ONH Pipeline Outline

Kevin Stachelek, Jennifer Aparicio

10/19/2017

The current pipeline takes a directory with gzipped fastq files (.fastq.gz) as input to a loader shell script (run_onh_pipeline.sh) which feeds them to a python script (onh_pipeline.py) which, using the python subprocess module executes shell commands in the order shown below:

1. Trimmomatic
   - Remove adapters
   - Remove leading low quality or N bases
   - Remove trailing low quality or N bases
   - Scan the read with a n-base wide sliding window, cutting when the average quality per base drops below k
   - Drop reads below a given length
2. BwaMem
   - local alignment

Picard

3. SamFormatConverter
   - Convert a BAM file to a SAM file, or SAM to BAM. Input and output formats are determined by file extension.
4. SortSam
   - Sorts a SAM or BAM file
5. MarkDuplicates
   - Identifies duplicate reads.
6. AddOrReplaceReadGroups
   - Replace read groups in a BAM file
7. BuildBamIndex
   - Generates a BAM index ".bai" file.
8. Mosdepth
   - fast BAM/CRAM depth calculation for WGS, exome, or targeted sequencing.

GATK

9. BaseRecalibrator
   - Detect systematic errors in base quality scores
10. PrintReads
    - Write out sequence read data (for filtering, merging, subsetting etc)
11. VariantFiltration
    - Filter variant calls based on INFO and FORMAT annotations
12. SelectVariants
    - Select a subset of variants from a larger callset
13. HaplotypeCaller
    - Call germline SNPs and indels via local re-assembly of haplotypes
14. GenotypeGVCFs

- Perform joint genotyping on gVCF files produced by HaplotypeCaller

15. VariantRecalibrator
    - Build a recalibration model to score variant quality for filtering purposes
16. ApplyRecalibration
    - Apply a score cutoff to filter variants based on a recalibration table
17. CalculateGenotypePosteriors
    - Calculate genotype posterior likelihoods given panel data
18. VariantAnnotator
    - Annotate variant calls with context information

---

19. TableAnnovar
    - takes an input variant file (such as a VCF file) and generate a tab-delimited output file with many columns, each representing one set of annotations. Additionally, if the input is a VCF file, the program also generates a new output VCF file with the INFO field filled with annotation information.
20. VcfAnno
    - vcfanno allows you to quickly annotate your VCF with any number of INFO fields from any number of VCFs or BED files. I am using it to annotate
        1. gnomad minor allele frequency
        2. dbsnp ids
21. Genmod
    - GENMOD is a simple to use command line tool for annotating and analyzing genomic variations in the VCF file format. GENMOD can annotate genetic patterns of inheritance in vcf:s with single or multiple families of arbitrary size.