Received: 13 September 2016 | Revised: 24 March 2017 | Accepted: 17 May 2017

DOI: 10.1002/humu.23265

SPECIAL ARTICLE

WILEY HGVS HUMAN GENOME VARIATION SOCIETY

Matching phenotypes to whole genomes: Lessons learned from four iterations of the personal genome project community challenges

Binghuang Cai[1] | Biao Li[2] | Nikki Kiga[1] | Janita Thusberg[2] | Timothy Bergquist[1] | Yun-Ching Chen[3] | Noushin Niknafs[3] | Hannah Carter[4] | Collin Tokheim[3] | Violeta Beleva-Guthrie[3] | Christopher Douville[3] | Rohit Bhattacharya[5] | Hui Ting Grace Yeo[3] | Jean Fan[3] | Sohini Sengupta[3] | Dewey Kim[3] | Melissa Cline[6] | Tychele Turner[7] | Mark Diekhans[6] | Jan Zaucha[8,9] | Lipika R. Pal[10] | Chen Cao[10,11] | Chen-Hsin Yu[10,11] | Yizhou Yin[10,11] | Marco Carraro[12] | Manuel Giollo[12,13] | Carlo Ferrari[13] | Emanuela Leonardi[14] | Silvio C.E. Tosatto[12,15] | Jason Bobe[16] | Madeleine Ball[16] | Roger A. Hoskins[17] | Susanna Repo[18] | George Church[16] | Steven E. Brenner[17] | John Moult[10,19] | Julian Gough[9] | Mario Stanke[20] | Rachel Karchin[3,21] | Sean D. Mooney[1]

[1]Department of Biomedical Informatics & Medical Education, University of Washington School of Medicine, Seattle, Washington

[2]The Buck Institute for Research on Aging, Novato, California

[3]Department of Biomedical Engineering and Institute for Computational Medicine, Johns Hopkins University, Baltimore, Maryland

[4]Department of Medicine, Division of Medical Genetics, Institute for Genomic Medicine and Moores Cancer Center, University of California San Diego, La Jolla, Califonia

[5]Department of Computer Science, Institute for Computational Medicine, Johns Hopkins University, Baltimore, Maryland

[6]Department of Biomolecular Engineering, University of California, Santa Cruz, California

[7]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, Maryland

[8]Department of Computer Science, University of Bristol, Bristol, UK

[9]Bristol Centre for Complexity Sciences, University of Bristol, Bristol, UK

[10]Institute for Bioscience and Biotechnology Research, University of Maryland, Rockville, Maryland

[11]Computational Biology, Bioinformatics and Genomics, Biological Sciences Graduate Program, University of Maryland, College Park, Maryland

[12]Department of Biomedical Sciences, University of Padova, Padova, Italy

[13]Department of Information Engineering, University of Padova, Padova, Italy

[14]Department of Woman and Child Health, University of Padova, Padova, Italy

[15]CNR Neuroscience Institute, Padova, Italy

[16]PersonalGenomes.org, Boston, Massachusetts

[17]Department of Plant and Microbial Biology, University of California, Berkeley, California

[18]European Bioinformatics Institute, Hinxton, UK

[19]Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, Maryland

[20]Institute of Mathematics and Computer Science, University of Greifswald, Greifswald, Germany

[21]Department of Oncology, The Johns Hopkins Medical Institutions, Baltimore, Maryland

Correspondence
Sean D. Mooney, Department of Biomedical Informatics & Medical Education, University of Washington School of Medicine, 850 Republican Street, Seattle, Washington 98109.
Email: sdmooney@uw.edu

Contract grant sponsors: NIH (U41 HG007346, R13 HG006650, R01LM0009722, UL1TR000423, U01 HG008473, R01MH105524); NSF (1458443).

For the CAGI Special Issue

Abstract

The advent of next-generation sequencing has dramatically decreased the cost for whole-genome sequencing and increased the viability for its application in research and clinical care. The Personal Genome Project (PGP) provides unrestricted access to genomes of individuals and their associated phenotypes. This resource enabled the Critical Assessment of Genome Interpretation (CAGI) to create a community challenge to assess the bioinformatics community's ability to predict traits from whole genomes. In the CAGI PGP challenge, researchers were asked to predict whether an individual had a particular trait or profile based on their whole genome. Several approaches were used to assess submissions, including ROC AUC (area under receiver operating characteristic curve), probability rankings, the number of correct predictions, and statistical significance simulations. Overall, we found that prediction of individual traits is difficult, relying on a strong knowledge of trait frequency within the general population, whereas matching genomes to trait profiles relies heavily upon a small number of common traits including ancestry, blood type, and eye color. When a rare genetic disorder is present, profiles can be matched when one or more pathogenic variants are identified. Prediction accuracy has improved substantially over the last 6 years due to improved methodology and a better understanding of features.

KEYWORDS

biomedical informatics, community challenge, critical assessment, genome, genome interpretation, open consent, personal genome project (PGP), phenotype

1 | INTRODUCTION

Sequencing of whole genomes has improved our ability to build new methods for predicting phenotype from genotype. These new methods will be instrumental when applied in the clinic to achieve precision medicine. Community challenges are a mechanism to assess how well the scientific community can solve certain problems in an unbiased manner (Friedberg, Wass, Mooney, & Radivojac, 2015). As part of the Critical Assessment of Genome Interpretation (CAGI) challenges (CAGI, 2016), we developed a genome-to-phenotype matching challenge to assess how well individual traits and phenotypic profiles could be predicted from whole-genome sequences. Using open consent model genomes (from individuals who are comfortable sharing them without any promises of privacy, confidentiality, or anonymity) (Ball et al., 2014; PGP, 2016), we asked the community to predict traits and/or profiles for whole genomes for which the matched phenotypes had not yet been publicly released (but were available to and had been held back by challenge organizers). We have iterated on this process four times over the past 6 years, in 2010, 2011, 2013, and 2015, with only slight deviations on the experimental design each year. Our aim was to provide an opportunity for researchers to compare new methods and approaches for interpreting whole-genome sequences into phenotype predictions.

