PacBio de novo genome assembly

Hands-on

Christophe Klopp
http://bioinfo.genotoul.fr/

Content

- The aims of the hands-on
- File formats
- Software packages (location)
- Data set :
  - Genome coverage
  - Read length histogram
  - File location
- Assembly challenge
- Polishing challenge

The aims of the hands-on

- Produce an assembled genomes from read file(s).
- Produce genomes with :
  - the awaited length,
  - few errors (INDELs).
- Use different software packages to produce the genomes.
- Understand the command line structure and some of the parameters used by the assemblers.
- Polish the genome to have good quality sequence.

File formats

- PacBio reads :
  - *.bas.h5 : reference file for a cell
  - *.bax.h5 : film files , usually 3 per cell
- fastq : read files with quality
- fasta : read or genome file without quality
- cmp.h5 : alignment file for Quiver
- bam : alignment file for Pilon

Where to find the data?

- The files you are going to use in this hands-on are on the virtual disk in the **/opt/byod-data/PacBio** folder

```
.
|-- Corrected_fasta
|   |-- Ecoli_corrected_10x.fasta
|   |-- Ecoli_corrected_15x.fasta
|   |-- Ecoli_corrected_20x.fasta
|   |-- Ecoli_corrected_32x.fasta
|   `-- Ecoli_corrected_5x.fasta
|-- Pilon
|   |-- composition3.fasta
|   |-- composition3.raw.sort.bam
|   `-- composition3.raw.sort.bam.bai
|-- Quiver
|   |-- composition3.cmp.h5
|   |-- composition3.fasta
|   `-- composition3.fasta.fai
|-- Raw
|   |-- Analysis_Results
|   |   |-- E01_1.fasta
|   |   |-- E01_1.fasta.length
|   |   |-- E01_1.fastq
|   |   |-- m141013_011508_sherri_c100709962550000001823135904221533_s1_p0.1.bax.h5
|   |   |-- m141013_011508_sherri_c100709962550000001823135904221533_s1_p0.2.bax.h5
|   |   |-- m141013_011508_sherri_c100709962550000001823135904221533_s1_p0.3.bax.h5
|   |   `-- m141013_011508_sherri_c100709962550000001823135904221533_s1_p0.bas.h5
|   `-- m141013_011508_sherri_c100709962550000001823135904221533_s1_p0.metadata.xml
|-- Raw_fastq
|   |-- E01_1_10x.fastq
|   |-- E01_1_135x.fastq
|   |-- E01_1_20x.fastq
|   |-- E01_1_40x.fastq
|   `-- E01_1_80x.fastq
|-- pilon-1.20.jar
`-- raw.vcf

6 directories, 26 files
```

Software packages

- Three software packages will be used for the assemblies:
  - CANU
  - Falcon
  - Miniasm
- Two software packages for the polishing
  - Quiver
  - CANU
- These packages have been developed by different teams and have different command lines and parameters.
- The implementation and the run time are quite different between packages.

Virtual environments (ve)

- Falcon and Quiver come with a ve.
- When you activate the ve your prompt changes and you get access to new commands

```
(fc-env)[root@vm0004 mydisk]#
```

- To activate the ve you have to source the activate file :
  - `source /ifb/FALCON-integrate/fc-env/bin/activate`
  - `source /ifb/ve-quiver/bin/activate`

CANU and Falcon

- For these two packages, because the read correction process is quite long you will only perform the final assembly (using corrected reads).
- **NB.** You have to find out in the parameters how to tell the software package that you are working with already corrected reads.

Where to find how to use software packages?

- Information about the packages can be found on their reference web sites
  - CANU https://github.com/marbl/canu
  - Falcon https://github.com/PacificBiosciences/FALCON
  - Miniasm https://github.com/lh3/miniasm
  - Quiver : https://github.com/PacificBiosciences/GenomicConsensus
  - Pilon : https://github.com/broadinstitute/pilon
- Other websites presenting how to use the packages exist and can be used.

