UNIVERSITÀ
DEGLI STUDI
DE L'AQUILA

DISIM
Dipartimento di Ingegneria
e Scienze dell'Informazione
e Matematica

Experiment Description

2021/2022

Course: **Bioinformatics**

University of L'Aquila

Federico Di Menna federico.dimenna@student.univaq.it Matr. 275332

Reference paper: https://www.sciencedirect.com/science/article/pii/S266591312030131X
Github repository: https://github.com/federix98/Bioinformatics

miRNA Sequencing Experiment

Identification of new epigenetic factors involved in the pathogenesis of hereditary-familial breast / ovarian cancer.

DESCRIPTION OF THE OBJECTIVES: Aim of the study is the analysis of the expression levels of circulating miRNAs in patients with hereditary-familial breast / ovarian cancer (BC / OC) with (BRCA) or without (non-BRCA) pathogenetic variants in the high penetrance genes BRCA1 and BRCA2, further stratified according to genetic risk (high, medium and low genetic risk). The expected results include the identification of circulating microRNAs differentially expressed in the groups of patients under study. The analysis of these non-invasive epigenetic biomarkers, identifiable by rapid laboratory tests, has a strong clinical impact, supporting the adoption of prevention, screening and personalized medicine programs for specific subgroups of patients, and their relatives, with a predisposition family-heir. The in silico analysis of the target genes of the identified microRNAs will also be able to investigate the molecular pathogenesis mechanisms of this category of breast / ovarian tumors.

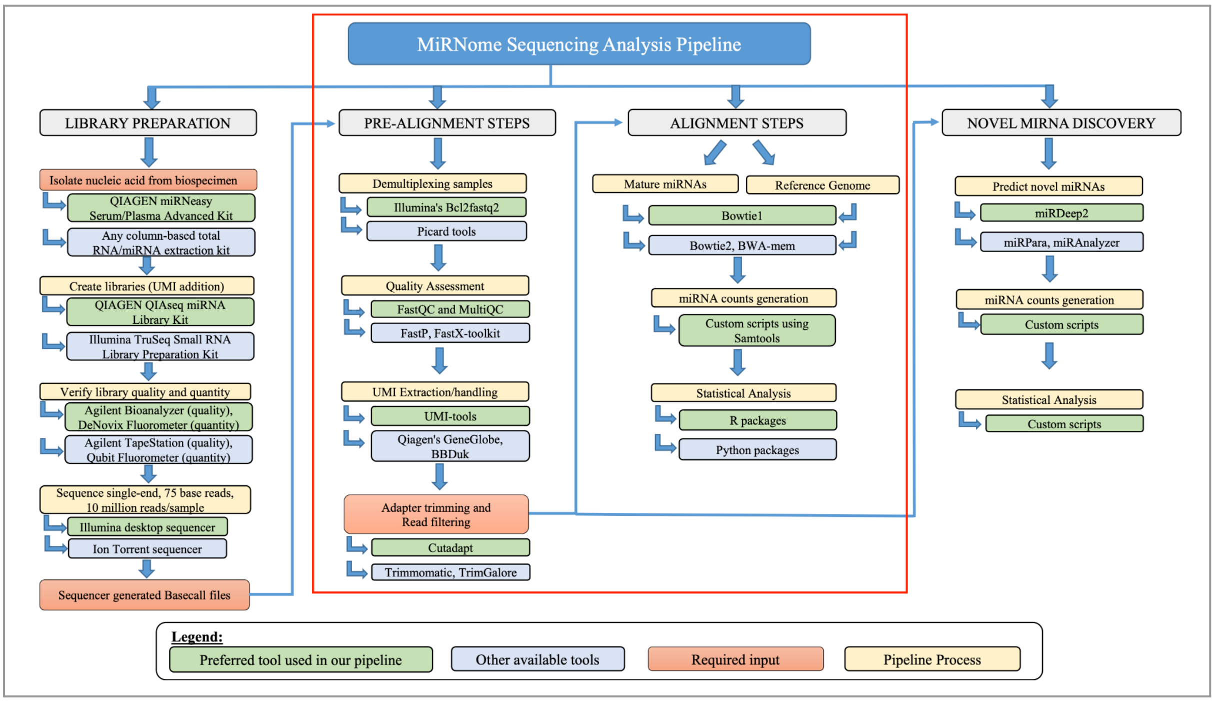
Samples

- 21 BC hereditary-familial cases (mean age 47.3yy) of which 14 BRCA and 7 non-BRCA
- 3 controls (mean age 48.3yy) age t-test p = 0.9

Goal:

Identify and analyze the differential expression between different groups (different experimental conditions).