The Personal Genome Project (PGP) is a research study that produces freely available scientific resources that bring together genomic, environmental, and human trait data donated by volunteers (Ball et al., 2014). More about the project PGP can be found in Ball et al. (2014) and PGP (2016). The PGP follows an open consent model to human genome sequencing and was the reason our prediction challenge was possible (Ball et al., 2014). On the PGP recruitment and enrollment website, individuals provide a public profile of their phenotypes by answering a number of trait questions in different categories. Selected individuals then have their samples collected and sequenced, and their assembled genomes made available on the PGP Website along with their survey answers. The PGP CAGI organizers decided that as genomes and phenotype profiles were released, the matching key between the genomes and the phenotypes would remain undisclosed to allow submitters to blindly predict matches between available profiles and released genomes. Note that the trait data are self-reported by PGP participants using the PGP project website and are updated on the CAGI challenge website at the beginning of the challenge.

The PGP challenge has been held in 2011, 2013, and 2015. In this manuscript, we present the methodologies of the challenge submissions as well as an assessment of each submitter's ability to predict individual traits and to match whole genomes to phenotype profiles. We also assess whether we observe improvement over the 6 years of hosting this challenge. Using simulations, we show that the best groups made predictions that are statistically better than random submissions. We also show that in order to obtain accurate predictions of individual traits, an estimate of the statistical priors must be known. We then show that 5%–10% of the PGP profiles can be matched based on a small number of common features including ancestry, blood type, and eye color. In specific cases, when a rare genetic disorder was present, profiles could be matched to a genome when one or more pathogenic variants were identified.

Moreover, through the four iterations of PGP challenges, we found that challenges are a good measure of our standing in predicting phenotypes from whole-genome sequences. The authors found it of interest that, at best, only 20%–25% of whole genomes (the best correct matching prediction in PGP 2015) could be matched to profiles

TABLE 1 Comparison of PGP 2015, PGP 2013, PGP 2011, and PGP 2010

	PGP 2015	PGP 2013	PGP 2013 (assessed)	PGP 2011	PGP 2010
Genomes	23	77	50	10	10
Phenotype profiles	101	291	183	-	-
Traits	239	239	239	40	40
Submissions	5	16	16	4	1
Groups	4	5	5	4	1

Notes: PGP 2010 and PGP 2011 requested submitters predict each of the 40 traits, and PGP 2013 and PGP 2015 requested submitters match genomes to entire profiles. Note in 2013, 27 genomes and 108 profiles were trivially removable as unmatchable since 27 profiles had published 23andMe data and 108 profiles had no indicated sample taken on the PGP website. This is shown in the two columns of PGP 2013, in which column "PGP 2013 (assessed)" shows the information of data used to do the assessment.

derived from participant reported phenotype, suggesting that our ability to identify individuals or individual traits from a genome is far less accurate than one might expect based on our significant advances in genome interpretation approaches over the last two decades.

2 | MATERIALS AND METHODS

In this section, we first give an overview of the four iterations of CAGI PGP challenges (i.e., PGP 2010, PGP 2011, PGP 2013, and PGP 2015), and then discuss in detail each iteration of the challenge, including data, submitters' methods, and assessment methods.

2.1 | Overview

In this subsection, we briefly introduce the four iterations of PGP challenges, with comparisons of the four PGP challenges shown in Table 1.

In 2010, challenge participants were asked to predict 32 binary traits of the "PGP10," which were the first 10 individuals sequenced by the PGP project. The survey answers were withheld until after the submission deadline to ensure blinded predictions by the submitters. Predictors for 2010 were told how many predictions they got right, but not which predictions they got right, and thus were essentially blind to the results. The 2010 experiment was deemed a trial run, and as the data remained unspoiled, it was repeated in 2011 (PGP 2011) with the same data. Only those results of PGP 2011 are reported here. In total, there were four submissions from four groups for that challenge. One of the submissions used a novel methodology, which generated a patent application (Chen et al., 2014).

By 2013, PGP had made a website available that included all the participants, their survey answers, and limited information about the status of the sequencing of their genome, including whether a sample had been collected from that individual and was awaiting sequencing. Since the survey answers were now available publicly, the challenge was changed to a blinded matching challenge where groups were asked to match a genome sequence to a phenotypic "profile." These profiles were a list of "yes" or "no" answers from the survey questions, based on whether the individual considered her/himself to have each of 239 phenotypes. Seventy-seven phenotype profiles were released that had a matched genome, whereas a number of other profiles were released for which the matched genome was not yet sequenced. The true matches between the profiles and genomes were kept undisclosed by PGP for the duration of the challenge, and challenge participants were asked to submit the probabilities each genome had of matching to each profile. In total, 16 submissions from five groups were received. In addition, there was an optional challenge in PGP 2013, which was to predict individual traits from genomes.

In 2015, the 2013 matching challenge experiment was repeated, and 23 genomes and 101 phenotype profiles were released. We received five submissions from four different teams. Assessment was performed similarly to that of 2013 and the results from both challenges were compared.

Below and in the Supp. Material, we describe the submissions of independently submitted challenge predictors who were willing to share their methods. These identified groups are named and the list is as follows.

Group1: the group of Rachel Karchin at Johns Hopkins University;
Group2: the group of Silvio Tosatto at the University of Padova;
Group3: the group of John Moult at the University of Maryland;
Group4: the group of Julian Gough at the University of Bristol; and
Group5: the group of Mario Stanke at the University of Greifswald.

To be consistent, we have also named four submitters, who declined to be included or were otherwise unreachable, as Group6, Group7, Group8, and Group9. For groups that had multiple submissions, we numbered them by adding _1, _2, _3, and so on after group name and provided a description of their differences.

2.2 | PGP 2011

Data provided in the PGP challenge for 2011 included genomes of 10 participants and 32 binary traits. The traits are listed in Table 2. The challenge was aimed at predicting the values of each binary trait based on genome sequence. In total, there were four submissions for the binary trait prediction challenge and the methodologies of the two identified methods and the assessment methods are described below.

2.2.1 | Submissions

In 2011, we received both identified submissions from Group1 and Group2 and anonymous submissions from Group6. Their methodologies are described in Supp. Material.

