford University School of Medicine, Stanford, CA 94305; [x]Robinson Bradshaw & Hinson, P.A., Chapel Hill, NC 27517; [y]Personalis, Inc., Palo Alto, CA 94301; [z]Cha Cancer Institute, Cha University of Medicine and Science, Seoul 135-081, Korea; [aa]Department of Genetics, Stanford University, Stanford, CA 94305; [bb]Department of Systems Biology, Harvard Medical School, Boston, MA 02115; [cc]Department of Genetics, Harvard Medical School and Howard Hughes Medical Institute, Boston, MA 02115; [dd]Macrogen, Seoul, Korea; and [ee]Department of Pathology, Harvard Medical School, Boston, MA 02115

This article is part of the special series of Inaugural Articles by members of the National Academy of Sciences elected in 2011.

Edited by C. Thomas Caskey, Baylor College of Medicine, Houston, TX, and approved June 11, 2012 (received for review February 1, 2012)

Rapid advances in DNA sequencing promise to enable new diagnostics and individualized therapies. Achieving personalized medicine, however, will require extensive research on highly reidentifiable, integrated datasets of genomic and health information. To assist with this, participants in the Personal Genome Project choose to forgo privacy via our institutional review board-approved "open consent" process. The contribution of public data and samples facilitates both scientific discovery and standardization of methods. We present our findings after enrollment of more than 1,800 participants, including whole-genome sequencing of 10 pilot participant genomes (the PGP-10). We introduce the Genome-Environment-Trait Evidence (GET-Evidence) system. This tool automatically processes genomes and prioritizes both published and novel variants for interpretation. In the process of reviewing the presumed healthy PGP-10 genomes, we find numerous literature references implying serious disease. Although it is sometimes impossible to rule out a late-onset effect, stringent evidence requirements can address the high rate of incidental findings. To that end we develop a peer production system for recording and organizing variant evaluations according to standard evidence guidelines, creating a public forum for reaching consensus on interpretation of clinically relevant variants. Genome analysis becomes a two-step process: using a prioritized list to record variant evaluations, then automatically sorting reviewed variants using these annotations. Genome data, health and trait information, participant samples, and variant interpretations are all shared in the public domain—we invite others to review our results using our participant samples and contribute to our interpretations. We offer our public resource and methods to further personalized medical research.

genome interpretation | genomic medicine | human genetics

As whole genome DNA sequencing costs plummet below the cost of standard diagnostic genetic testing, personal genomes promise dramatic changes for science, medicine, and society. A genome sequence can be a clinical diagnostic that lasts a lifetime, and personal genomes for every individual are likely to become standard components of health care. We now face challenging questions: How do we interpret genome data? Can we and should we regulate access to personal genetic data and/or interpretations? Can whole-genome data truly be considered anonymizable—even if not combined with other personal data? How strictly should a promise of privacy made to research subjects limit our ability to scientifically share their data with other researchers? The fact that combined genetic and phenotype data are so personal and reidentifiable creates a tension between standard commitments ensuring research subject privacy and the scientific need for verification and reproducibility of research findings (1).

The Personal Genome Project (PGP) explores one solution to these issues in its creation of a public resource where participants acknowledge and agree to the potential risk of reidentification. This public resource not only shares genome data publicly but brings these together with publicly shared phenotype information, genetic interpretations, and cell lines; such integrated data means the PGP can provide common ground for many types of genome research. Sharing reidentifiable data requires new instruments for informed consent, as participants explicitly waive their expectation of privacy to make personal biological and

Author contributions: M.P.B., J.V.T., A.W.Z., M.A., J. Bobe, M.F.C., S.M.D., P.W.E., J.E.L., D.B.V., H.L.R., and G.M.C. designed research; M.P.B., J.V.T., A.W.Z., T.C., A. M. Rosenbaum, X.W., W.K.C., P.W.E., A. M. Raman, K.R., C.E.S., and H.L.R. performed research; M.P.B., J.V.T., A.W.Z., T.C., A. M. Rosenbaum, X.W., J. Bhak, C.C., A.G., A.L., J.-H.L., B.C.K., Z.L., A. M. Raman, W.V., J.L.Y., L.Y., S.-J.K., J.B.L., L.P., and K.Z. contributed new reagents/analytic tools; M.P.B., J.V.T., A.W.Z., T.C., A. M. Rosenbaum, X.W., M.J.C., P.H., J.-I.K., M.F.M., G.B.N., B.A.P., H.Y.R., K.R., M.T.W., W.V., J.A., E.A.A., R.D., and J.-S.S. analyzed data; and M.P.B., J.V.T., A.W.Z., and G.M.C. wrote the paper.

Conflict of interest statement: G.M.C. has advisory roles in and research sponsorships from several companies involved in genome sequencing technology and personal genomics (http://arep.med.harvard.edu/gmc/tech.html).

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: The sequences reported in this paper are made available through the Personal Genome Project (http://www.personalgenomes.org/data/PGP12.05/).

See Profile on page 11893.

[1]M.P.B., J.V.T., and A.W.Z. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: gchurch@genetics.med.harvard.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1201904109/-/DCSupplemental.

health information public (2). This process, now called "open consent" (3), places a high value on the autonomy of individuals and on their ability to give open-ended consent for unknown risks. Our informed consent materials extensively discuss both risks associated with loss of privacy and the limited options for restoring privacy once data and cell lines are made public.

Although our goal is to have the broadest possible participation in the PGP, because of the novel nature of the risks and research the Committee on Human Studies (Boston, MA) encouraged us to initially enroll individuals with a master's-level degree or equivalent training in genetics. The "PGP-10" pilot group was chosen in 2006 from 10 such individuals who volunteered for the project. These individuals have chosen to publicly associate their names with their PGP accounts—participants may voluntarily self-identify in this way, but this is not required. Samples from these 10 individuals have since been used to pilot a variety of technologies within our groups and others, including whole-genome sequencing, induced pluripotent stem (iPS) cell line generation and genome engineering, allele-specific expression profiling, epigenetic profiling, and microbiome profiling (4–11). These data go beyond the genome sequence itself to create additional layers of information that move into the realm of associated environmental and trait profiling.

Beyond generating an initial public resource of linked genotype and phenotype data, a key goal of our pilot was to develop and prototype methods for interpreting genome information and making these interpretations public. Early versions of our methods have already been used by other groups in their own genome research and interpretations (9, 12–14). Unlike many published genome interpretation efforts, which have focused on discovery of novel pathogenic variants in patients with genetic disease (15–19), this pilot focuses on 10 individuals not believed to have such diseases. As cohorts with heritable medical conditions join the PGP, our research will extend to disease-focused interpretations. Nevertheless, interpretation methods for individuals not suspected of having genetic disease will be essential for integrating genome data into clinical practice as genome sequencing becomes increasingly routine.

