DUKE UNIVERSITY

Assignment 2 part 4

Tom Vamvanij

November 21, 2014
CBB 520

12 SINGLE BASE ERRORS

With the BAM file obtained from the previous section, `samtools mpileup` can be used to find variants between reads and the assembly (i.e. errors). A bash script was written to automate the work flow and filter only single base variants and high-quality reads as follows.

```
#!/usr/bin/env bash
reference_file="$1" # test 'abyss_80/abyss_80.fa'
bam_file="$2" # test 'abyss_80/abyss_80_little.bam'
prefix="${reference_file##*/}"
prefix="${prefix%.*}"
output_file="${prefix}_single_base_errors.txt"

printf "# Contig Position Quality Read Assembly\n\
#----------------------------------------\n" > $output_file
samtools faidx $reference_file # indexing req'd for mpileup
samtools view -u -F 12 -f 2 $bam_file |
samtools mpileup -uf $reference_file - | tee ${prefix}.bcf | #TODO remove tee?
bcftools view -vcg - |
awk '!/^#/' | while read line ;
do
  read chrom pos id ref alt qual _ <<< $line
  if (( ${#ref} == 1 && "${qual%.*}" >= 20 )) && [[ $ref != "N" ]]; then
    sequence=$(samtools faidx $reference_file $chrom |
      tail -n +2 |tr -d '\n') # remove ID & line breaks
```

```
      if (( $pos >= 11 )); then
        context_fore="${sequence:$pos-11:10}"
      else
        context_fore="${sequence:0:$pos-1}"
      fi
      if (( ${#sequence} >= $pos+10 )); then
        context_aft="${sequence:$pos:10}"
      else
        context_aft="${sequence:$pos}"
      fi
    assembly="${context_fore}__${ref}__${context_aft}"
    read_seq="${context_fore}__${alt}__${context_aft}"
    printf "$chrom $pos $qual $read_seq $assembly\n"
  fi
done |
column -t >> $output_file
head -n 22 $output_file
echo
echo More results in ${output_file}
read
```