TABLE 2 List of binary traits used for prediction in PGP 2011

Trait ID	Trait name
1	Asthma
2	Crohn's disease
3	Ulcerative colitis
4	Irritable bowel syndrome
5	Rheumatoid arthritis
6	Type II diabetes
7	Coronary artery disease
8	Long QT syndrome
9	Hypertrophic cardiomyopathy
10	Glaucoma
11	Color blindness
12	Bipolar disorder
13	Celiac disease
14	Psoriasis
15	Lupus
16	Breast cancer
17	Prostate cancer
18	Migraine
19	Lactose intolerance
20	Dyslexia
21	Autism
22	Osteoporosis
23	Incontinence
24	Kidney stones
25	Varicose veins
26	Sleep apnea
27	Tongue rolling (tube)
28	Phenylthiocarbamide tasting
29	Blood type: has A antigen?
30	Blood type: has B antigen?
31	Blood type: is Rh(D) positive?
32	Absolute pitch

2.2.2 | Assessments

We used several different approaches to evaluate the performances of the four submissions from PGP 2011, which included area under the curve (AUC) of receiver operator characteristics (ROC) curve (Fawcett, 2006) and simulations based on probability permutation. Each submission was assessed for accuracy of predicting the 40 binary traits and ROC and precision recall (PR) curves were generated. The ROC AUC was used as a mechanism to rank submissions. In order to assess statistical significance of a submission, a simulation-based permutation test was performed. Each submission was permuted 10,000 times and the resulting ROC AUC was determined and compared with the submission AUC. Details of each assessing method are discussed as follows.

ROC AUC: The ROC AUC is one of the most effective methods to evaluate the performance of prediction results (Fawcett, 2006). The ROC AUC of the prediction output of each submission was calculated based on the predicted probabilities and the true binary answers of 320 (10 genomes × 32 traits) genome trait pairs. We also plotted the PR curves based on these probabilities and the true answers. The ROC curves and PR curves were generated from the JAVA package "jstatplotter.jar" (developed by Kevin Van Bui at the University of Pittsburgh).

Simulations: We designed and developed simulations to test the hypothesis that permuted data derived from the actual submission could exceed the accuracy of the submitting group, and we used this to estimate the *P* value. The simulations were conducted in the following steps:

1. Generate 10,000 random results based on randomized permutation of the probabilities of the 320 genome and trait pairs of a submission;
2. Calculate the AUCs of the randomized results;
3. Calculate the distance of submission's AUC from the average AUCs of the permuted results;
4. Calculate *P* values based on the following equation:

$$P \text{ value} = \frac{\#\text{ of simulations with AUC greater than the submission}}{\#\text{ of simulations}} \quad (1)$$

Scoring: We also scored each submission based on the statistical analysis mentioned above, that is, ROC AUC and AUC distance from simulated results. By adding the two ranks of each submission together, we ranked the four submissions.

2.3 | PGP 2013

Data provided by CAGI for this challenge included 77 genomes and 291 phenotypic profiles, 214 of which were decoys (i.e., they did not match any genome). The phenotype profiles included a list of self-reported binary traits (e.g., asthma, breast cancer, lung cancer, colon polyps, and melanoma) and additional phenotypic and genetic information provided by PGP participants. Each profile included 239 traits (list of trait names that could be found in Supp. Material). The challenge required matching each genome to the appropriate phenotypic profile. There were 16 submissions from five groups. Each submission was a TSV file that contained probabilities for each participant genome to the 291 challenge profiles. Twenty-seven of the 77 genomes had additional genotypic data from 23andMe, and predictions on these were considered to be trivial and were removed from the analysis, which left 50 genomes being considered. The submissions and assessments are described as follows.

2.3.1 | Submissions

In 2013, we received both identified submissions from Group1, Group2, and Group3 as well as anonymous submissions from Group7 and Group8. Group1 and Group2 are groups that participated in PGP 2011, whereas Group3 was a new submitter group in PGP 2013. Their methodologies are described in Supp. Material, with overview shown in Table 3.

TABLE 3 Overview of methodology of PGP 2013 submissions

	Group7	Group2	Group3	Group1	Group8
Nongenomic information	None	None	None	None	Posts from PGP forum
Ancestry	No	No	1000 Genomes Project data	No	DIYDodecad 2.1
Traits used	107 traits from training set	All	Mendelian traits	All	Traits self-reported by certain participants
Filter variants for each trait	200	Selected nsSNPs from top 10 genes	In-house methods	HGMD and GWAS hits	N/A
Mathematical model	Naive Bayes	Correlation	Bayesian	Bayesian network	Trait specific
Parameter estimation	Estimate odds ratio of each variant from training set	Search PubMed in estimating disease similarity	From training set and guesswork	Literature and guesswork	N/A

2.3.2 | Assessments

As noted by one team (Group8), there were 108 profile that could be readily identified as being decoys, because the PGP website reported that no blood or saliva samples had been collected for these participants (Ball et al., 2014; PGP, 2016). However, only this team explicitly used such information to exclude these profiles in the prediction, which gave them an advantage that was not based on the merits of their genetic interpretation. Thus, we assessed the submissions only based on the 50 genomes without 23andMe data, and we excluded the 108 profile decoys in question from our evaluation, leaving 183. We used different approaches to evaluate the performances of the five different submissions, which included correct predictions and probability rankings.

Correct prediction: For this assessment, we considered a prediction for a genome to be correct if the trait profile assigned to the highest probability by the predictor was the correct trait profile for that individual. We counted the number of correct predictions by the following steps: (1) for each genome, determine which trait profile was assigned the highest probability; (2) for each genome, determine if this highest-ranked trait profile was correct; and (3) count for how many genomes, from a total of 50, the highest-ranked trait profile was the correct one. We also calculated the correct prediction rate by dividing the number of correct prediction by the total number of genomes, that is, 50.

Mean ranking of the true match: We calculated the mean ranking of the true match phenotype profile in each of 50 participants with genomes in each submission by the following steps: (1) rank all probabilities of the 183 phenotype profiles for a genome/participant; (2) identify the rank of the correct match for each participant; and (3) take the mean of the identified ranks of the correct matches.

Scoring: We also scored each submission based on the statistical analysis including criteria mentioned above, that is, the number/rate of correct predictions and the mean ranking of true matches. By adding the ranking numbers for the two above-mentioned scoring criteria together as the overall score of each submission, we got the overall placement of all 16 submissions.

2.4 | PGP 2015

The CAGI PGP challenge in 2015 was similar to 2013. We asked submitters to match 23 genomes to 101 phenotype profiles. The phenotype profiles remained largely unchanged from 2013. In CAGI 2015, there were 239 binary traits, which included both Mendelian traits and "complex" traits. The challenge included 78 additional decoy phenotype profiles that did not match any genome. The submissions and assessment methods are described below.