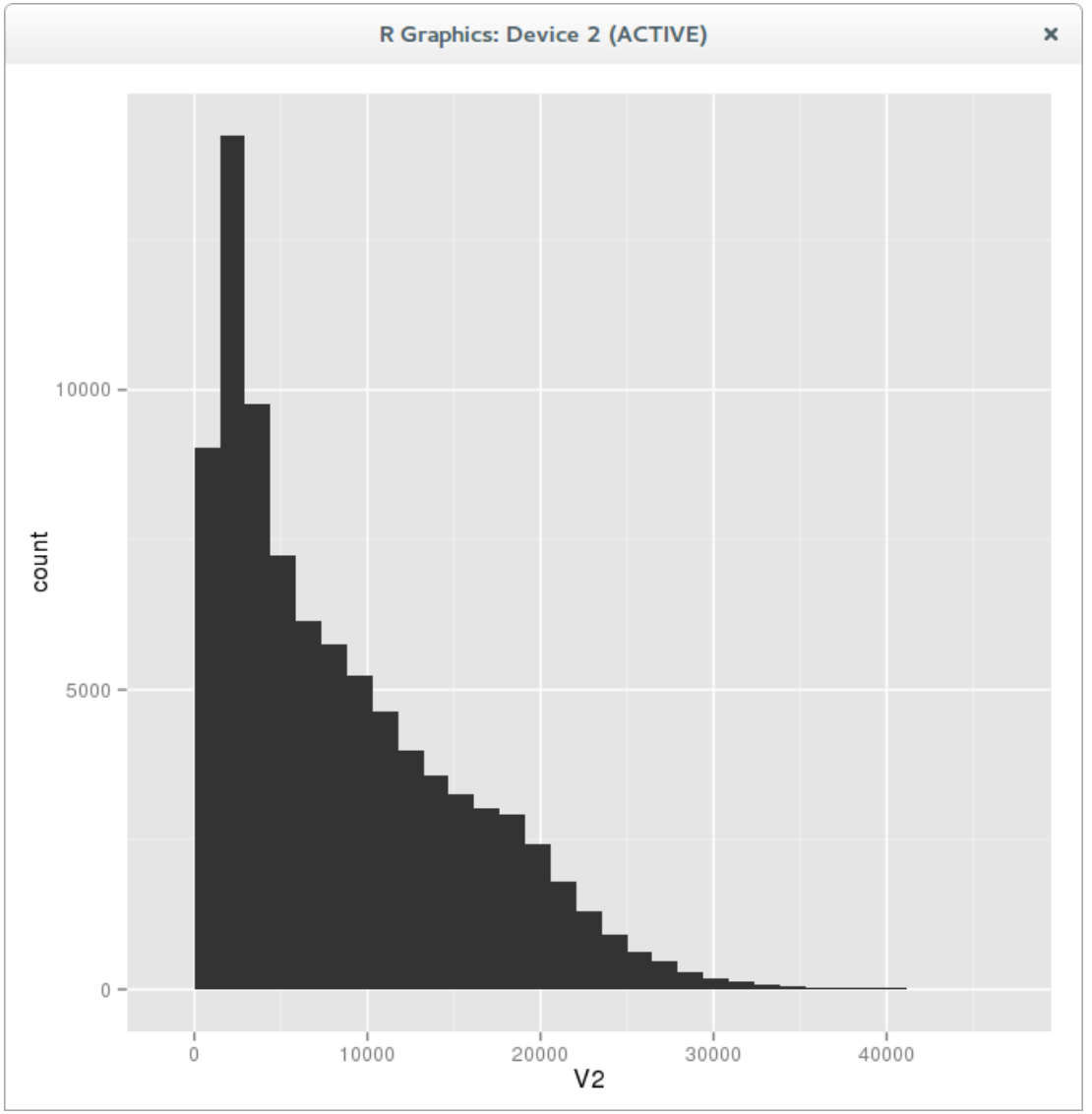
Hands-on data set

- We will use a :
  - public data set
  - made available by PacBio
  - from E coli
  - Which can be found here : https://github.com/PacificBiosciences/DevNet/wiki/Datasets

Genome coverage

- As presented in the course, the genome coverage has an impact on the final assembly.
- In this hands-on we will use different coverages to monitor the induced assembly metrics changes.
- The raw genome coverage of this data set is **135x**.

Read length histogram



Assembly challenge

- 3 teams / 3 colors
  - Red
  - Green
  - Blue
- One table to fill with the assembly results
- One hour to process as many data sets as possible : use all the CPU and memory you can have...
- And the winner is...

