UK Biobank Whole Exome Sequencing 300k Release: Analysis Best Practices

The current release of the UK Biobank exome sequencing data on 302,333 participants comprises single- and multi-sample variant data generated via the same protocols as were applied to the UKB 200k WES release in early 2021: all samples are processed with the OQFE mapping protocol, variants called with Deep Variant and aggregated into a multi-sample VCF with GLnexus (https://www.biorxiv.org/content/10.1101/2020.12.15.356360v1.full). The multi-sample VCF contains per-genotype metrics including depth and genotype qualities, allowing researchers to perform custom variant- and genotype-level as appropriate for their desired analyses. As such, a single unfiltered multi-sample VCF was provided for the 300k WES release along with the derived PLINK files. In response to feedback from the UK Biobank community, the 300k WES release also includes a set of auxiliary files to aide researchers in implementing basic best practices for genotype-phenotype association analyses.

UKB WES Filtering Best Practices for Genotype-Phenotype Association Analyses

The breadth and depth of UKB phenotypes provide researchers a broad landscape of association analyses, from single-variant tests to gene burden testing, across individual and aggregated phenotypes. While no one set of filtered genotypes can be optimized for all possible analyses, there are features fundamental to the UKB WES data that can lead to spurious association results if not accounted for.

Specifically, the UKB WES data was generated in two phases: the first 50k participants (Phase 1) and then the balance of the total 500k cohort (Phase 2). As described in the Phase 1 release manuscript (https://pubmed.ncbi.nlm.nih.gov/33087929/), the 50k release participants were selected to enrich for specific phenotypes. Given the non-random order of participant sequencing, variations in sequencing coverage that occur over long-term projects can manifest as spurious association results. The UKB community reported such spurious hits when single-variant tests were run on the unfiltered UKB WES 200k genotypes. As an example, Figure 1A shows all single-variant hits of the UKB WES 200k unfiltered genotypes tested against an asthma phenotype (PHE10_J45), indicating a large number of likely spurious variants with significant or near-significant P-values. Examination of these spurious hits in the UKB WES 200k unfiltered set indicates that these variants tend to be enriched for sample-genotypes with low per-genotype read depth.

As noted in the UKB WES 200k FAQ (https://www.ukbiobank.ac.uk/media/cfulxh52/uk-biobank-exome-release-faq_v9-december-2020.pdf, section 23.d), we suggest the inclusion of a batch covariate in association tests on these data to account for differences in oligo lots between Phases. These coverage heterogeneities can also be mitigated by a single variant-level filter requiring that at least 90% of all genotypes for a given variant (independent of variant allele zygosity) have a read depth of at least 10 (i.e. DP>=10). When this filter is applied to the UKB WES 200k data prior to association analysis, the results are largely devoid of the spurious hits (Fig. 1B).

Application of this depth filter (“90pct10dp”) is consistent across the UKB 200k and UKB 300k WES sets with respect to numbers of variants removed (Table 1). The filtering can also be performed directly on the multi-sample VCF with the bcftools (http://samtools.github.io/bcftools/bcftools.html) commands below:

```
bcftools norm -m - -f <reference> -Oz -o <normVCF> <inputVCF>
bcftools view -i 'F_PASS(DP>=10 & GT!="mis")> 0.9' -Oz -o <filtered_normVCF> <normVCF>
```

where <reference> can be found here:
https://ftp.ncbi.nlm.nih.gov/1000genomes/ftp/technical/reference/GRCh38_reference_genome/

Alternatively, with the provided helper file named 'filtered_variants.txt', which is a single-column text file containing variants failing the "90pct10dp" depth filter in the CHR:POS:REF:ALT format, the following command using PLINK 1.9 can be used to remove the filtered variants from the UKB 300k WES PLINK files:

```
plink --bfile <original> --out <filtered> --exclude filtered_variants.txt --keep-allele-order
```

