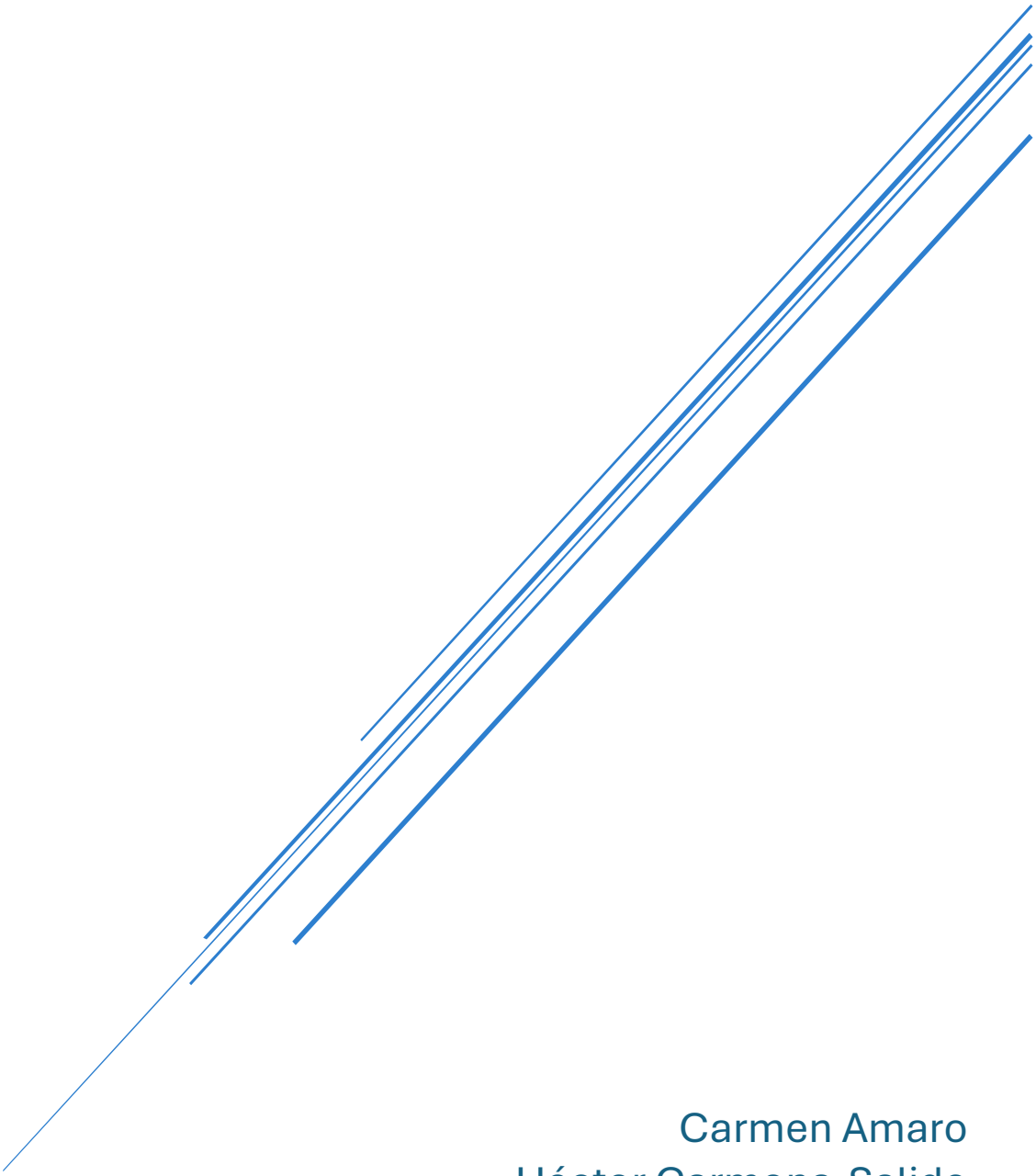
PRÀCTICA GALAXY:

Ens iniciem en la Bioinformàtica



Carmen Amaro
Héctor Carmona-Salido
Rubén Salvador-Clavell
Patogènesi Microbiana (BCB)

Contingut

Glossari

Anotació: procés pel qual se li adjudica informació biològica a les seqüències d'ADN.

Arxiu EMBL: anotació en format EMBL. Aquest format comença per un identificador ("ID") i se li poden afegir tantes línies d'anotació com calga. L'inici de la seqüència es marca amb un "SQ" i el final amb "//".

Arxiu FAA: seqüències aminoacídiques en format FASTA.

Arxiu FASTA: cadenes de caràcters (nucleòtids) que determinen les lectures de la seqüenciació. Estan compostos per dues línies: la primera (encapçalada pel símbol ">") és la informació de la seqüència o encapçalat; la segona és la cadena de caràcters com a tal.

Arxiu FASTQ: seqüències FASTA + detalls de qualitat de la informació (la "Q" és de *Quality*). Aquests arxius posseeixen més línies:

- Línia 1: encapçalat (en aquest cas comença per "@"), seguit per l'identificador de la lectura.
- Línia 2: la seqüència de nucleòtids.
- Línia 3: comença amb "+" (pot ser només aquest símbol o repetir la informació de l'encapçalat).
- Línia 4: informació de la qualitat de la seqüenciació de cada base. Aquesta qualitat es mesura amb una escala coneguda com a valor de qualitat Phred, que representa la probabilitat que una base siga incorrecta. Així, cada lletra o símbol representa la qualitat d'una base de la seqüència codificada en format ASCII.

La informació de qualitat es codifica en ASCII perquè això permet codificar un valor de qualitat en un sol caràcter. Per tant, les línies 2 i la 4 tenen el mateix nombre de caràcters. Aquests caràcters de qualitat ASCII, ordenats de menor a major qualitat, són:

!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~

Exemple d'arxiu FASTA (línia 1; línia 2):

```
>contig1 Vibrio vulnificus VvX
GTGTACTTCGTTCAGTTTATGACAGGTGTTGCTCGATTATCGCCTACCGTGACTTCGGATTCTATCGTG
```

Exemple d'arxiu FASTQ (línia 1; línia 2; línia 3; línia 4):

```
@1/1
GTGTACTTCGTTCAGTTTATCAGATGGGTGTTTAGCAGTTAATACACCTCACCGTGACTTCGGATTCTATCGTGT
+
&52429=;6787/.-01'&%&%$&$$,/3032*('&&(,)'&)*.%(%-()*(-89<>A8<9;<A?9976,&''-
```

Arxiu FFN: seqüències nucleotídiques FASTA dels diferents gens anotats.

Arxiu GBK o GBFF: anotació en format de GenBank. Aquest format conté un encapçalat amb informació sobre la seqüència anotada. Després, conté el nom de les diferents regions anotades, així com la posició. Finalment, conté la seqüència. La figura 1 mostra un exemple d'arxiu GBFF de la soca CECT 4999 de *Vibrio vulnificus*:

```
LOCUS       CP014636             3394464 bp    DNA     circular BCT 06-JUL-2017
DEFINITION  Vibrio vulnificus strain CECT 4999 chromosome I, complete sequence.
ACCESSION   CP014636
VERSION     CP014636.1
DBLINK      BioProject: PRJNA312018
            BioSample: SAMN04457466
KEYWORDS    .
SOURCE      Vibrio vulnificus
  ORGANISM  Vibrio vulnificus
            Bacteria; Pseudomonadota; Gammaproteobacteria; Vibrionales;
            Vibrionaceae; Vibrio.
REFERENCE   1  (bases 1 to 3394464)
  AUTHORS   Wu,K.-M. and Tsai,S.-F.
  TITLE     Direct Submission
  JOURNAL   Submitted (29-FEB-2016) Institute of Molecular and Genomic
            Medicine, National Health Research Institutes, 35, Keyan Road,
            Zhunan Town, Miaoli County 350, Taiwan
COMMENT     Annotation was added by the NCBI Prokaryotic Genome Annotation
            Pipeline (released 2013). Information about the Pipeline can be
            found here: https://www.ncbi.nlm.nih.gov/genome/annotation_prok/

            ##Genome-Assembly-Data-START##
            Assembly Date           :: 2011
            Assembly Method         :: Newbler v. 1.1.03.24
            Assembly Name           :: VVCECT49991.0
            Genome Representation   :: Full
            Expected Final Version  :: Yes
            Genome Coverage         :: 42.0x
            Sequencing Technology   :: 454
            ##Genome-Assembly-Data-END##

            ##Genome-Annotation-Data-START##
            Annotation Provider               :: NCBI
            Annotation Date                   :: 03/01/2016 18:43:11
            Annotation Pipeline               :: NCBI Prokaryotic Genome
                                                 Annotation Pipeline
            Annotation Method                 :: Best-placed reference protein
                                                 set; GeneMarkS+
            Annotation Software revision      :: 3.1
            Features Annotated                :: Gene; CDS; rRNA; tRNA; ncRNA;
                                                 repeat_region
            Genes (total)                     :: 4,633
            CDS (total)                       :: 4,485
            Genes (coding)                    :: 4,391
            CDS (coding)                      :: 4,391
            Genes (RNA)                       :: 148
            rRNAs                             :: 10, 9, 9 (5S, 16S, 23S)
            complete rRNAs                    :: 10, 9, 9 (5S, 16S, 23S)
            tRNAs                             :: 116
            ncRNAs                            :: 4
```

```
            Pseudo Genes (total)              :: 94
            Pseudo Genes (ambiguous residues) :: 0 of 94
            Pseudo Genes (frameshifted)       :: 39 of 94
            Pseudo Genes (incomplete)         :: 46 of 94
            Pseudo Genes (internal stop)      :: 16 of 94
            Pseudo Genes (multiple problems)  :: 5 of 94
            ##Genome-Annotation-Data-END##
FEATURES             Location/Qualifiers
     source          1..3394464
                     /organism="Vibrio vulnificus"
                     /mol_type="genomic DNA"
                     /strain="CECT 4999"
                     /serovar="E"
                     /host="Anguilla anguilla"
                     /culture_collection="CECT:4999"
                     /db_xref="taxon:672"
                     /chromosome="I"
                     /country="Spain"
                     /collection_date="1990"

     gene            complement(353..787)
                     /locus_tag="VVCECT4999_00005"
     CDS             complement(353..787)
                     /locus_tag="VVCECT4999_00005"
                     /inference="EXISTENCE: similar to AA
                     sequence:RefSeq:WP_017044116.1"
                     /note="An electron-transfer protein; flavodoxin binds one
                     FMN molecule, which serves as a redox-active prosthetic
                     group; Derived by automated computational analysis using
                     gene prediction method: Protein Homology."
                     /codon_start=1
                     /transl_table=11
                     /product="FMN-binding protein MioC"
                     /protein_id="ASJ37192.1"
                     /translation="MIHIITGSTLGGAEYVGDHLSDLLQEAGFETTIHNHPQLSEIPN
                     QGTWLIITSTHGAGEYPDNIQPFIHALQATPPKMSDVRYAVVAIGDSSYDTFCAAGLH
                     AHALLTDIGATAITQCLTIDVLSHTVPEDAAEVWLKQHIEQF"
     gene            complement(835..2223)
                     /gene="trmE"
                     /locus_tag="VVCECT4999_00010"
     CDS             complement(835..2223)
                     /gene="trmE"
                     /locus_tag="VVCECT4999_00010"
                     /inference="EXISTENCE: similar to AA
                     sequence:RefSeq:WP_011149046.1"
                     /note="in Escherichia coli this protein is involved in the
                     biosynthesis of the hypermodified nucleoside
                     5-methylaminomethyl-2-thiouridine, which is found in the
                     wobble position of some tRNAs and affects ribosomal
                     frameshifting; shows potassium-dependent dimerization and
                     GTP hydrolysis; also involved in regulation of
                     glutamate-dependent acid resistance and activation of
                     gadE; Derived by automated computational analysis using
                     gene prediction method: Protein Homology."
                     /codon_start=1
                     /transl_table=11
                     /product="tRNA modification GTPase"
                     /protein_id="ASJ37193.1"
                     /translation="MVVKTIGSFMTTDTIVAQATAPGRGGVGIIRVSGPQAAQVALEV
                     TGKTLKPRYAEYIPFKAQDGSELDQGIALFFPNPHSFTGEDVLELQGHGGPVVMDMLI
                     KRILTISGVRPARPGEFSERAFLNDKMDLTQAEAIADLIDASSEEAAKSALQSLQGQF
                     SKRIHTLVESLIHLRIYVEAAIDFPEEEIDFLADGKVAGDLQAIIDNLDAVRKEANQG
                     AIMREGMKVVIAGRPNAGKSSLLNALSGKDSAIVTDIAGTTRDVLREHIHIDGMPLHI
                     IDTAGLRDASDEVEKIGIERAWDEIRQADRVLFMVDGTTTDATDPKEIWPDFIDRLPE
                     QIGITVIRNKADQTQEQLGICHVSQPTLIRLSAKTGQGVDALRNHLKTCMGFSGNTEG
                     GFMARRRHLDALQRAAEHLLIGQEQLEGYMAGEILAEELRIAQQHLNEITGEFSSDDL
                     LGRIFSSFCIGK"
```

Figura 1. Exemple d'arxiu GBFF de la soca CECT4999 de V. vulnificus. *El quadre roig assenyala la zona on es troba l'anotació del gen* trmE. *Imatge pròpia.*

Enquadrat en color roig podem observar la informació associada a un dels gens anotats en el genoma. Aquest gen es troba en la cadena complementària, en les

posicions indicades entre parèntesis. En la seua part "CDS", tenim la informació relacionada amb la proteïna codificada.

Arxiu SAM: arxiu de text que arreplega la informació d'un mapatge de lectures d'un genoma. Consisteix en un encapçalat amb informació i una part principal en forma de taula. Cada fila conté informació d'una lectura i les columnes presenten el nom de la lectura, del cromosoma/*còntig* al qual s'assembla, posició en la qual es troba, qualitat del mapatge, si conté alguna mutació (SNP, inserció o deleció; aquest valor es diu CIGAR), etc. Per a obtindre-ne més informació, consulteu la figura 2.

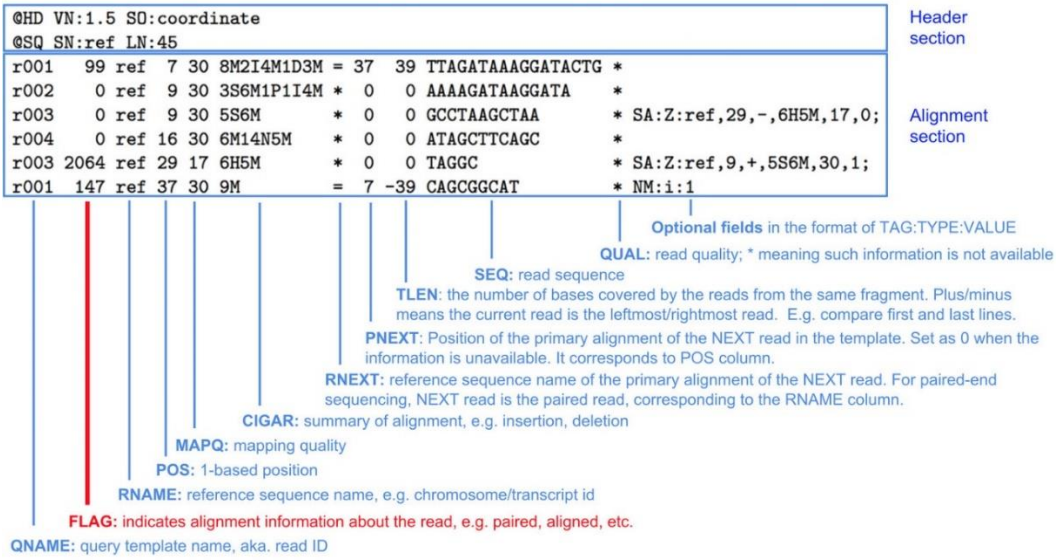


Figura 2. Exemple de format SAM. Es descriu cadascun dels camps que componen la taula del cos de l'arxiu. Extret de https://www.samformat.info/

Arxiu TSV: llistat a mode de taula dels gens anotats (TSV: versió separada per tabuladors).

CDS: regió codificant d'un gen que dona lloc a una proteïna. En el cas dels gens bacterians, es correspon amb el gen complet, ja que no contenen introns. L'expressió de la proteïna ha d'estar provada empíricament.

Dades crues, "*raw data*" o lectures crues: seqüències en format FASTQ que ens arriben després de tractar les dades de la plataforma de seqüenciació.

Assemblatge *de novo*: generació d'un genoma del qual no tenim referència prèvia.

Grafs de De Bruijn: estructura que permet representar superposicions entre cadenes de caràcters. Molts algorismes d'assemblatge es basen en aquests grafs per a reconstruir el genoma. En aquests casos, cada node es representa per una seqüència d'ADN de longitud variable (longitud k, anomenats k-mers) que es connecten si tenen seqüències iguals. Després, es busca el millor recorregut que unisca tots els nodes per a reconstruir la seqüència inicial.

AquestsAquests grafs després es poden visualitzar per a veure la qualitat de l'assemblatge. Un exemple el tenim en la figura 3.

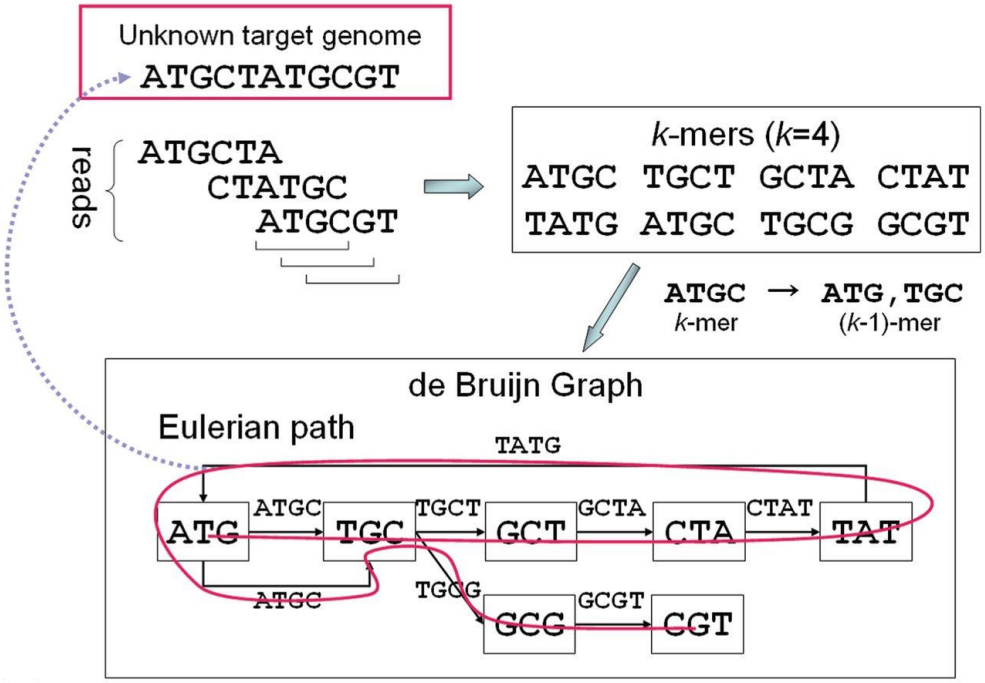


Figura 3. Grafs de De Bruijn. A partir de les lectures, es generen seqüències a l'atzar de grandària k (k-mers). Aquests k-mers formaran els nodes del graf que s'uniran si se superposen entre si. Finalment, es busca el millor camí que recórrega tots els nodes. Extret de [1].

L50: nombre de *còntigs* necessaris perquè en sumar les seues longituds done un resultat igual o superior al 50 % del total. Per a això, els *còntigs* s'ordenen de major a menor grandària. Una variació és l'L90, que seria per al 90 % de l'assemblatge. En la figura 4, atès que la grandària total són 5,32 Mpb, la meitat (aprox. 2,66 Mpb) s'aconsegueix amb 2 *còntigs*. Per tant, l'L50 és 2. En general, com més baix siga aquest valor, millor serà el nostre assemblatge.

