BIOC | BioInformatics Competence Center

Bioinformatics Analysis of RNA-sequencing

BIO-693

Unil EPFL

Allison Burns, Maxime Jan, Linda Mhalla,
Christian Iseli & Nicolas Guex
– summer 2023 –

This morning

- Connect to the cluster
- Collect and organize all the raw fastq data
- Prepare reference data
- Map fastq to reference
- Check mapping statistics
- Obtain counts

Getting started on helvetios

- Connect to helvetios and request an interactive session
  - `ssh <username>@helvetios.epfl.ch`
  - `module add tmux`
  - `tmux`, or `tmux attach`
  - `cd /scratch/<username>`
  - `Sinteract -a bio-693 -c 20 -t 08:00:00 -m 40g`
- On the cluster node
  - `cp /scratch/iseli/course/SraRunTable.txt .`
    - Do not forget/omit the dot (`.`)
  - `singularity shell -s /bin/bash -B /scratch/$USER -B /scratch/iseli /scratch/iseli/MicMap-2.4-f38.sif`

Example SRA – retrieve

- Retrieve all the samples

```
for n in `cut -d , -f 1 SraRunTable.txt | grep ^SRR`; do
  echo "Processing $n";
  fasterq-dump --split-3 --progress $n;
  pigz --fast ${n}*.fastq;
  sra-stat --quick $n >${n}_stat.txt;
done
```

- Check retrieved data

```
for f in SRR*_stat.txt; do
  n=`basename $f _stat.txt`;
  cntFQ=`perl -ne '($c)=$_=~/^[^:]+\|(\d+):/;$t+=$c;END{print "$t\n"}' $f`;
  linesFQ=`gzip -dc ${n}_1.fastq.gz | wc -l`;
  if [[ $cntFQ -ne $((linesFQ / 4)) ]]; then
    echo "We have a problem for $n - $cntFQ vs $((linesFQ / 4))";
  else
    echo "count is correct for $n - $cntFQ";
  fi;
done
```

Retrieving SRA data

```
helvetios_iseli: ls -alth
total 76G
drwxr-xr-x 2 iseli BICC-unit 4.0K Jun 17 07:42 fasterq.tmp.helvetios.4169460
drwxr-xr-x 3 iseli BICC-unit 4.0K Jun 17 07:36 .
-rw-r--r-- 1 iseli BICC-unit  39K Jun 17 07:36 SRR16219945_stat.txt
-rw-r--r-- 1 iseli BICC-unit 2.2G Jun 17 07:31 SRR16219945_1.fastq.gz
-rw-r--r-- 1 iseli BICC-unit 2.4G Jun 17 07:31 SRR16219945_2.fastq.gz
-rw-r--r-- 1 iseli BICC-unit  39K Jun 17 05:45 SRR16219944_stat.txt
-rw-r--r-- 1 iseli BICC-unit 2.3G Jun 17 05:39 SRR16219944_1.fastq.gz
-rw-r--r-- 1 iseli BICC-unit 2.5G Jun 17 05:39 SRR16219944_2.fastq.gz
-rw-r--r-- 1 iseli BICC-unit    0 Jun 17 03:44 SRR16219943_stat.txt
-rw-r--r-- 1 iseli BICC-unit 2.8G Jun 17 03:38 SRR16219943_2.fastq.gz
-rw-r--r-- 1 iseli BICC-unit 2.7G Jun 17 03:38 SRR16219943_1.fastq.gz
-rw-r--r-- 1 iseli BICC-unit  40K Jun 17 01:34 SRR16219940_stat.txt
-rw-r--r-- 1 iseli BICC-unit 2.5G Jun 17 01:28 SRR16219940_1.fastq.gz
-rw-r--r-- 1 iseli BICC-unit 2.8G Jun 17 01:28 SRR16219940_2.fastq.gz
-rw-r--r-- 1 iseli BICC-unit  40K Jun 16 23:25 SRR16219938_stat.txt
-rw-r--r-- 1 iseli BICC-unit 2.8G Jun 16 23:19 SRR16219938_2.fastq.gz
-rw-r--r-- 1 iseli BICC-unit 2.6G Jun 16 23:19 SRR16219938_1.fastq.gz
```

