Genomics Data Science with Galaxy Final Project

Shengyuan Wang

Mar 27, 2020

Instructions:

The zip file fastq_bundle.zip contains six fastq files. These files contain targeted re-sequencing data for a father mother and daughter trio (identified as NA12877, NA12878, and NA12880 respectively). The data consists of raw reads from an Illumina MiSeq sequencer sequenced as paired ends (R1/R2) to 125bp in length.

Create a Galaxy workflow to identify polymorphic sites in all three individuals. Your workflow will need to map the three sets of paired reads to the appropriate reference genome. You will then need to use a variant caller to identify sites that appear to have strong support for the presence of a polymorphism, and call the genotype at that site for each sample.

You should report your results in VCF (variant call format). You should only include sites where the chance of a false positive call is 1 in 10,000 or better according to the VCF qual field.

Using your resulting VCF determine 1) the number of single nucleotide variants, 2) the number of insertion/deletion variants, 3) the number of multi-nucleotide variants, 4) the number of variants with multiple alternate alleles, and 5) the names of the 5 genes with the largest number of polymorphic sites.

Procedures and Results:

1. Upload all 6 fastq files to Galaxy (https://usegalaxy.org/), choose the reference genome as Human Feb. 2009 (GRCh37/hg19) (hg19), filetype as fastqsanger.
2. Use FastQC (Version 0.72+galaxy1) to perform quality control, all 6 sequences have mean sequence quality score at 37, indicate they have good quality.
3. Mapped paired-end sequences to reference genome using Map with BWA-MEM (Version 0.7.17.1), output mapped reads in BAM format.
4. Assigns all the reads to a single new read-group using AddOrReplaceReadGroups (Version 2.18.2.1).
5. Merge all 3 mapping BAM files using MergeSamFiles (Version 2.18.2.1).
6. Remove low quality position using Filter (Version 1.1.0).
7. Remove duplicates using MarkDuplicates (Version 2.18.2.2).
8. Grooming the merged BAM files using CleanSam (Version 2.18.2.1).
9. Find single-nucleotide polymorphisms using FreeBayes (Version 1.3.1).
10. Keep sites where the change of a false positive call is 1 in 10,000 or better using VCFfilter (Version 1.0.0_rc3+galaxy3), set Specify filtering value as QUAL > 40.
11. Use VCFfilter again to determine the number of:

- Single-nucleotide variants by “TYPE = snp”: 2295
- Insertion/deletion by “TYPE = ins” and “TYPE = del”: 258
- Multi-nucleotide variants by “TYPE = mnp”: 7
- Variants with multiple alternate alleles by “TYPE = complex”: 84

12. Use SnpEff eff (Version 4.3+T.galaxy1) to annotate variants. Download the result in csv format. Sort the genes based on variant_impact_MODIFIER. The name of the genes with the largest number of polymorphic sites are RBFOX1, ABAT, ADCY9, CACNA1H, LMF1.