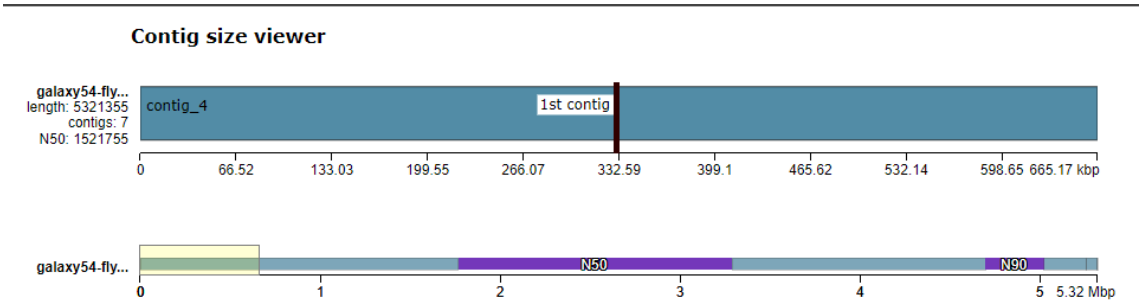


Figura 4. L50 i N50 d'un assemblatge obtingut mitjançant el programa QUAST. La part de dalt és una ampliació de la part en groc de baix. La figura inferior representa un assemblatge els còntigs del qual *han sigut ordenats de major a menor grandària. En morat, es representa l'N50 i l'N90. Imatge pròpia.*

Mapatge: procés que assigna a cada lectura una posició concreta en un assemblatge o genoma. En aquest cas, necessitarem una referència prèvia, generalment un assemblatge ja tancat.

N50: longitud que té el *còntig* de l'L50. En la figura 4, l' N50 seria la longitud del *còntig* número 2, que es correspon amb l' L50 i es destaca en morat. Igual que existeix l'L90, també existeix un N90 per al 90 % del genoma. Valors més alts d'aquesta estadística impliquen millor contigüitat.

ORF: prové de l'anglès *Open Reading Frame* (pauta de lectura oberta). És una regió que pot, putativament, contenir un gen. A diferència d'un CDS, no necessita haver sigut provat experimentalment. En el cas d'organismes eucariotes, inclou tant introns com exons.

Profunditat de seqüenciació (*depth of coverage*): nombre mitjà de vegades que se seqüencia cada base. No confondre amb rang o cobertura de seqüenciació (*breadth of coverage*) que representa el percentatge que es cobreix amb un mínim de lectures (figura 5).

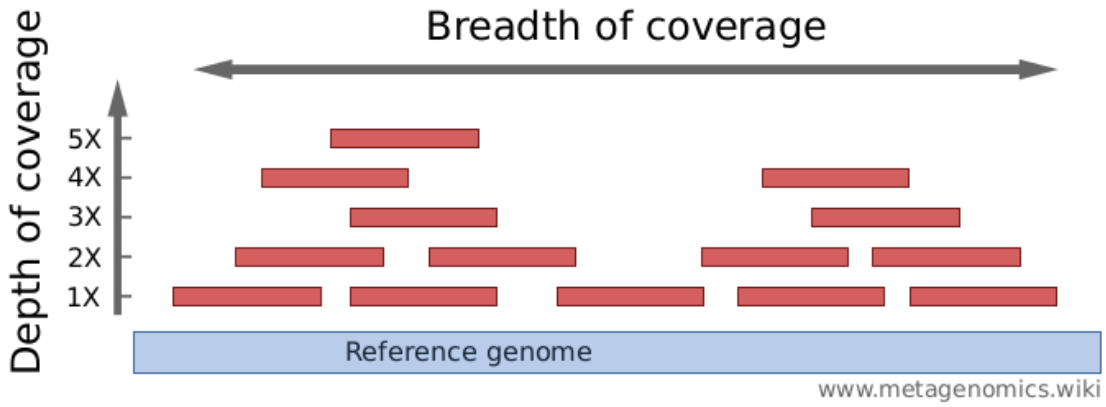


Figura 5. Representació esquemàtica de la profunditat de seqüenciació (Depth of coverage) i el rang de seqüenciació (Breadth of coverage). Extret de https://www.metagenomics.wiki/pdf/qc/coverage-read-depth.

Poliment del genoma: Com ja hem comentat, el fet d'usar lectures llargues implica afegir potencials errors d'assemblatge. Per això és útil combinar amb lectures curtes d'Illumina, que són més fiables i solen tenir més profunditat de seqüenciació. Malgrat això, en els assemblatges híbrids hi ha regions que únicament es generen a partir de lectures llargues i poden tenir algun error. Per a corregir això, se sol fer un "poliment" de l'assemblatge. Alguns assembladors com ara HGAP ja ho fan per defecte, uns altres, com ara Unicycler, no. Els programes com Pilon busquen discrepàncies entre les lectures curtes (o opcionalment les llargues) i l'assemblatge i les corregeix.

Introducció a la Bioinformàtica

Benvingudes, benvinguts i benvingutis al fantàstic món de la Bioinformàtica. En aquesta pràctica estudiarem unes pinzellades de manipulació de dades genòmiques d'una soca de *V. vulnificus*. En concret, a partir de les dades crues (lectures) provinents d'un seqüenciador, podrem visualitzar-los, analitzar-ne la qualitat i filtrar-los aplicant diferents paràmetres. A continuació, els assemblarem *de novo*, i n'obtindrem un arxiu (FASTA) amb el genoma bacterià. A partir d'aquest genoma podrem conèixer millor els gens que el componen (anotació). No obstant això, comencem pel principi.

La bioinformàtica és una ciència interdisciplinària que utilitza tècniques de diferents àrees com ara la computació o les matemàtiques (especialment l'estadística) per a l'anàlisi de dades biològiques. El terme *bioinformàtica* va ser encunyat inicialment per Hesper i Hogeweg el 1970 [2]. L'aparició, en els últims anys, de les tècniques de seqüenciació de l'ADN i, més recentment, de la intel·ligència artificial han suposat una gran revolució en aquesta àrea [3], [4].

Breu història de les *òmiques*

En un context biològic, el sufix *òmica* s'utilitza per a referir-se a l'estudi de grans conjunts de dades biològiques que representen tots els elements d'un tipus determinat, com per exemple *gens* (genòmica), *proteïnes* (proteòmica), entre altres. Hi ha termes que convé no confondre:

- Ciències òmiques: genòmica, transcriptòmica…
- Dades òmiques: lectures, matriu d'expressió…
- Tecnologies òmiques: seqüenciació d'ADN, *microarrays*…

Hi ha diferents òmiques en funció del camp de la biologia que tracta. En la figura 6, podem observar l'anomenada «cascada de les òmiques», la qual recull un resum dels diferents nivells d'aplicació d'aquestes tècniques per a obtindre una visió integral d'un organisme, sistema o mostra biològica. A més aquestes, en Microbiologia també s'utilitza el prefix *meta* quan estudiem comunitats microbianes complexes. Així, la metatranscriptòmica, per exemple, se centraria en l'estudi de tots els gens que s'estan transcrivint en una mostra concreta.

Per a aquesta sèrie de pràctiques, ens quedarem en la primera de les òmiques, la Genòmica. Aquesta ciència és molt útil en l'estudi de microorganismes i, més específicament, de bacteris patògens, com és el nostre cas [3].

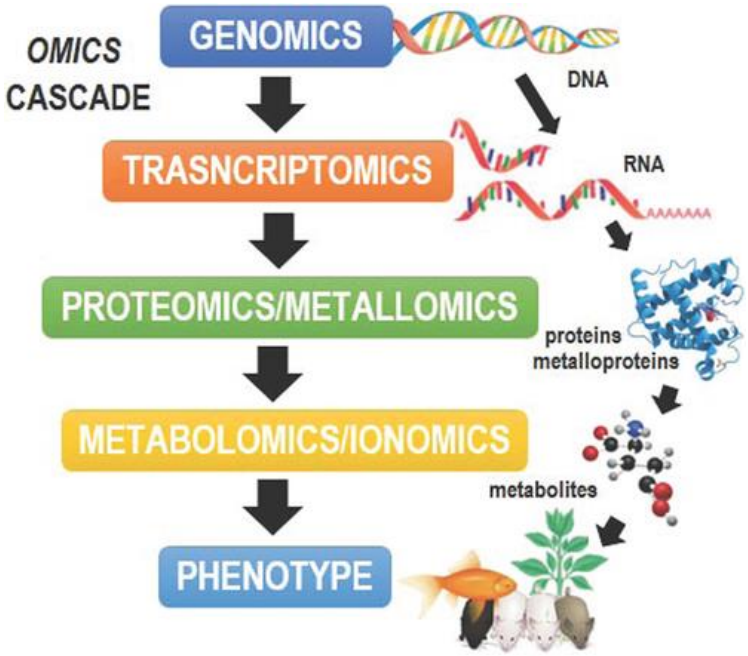


Figura 6. La cascada òmica. Extret de [5].

Genòmica i tècniques de seqüenciació

La genòmica és l'estudi d'un conjunt complet d'ADN (amb tots els seus gens) de qualsevol organisme, centrant-se en la funció, estructura, evolució, mapatge i edició. El genoma conté tota la informació necessària per al desenvolupament i el creixement d'aquest organisme. L'estudi del genoma ha suposat una revolució sense precedents, i ha ajudat el món de la recerca a entendre la interacció dels gens entre si i amb l'entorn, així com la manera en què sorgeixen certes malalties [6].

Ara bé, com arribem a obtindre-ne resultats? Per a això, necessitem la tècnica de la seqüenciació d'ADN, la qual ens determinarà l'ordre dels nucleòtids (els quatre components bàsics químics, anomenats *bases*) d'un fragment d'ADN donat. Des del primer genoma seqüenciat el 1976, el bacteriòfag MS2, passant pel primer genoma procariòtic (*Haemophilus influenziae*) el 1995, fins a la finalització del Projecte del Genoma Humà, les millores tecnològiques i l'automatització han augmentat la velocitat i reduït els costos fins al punt en què gens individuals poden seqüenciar-se de manera rutinària. Així, en alguns laboratoris es poden seqüenciar més de 100 bilions de bases a l'any, i, actualment, un genoma complet pot seqüenciar-se per tan sols uns quants centenars de dòlars/euros [7].

La figura 7 resumeix l'aparició i evolució de les tècniques de seqüenciació d'ADN. La primera tècnica de seqüenciació d'ADN va ser desenvolupada per Frederick Sanger, amb el mètode de terminació de cadena. Quasi al mateix temps, Maxam i Gilbert van desenvolupar una tècnica que, juntament amb la de Sanger, són

considerades les seqüenciacions de primera generació. Encara que avui en dia puguen semblar unes tècniques simples, van ser tota una revolució. De fet, va anar gràcies a aquestes tècniques que es va poder publicar el primer esborrany del genoma humà el 2001 [8], [9]. La tècnica de seqüenciació per Sanger s'ha anat optimitzant i automatitzant i encara s'utilitza avui , a causa del seu cost baix, la seua rapidesa i la seua fiabilitat.

Els esforços per a obtindre cada vegada més genomes en menys temps van donar lloc a una nova generació de tècniques de seqüenciació, conegudes com a segona generació, nova generació (*Next-Generation*) o tècniques d'alt rendiment (*High-Throughput*). Les tècniques *Next-Generation Sequencing* (o NGS) són un conjunt de noves tecnologies utilitzades per a la seqüenciació d'ADN i ARN i la detecció de variants/mutacions. Amb les NGS es poden seqüenciar centenars i milers de gens o tot el genoma en poc de temps. Les variants/mutacions de seqüència detectades per NGSs s'han utilitzat àmpliament per al diagnòstic de malalties, el pronòstic, la decisió terapèutica i el seguiment de pacients. La capacitat de la seua seqüenciació paral·lela massiva ofereix noves oportunitats per a la medicina de precisió personalitzada.

Com a seqüenciadors de segona generació tenim el sistema *454 Life Sciences* (piroseqüenciació) o *Ion Torrent*, en desús actualment. A gran escala, van aparèixer els primers seqüenciadors d'Illumina. No obstant això, aquests últims s'han anat actualitzant i millorant en gran manera, i formen part de les rutines de molts laboratoris de diagnòstic.

Juntament amb la versió actualitzada d'Illumina, també trobem una nova generació de tècniques de seqüenciació. Aquestes tècniques són conegudes per generar menys lectures, però de molta més longitud, la qual cosa les ha dotades del nom *molècula llarga*, *molècula única* o t*ercera generació*. Dins d'aquesta categoria, actualment es troba la tecnologia SMRT (*Single Molecule Real Time*) de Pacific Biosciences (PacBio) o la seqüenciació per nanopors d'Oxford Nanopore Technologies (ONT). La principal diferència entre Illumina i les altres dues tècniques de seqüenciació és que, mentre que la primera dona lloc a moltes lectures curtes (d'aproximadament 300 bases), les segones ofereixen menys lectures però més llargues (unes 15k bases de mitjana) [10].

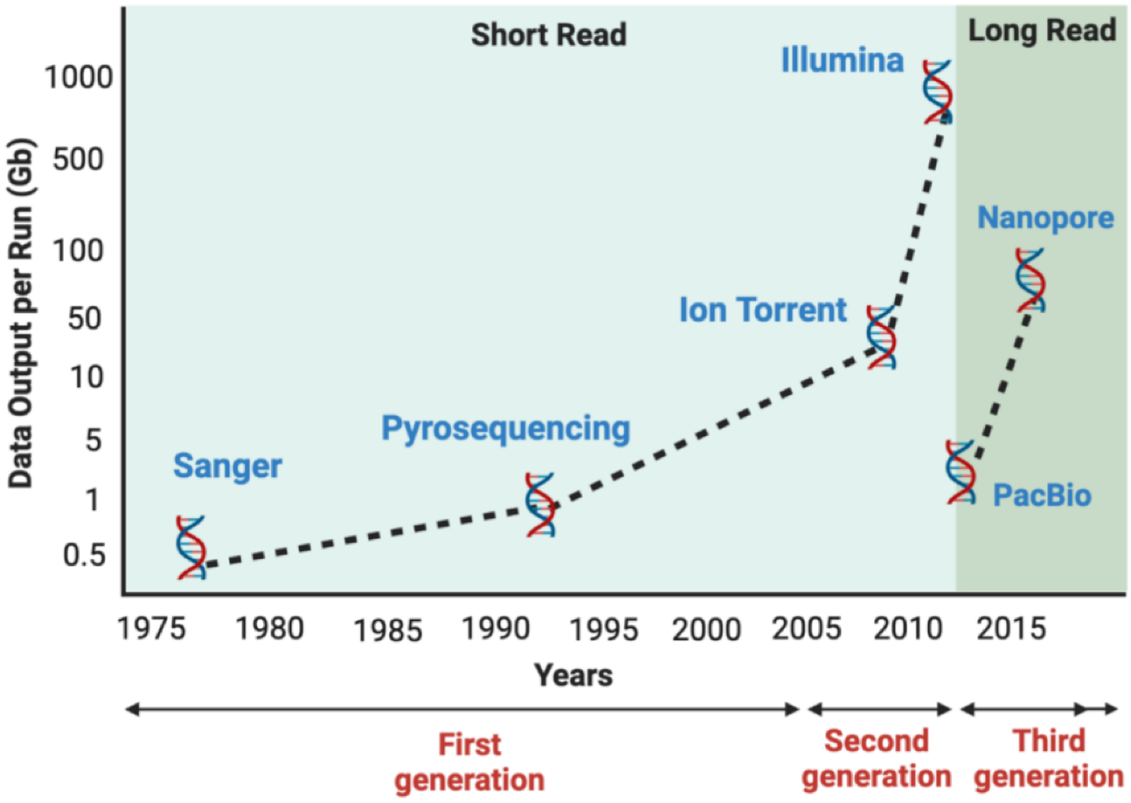


Figura 7. Evolució de les principals tecnologies de seqüenciació. Actualment, Illumina és la tecnologia que més quantitat de dades pot generar per cada tanda de seqüenciació. Extret de [11].

En definitiva, moltes d'aquestes noves tecnologies es van desenvolupar amb el suport del Programa de Tecnologia Genòmica de l'Institut Nacional d'Investigació del Genoma Humà (*National Human Genome Research Institute*, NHGRI) i les seues concessions per a Tecnologia de Seqüenciació d'ADN Avançada. Un dels objectius de l'NHGRI és promoure noves tecnologies que a la llarga puguen reduir el cost de la seqüenciació d'un genoma humà a menys de 1000 $ i ser de fins i tot una qualitat superior a la possible avui.

Objectiu de les pràctiques

Una vegada ens trobem en context, explicarem el desenvolupament de les pròximes sessions. Per a aquestes pràctiques, treballarem amb lectures curtes d'Illumina MiSeq i llargues d'Oxford Nanopore MinION de la soca Vv5 de *V. vulnificus*. Aquestes dades es troben en un repositori públic al web del *National Center for Biotechnology Information* (NCBI) sota el codi GenBank: JADIXT000000000.2. L'objectiu d'aquestes pràctiques serà aprendre a fer servir les eines necessàries per a treballar a partir de resultats de seqüenciació, des de l'anàlisi de les lectures fins a l'assemblatge i l'anotació del genoma final. Per a això, usarem l'entorn web gratuït Galaxy (https://usegalaxy.eu/), el qual conté totes les eines que emprarem.

Esquema de treball o *workflow*

1. Creació d'un compte gratuït i aprenentatge de l'entorn Galaxy.

2. Obtenció de les dades crues de treball de l'NCBI i primers passos en Galaxy: importació dels arxius amb l'eina *Faster Download and Extract Reads in FASTQ* i visualització de les dades (lectures).

3. Control de qualitat de les dades amb *FastQC* (per a lectures curtes d'Illumina) o "NanoPlot" (per a lectures llargues de Nanopore). La figura 8 mostra un exemple de gràfic que es pot obtindre:

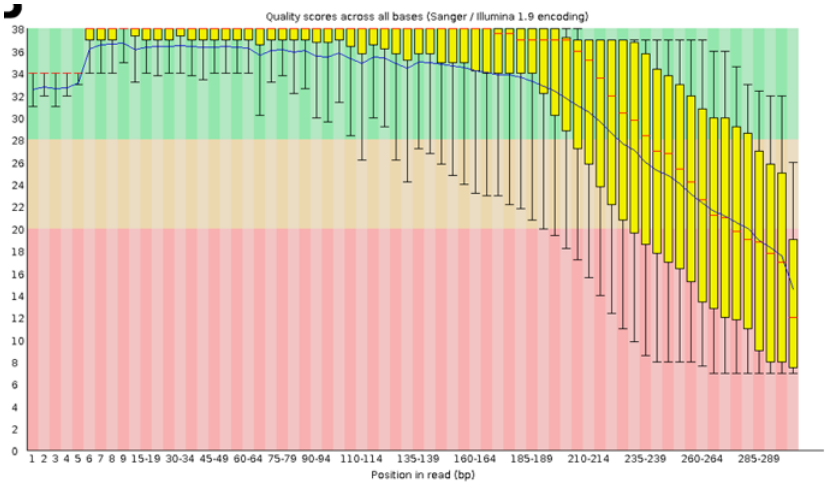


Figura 8. Exemple d'un gràfic obtingut amb FastQC de la qualitat de lectures. La línia blava marca la qualitat global per posició de la base en les lectures. S'observa que, sobretot en posicions finals de les lectures, la qualitat descendeix. Imatge pròpia.

Si després de la visualització de la qualitat global de les lectures es considera que es pot millorar, es procedeix al filtrat amb l'eina *PRINSEQ*, amb l'anàlisi posterior de la nova qualitat (figura 9).

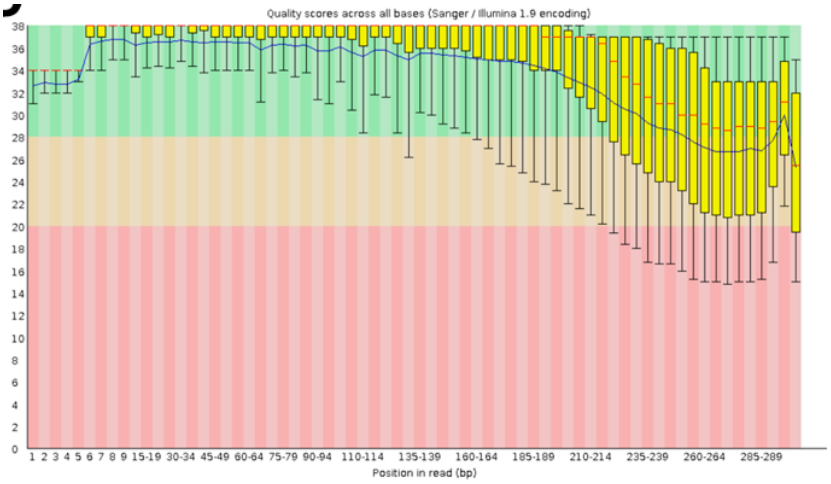


Figura 9. Exemple d'un gràfic obtingut amb FastQC de la qualitat de lectures després del filtrat per qualitat mitjançant PRINSEQ. Observem com les últimes posicions de les lectures, on abans hi havia una qualitat molt dolenta, tenim una millora notable. Imatge pròpia.

4. Assemblatge *de novo* mitjançant tres aproximacions diferents:
   - SPAdes: assemblatge amb les lectures curtes d'Illumina.
   - Flye: assemblatge amb les lectures llargues de MinION.
   - Unicycler: assemblatge híbrid, combinant-ne les lectures curtes i les llargues.

