NGS mapping tutorial

Introduction into NGS Mapping

Within this tutorial we will

- learn to understand what mapping is
- how different tools and parameters can influence the results
- which tools suits which experiments

bowtie 2

http://bowtie-bio.sourceforge.net/bowtie2/index.shtml

Bowtie 2 is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences. It is particularly good at aligning reads of about 50 up to 100s or 1,000s of characters, and particularly good at aligning to relatively long (e.g. mammalian) genomes. Bowtie 2 indexes the genome with an FM Index to keep its memory footprint small: for the human genome, its memory footprint is typically around 3.2 GB. Bowtie 2 supports gapped, local, and paired-end alignment modes.

In this tutorial we will map data of two experiments against a reference sequence.

The first data set is a paired end DNA sequence of a cassava genotype called TMEB117 and the second data set is a paired end RNA sequencing from a cassava genotype called TMEB419.

The reference sequence is chromosome01 from the latest cassava genome built V8 (https://phytozome-next.jgi.doe.gov/info/Mesculenta_v8_1 (Links to an external site.)).

The files we need for this tutorial have been provided to you in the directory for Day2 of the course under the Epicass_workshop folder. You would need to navigate into the subfolder called rawData

cd rawData

As you remember from the lecture, each NGS mapper uses indices to accelerate the mapping process, but unfortunately each software has it's own way to create these indices. Therefore we need to create first the index for bowtie2 of our reference sequence.

Create a directory for your references and indices within your working directory for the day2.

mkdir reference

As mentioned we will use only chromosome01 for the exercise and it is available "chromosome01.fasta"

Now we can create the index which we will call Bowtie2Cassava01Index and will be located in the reference directory:

bowtie2-build --threads 4 -o 3 rawData/chromosome01.fasta reference/Bowtie2Cassava01Index

You will get the following index

```
-rw-r--r-- 1 itbaag10 users    10713438 Sep 18 16:50 Bowtie2Cassava01Index.4.bt2
-rw-r--r-- 1 itbaag10 users         341 Sep 18 16:50 Bowtie2Cassava01Index.3.bt2
-rw-r--r-- 1 itbaag10 users    21426880 Sep 18 16:50 Bowtie2Cassava01Index.2.bt2
-rw-r--r-- 1 itbaag10 users    18479518 Sep 18 16:50 Bowtie2Cassava01Index.1.bt2
-rw-r--r-- 1 itbaag10 users    21426880 Sep 18 16:50 Bowtie2Cassava01Index.rev.2.bt2
-rw-r--r-- 1 itbaag10 users    18479518 Sep 18 16:50 Bowtie2Cassava01Index.rev.1.bt2
```

files:

Having the index we create a directory for the mapping results:

mkdir mapping

and then we can run the mapping:

bowtie2 -p 4 --very-sensitive-local -x reference/Bowtie2Cassava01Index -S mapping/TMEB117.sam -1 rawData/TMEB117_R1_frac.fastq -2 rawData/TMEB117_R2_frac.fastq 2>&1 | tee mapping/TMEB117.log

We will get a mapping file in the SAM format and the log file with the mapping statistics.

- mapping/TMEB117.sam
- mapping/TMEB117.log

Check the log and find how many unique hits of pairs we have.

Check the SAM file and check how many hits we have.

Time allowing, I propose to run a third test mapping either using the raw data of trimmed data by changing the mapping mode from local to end-to-end or from very-sensitive to fast, loosing some accuracy but reducing the mapping time.

Replace in the commands above *--very-sensitive-local* to *--very-sensitive* or from *--very-sensitive-local* to *--fast-local*.

Again, compare the output with the previous ones.

The data set TMEB117 is a DNA sequencing and bowtie2 the right tool for mapping and data analysis.

TMEB419 however is a RNA sequencing and we could run bowtie2 and check how the results look like. For RNA sequencing data exist dedicated aligners which find

splicing sites within the reads and guarantee an accurate reconstruction of the sequenced transcripts.

STAR

https://github.com/alexdobin/STAR (Links to an external site.)

STAR is a splicing site sensitive mapper for RNA sequencing data. We will not be covering RNAseq analysis, however, for understanding mapping strategies we will have a look at STAR.

The results we will direct into the existing mapping directory.

The data we will use is paired end RNA sedquencing and is already in the data directory (TMEB419RNA_frag_R1.fastq, TMEB419RNA_frag_R2.fastq).

Firstly we need to create the index for STAR and as mentioned we will be able to include the gene annotations so that the mapped reads can be associated to know genes/transcripts.

For that we need the annotation file "Mesculenta_671_v8.1.gene_exons.gtf" from copied to your reference directory.