Checking SRA data

```
Singularity> for f in SRR*_stat.txt; do n=`basename $f _stat.txt`; cntFQ=`perl -ne '($c)=$_=~/^[^:]+\|(\d+):/;$t+=$c;END{print
 "$t\n"}' $f`; linesFQ=`gzip -dc ${n}_1.fastq.gz | wc -l`; if [[ $cntFQ -ne $((linesFQ / 4)) ]]; then echo "We have a problem
for $n - $cntFQ vs $((linesFQ / 4))"; else echo "count is correct for $n - $cntFQ"; fi; done

count is correct for SRR16219921 - 41125131
count is correct for SRR16219922 - 31498370
count is correct for SRR16219923 - 37099819
count is correct for SRR16219924 - 35957634
count is correct for SRR16219925 - 38804161
count is correct for SRR16219926 - 33688333
count is correct for SRR16219928 - 39271560
count is correct for SRR16219929 - 42674023
count is correct for SRR16219930 - 30197829
count is correct for SRR16219931 - 34847378
count is correct for SRR16219932 - 34691862
count is correct for SRR16219933 - 33625186
count is correct for SRR16219935 - 40275240
count is correct for SRR16219936 - 34729325
count is correct for SRR16219937 - 37121145
count is correct for SRR16219938 - 38216703
count is correct for SRR16219940 - 38089443
count is correct for SRR16219941 - 37228589
count is correct for SRR16219942 - 36950994
count is correct for SRR16219943 - 40066968
count is correct for SRR16219944 - 34343035
count is correct for SRR16219945 - 32508162
count is correct for SRR16219946 - 34186012
count is correct for SRR16219947 - 36344402
count is correct for SRR16219948 - 35999453
count is correct for SRR16219949 - 33542412
count is correct for SRR16219950 - 39697257
count is correct for SRR16219951 - 38813449
count is correct for SRR16219952 - 33811791
count is correct for SRR16219954 - 33188661
count is correct for SRR16219955 - 33387404
count is correct for SRR16219956 - 37428628
Singularity>
```

Organize data

- Create a working directory in `/scratch/$USER`
- The container is available here :
  - `/scratch/iseli/MicMap-2.4-f38.sif`
- Data is available here :
  - `/scratch/iseli/course`
- Try to use symlinks when possible

Getting alignment (BAM) files

- STAR aligner – https://github.com/alexdobin/STAR
- Source of reference data
  - Gencode – https://www.gencodegenes.org/
  - EnsEMBL – https://www.ensembl.org/index.html
- Ideally use genome (FASTA) + gene annotation (gtf, gff, gff3)
  - Use primary assembly for the genome
  - Use comprehensive gene annotation (gencode) for the genes
  - Use `wget` to retrieve the data on the cluster

Preparing STAR index

- Use `genomeGenerate` command
  - `STAR --runMode genomeGenerate \`
    `--genomeDir STAR_mm39vM32/ \`
    `--genomeFastaFiles \`
    `GRCm39.primary_assembly.genome.fa \`
    `--sjdbGTFfile gencode.vM32.annotation.gtf`
- Beware the `genomeSAindexNbases` parameter for small genomes

Preparing STAR index files

```
Sinteract -a bio-693 -c 4 -t 08:00:00 -m 30g
```

```
Singularity> STAR --runMode genomeGenerate --genomeDir STAR_mm39vM32/ --genomeFastaFiles GRCm39.primary_assembly.genome.fa --s
jdbGTFfile gencode.vM32.annotation.gtf
        STAR --runMode genomeGenerate --genomeDir STAR_mm39vM32/ --genomeFastaFiles GRCm39.primary_assembly.genome.fa --sjdbGT
Ffile gencode.vM32.annotation.gtf
        STAR version: 2.7.10b   compiled: 2022-11-08T13:46:18+01:00 pccig4007.unil.ch:/root/src/github/STAR/source
Jun 19 10:32:47 ..... started STAR run
Jun 19 10:32:47 ... starting to generate Genome files
Jun 19 10:33:24 ..... processing annotations GTF
Jun 19 10:33:42 ... starting to sort Suffix Array. This may take a long time...
Jun 19 10:33:56 ... sorting Suffix Array chunks and saving them to disk...
Killed
Singularity>
```