Results

More than 1,000 Participants Enrolled Through Open Consent with Public Health Records. The PGP has piloted the use of an open consent format for collection of combined genome and phenotype data, allowing data to be shared publicly. PGP participants must understand and agree to the following: (*i*) any genome and health record data provided to us could be included in an open-access public database, (*ii*) no guarantees are made regarding anonymity, privacy, and confidentiality, (*iii*) participation may involve a risk of harm or privacy loss to themselves and their relatives, (*iv*) participation does not promise to benefit participants in any tangible way, and (*v*) withdrawal from the study is possible at any time, but complete removal of data that have been available in the public domain may not be possible. This process of making data public means that results are also returned to participants, and an ongoing relationship with these participants is maintained to monitor outcomes of participation prospectively.

On the basis of our experiences with the PGP-10, we created an enrollment system for volunteers that ensures they understand the risks entailed (Fig. 1). Volunteers are provided with a study guide to inform them of genetic concepts and privacy risks and are required to pass an entrance examination testing their understanding of human subjects research, PGP protocols, and basic genetics. Of volunteers meeting minimum eligibility criteria, 44% drop out at this step; 87% of those who successfully complete the examination go on to sign the full consent form and enroll in the project (*SI Appendix*, Fig. S1).

The examination and consent form are completed through an Internet-based system and are electronically signed by volunteers; more than 1,800 participants have enrolled through this process as of May 2012. Because participants are a self-selected group, they are not representative of the general population (*SI Appendix*, Fig. S2); however, we may prioritize participants from

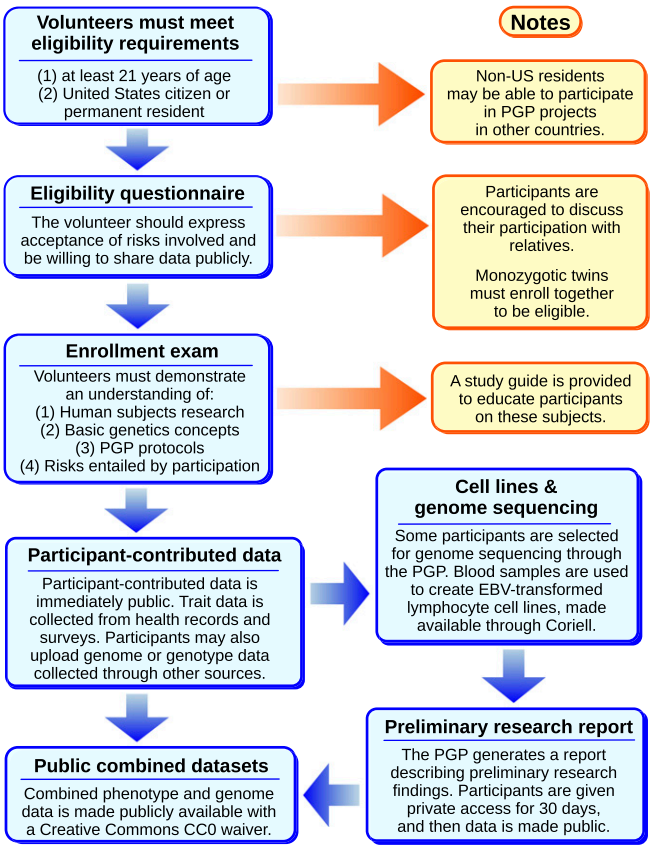


Fig. 1. PGP enrollment and data collection process. Enrollment in the PGP involves a series of steps meant to ensure informed consent for the public release of personal, reidentifiable genome and trait data. Current and historical copies of our consent forms are publicly available at http://www.personalgenomes.org/consent/.

underrepresented groups or who have particular traits and/or familial relationships to other participants. Participants are able to extend their profiles with a variety of personal data, including self-collected genetic data, listing enrolled relatives, health records and trait information, and answers to trait and ancestry surveys. These data are made publicly available immediately. As of May 2012, more than 1,000 participants have imported electronic health record data. In addition, as of May 2012, more than 800 participants have DNA samples derived from blood or saliva. These health record data and DNA samples represent the seed of a public resource integrating phenotype data with genotype data and include both common and rare diseases (phenotype data in *SI Appendix*, Dataset S1).

PGP-10 Pilot Cell Lines and Genome Sequence Data. To enable follow-up functional studies and genome sequence confirmation by third parties, cell lines are established for PGP participants and shared publicly alongside whole-genome data. Fibroblast and EBV-transformed lymphocyte cell lines were established with samples collected from the PGP-10 pilot cohort and have been made available through Coriell Cell Repositories (*SI Appendix*, Table S1). The PGP-10 genome data were produced using DNA purified from these cell lines, sequenced by Complete Genomics, Inc. (CGI) using their 2.0 pipeline (software version 2.0.1.5, matched against the build 37 reference genome). These genome data files have been shared publicly via our site (http://www.personalgenomes.org/data/PGP12.05/).

In addition to calling variants, CGI's genome files report which regions of the genome are confidently called as matching reference and which are "no-call" gaps that are insufficiently covered (and therefore not called as either variant or reference).

Using these data, we are able to assess what fraction of the genome has been successfully genotyped. On average, 96.5% of assembled reference genome positions were called homozygously in the CGI var files for the PGP-10 (*SI Appendix*, Table S2). Coverage is subject to systematic biases: positions called in one genome are much more likely to have been called in the other nine genomes (*SI Appendix*, Fig. S3). A position called in any given genome has a 92% chance of also being called in the other nine genomes, whereas a position not covered in that genome only has a 12% chance of being covered in all of the other nine.

The high quality of our pilot data is evident from analysis of several genomes derived from the same individual. PGP1 genomes were produced using DNA from three different cell lines: EBV-transformed lymphocytes, fibroblasts, and fibroblast-derived iPS cells. We use these data to assess overlap in variant calls because the underlying DNA sequences are expected to be mostly identical. When analysis is limited to positions explicitly called reference or variant in all three genomes (2,993,691 variant positions, 2.65 Gb total), 98.5% of variant positions are shared in all three genomes (Fig. 2*A*). When reference positions are taken into account, the three genomes have matching calls for 99.998% of these positions.

Which positions are sufficiently covered, and thus explicitly called as reference or variant, varies between genomes. Within one of the three PGP1 genomes 87% of variant positions, on average, are also called by the other two genomes. From this set of positions we can estimate the error rate due to random (rather than systematic) causes within a given genome: 99.6% of these variant positions are also called variant by at least one of the other two genomes (i.e., called variant by at least two out of three). When reference positions are included in analysis, 96% of positions called in one genome are called in all three, and 99.9994% of genotype calls in that genome match the call made by at least two out of the three genomes.