2.4.1 | Submissions

We received five submissions from four teams for the PGP Challenge of CAGI 2015. Each submission was formatted as a TSV file and contained probabilities for each participant genome to 101 profiles. The prediction methods of the identified submissions were from Group1, who applied an improved version of the method used in the 2013 challenge, Group4 and Group5. The latter two groups (i.e., Group4 and Group5) were new participants in the CAGI PGP challenges. We also received anonymous submissions from Group9. Their methods are described in Supp. Material.

2.4.2 | Assessments

We assessed the 2015 challenge using the assessing methods from the 2011 challenge and the 2013 challenge, which include ROC AUC, simulations, correct prediction, mean ranking of true match, with additional description for PGP 2015 introduced as below. Also, we included several other comparative analyses including an analysis of prediction performance of each participant, and an analysis of ancestry, blood type, and eye color of the participants in the challenge (Supp. Material).

ROC AUC: In PGP 2015, the ROC AUC of the prediction output of each submission was calculated based on the predicted probabilities and the true binary answers of 2,323 (23 genomes × 101 phenotype profiles) genome phenotype profile pairs. We also plotted the PR curves based on these probabilities and the true answers.

Correct prediction: For the assessment of PGP 2015, we also considered a prediction for a genome to be correct if the trait profile assigned the highest probability by the predictor was the correct trait profile for that individual. We computed the number and rate of correct predictions using the similar steps of the ones in PGP 2013, but with the total number of genomes as 23.

Mean ranking of the true match: We also calculated the mean ranking of the true match phenotype profile in each of 23 participants with genomes in each submission by the similar steps as the ones in PGP 2013, that is, (1) rank all probabilities of the 101 phenotype profiles for a genome/participant; and (2) identify the rank of the correct match for each participant; and (3) take the mean of the identified ranks of the correct matches.

Simulations: We also designed and developed simulations to test the hypothesis that permuted data derived from the actual submission could exceed the accuracy of the submitting group, and we used this to estimate a *P* value. The simulations were conducted in the similar steps of the ones in PGP 2011, that is, (1) generate 10,000 random results based on randomized permutation of the probabilities of the 2,323 genome and phenotype profile pairs of a submission; (2) calculate the AUCs of the randomized results; (3) calculate the distance of submission's AUC from the average AUCs of the permuted results; and (4) calculate *P* values based on Equation (1).

Scoring: We also scored each submission based on the statistical analysis including criteria mentioned above, that is, ROC AUC, the number/rate of correct predictions, the mean ranking of true matches, and simulations. By adding the ranking numbers for the four above-mentioned scoring criteria together as the overall score of each submission, we got the overall placement of all five submissions.

3 | RESULTS

Below are the assessment results of PGP 2011, PGP 2013, and PGP 2015.

3.1 | PGP 2011

The assessment results of PGP 2011 are shown in Figures 1–3 and Tables 4 and 5. Figure 1 is the ROC curves of the four binary prediction submissions of PGP 2011. Submissions of Group2_1, Group2_2, Group6, and Group1 have AUCs of 0.5121, 0.4631, 0.2785, and 0.8614, respectively, as shown in the second column of Table 4. Figure 2 shows the PR curves of the four submissions of PGP 2011. The simulation results of PGP 2011 are shown in the third and fourth columns of Table 4 and Figure 3. In Figure 3, the red line is the AUC of the submission from Group1, the green line is the average AUC of all the 10,000 permuted results for the Group1's submission, and the blue lines are the AUCs of all simulations. Note that only the Group1's submission is significant as it has a *P* value smaller than 1.0×10^{-5}. Table 5 shows the overall placement of all the four submissions based on their AUC and simulation results.

3.2 | PGP 2013

The assessment results of PGP 2013 are shown in Tables 6–8.

Table 6 shows the number of correct predictions and mean ranking of true matches. We can see that the submission of Group1 had the highest number of correct matches of 6 and the best mean rank of 25.4.

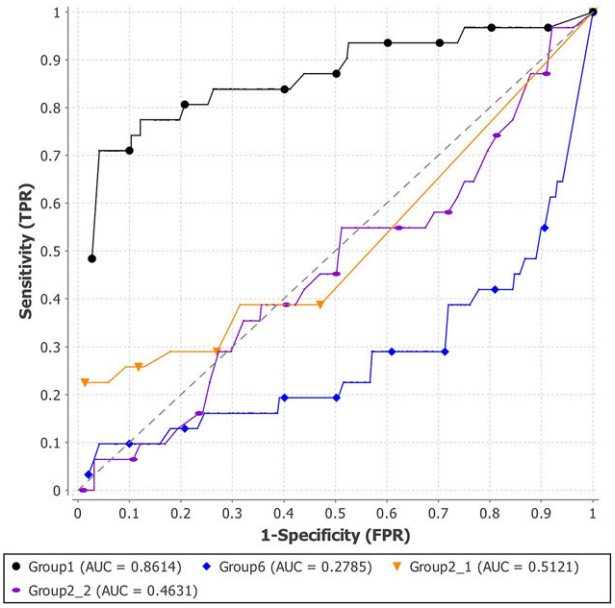

FIGURE 1 ROC AUC curves of the four submissions of PGP 2011

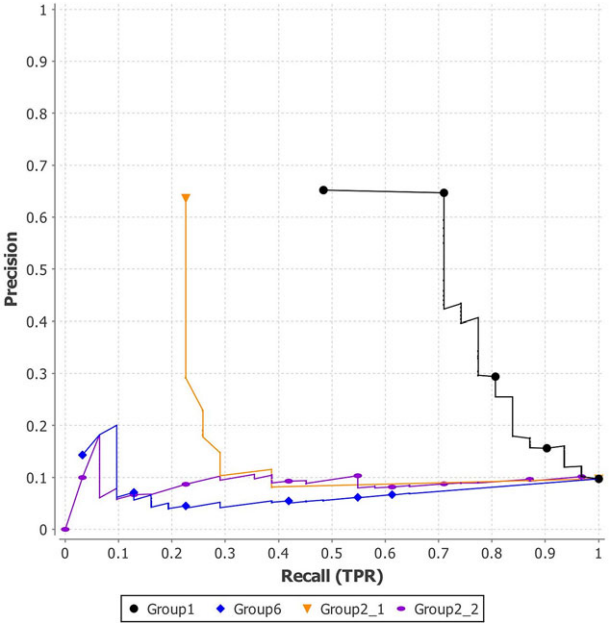

FIGURE 2 PR curves of the four submissions of PGP 2011

These are also shown in the overall ranking in Table 7. We also conducted simulations for the best submission from Group1. Specifically, for each genome, we randomly selected one phenotypic profile by uniform sampling. We made 10,000 such predictions for all 50 genomes. The largest number of genomes correctly predicted was four, implying that the observed prediction of six was statistically significant with a probability of less than 10^{-4}. Similarly, the best average rank from simulation was 75.2, nearly three times that of the observed rank from submission of Group1. Above all, the submission from Group1 group had the highest performance in the 2013 challenge.