Pipeline


MiRNome Sequencing Analysis Pipeline
LIBRARY PREPARATION
Isolate nucleic acid from biospecimen
QIAGEN miRNeasy Serum/Plasma Advanced Kit
Any column-based total RNA/miRNA extraction kit
Create libraries (UMI addition)
QIAGEN QIAseq miRNA Library Kit
Illumina TruSeq Small RNA Library Preparation Kit
Verify library quality and quantity
Agilent Bioanalyzer (quality), DeNovix Fluorometer (quantity)
Agilent TapeStation (quality), Qubit Fluorometer (quantity)
Sequence single-end, 75 base reads, 10 million reads/sample
Illumina desktop sequencer
Ion Torrent sequencer
Sequencer generated Basecall files
PRE-ALIGNMENT STEPS
Demultiplexing samples
Illumina's Bcl2fastq2
Picard tools
Quality Assessment
FastQC and MultiQC
FastP, FastX-toolkit
UMI Extraction/handling
UMI-tools
Qiagen's GeneGlobe, BBDuk
Adapter trimming and Read filtering
Cutadapt
Trimmomatic, TrimGalore
ALIGNMENT STEPS
Mature miRNAs
Reference Genome
Bowtie1
Bowtie2, BWA-mem
miRNA counts generation
Custom scripts using Samtools
Statistical Analysis
R packages
Python packages
NOVEL MIRNA DISCOVERY
Predict novel miRNAs
miRDeep2
miRPara, miRAnalyzer
miRNA counts generation
Custom scripts
Statistical Analysis
Custom scripts
Legend:
Preferred tool used in our pipeline
Other available tools
Required input
Pipeline Process

The pipeline has been implemented using a shell script available inside the repository (*pipeline.sh* file).

The pipeline is executed over all the .fastq.gz files we have by using a for loop in order to iterate them (the files have been renamed for sake of simplicity):

```
for i in 70459 70460 70461 70462 74665 74666 74667 74668 74669 74670 74671 74672
74673 74674 74675 74676 74677 74678 74679 74680 74681 74682 74683 74684
```

Where the variables have been instantiated in this way:

- **$INPUTDIR:** Directory which contains QCReports folder and all the ***.fastq.gz** input files.
- **$OUTPUTDIR**: Where to store the results
- **$i**: name of the file which we are analyzing. Note that this command is contained in a for statement.

Library Preparation

QIAseq miRNA library kit (QIAGEN, Hilden, Germany) has been used for library preparation following the manufacturer's instructions.

- 4 cp: Libraries were then prepared for sequencing and sequenced on single-end 150 bp mode on NovaSeq 6000
- 20 cp: Libraries were sequenced on single-end 75 bp mode on NextSeq 500

Primary bioinformatic analysis includes:

- Base calling and demultiplexing. Processing raw data for both format conversion and de-multiplexing by Bcl2Fastq 2.20 version of the Illumina pipeline
- FastQC

Pre-Alignment Steps

Quality Assessment

Quality assessment has been done using the **FastQC** tool. The tool has been launched as command line instruction as follows:

```
fastqc -o $INPUTDIR/QCReports $INPUTDIR/$i".fastq.gz" -t 16
```

Parameters:

-t --threads *Specifies the number of files which can be processed*
simultaneously. Each thread will be allocated 250MB of
memory so you shouldn't run more threads than your
available memory will cope with, and not more than
6 threads on a 32 bit machine

-o --outdir *Create all output files in the specified output directory.*
Please note that this directory must exist as the program
will not create it. If this option is not set then the
output file for each sequence file is created in the same
directory as the sequence file which was processed.

Results

All the results file are available within the repository here:
https://github.com/federix98/Bioinformatics/tree/main/miRNA_Seq/Fastq-files/QCReports

UMI Extraction / Handling

UMI Extraction has been performed using the **umi_tools** tool.