The indexing command for STAR will be the following:

mkdir reference/STARCassava01Index (the directory for the index must be created first)

STAR --runThreadN 4 --runMode genomeGenerate --genomeDir reference/STARCassava01Index --genomeFastaFiles rawData/chromosome01.fasta --sjdbGTFfile rawData/Mesculenta_671_v8.1.gene_exons.gtf --genomeSAindexNbases 13 --genomeChrBinNbits 16

and you will end up with a bunch of index file:

```
itbaag10@magilla:/biodata/ANDREAS/Course/Data$ ll STARCassava01Index/
total 439168
-rw-r--r-- 1 itbaag10 users       597 Sep 18 12:45 genomeParameters.txt
-rw-r--r-- 1 itbaag10 users        11 Sep 18 12:45 chrStart.txt
-rw-r--r-- 1 itbaag10 users        13 Sep 18 12:45 chrName.txt
-rw-r--r-- 1 itbaag10 users        22 Sep 18 12:45 chrNameLength.txt
-rw-r--r-- 1 itbaag10 users         9 Sep 18 12:45 chrLength.txt
-rw-r--r-- 1 itbaag10 users     56663 Sep 18 12:46 geneInfo.tab
-rw-r--r-- 1 itbaag10 users   1049777 Sep 18 12:46 exonGeTrInfo.tab
-rw-r--r-- 1 itbaag10 users    308872 Sep 18 12:46 transcriptInfo.tab
-rw-r--r-- 1 itbaag10 users    493312 Sep 18 12:46 exonInfo.tab
-rw-r--r-- 1 itbaag10 users    445523 Sep 18 12:46 sjdbList.fromGTF.out.tab
-rw-r--r-- 1 itbaag10 users    445523 Sep 18 12:46 sjdbList.out.tab
-rw-r--r-- 1 itbaag10 users    350053 Sep 18 12:46 sjdbInfo.txt
-rw-r--r-- 1 itbaag10 users  45996003 Sep 18 12:46 Genome
-rw-r--r-- 1 itbaag10 users 376054383 Sep 18 12:46 SA
-rw-r--r-- 1 itbaag10 users  24466875 Sep 18 12:46 SAindex
```

Files like geneInfo.tab, exonGeTrInfo.tab, transcriptInfo.tab, exonInfo.tab, sjdbList.fromGTF.out.tab, sjdbList.out.tab, sjdbInfo.txt shows you tat the indexing contains information of the predicted genes, transcripts, exons and splicing sites annotated in the Mesculenta_671_v8.1.gene_exons.gtf file.

Now we are ready to map the DNA sequencing files to the reference sequence:

STAR --runThreadN 4 --genomeDir reference/STARCassava01Index --outFilterMismatchNoverLmax 0.06 --outFilterMatchNminOverLread 0.2 --outFilterScoreMinOverLread 0.2 --alignIntronMax 20000 --alignMatesGapMax 10000 --outFileNamePrefix mapping/TMEB117_ --readFilesIn rawData/TMEB117_R1_frac.fastq rawData/TMEB117_R2_frac.fastq

Now we map the RNA sequencing files to the reference sequence:

STAR --runThreadN 4 --genomeDir reference/STARCassava01Index --outFilterMismatchNoverLmax 0.06 --outFilterMatchNminOverLread 0.2 --outFilterScoreMinOverLread 0.2 --alignIntronMax 20000 --alignMatesGapMax 10000 --outFileNamePrefix mapping/TMEB419RNA_ --readFilesIn rawData/TMEB419RNA_frag_R1.fastq rawData/TMEB419RNA_frag_R2.fastq

Your output will be the SAM file, report files and the list of predicted splicing sites as show below:

```
-rw-r--r-- 1 itbaag10 users   310242975 Sep 18 17:22 TMEB419RNA_Aligned.out.sam
-rw-r--r-- 1 itbaag10 users      105902 Sep 18 17:22 TMEB419RNA_SJ.out.tab
-rw-r--r-- 1 itbaag10 users         364 Sep 18 17:22 TMEB419RNA_Log.progress.out
-rw-r--r-- 1 itbaag10 users       15184 Sep 18 17:22 TMEB419RNA_Log.out
-rw-r--r-- 1 itbaag10 users        1669 Sep 18 17:22 TMEB419RNA_Log.final.out
```

Inspect the report file and check the hit rate and the average hit size.

Check the CIGAR and evaluate the soft clipping.

As an additional exercise, you could run the same command with the trimmed reads (from the earlier tutorial) and observe the change of soft clipped fragments.

You can convert the SAM file to a BAM file (important for large files as will be easier to transfer due to size reduction / compression) using samtools

To visualize the data you would then download the BAM file to your local machine and visualize on IGV (Integrated Genome Viewer)