Last updated: Apr 21, 2020

# MOAGR_RNASEQ (BGI)

# Introduction

Save Copy to Evernote

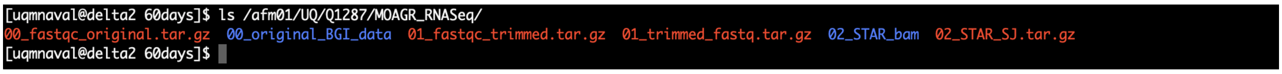
This is the BGI data on RNA-seq. It corresponds to paired end RNA-seq data 100bp + 100bp for old and young mice (2 months vs 22 months).
They contain the same data as TREV-seq from November.

In their report: around 80% of reads Q30 and 90% Q20!

**Directories:**

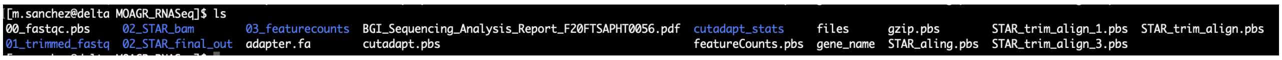
```
All the data is backed up in the RDM Q1287 belonging to the MOAGR project.
/afm01/UQ/Q1287/MOAGR_RNASeq

—> 00_original_BGI_data
        ->  19-04940-57
        ->  19-05010-24
        ->  19-05041-58
        -> 19-05059-75
—> trimmed_fastq (the data after cutadapt trimming NEXTERA adapter)
-> STAR_mapping
-> counts
```

RDM so far: (back up of fastq, fastq trimmed, bam aligned)



```
[uqmnaval@delta2 60days]$ ls /afm01/UQ/Q1287/MOAGR_RNASeq/
00_fastqc_original.tar.gz  00_original_BGI_data  01_fastqc_trimmed.tar.gz  01_trimmed_fastq.tar.gz  02_STAR_bam  02_STAR_SJ.tar.gz
[uqmnaval@delta2 60days]$
```

Delta working directory: (not backed up)



```
[m.sanchez@delta MOAGR_RNASeq]$ ls
00_fastqc.pbs     02_STAR_bam        03_featurecounts  BGI_Sequencing_Analysis_Report_F20FTSAPHT0056.pdf  cutadapt_stats     files      gzip.pbs         STAR_trim_align_1.pbs  STAR_trim_align.pbs
01_trimmed_fastq  02_STAR_final_out  adapter.fa        cutadapt.pbs                                       featureCounts.pbs  gene_name  STAR_aling.pbs   STAR_trim_align_3.pbs
```

