

PEC 1 - Anàlisi de Dades Òmiques

Vicent Caselles Ballester

2023-11-15

Contents

| | | |
|----------|------------------------------------|-----------|
| 1 | Introducció i objectius | 1 |
| 2 | Materials i mètodes | 3 |
| 2.1 | Presentació de l'informe | 3 |
| 2.2 | Dataset | 3 |
| 2.3 | Disponibilitat del codi | 3 |
| 2.4 | Comentari general | 3 |
| 3 | Resultats | 3 |
| 3.1 | Pregunta 2 | 3 |
| 3.2 | Pregunta 3 | 5 |
| 3.3 | Pregunta 4 | 7 |
| 4 | Discussió | 9 |
| 5 | Conclusions | 10 |
| 6 | Referències | 11 |
| 7 | Apèndix | 11 |

1 Introducció i objectius

L'objectiu d'aquesta PEC és consolidar els coneixements estudiats fins ara, referents a l'anàlisi de dades òmiques mitjançant eines informàtiques, concretament **Galaxy** (plataforma online) i el llenguatge de programació/*software* **R**, amb l'ús de Bioconductor.

En aquesta PEC, treballarem amb fitxers en format FASTQ. Aquest format neix a partir del format FASTA, que és un tipus especialitzat de fitxer de text per a la representació de seqüències de nucleòtids o aminoàcids.

Els fitxers FASTA solen començar amb un símbol > seguit d'una descripció, que sol correspondre a algún identificador per a la seqüència que hi ha a continuació (com per exemple el nom d'un gen, transcrit...). Després d'aquesta línia trobem la seqüència.

```
>NG_008679.1:5001-38170 Homo sapiens paired box 6 (PAX6)
ACCTCTTTTCTTATCATTGACATTTAACTCTGGGGCAGGTCCTCGCGTAGAACGCGGCTGTCAGATCT
GCCACTTCCCTGCGGAGCGCGGTGAGAAAGTGTGGGAACCGGCGCTGCCAGGCTCACCTGCCTCCCCGC
CCTCCGCTCCAGGTAACCGCCCGGCTCCGGCCCCGGCCCGGCTCGGGGCCCGCGGGCCTCTCCGCTG
CCAGCGACTGCTGTCCCCAAATCAAAGCCGCCCAAGTGGCCCCGGGCTTGATTTTGTCTTTTAAAG
GAGGCATACAAAGATGGAAGCGAGTTACTGAGGGAGGGATAGGAAGGGGGGTGGAGGAGGACTTGCTT
TGCCGAGTGTCTCTTCTGCAAAAGTAGCAAAATGTTCCACTCCTAAGAGTGGACTTCCAGTCCGCCCT
GAGCTGGGAGTAGGGGGCGGGAGTCTGCTGCTGCTGCTGCTAAAGCCACTCGCGACCGCGAAAAATGCA
GGAGGTGGGGACGCACTTTGCATCCAGACCTCCTCTGCATCGCAGTTCACGACATCCACGCTTGGGAAAG
TCCGTACCCGCGCCTGGAGCGCTTAAAGACACCCTGCCGCGGGTCGGGCGAGGTGCAGCAGAAGTTCCC
CGCGTTGCAAAGTGCAGATGGCTGGACCGCAACAAAGTCTAGAGATGGGGTTCGTTTCTCAGAAAGACGC
```

Figure 1: Exemple de format de fitxer FASTA. Font: <https://compgenomr.github.io/book/fasta-and-fastq-formats.html>

El format FASTQ neix a partir del format FASTA, i està dissenyat per a contenir, a més de la seqüència en si, mètriques de qualitat derivades de experiments de seqüenciació *Next Generation*.

```
Identifier — | @HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCAG/1
Sequence — | TTAATTGGTAAATAAATCTCCTAATAGCTTAGATNTTACCTTNNNNNNNNNTAGTTTCTTGAGA
+ sign & identifier — | +HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCAG/1
Quality scores — | efcffffcfeefffcfffffddfd`feed`]_Ba_^__[YBBBBBBBBBRTT\]] [] dddd`

                                     Base T
                                     phred Quality ] = 29
```

Figure 2: Exemple de format de fitxer FASTQ. Font: <https://compgenomr.github.io/book/fasta-and-fastq-formats.html>

És a dir, a més de la seqüència pròpiament dita, tenim informació sobre l'experiment, o els *reads*, que l'han generat. Aquest format utilitza quatre línies per seqüència. La primera, que comença per un símbol @ (anàlogament al > de FASTA), conté la descripció de la seqüència. Dintre de la informació que pot contenir aquesta descripció, trobem informació referent a la tecnologia que s'ha fet servir, com els números ID de les *flow cells* o les *lanes*.