Instruction:

```
umi_tools extract --stdin=$INPUTDIR/$i".fastq.gz"
--log=$OUTPUTDIR/$i/$i"-UMIextraction-fromrawreads.log"
--stdout=$OUTPUTDIR/$i/$i"-directUMIextracted.fastq" --extract-method=regex
--bc-pattern='.+(?P<discard_1>AACTGTAGGCACCATCAAT){s<=2}(?P<umi_1>.{12})(?P<discard_2>.+)'
```

Parameters:

--stdin: specify the path of the input file to analyze
--log: log file path (all the log files are available inside the repository under analysis/name_sample/.log)*
--stdout: where to write the output of the extraction (fastq file)
--extract-method: How to extract the umi +/- cell barcodes, Choose from 'string' or 'regex'
--bc-pattern: specify the barcode pattern

Log file for the first sample:
https://github.com/federix98/Bioinformatics/blob/main/miRNA_Seq/analysis/70459/70459-UMIextraction-fromrawreads.log

All the log files of this phase are in analysis/sample_name/ and they are named:
**-UMIextraction-fromrawreads.log.*

Adapter Trimming and Read Filtering

Adapter trimming has been done using the **cutadapt v4.1** tool. We only want to take the samples between 18 and 30 of length as specified in the paper.

Instruction:

```
cutadapt --minimum-length=$MINLENGTH --maximum-length=$MAXLENGTH -o
$OUTPUTDIR/$i/$i"-directUMIextracted-min"$MINLENGTH"max"$MAXLENGTH"L.fastq"
$OUTPUTDIR/$i/$i"-directUMIextracted.fastq" >
$OUTPUTDIR/$i/$i"-directUMIextracted-readlengthfilter-cutadapt[min"$MINLENGTH"max"$
MAXLENGTH"L].log"
```

Parameters:
--minimum-length: min length to consider in the filter (our case is 18)
--maximum-length: max length to consider in the filer (our case is 30)
-o : output d
> log inside the log file specified

Alignment

For the alignment phase I used **bowtie**.

Instruction:

```
   bowtie -n 0 -l 30 --norc --best --strata -m 1 --threads 16 mature
$OUTPUTDIR/$i/$i"-directUMIextracted-min"$MINLENGTH"max"$MAXLENGTH"L.fastq" --un
$OUTPUTDIR/$i/$i"-maturemiRNA-unalignedReads-bowtie1-beststratam1.fastq" -S
$OUTPUTDIR/$i/$i"-maturemiRNA-aligned-bowtie1-beststratam1.sam" 2>
$OUTPUTDIR/$i/$i"-bowtie-maturemiRNA-beststratam1.log"
```

best and strata parameters are about reporting:

--best *hits guaranteed best stratum; ties broken by quality*
--strata *hits in sub-optimal strata aren't reported (requires --best)*

We are specifying with this instruction that we don't want any dismatches among mature and our sequences specifying *-n 0 -l 30* (***-n/--seedmms <int> max mismatches in seed (can be 0-3, default: -n 2), -l/--seedlen <int> seed length for -n (default: 28)***).

Sort and Index

To sort and index aligned results we use the following two instructions

```
samtools sort $OUTPUTDIR/$i/$i"-maturemiRNA-aligned-bowtie1-beststratam1.sam" >
$OUTPUTDIR/$i/$i"-maturemiRNA-aligned-bowtie1-beststratam1.bam"
   samtools index $OUTPUTDIR/$i/$i"-maturemiRNA-aligned-bowtie1-beststratam1.bam"
```

