# Post analysis of SNV calls: Annotating, filtering and quality assessment

Yaron Butterfield*, Richard Corbett*, Steven JM Jones, Inanc Birol

Genome Sciences Centre, BC Cancer Agency, Vancouver, BC, Canada
* These authors contributed equally.

PASsiT
Post Alignment SNV Tools http://www.bcgsc.ca/platform/bioinfo/software/passit

With single nucleotide variations (SNVs) being identified in ever increasing numbers with next generation sequencing data, we are often interested in a breakdown of the effect of the mutations, and if a particular change is somatic or germline. We are also interested in high quality true positive mutations, which are increasingly important considering their use in clinical decisions. Here we present a suite of tools we developed to help us address these needs in our high-throughput sequencing and analysis environment at the British Columbia Genome Sciences Centre.

Before SNVs are annotated, a level of filtering can be applied to improve confidence and decrease false positives. SNVs are scored by a combination of allelic frequencies, strandedness, read positions, and SNV caller-reported quality. For each mutation, we count how many times the SNV or the reference allele is present, as well as which strand and read position the mutant base comes from. We then partition the read pileup file into groups of unique value for each of the metrics. By calculating the dbSNP concordance along the full scale of each of our metrics, we are able build a lookup table that to estimate this concordance for any SNV identified in a similar library. This allows us to take a pileup file and rank the identified SNVs by order of confidence.

SNVannotator is a python script part of the PASsiT package that takes a list of mutations from various sources such as Samtools' Pileup/varFilter, GFF, VarScan, VCF, or SNVMix, and classifies the effect of each SNV. Using reference data from Ensembl, SNVannotator outputs an annotated list of SNVs identifying if a mutation is intergenic or intragenic; and if the latter, what gene it is in, and if it's in the 5' or 3' UTR, intron or exon. If in an exon, the mutation is marked as synonymous or non-synonymous according to its effect. SNVs are identified as known or novel with respect to the list of polymorphisms recorded in the dbSNP repository. In addition, a list of novel non-synonymous SNVs is generated with the associated protein and its amino acid change resulting from the mutation. If a matched tumor/normal pair is given, SNVannotator identifies somatic non-synonymous mutations by comparing the SNV calls on both.

Using the called SNVs and their estimated zygosity states, we also identify regions of loss of heterozygosity (LOH). For each sample, genomic bins of consistent SNV zygosity states are used by a hidden Markov model (HMM) to identify genomic regions of consistent rates of heterozygosity. The HMM partitions each tumor genome into three states: normal heterozygosity, increased homozygosity (low), and total homozygosity (high), where the intermediate state of low homozygosity represents a genomic region where only a portion of the cellular population sampled had lost one of the alleles.