La segona línia conté la seqüència de DNA/RNA/aminoàcids. La tercera línia conté un signe positiu (+) que marca el final de la seqüència, i la quarta línia conté els quality scores en format ASCII. Aquests són molt importants, ja que ens donen informació sobre la qualitat de l'experiment de seqüenciació, i permeten tenir una mesura de la certesa amb la qual s'ha assignat un nucleòtid a una posició del *read* determinada. Òbviament, els quality scores han de tenir la mateixa longitud que la seqüència. Generalment, s'utilitzen els anomenats *Phred quality scores*, definits per $Q_{phred} = -10 \log_{10} e$, on e és la probabilitat de que la base hagi sigut assignada a un nucleòtid erròniament (per tant, si tens Q – que el trobes amb el caràcter ASCII del *Quality Score* – pots trobar $p_{error} = 10^{-Q/10}$).

A continuació, deixo una llista amb la interpretació dels *quality scores* que he trobat a Internet. Cada *threshold* correspon a una probabilitat d'error diferent, amb puntuacions més grans corresponent a probabilitats d'error menors.

- 10 correspon a 10% error (1/10)
- 20 correspon a 1% error (1/100)
- 30 correspon a 0.1% error (1/1000)
- 40 correspon a 0.01% error (1/10000)

Com veurem a continuació (Figura 6), el nostre fitxer no presenta *quality scores* de mitjana inferior a 20, corresponent a una probabilitat d'error inferior a 0.01.

2 Materials i mètodes

2.1 Presentació de l'informe

Tots els gràfics han sigut confeccionats amb R+Bioconductor (software lliure + paquet especialitzat en anàlisis bioinformàtics) i Galaxy (plataforma online que permet realitzar complexos anàlisis bioinformàtics sense coneixement de programació, i sense haver-te de preocupar d'instal·lar *software*, ni per les versions d'aquest – només necessites un navegador web i accés a Internet), tal i com s'indica a l'enunciat de la PEC. Aquest informe ha estat generat amb Markdown i L^AT_EX.

2.2 Dataset

En quant a les dades, tal i com s'ha dit al fòrum del campus, corresponen a un experiment de RNA-Seq amb mostres de transcriptoma complet o RNA vesicular de *Pseudomonas aeruginosa*. L'experiment ha sigut realitzat en format *paired-end*.

Al primer apartat de resultats he fet una mica d'especulació informal sobre el dataset, prèviament a tenir aquesta informació. He intentat seguir un procediment mental lògic, però en certes ocasions he fallat.

2.3 Disponibilitat del codi

El codi utilitzat per a generar aquest informe es troba a <https://github.com/vcasellesb/PEC1-Analisis-Dat-Omic>.

2.4 Comentari general

La veritat és que el meu coneixement tècnic és bastant limitat. Tinc una idea general de tot, però no tinc coneixements especialment increïbles ni de genòmica, matemàtiques/estadística ni programació. El que sí que sé fer és utilitzar Internet per a buscar informació. Per això, és possible que una gran part de lo que hi hagi en aquest informe sigui extret d'Internet. Intentaré citar tot el possible.

3 Resultats

3.1 Pregunta 2

3.1.1 Comptatge de seqüències

La primera part de la pregunta 2, on demana la descripció del format FASTQ, ha sigut resposta a l'apartat d'introducció. A continuació mostro dues maneres d'averiguar el número de seqüències que es demana a l'enunciat, mitjançant R i també mitjançant el terminal d'Unix.

```
# font per a aquest codi: https://www.biostars.org/p/9487218/  
require(Biostrings, quietly = T)  
fq <- readDNAStringSet('data/S07_Ves02_read1.fastq',format='FASTQ')  
length(fq)
```

```
## [1] 250000
```

```
(base) vicentcaselles@Vicents-MacBook-Pro pec1 % cat data/S07_Ves02_read1.fastq |  
wc -l | awk '{print $1/4}'  
250000
```

Figure 3: Demostració obtenció de número reads amb Unix terminal – el resultat s'ha de dividir entre 4, ja que cada seqüència requereix 4 línies

Com veiem, els resultats coincideixen.

3.1.2 Informació sobre el tipus de seqüenciació que es va dur a terme

Mirant per Internet, he trobat que, pel format de la primera línia de cada seqüència, segurament aquest fitxer provingui d'un instrument de la marca *Illumina*. A continuació mostro el tipus de format al que crec que pertany el nostre fitxer.

With Casava 1.8 the format of the '@' line has changed:

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
```

| | |
|---------|--|
| EAS139 | the unique instrument name |
| 136 | the run id |
| FC706VJ | the flowcell id |
| 2 | flowcell lane |
| 2104 | tile number within the flowcell lane |
| 15343 | 'x'-coordinate of the cluster within the tile |
| 197393 | 'y'-coordinate of the cluster within the tile |
| 1 | the member of a pair, 1 or 2 (<i>paired-end or mate-pair reads only</i>) |
| Y | Y if the read is filtered (did not pass), N otherwise |
| 18 | 0 when none of the control bits are on, otherwise it is an even number |
| ATCACG | index sequence |

Figure 4: Format de descripció de seqüències pertanyent a màquines de la casa comercial Illumina (NGS). Font: Wikipedia

He intentat esbrinar a quina espècie pertany el fitxer, però no he sigut capaç. He intentat fer un BLAT amb diferents seqüències o *reads*, perquè pensava que els *reads* dels fitxers FASTQ eren contiguus, però he trobat que no és així.

He intentat fer BLATs amb diferents *reads* triats arbitràriament (i intentant que tinguessin un bon *Quality Score*), però cap m'ha donat resultats en els quals tingui confiança. Una sospita que tinc es *Saccharomyces cerevisiae*, però tinc poques proves que aportar.

Actualització: Després de buscar per internet, convertir el fitxer FASTQ a FASTA amb *Galaxy* i fer diversos BLATs amb les seqüències que n'han resultat, resulta que ens han dit l'espècie al fòrum del Campus Virtual. És *Pseudomonas aeruginosa*, un bacil gram-negatiu.

En quant al tipus d'experiment que ha generat les dades del meu fitxer, entenc que, com que diu **read1**, que es deu tractar d'una seqüenciació *paired-end*.

D'acord amb el que he trobat (referència 2), això ho podem comprovar de la següent manera. Si les meves dades provenen d'un experiment de seqüenciació *paired-end*, llavors el fitxer corresponent als reads 2 hauria de tenir el mateix nombre de seqüències. Anem a comprovar-ho:

```
require(Biostrings, quietly = T)
fq2 <- readDNAStringSet('data/S08_Ves02_read2.fastq',format='FASTQ')
length(fq2)
```

```
## [1] 250000
```

Com veiem, això es compleix. Ho he mirat amb el fitxer **S06_Ves01_read1.fastq** i també es compleix, així que no tinc clar fins a quin punt això té cap rellevància, però bé, el raonament té sentit.

A continuació mostro les ids de les dues primeres reads dels fitxers **S07_Ves02_read1.fastq** i **S08_Ves02_read2.fastq** que, si no m'equivoco, haurien de provenir del mateix experiment.