   Anàlisi dels assemblatges utilitzant els arxius GRAFO de De Bruijn amb l'eina *Bandage Info*, i visualització dels còntigs amb *Bandage Image* Tot seguit, avaluació dels estadístics amb *QUAST* i *BUSCO*. A partir d'aquí, decidirem quina aproximació d'assemblatge és millor.
   La figura 10 mostra un exemple d'un genoma de *V. vulnificus* tancat, on visualitzem dos *còntigs*, els quals es corresponen amb els dos cromosomes del bacteri.

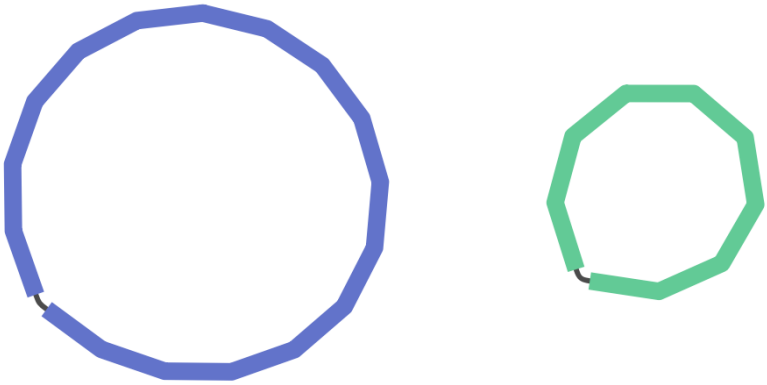


Figura 30. Imatge de l'assemblatge d'un genoma tancat i circularitzat de V. vulnificus*. S'hi observen només dos* còntigs*, els quals es corresponen amb els dos cromosomes d'aquest bacteri. A més, veiem que els cromosomes estan circularitzats, això és, estan complets de tal manera que el programa de treball entén que el final de la seqüència de cada* còntig *es correspon amb l'inici d'aquest* còntig*. La diferència de grandària indica, a més, que el cromosoma 1 (esquerra) és major que el cromosoma 2 (dreta). Imatge pròpia.*

5. Poliment de l'assemblatge amb l'eina *Pilon*.

6. Anàlisi de la Identitat Nucleotídica Mitjana (ANI) del genoma obtingut amb un altre genoma de referència de *V. vulnificus*.

7. Anotació del genoma amb *Bakta*.

Alternativament, en lloc de fer un assemblatge *de novo*, es pot fer un mapatge de les lectures davant d'un genoma de referència. Per a això, es poden usar diverses estratègies, com ara la que ofereix l'eina *Snippy*.

Finalment, amb l'arxiu FASTA definitiu del genoma assemblat podem jugar lliurement i utilitzar-lo per als nostres estudis, com a filogènia, recombinació, busca de pròfags, etc.

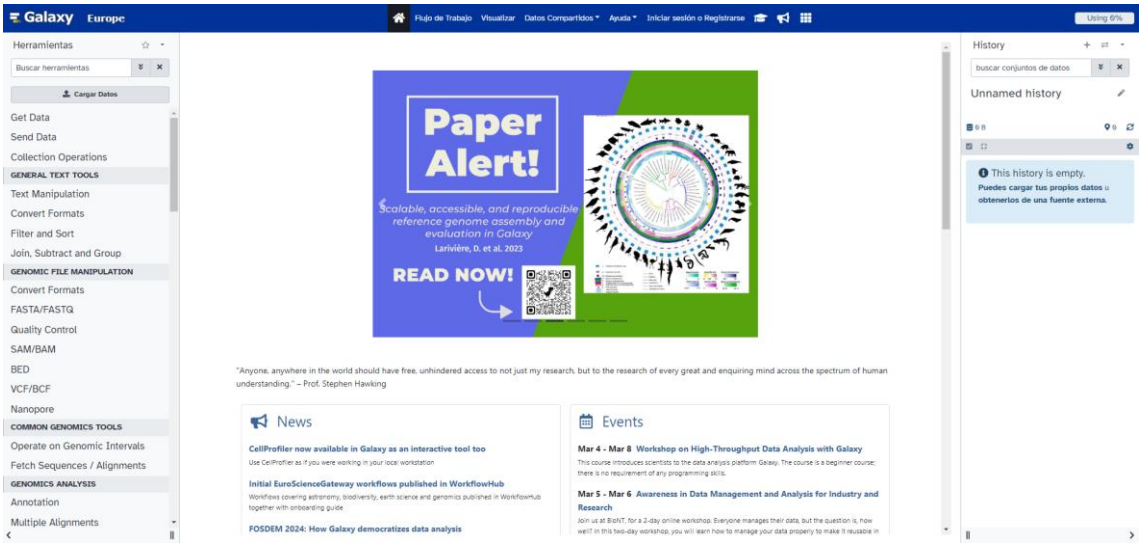
Pràctica 1: Ens endinsem en Galaxy

Treballar en Bioinformàtica implica, en molts casos, conèixer llenguatges de programació, com ara R o Python, i fins i tot treballar amb el sistema operatiu Linux. No obstant això, per a facilitar-nos el treball (almenys d'aquesta pràctica) existeix l'entorn de treball en línia Galaxy. En aquest entorn podem disposar de les diferents eines bioinformàtiques de les quals hem parlat anteriorment, sense necessitat d'instal·lar-les en el nostre ordinador. A més, treballar en remot, implica que no necessitarem un ordinador potent, ja que utilitzarem els recursos informàtics del propi servidor de l'entorn Galaxy.

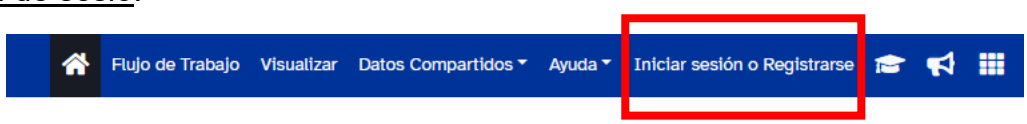
En definitiva, aquesta pràctica consisteix en un breu tutorial de Galaxy.

Seccions i eines de Galaxy

Enllaç al web: https://usegalaxy.eu/

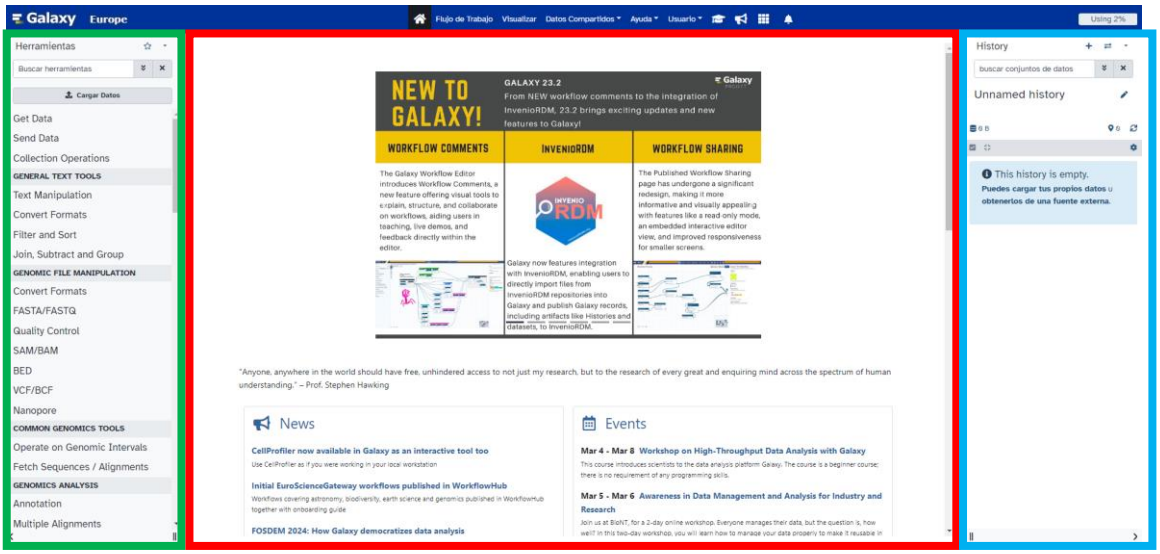


Inici de sesió:



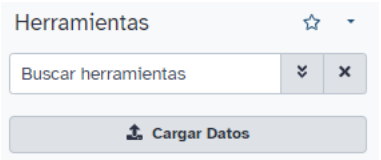
Hem de registrar-nos per a poder iniciar-hi sessió. No importa si empreu un correu institucional o el vostre correu electrònic personal.

Interfície una vegada hem iniciat sessió:



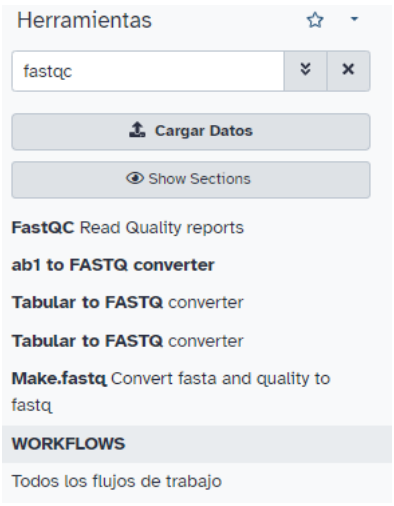
Eines de treball | Entorn de treball | Historial de treball

Eines de treball



En la part superior d'aquesta secció observem que podem buscar les eines bioinformàtiques que necessitem per a la realització del nostre treball (si les coneixem). A més a més, trobem el botó per a la càrrega de les dades. Aquí serà on, més tard, carregarem les dades crues a la plataforma per a treballar-hi.

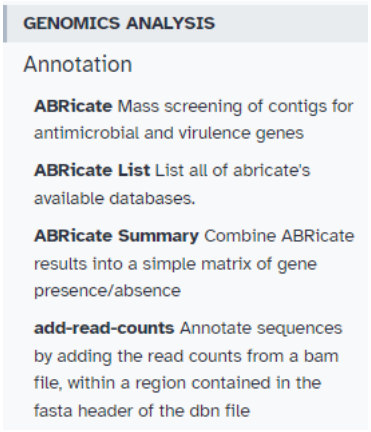
Cerquem, per exemple, l'eina FastQC:



Ens mostra les diferents opcions que contemplen la paraula emprada.

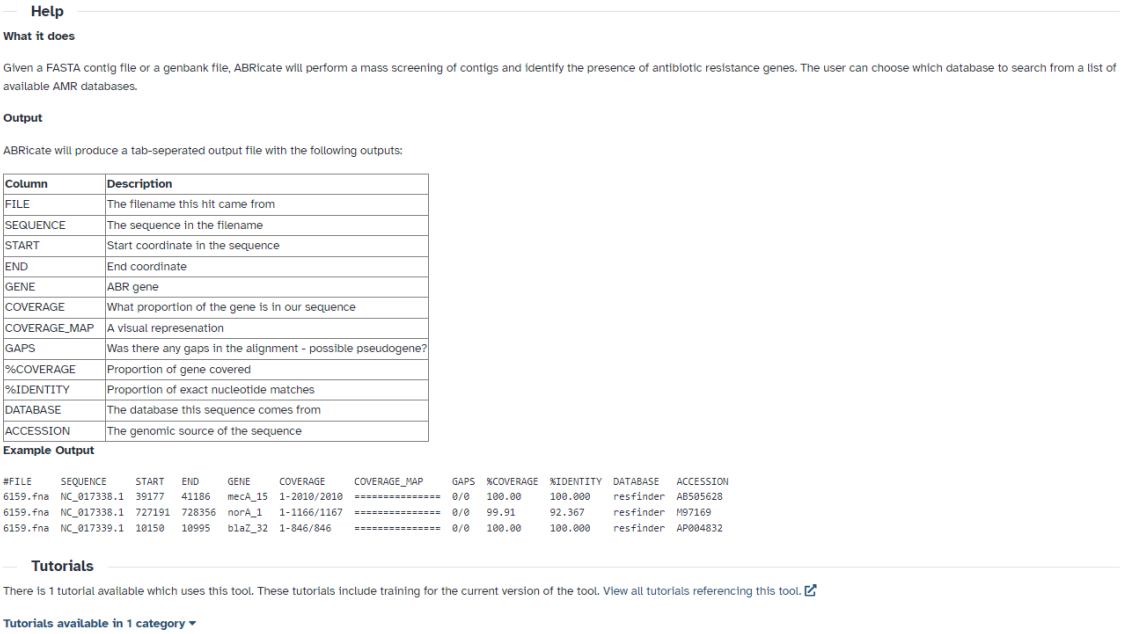
Tornant a la secció d'Eines, si ens desplacem amb la barra podem trobar agrupades les diferents accions bioinformàtiques per grups de treball: *General Text Tools*, *Genomic File Manipulation*, *Genomics Analysis*, etc.

Si accedim a la funció *Annotation* de *Genomics Analysis*, trobem les eines bioinformàtiques que serveixen per a anotar els genomes (ordenades alfabèticament).



Si accedim a la primera eina, ABRicate, a l'entorn de treball de la part central se'ns en mostra el funcionament:

Juntament amb informació important de l'eina: què fa, quina eixida o *output* ofereix (exemple), entre altres.



Help

What it does

Given a FASTA contig file or a genbank file, ABRicate will perform a mass screening of contigs and identify the presence of antibiotic resistance genes. The user can choose which database to search from a list of available AMR databases.

Output

ABRicate will produce a tab-seperated output file with the following outputs:

Column	Description
FILE	The filename this hit came from
SEQUENCE	The sequence in the filename
START	Start coordinate in the sequence
END	End coordinate
GENE	ABR gene
COVERAGE	What proportion of the gene is in our sequence
COVERAGE_MAP	A visual represenation
GAPS	Was there any gaps in the alignment - possible pseudogene?
%COVERAGE	Proportion of gene covered
%IDENTITY	Proportion of exact nucleotide matches
DATABASE	The database this sequence comes from
ACCESSION	The genomic source of the sequence

Example Output

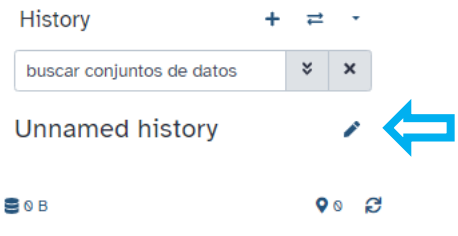
```
#FILE     SEQUENCE     START   END     GENE     COVERAGE     COVERAGE_MAP     GAPS  %COVERAGE  %IDENTITY  DATABASE   ACCESSION
6159.fna  NC_017338.1  39177   41186   mecA_15  1-2010/2010  ===============  0/0   100.00     100.000    resfinder  AB505628
6159.fna  NC_017338.1  727191  728356  norA_1   1-1166/1167  ===============  0/0   99.91      92.367     resfinder  M97169
6159.fna  NC_017339.1  10150   10995   blaZ_32  1-846/846    ===============  0/0   100.00     100.000    resfinder  AP004832
```

Tutorials

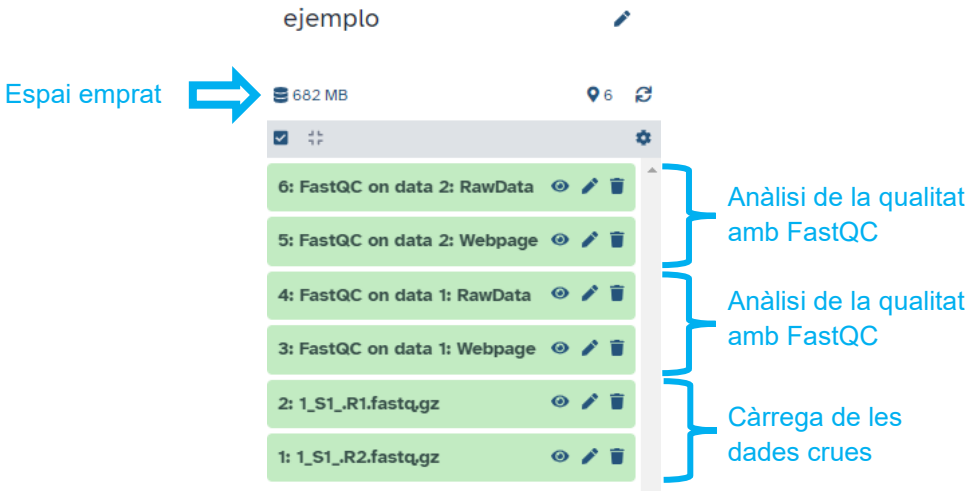
There is 1 tutorial available which uses this tool. These tutorials include training for the current version of the tool. View all tutorials referencing this tool.

Tutorials available in 1 category

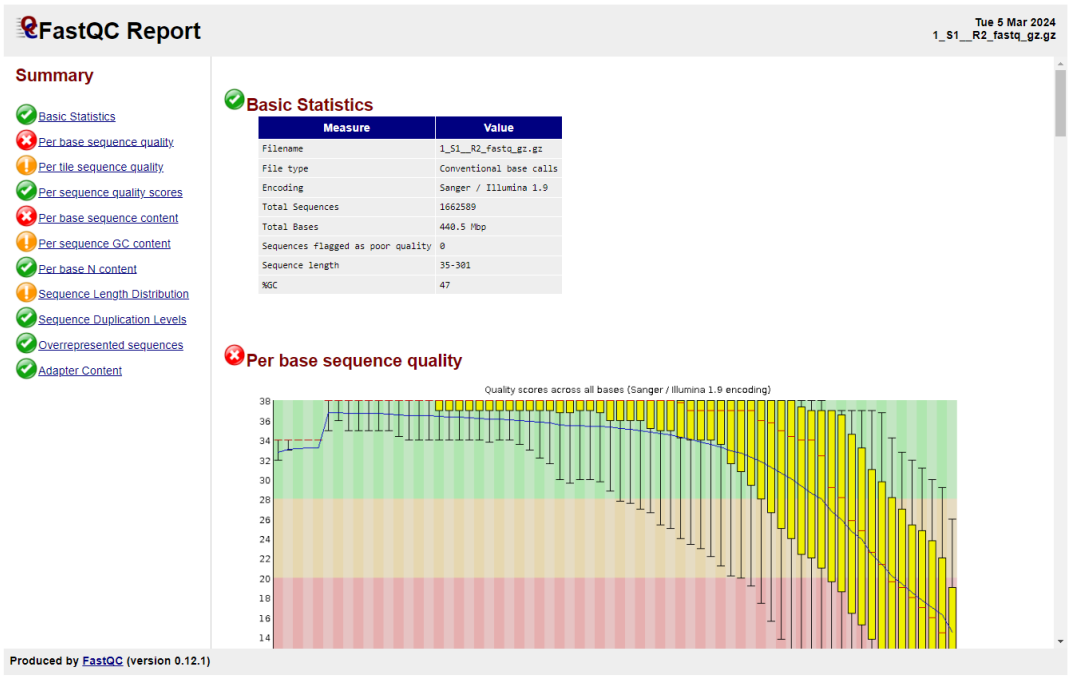
Historial de treball



També veiem a continuació les diferents accions que hem anat fent amb les nostres dades, des de la seua càrrega a l'entorn i el seu processament:



Si cliquem, per exemple, sobre de l'acció 3, visualitzem els resultats de l'anàlisi de qualitat efectuat:



Pràctica 2: Primers passos

És el nostre moment! Treballarem amb les dades en cru de la seqüenciació de la soca Vv5 de *V. vulnificus*. Breument, aquesta soca va ser aïllada d'un cas clínic en tilàpia del Nil (*Oreochromis niloticus*), en la zona costanera d'Israel. Donat el seu interès en el nostre camp, decidim seqüenciar-la mitjançant dues tècniques: Illumina MiSeq, per a obtindre-ne lectures curtes amb una alta profunditat de seqüenciació i molt fiable, i Oxford Nanopore MinION, que ens en dona lectures molt llargues amb una cobertura molt alta encara que amb alguns errors de seqüenciació. La combinació de tots dos tipus de lectura (assemblatge híbrid) ens permet obtindre un genoma d'elevada qualitat i fiabilitat, encara que, òbviament, ens haja costat més diners (inconvenient per a alguns grups de recerca).

Aquest genoma va ser publicat en l' NCBI, associat a un *Announcement*, en el qual es descriu la soca aïllada, i les dades crues de les seqüenciacions es van pujar al Sequence Read Archive, SRA [12].

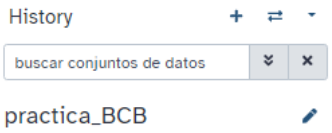
Passos previs

El primer pas, doncs, serà registrar-nos en Galaxy Europa, per a disposar de més espai d'emmagatzematge i treball a l'entorn virtual.

