Introduction

What is Krunch?

Krunch is a general purpose kmer based variant calling tool that will call SNPs, indels, and any size deletion. Most existing variant callers first align reads to a reference genome and then look for discordant reads. Krunch identifies reads containing kmers that don't match the reference then assemblies those reads into contigs. These contigs are then aligned to the reference genome and based on the pattern of alignment variants are called. In essence Krunch first identifies reads containing any sort of variant then resolves what sort of variant it is by assembling variant containing reads and comparison of the assembled contigs with the reference genome.

The two major parameters that affect performance are the kmer sized used and the cutoff number of reads required to call a variant. The default kmer size and cutoff is 30 bps and 4 reads. These values should work well for most whole genome sequencing projects. If sequencing reads longer than 200 bp reads than a larger kmer size will improve sensitivity but kmer sizes larger than 75 basepairs will degrade performance due to the increased chance of a sequencing error falling within the kmer. For calling variants using high depth sequencing > 50X coverage a larger read cutoff would be appropriate to reduce the number of false positives.

Quick Start Example:

/path_caller/runCaller.sh -r /home/user/Genomes/hg39.fa -f /home/user/experiment/read_library.fastq

The only two required parameters are a reference genome in fasta format and a read library in fastq format. The output will be called SnpCalls.vcf and indelCalls.vcf and will be placed in the same directory the tool was run in unless a working directory was specified.

Usage:

runCaller.sh -r < reference genome > -f <fastq file> [-s <kmer size>] [-c
<count cutoff>] [-o <output dir>] -t [<number of cores>] [-h]

Where:

-r (required) Reference genome in fasta format.

-f <fastq file>
(required) Location of the fastq file you wish to call variants in. If you have a paired end dataset, then merge the two fastq files containing read 1 and read 2 before calling variants. Currently Krunch does not use mate pair information when calling variants this will likely change in the future.

-s <kmer size>
The kmer size you wish to use. The default value is 30, larger values increase the sensitivity but only to a certain point. Kmers larger than half the read size will not work very well. Kmers larger than ~75 basepairs regardless of read size will also negatively affect performance. A good rule of thumb is to use a kmer size as large as possible that is less than half the read size and smaller than 75 basepairs.

-c <count cutoff>
The number of supporting reads required before a variant is called.
The default value is 4, larger values increase specificity at the cost of sensitivity. If you have particularly high coverage datasets ~>50X increasing this cutoff is a good idea.

-o <output dir>
The output directory and location to store intermediate working files.
The default value is the current working directory. Intermediate working files may be large so plan accordingly.

-t Number of Cores default value is 4

-h Displays usage information and then exits