In total, 3,815,237 different variant positions were reported in the three genomes, 77% of which were called in all three (Fig. 2*B*). When these diagrams are constructed separately by variant type, we find that more complex length-changing variant calls also have high consistency, with 99.0% of such variants in a given genome called as variant by at least two out of three (*SI Appendix*, Fig. S4). In Fig. 2*B*, most positions where variant calls do not match are due to differences in coverage or base call quality that result in a "no-call" in one or more of the three sequences, as opposed to actual inconsistency in the variant vs. reference calls. This demonstrates the importance of respecting the logical inequivalence between the predicates "is not called as variant" and "is called as reference" and the need for correspondingly precise bookkeeping, possibly through the use of three or four valued logics (20, 21).

All of the PGP-10 had genome data produced from EBV-transformed lymphocyte cell lines, and these are used in all remaining genome analyses. Because the cost difference between exomes and whole genomes is already small, and may eventually vanish entirely, we preferred whole-genome sequencing over targeted approaches. On average these genomes have 3.2 million substitution variant calls relative to the build 37 reference genome and 300,000 short length-changing variants (*SI Appendix*, Table S2). Each individual has on average 8,250 single base substitution variants predicted to be nonsynonymous in a canonical transcript from University of California, Santa Cruz Known Genes (Table 1) (22). Of these, almost all (99.97%) are found in either dbSNP (build 132) or Exome Variant Server data (ESP5400) (23, 24). Notably, this novel variant rate (0.03%) is lower than the rate of random error we would predict on the basis of PGP1 genome comparisons (Fig. 2 and *SI Appendix*, Fig. S4); this may be due to increased accuracy in coding regions or due to common errors shared by both our data and other databases.

Genome variant statistics can vary depending on a given genome's coverage and the stringency used to identify variations, but our data are generally similar to whole-genome sequencing numbers reported elsewhere. Our counts for the number of missense variants in a single individual are somewhat lower than in other publications: this may be due to differences in coverage, stringency in variant calls, or the transcript annotations used for predictions (25, 26). MacArthur et al. (27) reported, on average, 304 nonsense and frameshift variants per individual with European ancestry (compared with our average of 166); their count was reduced to 64 after filtering to increase both variant confidence

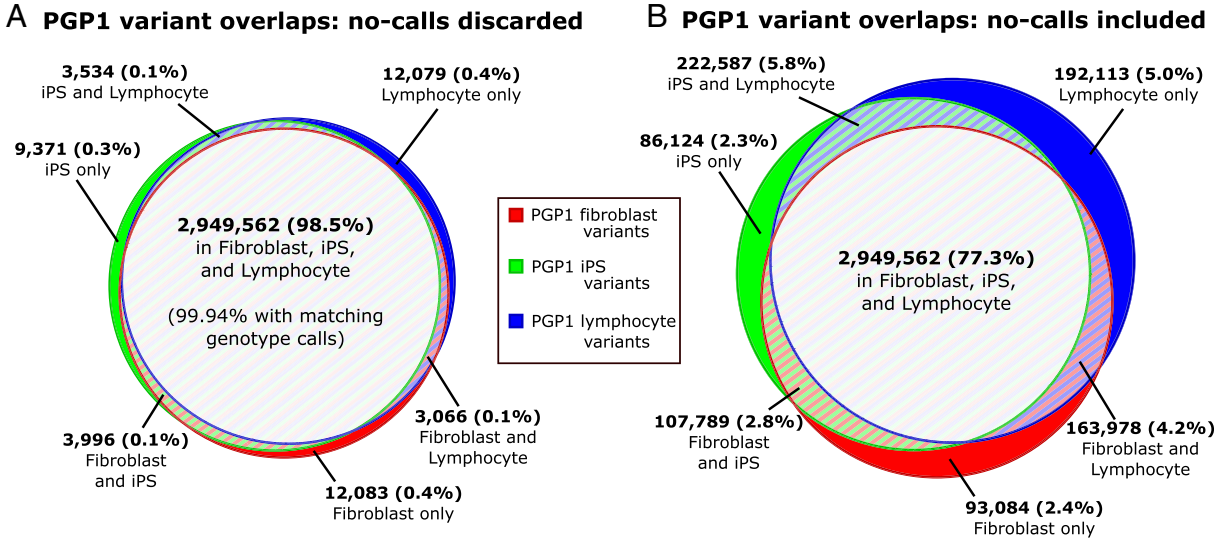


Fig. 2. Venn diagram comparisons of variant calls in PGP1 genomes. Analysis of PGP1 genome variant calls from three different tissues: fibroblast cells, fibroblast-derived iPS cells, and EBV-transformed lymphocyte cells. (*A*) Overlap of all variant calls, limited to positions that are explicitly called as reference or variant in all three genomes. Positions where any of the three genomes have a no-call (lacking coverage to make a confident call) are discarded from analysis. The low residual discordance consists of sequencing errors or real differences between these three tissues and indicates high sequence quality in each of these samples. (*B*) Overlap of all variant calls; positions not called in other genomes are included in the analysis. Most locations that were called as "variant" by one genome and not by other genomes were due to a lack of coverage in the other genomes. Reporting the regions confidently called as matching reference (as opposed to regions lacking sufficient coverage) is critical to genome interpretation and data comparisons.

Table 1. PGP-10 variants with potential functional consequences

Individual (huID)	No. of nonsyn. single base substitution (nsSNPs) variants	No. of nsSNPs not present in dbSNP132 or ESP5400	No. of nonsyn. with "probably damaging" Polyphen 2 prediction	No. of variants in PharmGKB or HuGENet	No. of nonsyn. variants in genes with clinical testing (GeneTests)	No. of nonsyn. variants matching OMIM entries	No. of nonsense and frameshift mutations	No. with prioritization score of 4 or more
PGP1 (hu43860C)	7,781	4	610	1,853	773	34	170	23
PGP2 (huC30901)	8,170	5	618	1,747	809	46	148	29
PGP3 (huBEDA0B)	7,899	1	582	1,793	780	46	169	35
PGP4 (huE80E3D)	8,042	2	606	1,820	829	47	147	26
PGP5 (hu9385BA)	8,312	1	652	1,865	873	53	172	27
PGP6 (hu04FD18)	8,008	3	596	1,810	832	46	167	29
PGP7 (hu0D879F)	8,380	4	649	1,917	868	48	167	25
PGP8 (huAE6220)	8,551	3	694	1,879	876	50	157	30
PGP9 (hu034DB1)	7,542	2	575	1,740	752	41	166	27
PGP10 (hu604D39)	9,810	2	769	1,723	956	42	199	37
Average	8,250	2.7	635	1,815	835	45	166	29

nonsyn., nonsynonymous; nsSNP, nonsynonymous SNP.

and likelihood of functional effect (i.e., not terminal or rescued by splice variants; this latter filtering was not performed by us).

Prioritization of Variants with Potential Clinical Relevance. Creating public methods for genome interpretation and returning interpreted results to participants are core goals of the PGP. Our system facilitates interpretation of whole-genome data by prioritizing variants for review. Preliminary versions of the system have been used in previous publications (9, 12–14). Here we apply the system to our pilot PGP-10 genomes.