# 00- FastQC Quality analysis:

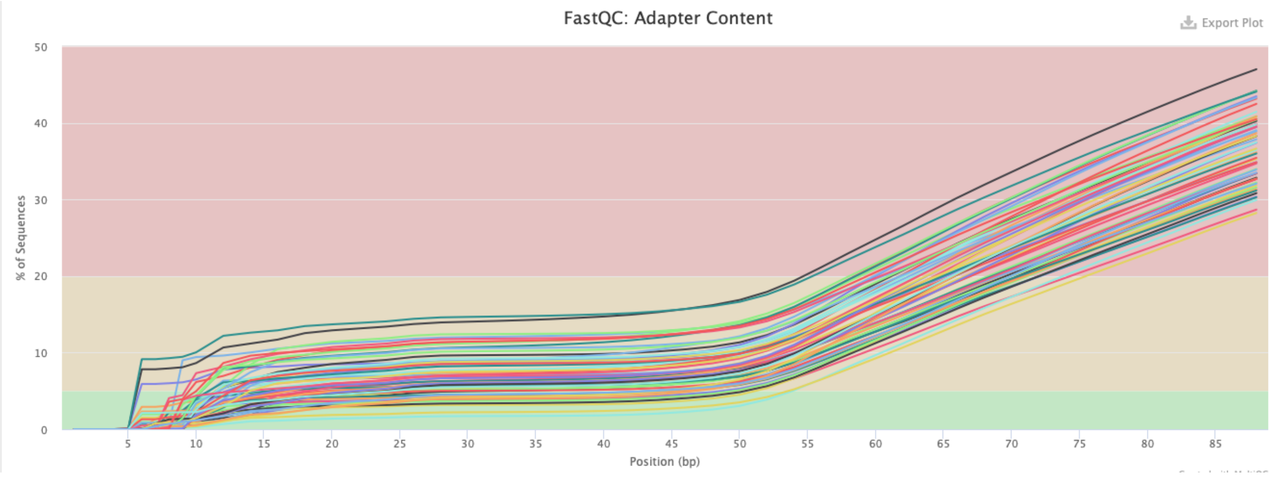
Good / MultiQC

R1_fastq_multiqc...tml
1.9 MB

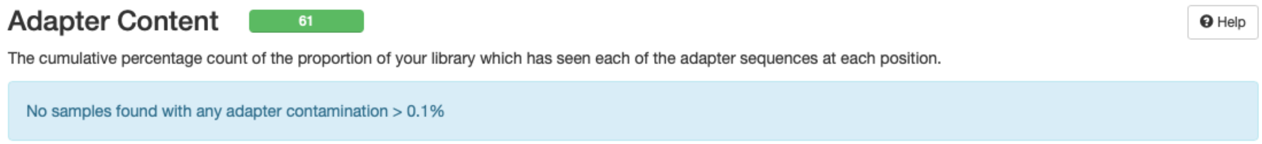
R2_fastq_multiqc...tml
1.8 MB

mmm… again we have quite a lot of adapters…. So the we have as in Trev-seq reduce the following paramenters in the mapping: ( I don't know if that is usually the case… at least we see that in the libraries prepared by Trevor… mmmm…)

As you can see… lot of adapter content! R1! That does not happen in R2!



R2 (below)



**Adapter Content** 61

The cumulative percentage count of the proportion of your library which has seen each of the adapter sequences at each position.

No samples found with any adapter contamination > 0.1%

# 01- Mapping with STAR

### Aligning with STAR: (done in delta)

```
[m.sanchez@delta MOAGR_RNASeq]$ cat STAR_aling.pbs
#!/bin/bash
#PBS -A UQ-IMB
#PBS -N STAR
#PBS -l walltime=02:00:00
#PBS -l select=1:ncpus=10:mem=40GB
#cd /30days/uqmnaval/RNAseq/01_demultiplexed/

cd /shares/common/users/m.sanchez/MOAGR_RNASEQ/MOAGR_RNASeq

module load STAR/2.4.2a


STAR --runThreadN 12 --runMode alignReads --genomeDir
/shares/common/users/m.sanchez/genomes/mm10/  --readFilesIn ${FILE}_1.fq.gz
${FILE}_2.fq.gz  --readFilesCommand zcat --outFileNamePrefix ${FILE}.annot.0.3  --
outSAMtype BAM SortedByCoordinate  --twopassMode Basic --sjdbOverhang 99 --
sjdbGTFfile /shares/common/users/m.sanchez/genomes/mm10/gencode.vM24.annotation.2.gtf
--outFilterMatchNminOverLread 0.3 --outFilterScoreMinOverLread 0.3
```

The following parameters have been tweaked to our data. As it contains high number of adapters....
We need to be more permissive on those parameters...
**--outFilterMatchNminOverLread 0.3 --outFilterScoreMinOverLread 0.3**

