Next Generation Sequencing data analysis with Subread and IGV. Example of a real-world SNP analysis.

Who Am I

Ghislain Bidaut

(ghislain.bidaut@inserm.fr)

IR Cibi Platform (CRCM Integrative Bioinformatics)

Web: http://cibi.marseille.inserm.fr

Forge: http://forgecrcm.marseille.inserm.fr/projects/cibi

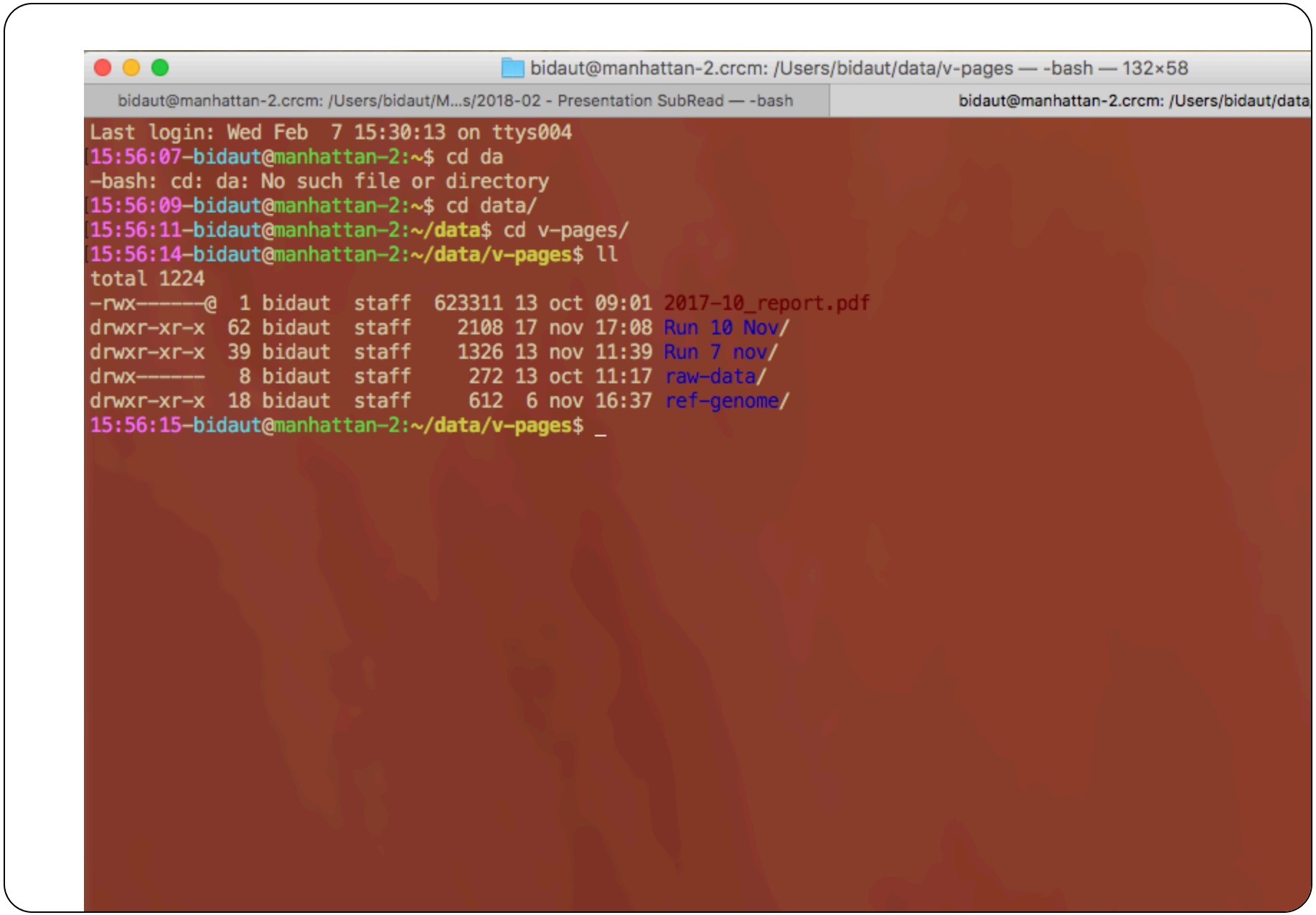
Goal

- SNP detection from NGS data with details on
  - Build a bioinformatics pipeline
  - How to run programs
  - How to interpret results
- This will be illustrated through an NGS data analysis in *E. coli* in order to extract variants.

