RNA sequencing analyses in families with multiple affected of bipolar disorder

Inés García-Ortiz

2022-11-29

The purpose of this vignette is to complement, describe and explain the scripts created for my master thesis project, in which RNAseq data from 31 different individuals undergo different analyses.

# How does the data look?

We have RNAseq data from 31 different individuals of 8 different families with high number of Bipolar Disorder affected relatives. They have two different IDs: the individual ID present in the databank of the group and the sample ID used for the RNAseq itself.

| Databank ID | RNAseq ID | Condition | Gender | Age | Family |
| --- | --- | --- | --- | --- | --- |
| 21873 | BWH011 | Affected | Male | adult | 210 |
| 12005 | BWH002 | Unaffected | Male | elder | 120 |
| 12001 | BWH003 | Unaffected | Male | adult | 120 |
| 12002 | BWH004 | Unaffected | Male | adult | 120 |
| 12003 | BWH005 | Affected | Female | adult | 120 |
| 12004 | BWH006 | Affected | Female | adult | 120 |
| 21769 | BWH007 | Affected | Male | adult | 172 |
| 21768 | BWH008 | Unaffected | Female | adult | 172 |
| 21770 | BWH009 | Affected | Female | young | 172 |
| 21771 | BWH010 | Affected | Female | young | 172 |
| 12504 | BWH012 | Unaffected | Male | elder | 125 |
| 12500 | BWH013 | Unaffected | Female | elder | 125 |
| 12506 | BWH014 | Affected | Male | adult | 125 |
| 2701 | BWH015 | Unaffected | Male | elder | 27 |
| 2700 | BWH016 | Affected | Female | elder | 27 |
| 2705 | BWH017 | Affected | Female | young | 27 |
| 2706 | BWH018 | Affected | Male | adult | 27 |
| 2707 | BWH019 | Unaffected | Female | adult | 27 |
| 7203 | BWH000 | Affected | Male | young | 72 |
| 7204 | BWH001 | Affected | Male | young | 72 |
| 7201 | BWH020 | Unaffected | Male | elder | 72 |
| 7200 | BWH021 | Unaffected | Female | elder | 72 |
| 7202 | BWH022 | Affected | Male | young | 72 |
| 12709 | BWH023 | Unaffected | Male | elder | 127 |
| 12700 | BWH024 | Unaffected | Female | elder | 127 |
| 12708 | BWH025 | Affected | Male | young | 127 |
| 6504 | BWH026 | Affected | Male | elder | 65 |
| 6506 | BWH027 | Unaffected | Female | elder | 65 |
| 6502 | BWH028 | Unaffected | Male | adult | 65 |
| 6503 | BWH029 | Unaffected | Female | adult | 65 |
| 6500 | BWH030 | Affected | Male | adult | 65 |

The pedigree family structures is the following:

(insertar imagen pedigris)

# Data processing in bash

## Preprocessing of reads

The original fastq files come in three files for running files, that had to be concatenated.

#!/bin/bash  
cd /home/proyectos/genpsych/RNAseq/RNAseq\_FastaQ/  
for d in BWH--[0-9][0-9][0-9]  
do  
 cd $d  
 #Uncompress  
 gunzip -c \*R1.fastq.gz > $d\_R1.fastq  
 gunzip -c \*R2.fastq.gz > $d\_R2.fastq  
 #Move to my folder  
 mv \*fastq ../fastqInes   
 cd ..  
 echo continuing  
done

A first fastQC assay was done:

#!/bin/bash  
module load fastqc/0.11.9  
fastqc \*fastq -t 20 -o FASTQC

Though the overall quality of the reads was very good, I decided to trim a little bit the ends using Trimmomatic. To generate the commands for each sample the script used was:

#!/bin/bash  
for i in `ls -1 \*R1.fastq | sed 's/\\_R1.fastq//'`;   
do   
echo java -jar /usr/local/trimmomatic/0.39/trimmomatic-0.39.jar PE -threads 20   
-phred33 $i\\_R1.fastq $i\\_R2.fastq $i\\_R1\_paired.fq.gz $i\\_R1\_unpaired.fq.gz $i\\_R2\_paired.fq.gz $i\\_R2\_unpaired.fq.gz   
SLIDINGWINDOW:4:20 MINLEN:50 >> trim\_cmd;   
done

The command first has to call to java and declare that it is a paired end (*PE*) experiment. The sliding window (*SLIDINGWINDOW:4:20*) in trimmomatic means that it will scan the read from the 5’ end in groups of **4** and when the average quality drops from **20** it will clip it. We will only accept reads with a minimum length (*MINLEN:50*) of 50 basepairs.

Giving the following whole script:

#!/bin/bash  
module load trimmomatic/0.39  
java -jar /usr/local/trimmomatic/0.39/trimmomatic-0.39.jar PE -threads 20 -phred33 BWH--000\_R1.fastq BWH--000\_R2.fastq BWH--000\_R1\_paired.fq.gz BWH--000\_R1\_unpaired.fq.gz BWH--000\_R2\_paired.fq.gz BWH--000\_R2\_unpaired.fq.gz SLIDINGWINDOW:4:20 MINLEN:50  
java -jar /usr/local/trimmomatic/0.39/trimmomatic-0.39.jar PE -threads 20 -phred33 BWH--001\_R1.fastq BWH--001\_R2.fastq BWH--001\_R1\_paired.fq.gz BWH--001\_R1\_unpaired.fq.gz BWH--001\_R2\_paired.fq.gz BWH--001\_R2\_unpaired.fq.gz SLIDINGWINDOW:4:20 MINLEN:50  
java -jar /usr/local/trimmomatic/0.39/trimmomatic-0.39.jar PE -threads 20 -phred33 BWH--002\_R1.fastq BWH--002\_R2.fastq BWH--002\_R1\_paired.fq.gz BWH--002\_R1\_unpaired.fq.gz BWH--002\_R2\_paired.fq.gz BWH--002\_R2\_unpaired.fq.gz SLIDINGWINDOW:4:20 MINLEN:50  
java -jar /usr/local/trimmomatic/0.39/trimmomatic-0.39.jar PE -threads 20 -phred33 BWH--003\_R1.fastq BWH--003\_R2.fastq BWH--003\_R1\_paired.fq.gz BWH--003\_R1\_unpaired.fq.gz BWH--003\_R2\_paired.fq.gz BWH--003\_R2\_unpaired.fq.gz SLIDINGWINDOW:4:20 MINLEN:50  
java -jar /usr/local/trimmomatic/0.39/trimmomatic-0.39.jar PE -threads 20 -phred33 BWH--004\_R1.fastq BWH--004\_R2.fastq BWH--004\_R1\_paired.fq.gz BWH--004\_R1\_unpaired.fq.gz BWH--004\_R2\_paired.fq.gz BWH--004\_R2\_unpaired.fq.gz SLIDINGWINDOW:4:20 MINLEN:50  
java -jar /usr/local/trimmomatic/0.39/trimmomatic-0.39.jar PE -threads 20 -phred33 BWH--005\_R1.fastq BWH--005\_R2.fastq BWH--005\_R1\_paired.fq.gz BWH--005\_R1\_unpaired.fq.gz BWH--005\_R2\_paired.fq.gz BWH--005\_R2\_unpaired.fq.gz SLIDINGWINDOW:4:20 MINLEN:50  
java -jar /usr/local/trimmomatic/0.39/trimmomatic-0.39.jar PE -threads 20 -phred33 BWH--006\_R1.fastq BWH--006\_R2.fastq BWH--006\_R1\_paired.fq.gz BWH--006\_R1\_unpaired.fq.gz BWH--006\_R2\_paired.fq.gz BWH--006\_R2\_unpaired.fq.gz SLIDINGWINDOW:4:20 MINLEN:50  
java -jar /usr/local/trimmomatic/0.39/trimmomatic-0.39.jar PE -threads 20 -phred33 BWH--007\_R1.fastq BWH--007\_R2.fastq BWH--007\_R1\_paired.fq.gz BWH--007\_R1\_unpaired.fq.gz BWH--007\_R2\_paired.fq.gz BWH--007\_R2\_unpaired.fq.gz SLIDINGWINDOW:4:20 MINLEN:50  