To assist discovery of variants with potential phenotypic effects, potential amino acid changes are predicted for all variants occurring within gene coding regions. Variants are then matched against a variety of publicly available datasets: allele frequency data from 1,000 Genomes Project and Exome Variant Server data (24, 28), Polyphen 2 predictions (29), Human Genome Epidemiology Network (HuGENet) (30), Pharmacogenetics Knowledge Base (PharmGKB) (31), GeneTests (32), and Online Mendelian Inheritance in Man (OMIM) (33). After processing, there are many variants that potentially have clinically important consequences (Table 1). On average 635 variants are predicted as "probably damaging" by Polyphen 2, and another 166 are predicted to be severely disruptive nonsense or frameshift variants. When matching variants against our imported databases, each genome on average was found to have 1,815 variants with dbSNP IDs matched to a PharmGKB or HuGENet entry, 45 nonsynonymous variants matched to an OMIM entry, and 835 nonsynonymous variants occurring within genes that have clinical testing available (GeneTests). In total, these variants represented thousands of locations of potential significance when searching a presumed-healthy genome for clinically significant findings.

More complete evaluation of these variants requires incorporating information from the literature, but there are too many variants to do this comprehensively; variant interpretation is inefficient because automatic literature interpretation is computationally refractory—literature analysis requires human attention. To address this, we sought to prioritize variants for review. Review prioritization is implemented through an automatic "prioritization score" heuristic that uses these data to score variants in three categories: computational information, published gene-specific information, and published variant-specific information (*SI Appendix*, Table S3). Each category assigns up to two points, for a total of up to six points for a given variant. On average we found that each of the PGP-10 genomes had 29 variants with prioritization scores of 4 or more, and 131 variants with scores of 3 or more. Because our system accumulates data (see below), the burden of variant review drops dramatically when evaluations from prior genome interpretations can be reused: after 64 genomes we find that there are on average only 8 variants with a prioritization score of 4 or more, and 44 with a score of 3 or more (Fig. 3).

To test how well prioritization scores performed in prioritizing known disease-causing variants, we evaluated the prioritization scores that would be assigned to variants taken from a variety of disease-causing mutation databases (34–38) (lists downloaded September 2011). Although the findings reported in these databases may also be found in the databases used by our prioritization calculation (OMIM, Genetests, PharmGKB, and HuGENet), they are otherwise independent and are not themselves used in generating prioritization scores. We compared the prioritization scores assigned to variants from these databases with scores given to all nonsynonymous variants in PGP genomes (Fig. 4). On average, 44.0% of variants from these disease databases had prioritization scores or 4 or more, and 90.2% had scores of 3 or more. In contrast, only 0.22% of nonsynonymous variants in the PGP-10 have scores of 4 or more, and 1.1% have scores of 3 or more.

We applied our prioritization score system to prioritize genetic variants within the PGP-10 genomes for review. Our analysis focused on the discovery of unexpected variants predicted to have clinically significant consequences with moderate or high penetrance, because these potentially actionable variants were seen as the most important to return. Using the prioritization scores and presence in databases to guide our review of rare variants, we found 10 variants predicted to cause notable traits or pathogenic effects with moderate or high penetrance (*SI Appendix*, Table S4) and 21 variants predicted to cause moderate or severe disease in a recessive manner (*SI Appendix*, Table S5).

Follow-Up of Findings in the PGP-10. In the course of our review of the PGP-10 variants we observed multiple instances in which literature reports suggested that highly penetrant pathogenic phenotypes were caused by, or associated with, variants in the PGP-10 genomes. We found that such reports must be carefully appraised. Although some of these can be discarded because of clear phenotype discordance or unusual allele frequencies, some variants are rare and predict severe late-onset disease: participants could have undetected early stages of possibly clinically serious conditions. Because the PGP-10 genome analyses were not driven by medical or family history, follow-up evaluation of such findings entails issues very similar to follow-up of "incidental" findings; this potentially leads participants to incur unnecessary medical procedures, risks, and costs (39). However, after considerable discussion within the PGP team, we pursued additional communications and noninvasive clinical testing, with the thought that the public nature of our data and interpretations would inform researchers and clinicians who have similar findings in the future.

Focused follow-up was performed for one of the first variants found, MYL2-A13T in PGP6, which has been reported to cause familial hypertrophic cardiomyopathy in a dominant manner (40–44). Because this disease is potentially lethal and because there

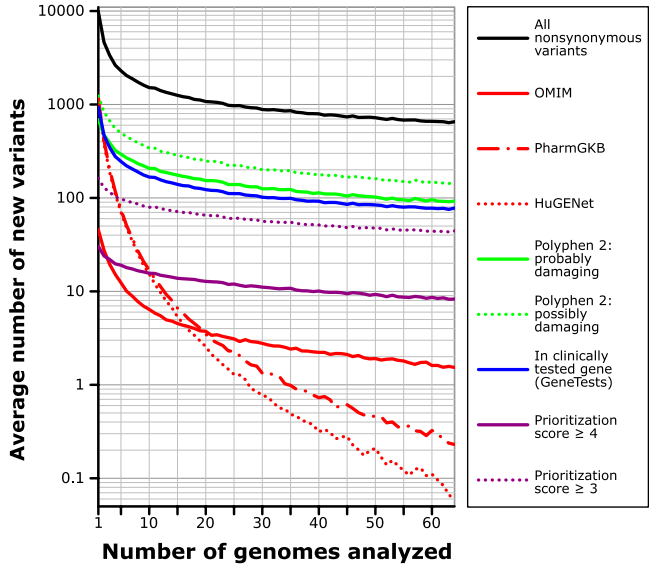


Fig. 3. Drop in number of new variants in each additional genome. For each new genome that is analyzed, the number of new variants not already seen in a previous genome falls dramatically. If editors record variant evaluations, the process of genome evaluation becomes easier as the number of new variants that are prioritized within each new genome is reduced. Data represent the average of 1,000 simulations using random orderings of a combined set of 64 genomes (the PGP-10 and 54 unrelated public genomes released by CGI).

were several publications supporting a pathogenic effect for the variant (*SI Appendix*, Fig. S5*A*), we confirmed the presence of this variant in a Clinical Laboratory Improvement Amendments-approved laboratory and consulted with researchers at the Laboratory for Molecular Medicine (LMM). LMM's internal data contained an additional pedigree of hypertrophic cardiomyopathy involving this variant (*SI Appendix*, Fig. S5*B*). Combined with published pedigrees the familial evidence is weak: both this pedigree and one of the published pedigrees each had one affected individual who was not a carrier of the MYL2-A13T variant, demonstrating that segregation of the variant was inconsistent with disease and significantly weakening the pathogenic hypothesis.