13 PATTERNS

Results from the two assemblies (Velvet_61 and Abyss_80) reveal no obvious pattern. A fair amount of repetition can be observed before the errors, but no so much that pure chance may be ruled out as the cause. The first 20 errors of each assembly are shown below.

```
[bitnami@linux gene]$ head ./velvet_61_single_base_errors.txt -n 20
# Contig Position Quality Read Assembly
#-----------------------------------------
NODE_3_length_2835_cov_4930.320801     2       187   T__G__TGATGCCCTT            T__T__TGATGCCCTT
NODE_11_length_984_cov_5482.786621     100     225   TGGTATGATA__G__TTTGCAAGTA   TGGTATGATA__T__TTTGCAAGTA
NODE_11_length_984_cov_5482.786621     1039    222   ACTCAATAAG__T__ATCTT        ACTCAATAAG__C__ATCTT
NODE_12_length_873_cov_488.054993      73      155   GAAAATGAAA__T__CCTGTTCTTT   GAAAATGAAA__C__CCTGTTCTTT
NODE_12_length_873_cov_488.054993      222     225   CTTTTTATAG__G__TTGTCTTTTT   CTTTTTATAG__A__TTGTCTTTTT
NODE_12_length_873_cov_488.054993      375     99    TACATAATCA__G__TGACTTTCGT   TACATAATCA__C__TGACTTTCGT
NODE_12_length_873_cov_488.054993      564     63    GCCAAATCAA__C__CCAATGTGGT   GCCAAATCAA__G__CCAATGTGGT
NODE_18_length_6542_cov_1355.186646    105     222   TATATTAAAG__T__ATTGTAGAGA   TATATTAAAG__G__ATTGTAGAGA
NODE_20_length_3536_cov_1376.391968    1953    222   TTCATAATAG__T__ACTCTTGGTG   TTCATAATAG__A__ACTCTTGGTG
NODE_27_length_75_cov_217.919998       52      34    GTGTTGCCGC__T__ATCGCTGCCG   GTGTTGCCGC__C__ATCGCTGCCG
NODE_27_length_75_cov_217.919998       76      140   TTGCCGCTGC__T__CCAGCCACTA   TTGCCGCTGC__C__CCAGCCACTA
NODE_27_length_75_cov_217.919998       89      181   AGCCACTACC__A__CTCTATCTCC   AGCCACTACC__C__CTCTATCTCC
NODE_27_length_75_cov_217.919998       112     131   CTGACGAAAG__A__GTCAACTTGG   CTGACGAAAG__G__GTCAACTTGG
NODE_33_length_757_cov_403.033020      566     225   GCAAAATGGA__C__CATCCACATA   GCAAAATGGA__T__CATCCACATA
NODE_33_length_757_cov_403.033020      702     225   TCATTTGAAG__A__AAAGAATCGT   TCATTTGAAG__G__AAAGAATCGT
NODE_33_length_757_cov_403.033020      714     225   AAGAATCGTT__T__TCCAGATACT   AAGAATCGTT__C__TCCAGATACT
NODE_46_length_105_cov_265.847626      9       134   GAATGAAA__C__ACATATTTTA     GAATGAAA__T__ACATATTTTA
NODE_47_length_172_cov_342.523254      55      143   CCAATGTGGA__A__AACCTTTCGA   CCAATGTGGA__C__AACCTTTCGA
NODE_47_length_172_cov_342.523254      94      40    CCCTTATGTT__T__GGTGTCACTG   CCCTTATGTT__C__GGTGTCACTG
```

```
[bitnami@linux gene]$ head ./abyss_80/abyss_80_single_base_errors.txt -n 20
# Contig Position Quality Read Assembly
#-----------------------------------------
  5     80      165   AAAATGATGG__T__GTGGAAGACA         AAAATGATGG__G__GTGGAAGACA
  62    80      221   ACACCCAAAC__A__CTCGCATAGA         ACACCCAAAC__C__CTCGCATAGA
  72    80      222   GACCGTAAGG__T__CGGGTCGAAC         GACCGTAAGG__G__CGGGTCGAAC
  79    56      26    TACCCTAATC__T__AACCCAGATC         TACCCTAATC__C__AACCCAGATC
  92    80      74    TAGTATTAGG__T__AGTCAGATGA         TAGTATTAGG__G__AGTCAGATGA
  111   80      89    CTGGGGGGAG__T__ATGGTCGCAA         CTGGGGGGAG__G__ATGGTCGCAA
  112   80      78    CTGGGGGGAG__T__ATGGTCGCAA         CTGGGGGGAG__A__ATGGTCGCAA
  116   80      222   ATCTTTCGGG__T__CCCAACAGCT         ATCTTTCGGG__G__CCCAACAGCT
  119   80      222   TAAGGTCGGG__T__CGAACGGCCT         TAAGGTCGGG__G__CGAACGGCCT
  123   80      222   GGTGACGGAG__T__GCGCTGGTCA         GGTGACGGAG__G__GCGCTGGTCA
  124   80      222   TTATGTCTTG__T__GTGATAATTT         TTATGTCTTG__G__GTGATAATTT
  127   80      159   TGAGGACAGC__G__ACACGTGCAT         TGAGGACAGC__A__ACACGTGCAT
  128   80      25    GGTTCCGGGG__T__CGAAATGACC         GGTTCCGGGG__G__CGAAATGACC
  142   78      49    TGTGGTGGGG__T__GGTATAGTCC         TGTGGTGGGG__G__GGTATAGTCC
  154   80      21    ATTCCCAGAG__T__GTGTTTCTTT         ATTCCCAGAG__A__GTGTTTCTTT
  164   80      222   CGAATTTGGG__T__ATAGGGGCGA         CGAATTTGGG__G__ATAGGGGCGA
  179   80      68    ACTGCCAGCA__T__CAACGTCAGG         ACTGCCAGCA__C__CAACGTCAGG
  189   80      54    AGTGTTTGGG__T__GTAAAACCCA         AGTGTTTGGG__G__GTAAAACCCA
```