Insufficient RAM

```
Sinteract -a bio-693 -c 4 -t 08:00:00 -m 80g
```

```
Singularity> STAR --runMode genomeGenerate --genomeDir STAR_mm39vM32/ --genomeFastaFiles GRCm39.primary_assembly.genome.fa --s
jdbGTFfile gencode.vM32.annotation.gtf
        STAR --runMode genomeGenerate --genomeDir STAR_mm39vM32/ --genomeFastaFiles GRCm39.primary_assembly.genome.fa --sjdbGT
Ffile gencode.vM32.annotation.gtf
        STAR version: 2.7.10b   compiled: 2022-11-08T13:46:18+01:00 pccig4007.unil.ch:/root/src/github/STAR/source
Jun 19 11:00:12 ..... started STAR run
Jun 19 11:00:12 ... starting to generate Genome files
Jun 19 11:00:47 ..... processing annotations GTF
Jun 19 11:01:04 ... starting to sort Suffix Array. This may take a long time...
Jun 19 11:01:17 ... sorting Suffix Array chunks and saving them to disk...
Jun 19 12:04:55 ... loading chunks from disk, packing SA...
Jun 19 12:05:44 ... finished generating suffix array
Jun 19 12:05:44 ... generating Suffix Array index
Jun 19 12:08:42 ... completed Suffix Array index
Jun 19 12:08:42 ..... inserting junctions into the genome indices
Jun 19 12:16:08 ... writing Genome to disk ...
Jun 19 12:16:08 ... writing Suffix Array to disk ...
Jun 19 12:16:14 ... writing SAindex to disk
Jun 19 12:16:15 ..... finished successfully
Singularity>
```

Successful

Run the aligner

```
n=SRR16219925
STAR --runMode alignReads \
  --runThreadN 20 \
  --genomeDir STAR_mm39vM32 \
  --twopassMode Basic \
  --outFileNamePrefix ${n}_mm39vM32/ \
  --outSAMtype BAM SortedByCoordinate \
  --outSAMattributes All \
  --outSAMattrRGline ID:$n LB:$n SM:$n PL:ILLUMINA \
  --readFilesCommand "gzip -dc" \
  --quantMode GeneCounts \
  --limitBAMsortRAM 16000000000 \
  --readFilesIn ${n}_?.fastq.gz
```

Example run

```
Singularity> STAR --runMode alignReads --runThreadN 20 --genomeDir STAR_mm39vM32 --twopassMode Basic --outFileNamePrefix ${n}_
mm39vM32/ --outSAMtype BAM SortedByCoordinate --outSAMattributes All --outSAMattrRGline ID:$n LB:$n SM:$n PL:ILLUMINA --readFi
lesCommand "gzip -dc" --quantMode GeneCounts --limitBAMsortRAM 16000000000 --readFilesIn ${n}_?.fastq.gz

        STAR --runMode alignReads --runThreadN 20 --genomeDir STAR_mm39vM32 --twopassMode Basic --outFileNamePrefix SRR1621992
5_mm39vM32/ --outSAMtype BAM SortedByCoordinate --outSAMattributes All --outSAMattrRGline ID:SRR16219925 LB:SRR16219925 SM:SRR
16219925 PL:ILLUMINA --readFilesCommand "gzip -dc" --quantMode GeneCounts --limitBAMsortRAM 16000000000 --readFilesIn SRR16219
925_1.fastq.gz SRR16219925_2.fastq.gz
        STAR version: 2.7.10b   compiled: 2022-11-08T13:46:18+01:00 pccig4007.unil.ch:/root/src/github/STAR/source
Jun 19 14:17:37 ..... started STAR run
Jun 19 14:17:37 ..... loading genome
Jun 19 14:17:58 ..... started 1st pass mapping
Jun 19 14:19:36 ..... finished 1st pass mapping
Jun 19 14:19:37 ..... inserting junctions into the genome indices
Jun 19 14:20:52 ..... started mapping
Jun 19 14:22:46 ..... finished mapping
Jun 19 14:22:47 ..... started sorting BAM
Jun 19 14:23:39 ..... finished successfully
Singularity>
```