```
names(fq2[1])
```

```
## [1] "D00733:159:CA65UANXX:8:2210:1161:2419 2:N:0:AGTCAA"
```

```
names(fq[1])
```

```
## [1] "D00733:159:CA65UANXX:8:2210:1161:2419 1:N:0:AGTCAA"
```












Com veiem, els dos reads tenen el mateix id excepte que el id provenent del fitxer que acaba en `read2` presenta un 2 al caràcter que indica quin membre del *pair* és, fet que és molt coherent. A més, també podem observar que les coordenades del clúster es corresponen, reforçant la meua teoria de que aquests dos fitxers corresponen als dos reads d'un experiment paired-end.

Com podem veure, a més a més, els ids dels reads dels dos fitxers presenten el mateix índex (`AGTCAA`), fet que sembla indicar (amb el coneixement que tinc ara mateix) que es tracta d'un experiment de *Single Indexed Sequencing* (https://support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/miseq/indexed-sequencing-overview-guide-15057455-08.pdf).

3.2 Pregunta 3

3.2.1 Qualitat de la seqüenciació

A continuació, mostro els resultats principals de córrer l'eina FASTQC a *Galaxy*. Primer de tot, mostro les estadístiques bàsiques i els resultats dels *tests* que corre FASTQC.

| Summary | |
|--|--|
|  Basic Statistics | |
|  Per base sequence quality | |
|  Per tile sequence quality | |
|  Per sequence quality scores | |
|  Per base sequence content | |
|  Per sequence GC content | |
|  Per base N content | |
|  Sequence Length Distribution | |
|  Sequence Duplication Levels | |
|  Overrepresented sequences | |
|  Adapter Content | |

| Measure | Value |
|-----------------------------------|-------------------------|
| Filename | S07_Ves02_read1_fastq |
| File type | Conventional base calls |
| Encoding | Sanger / Illumina 1.9 |
| Total Sequences | 250000 |
| Total Bases | 12.5 Mbp |
| Sequences flagged as poor quality | 0 |
| Sequence length | 50 |
| %GC | 55 |

Figure 5: Summary i estadístiques bàsiques obtingudes en el report de FASTQC