Counting Steps

The following instructions are the sub-pipeline for the counting steps:

```
umi_tools count --method=unique --per-contig -I
$OUTPUTDIR/$i/$i"-maturemiRNA-aligned-bowtie1-beststratam1.bam" -L
$OUTPUTDIR/$i/$i"_counts-uniquemethod-maturemiRNA.log" -S
$OUTPUTDIR/$i/$i"_counts-finaloutput-uniquemethod-maturemiRNA.txt"
   echo "Deduplicating the aligned BAM"
   umi_tools dedup --method=unique -I
$OUTPUTDIR/$i/$i"-maturemiRNA-aligned-bowtie1-beststratam1.bam" -S
$OUTPUTDIR/$i/$i"_deduplicated-matureMirna-uniquemethod-beststratam1.bam" -L
$OUTPUTDIR/$i/$i"-deduplicate-matureMirna-uniquemethod-beststratam1.log"
   echo "Indexing the BAM output for finding counts"
   samtools index
$OUTPUTDIR/$i/$i"_deduplicated-matureMirna-uniquemethod-beststratam1.bam"
   echo "Generation of miRNA counts for file $i"
   samtools idxstats
$OUTPUTDIR/$i/$i"_deduplicated-matureMirna-uniquemethod-beststratam1.bam" | cut
-f1,3 - | sed "1s/^/miRNA\t${i}-miRNAcount\n/" - >
$OUTPUTDIR/maturemiRNAcounts/$i"-maturemiRNA-counts.txt"
```

Genome

We need now to repeat the alignment part with the genome with respect to the unaligned sequences we have. We can to this using the following sub pipeline (similar to the previous instructions)

```
bowtie -n 1 -l 32 --norc --best --strata -m 1 --threads 16
GCA_000001405.15_GRCh38_no_alt_analysis_set
$OUTPUTDIR/$i/$i"-maturemiRNA-unalignedReads-bowtie1-beststratam1.fastq" --al
$OUTPUTDIR/$i/$i"-genome-alignedReads-bowtie1-beststratam1.fastq" -S
$OUTPUTDIR/$i/$i"-genomeaftermiRNA-aligned-bowtie1-beststratam1.sam" 2>
$OUTPUTDIR/$i/$i"-bowtie-genomeaftermiRNA-beststratam1.log"
   echo "Sorting and indexing again for downstream analyses"
   samtools sort
$OUTPUTDIR/$i/$i"-genomeaftermiRNA-aligned-bowtie1-beststratam1.sam" >
$OUTPUTDIR/$i/$i"-genomeaftermiRNA-aligned-bowtie1-beststratam1.bam"
   samtools index
$OUTPUTDIR/$i/$i"-genomeaftermiRNA-aligned-bowtie1-beststratam1.bam"
   # rm -rf $OUTPUTDIR/$i/$i"-genomeaftermiRNA-aligned-bowtie1-beststratam1.sam"
   echo "Deduplicating the BAM file before counting through custom scripts"
   umi_tools dedup --method=unique -I
$OUTPUTDIR/$i/$i"-genomeaftermiRNA-aligned-bowtie1-beststratam1.bam" -S
$OUTPUTDIR/$i/$i"_deduplicated-genomeaftermiRNA-uniquemethod.bam" -L
$OUTPUTDIR/$i/$i"-deduplicate-genomeaftermiRNA-uniquemethod.log"
   echo "Index the deduplicated BAM file"
   samtools index $OUTPUTDIR/$i/$i"_deduplicated-genomeaftermiRNA-uniquemethod.bam"
   echo "Actual filtering sam step only for overlapping microRNA locations"
   tagBam -i $OUTPUTDIR/$i/$i"_deduplicated-genomeaftermiRNA-uniquemethod.bam"
-files hsa-genome-miRBase22v-onlymiRNAs-convforTagBAM.bed -names -tag XQ >
$OUTPUTDIR/$i/$i"-tagged.bam"
```

We can notice here that they increased the leniency by a bit using 1 mismatch instead of 0.

After the execution of this part of the pipeline we have to run the file ***get-count.sh*** in order to make the real count. Indeed in the explanation document they write:

14. For each of the 2656 miRNAs per sample, use command 'Samtools view tagged.bam | grep -c "miRNAname" >> genome-basedcounts.tsv'. Apply 'sort -k1 | uniq' on the counts file to retain only the unique miRNA counts.

Statistical Analysis

At this point we have all the counts for both the genome and mature alignment and we can use **Deseq2** for generating statistical plots and data. The results are described in the powerpoint presentation.

Tools

QC	FASTQC	v0.11.9
	UMI TOOLS	v.1.1.2
	CUTADAPT	v4.1
BOW TIE	BOWTIE	v1.2.2
	SAMTOOLS	v1.16.1