COMP3353 Bioinformatics

Lab Session 2: Alignment

Date: Tuesday, Oct. 09, 2018 10:30-12:20
Location: HW311

Please try your best to answer all questions within the lab session. Please paste the codes and results to the checkpoints in Moodle. For questions please ask the TA or the postgraduate helpers. You can discuss with other students for ideas but please don't copy others' code or results. If you cannot finish the questions in class, please finish them after class and submit the code and results no later than Friday 11:59 pm.

Preparation 1: Install software

1.1 Install miniconda

Conda is an app store for both Linux and Mac. While the system's default app installers including both the "apt-get" in Ubuntu and "Yum" in RedHat require administrator (root) privilege, Conda installs apps in your home folder and solves the libraries dependencies automatically. One thing to note is, in Conda, not all apps that are available in Linux would also be available in Mac, so if Conda can't help you with installing a tool in MacOS, you are always welcomed to use the CS academy server or install a virtual Linux with VirtualBox for Mac.

To install Conda, please follow the guide in this page. If you use the CS academy server, you can use the below commands:

- `curl https://repo.continuum.io/miniconda/Miniconda2-latest-Linux-x86_64.sh > Miniconda2-latest-Linux-x86_64.sh`
- `sh Miniconda2-latest-Linux-x86_64.sh`

Use the down arrow to go through the license, type **"`yes`"** to agree with the license. When being prompt for the installation path, press enter to install Conda to your home folder by default. When being asked "Do you wish the installer to prepend the Miniconda2 install location to PATH in your …", type **"`yes`"** to agree. After the installation, run command **"`source ~/.bashrc`"** to load the updated environment variable **"`$PATH`"**. Now you should be able to use `conda` by using the command **"`conda`"**.

1.2 Add the "bioconda" channel to Conda

"bioconda" is one of the many available Conda channels. It hosts most of the commonly used bioinformatics tools. Please use the following commands to add the "bioconda" channel to Conda:

- `conda config --add channels defaults`
- `conda config --add channels conda-forge`
- `conda config --add channels bioconda`

For more details of bioconda, please visit this page.

1.3 Install BWA

- `conda install -c bioconda bwa`

1.4 Install samtools

- `conda install -c bioconda samtools`

Go through this samtools tutorial and go back later when needed

1.5 Install IGV

Complete the IGV tutorial written by Obi Griffith.

For both the Linux and Mac user, please install `IGV` by following the instructions here. Note that IGV requires Java 8, thus for the CS academy server users, please finish the following steps to install Java 8 to your home folder:

- `cd ~`
- `curl http://www.bio8.cs.hku.hk/comp3353/jre-8u181-linux-x64.tar.gz | tar -zxf -`

Now you should be able to see a folder named `jre1.8.0_181` in your home folder. Then, add the folder to the $PATH environment variable and make it permanent by writing the command into "~/.bashrc" (a program that will be run automatically every time you open a new terminal). Please read all the steps below carefully before you can proceed:

- `cat >> ~/.bashrc # Press enter the command will run but no prompt will appear because the command is waiting for your input to be append to the file, notice that I used two redirection symbols, it means append to the end of the file, while one symbols means overwrite the file entirely, please don't make a mistake!`

Now copy the following one line from and paste to the terminal

```
export PATH="~/jre1.8.0_181/bin:$PATH"
```

Press Enter, then press Ctrl+D. Now your "~/.bashrc" file should be appended with the new line you've just pasted. You can give it a check by using the tail command.

To activate you changes, close the terminal and open again, or, use the following command:

- `source ~/.bashrc`

Finally, check if the current Java version is what we want by command:

- `java -version`

You should see Java version "1.8.0_181".