We informed PGP6 of our findings, reviewed the literature with him, and recommended cardiac follow-up for a noninvasive, nonurgent, baseline echocardiogram—this echocardiogram proved to be normal. Because he seems to be unaffected and his parents had no medical history of cardiac disease, this rare variant could be interpreted as a false-positive finding. However, familial hypertrophic cardiomyopathy is known to have incomplete penetrance, and the participant reports maternal and paternal uncles with early cardiac disease—uncertainty remains regarding the effect of this variant. It remains possible that this participant will develop symptoms at some later date; because the PGP maintains ongoing relationships with participants, such health updates can be added to participant records.

Less intensive follow-up, in the form of self-reported personal and family medical history, was performed for other variants that had reported or predicted strong phenotype effects. In the case of SERPINA1 variants found in PGP1 (who is compound heterozygous for variants predicted to result in E366K and E288V substitutions), the participant would be predicted to have increased susceptibility to developing chronic obstructive pulmonary disease (COPD) in response to smoking—the participant has no history of smoking and no diagnosis of COPD. To minimize the influence of confirmation bias, the remaining findings were combined into a single questionnaire given to all participants, without any specific efforts to alert them to which variants came from which individual (*SI Appendix*, Table S6). As with other participant trait surveys, the results of this targeted questionnaire are now publicly associated with the participant profiles. None of the participants reported traits or family histories consistent with the potential findings.

The SCN5A-G615E variant predicted in PGP9 was of particular concern: although we assessed the published findings as lacking statistical significance, it is included in a commercial genetic test for Long-QT syndrome (which can cause sudden death) (45). Subsequent to the survey, we contacted the participant (a 50-y-old woman) to notify her of our findings. In addition to no personal diagnosis of Long-QT syndrome and no family history of sudden death or Long-QT syndrome, she has had electrocardiogram tests performed in 2010 and 2012, with normal results.

On the whole, despite the presumed-healthy status of our 10 pilot participants, we found many apparently erroneous hypotheses in the literature whereby rare variants were predicted to have an inconsistent (and sometimes severe) phenotype. We also found that the process of genome interpretation involved a high amount of labor, much of which could potentially be reused in later genomes. These issues led us to extend our genome interpretation system to facilitate and record standardized variant evaluations with stringent evidence requirements.

Open Software for Variant Detection, Genome Reports, and Assisted Evaluation. Our Genome-Environment-Trait Evidence (GET-Evidence) system records variant evaluations using a peer production system and is integrated with our automatic genome processing and variant prioritization. Variant interpretations can be recorded by editors, categorized, and scored according to strength of evidence and clinical effect, and relevant papers can be added using PubMed identifiers. GET-Evidence variant pages contain links to external data sources where available, including

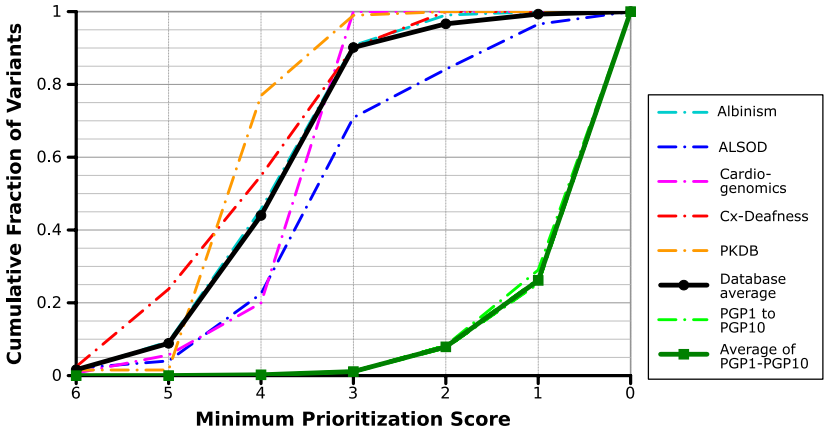


Fig. 4. Assessment of prioritization scores using disease-specific mutation databases. To demonstrate successful prioritization of variants with our prioritization score, we calculated the prioritization scores assigned to variant lists from a variety of disease-specific mutation databases: the Albinism Database (Albinism), the ALS Online Genetics Database (ALSOD), the Cardiogenomics Sarcomere Protein Gene Mutation Database (Cardiogen), the Connexins and Deafness Homepage (Cx-Deafness), and the Autosomal Dominant Polycystic Kidney Disease Mutation Database (PKDB). A variety of factors contribute to variation in performance for these lists: some diseases, for example, are more likely to be caused by severe frameshift or nonsense mutations (which we score highly), and some lists may include genes that are not yet used in clinical testing.

OMIM (33), GeneTests (32), dbSNP (23), PharmGKB (31), HuGENet (30), and PubMed (46).

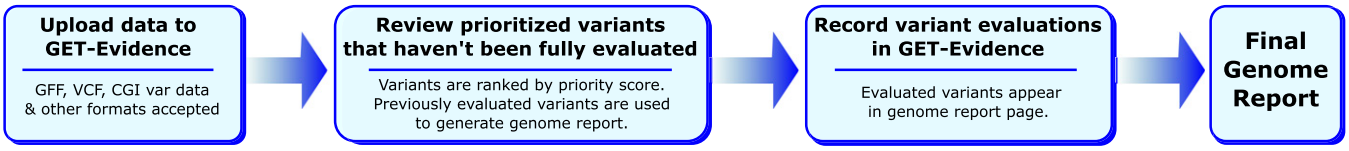
GET-Evidence facilitates whole-genome interpretation by creating an interpretation pipeline that combines genome data processing, prioritization of variants for review, and recording of variant evaluations (Fig. 5*A*). When a genome data file is uploaded, the genome analysis system calculates the prioritization scores for all variants in an uploaded genome and matches these variants against the existing database. Two major reports are provided: an "insufficiently evaluated variants" report and a "genome report" (Fig. 5 *B* and *C*, respectively). The "genome report" lists all variants within the genome that have been sufficiently evaluated within GET-Evidence—variants initially seen here have likely been seen and evaluated in a genome previously analyzed through GET-Evidence. The "insufficiently evaluated variants" report contains all novel and unevaluated variants, sorted by prioritization score and accompanied by information that may guide evaluation (e.g., allele frequency, presence in databases, Polyphen 2 results, and number of article links added). Editors may then record or update evaluations of variants; once a variant is sufficiently evaluated, it is displayed within the genome report.

Variant evaluations record diverse information about variants that contribute to genome interpretation (Fig. 6). Editors can classify variants according to phenotypic effect (pathogenic, protective, pharmacogenetic, or benign) and inheritance pattern (dominant, recessive, or other). Papers may be added by using PubMed identifiers, creating new fields for entering case/control data and a field for adding notes regarding what evidence the paper has regarding the variant. To highlight important findings from a publication and to gather standardized information for later development of automatic interpretation, the abstracts of linked publications can be annotated through highlighting evidence features using the BioNotate platform (47). Finally, to record the overall interpretation of the variant and any additional relevant information, short summary and longer summary sections provide regions for free text summary of the variant's effect and evidence.