Talk

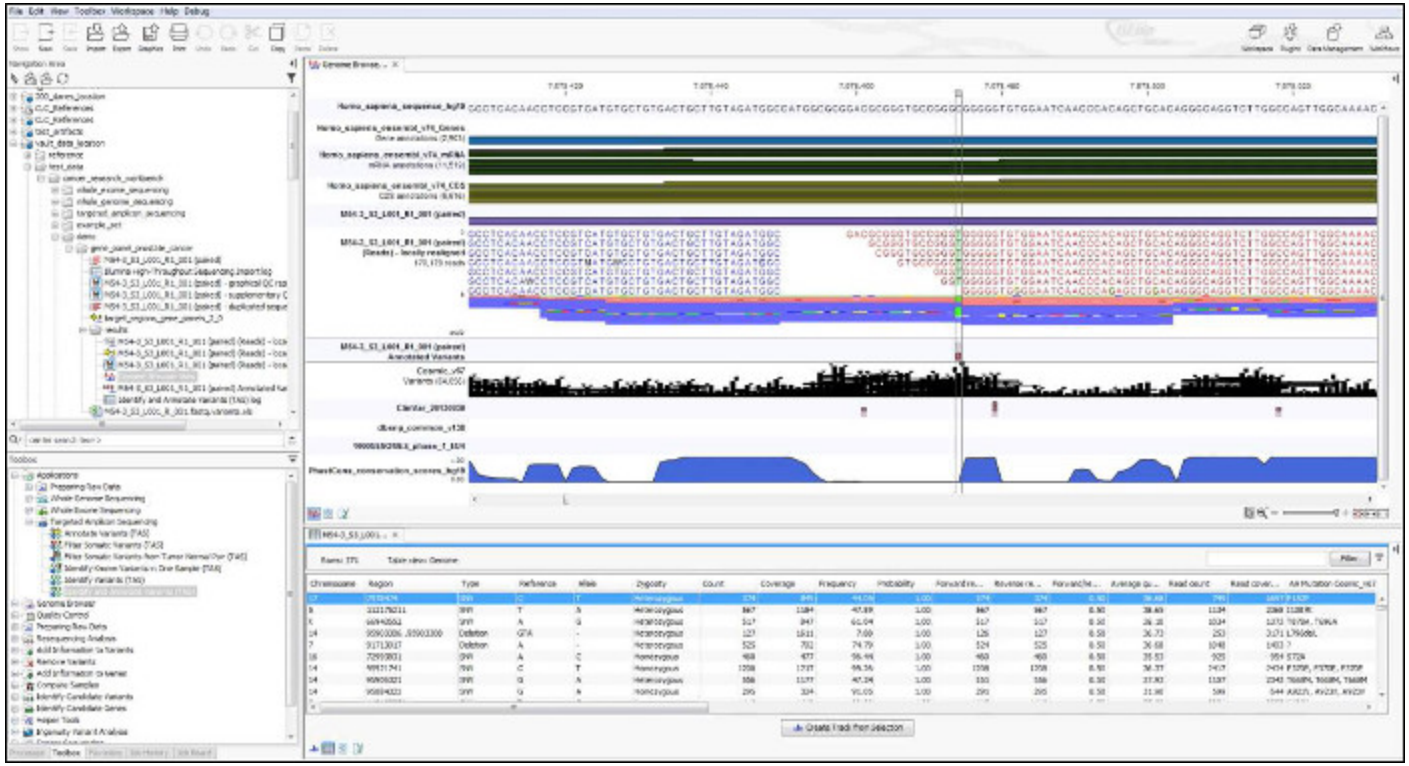
- Prerequisites
  - Have access to a Linux/Mac/Unix machine: (We have many at the CRCM - maybe you are using one *now*)
  - Know how to use Unix command line.
  - Know how access and read the documentation.
- Plan
  - **Install** the necessary programs
  - **Write** a pipeline
  - **Visualize** the data for analysis.