samtools

- Need to get familiar with this tool to work with BAM files
- https://github.com/samtools/samtools

After STAR alignment completes

- Examine the directory `${n}_mm39vM32/`
- Examine the logs, counts and BAM files
  - `cat Log.final.out`
  - `less ReadsPerGene.out.tab`
  - `sort -k 4,4nr ${n}_mm39vM32/ReadsPerGene.out.tab | less`
  - `grep ENSMUSG00000064351.1 gencode.vM32.annotation.gtf | less`
  - `samtools view -h Aligned.sortedByCoord.out.bam | less`
  - `samtools index Aligned.sortedByCoord.out.bam`
  - ```
    samtools tview \
      -p chr1:5153351 \
      ${n}_mm39vM32/Aligned.sortedByCoord.out.bam \
      GRCm39.primary_assembly.genome.fa
    ```
  - Chr1:5171281 – probable alt splicing
  - Chr1:5194491 – variants
    ```

Get a feel for duplicates

- ```
  java -jar /usr/local/share/java/picard.jar MarkDuplicates \
    -I ${n}_mm39vM32/Aligned.sortedByCoord.out.bam \
    -O ${n}_mm39vM32/Aligned.sortedByCoord.MD.bam \
    -M ${n}_mm39vM32/marked_dup_metrics.txt
  ```
- ```
  grep -A 1 ^LIBRARY \
    ${n}_mm39vM32/marked_dup_metrics.txt | column -t
  ```

SAM / BAM file format

- SAM is a text file with
  - A header with lines starting by @
  - The rest with at least 11 columns, tab-delimited
- BAM is a compressed version of SAM

1	2	3	4	5	6	7	8	9	10	11
Name	Flag	Chr	Pos	MpQ	Cigar	PChr	Ppos	Len	Seq	Qual
String	Int	String	Int	Int	String	String	Int	Int	String	String

- To view:
  - `samtools view <file>.bam | less -S`

SAM/BAM flags

Decimal	Hex	Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing filters, such as platform/vendor quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment

Secondary and duplicates

- ```
  samtools view \
    ${n}_mm39vM32/Aligned.sortedByCoord.out.bam \
    chr1:5153351-5153600 | less
  ```
- ```
  samtools view \
    ${n}_mm39vM32/Aligned.sortedByCoord.MD.bam \
    chr1:5153351-5153600 | less
  ```
- Can play with `-f 256` and `-f 1024` to filter reads

Get raw counts

- Several possibilities – `htseq-count`, `featureCounts`
  - Htseq-count – https://doi.org/10.1093%2Fbioinformatics%2Fbtu638
  - featureCounts – https://doi.org/10.1093/bioinformatics/btt656

Get raw counts using `featureCounts`

- ```
  featureCounts -T 20 \
    --primary -s 2 -p -B -Q 10 \
    -a gencode.vM32.annotation.gtf \
    -o featureCounts.txt \
    SRR162199*_mm39vM32/Aligned.sortedByCoord.out.bam \
    2>featureCounts_log.txt
  ```
- ```
  sed 's/_mm39vM32\/Aligned.sortedByCoord.out.bam//g' \
    featureCounts.txt.summary | column -t
  ```
- ```
  cut -f1,7-  featureCounts.txt | sed 1d | \
    sed 's/_mm39vM32\/Aligned.sortedByCoord.out.bam//g' \
    > final_counts.txt
  ```
  ```