Download IGV

- `cd ~`
- `curl http://data.broadinstitute.org/igv/projects/downloads/2.4/IGV_2.4.14.zip >IGV_2.4.14.zip`
- `unzip IGV_2.4.14.zip`
- `cd IGV_2.4.14`

Run IGV

```
java -Xmx1500m -jar lib/igv.jar
```

Section 1: Alignment using BWA

Download Data

1. Download the FASTA file for the E. coli reference genome:

```
curl
ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/005/845/GCF_000005845.2_ASM584v2/GCF_000005845.2_ASM584v2_genomic.fna.gz > ecoli.fa.gz
gunzip ecoli.fa.gz
```

2. Download the FASTQ files of SRR2584857:

```
curl
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR258/007/SRR2584857/SRR2584857_1.fastq.gz > SRR2584857_1.fastq.gz
```

```
curl
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR258/007/SRR2584857/SRR2584857_2.fastq.g
z > SRR2584857_2.fastq.gz
```

Run Alignment

1. Build index

```
bwa index ecoli.fa
```

2. Alignment (Upload the running logs)

```
bwa mem -t 4 ecoli.fa SRR2584857_1.fastq.gz SRR2584857_2.fastq.gz | samtools
sort -o SRR2584857.bam -
```

3. Have a look at the result.

```
samtools view SRR2584857.bam | head -n 2
```

4. Delete FASTQ files to release space

The student account has limited disk quota on the academy server, so let's delete the FASTQ files. We can download them again if we need them in the future.

```
rm SRR2584857_1.fastq.gz SRR2584857_2.fastq.gz
```

Section 2: Know more about the alignment result

Download data

Download the sorted BAM with human data and its corresponding index file:

```
curl www.bio8.cs.hku.hk/comp3353/sample.sorted.bam > sample.sorted.bam
curl www.bio8.cs.hku.hk/comp3353/sample.sorted.bam.bai >
sample.sorted.bam.bai
```

Question #1

Is the BAM file `'sample.sorted.bam'` sorted or unsorted? How do you know?

Question #2

How many alignments are in `'sample.sorted.bam'`? Show your work.

Question #3

How many alignments in `'sample.sorted.bam'` overlap chromosome `1` from position `12398392` to `12399392`? Note that this can be done with a single `'samtools view'` command making use of the BAM index (`sample.sorted.bam.bai`). Show your work.

Question #4

The `'sample.sorted.bam'` file includes paired-end alignments. For a given pair, it is often the case that the `TLEN` is positive for one end of the pair and negative for the other end. Why is this?

Question #5

Using only the positive `TLEN` for each paired-end alignment, what is the min, max, and the most frequent `TLEN` observed in the BAM file? Show your work.

Question #6

What fraction of the paired-end alignments are `'properly paired'`?
Hint: use the `'samtools flagstat'` command. Does this seem like a reasonable fraction for a high-quality DNA fragment library?

Question #7

How many alignments are on the positive strand and on the negative strand?
Hint: you need to test the `FLAG` field. Show your work.

Question #8

Use IGV to visualize the `'sample.sorted.bam'` and figure out how to display pair-end reads in IGV. Go to locus `chr17:41,244,367-41,244,497` and make a screenshot and upload it to Moodle.

Question #9

Figure out how to go to the region of gene `'ATM'` when using IGV to view the `'sample.sorted.bam'`. Make a screenshot and upload it to Moodle.

Optional IGV questions:

These are optional questions that you don't need to submit your results to Moodle. But I strongly suggest you solve the questions if you have additional times. IGV is like a swiss knife in bioinformatics visualization. It is very important for biology and medical student to master the skill of using IGV. Please feel free to book a slot in the consultation hours for any problems.

Question #a1

Using IGV, how many sequences aligned to exon `17` of the `ATM` gene?
NOTE: make sure you set the reference genome to `'Human hg19'`

Question #a2

At what position is the apparent SNP in exon 17 of the ATM gene? What allele is observed for this individual?

Question #a3

Following question #a2, what do you think the individual's genotype is at this site based on the alignments? You may need to refer back to earlier lectures if you have forgotten about genotypes.

Question #a4

What do you think the individual's genotype is at the SNP in following locus: `chr17:41,244,367-41,244,497`? What are the `Phred quality scores` associated with the observed alternate allele? Does this make you more or less confident in your genotype assessment?

Question #a5

What is your assessment of the two apparent SNPs at this locus: `chr7:55,227,815-55,228,328`? Are they any less or more confident than the SNP in question #a4? What else can you say about the inheritance of these two SNPs?

Question #a6

Is there a SNP at position `chr7:55,224,334`? Why or why not?

End of the Lab Session 2