The assembly table to fill

miniasm	10x	20x	40x	80x	135x
Nb contigs					
Total length					

CANU	5x	10x	15x	20x	32x
Nb contigs					
Total length					

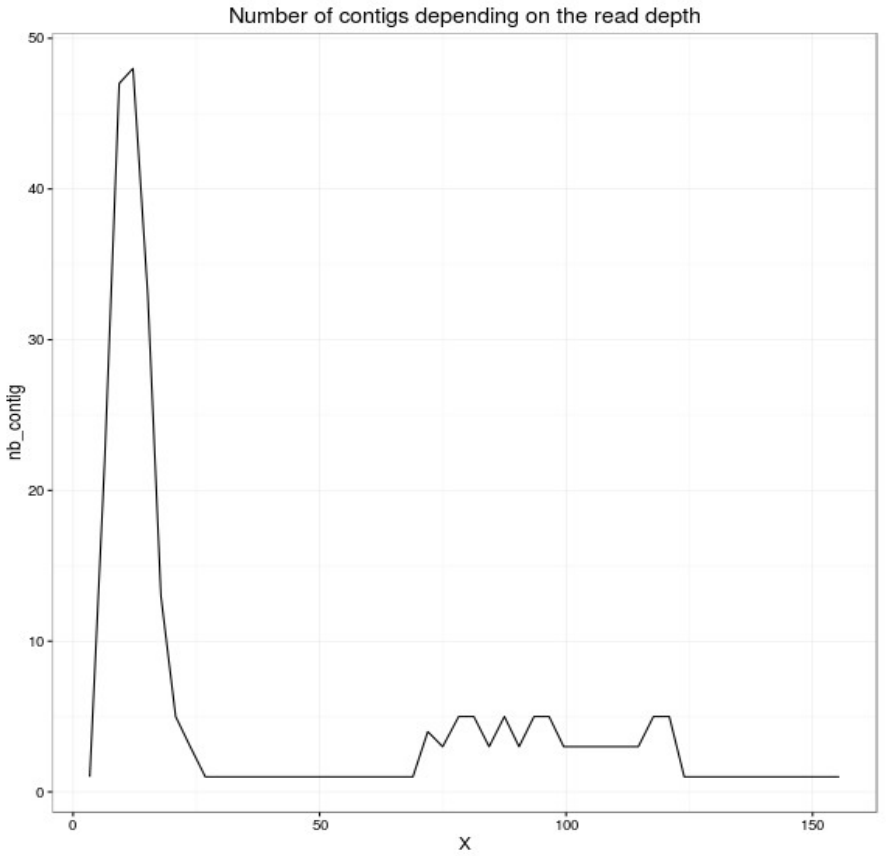
Falcon	5x	10x	15x	20x	32x
Nb contigs					
Total length					

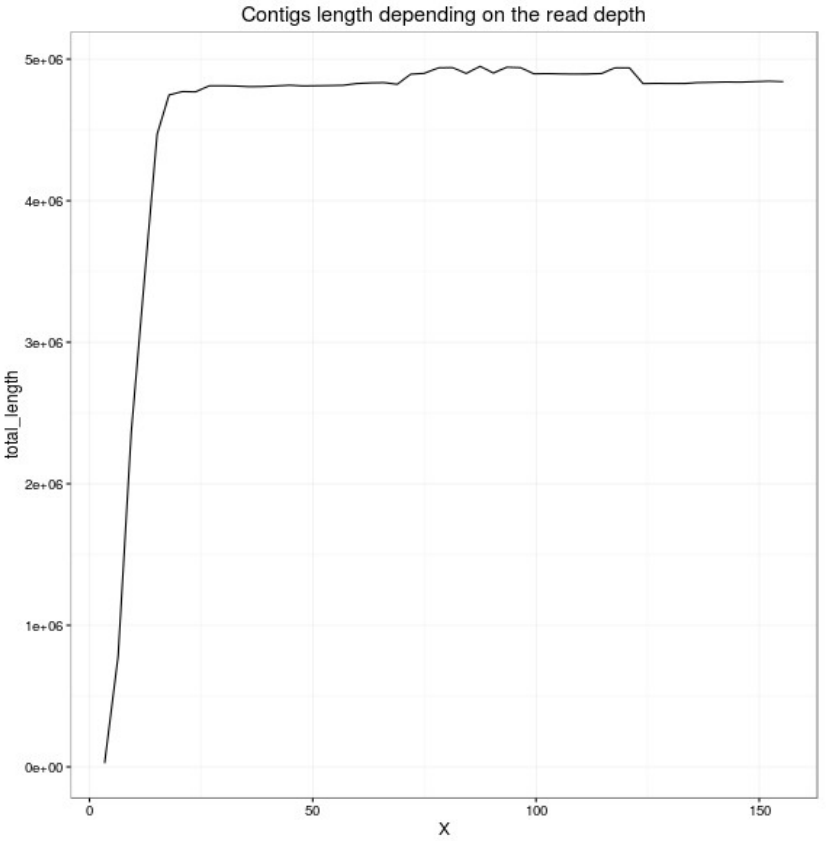
The challenge is running for an hour !

- You can post questions on :

  https://mensuel.framapad.org/p/elixir_ljubljana

Evolution of the assembly metrics





Assemblies made with miniasm.

Polishing challenge

- Once you have assembled the genome the work is partly done...you have to polish your assembly.
- To polish the genome you have to generate an alignment file :
  - With pbalign for Quiver
  - With blasr for pilon

The aim of the challenge

- To produce a polished reference with both software packages
- To see the polishing effect on the different contigs

Polishing dataset

- In the Quiver and Pilon directories you will find the alignment files and the reference genome
- The reference genome is a short version build with contigs coming from different assemblers

Falcon	199,900 bp
CANU1	119,388 bp
CANU2	80,400 bp
Miniasm	207,920 bp

The polishing table to fill

Per contig

	Contig length	Nb insertions	Nb deletions
Quiver			
Pilon			

The challenge is running for an hour !

- You can post questions on : https://semestriel.framapad.org/p/ELIXIR_pacbio_assembly_course