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

B.T. Naughton¹, A. Chowdry¹, J.M. Macpherson¹, G. Benton¹
²23andMe, 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).

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

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.

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.

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-penetration 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 evidenced 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