Pre- and Post-filtering association results using UKB 200k WES data

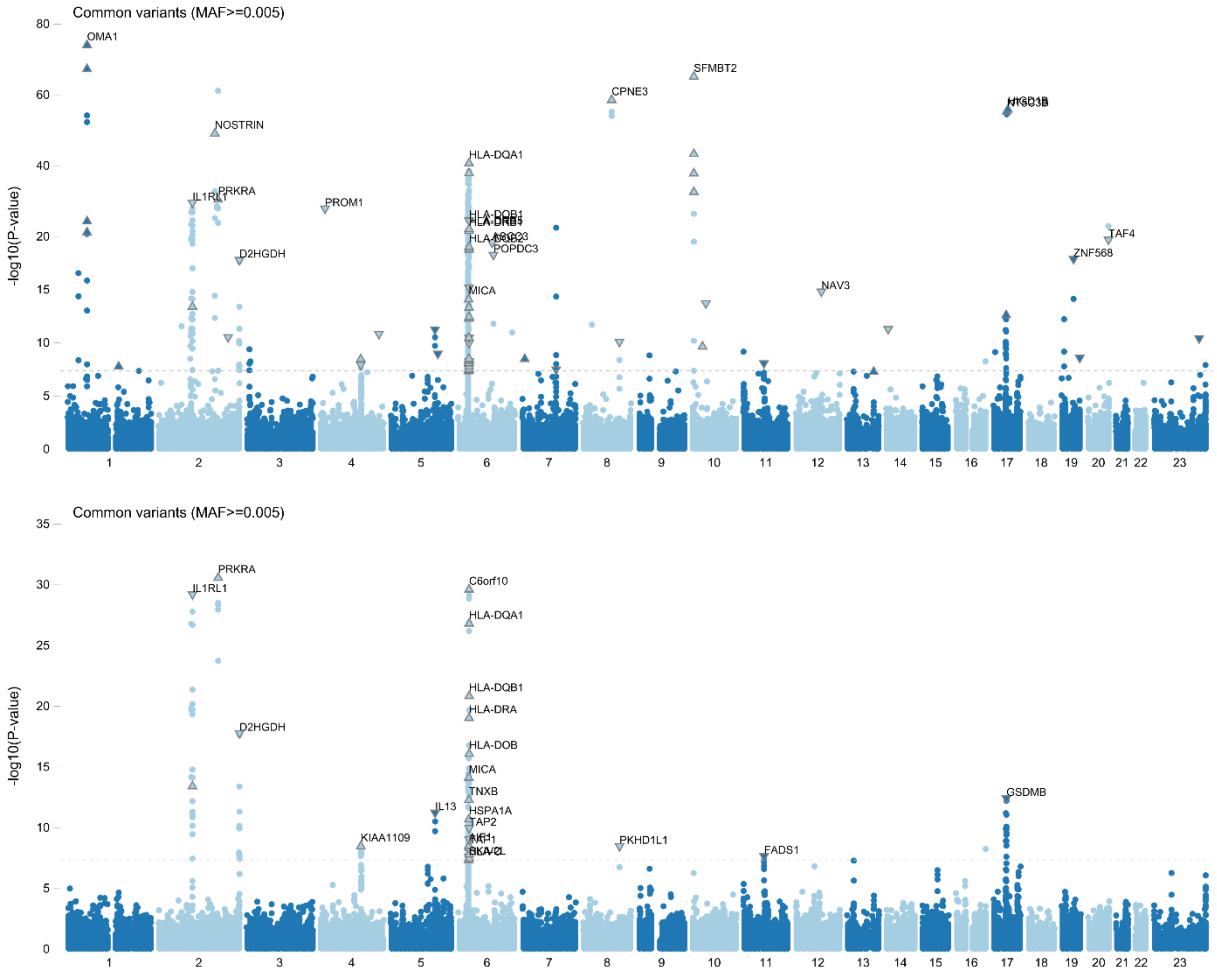

Figure 1*: Pre- and post-filtering UKB WES 200k association results with asthma phenotype (Phe10_J45. Subfigures A and B show results on the unfiltered UKB WES 200k genotypes and the 90% DP>10 variant-filtered genotypes, respectively. The tests were logistic regressions performed with standard covariates (10 PCs, age, sex, age^2, age_x_sex).*

Table 1

% Filtered Variants	SNP	Indel
UKB 200k	1.52%	5.54%
UKB 300k	1.57%	5.20%