In addition to these classifications and text summaries, GET-Evidence uses a series of scored categories to facilitate automatic filtering and scoring of variants (*SI Appendix*, Table S7). These categories are divided into two major sections: (*i*): variant evidence scores, which assess how strongly various lines of evidence support the variant having a hypothesized effect, and (*ii*) clinical importance scores, which assess clinical aspects of the variant's hypothesized effect (Fig. 6). Variant evidence scores and clinical importance scores are used to generate an overall assessment of evidence (uncertain, likely, or well-established) and clinical importance (low, moderate, or high) (*SI Appendix*, Tables S8 and S9). Notably, variants are only considered "likely" or "well-established" if they meet minimum statistical significance requirements in either case/control or familial categories (described in *SI Appendix*, Table S9). By segregating evidence from severity we are able to distinguish between a well-established variant with a weak pathogenic effect ("well-established pathogenic, low clinical importance") from a poorly understood but potentially severe variant ("uncertain pathogenic, high clinical importance").

After evaluating all variants in GET-Evidence, almost all variants we found with potentially strong phenotypic consequences were evaluated as "uncertain" (Table 2 and *SI Appendix*, Table S10). Although it is always possible that one or more of these variants does cause disease with incomplete penetrance or late onset, there are clearly some erroneous associations listed in Table 2 and *SI Appendix*, Table S4. Introducing stringent evidence requirements for interpreting published data successfully addresses this issue with incidental findings. In addition, GET-Evidence's peer production model for variant evaluation assists genome interpretation by allowing the reuse of variant evaluations by later genome evaluations, thereby minimizing duplication of effort. By creating such a shared central resource for recording interpretations,

A



B

Evaluated variants | Insufficiently evaluated variants | Search:

Variant	Zyg.	Allele freq.	Num. of articles	Prior. Score	Prioritization reasons
SERPINA1-E366K	Het	1.2%	5	6	In OMIM, unevaluated web hits, Polyphen 2: probably damaging, Clinically tested gene with associated GeneReview
SERPINA1-E288V	Het	3.0%	6	6	In OMIM, unevaluated web hits, Polyphen 2: probably damaging, Clinically tested gene with associated GeneReview
DSP-R1537C	Het	1.0%	1	5	Unevaluated web hits, Polyphen 2: probably damaging, Clinically tested gene with associated GeneReview
TYR-S192Y	Hom	27%	2	5	In OMIM, HuGENet, PharmGKB, unevaluated web hits, Polyphen 2: probably damaging, Clinically tested gene with associated GeneReview
SLC45A2-L374F	Hom	69%	4	5	In OMIM, unevaluated web hits, Polyphen 2: probably damaging, Clinically tested gene with associated GeneReview
DNA5-Q1464 Frameshift	Het	?	0	4	Frameshift, Clinically tested gene with associated GeneReview
LYST-R159K	Het	?	2	4	Polyphen 2: probably damaging, Clinically tested gene with associated GeneReview
WFS1-C426Y	Het	0.12%	2	4	Confirmed web hits, Polyphen 2: possibly damaging, Clinically tested gene with associated GeneReview
BBS7-D412G	Het	0.2%	2	4	Polyphen 2: probably damaging, Clinically tested gene with associated GeneReview
...					

C

Evaluated variants | Insufficiently evaluated variants | Search:

Variant	Clinical import.	Impact	Allele freq.	Short summary
SERPINA1-E366K	High	Well-established pathogenic Recessive, Carrier (Het)	1.2%	This is also called the "Pi Z" or "Z" allele. When homozygous (acting in a recessive manner) this variant is the major cause of severe alpha-1-antitrypsin deficiency (95% of cases) which often leads to emphysema or chronic obstructive pulmonary disease (COPD) and liver disease in adults and children. Heterozygosity for this variant may also be associated with increased rate of lung or liver problems, especially when combined with another variant with reduced function (compound heterozygous).
BBS7-D412G	High	Uncertain pathogenic Recessive, Carrier (Het)	0.2%	Predicted to have damaging effect, other mutations in this gene have been implicated in causing Bardet-Biedl syndrome in a recessive manner.
RYR2-G1885E	High	Uncertain pathogenic Recessive, Carrier (Het)	1.8%	Reported to cause arrhythmogenic right ventricular cardiomyopathy when compound heterozygous with G1886S, although this finding is weakened after correcting for multiple hypotheses and it is unclear what penetrance such a genotype might have, if it is causal.
C3-R102G	Moderate	Likely pathogenic Complex/other, Heterozygous	5.2%	This variant (also called C3F) is common in Europeans (17% allele frequency), and is associated with age-related macular degeneration. In the US, 1.5% of adults over 40 have the disease, but the incidence increases strongly with age (>15% in women over 80). Assuming an average lifetime risk of ~10%, we estimate heterozygotes have a ~13% risk and homozygous have ~20%.
WFS1-C426Y	Moderate	Uncertain pathogenic Dominant, Heterozygous	0.12%	Reported in a single case of familial depression, but the case had no familial linkage data and no statistical significance.
...				

Fig. 5. GET-Evidence and genome reports. (*A*) Using GET-Evidence involves genome upload followed by review of prioritized insufficiently evaluated variants. Combining these reviews with previously reviewed variants produces the final genome report. (*B*) Insufficiently evaluated variants are ranked according to prioritization score and are listed with additional information of interest (allele frequency, number of associated articles, presence in databases, and computational predictions). (*C*) Sufficiently evaluated variants are presented in the genome report with summary information regarding variant effect, severity, and evidence.

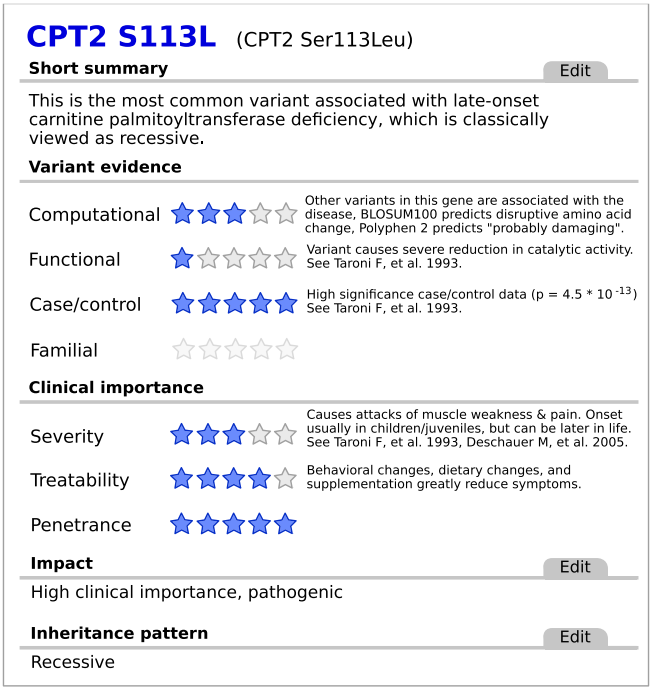


Fig. 6. Sample GET-Evidence variant report. Variant report pages on GET-Evidence allow editors to record and organize information relevant to variant interpretation. A scoring system is used for variant evidence and clinical importance categories to allow automatic sorting of interpreted variants. On the basis of the strong case/control evidence and high treatability and penetrance, this recessive pathogenic variant carried by PGP4 (listed in Table 2) was evaluated as well-established and high clinical importance.

