Interpretation of Variants of Unknown Significance in a Large Database of Genotyped and Phenotyped Individuals

B.T. Naughton[1], A. Chowdry[1], J.M. Macpherson[1], G. Benton[1]

[1]23andMe, Inc, Mountain View, CA

Introduction

The interpretation of variants of unknown significance (VUS) from whole-genome sequence data is a substantial challenge in genetics. VUS are usually too rare to be amenable to genome-wide association studies and so traditionally have been interpreted with reference to the primary literature (especially for high-penetrance or Mendelian mutations) or by computational methods (e.g., SIFT, PolyPhen). While these methods can provide useful insights, they are often limited by a lack of data and presence of false positives in the primary literature and by algorithms that inform about the effect of the variant on the protein and not on the disease state.

Here we present data demonstrating how the 23andMe database can be used to empirically determine the phenotypic effect of VUS. We present statistics on five previously-characterized variants in *BRCA1* and *BRCA2.* In a real-world example, we analyze a VUS in *MLH1* from a sequenced exome that was suspected to be cancer-causing.

Methods

23andMe, a personal genomics company (*http://www.23andme.com*), has assembled a database of genotypes for over 180,000 individuals, over 100,000 of whom have consented to participate in research and answered at least one research question. Participants answer research questions on the 23andMe website on topics as diverse as their medical and drug history, personality, lifestyle and exercise (Fig. 1).

