Sequencing technologies

Principal technologies:

454 Life Sciences/Roche

Reads size: 0.5-1kb

Reads nb: ~10^6

Total seq: 0.7 Gb

https://en.wikipedia.org/wiki/DNA_sequencer

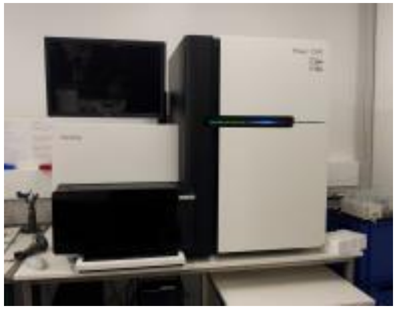
Illumina

Reads size: 2*150

Reads nb: ~$6*10^9$-$20*10^9$

Total seq: 600Gb

https://emea.illumina.com/systems/sequencing-platforms.html

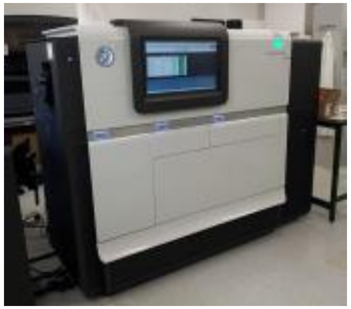


PacBio

Reads size: 30 kb

Total seq: 20Gb

https://www.pacb.com/products-and-services/sequel-system/

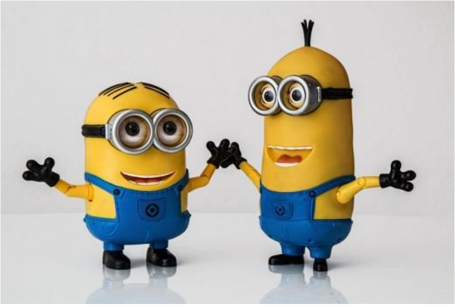


Oxford nanopore

Reads size: 30 kb

Total seq: 15Tb

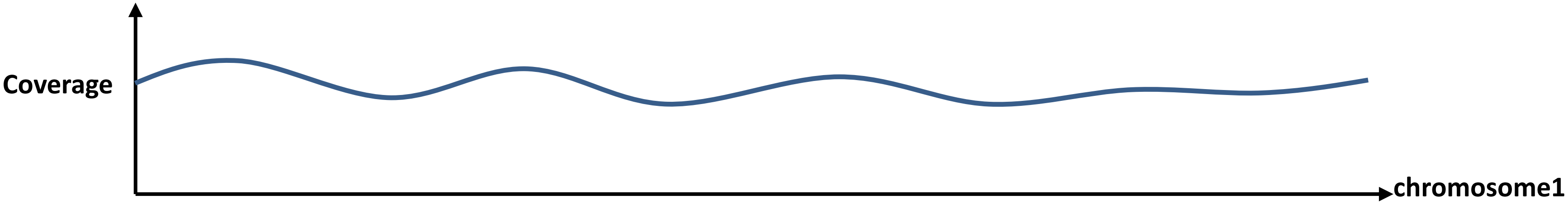
https://nanoporetech.com/products/promethion



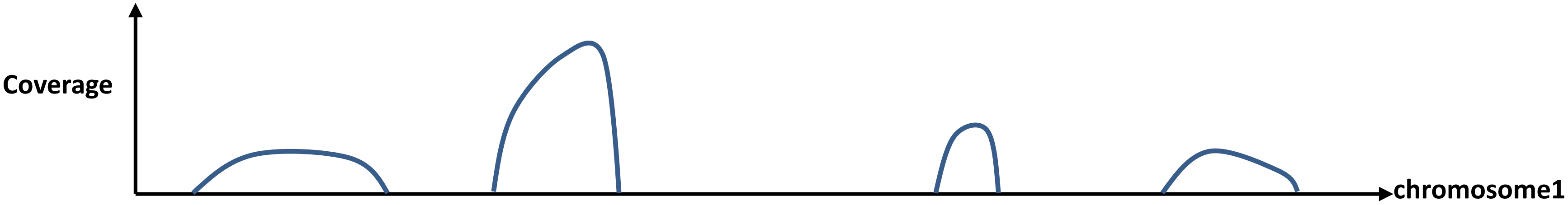
From the output of sequencing to the variant calling file