GET-Evidence can act as a forum for building consensus on interpretation. The analysis system and variant interpretations, along with our public genome interpretations, are available at http://evidence.personalgenomes.org.

Discussion

With the advent of low-cost whole-genome sequencing and growing interest in personalized medicine, the research community is faced with the challenge of developing tools for interpreting genome data and using these data to inform lifestyle choices and clinical care in an effective manner. Doing so will require large, highly personal datasets: whole-genome data combined with health records, traits, and personal medical histories. Because such data are highly reidentifiable, building these datasets results in a tension between privacy protection and the desire to share and reuse data.

The approach the PGP takes is a highly public option: enrolling participants who agree to the hypothetical and unknown risks associated with making personal biological data public through an open consent format. Our public resource enables the process of scientific discovery and clinical use of genomes. In addition, we share our open consent documents and methods to enable other researchers who wish to produce public data in their own research studies.

As part of these integrated public datasets, the PGP has also created a public software tool for genome interpretation and a public database of variant interpretations. Because these records are freely editable by any registered user, the database provides a forum for achieving a public consensus interpretation of genetic variants. Other groups may freely use the GET-Evidence system, and we encourage others to contribute their interpretations of genetic variants in the public database. These edits and other data within GET-Evidence are shared, in turn, as public domain under a CC0 waiver and may be used by academic and commercial genome interpretation efforts. Future development of the GET-Evidence system should move closer toward our goal of a richly interconnected dataset of genomes, environments, and traits. Planned improvements include coded phenotypes for genetic variants as well as participant health records, genome analysis for compound heterozygosity, splicing mutations, copy number variants, and tracking the biological and computational provenance of public data.

Our genome interpretation findings highlight one of the ethical issues raised when working toward clinical utilization of whole genomes: what should be done if potentially severe pathogenic mutations are found within whole-genome sequence data? Although stringent evidence guidelines help by classifying

Table 2. Evaluation of variants reported or predicted to have strong phenotype effects

Variant (heterozygous unless otherwise noted)	Predicted phenotype	Allele frequency (%)	Prioritization score	Evidence assessment in GET-Evidence	Clinical importance assessment in GET-Evidence
SERPINA1-E366K/ SERPINA1-E288V (compound het)	Moderate α-1 antitrypsin deficiency	1.2 and 3.0	5	Well-established/ well-established	High/low
WFS1-C426Y	Familial depression	0.1	5	Uncertain	Moderate
FLG-S761fs	Palmar hyperlinearity and keratosis pilaris (ichthyosis vulgaris in recessive manner)	Unknown	4	Uncertain	Moderate (for ichthyosis vulgaris)
PKD1-R4276W	Autosomal dominant polycystic kidney disease	0.2	4	Uncertain	High
MYL2-A13T	Hypertrophic cardiomyopathy	0.02	5	Uncertain	High
SCN5A-G615E	Long-QT Syndrome	0.03	4	Uncertain	High
PKD2-S804N	Autosomal dominant polycystic kidney disease	0.3	5	Uncertain	High
SLC9A3R1-R153Q	Kidney stones	0.3	4	Uncertain	Moderate
RHO-G51A	Autosomal dominant retinitis pigmentosa	0.2	4	Uncertain	Moderate
EVC-R443Q	Ellis-van Creveld syndrome	7.9	3	Reevaluated as benign	Reevaluated as benign

Additional data regarding these variants, including PGP participant identifiers and Pubmed identifiers for related literature, are available in *SI Appendix*, Table S4.

many findings as uncertain, effects could manifest later in life. Withholding information from patients is becoming less acceptable in clinical practice and may become less acceptable for research data as well. Continuing work with PGP participants will provide insights into how genome data may be integrated more generally into both research and clinical settings.

We maintain an ongoing relationship with participants to monitor the outcomes of publicly sharing personal data. Many participants are interested in making an ongoing contribution to science—as part of our study, we can invite participants to take part in additional research. Thus, subsets of participants may choose to contribute to disease-specific research and novel profiling methods (e.g., allele-specific expression, epigenetic, metabolomic, proteomic, or microbiome profiling). In addition, biobanked tissues and cell lines may be used by researchers for additional characterization, follow-up functional studies, and genome engineering. Each additional study benefits from all previous data for the same participant, building a further-enriched dataset and contributing to the development of new personalized medical diagnostics and therapies. Currently approved for studying up to 100,000 participants, the PGP has the potential to be a widely used ongoing resource—a large, rich, public set of well-characterized individuals with extensive biological data and an ongoing interest in contributing to research.

Materials and Methods

SI Appendix, SI Materials and Methods provides full details of our enrollment process and open consent protocols. Additional details of Continuity of Care Record format health record data, cell lines, samples, genome sequencing, and quality assessment, as well as prioritization score assessment using disease-specific mutation databases, are also presented. Finally, we elaborate on the GET-Evidence data processing and editing platform; its development is facilitated through use of a shared computational and storage infrastructure (48).

ACKNOWLEDGMENTS. We thank all members of the G.M.C. laboratory and other members of the Personal Genome Project Community, Ting Wu, and other members of the Personal Genetics Education Project for their help and advice; and Gerard T. Berry, Gerald Cox, Dongliang Ge, Ho Ghang, Taehyung Kim, Min Seob Lee, Sunghoon Lee, Stephen Quake, Kevin V. Shianna, and Anne West for contributions of data, assistance in analyses, and advice to our previous genome analysis efforts. This work was supported in part by National Institutes of Health Grants P50HG005550 (National Human Genome Research Institute) and R01HL094963 (National Heart, Lung, and Blood Institute), and by PersonalGenomes.org.