In addition, Group1 and Group2 submitted predictions for individual traits based on genomes to an optional challenge of PGP 2013. In the main challenge, the submission from Group1 correctly identified

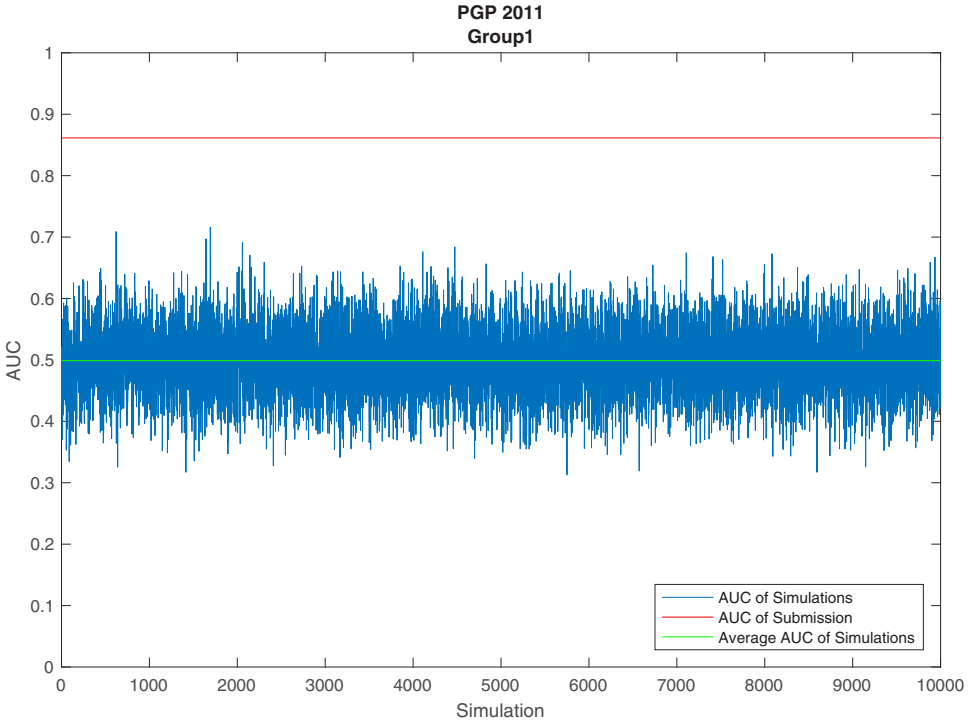

FIGURE 3 AUC comparison of simulative results and submission for the best submission from Group1 of PGP 2011

TABLE 4 Comparison of AUC, simulation results for the four submissions of PGP 2011

	AUC	AUC distance	*P* value
Group1	0.8614	0.3611	$< 1.0 \times 10^{-5}$
Group2_1	0.5121	0.0227	0.3158
Group2_2	0.4631	−0.0336	0.7332
Group6	0.2785	−0.2201	0.9999

TABLE 5 Ranking of the four PGP 2011 submissions based on several metrics

	Ranking via AUC	Ranking via AUC distance	Overall score	Overall placement
Group1	1	1	2	1
Group2_1	2	2	4	2
Group2_2	3	3	6	3
Group6	4	4	8	4

27 genomes with 23andMe data and also successfully tied five genomes to corresponding phenotypic profiles. We therefore focus on the predictions on these 32 participants/genomes. Empirically, we defined three risk categories (i.e., risk of a participant/genome having a phenotype) based on predicted probabilities: low risk with probability lying between 0.5 and 0.7, middle risk with probability lying between 0.7 and 0.9, and high risk with a probability greater than 0.9. For each participant, we counted the number of traits to which the participant provided a negative response in the survey. Generally, the submission made sparse predictions on disease risk for these 32 participants compared with reported number of positive traits. For all 32 genomes

TABLE 6 AUC, correct matches, and mean ranks of each submission for PGP 2013

	Number of correct predictions out of 50	Mean rank
Group1	6	25.4
Group8_2	5	35.2
Group8_4	5	37.3
Group8_3	5	37.9
Group8_1	4	39.8
Group3_2	2	37.1
Group3_3	2	45.1
Group3_1	1	35.8
Group2_4	1	68.5
Group7_1	0	59.4
Group2_1	0	66.6
Group2_2	0	67.0
Group2_5	0	69.2
Group2_3	0	79.3
Group2_6	0	84.5
Group7_2	0	183.0

correctly predicted in the main challenge, the submitters predicted that each genome carried, on average, an additional 2.4 low-risk traits, 1.1 middle-risk traits, and 0.2 high-risk traits, on top of the average 11.2 reported phenotype traits. Table 8 lists the number of reported and predicted traits for five participants without 23andMe data. From the

table, we can see that the predictions estimated that each participant potentially carried 4.4 additional traits on top of the reported traits.

3.3 | PGP 2015

The assessment results of PGP 2015 are shown in Figures 4–6 and in Tables 9 and 10. Figure 4 and the second column of Table 9 show the ROC curves and AUC of all the five submissions. We observed that the submission of Group1 had the highest AUC of 0.8531, whereas Group5_1 and Group5_2 had AUCs of 0.7982 and 0.7976, respectively. The AUC of the predictions of Group4 and Group9 were lower (i.e., 0.5170 and 0.6132). In addition, Figure 5 shows the PR curves of the five submissions. The figure demonstrates the higher performance of the submissions of Group1 and Group5.