Principles types of sequencing:

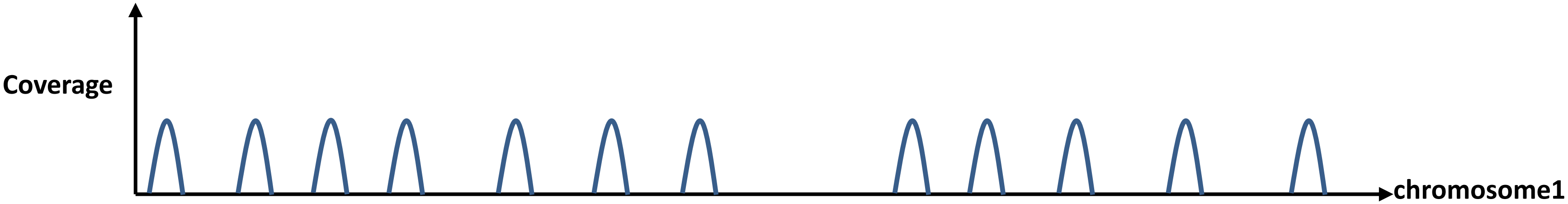
- **W**hole **G**enome **S**equencing (WGS): All the genome is “uniformly” sampled (some biases exist depending to sequencing technologies).



- messenger **RNA seq**uencing (RNAseq): mRNA are sequenced after a step of cDNA complementation

