# Bi188 2013

# Computational exercise 1

**Due: 3pm on April 26th, 2013**

## Warning #1!!!

Do not under any circumstances think that you will be able to complete this assignment in less than 48 hours. If you start the day before the deadline, you will almost certainly not be able to complete the computation on time, let alone the analysis. The data can be run through the pipeline in less than 12 hours if the aligner is given 4-8 CPUs, however, if everyone starts running jobs in the same time on the same machine, the result will be that the machine will either crash or slow down to a crawl and nobody will be able to complete their assignment. And that is only one part of the exercise - the more time-consuming for you part will be writing the code to analyze the results and then interpreting the output. For these reasons, we strongly urge you to start processing the data as soon as possible, and not wait until the last moment.

## Warning #2!!!

Computer memory is a limited resource that can be easily exhausted if one does some variation of storing the whole of dbSNP in a giant dictionary or creating a dictionary that has individual genomic positions as keys and contains A LARGE FRACTION of the genome. Doing this may cause the system to at worst crash and at best slow down to a crawl. There are situations in which such brute force algorithmic approaches are inevitable, but none of what you will be doing here will require you to adopt them, so pay attention to the efficiency of your code.

## Disclaimer

The data you will be working with here is a toy example that has been stripped out of a lot of the complexity that is present in real-life data. As a result the pipeline you will be running has been considerably simplified compared to what the real pipelines professional bioinformaticians run to analyze actual data. The reason we have made the decision to simplify things is that it is more important for you to understand how data is analyzed and interpreted on a more of a conceptual level rather than having to go deep down into the weeds of dealing with the imperfections of high-throughput sequencing data (which are likely to change in the future as new sequencing technologies are developed and adopted, leaving a lot of those details irrelevant). However, if by any chance you happen to know a lot about analyzing genome resequencing data, what you will be doing here may potentially confuse you, which is why we are starting with this disclaimer.

## 1 The case

You are presented with a case of a 17-year old patient presenting the following symptoms: movement coordination difficulties, short stature and low weight, delayed puberty, slurred speech, and prematurely aged skin; in addition, the patient has developed leukemia in the past year. Neither of the parents shows symptoms of the disease. You are aware that a variety of mendelian disorders with similar symptoms have been described, however, you are unable to distinguish between the possibilities based on the symptoms alone and in addition to this, the precise genetic basis for some of them is not known. Empowered by the tools of modern genomics, you decide to diagnose the case using whole exome sequencing. You isolate genomic DNA from the patient, shear it to an average size of 150-200bp, and take it through an exome capture (using 50bp oligonucleotide arrays) and library-building protocol. Finally, you sequence the library using a high-throughput sequencing platform, generating single-end reads of 75bp length.

## 2 Data and tools

The sequencing data in FASTQ format can be found here:

```
/woldlab/bostau/data00/pub/georgi/Bi188/Exercise1/Bi188-Exercise1.reads.fastq.bz2
```

In addition, you might or will need the following bioinformatic tools and files:

1. Short-read aligner program:

   ```
   /woldlab/bostau/data00/software/bowtie2-2.1.0/bowtie2-align
   ```

2. The human genome in the form of a FASTA file and a pre-compiled Burrows-Wheeler index (note that we will be using the male version for alignment, there are also the so called “random” chromosomes, which represent pieces of sequence that have been assembled but not accurately place within chromosomes; those we will ignore for now)

   ```
   /woldlab/bostau/data00/software/fasta/hg19-male.fa
   /woldlab/bostau/data00/software/fasta/hg19-female.fa

   /woldlab/bostau/data00/software/bowtie2-indexes/hg19-male
   /woldlab/bostau/data00/software/bowtie2-indexes/hg19-female
   ```

3. A copy of `samtools`:

   ```
   /usr/bin/samtools
   ```

4. A copy of dbSNP (note that we have given you a version of dbSNP that only contains SNVs with no known medical implications) in VCF format

   ```
   /woldlab/bostau/data00/pub/georgi/Bi188/dbSNP137_common_no_known_medical_impact_20130226.vcf.bz2
   ```

5. A copy of the refSeq annotation for the human genome in the form of a GTF file.

   ```
   /woldlab/bostau/data00/software/genes/hg19-refSeq.withNames.gtf
   ```

   Some snippets of code you might find useful:

   ```
   /woldlab/bostau/data00/pub/georgi/Bi188/Exercise1/Exercise1_functions.py
   ```

And a copy of the UCSC Genome Browser utilities you may find useful too:

```
/woldlab/bostau/data00/software/ucsc
```

Please refer to the bioinformatics manual on the class website for description of file formats, suggested alignment parameters and other tips you might find helpful. Note that if you use different tools/pipelines than the ones we have suggested you might get somewhat different results, therefore it is recommended that you stick to what is in the manual.

# 3 Questions

## 3.1 Coverage estimation

How good of a coverage of the exome did you get? Explain your algorithm/calculation.

## 3.2 Variant detection

How many SNVs do you detect in the sample? How many of them are novel? How would you classify them according to their expected effect on protein sequence and how many in each class do you see? Ignore indels for the purpose of this exercise.

## 3.3 Identification of causal variants

Now you want to find the most likely causal variant(s) for the condition. Explain how you would identify them and list the variant(s) and gene(s) you think are most likely to be mutated in this patient. Can you figure out which disease the patient has based on what you can find in the literature about you top candidate gene? Again, ignore indels for the purpose of this exercise. Use the following (tab-delimited) format for submitting your putative causal variants (use whatever categories you find appropriate to classify variants):

```
#CHROM  POS        REF  ALT  GT   GeneName  GeneID     TranscriptName  TranscriptID  Effect
chr1    247007112  T    C,A  1/1  AHCTF1    NM_015446  AHCTF1          NM_015446     missense
chr1    110603213  G    A    1/1  ALX3      NM_006492  ALX3            NM_006492     UTR3
....
```

Submit both answers to the questions (preferably by e-mail in pdf format) and a link to the code you wrote and your output files on the cluster.