Statistical Modelling of Population-level Exonic Variant Frequency Data with an Emphasis on Rare Variants

Yining Shi¹ and Shelley Bull^{1,2}

¹ Dalla Lana School of Public Health, University of Toronto, Toronto, Canada

² Lunenfeld-Tanenbaum Research Institute, Sinai Health, Toronto, Canada

E-mail: yiningsyn.shi@mail.utoronto.ca

Received May 15, 2020

Accepted for publication June 1, 2020

Published online September 5, 2021

Abstract

Introduction: Rare variants with allele frequency smaller than 1% are postulated to be associated with disease susceptibility. Since allele frequencies vary globally, the use of population control data that does not match the study population can produce bias. The research question is to identify factors that explain variation in allele frequency across populations. The secondary question is to evaluate the potential bias in using population as control data when studying variants. We use data from gnomAD (Genome Aggregation Database) to answer these questions.

Methods: We apply each of three model formulations: linear, logistic, and Poisson to explain how the frequency or count of variants depends on population subgroup/ancestry, functional annotation, sex, and disease status. We also evaluate interactions between population subgroups and functional annotation.

Results: For very rare variants (allele frequency $< 0.1\%$), likelihood ratio testing (LRT) provides evidence that allele frequencies vary with functional annotation and population in all three model formulations. By LRT, interactions of population and functional annotation are significant in the logistic model and the Poisson model. The goodness-of-fit statistics show a better fit in the linear model compared to low frequency variants.

Conclusion: We observe that population and functional annotation affect variant frequencies, and conclude that detection of differences across populations and annotations is model scale-dependent, especially for different degrees of rareness. Therefore, statisticians need to carefully consider the potential for bias when using gnomAD as control data. Moreover, gnomAD is a great resource for studies dealing with rare variants.