java -jar /usr/local/trimmomatic/0.39/trimmomatic-0.39.jar PE -threads 20 -phred33 BWH--008\_R1.fastq BWH--008\_R2.fastq BWH--008\_R1\_paired.fq.gz BWH--008\_R1\_unpaired.fq.gz BWH--008\_R2\_paired.fq.gz BWH--008\_R2\_unpaired.fq.gz SLIDINGWINDOW:4:20 MINLEN:50  
java -jar /usr/local/trimmomatic/0.39/trimmomatic-0.39.jar PE -threads 20 -phred33 BWH--009\_R1.fastq BWH--009\_R2.fastq BWH--009\_R1\_paired.fq.gz BWH--009\_R1\_unpaired.fq.gz BWH--009\_R2\_paired.fq.gz BWH--009\_R2\_unpaired.fq.gz SLIDINGWINDOW:4:20 MINLEN:50  
java -jar /usr/local/trimmomatic/0.39/trimmomatic-0.39.jar PE -threads 20 -phred33 BWH--010\_R1.fastq BWH--010\_R2.fastq BWH--010\_R1\_paired.fq.gz BWH--010\_R1\_unpaired.fq.gz BWH--010\_R2\_paired.fq.gz BWH--010\_R2\_unpaired.fq.gz SLIDINGWINDOW:4:20 MINLEN:50  
java -jar /usr/local/trimmomatic/0.39/trimmomatic-0.39.jar PE -threads 20 -phred33 BWH--011\_R1.fastq BWH--011\_R2.fastq BWH--011\_R1\_paired.fq.gz BWH--011\_R1\_unpaired.fq.gz BWH--011\_R2\_paired.fq.gz BWH--011\_R2\_unpaired.fq.gz SLIDINGWINDOW:4:20 MINLEN:50  
java -jar /usr/local/trimmomatic/0.39/trimmomatic-0.39.jar PE -threads 20 -phred33 BWH--012\_R1.fastq BWH--012\_R2.fastq BWH--012\_R1\_paired.fq.gz BWH--012\_R1\_unpaired.fq.gz BWH--012\_R2\_paired.fq.gz BWH--012\_R2\_unpaired.fq.gz SLIDINGWINDOW:4:20 MINLEN:50  
java -jar /usr/local/trimmomatic/0.39/trimmomatic-0.39.jar PE -threads 20 -phred33 BWH--013\_R1.fastq BWH--013\_R2.fastq BWH--013\_R1\_paired.fq.gz BWH--013\_R1\_unpaired.fq.gz BWH--013\_R2\_paired.fq.gz BWH--013\_R2\_unpaired.fq.gz SLIDINGWINDOW:4:20 MINLEN:50  
java -jar /usr/local/trimmomatic/0.39/trimmomatic-0.39.jar PE -threads 20 -phred33 BWH--014\_R1.fastq BWH--014\_R2.fastq BWH--014\_R1\_paired.fq.gz BWH--014\_R1\_unpaired.fq.gz BWH--014\_R2\_paired.fq.gz BWH--014\_R2\_unpaired.fq.gz SLIDINGWINDOW:4:20 MINLEN:50  
java -jar /usr/local/trimmomatic/0.39/trimmomatic-0.39.jar PE -threads 20 -phred33 BWH--015\_R1.fastq BWH--015\_R2.fastq BWH--015\_R1\_paired.fq.gz BWH--015\_R1\_unpaired.fq.gz BWH--015\_R2\_paired.fq.gz BWH--015\_R2\_unpaired.fq.gz SLIDINGWINDOW:4:20 MINLEN:50  
java -jar /usr/local/trimmomatic/0.39/trimmomatic-0.39.jar PE -threads 20 -phred33 BWH--016\_R1.fastq BWH--016\_R2.fastq BWH--016\_R1\_paired.fq.gz BWH--016\_R1\_unpaired.fq.gz BWH--016\_R2\_paired.fq.gz BWH--016\_R2\_unpaired.fq.gz SLIDINGWINDOW:4:20 MINLEN:50  
java -jar /usr/local/trimmomatic/0.39/trimmomatic-0.39.jar PE -threads 20 -phred33 BWH--017\_R1.fastq BWH--017\_R2.fastq BWH--017\_R1\_paired.fq.gz BWH--017\_R1\_unpaired.fq.gz BWH--017\_R2\_paired.fq.gz BWH--017\_R2\_unpaired.fq.gz SLIDINGWINDOW:4:20 MINLEN:50  
java -jar /usr/local/trimmomatic/0.39/trimmomatic-0.39.jar PE -threads 20 -phred33 BWH--018\_R1.fastq BWH--018\_R2.fastq BWH--018\_R1\_paired.fq.gz BWH--018\_R1\_unpaired.fq.gz BWH--018\_R2\_paired.fq.gz BWH--018\_R2\_unpaired.fq.gz SLIDINGWINDOW:4:20 MINLEN:50  
java -jar /usr/local/trimmomatic/0.39/trimmomatic-0.39.jar PE -threads 20 -phred33 BWH--019\_R1.fastq BWH--019\_R2.fastq BWH--019\_R1\_paired.fq.gz BWH--019\_R1\_unpaired.fq.gz BWH--019\_R2\_paired.fq.gz BWH--019\_R2\_unpaired.fq.gz SLIDINGWINDOW:4:20 MINLEN:50  
