Further information on the alt-contig mapping issue. This note was prepared by the Regeneron Genetics Center.

11 December 2019

UK Biobank 50k WES FE Dataset – alt-contig mapping issue

The UK Biobank 50k WES FE dataset is incorrectly mapped in a non-alt-aware manner. As specified by the Functionally Equivalent protocol (PMID: 30279509), FE crams are mapped with BWA to the full GRCh38 reference genome, which includes alternative contigs. These alternative contigs are derived from regions of the primary assembly (i.e. autosomal and sex chromosomes) and described in the ".alt" file included with the reference genome. The BWA mapping algorithm uses this reference alt file to accurately map reads to both the primary assembly and the alternative contigs (https://github.com/lh3/bwa/blob/master/README-alt.md). Specifically, inclusion of the reference alt file ensures that map quality scores are not penalized for when a read maps to both the primary assembly and a known alternative contig. The FE specification requires alt-aware mapping via inclusion of the reference alt file.

The UKB 50k WES FE CRAMs were incorrectly mapped to a version of the GRCh38 reference that does not include the reference alt file. Reads that map to both the primary assembly and an alternative contig will have deflated and often zero map qualities. Low-map-quality reads are generally ignored by variant callers and thus this error can result in an under-calling of variants in affected regions. Thus, all UKB 50k FE WES data (CRAMs, gVCFs, and PLINKs) are affected.

To facilitate existing and continued analysis of the existing FE WES UKB data while we work to provide the corrected data set, we here provide a BED file describing WES target regions impacted by the lack of alt-aware mapping. The capture design for the UKB WES data comprises 204,829 targets (39.20 MBp), 6,784 of which (1.27 MBp) overlap the alt-derived regions described in the reference alt file (provided as a BED file here). An additional 770 targets (0.26 MBp) were observed to have changes in either the number of reads mapped to the target or the average read-mapping quality in test samples when compared between alt-aware and non-alt-aware mappings.

This annotated BED file (xgen_plus_spikein_b38_alt_affected.bed) contains these 7,554 targets affected by the non-alt-aware mapping. We recommend that these regions be excluded from any analysis of the FE data and that all researchers consider how this mapping error might impact any results derived from the FE data.