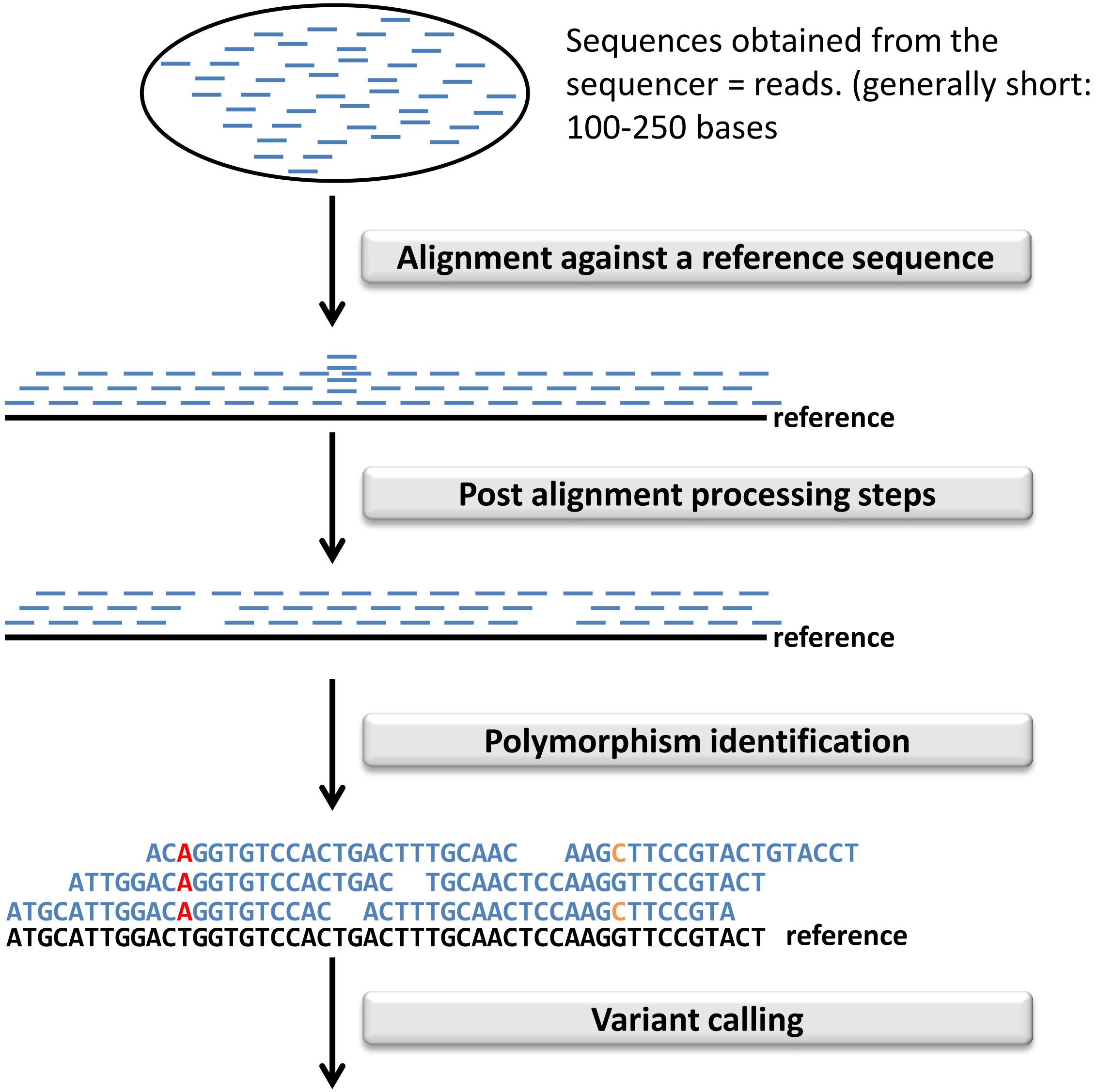


- **G**enotyping **B**y **S**equencing (GBS): The genome is sampled and only part of it is sequenced.



From the output of sequencing to the variant calling file

Standard workflow:



Sequences obtained from the sequencer = reads. (generally short: 100-250 bases

Alignment against a reference sequence

Post alignment processing steps

Polymorphism identification

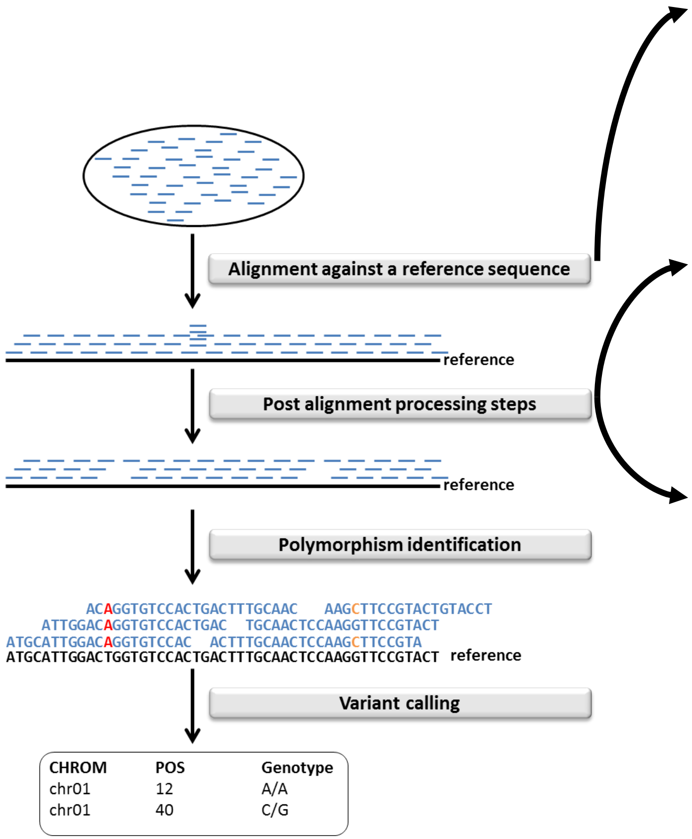
```
         ACAGGTGTCCACTGACTTTGCAAC   AAGCTTCCGTACTGTACCT
    ATTGGACAGGTGTCCACTGAC  TGCAACTCCAAGGTTCCGTACT
ATGCATTGGACAGGTGTCCAC  ACTTTGCAACTCCAAGCTTCCGTA
ATGCATTGGACTGGTGTCCACTGACTTTGCAACTCCAAGGTTCCGTACT  reference
```

Variant calling

CHROM	POS	Genotype
chr01	12	A/A
chr01	40	C/G

From the output of sequencing to the variant calling file

Depending on the sequencing technologies: steps from the sequencing data to the variant calling format are distinct


Alignment against a reference sequence
reference
Post alignment processing steps
reference
Polymorphism identification

