Lab Notebook

Encelia Research

Sonal Singhal

sonal.singhal1@gmail.com

Beginning 6 October 2014

Contents

Monday, 6 October 2014

1 Variant Calling

GATK did a horrible job generating variants, and it was very very slow. So, all future work comparing variants will be just comparing Samtools (as generated by BWA and Bowtie) with Platypus (as generated by BWA and Bowtie). In general, it looks like Platypus generates more variants, so Samtools might be the more conservative set.

I will want to compare the number of variants found and the genotypes called.

Tuesday, 7 October 2014

1 Variant Calling

I wanted to get average coverage across the entire contig for each contig for each assembly. I did this by writing the script `get_depth.py`, which is a simple wrapper around the `samtools depth` script. I calculated depths using the BWA-generated BAM files. For population genomics analyses, I will drop any contigs which have lower than 5× coverage, though I suspect not many annotated contigs will have that low of coverage.

Wednesday, 8 October 2014

1 Variant Calling

In order to do population genomics between *E. palmeri* and *E. ventorum*, I am going to align reads from both populations to the same reference. I decided to use *E. palmeri* because it seemed marginally more complete. I modified the alignment script to do this, and it is called `5alignment_bwa_sameref.py`. All BAM files from this analysis will be called `*sameref*`.

Thursday, 9 October 2014

1 Variant Calling

I did some tests comparing Bowtie / BWA for alignment and Platypus / GATK / Samtools for variant calling.

- Bowtie / samtools (which had been my go to) does really poorly!
- Anything with GATK is a joke – misses so many really good variants.
- Anything with Platypus seems to find a lot of spurious variants.
- BWA / samtools looks like the new best pipeline.

But, it weirded me out that there is such seeming inconsistency across the alignment and variant callers for SNPs found. So I did a few things.

- Looked at if there is more consistency when depth is high. There is, but it is marginal.
- Looked at if there is more consistency when contigs are annotated. There is, but it is marginal.
- Looked at if depth is connected to annotation, and yes, annotated contigs need to have higher depth.
- Looked at the filter values for SNPs in the winning approach (bwa / samtools) and compared values across SNPs that were good (i.e., found in another variant call for the same raw data) or bad (i.e., unique to the given set).
  - looks like the only filter worth considering is MQ, or mapping quality
  - note that 24991 are bad
  - note that 113730 are good
  - getting rid of SNPS that are MQ <= 20 will lose very few real SNPs likely, so the only filter worth considering

These analyses are all on my desktop, under `/Users/singhal/encelia/analyses/coverage_snp/` and `/Useres/singhal/encelia/analyses/snp_analysis/`.