Terminal(Mac) or xterm(Linux)


bidaut@manhattan-2.crcm: /Users/bidaut/data/v-pages — -bash — 132×58
bidaut@manhattan-2.crcm: /Users/bidaut/M...s/2018-02 - Presentation SubRead — -bash
bidaut@manhattan-2.crcm: /Users/bidaut/data
Last login: Wed Feb 7 15:30:13 on ttys004
[15:56:07-bidaut@manhattan-2:~$ cd da
-bash: cd: da: No such file or directory
[15:56:09-bidaut@manhattan-2:~$ cd data/
[15:56:11-bidaut@manhattan-2:~/data$ cd v-pages/
[15:56:14-bidaut@manhattan-2:~/data/v-pages$ ll
total 1224
-rwx------@ 1 bidaut staff 623311 13 oct 09:01 2017-10_report.pdf
drwxr-xr-x 62 bidaut staff 2108 17 nov 17:08 Run 10 Nov/
drwxr-xr-x 39 bidaut staff 1326 13 nov 11:39 Run 7 nov/
drwx------ 8 bidaut staff 272 13 oct 11:17 raw-data/
drwxr-xr-x 18 bidaut staff 612 6 nov 16:37 ref-genome/
15:56:15-bidaut@manhattan-2:~/data/v-pages$ _

Why using command line ?

- Command line is
  - **one-dimensional**
  - **scriptable and documentable**
  - **portable** in most cases
  - **repeatable**
- As opposed to GUI (Graphics User Interface), which is
  - **Two -dimensional**, hence more complex
  - **Not scriptable**, **not easily documentable**
  - **Impossible** to repeat same analysis



The data

Bioinformatics usually start with sequence data from sequencing platform.

In the present case, we are using Fastq files generated by Illumina

We assume QC is done.

```
1 $ cd data/v-pages/raw-data/Clean/151
2 $ ls
3 total 1459888
4 -rwx------  1 bidaut  staff  344158454 13 oct 10:35 FCHL5CWBBXX_L1_ECOijqRAADFAAPEI-97_1.f
5 -rwx------  1 bidaut  staff  403297004 13 oct 10:34 FCHL5CWBBXX_L1_ECOijqRAADFAAPEI-97_2.f
```

SubRead

- Supported, published and documented *all-in-one* package
  - **Fast general purpose aligner**
  - **Read aligner with exon junction detection**
  - **Count features** (For Chip-Seq and RNA-Seq)
  - **SNP detection** with exactSNP
  - Also runs under **R** (RSubread)
  - See subread page http://subread.sourceforge.net/
- Liao Y, Smyth GK and Shi W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Research*, 41(10):e108, 2013
- Liao Y, Smyth GK and Shi W. featureCounts: an efficient general-purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923-30, 2014

Installation with Conda

Conda is a general purpose installer for Linux Mac and windows featuring a bioinformatics oriented repository (Bioconda).

It creates *environnements* in which you can specifies versions of programs that will live in that environment.

Conda will download needed programs from repositories and fix dependancies.

A specific bioinformatics repository exist: **Bioconda**.

- See Conda page at https://conda.io/docs/
- See Bioconda page at https://bioconda.github.io/ and the corresponding Nature article

Creation of an environnement

```
 1 $ conda config --add channels defaults
 2 $ conda config --add channels conda-forge
 3 $ conda config --add channels bioconda
 4
 5 $ conda create --name subreadalign
 6 Fetching package metadata ..............
 7 Solving package specifications:
 8 Package plan for installation in environment /Users/bidaut/anaconda3/envs/subreadalign:
 9
10 Proceed ([y]/n)? yes
11
12 #
13 # To activate this environment, use:
14 # > source activate subreadalign
15 #
16 # To deactivate an active environment, use:
17 # > source deactivate
18 #
```