Com podem observar, hi ha diversos *tests* que no supera l'experiment. Mostro els gràfics més importants, i em centro especialment en aquells que han fallat segons el *summary*. També, com podem veure, el número de seqüències total coincideix amb el calculat *programàticament*.

A la següent figura podem veure que les bases que, de mitjana, presenten una pitjor qualitat són les primeres de cada read, pero cap d'aquestes mostra una mitjana inferior a 30.

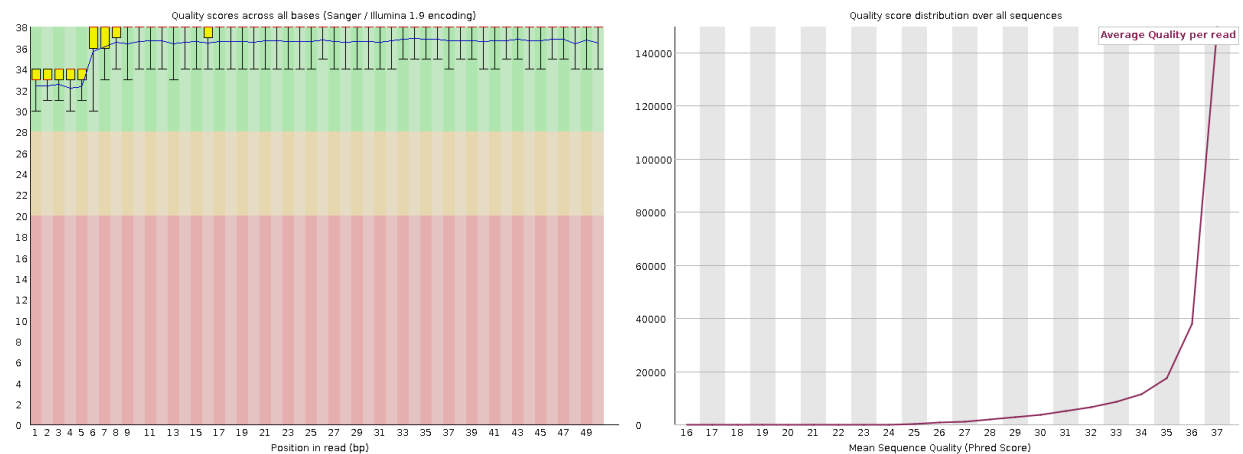


Figure 6: Quality scores per a cada posició del read

A continuació mostro dos gràfics referents al contingut de les bases, tant en percentatge total de les 4 bases com en percentatge de GC.

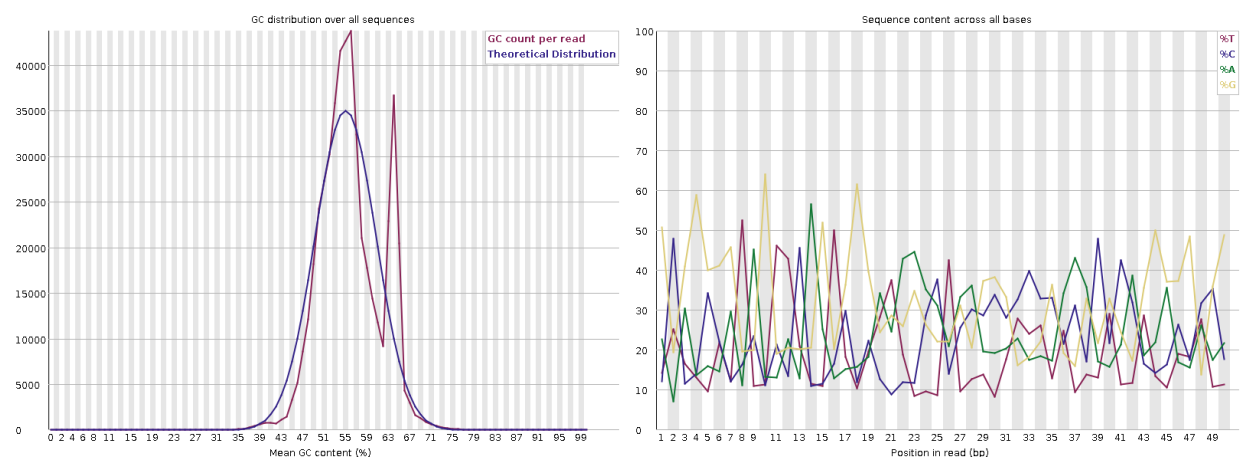


Figure 7: Contingut de les bases (% de nucleòtids) segons la seva posició

A continuació mostro un gràfic que mostra la quantitat de duplicació que hi ha hagut a l'experiment i el contingut d'adaptador que s'ha observat segons la posició dins del read.

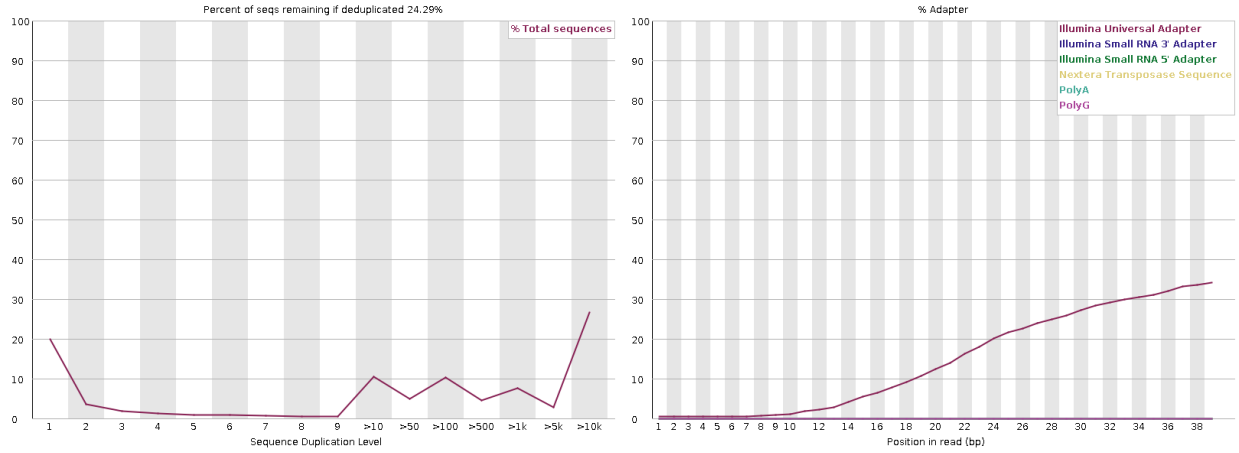


Figure 8: Nivells de dupliació de seqüències i percentatge de contingut d'adaptadors segons la posició

Finalment, mostro les seqüències més sobre-representades de l'experiment.

| Sequence | Count | Percentage | Possible Source |
|--|-------|--------------------|-----------------|
| GCAGCGGTAGTTCAGTCGGTTAGAATACCGGCTGTACGCGCGGGGGTGG | 31447 | 12.578800000000001 | No Hit |
| GCGGGCGTCGTATAATGGCATTACCTGAGCTTCCCAAGCTCATGACGAGG | 19672 | 7.868799999999999 | No Hit |
| GCGGGTATAGTTCAGTGGTAGAACCTCAGCCTTCCAAGCTGATGTCGG | 15965 | 6.386 | No Hit |
| GGGGCTATAGCTCAGCTGGGAGAGCGCTTGCATGGCATGCAAGAGGTCGA | 7283 | 2.9132000000000002 | No Hit |

Figure 9: Seqüències més sobre-representades, número de vegades que apareixen i percentatge que suposa del total de seqüències.

3.3 Pregunta 4

A continuació, vaig a dur a terme el *Quality Control* utilitzant el paquet *Rqc* de Bioconductor. Primer creo l'objecte que necessito, i després vaig creant els *plots* un a un. No mostro el codi.

