Gene-specific allele frequency thresholds for benign evidence to empower variant interpretation

Dajun Qian, Shuwei Li, Brice Sarver, Yuan Tian, Aaron Elliott, Hsiao-Mei Lu, Mary Helen Black

Allele frequency is often used as evidence of whether a variant is likely to be causative for a rare disease. However, current assessments of allele frequency for variant classification rely on either fixed or prevalence-adjusted thresholds and make no use of gene-specific information on variant pathogenicity. The publicly available Genome Aggregation Database provides an unprecedented spectrum of population-based human genetic variation that can be leveraged to examine the relationship between allele frequency and variant pathogenicity in classified variants on a gene-by-gene basis. Using a cohort of 176,280 patients who underwent genetic testing at a single diagnostic laboratory in 2012-2016, we assembled a training dataset of 1,388 classified missense variants in 10 genes (*BRCA1*, *BRCA2*, *CDH1*, *PALB2*, *PTEN*, *TP53*, *MLH1*, *MSH2*, *MSH6* and *PMS2)* included on hereditary cancer panels. The number of variants per gene ranged 44 to 438. We developed a constrained distribution fitting (CDF) approach to quantify gene-specific allele frequency thresholds (AFT) using data mining techniques of bounded constraints, monotonic distribution fitting and bootstrap sampling, each targeted to conservative estimates in case of uncertainty. Across variants in all 10 genes, positive predictive values (PPVs) for benign evidence were 0.40 ± 0.33, 0.43 ± 0.33 and 0.47 ± 0.33 using AFTs designated by a fixed 1% cutoff, prevalence-adjusted method, and CDF method, respectively. Negative predictive values (NPVs) were 1.00 ± 0.00 for fixed 1% cutoff and CDF methods, and 0.88 ± 0.31 for prevalence-adjusted method. Thus, the AFTs estimated by CDF showed 9% to 17% higher PPV than the other two methods and 12% higher NPV than the prevalence-adjusted method. Notably, differences between gene-specific AFTs estimated by CDF vs. other methods were striking for several genes. For example, using the CDF method, the AFT for benign evidence was as low as 0.0019% for *CDH1*, due to extremely rare pathogenic/likely pathogenic variants and as high as 0.43% for *PMS2* due to pathogenic/likely pathogenic variants having a wide range of frequencies. In contrast, AFTs estimated by the prevalence-adjusted method were nearly identical from 0.060% to 0.063% for both genes. Our results underscore the tremendous need for and practical usefulness of gene-specific allele frequency thresholds for benign evidence to empower variant interpretation.

Key words: variant classification; allele frequency; gnomAD; data mining

Characters with spaces: 2,427