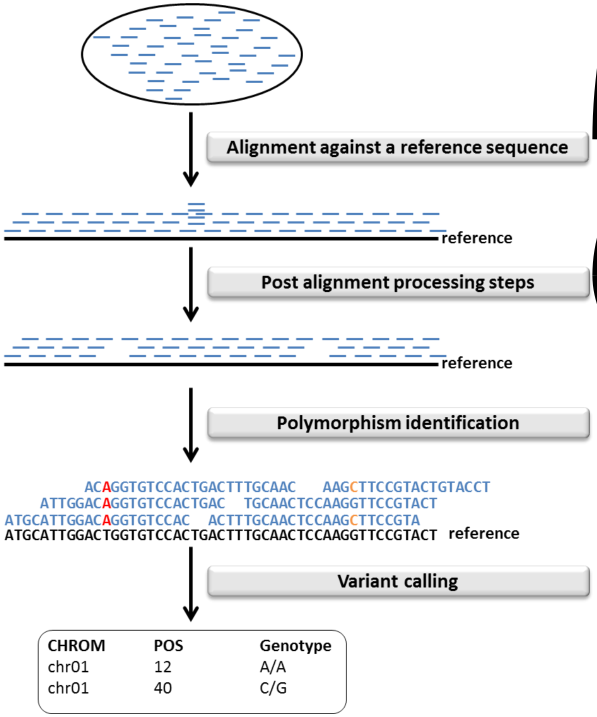
```
          ACAGGTGTCCACTGACTTTGCAAC     AAGCTTCCGTACTGTACCT
     ATTGGACAGGTGTCCACTGAC    TGCAACTCCAAGGTTCCGTACT
ATGCATTGGACAGGTGTCCAC   ACTTTGCAACTCCAAGCTTCCGTA
ATGCATTGGACTGGTGTCCACTGACTTTGCAACTCCAAGGTTCCGTACT  reference
```

CHROM	POS	Genotype
chr01	12	A/A
chr01	40	C/G

- **RNAseq: Aligner should take into account mRNA splicing.**
- **PCR duplicates are usually removed because they biased allelic ratio. It is not possible for GBS du to the approach… (see latter)**
- **RNAseq: Read overlapping splicing sites should be split.**

From the output of sequencing to the variant calling file

Depending on the sequencing technologies: steps from the sequencing data to the variant calling format are distinct



Alignment against a reference sequence

reference

Post alignment processing steps

reference

Polymorphism identification

```
            ACAGGTGTCCACTGACTTTGCAAC     AAGCTTCCGTACTGTACCT
     ATTGGACAGGTGTCCACTGAC   TGCAACTCCAAGGTTCCGTACT
ATGCATTGGACAGGTGTCCAC   ACTTTGCAACTCCAAGCTTCCGTA
ATGCATTGGACTGGTGTCCACTGACTTTGCAACTCCAAGGTTCCGTACT  reference
```