```
require(Rqc, quietly=T)
qa <- rqcQA('data/S07_Ves02_read1.fastq', workers=1)

require(dplyr)
require(kableExtra)
knitr::kable(perFileInformation(qa), caption = "Estadístiques bàsiques de l'experiment") %>%
  kable_styling(latex_options = "HOLD_position")
```

Table 1: Estadístiques bàsiques de l'experiment

| filename | pair | format | group | reads | total.reads | path |
|-----------------------|------|--------|-------|--------|-------------|------|
| S07_Ves02_read1.fastq | 1 | FASTQ | None | 250000 | 250000 | data |

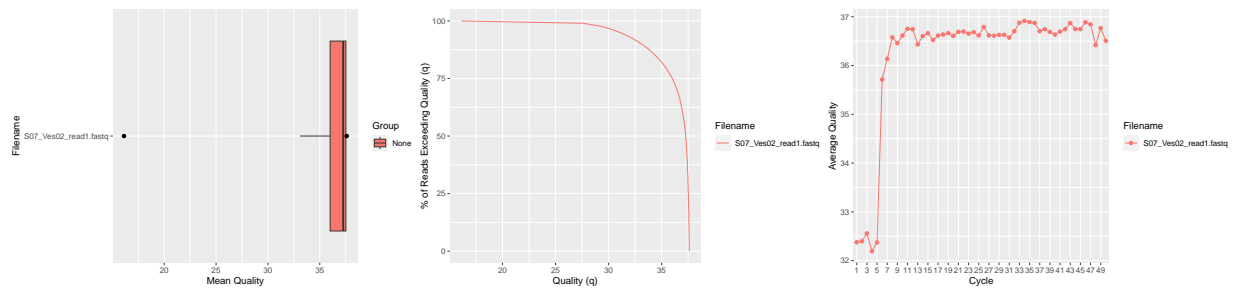


Figure 10: Gràfics referents a la qualitat dels base calls

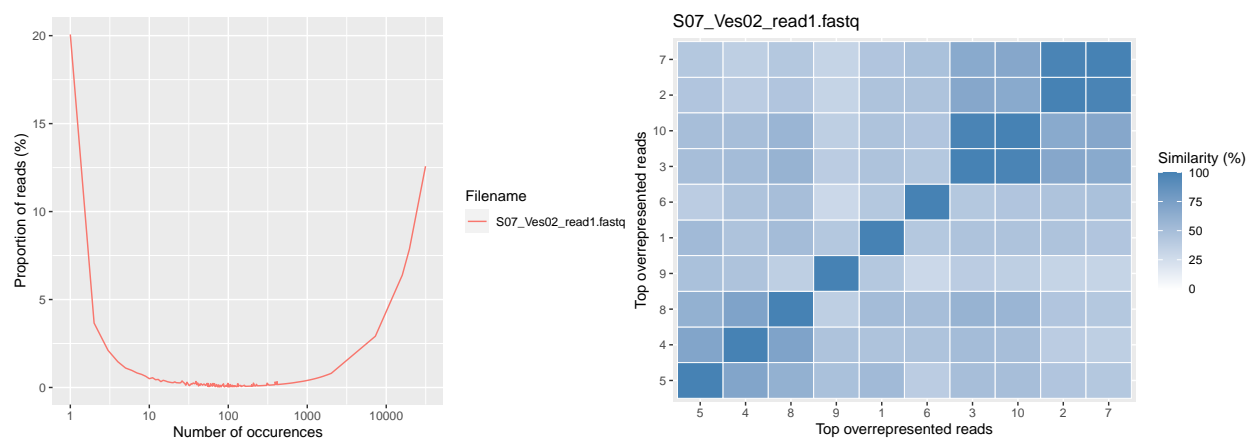


Figure 11: Sobrerepresentació dels reads i similaritat entre les seqüències més sobrerepresentades

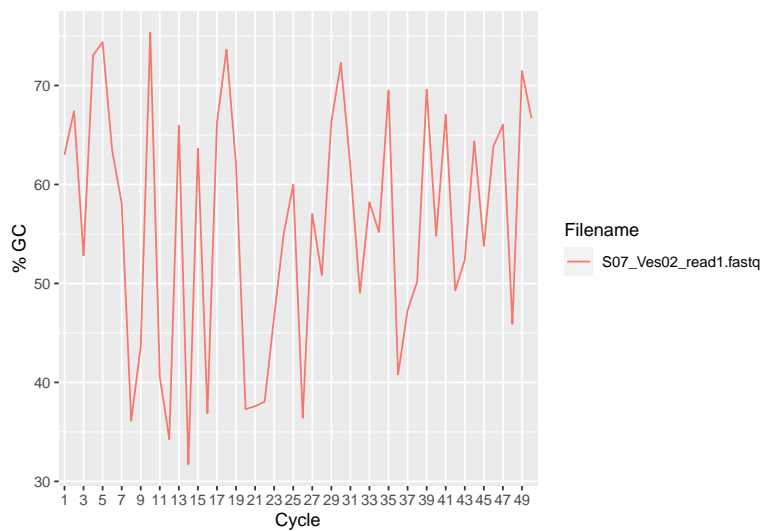


Figure 12: Percentatge de GC per posició

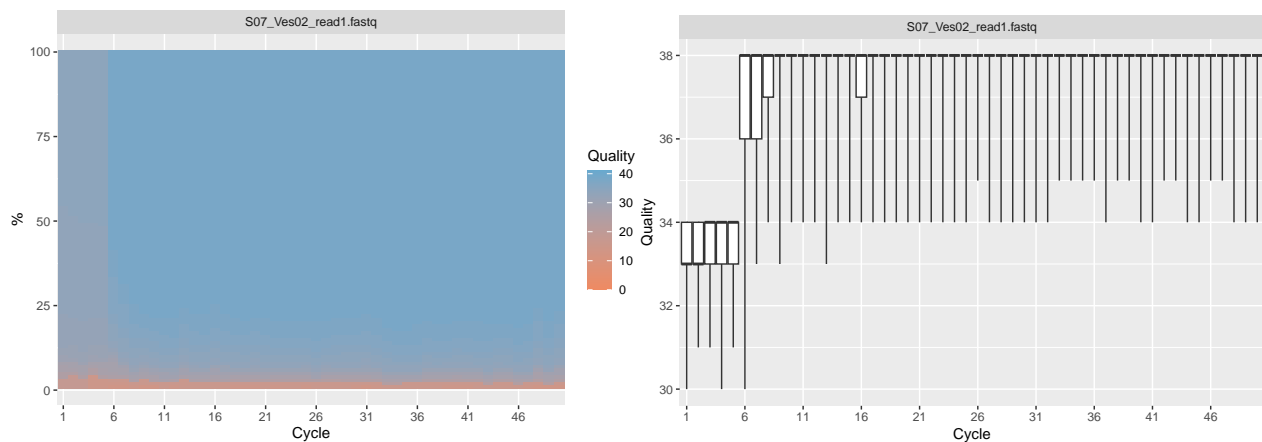


Figure 13: Quality plots per cycle

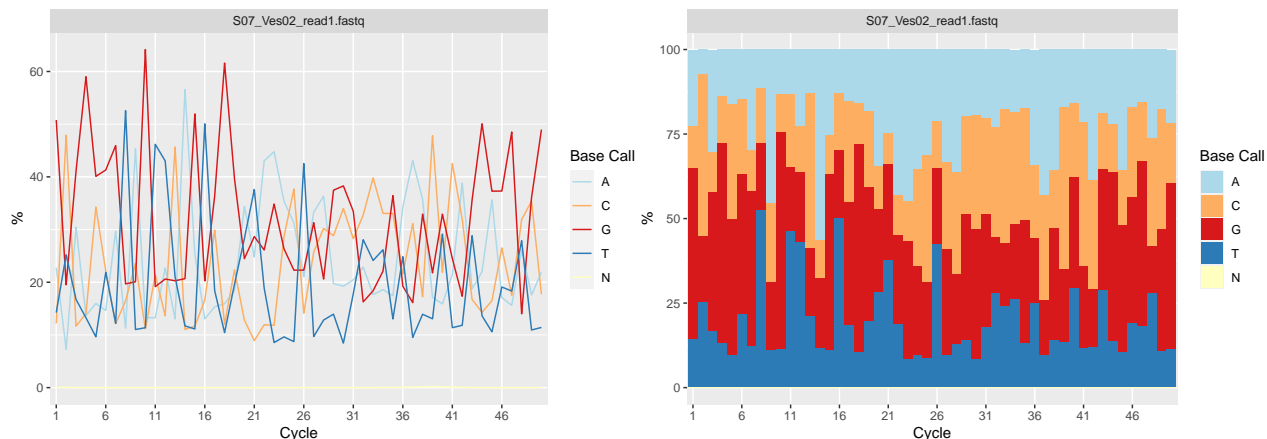


Figure 14: Bases assignades a cada posició dels reads

4 Discussió

Com podem veure, els resultats són molt semblants entre *Galaxy* i *R/Bioconductor*. En general, els *quality scores* són molt bons per a totes les bases o *cycles*. Un fet que m'he trobat a bastants tutorials/*lessons* a la web és que, generalment, la qualitat de les bases té una tendència a reduir-se cap al final del *read*, degut a problemes com el *signal decay*, on la senyal fluorescent va decaient amb el número de cicles del *read*; o com el *phasing* (referència 4). Això, en el nostre cas no es dona.

Per altra banda, una observació que em sembla preocupant és que sembla que hi ha un percentatge bastant alt de seqüències sobrerrepresentades, amb més de 10000 ocurrences en l'experiment/fitxer FASTQ, suposant això més d'un 10% de les seqüències. Veiem que aquest fet es pot observar tant als gràfics de *Galaxy* (fig. 9), com als gràfics de *Rqc* (fig. 11).

Això he estat buscant i es diu *enrichment bias*, i pot ser que realment aquests gens estiguin sobre-expressats biològica/natural -ment, o pot ser que hagin sigut generats artificialment, degut a l'amplificació selectiva de certes seqüències per sobre de les altres durant el pas de PCR.

A la figura 9 mostro les seqüències més sobrerrepresentades segons l'eina FASTQC de *Galaxy*. Aquí, podem veure que els *reads* que apareixen més vegades (o han sigut seqüenciats més vegades) no corresponen a cap *adapter sequence* reconegut per FASTQC. D'aquesta manera, o bé son transcrits que realment estan molt expressats a la mostra, o bé son productes d'amplificació selectiva.

Per a investigar això, podríem fer un BLAST amb aquestes seqüències, per a veure si corresponen a algún gen interessant. Un exemple de lloc on podríem fer aquesta cerca és la següent pàgina: <https://www.pseudomonas.com/blast/setnblast>. Ho he fet, però no he trobat resultats que, amb el meu coneixement actual, siguin mereixedors de mostrar-los en aquest informe, però crec que el procediment lògic seria aquest.