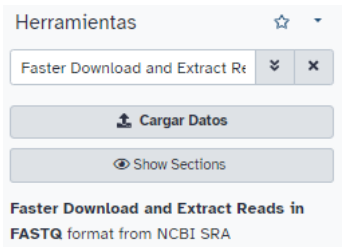
A continuació, accedirem a l'*Announcement*, per a poder disposar dels codis del SRA: **SRR13236868** (lectures curtes d'Illumina MiSeq) y **SRR13453794** (lectures llargues de Oxford Nanopore MinION).

Treballant en Galaxy: importació i visualització de les dades crues

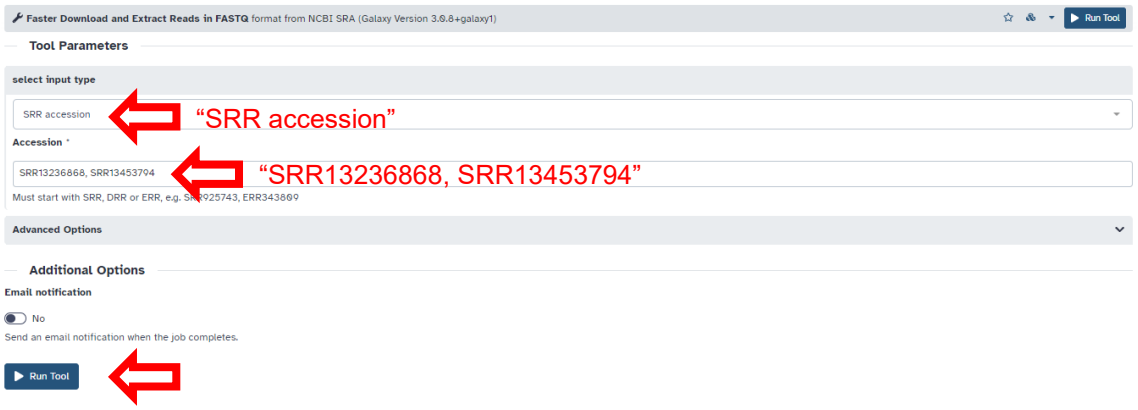
Una vegada hem accedit a l'entorn Galaxy, ens crearem un Historial de treball. Podem anomenar-lo com desitgem:



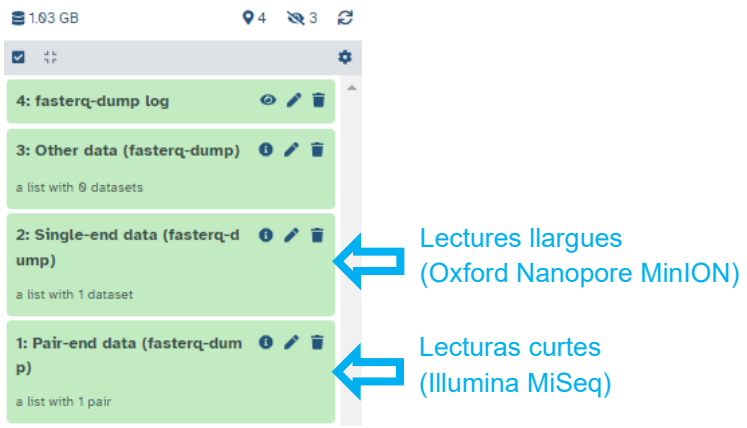
Ara carregarem les dades crues, tant d'Illumina com de Nanopore. L'opció més lenta (i que ocupa espai directament en el nostre ordinador) és descarregar-nos de l'SRA els arxius FASTQ (ocupen molt d'espai) i després carregar-los en Galaxy; una altra opció és carregar-los directament de l'SRA amb l'eina *Faster Download and Extract Reads in FASTQ* (és el que farem):



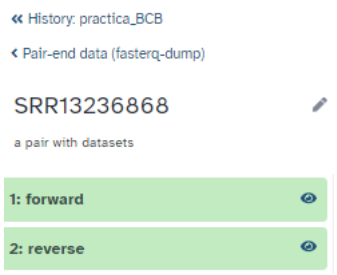
Clicant sobre l'eina, podrem indicar en el requadre «Accession» el número de l'SRR associat a les nostres lectures:



La càrrega dels arxius pot durar diversos minuts (fins i tot hores, segons disponibilitat del servidor; paciència!).



Si cliquem sobre «1: Pair-end data» i posteriorment sobre «1: SRR13236868», accedim a les dades crues d'Illumina. Observem que hi ha dues lectures diferents: *forward* i *reverse*, pel fet que aquest mètode de seqüenciació d'ADN proporciona aquestes lectures per separat.



Si ara cliquem sobre l' de les lectures *forward*, se'ns mostrarà el contingut d'aquest arxiu, és a dir, el FASTQ, amb els seus encapçalats, seqüències de nucleòtids + qualitat de cada base:

Per a accedir a les lectures de MinION, haurem de clicar sobre «2: Single-end data» i sobre l' d' «1: SRR13453794». En aquest cas, aquest mètode de seqüenciació només proporciona les seqüències en sentit 5'—3'.

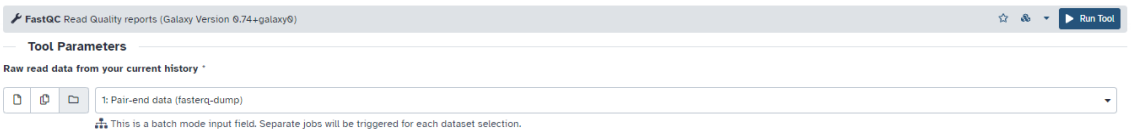
Ja tenim les nostres lectures! Ara toca fer un control de qualitat.

Pràctica 3: Control de qualitat de les lectures

Tot i que les tècniques de seqüenciació són molt útils, els processos necessaris fins a obtindre les dades finals (extracció i fragmentació de l'ADN, preparació de llibreries…) poden resultar en una degradació de l'ADN, i donen lloc a unes seqüències de baixa qualitat. A més, la tecnologia també introdueix uns certs biaixos en les seqüències, per exemple, una sobrerepresentació de seqüències amb un cert contingut en GC [13]. També pot ser necessari llevar possibles adaptadors usats per a la seqüenciació. Per a tot això, usarem dues eines molt útils en el control de qualitat de les seqüències. La primera d'elles és FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), que ens servirà per a veure l'estat de les nostres lectures. La segona serà Prinseq [14] que servirà per a llevar-nos aquelles seqüències que ens aporten molt de soroll.

Així doncs, amb les dades crues d'Illumina i Nanopore carregades, visualitzarem la seua qualitat. Per a això, utilitzarem l'eina que ja coneixem: FASTQC.

Una vegada dins de la pròpia eina, arrosseguem les lectures *1: Pair-end data* cap al rectangle a *Raw read data from your current history*, i li donem a *Run tool* per a executar-la:



Repetirem el mateix procés, però amb les lectures *2: Single-end data*.

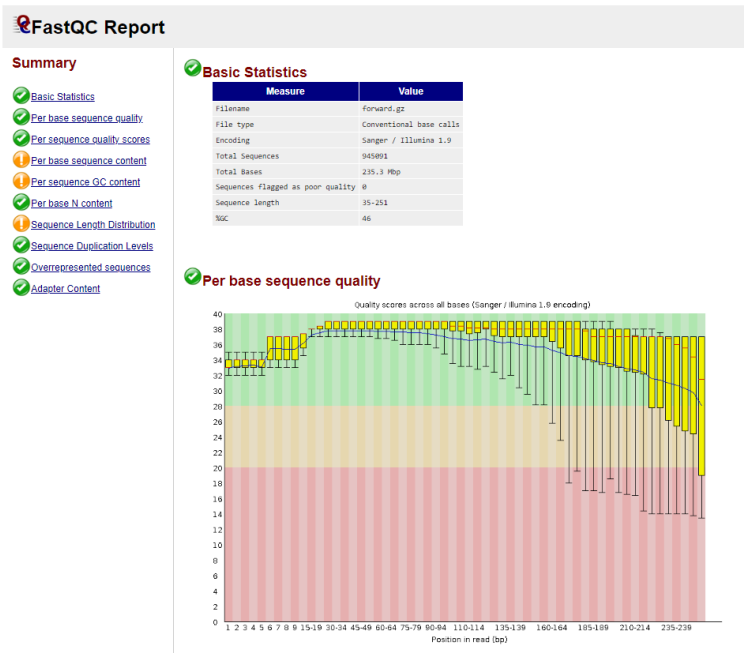
I obtenim com a resultat:

el

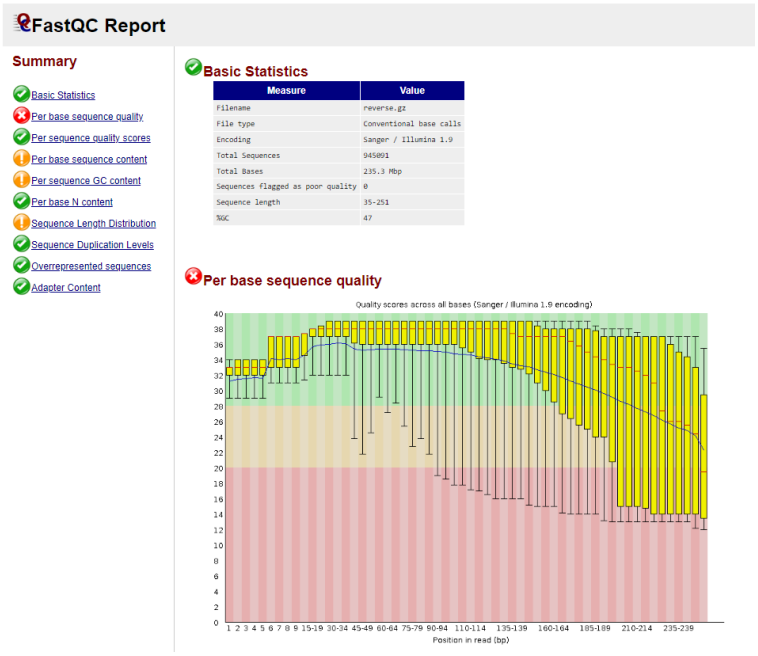
la

Sent el fitxer 8: *FastQC on collection 1: Webpage* visualitzador de la qualitat de les lectures *forward* i *reverse* (per separat) d'Illumina, i el fitxer *14: FastQC on collection 2: Webpage*, el de qualitat de les lectures de MinION.

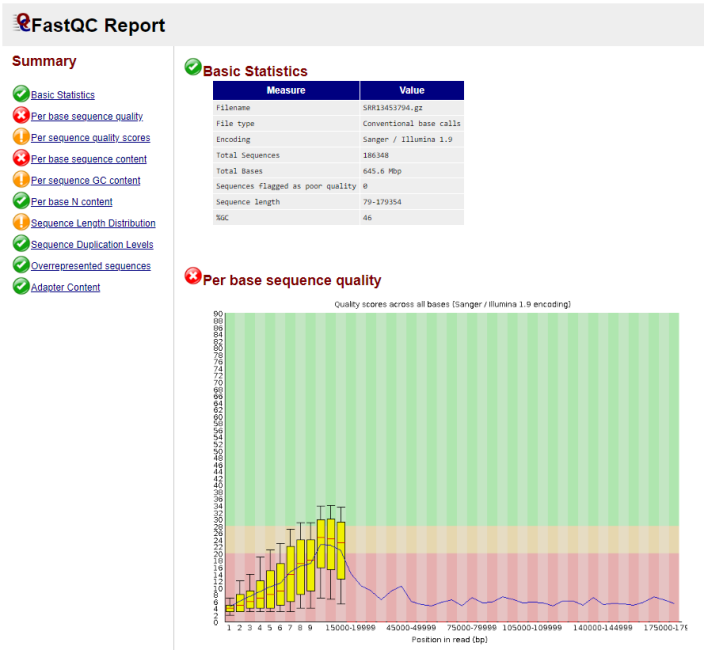
En primer lloc, visualitzarem la qualitat de les lectures *forward* d'Illumina:



Tot seguit, veurem la qualitat de les *reverse* d'Illumina:

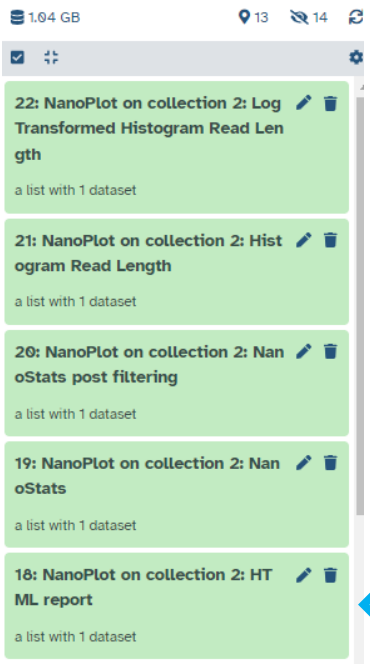


Finalment, observem la qualitat de les lectures de MinION:



Exercici opcional: Comentar les diferències que s'observen en el gràfic *Per base sequence quality* entre les lectures curtes i les llargues.

Veiem que el gràfic de FastQC de les lectures de MinION no és molt correcte. Això és pel fet que aquesta eina està especialitzada en lectures curtes. Per a lectures llargues, tenim altres eines d'anàlisis de qualitat com NanoPlot [15]. Busquem l'eina i l'executem amb les lectures *2: Single-end data*.



Els arxius HTML solen ser els informes de resultats

En visualitzem el resultat:

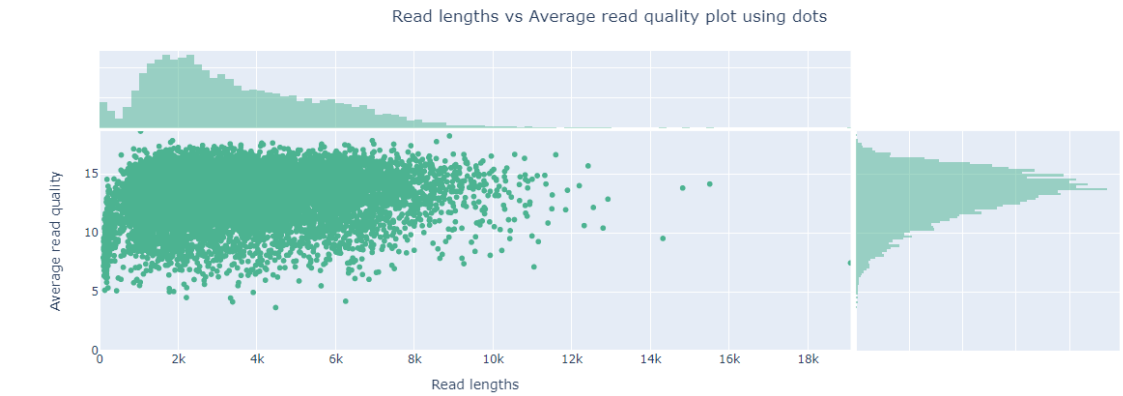
Summary Statistics | Plots | Report issue on Github

NanoPlot reports

Summary statistics

Metrics	dataset
number_of_reads	186348
number_of_bases	645611372.0
median_read_length	2925.0
mean_read_length	3464.5
read_length_stdev	2216.6
n50	4712.0
mean_qual	13.1
median_qual	13.4
longest_read_(with_Q):1	179354 (4.8)
longest_read_(with_Q):2	43459 (5.2)
longest_read_(with_Q):3	27503 (6.8)
longest_read_(with_Q):4	23895 (14.4)
longest_read_(with_Q):5	19522 (11.7)
highest_Q_read_(with_length):1	19.3 (1049)
highest_Q_read_(with_length):2	18.9 (3678)
highest_Q_read_(with_length):3	18.8 (3291)
highest_Q_read_(with_length):4	18.8 (6482)
highest_Q_read_(with_length):5	18.7 (5257)
Reads >Q5:	186237 (99.9%) 645.1Mb
Reads >Q7:	184417 (99.0%) 639.9Mb
Reads >Q10:	167803 (90.0%) 592.0Mb
Reads >Q12:	135445 (72.7%) 487.9Mb
Reads >Q15:	34565 (18.5%) 133.4Mb

Read lengths vs Average read quality plot using dots



Observem que el mètode MinION dona com a resultat un ventall molt gran de longituds de lectura: des de grandàries molt xicotetes, de pràcticament 1 base, fins a grandàries molt grans, majors de 18.000 bases. No obstant això, la qualitat és prou baixa, amb una mitjana de 12.

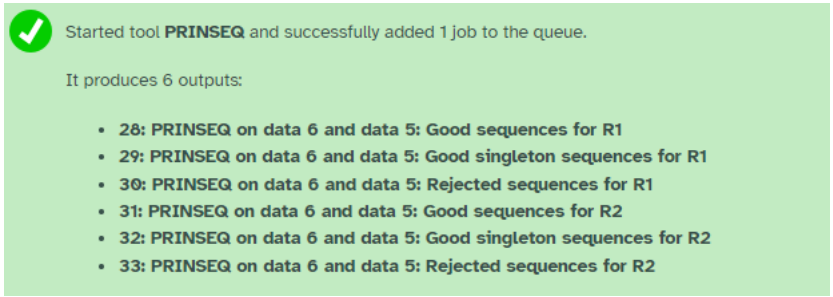
Després de la visualització de la qualitat de les lectures (curtes i llargues), procedim al filtrat aplicant diferents paràmetres: qualitat mínima i que ens *trimege* les últimes posicions de les lectures de mala qualitat. Utilitzarem l'eina PRINSEQ, i definirem que les llibreries són *Paired-end* (Illumina) o *Single-end* (MinION). I arrossegarem els arxius de les lectures al/als requadre/s.

Per a les lectures curtes, en modificarem alguns paràmetres (la resta, els podem deixar per defecte, o jugar amb els valors). En aquesta pràctica farem:

- "Filter sequences based on their mean score?" → YES
  "Filter sequences based with too small mean score?" → YES
  "Minimum mean score threshold to conserve sequences" → 25
- "Trim sequence by quality score?" → YES
  "Trim sequence by quality score from the 3'-end?" → YES
  "Quality score threshold to trim positions" → 20

Exercici opcional: Usant les lectures d'Illumina, empreu una configuració diferent i compareu-ne els resultats. Per exemple, es filtren més o menys lectures? Podríem ser més restrictius?

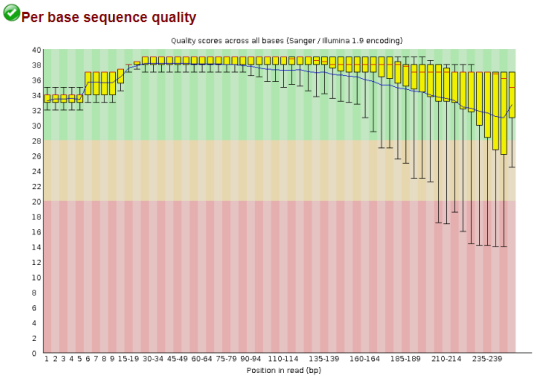
Resultat:



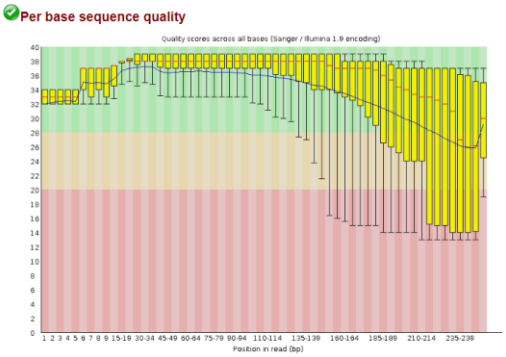
Els arxius que ens interessen (FASTQ filtrats) són el ***28: PRINSEQ on data 6 and data 5: Good sequences for R1*** (per a les lectures *forward*) i el ***31: PRINSEQ on data 6 and data 5: Good sequences for R2*** (per a les *reverse*).

Amb FASTQC en visualitzem la nova qualitat.

Per a les R1 (*Forward*):



I per a les R2 (*Reverse*):



Veiem una millora considerable respecte de les dades crues. Així doncs, aquests arxius «28» i «31» seran les nostres lectures filtrades definitives, que usarem per a assemblar i obtindre el nostre genoma.

Per al cas de les lectures llargues de MinION, podríem usar també PRINSEQ, però hi ha una altra eina anomenada NanoFilt molt més especialitzada en lectures llargues de Nanopore, igual que ocorria amb la visualització de la qualitat d'aquestes lectures, que usàvem NanoPlot. No obstant això, en l'entorn Galaxy no trobem l'eina NanoFilt.