Variant calling

CHROM	POS	Genotype
chr01	12	A/A
chr01	40	C/G

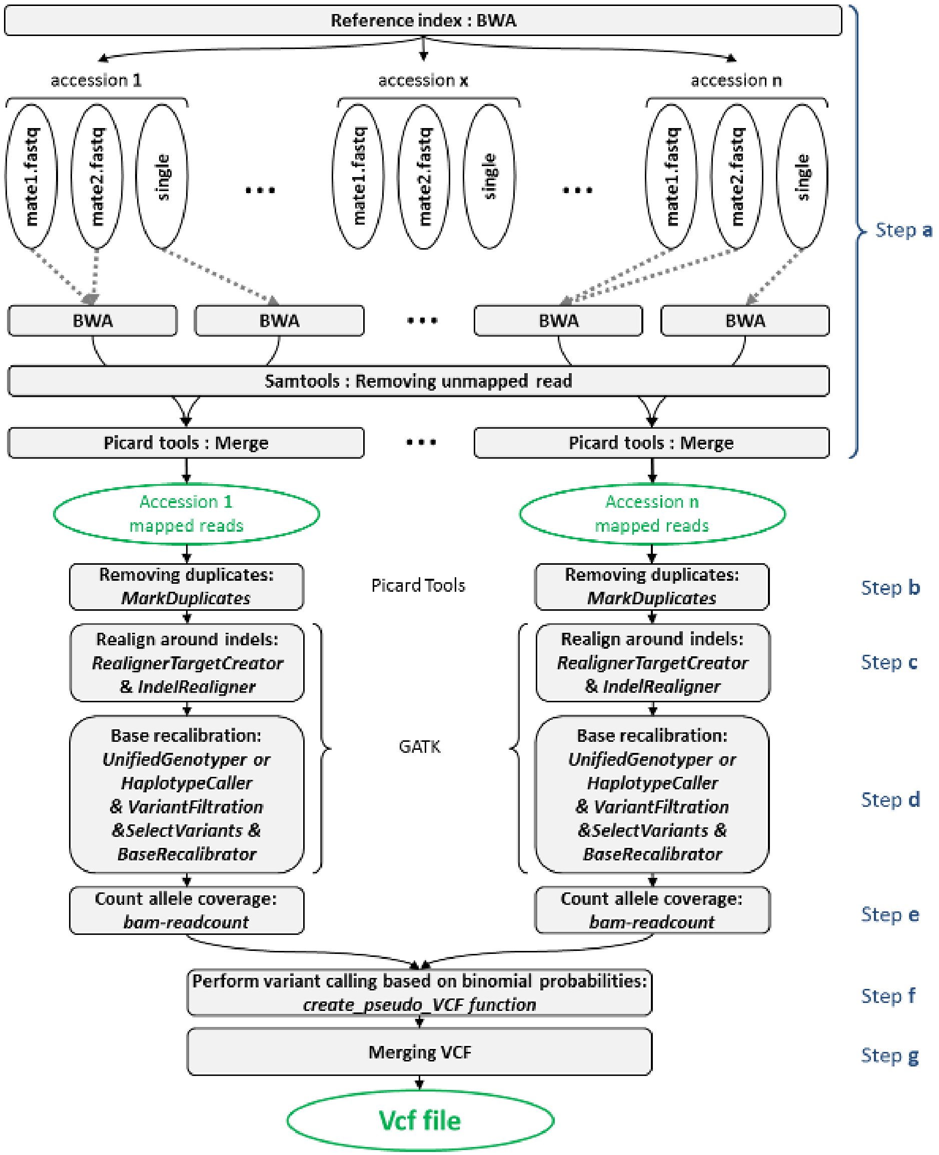
- RNAseq: Aligner should take into account mRNA splicing.
- PCR duplicates are usually removed because they biased allelic ratio. It is not possible for GBS du to the approach… (see latter)
- RNAseq: Read overlapping splicing sites should be split.

Several workflow exists:

- TOGGLe: https://github.com/SouthGreenPlatform/TOGGLE
- GATK best practice: https://software.broadinstitute.org/gatk/best-practices/
- VcfHunter: https://github.com/SouthGreenPlatform/VcfHunter

From the output of sequencing to the variant calling file

VcfHunter detailed workflow (Developped under GenomeHarvest) for WGS and GBS:



Reference index : BWA

accession 1 — mate1.fastq, mate2.fastq, single

…

accession x — mate1.fastq, mate2.fastq, single

…

accession n — mate1.fastq, mate2.fastq, single

BWA — BWA — … — BWA — BWA

Samtools : Removing unmapped read

Picard tools : Merge … Picard tools : Merge

Step a

Accession 1 mapped reads — Accession n mapped reads

Removing duplicates: *MarkDuplicates* (Picard Tools) — Step b

Realign around indels: *RealignerTargetCreator & IndelRealigner* (GATK) — Step c

Base recalibration: *UnifiedGenotyper or HaplotypeCaller & VariantFiltration &SelectVariants & BaseRecalibrator* (GATK) — Step d

Count allele coverage: *bam-readcount* — Step e

Perform variant calling based on binomial probabilities: *create_pseudo_VCF function* — Step f

Merging VCF — Step g

Vcf file

For WGS only

Possible but not recommended:

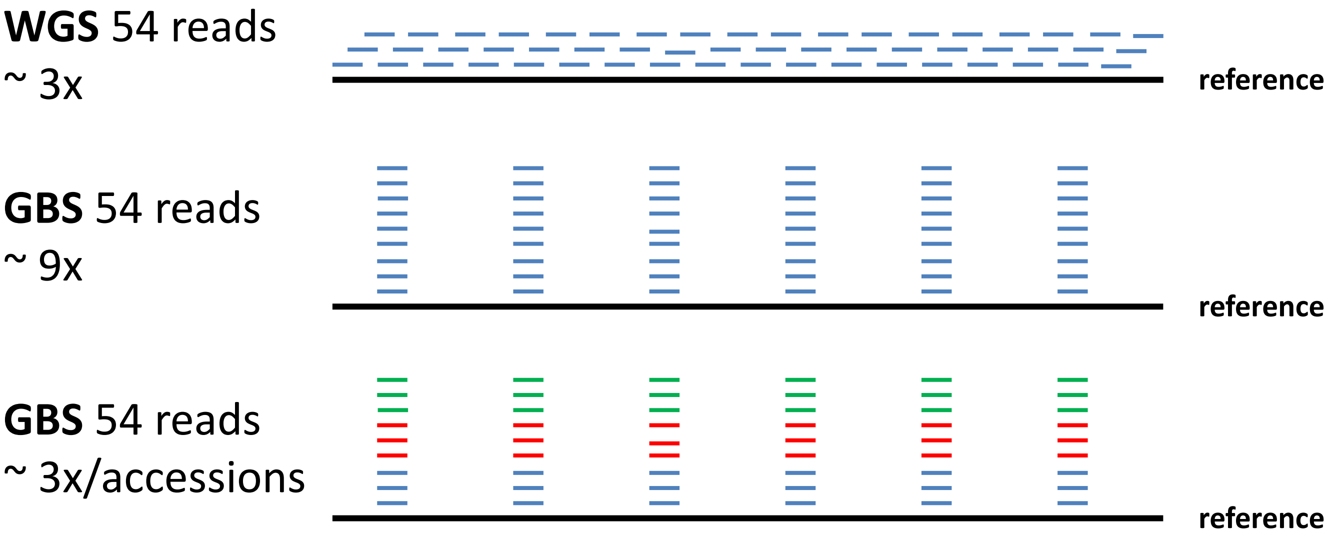
- High computation time
- Result not so good

The Genotyping By Sequencing in detail

- Principle: sequencing a constant part of the genome in several accessions

- Why?
  - ✓ The amount of reads obtained per sequencing run is constant
  - ✓ Necessity to have enough coverage to have a confident genotype calling
  - ✓ Several accessions can be sequenced in one run

Sequencing a sample of the genome which is a constant part ➔ allow to sequence more accessions in a run and to keep the same coverage



