SNP Calling Pipeline

Lukas Endler
10.08.2017

Overall Layout

- Preprocessing
- Mapping
- Alignment processing
- SNP calling
- Variant Filtering

Preprocessing

- Using bbtools
  - http://jgi.doe.gov/data-and-tools/bbtools/
- Mainly bbduk
  - Trimming
    - Only 5’ end, min base qual. 15
  - Removing adapter sequences (overlapping reads)
  - Removing primer sequences (amplicon approach)
- Also tried
  - Joining read pairs

Mapping

- BWA mem with standard parameters against both mouse and LCMV
- Only keep reads mapped to virus as proper pairs, no split alignments, no secondary alignments
- Also tried
  - BWA mem with high gap openening cost
  - Bowtie2
  - Mosaic
  - Some had lower InDel and SNP rates, but had no good way of assessing performance

Alignment Postprocessing

- Removing overlap of read pairs
  - BamUtils, just clips bases with lowest average BQ
    - https://genome.sph.umich.edu/wiki/BamUtil
    - Could be better – like samtools does
    - Needed as some tools are a bit dumb
- Downsampling to at most ~ 10K coverage
- Add InDel qualities/BAQ (lofreq dindel)
- Viterbi realignment method from lofreq (does not change anything much)
- To do
  - Try to do Base Quality Recalibration
  - Better realignment (maybe Indelrealigner)

SNP Calling

- Freebayes
  - Not directly useful for this purpose (not good undefined ploidy)
  - Looks at local haplotypes over certain windowsizes
  - Lots of metrics for filtering
- Varscan2
  - Simple algorithm
  - Use strandfilter
- LoFreq
  - Mainly uses BQ
  - Claims to reliably call low frequencies, already uses filters internally
  - MQ,BQ > = 20, mnin DP 75
- VPhaser2
  - Needed some hacks to be usable
  - Does not call variants fixed in a sample
  - Specifically for viral samples

Variant Filtering

- Did not perform thorough analysis up to now
  - Need truth and at least uncertain set of variants
  - Maybe use technical and biological replicates and overlaps between callers for true set
- Thresholds to define for
  - Calling Quality
  - Metrics like depths, strand bias, read position bias, homopolymer runs
  - Can be defined rigorously or just tentatively using an explorative approach (done for pooled sequencing removal of artifacts)

Hard Filtering

- Used for Freebayes results only
  - Strand bias
    - at least 10% of alternative allele counts supported by less common strand
    - P of Fishers Exact Test on strands of ref. to alternative allele > 0.01
  - Read position bias (RPP) < 30
  - Minimal count for a variant allele: 3

Legend

Green: synonymous_variant

Red:missense_variant

Orangered: stop lost

Orange stop gained

Orangered4: start lost

Pink: inframe deletion

Purple:: frameshift

Purple4: disruptive frameshift