java -jar /usr/local/trimmomatic/0.39/trimmomatic-0.39.jar PE -threads 20 -phred33 BWH--020\_R1.fastq BWH--020\_R2.fastq BWH--020\_R1\_paired.fq.gz BWH--020\_R1\_unpaired.fq.gz BWH--020\_R2\_paired.fq.gz BWH--020\_R2\_unpaired.fq.gz SLIDINGWINDOW:4:20 MINLEN:50  
java -jar /usr/local/trimmomatic/0.39/trimmomatic-0.39.jar PE -threads 20 -phred33 BWH--021\_R1.fastq BWH--021\_R2.fastq BWH--021\_R1\_paired.fq.gz BWH--021\_R1\_unpaired.fq.gz BWH--021\_R2\_paired.fq.gz BWH--021\_R2\_unpaired.fq.gz SLIDINGWINDOW:4:20 MINLEN:50  
java -jar /usr/local/trimmomatic/0.39/trimmomatic-0.39.jar PE -threads 20 -phred33 BWH--022\_R1.fastq BWH--022\_R2.fastq BWH--022\_R1\_paired.fq.gz BWH--022\_R1\_unpaired.fq.gz BWH--022\_R2\_paired.fq.gz BWH--022\_R2\_unpaired.fq.gz SLIDINGWINDOW:4:20 MINLEN:50  
java -jar /usr/local/trimmomatic/0.39/trimmomatic-0.39.jar PE -threads 20 -phred33 BWH--023\_R1.fastq BWH--023\_R2.fastq BWH--023\_R1\_paired.fq.gz BWH--023\_R1\_unpaired.fq.gz BWH--023\_R2\_paired.fq.gz BWH--023\_R2\_unpaired.fq.gz SLIDINGWINDOW:4:20 MINLEN:50  
java -jar /usr/local/trimmomatic/0.39/trimmomatic-0.39.jar PE -threads 20 -phred33 BWH--024\_R1.fastq BWH--024\_R2.fastq BWH--024\_R1\_paired.fq.gz BWH--024\_R1\_unpaired.fq.gz BWH--024\_R2\_paired.fq.gz BWH--024\_R2\_unpaired.fq.gz SLIDINGWINDOW:4:20 MINLEN:50  
java -jar /usr/local/trimmomatic/0.39/trimmomatic-0.39.jar PE -threads 20 -phred33 BWH--025\_R1.fastq BWH--025\_R2.fastq BWH--025\_R1\_paired.fq.gz BWH--025\_R1\_unpaired.fq.gz BWH--025\_R2\_paired.fq.gz BWH--025\_R2\_unpaired.fq.gz SLIDINGWINDOW:4:20 MINLEN:50  
java -jar /usr/local/trimmomatic/0.39/trimmomatic-0.39.jar PE -threads 20 -phred33 BWH--026\_R1.fastq BWH--026\_R2.fastq BWH--026\_R1\_paired.fq.gz BWH--026\_R1\_unpaired.fq.gz BWH--026\_R2\_paired.fq.gz BWH--026\_R2\_unpaired.fq.gz SLIDINGWINDOW:4:20 MINLEN:50  
java -jar /usr/local/trimmomatic/0.39/trimmomatic-0.39.jar PE -threads 20 -phred33 BWH--027\_R1.fastq BWH--027\_R2.fastq BWH--027\_R1\_paired.fq.gz BWH--027\_R1\_unpaired.fq.gz BWH--027\_R2\_paired.fq.gz BWH--027\_R2\_unpaired.fq.gz SLIDINGWINDOW:4:20 MINLEN:50  
java -jar /usr/local/trimmomatic/0.39/trimmomatic-0.39.jar PE -threads 20 -phred33 BWH--028\_R1.fastq BWH--028\_R2.fastq BWH--028\_R1\_paired.fq.gz BWH--028\_R1\_unpaired.fq.gz BWH--028\_R2\_paired.fq.gz BWH--028\_R2\_unpaired.fq.gz SLIDINGWINDOW:4:20 MINLEN:50  
java -jar /usr/local/trimmomatic/0.39/trimmomatic-0.39.jar PE -threads 20 -phred33 BWH--029\_R1.fastq BWH--029\_R2.fastq BWH--029\_R1\_paired.fq.gz BWH--029\_R1\_unpaired.fq.gz BWH--029\_R2\_paired.fq.gz BWH--029\_R2\_unpaired.fq.gz SLIDINGWINDOW:4:20 MINLEN:50  
java -jar /usr/local/trimmomatic/0.39/trimmomatic-0.39.jar PE -threads 20 -phred33 BWH--030\_R1.fastq BWH--030\_R2.fastq BWH--030\_R1\_paired.fq.gz BWH--030\_R1\_unpaired.fq.gz BWH--030\_R2\_paired.fq.gz BWH--030\_R2\_unpaired.fq.gz SLIDINGWINDOW:4:20 MINLEN:50