També, com podem veure a la figura 14 (esquerra) o a la figura 7 (que es correlacionen molt bé), sembla que hi ha una mica de *bias* per al nucleòtid G al principi i al final de cada *read*, encara que sembla que no és molt pronunciat. Cal destacar que he llegit en bastants fonts que, freqüentment, en els experiments de seqüenciació de RNA, s'observen diferències en la distribució del contingut nucleotídic al principi dels *reads*, degut a un fenomen que ocorre durant la fase de preparació de llibreria anomenat *priming with random hexamers* (referència 2).

Tot i això, en general, veiem un gràfic molt caòtic, fet que repercuteix en el fet de que l'experiment no passi el *test* corresponent. Hi ha molta variabilitat (el gràfic oscil·la molt a mesura que avances pel *read*), amb canvis grossos de contingut (aprox 30%). Això contrasta amb el que he observat generalment a la bibliografia (referències 2 i 5), on aquest gràfic s'estabilitzava a partir del cicle 15, aproximadament. Això no sé a què pot ser degut.

En quant al percentatge de GC, aquest és de 55% (fig. 5) de mitjana. Buscant a internet, he trobat que el percentatge de GC del genoma de *P. aeruginosa* oscil·la entre un 65 i un 67% (<https://doi.org/10.3389%2Ffmicb.2011.00150>). Això entenc que és algo a tenir en compte. Com podem veure a la distribució de GC (fig. 7), la distribució del nostre experiment dista considerablement de la distribució teòrica. Veiem un pic important a prop del 65%, fet curiós donat al fet que comento al principi del paràgraf. Podria ser degut al contingut específic del genoma de *P. aeruginosa*? Segons la bibliografia, aquests canvis en la distribució de GC solen ser deguts a contaminació o *bias* en la seqüenciació (podria –i segurament ho està– estar lligat a les seqüències sobre-representades). Segons la referència 7, la presència de “pics alternatius” pot voler dir diferents coses: pics pronunciats i estrets podrien ser deguts a la presència d'un contaminant específic (com adaptadors), mentre que pics més extensos indiquen contaminació d'una altra espècie.

Nosaltres tenim un major pic al voltant de 55%, i un pic menor al voltant de 65%. Podria voler dir això que hi ha contaminació? És realment el pic a 65% el que correspondria tenint en compte la espècie amb la que estem tractant? Trobo que, si és així, hi hauria hagut molta contaminació.

En quant al contingut d'adaptador (fig. 8), que mesura la presència d'aquests artefactes als nostres *reads*, segons he llegit (referència 6), és un fenomen bastant comú. En el nostre cas, veiem que aquest supera el 30% per al *Illumina Universal Adapter* en les últimes posicions dels *reads*. Això podria tenir-se en compte en els següents passos (fer un *trimming* de les *reads*)

