Toward understanding inflated type I errors in rare-variant analyses using publicly available sequence resources

Jung Kim[1], Stephen W. Hartley[1], Danielle M. Kayardi[1], Mingyi Wang[2,3], Dongjian Wu[2,3], Lei Song[1], Bin Zhu[1], Gregory T. Armstrong[4], Smita Bhatia[5], Leslie L. Robison[4], Yutaka Yasui[4], Brian Carter[6], Neal Freedman[1], Stephen J. Chanock[1], Lindsay M. Morton[1], Sharon A. Savage[1], Douglas R. Stewart[1]

[1]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD; [2]Cancer Genomics Research Laboratory, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD; [3]Leidos Biomedical Research, Inc., Frederick National Laboratory for Cancer Research, Frederick, MD; [4]Department of Epidemiology and Cancer Control, St.Jude Children's Research Hospital, Memphis, TN; [5]Institute for Cancer Outcomes and Survivorship, University of Alabama at Birmingham, Birmingham, AL; [6]Department of Population Science, American Cancer Society, Atlanta, GA

Introduction. The availability of large-scale, publicly available population-based germline exome- and genome-sequenced cohorts (e.g. ESP, 1000 Genomes, gnomAD, UK10K) holds tremendous promise for advancing genetic research, and can be used as a reference for orthogonal analyses. Generally, each cohort utilizes distinct analytic pipelines and individual-level data (BAM/VCF) are frequently unavailable. The inability to match calling and filtering pipelines between a newly sequenced study and controls derived from a publicly available cohort can lead to inflated type I error.

Methods. We utilized large-scale data from two cohorts sequenced in the same laboratory (4300 children with cancer [cases], 597 cancer-free adults [controls]) to systematically examine the impact of both laboratory components (capture, library prep and sequencer) and variant-calling pipeline (single *vs.* multiple callers, joint *vs.* separate calling) elements on potential inflation, which can be large. We tested differences in the distribution of rare (MAF<1%) synonymous variants using Fisher's exact test, expecting null results as synonymous variants are unlikely to

be associated with cancer risk. To quantify type I error, we constructed quantile-quantile plots and determined lambda delta95 ($\lambda_{\Delta 95}$), which adjusts for the large number of variants with p=1.00.

Results. When cases and controls are called *using the same variant-calling and filtering pipeline*, we observed minimal deviation of genes from the null distribution ($\lambda_{\Delta 95}$=1.04), even if laboratory components were different, and calling was either joint *vs.* separate. To *investigate the effect of the variant-calling pipeline*, we separated cases into two groups (n=2000 each) where all samples have same the laboratory components. We observed major inflation ($\lambda_{\Delta 95}$=1.16), which was diminished by using HaplotypeCaller only with the same post-calling filters on both groups ($\lambda_{\Delta 95}$=0.99). Given the HaplotypeCaller-only diminished inflation, we then compared controls (using HaplotypCaller only) with gnomAD. In this analysis, we once again observed substantial inflation ($\lambda_{\Delta 95}$=1.10), likely due to differences in post-calling variant filters, random forest for gnomAD and GATK hard filters for cases.

Conclusion. Comparing variant frequencies between genomic datasets without implementing the same variant-calling and post-calling variant filtering pipeline can be problematic. Direct statistical comparisons using controls from publicly available sequenced cohorts need to be carefully considered and interpreted. Optimally, raw data from controls should be obtained and processed with cases in a dedicated pipeline.