WGS 54 reads
~ 3x

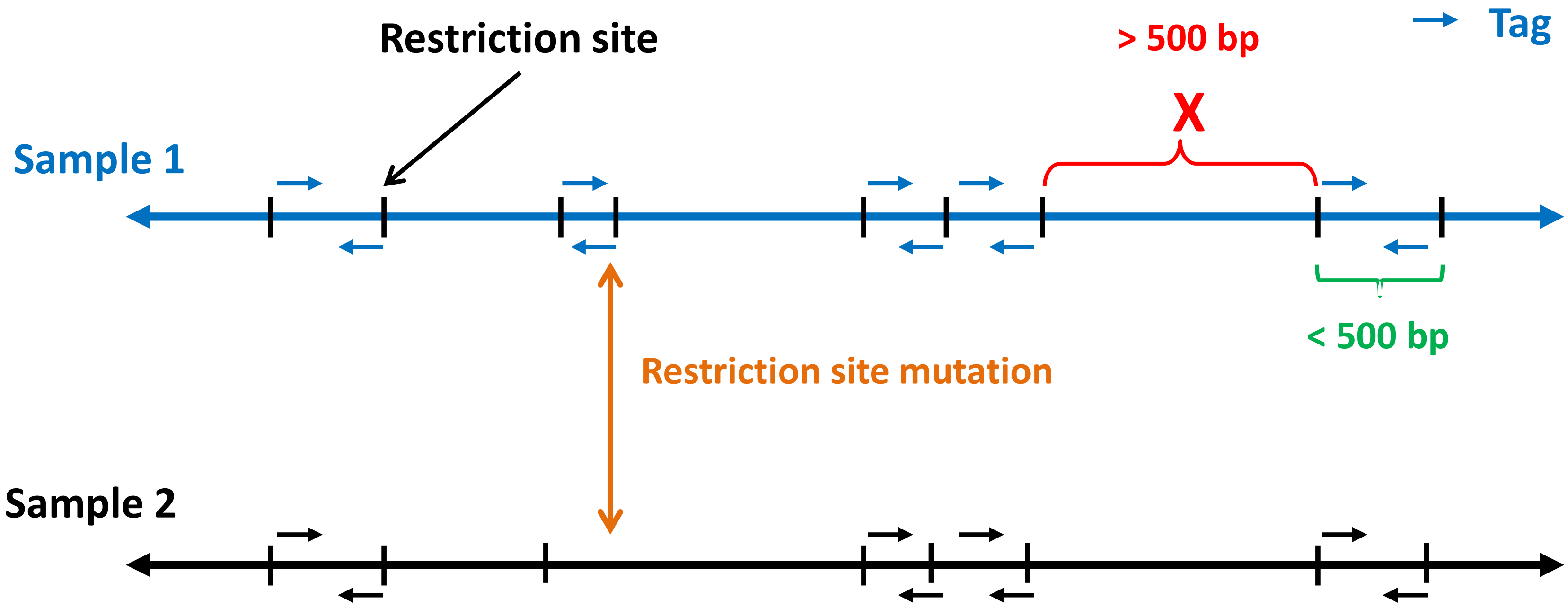
GBS 54 reads
~ 9x

GBS 54 reads
~ 3x/accessions

The Genotyping By Sequencing in detail

Slide adapted from Pierre Mournet

- Cutting the genome with restriction enzymes
- Selection of “short” fragments (<500)
- Sequencing of extremities of selected fragments
- Relative constant sampling of regions in distinct samples (exception if mutation in restriction sites)
- Single or combination of restriction enzyme(s)



The Genotyping By Sequencing in detail: combination of two enzymes (pstI & mseI)

pstI *mseI*

5’…TCCTCTTACAGGATCCTGCAGCAACAAGGGTTAAGAATTATAAGCA…3’
3’…AGGAGAATGTCCTAGGACGTCGTTGTTCCCAATTCTTAATATTCGT…5’

Enzymatic restriction

DNA insert
5’… GCAACAAGGGT …3’
3’… ACGTCGTTGTTCCCAAT …5’

+

Barcode adapter
5’-ACACTCTTTCCTACACGACGCTCTTCCGATCTXXXXTGCA
3’-TGTGAGAAAGGATGTGCTGCGAGAAGGTAGAYYYY
Barcode (unique to each individual)

Common adapter
TACAGATCGGAAGAGCGGTTCAGCAGGAATGCCGA-3’
GTCTAGCCTTCTCGCCAACTCGTCCTTACGGCT-5’

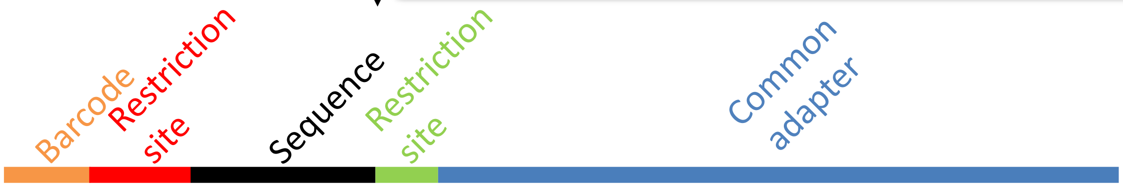
Ligation

5’-ACACTCTTTCCTACACGACGCTCTTCCGATCTXXXXTGCAGCAACAAGGGTTACAGATCGGAAGAGCGGTTCAGCAGGAATGCCGA-3’
3’-TGTGAGAAAGGATGTGCTGCGAGAAGGTAGAYYYYACGTCGTTGTTCCCAATGTCTAGCCTTCTCGCCAACTCGTCCTTACGGCT-5’

5’ -> 3’ fragment selection for sequencing

Illumina Primer
5’-ACACTCTTTCCTACACGACGCTCTTCCGATCTXXXXTGCAGCAACAAGGGTTACAGATCGGAAGAGCGGTTCAGCAGGAATGCCGA-3’