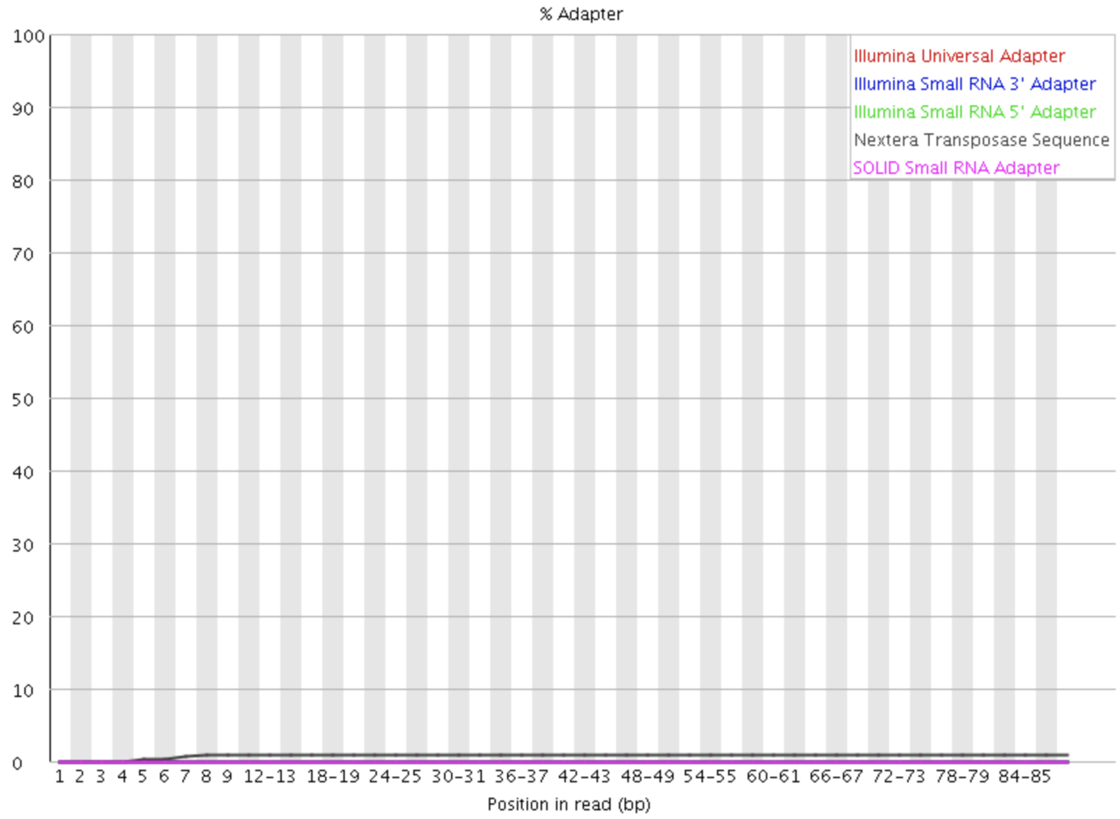
## 02- Cutting adapters

**#### CUTADAPT NEXTERA TRANSPOSASES Sequence**

```
#!/bin/bash
#PBS -l walltime=08:40:00
#PBS -l select=1:ncpus=10:mem=14GB
module load bwa
module load samtools
module load cutadapt/2.4
#cd /shares/common/users/m.sanchez/Trevseq/01_demultiplexed
cd /shares/common/users/m.sanchez/MOAGR_RNASEQ/MOAGR_RNASeq

#cd /shares/common/users/m.sanchez/MOAGR/ATAC/F19FTSAPHT1645_LIBydiR_1
### Illumina RNA PCR Primer
cutadapt -b file:adapter.fa  --out ${FILE}.trim.fq  ${FILE}
```

**#recheck in fastqC**

### ✓ Adapter Content



**STAR alignment on them...**

**Does it change the number of mapped reads?**
**# What we see if after Nextera Transposase Sequence trimming we increase more than 5% mapping!**

In fact for one sample is 6.5 of which 5% of the increase comes from 'uniquely mapped' and only 1% of multiple mapping. We also reduce form 23 to 17% the too short mapping!

**Thus, we conclude that we will trim Cutadapt for them and then perform STAR mapping!**

**Good!**

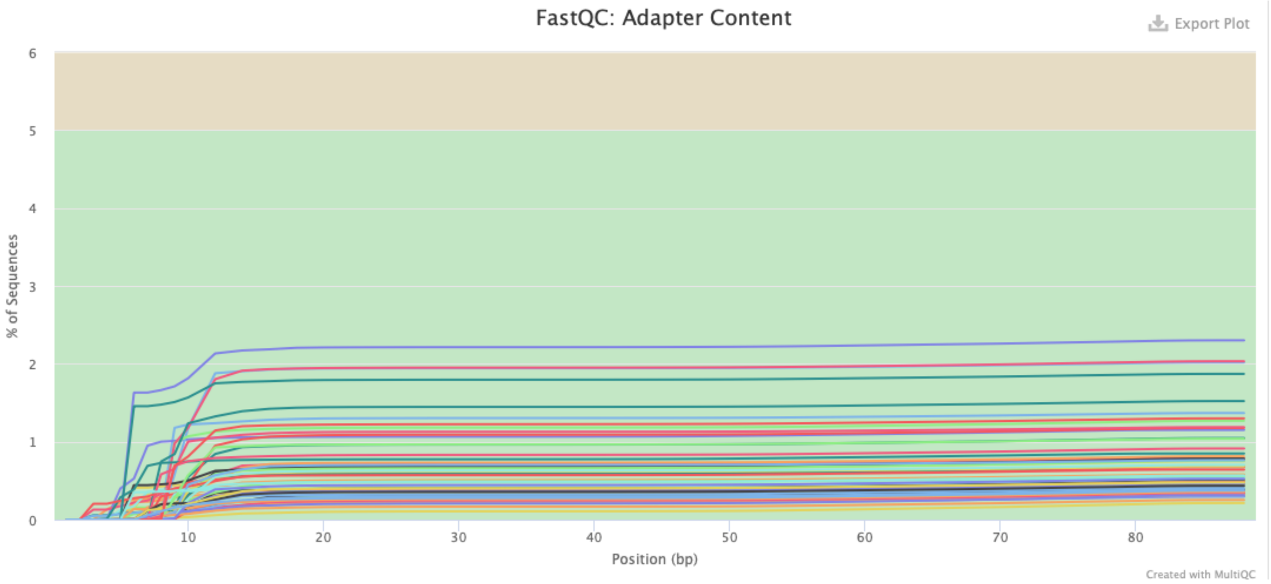
**##FASTQC / MultiQC after adapter trimming**