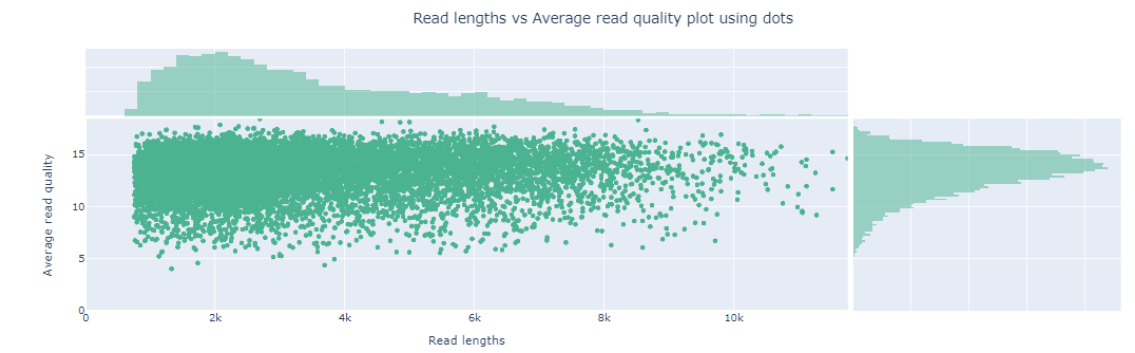
Així doncs utilitzarem una altra eina de filtrat d'arxius FASTQ com ara *Filter FASTQ*. Aquesta eina permet filtrar per qualitat i per longitud de les lectures. En el nostre cas, estem usant unes dades crues reals, no manipulades, per la qual cosa, casualment, tenen una qualitat prou baixa (com ja vam veure en la representació de la qualitat). Per això, filtrarem només per longitud de lectura, i ens quedarem amb aquelles lectures la longitud de les quals comprèn entre 750 i 12.000 bases. Obtenim l'arxiu ***38: Filter FASTQ on collection 2***.

Després de la visualització per NanoPlot de la nova qualitat:

Summary statistics

	dataset
Metrics	
number_of_reads	177859
number_of_bases	639826283.0
median_read_length	3048.0
mean_read_length	3597.4
read_length_stdev	2093.1
n50	4713.0
mean_qual	13.2
median_qual	13.5
longest_read_(with_Q):1	11998 (13.2)
longest_read_(with_Q):2	11995 (12.3)
longest_read_(with_Q):3	11990 (9.7)
longest_read_(with_Q):4	11985 (5.6)
longest_read_(with_Q):5	11985 (16.0)
highest_Q_read_(with_length):1	19.3 (1049)
highest_Q_read_(with_length):2	18.9 (3678)
highest_Q_read_(with_length):3	18.8 (3291)
highest_Q_read_(with_length):4	18.8 (6482)
highest_Q_read_(with_length):5	18.7 (5257)
Reads >Q5:	177755 (99.9%) 639.5Mb
Reads >Q7:	176314 (99.1%) 634.6Mb
Reads >Q10:	162813 (91.5%) 587.6Mb
Reads >Q12:	133091 (74.8%) 484.8Mb
Reads >Q15:	34387 (19.3%) 133.0Mb

Read lengths vs Average read quality plot using dots



Observem una homogeneïtat superior en les lectures. Ens quedarem amb aquest arxiu «**38**» com a lectures llargues filtrades per a l'assemblatge final.

Pràctica 4: Assemblatge *de novo*

Ja tenim les nostres lectures curtes i llargues filtrades i llistes per a assemblar-les. Recordem quins arxius són importants:

- Lectures curtes:
  - R1 (*forward*): «**28: PRINSEQ on data 6 and data 5: Good sequences for R1**»
  - R2 (*reverse*): «**31: PRINSEQ on data 6 and data 5: Good sequences for R2**»
- Lectures llargues: «**38: Filter FASTQ on collection 2**»

Hi ha multitud de mètodes d'assemblatge, cadascun basat en un algorisme diferent. És molt important provar diferents mètodes per a esbrinar quin s'ajusta millor a les nostres dades. En aquesta pràctica proposarem tres assemblatges diferents, amb la condició de visualitzar-ne les diferències i quedar-nos amb el millor assemblatge:

- Assemblatge amb les lectures curtes d'Illumina
- Assemblatge amb les lectures llargues de MinION
- Assemblatge híbrid, combinant les lectures curtes i les llargues

Per a l'assemblatge de les lectures curtes s'utilitza l'eina SPAdes [16] ja que sol produir els millors resultats.

La resta dels camps els deixem per defecte (per a aquesta pràctica).

El resultat que n'obtenim:

Podem descarregar-nos l'arxiu FASTA clicant sobre el resultat «52» i donant-li a «Save».

Per a l'**assemblatge de lectures llargues**, usem Flye [17]. Arrosseguem les lectures de MinION filtrades («38») al requadre de FASTQ i determinem el mode de l'assemblatge, és a dir, quin tipus de lectura estem introduint en l'eina, en el nostre cas, *Nanopore raw (--nano-raw)*. La resta dels camps també els deixem per defecte (per a aquesta pràctica).

El resultat que n'obtenim:

D'igual manera, podem descarregar-nos l'arxiu FASTA clicant sobre el resultat «54» i donant-li a *Save*.

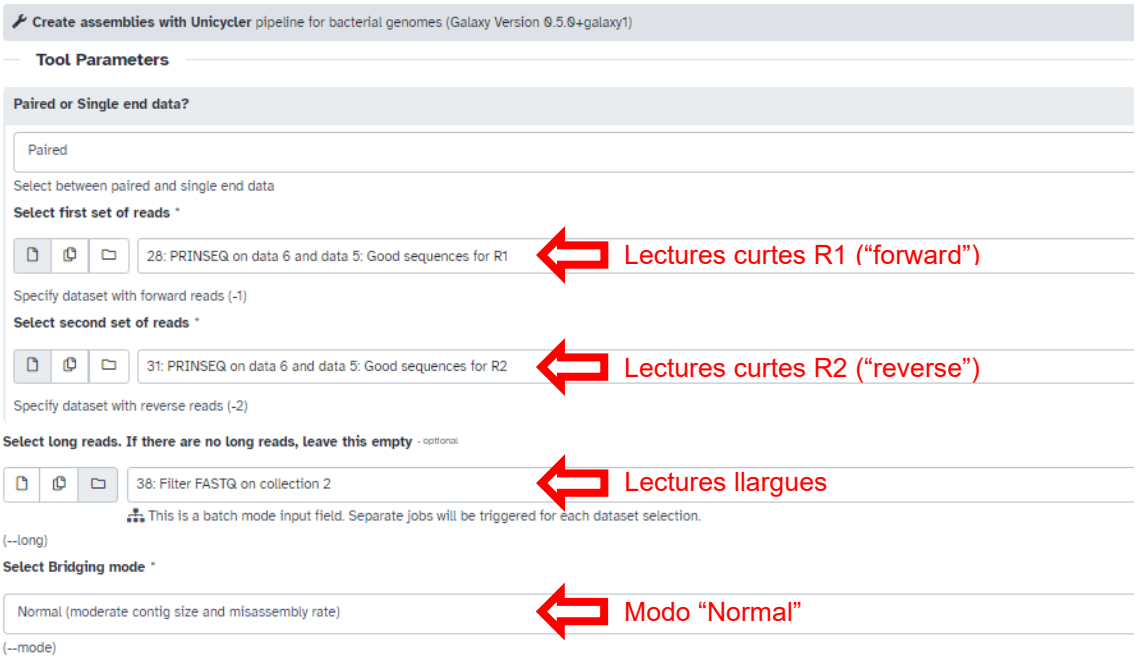
Finalment, per a l'**assemblatge híbrid**, utilitzarem l'eina Unicycler [18]. Aquesta aplicació permet assemblar lectures curtes, lectures llargues i una combinació dels dos tipus, obtenint perquè un assemblatge és molt més fiable i de major qualitat. En comptar tant amb lectures llargues com lectures curtes, podem obtindre millors resultats, ja que totes dues tecnologies es complementen i els avantatges d'una ajuden a suplir les mancances de l'altra.

Així doncs, introduirem tant les lectures curtes *forward* i *reverse* com també les llargues de MinION. A més, Unicycler permet realitzar l'assemblatge seguint tres modes diferents:

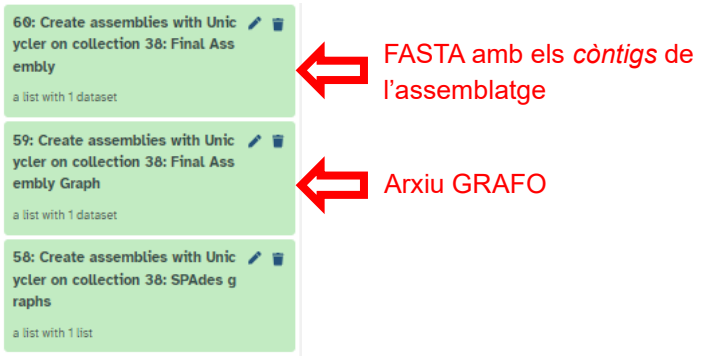
- Normal: genera *còntigs* de grandària moderada i comporta una taxa d'errors d'assemblatge.
- Conservatiu: genera *còntigs* més curts, però amb un error d'assemblatge menor.
- Robust: genera *còntigs* de major grandària, amb una taxa d'error d'assemblatge major.

En la nostra pràctica, utilitzarem el mode normal, però sentiu-vos lliures de jugar amb els modes d'assemblatge en un futur, i compareu-ne els resultats entre si.

La resta dels paràmetres els deixem per defecte (per a aquesta pràctica).



El resultat que n'obtenim:

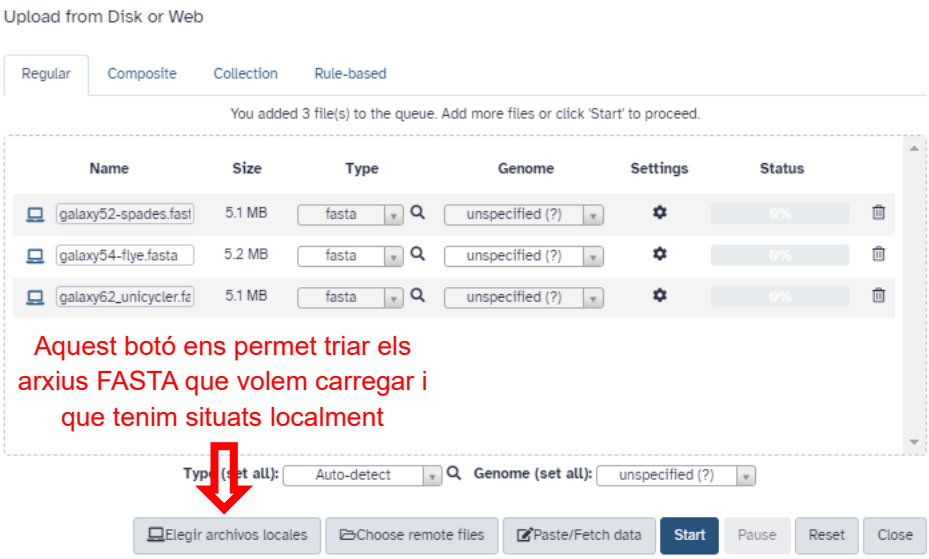


Finalment, descarreguem l'arxiu FASTA clicant sobre el resultat «60» i donant-li a *Save*.

Ja tenim els nostres tres assemblatges del genoma de la soca Vv5 de *V. vulnificus*. Amb ells podem treballar d'ara en avant.

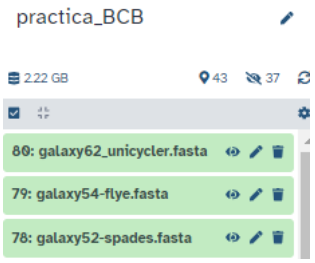
Cal dir que, com que amb el temps de la pràctica no podrem fer els tres assemblatges (són diverses hores d'execució de les eines), carregarem directament els arxius FASTA que ja hem obtingut anteriorment (i els tindreu a l'aula virtual).

Per a això, usarem la funció de Galaxy *Carregar Dades*, la qual trobem en la secció «Eines». Se'ns obrirà una finestra emergent:



Després de seleccionar els tres FASTA, canviem el tipus d'arxiu (*Type*) a *fasta*. Li donem a «Start». Una vegada carregats, podem tancar la finestra emergent.

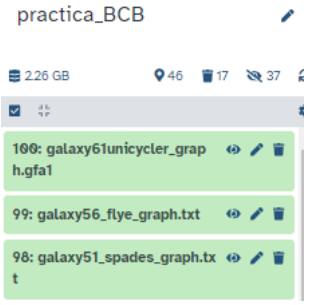
Ja tenim carregats els nostres FASTA:



Així mateix, descarregarem els arxius GRAF i els carregarem també a Galaxy. Aquests arxius ens permetran visualitzar amb l'eina Bandage la disposició espacial dels còntigs del nostre genoma.

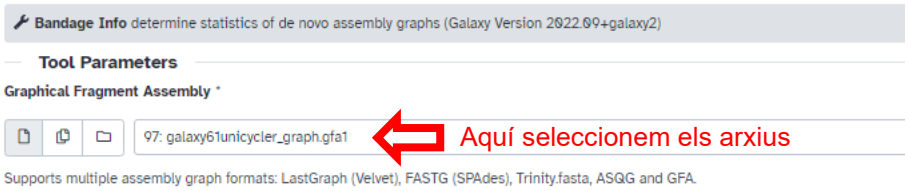
Usem també l'acció «Carregar Dades» de la secció d'Eines. Seleccionem els tres arxius GRAF, i en aquest cas no és necessari assignar el tipus d'arxiu en «Type». Li donem a «Start». Una vegada carregats, podem tancar la finestra emergent.

Ja tenim carregats els nostres GRAFOS:



Pràctica 5: Comparació dels assemblatges

Com a resultat de la pràctica anterior hem obtingut diversos assemblatges diferents. En aquesta pràctica, utilitzarem Bandage [19] per a visualitzar-ne els grafs obtinguts i Quast [20] per a comprar diferents estadístics. Això ens permetrà seleccionar el millor de tots. En primer lloc, visualitzarem la informació relacionada amb els nostres assemblatges. Usarem els arxius GRAFO en l'eina *Bandage Info*, que trobarem en el panell d' «Eines»:



Resultat per al GRAFO d' SPAdes:

Column 1	Column 2
Node count:	428
Edge count:	570
Smallest edge overlap (bp):	127
Largest edge overlap (bp):	127
Total length (bp):	5263714
Total length no overlaps (bp):	5209612
Dead ends:	4
Percentage dead ends:	0.46729%
Connected components:	10
Largest component (bp):	5195785
Total length orphaned nodes (bp):	978
N50 (bp):	198295
Shortest node (bp):	128
Lower quartile node (bp):	158
Median node (bp):	268
Upper quartile node (bp):	1539
Longest node (bp):	331696
Median depth:	36.6602
Estimated sequence length (bp):	5538437

Resultat per al GRAFO de Flye:

Column 1	Column 2
Node count:	11
Edge count:	20
Smallest edge overlap (bp):	0
Largest edge overlap (bp):	0
Total length (bp):	5293417
Total length no overlaps (bp):	5293417
Dead ends:	0
Percentage dead ends:	0%
Connected components:	4
Largest component (bp):	3469391
Total length orphaned nodes (bp):	0
N50 (bp):	1514179
Shortest node (bp):	919
Lower quartile node (bp):	2666
Median node (bp):	57460
Upper quartile node (bp):	859181
Longest node (bp):	1761235
Median depth:	0
Estimated sequence length (bp):	0

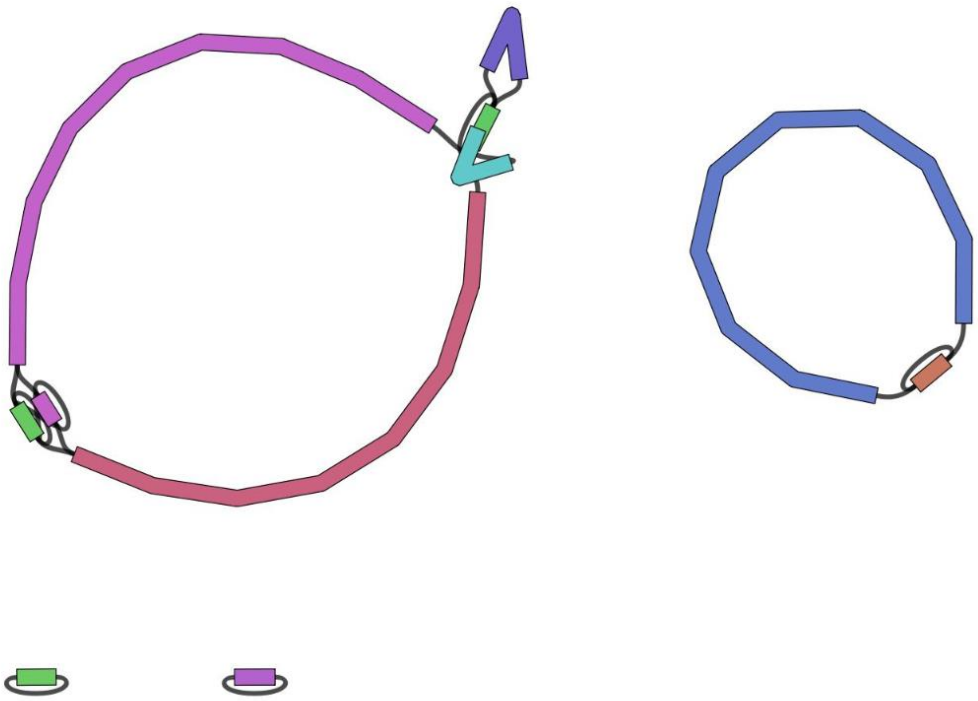
Resultat per al GRAFO d'Unicycler:

Column 1	Column 2
Node count:	67
Edge count:	78
Smallest edge overlap (bp):	0
Largest edge overlap (bp):	0
Total length (bp):	5313719
Total length no overlaps (bp):	5313719
Dead ends:	0
Percentage dead ends:	0%
Connected components:	5
Largest component (bp):	3473921
Total length orphaned nodes (bp):	0
N50 (bp):	2910955
Shortest node (bp):	1
Lower quartile node (bp):	44
Median node (bp):	111
Upper quartile node (bp):	559
Longest node (bp):	2910955
Median depth:	0
Estimated sequence length (bp):	0

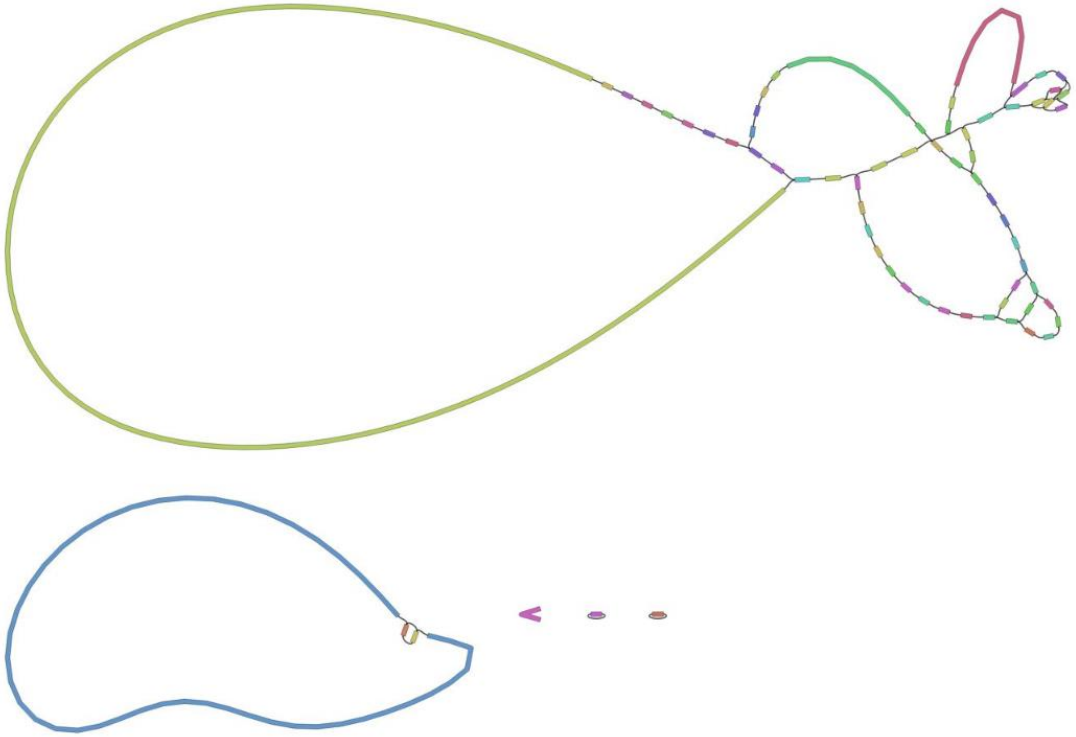
A continuació, visualitzem l'assemblatge com a tal, mitjançant l'eina *Bandage Image*; d'igual manera que en el cas anterior, seleccionem els arxius GRAFO carregats, i la resta dels paràmetres els deixem per defecte.