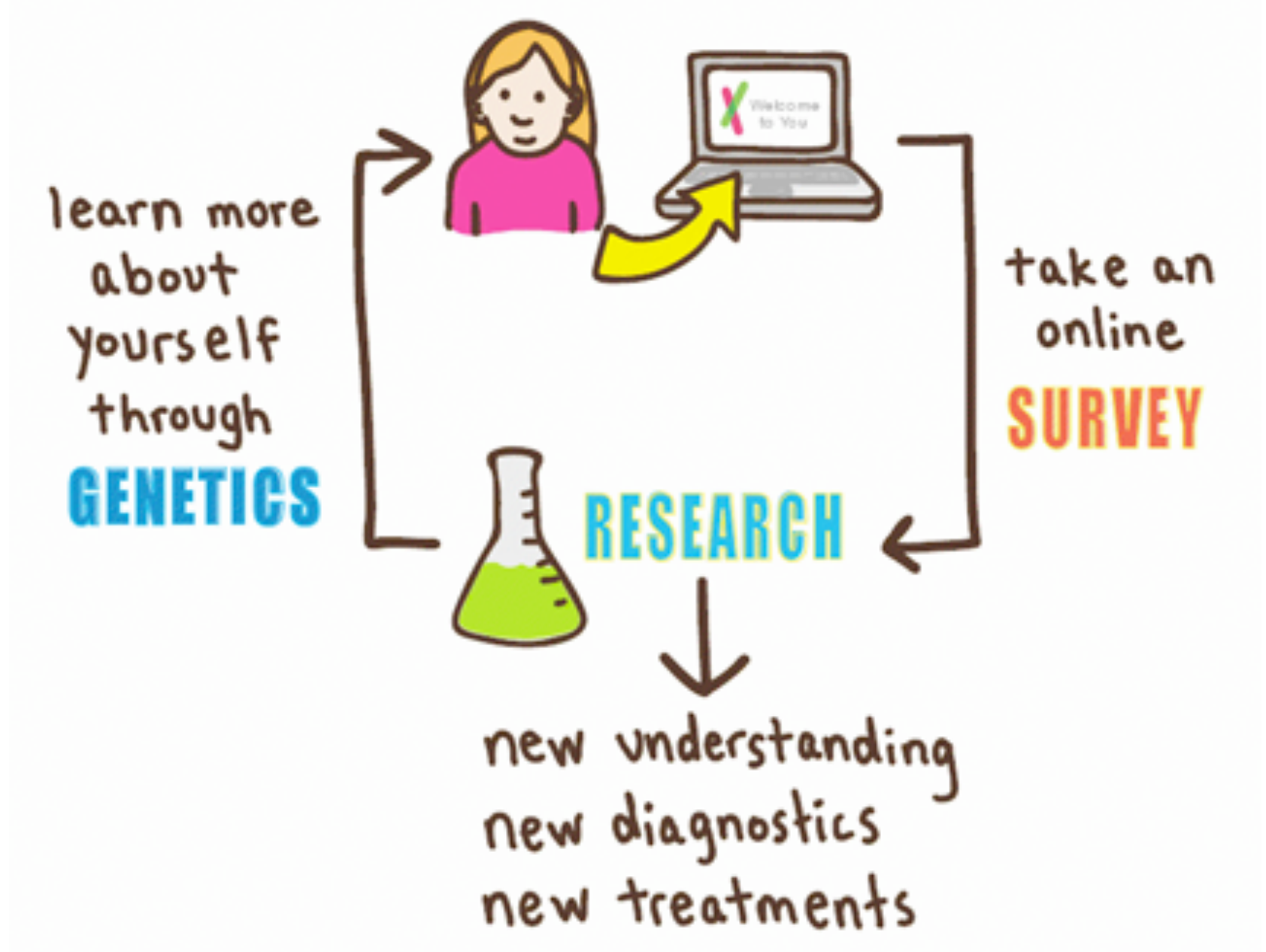

Figure 1. 23andMe's research platform engages consumers in research through the 23andMe website

23andMe uses this database to conduct genome-wide association studies (GWAS) [1,2] and phenome-wide association studies (PheWAS) (Fig. 2). 23andMe's database is notable for the breadth of phenotypic data it contains.

