Project Proposal

General: Pipeline for very-low frequency variant calling of somatic mutations

In general, somatic variant callers have been directed at detection of non-reference alleles (NRAs) present in tumor from surgically acquired tissue. Currently, the common practice is to limit variant detection to allele frequency >5% because artifacts associated with NGS (e.g. PCR errors, sequencing errors) have restricted the use of lower VAF thresholds. Thus, commonly used variant callers (e.g., Strelka2, Mutect2, LoFreq) for “low” frequency variant detection do not interrogate NRAs with a variant allele frequency (VAF) < 1-3%. The development of a variant caller pipeline to adjudicate all NRAs with a VAF < 1% based on pathogenicity of the variant rather than VAF has applications in understanding tumor phylogeny and the overall mutational landscape in solid tumor DNA through improved genotyping. In addition, a pipeline to detect very-low frequency variants would have applications in tumor-derived cell-free DNA from plasma (i.e., the liquid biopsy) where VAF is commonly <0.5%, particularly in detection of circulating tumor DNA associated with non-metastatic solid tumors.

Here, we propose to develop a pipeline using a python wrap around each module to handle the input/output between modules. To evaluate performance, sensitivity will be measured using a BAM files of cell-free DNA (coverage: ~3000X) from a healthy control where COSMIC exonic variants were synthetically injected across a range of VAFs. A .vcf of the injected positions and associated variants will be used as reference to determine sensitivity. Specificity will be measured using a BAM file from a separate healthy control sequenced under identical conditions. For all data sets, germline DNA is available as a reference to identify SNPs to support the internal variant caller. Analysis will be restricted to the exons and variant identification criteria at the end of the pipeline will be optimized to maximize sensitivity for a false discovery rate of <5%. The following workflow is proposed:

(input) BAM --> GATK\* (variant caller) to generate initial .vcf--> Isolation of exonic positions --> Snpeff\*\* (variant effect prediction) --> variant adjudication and generation of .vcf file\*\*\*

\*consider Freebayes as it has options that enable all NRAs to be identified

\*\* consider VEP if available as a stand-alone module

\*\*\* initial adjudication may consist of pathogenicity based on SIFT and/or Polyphen predictive algorithms