Resultat per al GRAFO d'SPAdes:

Resultat per al GRAFO de Flye:



Resultat per al GRAFO d'Unicycler:



Si en fem memòria, el nombre de *còntigs* del genoma ha de ser igual al nombre de cromosomes (+ plasmidis o altres elements genètics mòbils, si és el cas) del bacteri; en cas de no tenir un assemblatge així, el nombre de *còntigs* hauria de ser el menor possible, i preferiblement units entre si en el GRAFO.

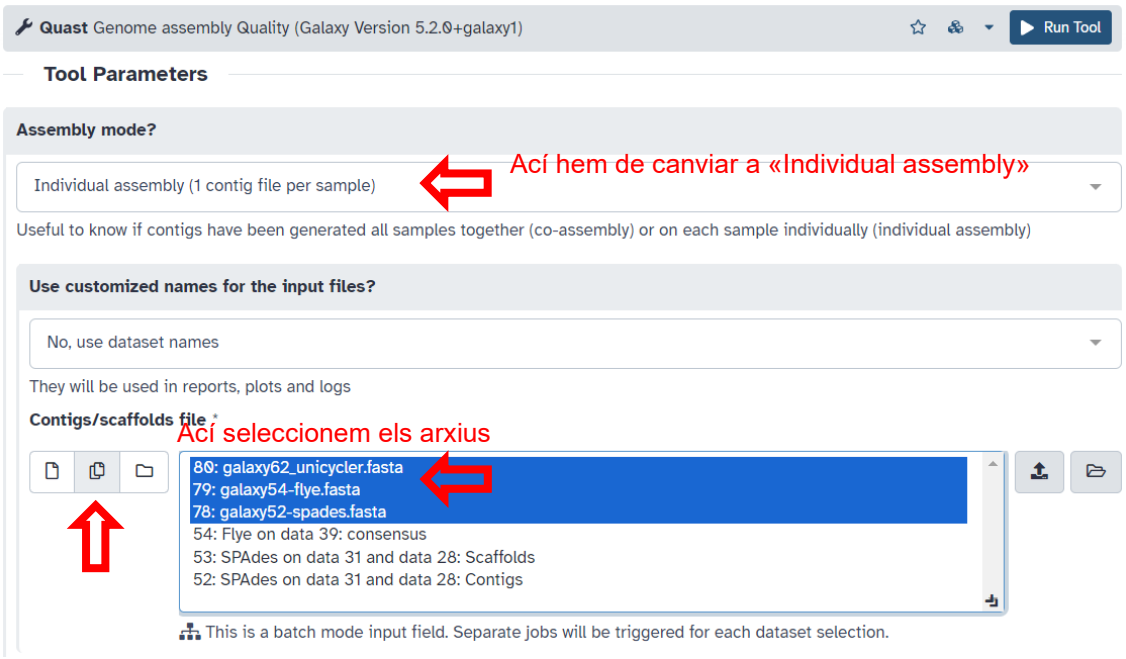
Així doncs, observant-ne els resultats obtinguts, podríem pensar que l'eina Flye ens ha donat un assemblatge del genoma millor, ja que la seua eixida ha donat 11 nodes (equivalents al número de *còntigs*), la majoria dels quals estan connectats amb els seus possibles cromosomes, davant dels 428 i 67 nodes d'Illumina i Unicycler, respectivament. No obstant això, farem altres anàlisis de qualitat dels assemblatges resultants per a comparar-los i decidir amb quin ens quedem.

En primer lloc, obtindrem unes estadístiques generals de cada assemblatge, com l' N50 o l' L50. Una bona eina per a això és *Quast*. Però, abans, un breu recordatori dels arxius a utilitzar:

- Assemblatge amb lectures curtes: "**78: galaxy52-spades.fasta**"
- Assemblatge amb lectures llargues: "**79: galaxy54-flye.fasta**"
- Assemblatge híbrid: "**80: galaxy62_unicycler.fasta**"

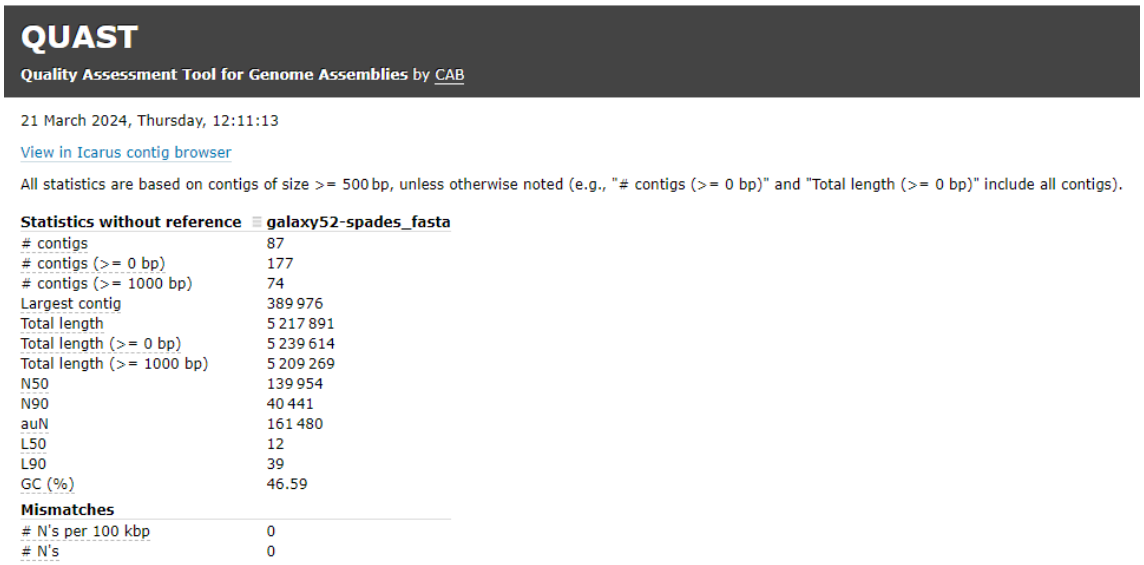
El primer que hem de fer és canviar el mode d'assemblatge a assemblatge individual, ja que en el nostre cas cada assemblatge està en un arxiu.

Seleccionarem també cadascun dels assemblatges, per a això, primer clicarem en «múltiples datasets» i seleccionarem cadascun dels arxius mantenint polsada la tecla «Control/Ctrl». D'aquesta manera, el programa s'executarà alhora per a cadascun dels assemblatges.



Addicionalment, es pot fer una estimació del nombre de gens o la busca d'ortòlegs amb BUSCO. Tanmateix, això ho farem més endavant. Una altra opció és canviar l'arxiu d'eixida perquè genere l'informe en altres formats (PDF per exemple). Nosaltres ho deixarem per defecte.

Una vegada tenim les opcions que volem, fem servir l'eina, la qual ens generarà el següent informe:



QUAST

Quality Assessment Tool for Genome Assemblies by CAB

21 March 2024, Thursday, 12:11:13

View in Icarus contig browser

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

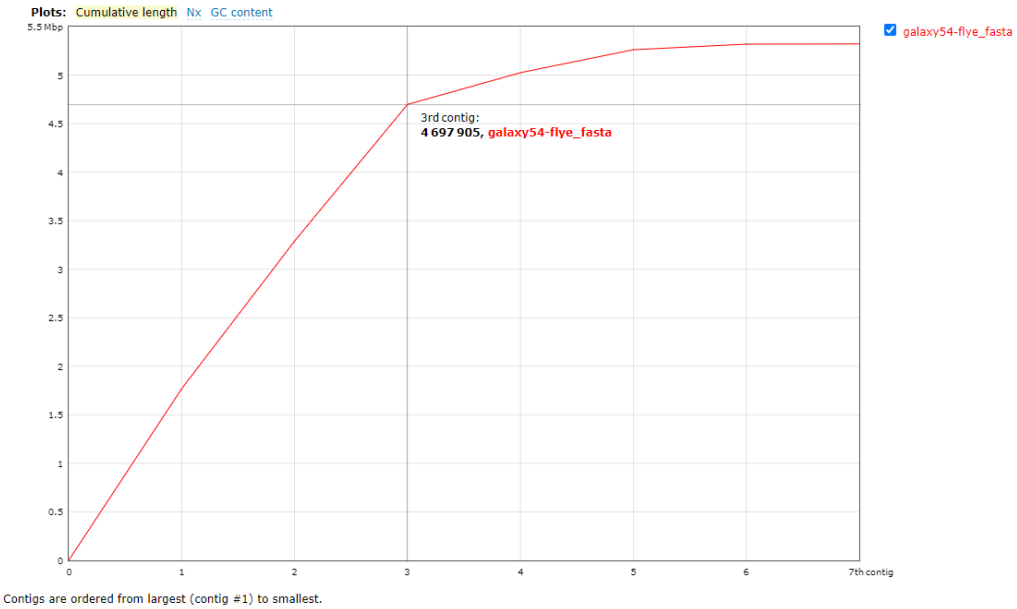
Statistics without reference	galaxy52-spades_fasta
# contigs	87
# contigs (>= 0 bp)	177
# contigs (>= 1000 bp)	74
Largest contig	389 976
Total length	5 217 891
Total length (>= 0 bp)	5 239 614
Total length (>= 1000 bp)	5 209 269
N50	139 954
N90	40 441
auN	161 480
L50	12
L90	39
GC (%)	46.59
Mismatches	
# N's per 100 kbp	0
# N's	0

Si els comparem, n'obtenim açò:

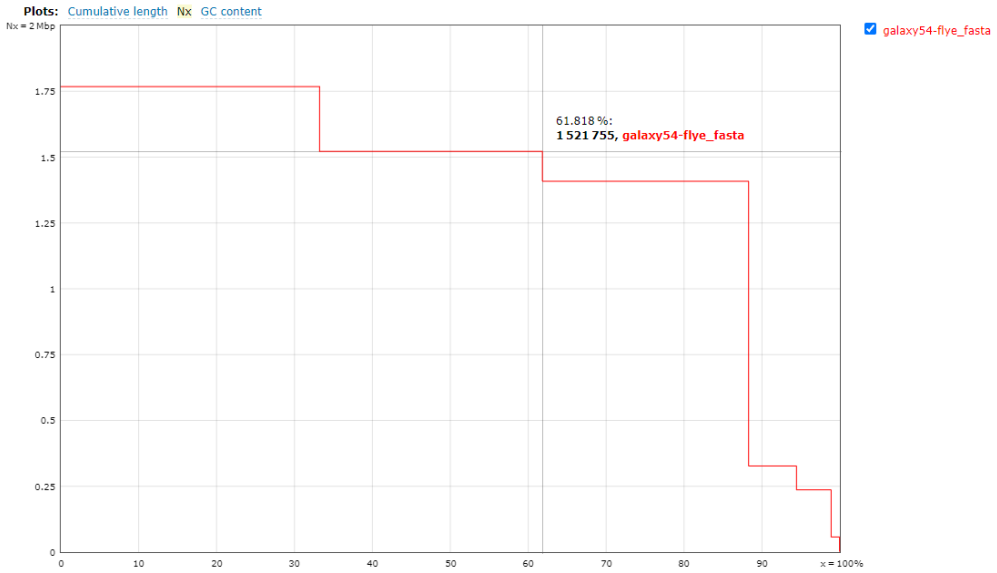
Statistics without reference	galaxy52-spades_fasta	galaxy54-flye_fasta	galaxy62_unicycler_fasta
# contigs	87	7	18
# contigs (>= 0 bp)	177	7	36
# contigs (>= 1000 bp)	74	7	15
Largest contig	389 976	1 767 815	2 910 055
Total length	5 217 891	5 321 355	5 309 019
Total length (>= 0 bp)	5 239 614	5 321 355	5 312 465
Total length (>= 1000 bp)	5 209 269	5 321 355	5 307 255
N50	139 954	1 521 755	2 910 055
N90	40 441	327 088	285 508
auN	161 480	1 426 432	2 212 148
L50	12	2	1
L90	39	4	3
GC (%)	46.59	46.67	46.64
Mismatches			
# N's per 100 kbp	0	0	0
# N's	0	0	0

D'aquest informe, hi ha diversos estadístiques molt importants que ens ajudaran a triar el millor assemblatge d'entre els tres que hem generat. Aquests són: número de *còntigs*, N50/N90 i L50/L90. A més, també és convenient revisar el contingut en GC, perquè sol ser prou constant dins de cada espècie.

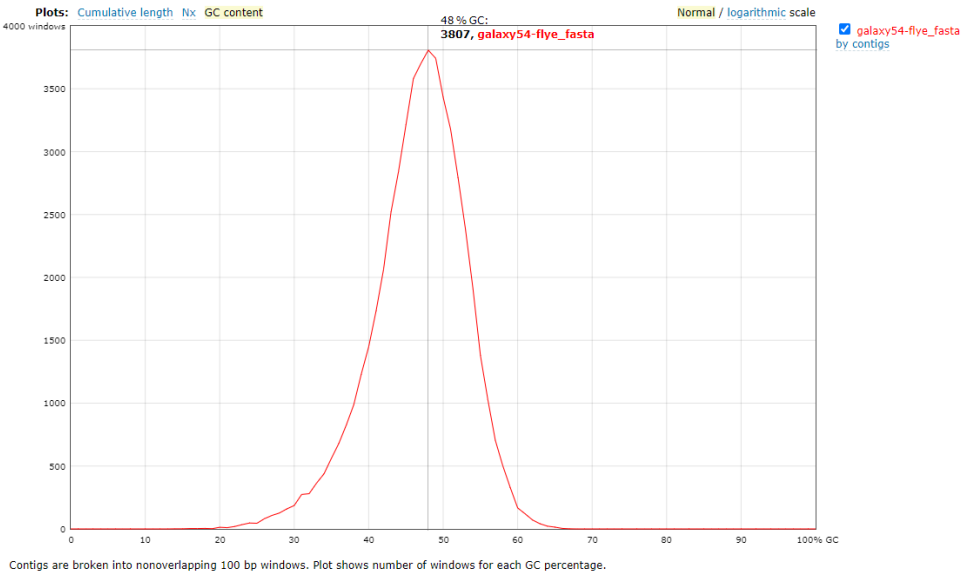
A més, és interessant revisar els gràfics que genera el programa. Per a entendre la informació que proporciona, ens quedem amb l'assemblatge de Flye. El primer d'ells, el de llargària acumulada, té una forma corba que s'estabilitza al voltant d'un valor. Aquest gràfic es basa en l'ordenació per grandària dels *còntigs* que hem vist per a l'N50 i l'L50. Representa la grandària de l'assemblatge segons afegim més *còntigs*. En la imatge de sota podem veure que si col·loquem el cursor per a cada valor de l'eix X (és a dir, cada *còntig*) podem veure'n la grandària equivalent; així, en aquest cas, amb tres *còntigs* tenim una grandària total de 4.697.905 pb. Per tant, el valor en el qual s'estabilitza és la grandària total de l'assemblatge, que serà el resultat de la suma de tots els *còntigs*.



El segon gràfic, l' Nx, també es basa en l'ordenació per grandària dels còntigs, però en aquest cas representa la grandària de cada *còntig* i el percentatge acumulat del genoma que suposen. Per exemple, en el cas de la figura de sota, els dos *còntigs* més grans suposen un 61,8 % del total de la grandària de l'assemblatge i el segon té una grandària aproximada de 1,5 Mpb.



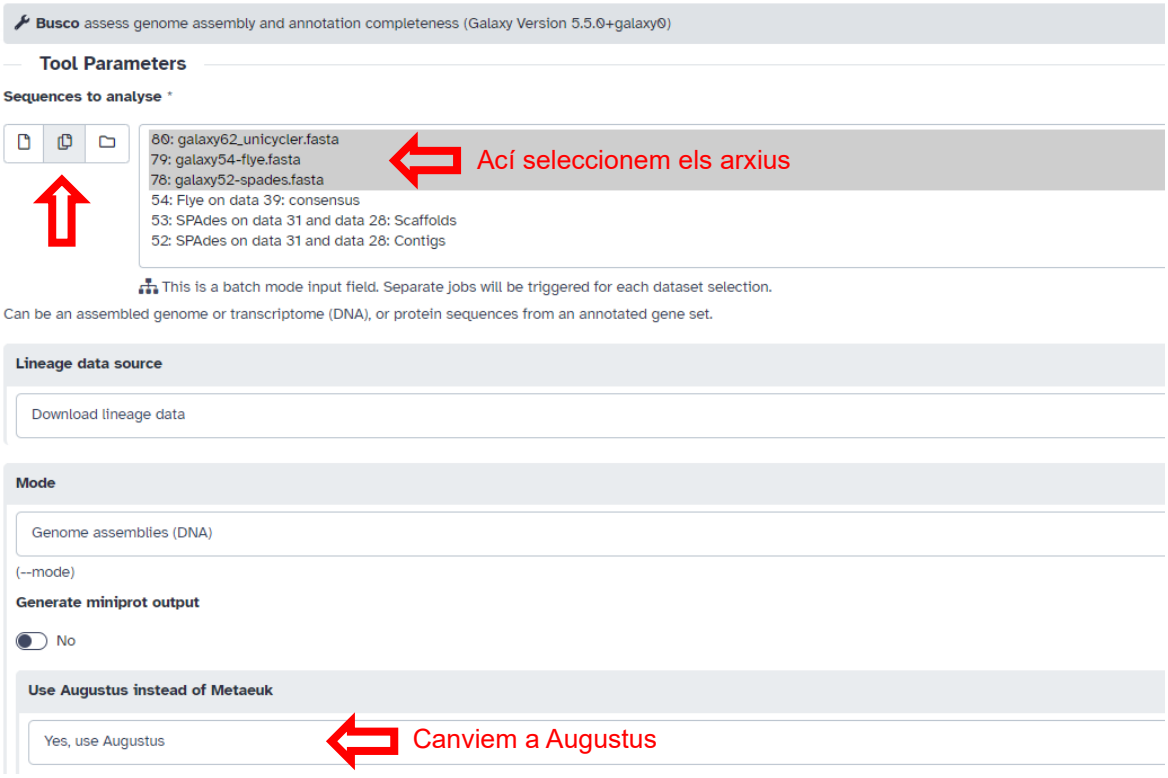
El tercer gràfic també és molt informatiu, especialment si tenim alguna contaminació o un mal assemblatge. Per a generar aquesta imatge, el programa trenca l'assemblatge en "finestres" de 100 pb sobre les quals calcula el contingut en GC. Després, representa la quantitat de finestres que hi ha per a cada percentatge en GC. En l'histograma d'aquest assemblatge, veiem que el màxim es troba en el 48 %, amb 3.807 finestres amb aquest percentatge. Com veiem, aquest valor està molt prop del contingut total de l'assemblatge, que és 46,67 %. Si tinguérem cap mena de contaminació, veuríem un altre màxim relatiu en valors més allunyats.

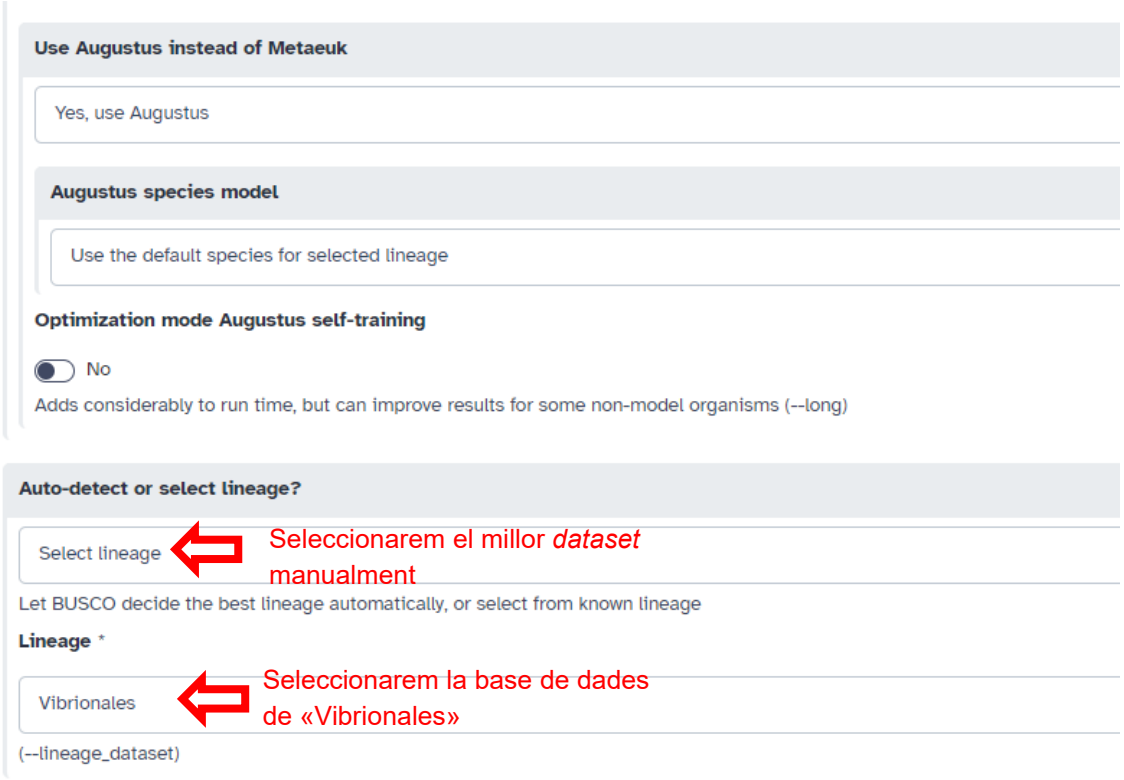


Exercici opcional: Revisar i comentar les gràfiques que es generen per a la resta dels assemblatges. Són comparables?