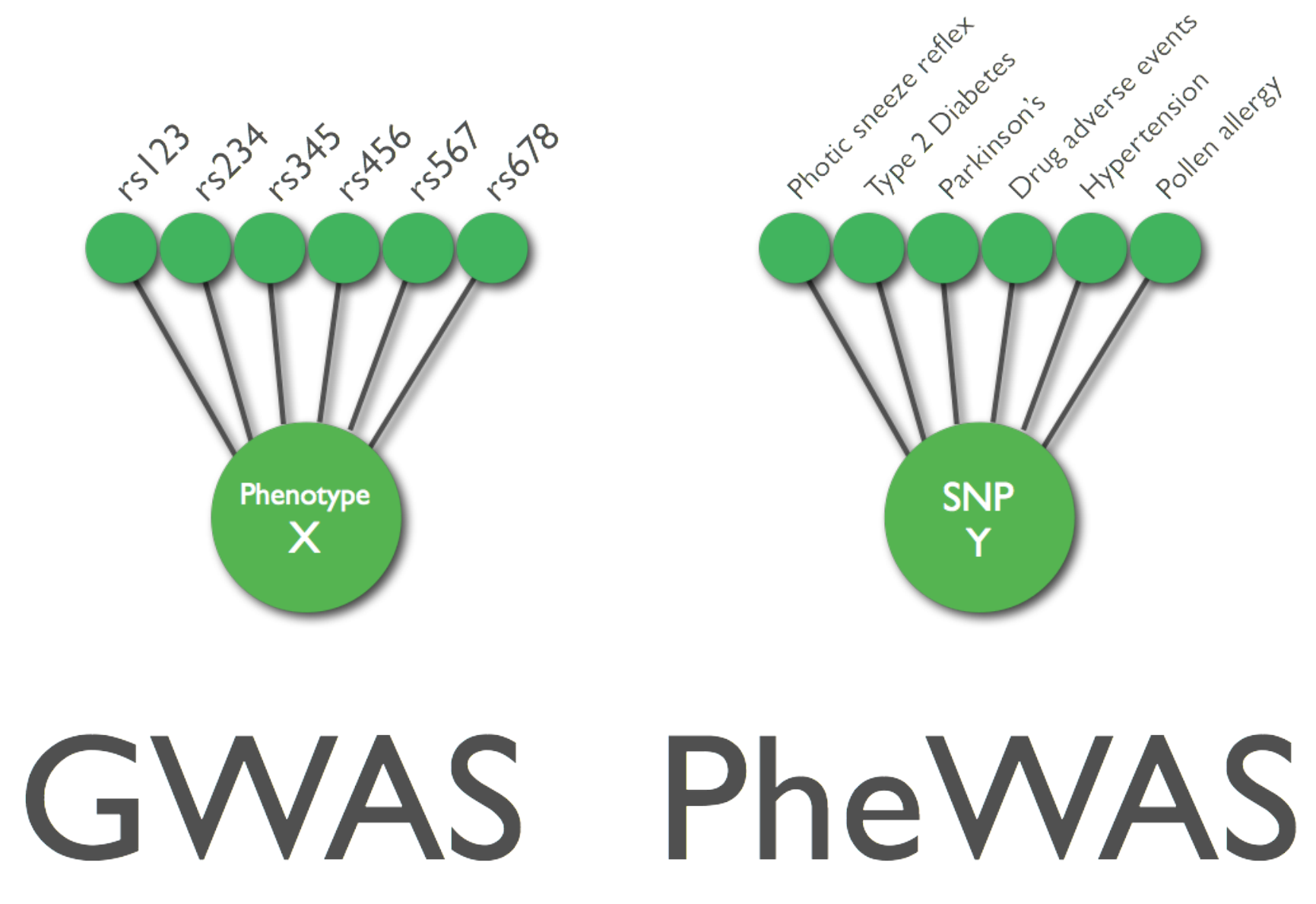

Figure 2. In a genome-wide association study (GWAS), a phenotype is tested for association with up to millions of SNPs. In a phenome-wide association study (PheWAS) a single variant is tested against a broad set of phenotypes.

Table 1. The 23andMe Research Platform: Key Statistics
>180,000 genotyped customers
>150,000 genotyped customers consented for research
1,000,000 SNP custom chip
30,000 Mendelian disease variants
>1000 different phenotypes surveyed
>80 million total phenotypic data-points
1 million new phenotypic data-points being added each week

Results

BRCA1 / BRCA2

We analyzed five variants in *BRCA1* and *BRCA2* using self-reported data from the 23andMe database (Table 2). We show that the BRCA mutations 5382insC and 6174delT are significantly associated with an increased risk for breast cancer. Conversely, we show that the BRCA mutations R841W and S1040N are likely benign variants that are not associated with increased breast cancer risk [4]. Though suggestive, the 185delAG mutation is not significantly associated with breast cancer. Our power calculations show that given the sample size there was a good chance we would not detect the effect.

Table 2. Case Study: Evaluating *BRCA1* and *BRCA2* Mutations with the 23andMe Database

Gene and variant	Number of breast cancers reported	Total number of variant carriers	Permutation test empirical p-value	Power to detect the effect
BRCA1 185delAG	2	9	0.061	0.60
BRCA1 5382insC	4	7	<0.001	0.84
BRCA2 6174delT	8	18	<0.001	0.79
BRCA1 R841W	8	101	0.972	>0.99
BRCA1 S1040N	19	407	0.852	>0.99

All data are from female 23andMe customers of European ancestry, over age 40, who have completed a survey on their cancer status. The "empirical p-value" column represents the fraction of times in 10,000 iterations we see a greater number of breast cancers in a random, matched subset of the database than in this set of carriers. Individuals are matched on sex, age and ancestry. The penetrances of 185delAG, 5283insC and 6174delT for breast cancer have been estimated at 46%, 46% and 43% respectively [3]. We use the more conservative 43% number for the known-neutral R841W and S1040N variants. These penetrances are used to calculate, by simulation, our power to detect the effect at $\alpha=0.05$.

MLH1 P603R

An individual with a family history of pancreatic cancer had his exome sequenced and was found to carry the P603R mutation in the cancer-related gene, *MLH1*. Based on the available data, this mutation was believed to be the most likely cause of cancer in his family. However, at least one journal article had suggested that the variant was neutral [5]. The *MLH1* P603R variant had been previously curated by 23andMe and was present on the 23andMe genotyping chip.

