CSCI2950-C: Spring 2013. Computer Assignment

http://cs.brown.edu/courses/csci2950-c/Spring2013/index.html

Due: Monday February 25, 11:59pm

Aligned reads from a sequencing experiment are stored in Sequence Alignment Map (SAM) format, and the corresponding binary (BAM) format. Details of this format are here: http://samtools.sourceforge.net/SAM1.pdf

In this assignment you and a partner will examine (paired) reads from a simulated mixture of normal cells and tumor cells (derived from cell line data). This data comes from an ongoing benchmarking exercise for The Cancer Genome Atlas (TCGA), the largest public cancer genome sequencing effort. See details below.

Each pair of students will be assigned a different type of genomic variant – single nucleotide variant (SNV), copy number aberration, or structural variant -- and corresponding software package that detects that type of variant. Students will be responsible for downloading and running their assigned program on two different mixtures (BAM files). Each pair of students will hand in a brief document that addresses the following points:

1. Give a brief overview (1 paragraph) that summarizes the algorithm/code you were assigned.
2. What parameters did you use when running? If you used default parameters – list what values that includes.
3. How many variants did you detect for each mixture?
4. What differences did you see in your predictions between the two different mixtures? (To answer this question you may need to implement your own code to compare the output from the different mixtures.)
5. What, if any, problems did you have while running your algorithm or interpreting the output? How did you address/resolve these problems?

If we obtain good results (either individually or as a group), we may consider refining them and submitting them to the benchmark exercise.

Variant types/Algorithms

1. Single Nucleotide Variants (SNVs) / VarScan 2
   Paper: http://www.ncbi.nlm.nih.gov/pubmed/22300766
   Code/Website: http://varscan.sourceforge.net/ (use the somatic option)
   Input Details: This code does not run directly on a BAM file, instead it requires a "pileup" file created using a program called samtools. Creation of the "pileup" file requires the use of a fasta file for the human genome. This is the fasta file you should use:
   /data/compbio/csci2950-c/data/UCSCGenomeBrowser/hg19/hg19.fa
   Other: VarScan2 (run using the somatic option) allows for optional parameters that list the purity of the tumor and normal sample. Try running with these parameters appropriately set.

2. Copy Number Variants (CNVs) / BIC-Seq
   Paper: http://www.pnas.org/content/108/46/E1128/1.full
   Code/Website: http://compbio.med.harvard.edu/Supplements/PNAS11.html
   Input Details: This code does not run directly on a BAM file, instead it requires some amount of preprocessing. The download package includes a modified version of samtools that will run directly on the BAM file to output the necessary input files.
   Other: The package includes both a perl and an R module for running the code. You can choose whichever you are more familiar with to run the code. Also, the code requires a specific input flag when run with paired read data.

3. Structural Variants (SVs) / GASV
   Paper: http://bioinformatics.oxfordjournals.org/content/25/12/i222
   Code/Website: http://code.google.com/p/gasv/
   Input Details: The GASV code does not run directly on a BAM file, instead it requires preprocessing using a program called BAMToGASV (also included in the download).
   Other: The download also includes a program called GASV-pro. However, you only need to run GASV.

Data

In this assignment, you will analyze two different tumor mixtures from the TCGA Mutation Calling Benchmark 4 dataset: http://hgwdev.soe.ucsc.edu/~ewingad/benchmark4/instructions.pdf

Tumor1 (95% Cancer, 5% Normal)
/data/compbio/csci2950-c/data/TCGA_Benchmark4/tumorNormalMix/HCC1143_5N_95T/HCC1143.mix1.n5t95.bam

Tumor2 (60% Cancer, 40% Normal)
/data/compbio/csci2950-c/data/TCGA_Benchmark4/tumorNormalMix/HCC1143_40N_60T/HCC1143.mix1.n40t60.bam

Matched Normal (100% Normal) – Some algorithms require data from a matched normal sample in addition to the tumor sample.
/data/compbio/csci2950-c/data/TCGA_Benchmark4/tumorNormalMix/HCC1143_Normal/HCC1143.NORMAL.30x.compare.bam

Logistics

- Please place any code you download or write into the following directory:
  /data/compbio/csci2950-c/code
- BAM files can be rather large. Some of the programs make take a long time to run or require large amounts of memory. Utilizing the CS grid can be extremely helpful in these situations. http://cs.brown.edu/system/hardware/cluster/gridengine.html

Additional Resources

- **Samtools –** Since BAM files are stored in a binary format, they are not directly human-readable. Samtools is a program that includes many features for working with BAM files (it is required by at least one of the above variant calling programs). http://samtools.sourceforge.net/

- **Picard –** This is another set of tools for working with BAM files. It also provides an API for incorporating BAM file operations into Java code. http://picard.sourceforge.net/
- It is important to remember that these datasets are part of a mutation calling challenge from the TCGA. That means we don't know the true set of mutations – you job is to get a flavor for what it is like to work with cancer sequencing data (and maybe help us figure out correct set of mutations). If you run into problems or have questions don't hesitate to ask for help.