1. Greenbaum D, Sboner A, Mu XJ, Gerstein M (2011) Genomics and privacy: Implications of the new reality of closed data for the field. *PLOS Comput Biol* 7:e1002278.
2. Church GM (2005) The Personal Genome Project. *Mol Syst Biol* 1:2005.0030.
3. Lunshof JE, Chadwick R, Vorhaus DB, Church GM (2008) From genetic privacy to open consent. *Nat Rev Genet* 9:406–411.
4. Ball MP, et al. (2009) Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat Biotechnol* 27:361–368.
5. Zhang K, et al. (2009) Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. *Nat Methods* 6:613–618.
6. Sommer MOA, Dantas G, Church GM (2009) Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science* 325:1128–1131.
7. Li JB, et al. (2009) Multiplex padlock targeted sequencing reveals human hypermutable CpG variations. *Genome Res* 19:1606–1615.
8. Lee JH, et al. (2009) A robust approach to identifying tissue-specific gene expression regulatory variants using personalized human induced pluripotent stem cells. *PLoS Genet* 5:e1000718.
9. Drmanac R, et al. (2010) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* 327:78–81.
10. Sullivan GJ, et al. (2010) Generation of functional human hepatic endoderm from human iPS cells. *Hepatology* 51:329–335.
11. Gore A, et al. (2011) Somatic coding mutations in human induced pluripotent stem cells. *Nature* 471:63–67.
12. Kim JI, et al. (2009) A highly annotated whole-genome sequence of a Korean individual. *Nature* 460:1011–1015.
13. Ashley EA, et al. (2010) Clinical assessment incorporating a personal genome. *Lancet* 375:1525–1535.
14. Dewey FE, et al. (2011) Phased whole-genome genetic risk in a family quartet using a major allele reference sequence. *PLoS Genet* 7:e1002280.
15. Choi M, et al. (2009) Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci USA* 106:19096–19101.
16. Lupski JR, et al. (2010) Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N Engl J Med* 362:1181–1191.
17. Ng SB, et al. (2010) Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* 42:30–35.
18. Rope AF, et al. (2011) Using VAAST to identify an X-linked disorder resulting in lethality in male infants due to N-terminal acetyltransferase deficiency. *Am J Hum Genet* 89:28–43.
19. Yandell M, et al. (2011) A probabilistic disease-gene finder for personal genomes. *Genome Res* 21:1529–1542.
20. Belnap N (1977) *Modern Uses of Multiple-Valued Logic*, eds Dunn M, Epstein G (Springer, New York).
21. Fitting M (1994) Kleene's three valued logics and their children. *Fundam Inf* 20:113–131.
22. Hsu F, et al. (2006) The UCSC Known Genes. *Bioinformatics* 22:1036–1046.
23. Sherry ST, et al. (2001) dbSNP: The NCBI database of genetic variation. *Nucleic Acids Res* 29:308–311.
24. NHLBI Exome Sequencing Project (2012) Exome Variant Server. Available at: http://evs.gs.washington.edu/EVS/. Accessed May 15, 2012.
25. Chen R, et al. (2012) Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* 148:1293–1307.
26. Tennessen JA, et al. (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, PMID:22604720.
27. MacArthur DG, et al.; 1000 Genomes Project Consortium (2012) A systematic survey of loss-of-function variants in human protein-coding genes. *Science* 335:823–828.
28. 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature* 467:1061–1073.
29. Adzhubei IA, et al. (2010) A method and server for predicting damaging missense mutations. *Nat Methods* 7:248–249.
30. Yu W, Gwinn M, Clyne M, Yesupriya A, Khoury MJ (2008) A navigator for human genome epidemiology. *Nat Genet* 40:124–125.
31. Klein TE, et al.; Pharmacogenetics Research Network and Knowledge Base (2001) Integrating genotype and phenotype information: an overview of the PharmGKB project. *Pharmacogenomics J* 1:167–170.
32. Pagon RA (2006) GeneTests: An online genetic information resource for health care providers. *J Med Libr Assoc* 94:343–348.
33. McKusick-Nathans Institute of Genetic Medicine; Johns Hopkins University; National Center for Biotechnology Information, National Library of Medicine (2009) Online Mendelian Inheritance in Man. Available at: http://www.ncbi.nlm.nih.gov/omim. Accessed June 1, 2009.
34. University of Minnesota (2011) Albinism database. Available at: http://albinismdb.med.umn.edu/. Accessed December 1, 2011.
35. Lill CM, Abel O, Bertram L, Al-Chalabi A (2011) Keeping up with genetic discoveries in amyotrophic lateral sclerosis: The ALSoD and ALSGene databases. *Amyotroph Lateral Scler* 12:238–249.
36. NHLBI Program for Genomic Applications, Harvard Medical School (2011) Genomics of cardiovascular development, adaptation, and remodeling. Available at: http://www.cardiogenomics.org. Accessed May 26, 2010.
37. Ballana E, Ventayol M, Rabionet R, Gasparini P, Estivill X (2011) Connexins and deafness homepage. Available at: http://davinci.crg.es/deafness/. Accessed December 1, 2011.
38. PKD Foundation (2011) The autosomal dominant polycystic kidney disease mutation database. Available at: http://pkdb.mayo.edu/. Accessed December 1, 2011.
39. Kohane IS, Masys DR, Altman RB (2006) The incidentalome: A threat to genomic medicine. *JAMA* 296:212–215.
40. Poetter K, et al. (1996) Mutations in either the essential or regulatory light chains of myosin are associated with a rare myopathy in human heart and skeletal muscle. *Nat Genet* 13:63–69.
41. Szczesna D, et al. (2001) Familial hypertrophic cardiomyopathy mutations in the regulatory light chains of myosin affect their structure, Ca2+ binding, and phosphorylation. *J Biol Chem* 276:7086–7092.
42. Andersen PS, et al. (2001) Myosin light chain mutations in familial hypertrophic cardiomyopathy: Phenotypic presentation and frequency in Danish and South African populations. *J Med Genet* 38:E43.
43. Szczesna-Cordary D, Guzman G, Ng SS, Zhao J (2004) Familial hypertrophic cardiomyopathy-linked alterations in Ca2+ binding of human cardiac myosin regulatory light chain affect cardiac muscle contraction. *J Biol Chem* 279:3535–3542.
44. Hougs L, et al. (2004) One third of Danish hypertrophic cardiomyopathy patients have mutations in MYH7 rod region. *Eur J Hum Genet* 13:161–165.
45. Kapplinger JD, et al. (2009) Spectrum and prevalence of mutations from the first 2,500 consecutive unrelated patients referred for the FAMILION long QT syndrome genetic test. *Heart Rhythm* 6:1297–1303.
46. National Center for Biotechnology Information. (2011) PubMed. Available at: http://www.ncbi.nlm.nih.gov/pubmed/. Accessed December 2, 2011.
47. Cano C, Monaghan T, Blanco A, Wall DP, Peshkin L (2009) Collaborative text-annotation resource for disease-centered relation extraction from biomedical text. *J Biomed Inform* 42:967–977.
48. Zaranek AW, Clegg T, Vandewege W, Church GM (2008) Free Factories: Unified infrastructure for data intensive web services. *Proc USENIX Annu Tech Conf* 2008:391–404.