R1_trim_fastq_m...tml
1.9 MB

R2_trim_fastq_m...tml
1.8 MB

We have reduced the amount of NEXTERA transposes compared to original Fastq data.

Before 30-50% of samples contained adapter…. Now it has been reduced to 2%. This affects % uniquely mapped reads.. (see below)



# 03- Mapping STAR without adapters:

```
for i in `cat files`; do qsub -v FILE=$i STAR_trim_align.pbs ;done
```

**#Summary mapping stats**

```
[m.sanchez@delta MOAGR_RNASeq]$ cat STAR_trim_align.pbs
#!/bin/bash
#PBS -A UQ-IMB
```

```
#PBS -N STAR
#PBS -l walltime=02:00:00
#PBS -l select=1:ncpus=10:mem=40GB
#cd /30days/uqmnaval/RNAseq/01_demultiplexed/

cd /shares/common/users/m.sanchez/MOAGR_RNASEQ/MOAGR_RNASeq


module load STAR/2.4.2a
#source /gpfs1/homes/uqmnaval/miniconda/etc/profile.d/conda.sh

##The one used
STAR --runThreadN 12 --runMode alignReads --genomeDir
/shares/common/users/m.sanchez/genomes/mm10/  --readFilesIn ${FILE}_1.fq.gz.trim.fq
${FILE}_2.fq.gz.trim.fq  --readFilesCommand zcat --outFileNamePrefix
${FILE}.annot.0.3  --outSAMtype BAM SortedByCoordinate  --twopassMode Basic --
sjdbOverhang 99 --sjdbGTFfile
/shares/common/users/m.sanchez/genomes/mm10/gencode.vM24.annotation.2.gtf --
outFilterMatchNminOverLread 0.3 --outFilterScoreMinOverLread 0.3
```

**Mapping Summary stats for RNA-seq:**

STAR_mapping_...lsx
30 KB

## 04- Feature Counts

We make use of latest gencode version mm10 -> vM24
We count for gene_name.

```
[m.sanchez@delta MOAGR_RNASeq]$ cat featureCounts.pbs
#!/bin/bash
#PBS -A UQ-IMB
#PBS -N featurecounts
#PBS -l walltime=03:00:00
#PBS -l select=1:ncpus=10:mem=40GB

#cd /shares/common/users/m.sanchez/Trevseq/01_demultiplexed

cd /shares/common/users/m.sanchez/MOAGR_RNASEQ/MOAGR_RNASeq


/shares/common/users/m.sanchez/subread-2.0.0-Linux-x86_64/bin/featureCounts -t exon -
g gene_name -M -a
/shares/common/users/m.sanchez/genomes/mm10/gencode.vM24.annotation.2.gtf -o
${FILE}.counts.txt  ${FILE}
```

```
for i in `ls */*/*annot.0.3Aligned.sortedByCoord.out.bam`; do qsub -v FILE=$i
featureCounts.pbs ;done
```

**### We put data together / Final counts:**

all_counts_BGI.txt
8.7 MB

**#### Now we need to know what is what in each header/file name!**

## 05- Processing of files / DESEQ or EDGER???

**—> Assess for bias on the data? Batch effect?**
**--> Maybe we should have the names of the data… better labels, tissue cell type…?**

To perform or to have a PCA?