23andMe's database shows lack of association to cancer for P603R

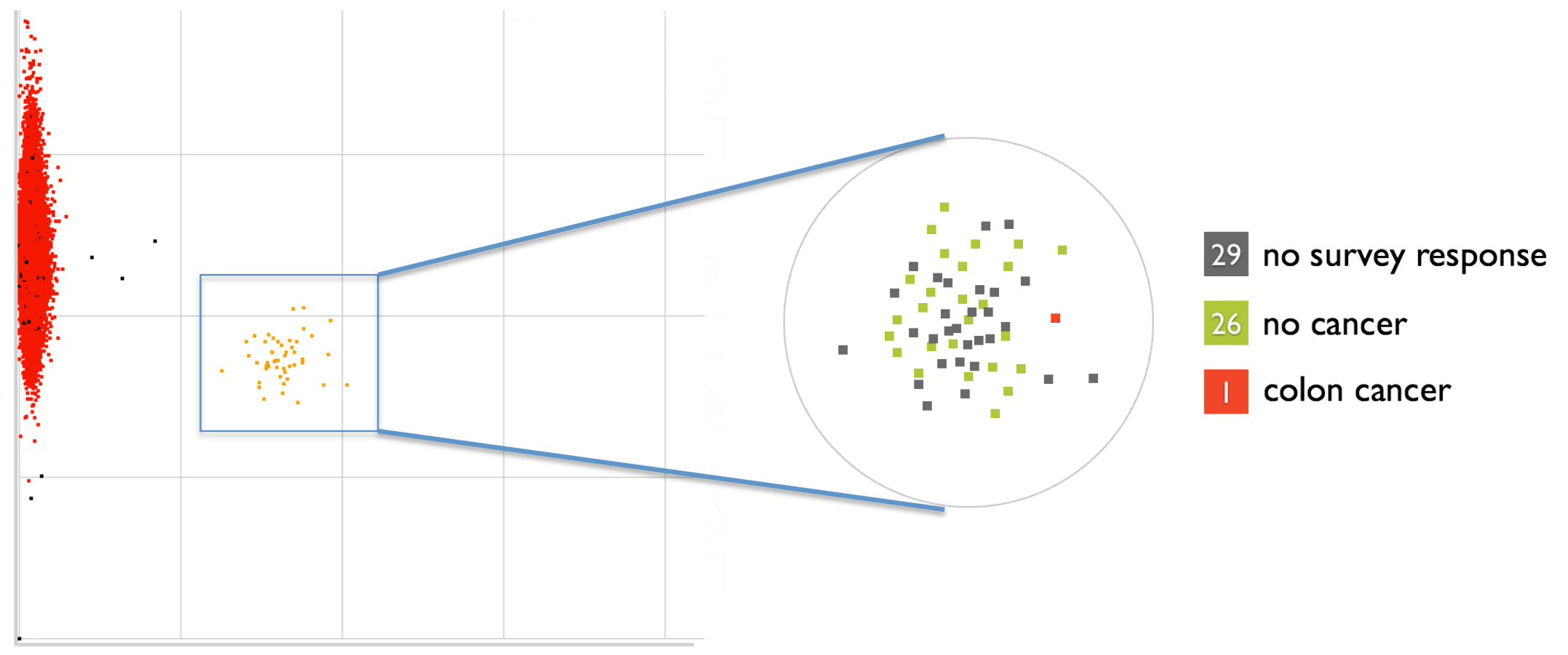

Figure 3. A genotyping cluster plot of the 23andMe database, including 56 carriers of MLH1 P603R. In the zoomed-in inset we show in illustration the breakdown of carriers by survey response.

The P603R variant was analyzed in the same manner as the BRCA1 and BRCA2 variants above. 56 individuals in the database were found to carry the P603R variant. 27 of 56 had answered a cancer survey on 23andMe. Only 1 out of 27 declared that they have cancer or have had cancer in the past. A permutation test showed no evidence for an increased probability of cancer given these data, in agreement with the literature [5].

Discussion

Due to the extensive phenotyping of our cohort (over 80 million phenotypic data points) and the large number of rare variants curated on our custom genotyping chip, the method described here is applicable to a large number of variants and phenotypes. Our proof-of-principle experiments show that, given sufficient sample size, our database can help uncover the phenotypic effects of high-penetrance variants.

Since curated mutation databases are believed to have high error rates [6], we believe that a primary use-case for the 23andMe database, as evinced by the P603R example, is to show evidence for a lack of association to a putatively-associated disease.

Acknowledgments

We thank 23andMe's customers who consented to participate in research for enabling this study. We also thank the employees of 23andMe who contributed to the development of the infrastructure that made this research possible.

References and Resources

1. Eriksson N. et al. Web-based, participant-driven studies yield novel genetic associations for common traits. *PLOS Genetics* 2010 **6** (6):e1000993
2. Do C. B. et al. Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease *PLOS Genetics* 2011 **7**(6):e1002141
3. Chen S. Characterization of BRCA1 and BRCA2 mutations in a large United States sample. *J Clin Oncol* 2006 **24**(6):863-71
4. Goldgar D. E. et al. Integrated evaluation of DNA sequence variants of unknown clinical significance: application to BRCA1 and BRCA2 *Am J Hum Genet* 2004 **75**(4):535-44
5. Muller-Koch Y. et al. Sixteen rare sequence variants of the hMLH1 and hMSH2 genes found in a cohort of 254 suspected HNPCC (hereditary non-polyposis colorectal cancer) patients: mutations or polymorphisms? *Eur J Med Res* 2001 **6**(11):473-82
6. Tong M. T. et al. Automated validation of genetic variants from large databases: ensuring that variant references refer to the same genomic locations *Bioinformatics* 2011 **27**(6):891-3