A priori, el millor assemblatge seria el que hem obtingut amb Flye, perquè té un nombre de *còntigs* inferior. No obstant això, el que hem obtingut amb Unicycler té un L50 i un L90 més baixos, i l'N50 més alt. Per tant, no convé descartar-lo encara. Per a prendre una millor decisió sobre quin genoma mantenir, valorarem el grau de completesa de cadascun. Per a això, usarem BUSCO, el qual busca en cada genoma una sèrie de gens que estan conservats en diferents grups.

Una vegada tenim l'eina, el primer, com sempre, serà seleccionar els arxius sobre els quals treballarem. Després, passem a triar l'algorisme per a la busca de gens. BUSCO ofereix dues alternatives: Metaeuk (per defecte) i Augustus. La diferència entre ambdues és que Metaeuk és més ràpid però un poc menys sensible que Augustus. Aquesta diferència és més notòria en organismes eucariotes, on la busca de gens és més complexa a causa de la presència d'introns i exons, a més, els genomes són molt més grans. Atès que el genoma del nostre *V. vulnificus* no és gran, usarem Augustus.





Quan ho tinguem tot, fem servir l'eina. Això ens generarà dues eixides: una taula completa i un breu resum. Encara que la taula completa té més informació, per a obtindre els estadístiques, simplement ens anirem al breu resum, que tindrà aquesta forma:

```
***** Results: *****

C:99.9%[S:99.8%,D:0.1%],F:0.1%,M:0.0%,n:1445
1443    Complete BUSCOs (C)
1442    Complete and single-copy BUSCOs (S)
1       Complete and duplicated BUSCOs (D)
1       Fragmented BUSCOs (F)
1       Missing BUSCOs (M)
1445    Total BUSCO groups searched
```

També genera estadístiques de l'assemblatge, però per a això tenim els de Quast. Ens interessa el nombre de gens que ha trobat. Si comparem els tres assemblatges, veiem que, encara que l'assemblatge per Flye semblava millor en termes de contigüitat, realment té més errors d'assemblatge.

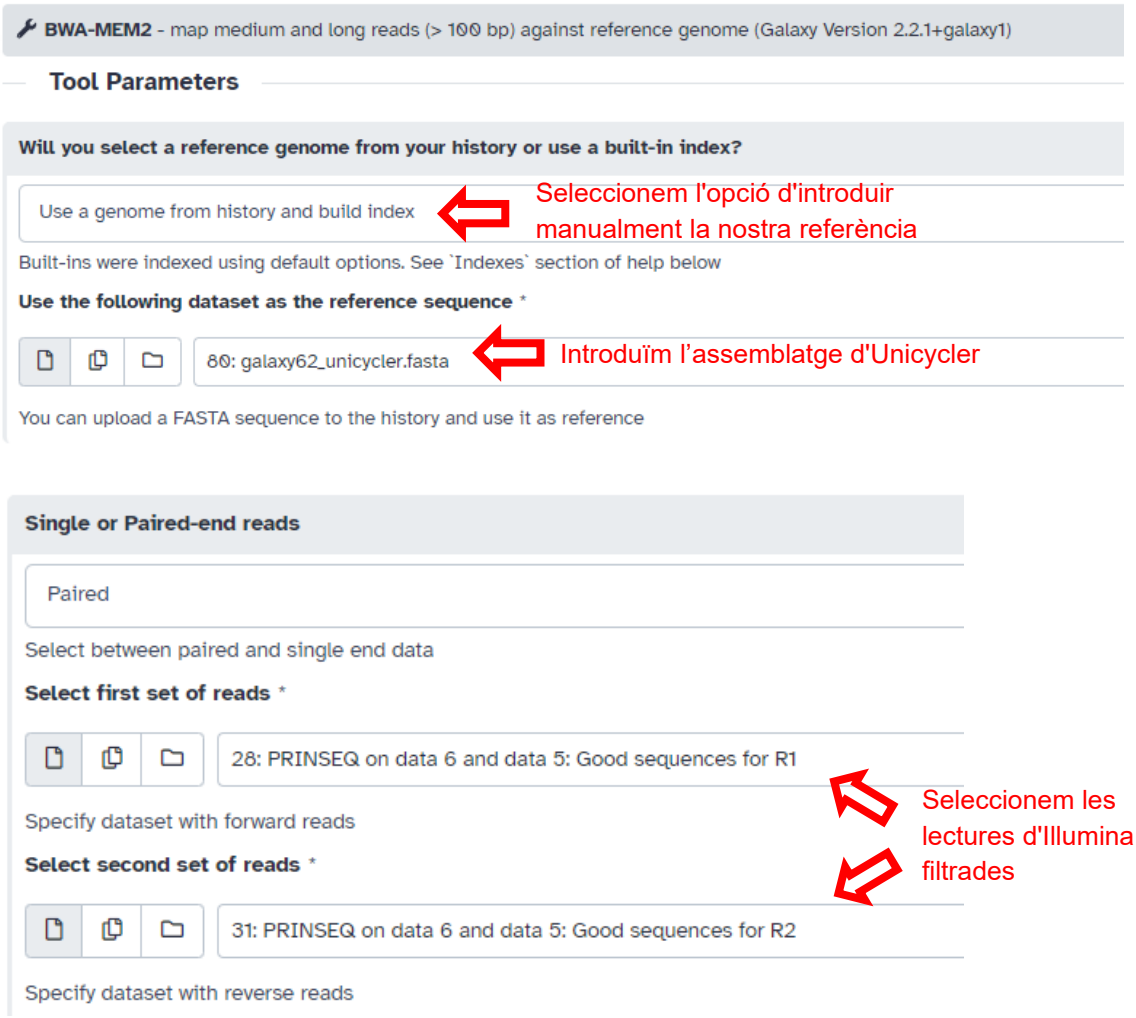
```
***** Results: *****

C:94.5%[S:94.4%,D:0.1%],F:3.5%,M:2.0%,n:1445
1365    Complete BUSCOs (C)
1364    Complete and single-copy BUSCOs (S)
1       Complete and duplicated BUSCOs (D)
51      Fragmented BUSCOs (F)
29      Missing BUSCOs (M)
1445    Total BUSCO groups searched
```

Una explicació d'aquest fenomen és la taxa d'error superior de les lectures de MinION, ja que en l'assemblatge híbrid es troben quasi tots els gens. Per això és important generar diversos assemblatges i comparar-los entre si.

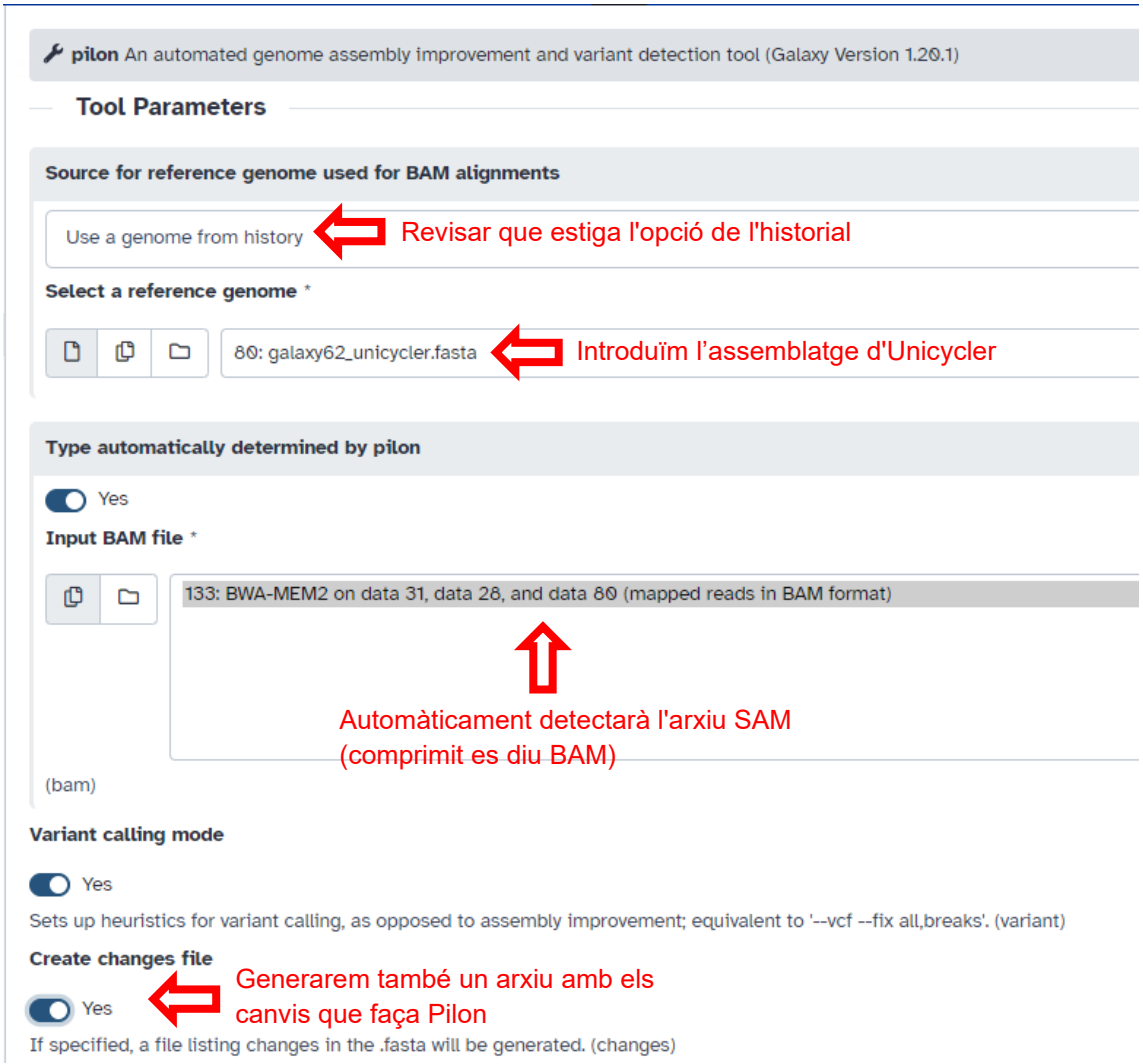
Ara ja comptem amb prou informació per a decidir quin és el millor assemblatge de tots els generats. El que menys errors té fins i tot mantenint una bona contigüitat és l'híbrid amb Unicycler, que serà el que usarem a partir d'ara.

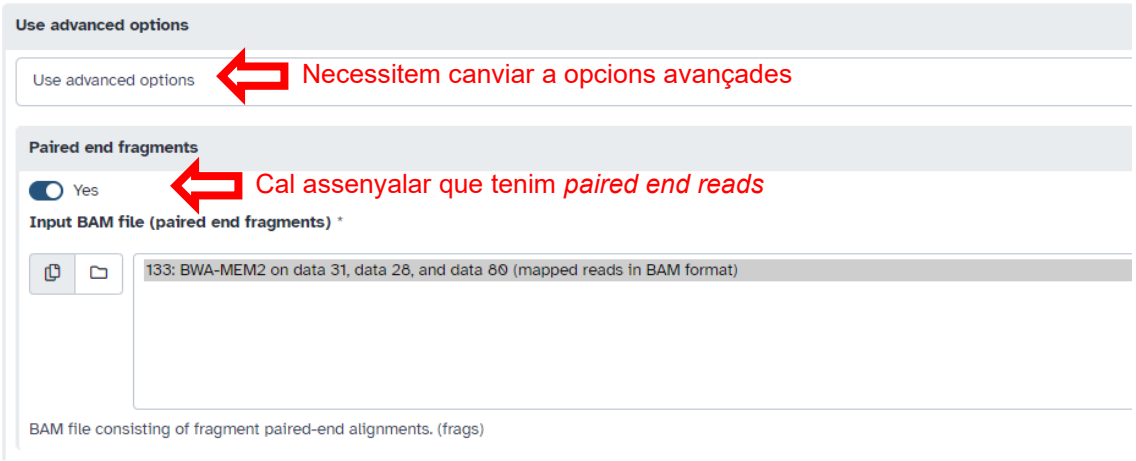
Pràctica 6: Poliment d'assemblatges

Com ja hem comentat, el fet d'usar lectures llargues implica afegir-hi potencials errors d'assemblatge. Per a això, *polirem* el nostre assemblatge amb Pilon. Per tant, el primer que farem serà fer el mapatge de les nostres lectures d'Illumina davant de l'assemblatge amb BWA-MEM2.



Una vegada tenim l'arxiu SAM, podem anar-nos a Pilon.





Amb totes les opcions canviades, fem servir l'eina, la qual ens generarà tres arxius. El primer és un arxiu que es denomina *variant calling file* (VCF), però aquest representa tot el genoma i tenim un resum més concís en l'arxiu de canvis en FASTA, que és aquest:

```
1:141962 1_pilon:141962 C T
1:142020 1_pilon:142020 C T
1:328482 1_pilon:328482 C T
1:328647 1_pilon:328647 C T
1:328700 1_pilon:328700 A T
1:328749 1_pilon:328749 A G
1:328767 1_pilon:328767 T G
1:328770 1_pilon:328770 G C
1:328797 1_pilon:328797 T A
1:328810 1_pilon:328810 A G
1:328812 1_pilon:328812 A G
1:328884 1_pilon:328884 C A
1:328890 1_pilon:328890 G A
1:328894 1_pilon:328894 C T
1:328899 1_pilon:328899 T C
1:328932 1_pilon:328932 T A
1:328951 1_pilon:328951 T C
1:329025 1_pilon:329025 A G
1:329073 1_pilon:329073 C T
1:329143 1_pilon:329143 T C
1:329197 1_pilon:329197 C T
1:329248 1_pilon:329248 C T
1:329265 1_pilon:329265 C T
1:329266 1_pilon:329266 C A
1:329275 1_pilon:329275 T G
1:329371 1_pilon:329371 A T
1:329396 1_pilon:329396 A G
1:329398 1_pilon:329398 T C
1:329416 1_pilon:329416 C T
1:511741 1_pilon:511741 C T
1:633957-633958 1_pilon:633957 TT .
1:1324986-1324987 1_pilon:1324984 TT .
1:1675262 1_pilon:1675258 A .
1:1855366 1_pilon:1855361 A C
1:1864930 1_pilon:1864925 A .
1:2071713 1_pilon:2071707 T C
1:2071755 1_pilon:2071749 G A
1:2071962 1_pilon:2071956 A G
1:2072327 1_pilon:2072321 C T
1:2072395 1_pilon:2072389 C T
1:2072448 1_pilon:2072442 A G
1:2072472 1_pilon:2072466 G A
1:2782359 1_pilon:2782353 A .
1:2786541-2786542 1_pilon:2786534 AA .
1:2824881 1_pilon:2824872 A G
1:2825106 1_pilon:2825097 G A
1:2905919 1_pilon:2905910 G A
2:465816 2_pilon:465816 G .
2:567035 2_pilon:567034 A G
2:1636444 2_pilon:1636443 A C
3:144041-144055 3_pilon:144041-144055 GCCGCTATTAACGAC AGTGCTATTAACACT
3:166777 3_pilon:166777 . A
3:250057 3_pilon:250058 G C
4:97803 4_pilon:97803 G A
```

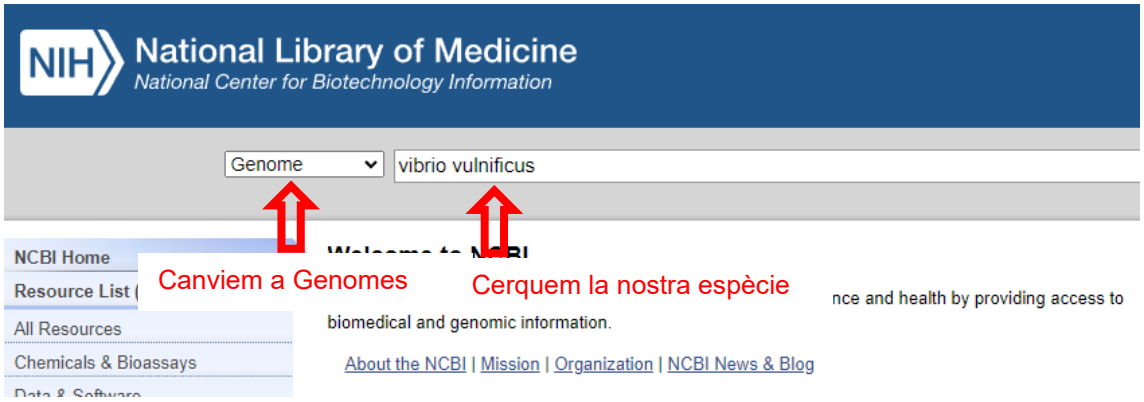
Per ordre, el primer número assenyala el *còntig* on s'ha produït el canvi, seguit de dos punts, la posició, el nou nom que té el *còntig* i el canvi (primer l'estat inicial i després el canvi). Si ens hi fixem, en la seua majoria són canvis menors, però el primer canvi del *còntig* 3 sí que és un canvi major. Aquests canvis suposen que de noves lectures es puguen fer mapatges sobre el genoma i porten, a més, canvis. Per això, es solen fer diverses rondes de Pilon fins que només hi ha un nombre mínim de canvis. Per a aquest genoma, si fem una nova ronda, observarem que no hi ha canvis, cosa per la qual ja tenim el nostre assemblatge polit i a punt per a ser anotat. Però abans, farem un últim control de qualitat.

Exercici opcional: Fer un poliment de l'assemblatge amb Flye i, posteriorment, comparar els valors de BUSCO per a veure si milloren.

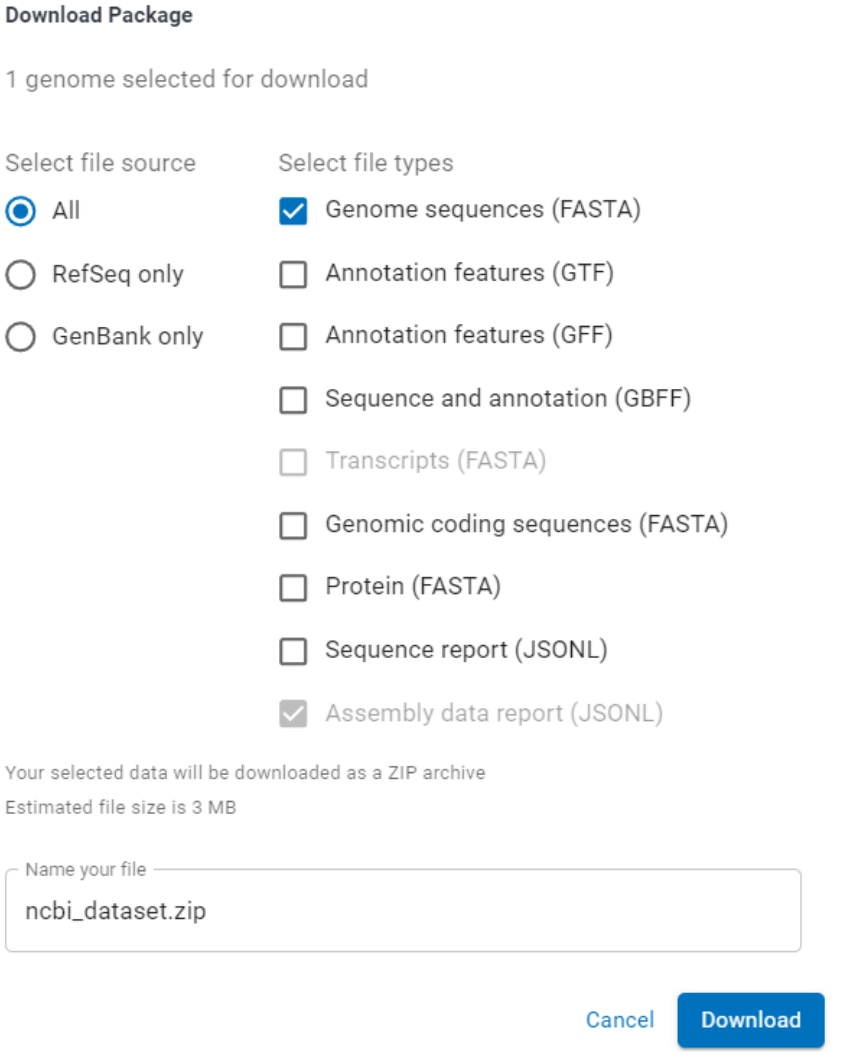
Pràctica 7: Identitat Nucleotídica Mitjana (ANI)

Per a verificar que la nostra soca efectivament correspon a una soca de *V. vulnificus*, compararem el genoma de la soca Vv5 amb uns altres de l'espècie. Després, calcularem la identitat nucleotídica mitjana (ANI, per les seues sigles en anglès) i comprovarem si és superior al 95 %, el límit estandarditzat per a la diferenciació d'espècies [21]. En primer lloc, ens descarregarem el nostre genoma polit.