Un comentari respecte al paquet *Rqc*, i és algo personal i que em sembla una arma de doble *filo* més que una crítica, és que abstrau moltíssim tot el procediment de generació dels gràfics de manera que, en el meu cas, preferiria que m'ensenyessin com generen els gràfics exactament, per així tenir opció de fer algunes modificacions. Per exemple, tots els gràfics apareixen amb una llegenda corresponent al nom dels fitxers que han entrat com a input a la funció *rqcQA*. En el nostre cas, només tenim un fitxer, i m'agradaria desfer-me de la llegenda que frustrantment apareix contínuament, precisament per a poder tenir més espai per els gràfics i així poder ajuntar-los en figures. Potser si que hi és la opció i no la he sabut trobar, però m'agrada tenir més control.

5 Conclusions

Com a conclusions, podem afirmar que les probabilitats d'error en el *base calling* són força baixes en general. Tot i això, s'hauria d'investigar a profunditat en els següents passos de l'anàlisi els motius pels que tenim seqüències tan sobre-representades, perquè hi ha tanta variabilitat en el % de nucleòtids d'acord amb la posició de la base, i el perquè de la desviació de la distribució del contingut de GC.

6 Referències

- <https://compgenomr.github.io/book/>
- https://bioinformatics.ccr.cancer.gov/docs/b4b/Module2_RNA_Sequencing/Lesson10/
- <https://doi.org/10.3389%2Ffmicb.2011.00150>
- <https://scienceparkstudygroup.github.io/rna-seq-lesson/03-qc-of-sequencing-results/index.html>
- https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon/lessons/qc_fastqc_assessment.html
- <https://gtpb.github.io/ELB18S/pages/L09>
- <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/5%20Per%20Sequence%20GC%20Content.html>

7 Apéndice

Per curiositat i com que al campus s'ha especificat que el genoma al que pertanyen els fitxers FASTQ és el del bacteri *Pseudomonas aeruginosa* (concretament la soca PAO1, disponible a: <https://www.pseudomonas.com/strain/download>), he volgut provar de fer el alineament amb aquest genoma. Per a fer-ho, he seguit el tutorial de la web següent: <https://scienceparkstudygroup.github.io/rna-seq-lesson/03-qc-of-sequencing-results/index.html>.

Bàsicament, el tutorial consisteix en alinear una sèrie de fitxers FASTQ directament a un genoma prèviament indexat amb l'eina **STAR** (<https://github.com/alexdobin/STAR>). Aquest tutorial utilitza **Docker**, que és una eina molt útil per a correr eines específiques en entorns aïllats (*containers*) de manera que és molt fàcil instal·lar paquets específics en qualsevol màquina sense preocupar-se de compatibilitat o espatllar la teua pròpia màquina.

A continuació mostro els resultats de dur a terme el meu *alignment*, que també l'he fet amb l'eina *STAR*. Qualsevol dubte sobre com he fet això, si-us-plau contacteu-me.

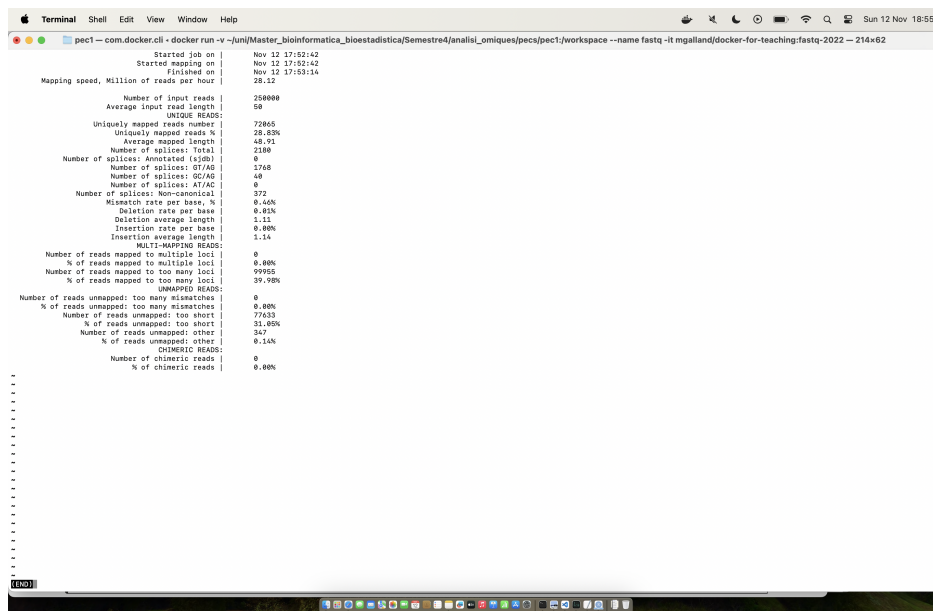


Figure 15: Resultat de l'alineament del fitxer FastQ amb el suposat genoma de l'espècie corresponent

Els resultats, sincerament, no els sé interpretar. Espero haver-ho après al acabar la assignatura.