The last two columns of Table 9 and Figure 6 show the simulation results. From Figure 6, we observed that the AUC of Group1's submission was higher than the permuted results. The last two columns of Table 9 show the distances between the submissions' AUCs and the average simulated AUCs, and the *P* values demonstrating significant differences from the permuted results. From these two columns, we can see that all the submissions were significantly different from random prediction, especially submissions from Group1 and Group5, with small *P* values (smaller than 1.0×10^{-5}).

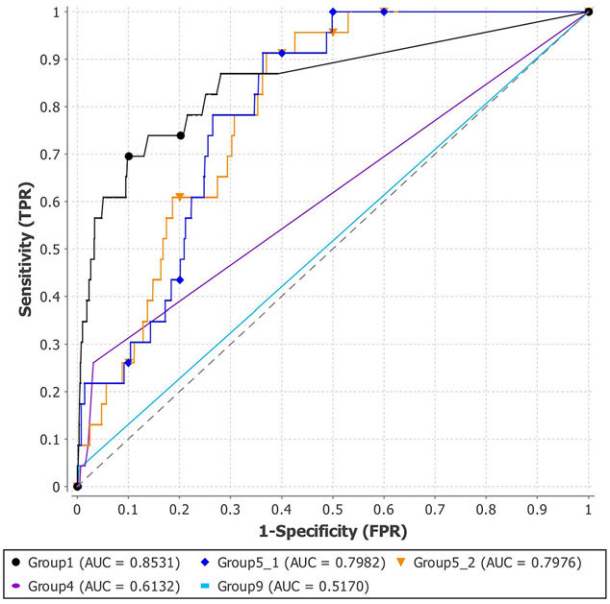

FIGURE 4 ROC AUC curves of the five submissions of PGP 2015

TABLE 7 Ranking of the 16 submissions in the PGP 2013 challenge

	Ranking via number of correct predictions	Ranking via mean rank	Overall score	Overall placement
Group1	1	1	2	1
Group8_2	2	2	4	2
Group8_4	2	5	7	3
Group8_3	2	6	8	4
Group3_2	6	4	10	5
Group3_1	8	3	11	6
Group8_1	5	7	12	7
Group3_3	6	8	14	8
Group7_1	10	9	19	9
Group2_4	8	12	20	10
Group2_1	10	10	20	10
Group2_2	10	11	21	12
Group2_5	10	13	23	13
Group2_3	10	14	24	14
Group2_6	10	15	25	15
Group7_2	10	16	26	16

TABLE 8 The number of reported and predicted traits for five participants without 23andMe data in PGP 2013

		Number of predicted traits		
huID	Number of reported traits	Low risk	Mid risk	High risk
hu619F51	3	2	1	1
huA05317	5	6	2	0
huEA4EE5	7	2	0	0
hu661AD0	4	3	1	0
hu5CD2C6	7	3	1	0

The third column of Table 9 shows the results of the correct predictions for each submission. From the column, we can see that Group1 got five correct predictions out of 23 true matches, whereas Group5 had three correct predictions. Other submissions had one correct prediction.

The fourth column of Table 9 shows the mean ranking of the true match phenotype profile for each of the 23 genomes/participant. From this column, we can see that Group1 obtained the true matches in the mean ranking of 16.93 of 101 for the 23 genomes, and Group5's submissions had 21.54 and 23.65 out of 101.

In addition, we also added the ranking of the correct matches among the five submissions to get the ranking of each participant/genome. From the results shown in the first two columns of Supp. Table S1, the best predicted participant was participant 14 (with a very small average ranking as compared with other participants), and the worst predicted participants were participants 16 and 21. We also listed profile information of each participant in the order of average ranking of each participant (Supp. Table S1). From the table, participant 14 was the only person with AB+ blood type among the 23 participants. In addition, from the ancestry analysis in Supp. Figure S1, we can see the two worst predicted participants are in the native American ancestry group, whereas the best predicted participants are across different groups.

3.4 | Comparisons of PGP challenges

Here, we compare the three PGP challenges, that is, PGP 2015, PGP 2013, and PGP 2011. From Table 1, we can see that PGP 2013 had more genomes and phenotype profiles, as well as more submissions,

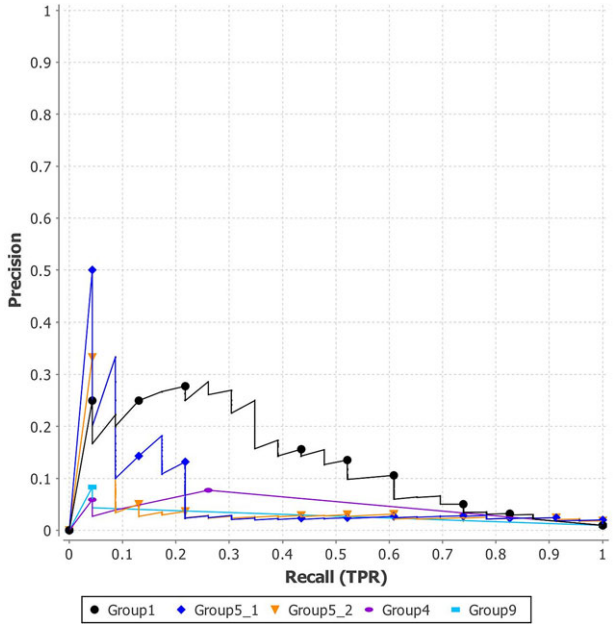

FIGURE 5 PR curves of the five submissions of PGP 2015

when compared with PGP 2011 and PGP 2015. From Tables 6 and 9, we can see that the highest proportion of correct predictions was in PGP 2015, with five out of 23 correct (around 21.74%), as compared with PGP 2013, where the best prediction was 6 out of 50 correct (around 12.00% correct). The top-ranked submission from Group1 in PGP 2015 had improved prediction performance with respect to the top-ranked submission, also from Group1, in PGP 2013, from 12.00% to 21.74% in terms of correct prediction. From the overall placements of submissions in three challenges in Tables 5, 7, and 10, we can see Group1 group had the best performance over the three challenges.

4 | DISCUSSION

We live in an era of genetic testing, either performed in the clinic or by direct-to-consumer (DTC) genetics services. Many results of genetic testing are difficult to interpret (Hudson, Javitt, Burke, Byers, & ASHG Social Issues Committee, 2007), though this is not always made sufficiently clear to patients or customers. It is important to understand the accuracy and limitations of currently available methodologies for interpreting the relationship between a patient's genetic information and phenotypic traits.