Enter environment

```
 1 $ source activate subreadalign
 2 (subreadalign) 11:23:28-bidaut@manhattan-2:~$
 3 $ conda list
 4 # packages in environment at /Users/bidaut/anaconda3/envs/subreadalign:
 5 #
 6 $ conda search subread
 7 Fetching package metadata ...............
 8 bioconductor-rsubread        1.22.1                r3.2.2_0  bioconda
 9 1.23.0                  r3.3.1_0  bioconda
10 1.25.2                  r3.3.1_0  bioconda
11 1.25.2                  r3.3.2_0  bioconda
12 1.25.2                  r3.4.1_0  bioconda
13 1.26.1                  r3.4.1_0  bioconda
14 1.28.0                  r3.4.1_0  bioconda
15 subread                      1.5.0p3                      0  bioconda
16 1.5.0                        0  bioconda
17 1.5.0.post3                  0  bioconda
18 1.5.2                        0  bioconda
19 1.5.3                        0  bioconda
20 1.5.3                        1  bioconda
21 1.6.0                        0  bioconda
22 1.6.0                        1  bioconda
23 1.6.0                        2  bioconda
```

Install subread

```
 1 $ conda install subread
 2 Fetching package metadata ...............
 3 Solving package specifications: .
 4
 5 Package plan for installation in environment /Users/bidaut/anaconda3/envs/subreadalign:
 6
 7 The following NEW packages will be INSTALLED:
 8
 9 subread: 1.6.0-2  bioconda
10 zlib:    1.2.11-0 conda-forge
11
12 Proceed ([y]/n)?
13 zlib-1.2.11-0. 100% |################################################################
14 subread-1.6.0- 100% |################################################################
```

Test installation

```
1 $ conda list
2 # packages in environment at /Users/bidaut/anaconda3/envs/subreadalign:
3 #
4 subread                  1.6.0                        2    bioconda
5 zlib                    1.2.11                       0    conda-forge
6 $  subread-align -version
7
8 Subread-align v1.6.0
```

Subread est installé!

Reference genome and annotations

Ref Genome for alignment: ftp://ftp.ensemblgenomes.org/pub/bacteria/release-37/fasta/bacteria_0_collection/escherichia_coli_str_k_12_substr_mg1655/dna/

```
! bash
$ mkdir ref-genome
$ cd ref-genome
$ wget ftp://ftp.ensemblgenomes.org/pub/bacteria/release-37/fasta/bacteria_0_collection/escherichia_co
```

Annotations used for **IGV**: ftp://ftp.ensemblgenomes.org/pub/bacteria/release-37/gff3/bacteria_0_collection/escherichia_coli_str_k_12_substr_mg1655

```
$ wget ftp://ftp.ensemblgenomes.org/pub/bacteria/release-37/gff3/bacteria_0_collection/escherichia_col
```

Index Genome

```
1 $  subread-buildindex ref_genome/Escherichia_coli_str_k_12_substr_mg1655.ASM584v2.dna.chro
```

Alignment and indexing

```
$ cd $(MYFASTADIR)
$ ls
FCHL5CWBBXX_L1_ECOijqRAADFAAPEI-97_1.fq.gz  FCHL5CWBBXX_L1_ECOijqRAADFAAPEI-97_2.fq.gz

$ subread-align -T 20 --sv -i ref_genome/Escherichia_coli_str_k_12_substr_mg1655.ASM584v2.dna.chromoso
$ ls
$ 151.bam
$
$ samtools sort 151.bam > 151.sorted.bam
$ samtools index 151.sorted.bam
$ ls
151.bam         151.sorted.bam       151.sorted.bam.bai
```

SNP detection

```
1 exactSNP -T 20 -f 0.3  -b -i 151.sorted.bam.bai -g ref_genome/Escherichia_coli_str_k_12_su
```