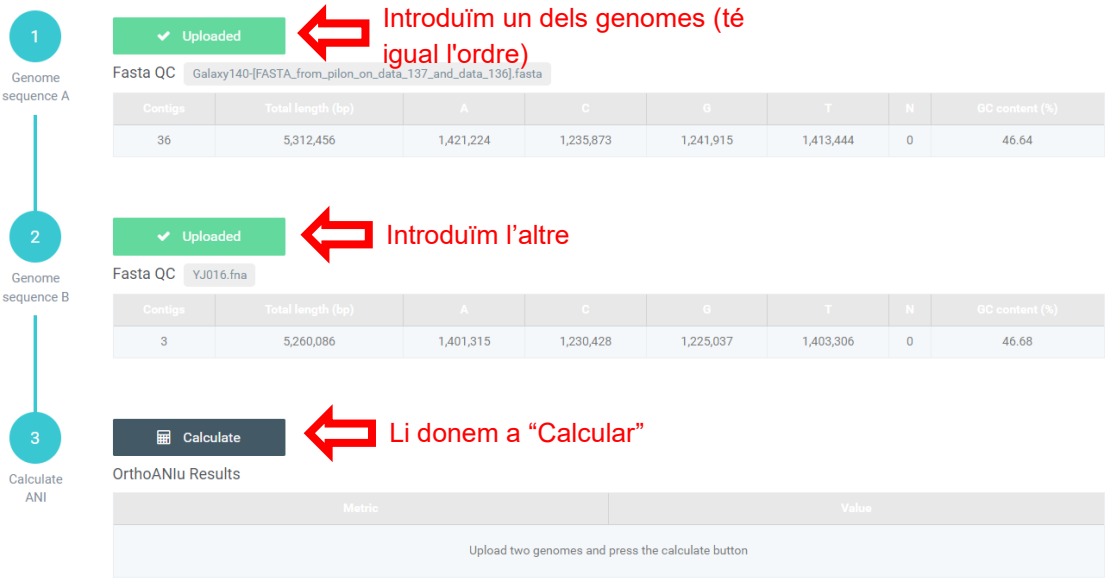
Posteriorment, anirem al web de l'NCBI (https://www.ncbi.nlm.nih.gov/) i descarregarem un genoma depositat de l'espècie.




Vibrio vulnificus
Vibrio vulnificus is a species of g-proteobacteria in the family Vibrionaceae.
Browse taxonomy
Current scientific name
Vibrio vulnificus
Taxonomic rank
species
NCBI Taxonomy ID
672
Genome
Browse all 2,222 genomes
Visualitzem els genomes disponibles
Assembly
GenBank
RefSeq
Scientific name
Modifier
Annotation
Action
ASM974v1
En seleccionem un
Vibrio vulnificus YJ016
YJ016 (strain)
ena-yuan-GCF_000009745.1
GCA_963860095.1
Vibrio vulnificus YJ016
Descarreguem
View details
Download


Download Package
1 genome selected for download
Select file source
All
RefSeq only
GenBank only
Select file types
Genome sequences (FASTA)
Annotation features (GTF)
Annotation features (GFF)
Sequence and annotation (GBFF)
Transcripts (FASTA)
Genomic coding sequences (FASTA)
Protein (FASTA)
Sequence report (JSONL)
Assembly data report (JSONL)
Your selected data will be downloaded as a ZIP archive
Estimated file size is 3 MB
Name your file
ncbi_dataset.zip
Cancel
Download

En l'última finestra podrem afegir arxius d'anotació en diferents formats, per a la nostra anàlisi només necessitarem l'arxiu FASTA (que tindrà extensió ".fna". Ho descomprimim i ens anirem al web de calculadora d'ANI en línia: https://www.ezbiocloud.net/tools/ani



Després d'una estona, n'obtindrem açò:

Metric	Value
OrthoANIu value (%)	98.08
Genome A length (bp)	5,298,900
Genome B length (bp)	5,258,100
Average aligned length (bp)	3,435,194
Genome A coverage (%)	64.83
Genome B coverage (%)	65.33

El més important d'aquesta taula és la primera fila, la qual ens diu el valor de l'ANI. En aquest cas és de 98,08 (> 95 %), cosa per la qual podem assegurar que la soca Vv5 efectivament és *V. vulnificus*.

Pràctica 8: Anotació

Després d'aquest breu respir de Galaxy, podem tornar amb l'anotació. L'anotació consisteix en l'acotació de les regions codificants en el nostre genoma. Hi ha diversos programes per a anotar el nostre genoma. Prokka [22] és considerat un dels anotadors estàndard en genomes procariotes. No obstant això, des de 2019 va deixar d'actualitzar-se i el mateix autor recomana ara usar Bakta [23]. Les bases per a l'assignació de regions en Bakta són aquestes:

- Identificar les fraccions del genoma que no codifiquen per a proteïnes (per exemple: tRNAs o CRISPR *arrays*).
- Identificar els elements genètics en el genoma (principalment la predicció de gens).
- Adjuntar la informació biològica disponible a aquests elements, buscant seqüències homòlogues en bases de dades.

Aquest nou anotador té com a afegit la busca d'homòlegs en bases de dades tan importants com ara CARD (The Comprehensive Antibiotic Resistance Database) per a trobar gens de resistència a antibiòtics o VFDB (Virulence Factor Database) per a detectar gens de virulència entre moltes altres. A més, ambdues estan en Galaxy!

L'anotació serà tan senzilla com indicar que assemblatge usarem (el poliment amb Pilon) i, opcionalment, podem afegir el nom del gènere, espècie, soca… Això serà necessari per a depositar els arxius en bases de dades públiques (com ara el NCBI). També pot ser interessant obtindre un arxiu resumen en TXT i, sobretot, un arxiu amb format GenBank, perquè molts programes de visualització de genomes (com SnapGeneViewer) ho necessiten.

L'eixida d'aquest programa és la següent:

El resum de l'anotació ens ofereix aquesta informació:

```
Sequence(s):
Length: 5312456
Count: 36
GC: 46.6
N50: 2910046
N ratio: 0.0
coding density: 87.5

Annotation:
tRNAs: 103
tmRNAs: 1
rRNAs: 20
ncRNAs: 32
ncRNA regions: 28
CRISPR arrays: 1
CDSs: 4699
pseudogenes: 5
hypotheticals: 99
signal peptides: 0
sORFs: 2
gaps: 0
oriCs: 2
oriVs: 0
oriTs: 0

Bakta:
Software: v1.9.2
Database: v5.0, full
DOI: 10.1099/mgen.0.000685
URL: github.com/oschwengers/bakta
```

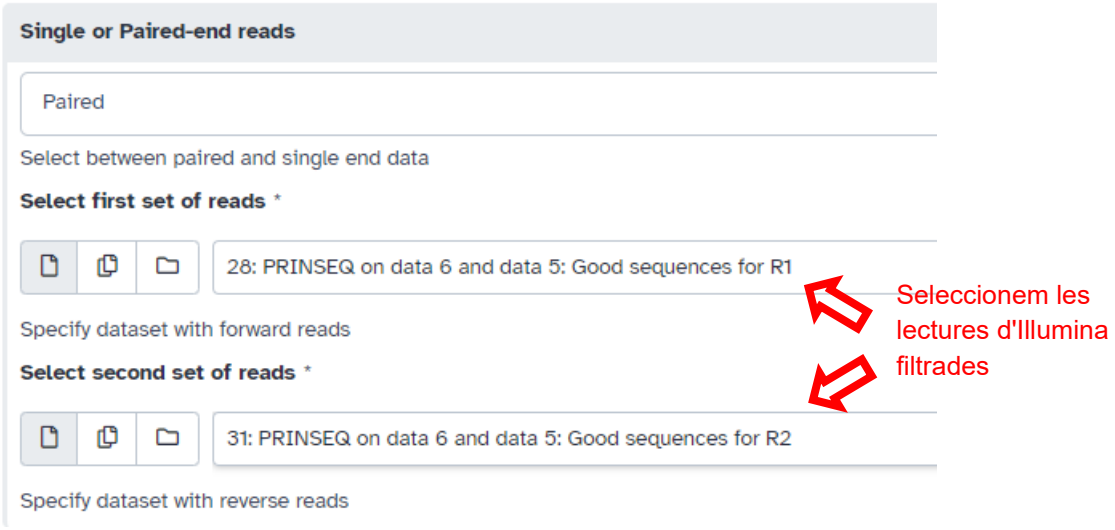
Amb això ja tindríem el nostre genoma anotat per a futures anàlisis!

Exercici opcional: Reviseu els diferents arxius d'eixida de Bakta. La imatge que genera té quatre anells interiors, en ordre representen un GC *skew* (en blau i taronja), el contingut en GC (en verd i roig), els diferents CDS (*Coding Sequences*) de la cadena negativa i els de la cadena positiva. Quina rellevància biològica té eixe biaix d'unes certes regions en el GC *skew*?

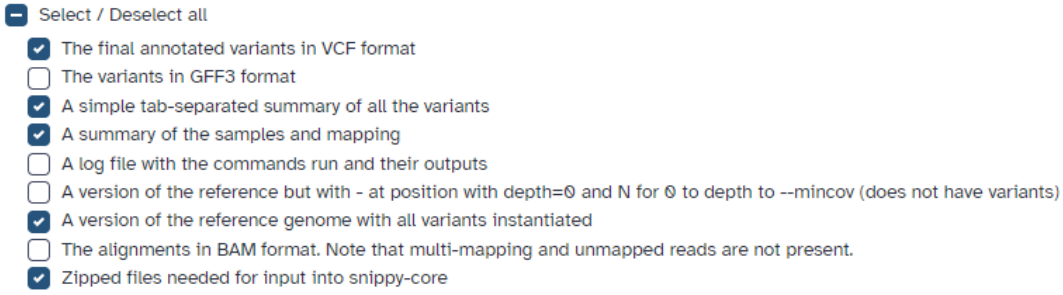
Pràctica opcional: Mapatge

Tots els passos explicats fins ara han sigut per a obtindre un genoma *de novo*, és a dir, del qual no tenim cap referència. Una altra opció és obtindre-lo a partir d'un mapatge de les nostres lectures contra un genoma de referència. Aquest procés pot ser més ràpid, però implica una sèrie de desavantatges; per exemple, la pèrdua d'informació continguda en el genoma seqüenciat que no apareix en el que usem de referència (des d'uns certs gens concrets fins a fags, plasmidis o altres elements genètics mòbils). No obstant això, pot ser una forma ràpida d'obtindre diferències en regions comunes si volguérem, per exemple, fer una filogènia després. De manera clàssica, el mapatge es sol fer amb BWA-MEM (com hem fet per al poliment), seguit d'un *anomenat de variants* mitjançant GATK o Freebayes. Per sort, comptem amb una eina que automatitza tot el procés anomenada Snippy (https://github.com/tseemann/snippy).

Una vegada seleccionem l'arxiu del nostre ordinador, ens eixirà açò:



Podem marcar els arxius d'eixida que vulguem, però els interessants serien el resum de les variants trobades i la versió del genoma de referència amb les variants incloses. Si tinguérem diversos genomes, podríem trobar fàcilment totes les variants amb snippy-core, per a fer anàlisis filogenètiques, per exemple. També és útil consultar la posició de les variants en l'arxiu VCF per a veure quins gens han canviat entre la nostra referència i el genoma seqüenciat.



Finalment, donem una ullada al resum que ens genera Snippy:

Column 1	Column 2
ReferenceSize	5260086
Software	snippy 4.6.0
Variant-COMPLEX	7692
Variant-DEL	233
Variant-INS	229
Variant-MNP	771
Variant-SNP	49472
VariantTotal	58397

Amb això acabem la pràctica! Una vegada tenim el nostre(s) genoma(es) ja podem fer les anàlisis que necessitem, des d'estudiar la presència de gens de resistència a antibiòtics fins a comprendre millor l'evolució i disseminació de microorganismes patògens. Però això queda ja a les vostres mans. Les possibilitats són infinites!

Resum

Ara que ja sabem els diferents tipus d'arxius amb els quals solem treballar en bioinformàtica, farem una recapitulació de cara a possibles projectes futurs de seqüenciació bacterians.

0. Triem el mètode de seqüenciació d'ADN. En el nostre cas, optem tant per lectura curta com per llarga, amb la finalitat de tindre un genoma tancat i poliment. Això dependrà principalment dels recursos disponibles.

1. Obtenim les lectures o dades crues que ens genere el seqüenciador. Toca estudiar la seua qualitat, a partir de la seua visualització en un gràfic de qualitat/base. Per a això, utilitzem una eina bioinformàtica anomenada FastQC, la qual analitza la qualitat dels arxius FASTQ.
   Si considerem que la qualitat global de les lectures pot millorar, podem filtrar les dades crues aplicant diferents criteris:
   - Qualitat: ens quedem amb les lectures amb qualitat mínima de 25, per exemple.
   - *Trimejat* dels extrems (o només de la regió final de les lectures).
   - Longitud: d'aquelles que no arriben a una certa grandària.

   Hi ha diferents eines per a aquest pas, com ara AfterQC, PrinSeq…
   Després del filtrat i/o *trimejat*, confirmem que la qualitat haja millorat.

2. Assemblatges; obtenim arxius FASTA. Disposem de nombroses eines per als assemblatges de les lectures, amb l'objectiu d'intentar tenir un genoma final òptim, si pot ser tancat (això és, tants *còntigs* o encapçalats+seqüència en el FASTA com a cromosomes/plasmidis tinga la soca en estudi). Enumerem algunes d'aquestes eines:
   - SPAdes: per a lectures curtes.
   - Unicycler: per a lectures curtes, lectures llargues o combinació de les dues (açò s'anomena assemblatge híbrid).
   - Flye: per a lectures llargues.
   - HGAP: per a lectures llargues (del sistema de seqüenciació PacBio).

3. Visualització dels assemblatges. Utilitzem l'eina Bandage, la qual representa els GRAFS de De Bruijn de l'assemblatge pròpiament dit. Aquests GRAFS són fitxers amb més informació que la continguda en els FASTA, la qual ajuda a representar en un visualitzador els *còntigs* del genoma seqüenciat.

4. Avaluació dels assemblatges. A partir de diferents estadístiques obtingudes amb els programes Quast i BUSCO, decidirem quina aproximació és la millor.

5. Poliment del genoma. Obtindrem un genoma de millor qualitat sense els errors que puguen introduir les lectures llargues. Alguns assembladors (com ara HGAP) el porten de sèrie, però una molt comuna és Pilon.

6. Anotació del genoma. Hi ha diferents eines bioinformàtiques que, basades en bases de dades pròpies, aconsegueixen anotar els gens que va trobant al llarg de l'arxiu FASTA que tenim. Exemple d'anotadors tenim: Prokka, Bakta…

7. Finalment, amb l'arxiu FASTA del genoma assemblat podem jugar lliurement i utilitzar-lo per als nostres estudis, com ara a filogènia, genòmica comparada, busca de pròfags, etc.

Referències

[1] T. Namiki, T. Hachiya, H. Tanaka, and Y. Sakakibara, 'MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads', *Nucleic Acids Res.*, vol. 40, no. 20, p. e155, Nov. 2012, doi: 10.1093/nar/gks678.

[2] B. Hesper and P. Hogeweg, 'Bio-informatics: a working concept. A translation of "Bio-informatica: een werkconcept" by B. Hesper and P. Hogeweg', Nov. 23, 2021, *arXiv*: arXiv:2111.11832. doi: 10.48550/arXiv.2111.11832.

[3] E. V. Koonin, K. S. Makarova, and Y. I. Wolf, 'Evolution of microbial genomics: conceptual shifts over a quarter century', *Trends Microbiol.*, vol. 29, no. 7, pp. 582–592, 2021.

[4] E. J. Topol, 'High-performance medicine: the convergence of human and artificial intelligence', *Nat. Med.*, vol. 25, no. 1, pp. 44–56, Jan. 2019, doi: 10.1038/s41591-018-0300-7.

[5] M. Á. García-Sevillano, T. García-Barrera, and J. L. Gómez-Ariza, 'Environmental metabolomics: Biological markers for metal toxicity', *ELECTROPHORESIS*, vol. 36, no. 18, pp. 2348–2365, 2015, doi: 10.1002/elps.201500052.

[6] E. V. Koonin and Y. I. Wolf, 'Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world', *Nucleic Acids Res.*, vol. 36, no. 21, pp. 6688–6719, 2008.

[7] B. M. Perez-Sepulveda *et al.*, 'An accessible, efficient and global approach for the large-scale sequencing of bacterial genomes', *Genome Biol.*, vol. 22, pp. 1–18, 2021.

[8] E. Lander *et al.*, 'Initial sequencing and analysis of the human genome', *Nature*, vol. 409, no. 6822, pp. 860–921, 2001.

[9] J. C. Venter *et al.*, 'The sequence of the human genome', *science*, vol. 291, no. 5507, pp. 1304–1351, 2001.

[10] J. Sohn and J.-W. Nam, 'The present and future of de novo whole-genome assembly', *Brief. Bioinform.*, vol. 19, no. 1, pp. 23–40, 2018.

[11] H. Satam *et al.*, 'Next-generation sequencing technology: Current trends and advancements', *Biology*, vol. 12, no. 7, p. 997, 2023.

[12] H. Carmona-Salido *et al.*, 'Draft Genome Sequences of Vibrio vulnificus Strains Recovered from Moribund Tilapia', *Microbiol. Resour. Announc.*, vol. 10, no. 22, pp. e00094-21, Jun. 2021, doi: 10.1128/MRA.00094-21.

[13] S. Tarazona *et al.*, 'Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package', *Nucleic Acids Res.*, vol. 43, no. 21, p. e140, Dec. 2015, doi: 10.1093/nar/gkv711.

[14] R. Schmieder and R. Edwards, 'Quality control and preprocessing of metagenomic datasets', *Bioinformatics*, vol. 27, no. 6, pp. 863–864, 2011.

[15] W. De Coster and R. Rademakers, 'NanoPack2: population-scale evaluation of long-read sequencing data', *Bioinformatics*, vol. 39, no. 5, p. btad311, 2023.

[16] A. Bankevich *et al.*, 'SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing', *J. Comput. Biol.*, vol. 19, no. 5, pp. 455–477, 2012.

[17] M. Kolmogorov, J. Yuan, Y. Lin, and P. A. Pevzner, 'Assembly of long, error-prone reads using repeat graphs', *Nat. Biotechnol.*, vol. 37, no. 5, pp. 540–546, 2019.

[18] R. R. Wick, L. M. Judd, C. L. Gorrie, and K. E. Holt, 'Unicycler: resolving bacterial genome assemblies from short and long sequencing reads', *PLoS Comput. Biol.*, vol. 13, no. 6, p. e1005595, 2017.

[19] R. R. Wick, M. B. Schultz, J. Zobel, and K. E. Holt, 'Bandage: interactive visualization of de novo genome assemblies', *Bioinformatics*, vol. 31, no. 20, pp. 3350–3352, 2015.

[20] A. Gurevich, V. Saveliev, N. Vyahhi, and G. Tesler, 'QUAST: quality assessment tool for genome assemblies', *Bioinformatics*, vol. 29, no. 8, pp. 1072–1075, 2013.

[21] K. T. Konstantinidis and J. M. Tiedje, 'Genomic insights that advance the species definition for prokaryotes', *Proc. Natl. Acad. Sci.*, vol. 102, no. 7, pp. 2567–2572, 2005.

[22] T. Seemann, 'Prokka: rapid prokaryotic genome annotation', *Bioinformatics*, vol. 30, no. 14, pp. 2068–2069, 2014.

[23] O. Schwengers, L. Jelonek, M. A. Dieckmann, S. Beyvers, J. Blom, and A. Goesmann, 'Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification', *Microb. Genomics*, vol. 7, no. 11, p. 000685, 2021.