4.1 | PGP 2011

In the assessment of PGP 2011, we focused primarily on assessing binary trait predictions. Interestingly, one team had a vastly higher ROC AUC than the other teams. We hypothesized that Group1's submission was superior due to their accurate modeling of the overall rates of traits in the general population, a challenging problem due to the lack of available data and the educated guesswork that goes into building predicted population rates. This is an important issue and needs to be addressed in order to make accurate genomic predictions. Further, while AUC is relatively high, overall precision remains low for all teams, suggesting that individual trait prediction is still very difficult. The prediction of Group1 was, by our assessment, the only statistically significant submission.

4.2 | PGP 2013

Four groups (i.e., Group1, Group2, Group3, and Group7) used systematic approaches to modeling the probability of a trait given genomic data. One group (i.e., Group8) found participants with self-reported

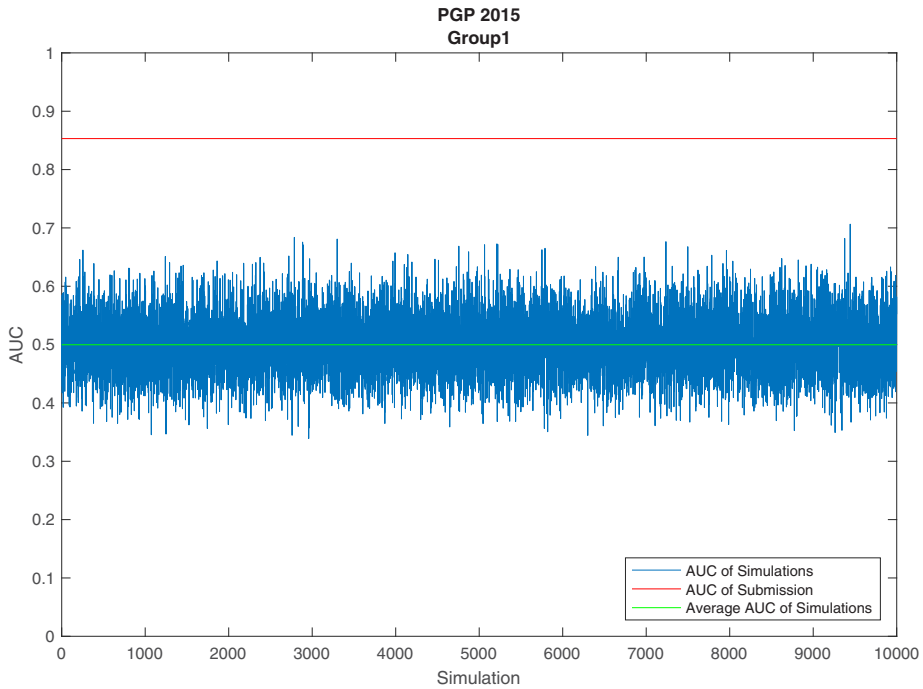

FIGURE 6 AUC comparison of simulative results and submission for the best submission from Group1 of PGP 2015

TABLE 9 AUC, correct matches, mean ranks, and simulation results of each submission for PGP 2015

	AUC	Number of correct predictions out of 23	Mean rank	AUC distance	*P* value
Group1	0.8531	5	16.93	0.3534	$< 1.0 \times 10^{-5}$
Group5_1	0.7982	3	23.65	0.2981	$< 1.0 \times 10^{-5}$
Group5_2	0.7976	3	21.54	0.2971	$< 1.0 \times 10^{-5}$
Group4	0.6132	1	39.59	0.1134	0.0001
Group9	0.5170	1	49.30	0.017	0.0387

TABLE 10 Ranking of the five submissions of PGP 2015

	Ranking via AUC	Ranking via number of correct predictions	Ranking via mean rank	Ranking via AUC distance	Overall score	Overall placement
Group1	1	1	1	1	4	1
Group5_1	2	2	3	2	9	2
Group5_2	3	2	2	3	10	3
Group4	4	4	4	4	16	4
Group9	5	4	5	5	19	5

traits and then searched the genetic data for specific genotypes of these traits. Their mathematical model contained multiple unknown parameters, such as variant frequency and penetrance. For many of the phenotypes, no accurate estimation or sufficient data exits for such parameters. This difficulty caused three groups to resort to intuitive guesswork when developing the models, which is hard to justify and which varied from group to group.

Interestingly, one team revealed that for certain participants, several discrepancies were observed between the provided traits and what genomic data could indicate. Such discrepancies, among other difficulties mentioned above, could have reduced the performance of complex models built for this challenge. On the other hand, some simple models, such as the Naive Bayesian method, may have been insufficient to capture weak statistical signal embedded in the highly heterogeneous data.

The two top performing groups successfully matched six and five genomes to their phenotypic profiles, although they used distinct approaches: one used a Bayesian network and the other used trait-specific models. The relatively accurate results from the submission of Group8_2 can be significantly attributed to the elimination of unlikely phenotypic profiles using additional information available from manual sleuthing on the Internet, including PGP participant discussion forums. Therefore, a straightforward way to improve the prediction for this challenge may be to apply the Bayesian network modeling (used by the top performing group, i.e., Group1) to curated data (which can be obtained through the similar approach used in the submission from Group8_2).

4.3 | PGP 2015

In our assessment, we also analyzed and compared ancestry, blood type, and eye color of the 101 participants (Supp. Material). These features are likely the source of a number of identified matches and appear to be driving predictive factors when identifying participants. For instance, the participant whom most teams correctly identified had the rare AB+ blood type. This feature could be one of the main factors in identification for this participant.

5 | CONCLUSION

We show that the ability to predict a broad profile of phenotypes from genotype is improving, and highly statistically significant, but still lacks the accuracy to be definitive. We also show that several major factors, including blood type and ancestry, increase the rate of correctly identifying individuals when compared with identification methods using rare traits. Notably, Group1, which performed well throughout each of the experiments, improved from predicting six out of 50 correctly (12%) in 2013 to five out of 23 (~21%) in 2015, a more than twofold improvement in performance. While these results are limited by small dataset, it is clear that we can still greatly improve our ability to definitively match an individual's phenotype with their genotype. In the future, we would like to include more genomes and trait profiles in the challenge and hope to see improvement of prediction accuracy.