SNP Results

```
 1 $ less 151.vcf
 2 ##fileformat=VCFv4.0
 3 ##comment=The QUAL values for the SNPs in this VCF file are calculated as min(40, - log_1
 4 ##INFO=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
 5 ##INFO=<ID=BGMM,Number=1,Type=Integer,Description="Number of mismatched bases in the back
 6 ##INFO=<ID=BGTOTAL,Number=1,Type=Integer,Description="Total number of bases in the backgr
 7 ##INFO=<ID=MM,Number=1,Type=String,Description="Number of supporting reads for each alter
 8 ##INFO=<ID=INDEL,Number=0,Type=Flag,Description="Indicates that the variant is an INDEL."
 9 ##INFO=<ID=SR,Number=1,Type=Integer,Description="Number of supporting reads (for INDEL on
10 #CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO
11 Chromosome      26333   .       C       T       40.0000 .       DP=93;MM=93;BGTOTAL=936;B
12 Chromosome      363282  .       G       T       40.0000 .       DP=206;MM=87;BGTOTAL=2058
13 Chromosome      363761  .       T       C       40.0000 .       DP=213;MM=99;BGTOTAL=2119
14 Chromosome      366409  .       C       T       2.3222  .       DP=2;MM=2;BGTOTAL=19;BGMM
15 Chromosome      807318  .       G       C       40.0000 .       DP=42;MM=42;BGTOTAL=256;B
16 Chromosome      807830  .       C       T       40.0000 .       DP=107;MM=107;BGTOTAL=107
17 Chromosome      807856  .       A       G       40.0000 .       DP=110;MM=110;BGTOTAL=109
18 Chromosome      1207789 .       C       G       40.0000 .       DP=122;MM=39;BGTOTAL=1147
19 Chromosome      1310570 .       C       T       40.0000 .       DP=143;MM=53;BGTOTAL=1211
20 Chromosome      1310588 .       G       T       40.0000 .       DP=141;MM=79;BGTOTAL=1179
21 Chromosome      2173362 .       CCCG    CG      1.0     .       INDEL;DP=98;SR=196
22 Chromosome      3560455 .       CC      CGC     1.0     .       INDEL;DP=110;SR=221
23 Chromosome      3945914 .       G       A       2.5453  .       DP=2;MM=2;BGTOTAL=25;BGMM
24 Chromosome      4296381 .       CC      CGCC    1.0     .       INDEL;DP=74;SR=146
25 151.filtered.vcf (END)
```

Getting help

```
$ exactSNP
Version 1.6.0

Usage:

./exactSNP [options] -i input -g reference_genome -o output

Required arguments:

-i <file>  Specify name of an input file including read mapping results. The
[-b if BAM] format of input file can be SAM or BAM (-b needs to be specified
if a BAM file is provided).
...
```

IGV visualisation

The **Integrated Genomics Viewer** (IGV) is a genomic browser available at the Broad Institute. It is already installed on Disc servers.

For large data volume analysis, it is recommended to run it from a visualisation server.

The Disc platform Web site give instructions on how to proceed.

IGV Web site: https://software.broadinstitute.org/software/igv/UserGuide

Some Linux Command lines

```
 1 # Create directory
 2 mkdir my_dir
 3 # change dir
 4 cd my_dir
 5 # list files
 6 $ ls
 7 # explore an ascii file
 8 $ less my_file.txt
 9 # seek help on a particular command
10 $ man ls
11 # unzip file
12 $ unzip <my_file.zip>
```

Conclusion

We have detailed variant search step with Subread.

- Initial data: **Fastq** format
- **Download** reference genome (**fa** format)
- **Alignment** with Subread
- **Sort and Index** BAM with samtools
- **SNP detection** with SNP
- **Visualisation** with IGV

Other technologies (RNA-Seq, Chip-Seq) can be processed using the same general analysis.

Going Further

- Help Conda: https://conda.io/docs/user-guide/index.html
- Help Subread:http://bioinf.wehi.edu.au/subread-package/SubreadUsersGuide.pdf
- Other functionalities: Chip-Seq, RNA Seq, etc...
- IGV: https://software.broadinstitute.org/software/igv/UserGuide
- Doc Linux: *man* command. *Presentation created with Python Landslide*
- The data has been generated by Vincent Pagès.

Thanks !