The next fastQC shows that the ends of the reads now present a much better quality:

#!/bin/bash  
module load fastqc/0.11.9  
fastqc \*paired.fq.gz -t 20 -o FASTQC\_afterTrim

## Genome download

I will use the reference genome GRCh38 from Ensembl as one of my future plans is to make a new transcript analysis and an alternative splicing analysis, though this is beyond the scope of this master thesis project. First the fasta file has to be downloaded chromosome by chromosome and store in a single file:

#!/bin/bash  
for chr in 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 MT X Y  
  
do  
echo Starting with chromosome ${chr}  
  
wget http://ftp.ensembl.org/pub/current\_fasta/homo\_sapiens/dna/Homo\_sapiens.GRCh38.dna.chromosome.${chr}.fa.gz  
cat Homo\_sapiens.GRCh38.dna.chromosome.${chr}.fa.gz >> Homo\_sapiens.GRCh38.dna.chromosome.all.fa.gz  
#Have a marker to check the process:  
echo Added the following bytes to general fasta:  
stat -c %s Homo\_sapiens.GRCh38.dna.chromosome.all.fa.gz  
  
rm Homo\_sapiens.GRCh38.dna.chromosome.${chr}\*  
echo Finishing with chromosome ${chr}  
done

Then the annotation file is downloaded:

