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 $5\times$ 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 `/Users/singhal/encelia/analyses/snp_analysis/`.

Monday, 13 October 2014 - Friday, 17 October 2014

1. Read Backed Phasing: read backed phasing didn't really appear to work; maybe because didn't build on haplotypecaller? don't want to use haplotypecaller because that's not trustworthy, so no phasing, I guess.

2. Sequence Divergence: estimated D_{xy} and D_a for Encelia palmeri and ventorum. Results are plotted locally under analysis. 0.76%: average Da, 0.97%: average Dxy

3. Ancestral allele: made massive mpileup file that I will parse later to create the ancestral sequence for use in ANGSD. Run via: `samtools mpileup -f ~/encelia/annotation/palmer`

4. SNP calls: redid them, because it became clear that the mapping filter was too stringtent. So, I redid them with no BAQ and no adjustments for mapping quality. I am still concerned about these results, because definitely gets some SNPs that aren't legit, which can be easily seen as the sites that are uniformly heterozygous. Need to revisit this, though I did end up using these results for the pop genomic stuff. I think these SNP calls are okay depending on what I am trying to do.

5. SNP Filtering 1: 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 57458 are bad
   - note that 48829 are good
   - getting rid of SNPS
     a) MQSB - filter less than 0.2
     b) SGB - filter less than -500
     c) MQ0F - filter greater than 0.2
     d) MQ - filter less than 40
   - using these filters, will lose very few real SNPs likely
   - gives: 121399: fixed SNPs, 78691: shared SNPs, 111270: polymorphic SNPs in palmeri, 63069: polymorphic SNPS in ventorum

6. Admixture: best result was 2 populations, no evidence for admixture

7. SNP Filtering 2: To filter for coverage, did the following:

```
~/bin/bedtools2/bin/bedtools intersect -a ../variants/Encelia_palmeri.bwa_sameref.sa
~/bin/bedtools2/bin/bedtools intersect -a ../variants/Encelia_ventorum.bwa_sameref.s
```

Week of 20 October - 24 October 2014

1. ancestral allele sequence identification: I had aligned the four outgroups to E. palmeri, and then defined the ancestral sequence using an in house script. It definitely seems like the polarization isn't quite to be trusted, ad it tended to call the palmeri allele ancestral over the ventorum. (Of course, this could be possible given E. ventorum's ẗhyper-derivedstatus.)

2. ANGSD: started running it. My first runs were super weird, which could be for a two main reasons: (1) the `only_proper_pairs` flag or (2) I restricted calling to only high coverage areas, rather than using the UTR regions. The code I ran was the following:

```
# identify the sites that are variable in either lineage
# -P is the number of threads
/home/ssinghal/bin/angsd0.612/angsd -GL 1 -b /home/ssinghal/encelia/analysis/angsd/v
/home/ssinghal/bin/angsd0.612/angsd -GL 1 -b /home/ssinghal/encelia/analysis/angsd/p
# get the SFS for each pop
# first argument is saf file, second argument is the number of chromosomes, -P 24 is
/home/ssinghal/bin/angsd0.612/realSFS /home/ssinghal/encelia/analysis/angsd/ventorum
/home/ssinghal/bin/angsd0.612/realSFS /home/ssinghal/encelia/analysis/angsd/palmeri.
# want to identify sites that are variable in either of the populations
gunzip -c /home/ssinghal/encelia/analysis/angsd/ventorum.saf.pos.gz /home/ssinghal/e
# want to estimate genotypes at variable sites for both chromosomes
/home/ssinghal/bin/angsd0.612/angsd -GL 1 -b /home/ssinghal/encelia/analysis/angsd/v
/home/ssinghal/bin/angsd0.612/angsd -GL 1 -b /home/ssinghal/encelia/analysis/angsd/p
# estimate joint frequency
/home/ssinghal/bin/angsd0.612/realSFS 2dsfs /home/ssinghal/encelia/analysis/angsd/ve
```

3. to run unfolded ANGSD

```
# can also run with -fold 1 to get folded spectrum (add this to all angsd commands)
# if running with -fold 1, then the number supplied to realSFS becomes the number of
/home/ssinghal/bin/angsd0.612/angsd -GL 1 -b /home/ssinghal/encelia/analysis/angsd/v
/home/ssinghal/bin/angsd0.612/angsd -GL 1 -b /home/ssinghal/encelia/analysis/angsd/p
/home/ssinghal/bin/angsd0.612/realSFS /home/ssinghal/encelia/analysis/angsd/ventorum
/home/ssinghal/bin/angsd0.612/realSFS /home/ssinghal/encelia/analysis/angsd/palmeri_
gunzip -c /home/ssinghal/encelia/analysis/angsd/ventorum_folded.saf.pos.gz /home/ssi
/home/ssinghal/bin/angsd0.612/angsd -GL 1 -b /home/ssinghal/encelia/analysis/angsd/v
/home/ssinghal/bin/angsd0.612/angsd -GL 1 -b /home/ssinghal/encelia/analysis/angsd/p
```

```
/home/ssinghal/bin/angsd0.612/realSFS 2dsfs /home/ssinghal/encelia/analysis/angsd/ve
```

4. so, rerunning ANGSD with these new parameters, and will see if the SFS makes more sense. If not, maybe worth pursuing just defining it myself using the SAMTOOLS calls.