ACKNOWLEDGMENT

The authors would like to thank anonymous predictors and other CAGI PGP participants for their contributions to the challenges. The CAGI experiment coordination is supported by NIH U41 HG007346 and the CAGI conference by NIH R13 HG006650. This work was also funded by NIH funding R01LM0009722, UL1TR000423, R01MH105524, U01 HG008473 and NSF funding 1458443.

REFERENCES

1000 Genomes Project Consortium. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319)1061–1073.

Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Current Protocols in Human Genetics*, Chapter 7:Unit7.20.

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*, 9

Anstee, D. J. (2009). Red cell genotyping and the future of pretransfusion testing. *Blood*, *114*(2), 248–256.

Ball, M. P., Bobe, J. R., Chou, M. F., Clegg, T., Estep, P. W., Lunshof, J. E., … Church, G. M. (2014). Harvard Personal Genome Project: Lessons from participatory public research. *Genome Medicine*, *6*(2), 10 (1–7).

Bodmer, W., & Bonilla, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. *Nature Genetics*, *40*(6), 695–701.

Burdett, T., Hall, P. N., Hastings, E., Hindorff, L. A., Junkins, H. A., Klemm, A. K., … Welter, D. (2015). The NHGRI-EBI Catalog of published genome-wide association studies. Retrieved from http://www.ebi.ac.uk/gwas

CAGI (Critical Assessment of Genome Interpretation). (2016). Welcome to the CAGI experiment. Retrieved from genomeinterpretation.org

Carter, H., Douville, C., Stenson, P. D., Cooper, D. N., & Karchin, R. (2013). Identifying Mendelian disease genes with the Variant Effect Scoring Tool. *BMC Genomics*, *14*(Suppl 3), S3 (1–16).

Chen, Y. C., Douville, C., Wang, C., Niknafs, N., Yeo, G., Beleva-Guthrie, V., … Karchin, R. (2014). A probabilistic model to predict clinical phenotypic traits from genome sequencing. *PLoS Computational Biology*, *10*(9), e1003825 (1–11).

Cooper, D. N., & Krawczak, M. (1996). Human Gene Mutation Database. *Human Genetics*, *98*(5), 629.

Douville, C., Masica, D. L., Stenson, P. D., Cooper, D. N., Gygax, D. M., Kim, R., … Karchin, R. (2016). Assessing the pathogenicity of insertion and deletion variants with the Variant Effect Scoring Tool (VEST-Indel). *Human Mutation*, *37*(1), 28–35.

Exome Aggregation Consortium (ExAC). (2015). ExAC Browser. Cambridge, MA. Retrieved from http://exac.broadinstitute.org.

Fang, H., & Gough, J. (2013). dcGO: Database of domain-centric ontologies on functions, phenotypes, diseases and more. *Nucleic Acids Research*, *41*(Database issue), D536–D544.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*, 861–874.

Friedberg, I., Wass, M. N., Mooney, S. D., & Radivojac, P. (2015). Ten simple rules for a community computational challenge. *PLoS Computational Biology*, *11*(4), e1004150 (1–5).

Gage, B. F., Eby, C., Johnson, J. A., Deych, E., Rieder, M. J., Ridker, P. M., … McLeod, H. L. (2008). Use of pharmacogenetic and clinical factors to predict the therapeutic dose of warfarin. *Clinical Pharmacology & Therapeutics*, *84*(3), 326–331.

Gonzalez-Perez, A., & Lopez-Bigas, N. (2011). Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. The *American Journal of Human Genetics*, *88*(4), 440–449.

Hindorff, L. A., MacArthur, J., Morales, J., Junkins, H. A., Hall, P. N., Klemm, A. K., & Manolio, T. A. (2015). A catalog of published genome-wide association studies. Retrieved from http://www.genome.gov/gwastudies.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(23), 9362–9367.

Hudson, K., Javitt, G., Burke, W., & Byers, P., & ASHG Social Issues Committee. (2007). ASHG statement on direct-to-consumer genetic testing in the United States. *The American Journal of Human Genetics*, *81*(3), 635–637.

Karolchik, D., Hinrichs, A. S., & Kent, W. J. (2009). The UCSC Genome Browser. *Current Protocols in Bioinformatics*, Chapter 1:Unit 1.4.

McKusick, V. A. (2007). Mendelian inheritance in man and its online version, OMIM. *The American Journal of Human Genetics*, *80*(4), 588–604.

Ng, P. C., & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research*, *31*(13), 3812–3814.

Oates, M. E., Stahlhacke, J., Vavoulis, D. V., Smithers, B., Rackham, O. J., Sardar, A. J., … Gough, J. (2015). The SUPERFAMILY 1.75 database in 2014: A doubling of data. *Nucleic Acids Research*, *43*(Database issue), D227–D233.

PGP (Personal Genome Project). (2016). Personal Genome Project: Harvard. Retrieved from personalgenomes.org.

Rich-Edwards, J. W., Colditz, G. A., Stampfer, M. J., Willett, W. C., Gillman, M. W., Hennekens, C. H., … Manson, J. E. (1999). Birthweight and the risk for type 2 diabetes mellitus in adult women. *Annals of Internal Medicine*, *130*(4 Pt 1), 278–284.

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L., Edwards, K. J., … Gaunt, T. R. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Human Mutation*, *34*(1), 57–65.

Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., … NHLBI Exome Sequencing Project. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, *337*(6090), 64–69.

The International HapMap Consortium. (2003). The International HapMap Project. *Nature*, *426*, 789–796.

Walsh, S., Liu, F., Wollstein, A., Kovatsi, L., Ralf, A., Kosiniak-Kamysz, A., … Kayser, M. (2013). The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. *Forensic Science International: Genetics*, *7*, 98–115.

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164 (1–7).

Yue, P., Melamud, E., & Moult, J. (2006). SNPs3D: Candidate gene and SNP selection for association studies. *BMC Bioinformatics*, *7*, 166 (1–15).

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Cai B, Li B, Kiga N, et al. Matching phenotypes to whole genomes: Lessons learned from four iterations of the personal genome project community challenges. *Human Mutation*. 2017;00:1–11. https://doi.org/10.1002/humu.23265