#!/bin/bash  
wget http://ftp.ensembl.org/pub/release-107/gtf/homo\_sapiens/Homo\_sapiens.GRCh38.107.chr.gtf.gz  
gunzip Homo\_sapiens.GRCh38.107.chr.gtf.gz

## Read alignment

The next step is to align the reads to the reference genome:

#!/bin/bash  
module load hisat2/2.1.0  
module load samtools/1.9  
  
#Build index  
hisat2-build Homo\_sapiens.GRCh38.dna.chromosome.all.fa hg38  
  
#Align each sample  
for index in 00 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30  
  
do  
 echo Starting with sample ${index}  
 hisat2 -q --rna-strandness RF -k 1 -p 4 -x hg38 -1 trimmed\_paired\_fastq/BWH0${index}\_R1\_paired.fq -2 trimmed\_paired\_fastq/BWH0${index}\_R2\_paired.fq -S BWH0${index}.sam 2>> summary\_alignment.txt  
  
 echo Getting into samtools, sample ${index}  
 samtools view -@ 4 -bo BWH0${index}.sam BWH0${index}.bam  
 samtools sort -@ 4 -o BWH0${index}\_sorted.bam BWH0${index}.bam  
 samtools index BWH0${index}\_sorted.bam  
 echo Finishing with sample ${index}  
done

In the first paragraph of the loop the alignment as it is is performed, using hisat2. The second part of the loop uses samtools to transform the sam files into bam files (much smaller in size) and sorting and indexing of the reads by their coordinates.

The results of this process is very satisfactory:

grep overall summary\_alignment.txt  
98.14% overall alignment rate  
98.26% overall alignment rate  
98.29% overall alignment rate  
98.30% overall alignment rate  
97.86% overall alignment rate  
97.64% overall alignment rate  
97.67% overall alignment rate  
97.59% overall alignment rate  
96.79% overall alignment rate  
97.87% overall alignment rate  
97.81% overall alignment rate  
98.03% overall alignment rate  
98.17% overall alignment rate  
97.68% overall alignment rate  
97.39% overall alignment rate  
97.18% overall alignment rate  
96.08% overall alignment rate  
97.01% overall alignment rate  
96.83% overall alignment rate  
97.40% overall alignment rate  
97.50% overall alignment rate  
97.70% overall alignment rate  
96.89% overall alignment rate  
97.34% overall alignment rate  
97.73% overall alignment rate  
97.34% overall alignment rate  
96.92% overall alignment rate  
97.40% overall alignment rate  
97.42% overall alignment rate  
97.49% overall alignment rate  
97.33% overall alignment rate  
  
grep overall summary\_alignment.txt | wc -l  
31

## Gene quantification

The gene quantification was done using htseq-count, a Python package.

#!/bin/bash  
module load miniconda/3.7  
ls -l \*sorted.bam | sed 's/\_sorted.bam//g' > lista.txt  
  
cat lista.txt | while read index  
do  
 echo Starting with ${index}  
 htseq-count -f bam -r pos -m intersection-strict --stranded reverse   
 --minaqual 1 -t gene --idattr gene\_id ${index}\_sorted.bam   
 Homo\_sapiens.GRCh38.107.chr.gtf > ${index}.tsv  
  
 echo Finishing with ${index}  
done

Out of the arguments of this command arguably the most interesting is *-m intersection-strict*, which specifies the way the overlaps are managed. If the read is in the range of a coding exon but it is present outside its boundaries, it is interpreted as if this read has no gene match.

# Differential Gene Expression analysis

The DGE analysis aims to detect if there are any differentially expressed genes between the affected and unaffected individuals.

summary(cars)

## speed dist   
## Min. : 4.0 Min. : 2.00   
## 1st Qu.:12.0 1st Qu.: 26.00   
## Median :15.0 Median : 36.00   
## Mean :15.4 Mean : 42.98   
## 3rd Qu.:19.0 3rd Qu.: 56.00   
## Max. :25.0 Max. :120.00

## Including Plots

You can also embed plots, for example:



Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.