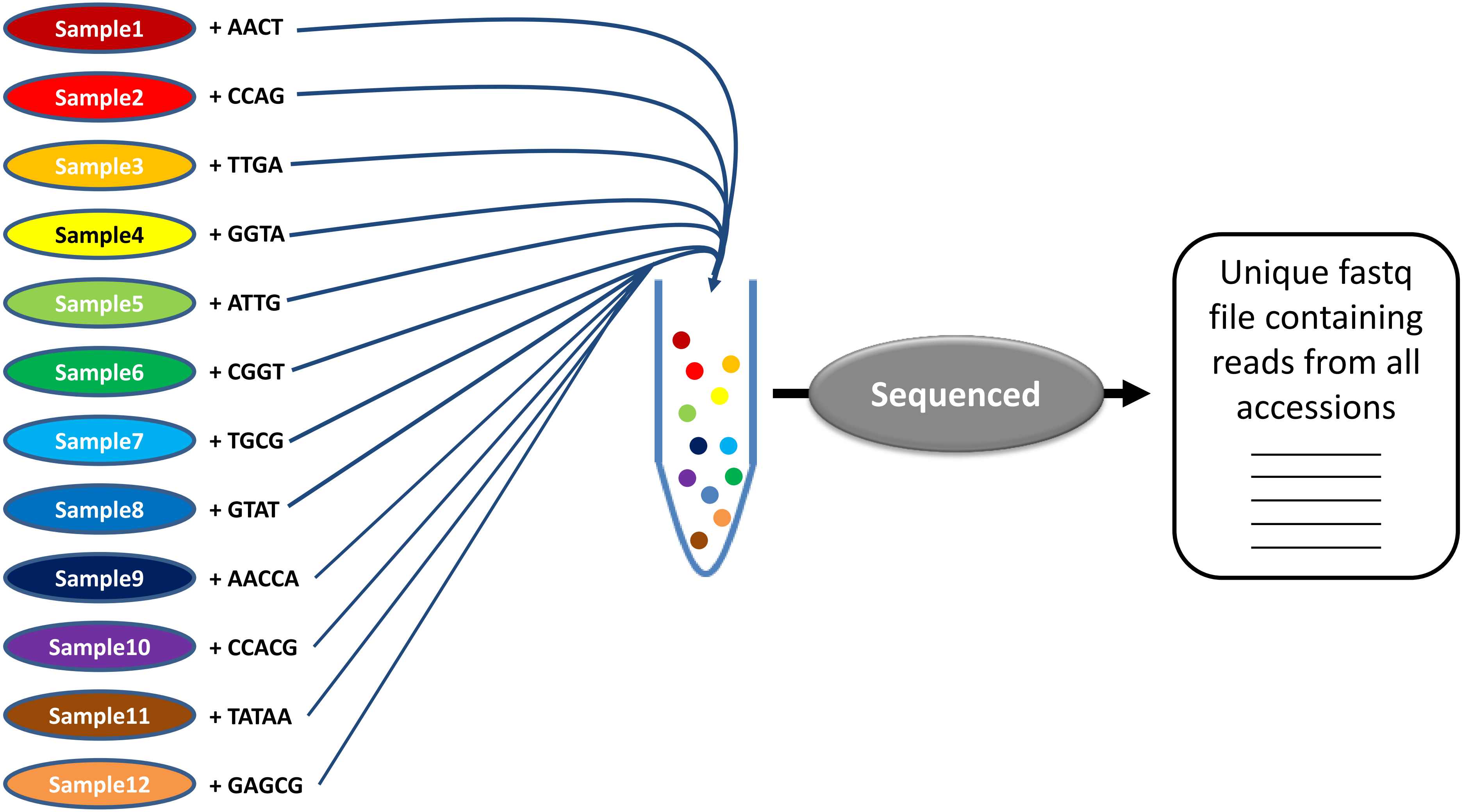
Sequencing



5’-XXXXTGCAGCAACAAGGGTTACAGATCGGAAGAGCGGTTCAGCAGGAATGCCGA-3’

From the output of GBS sequencing to the variant calling file: in command line

- We have generated a small GBS dataset comprising 12 samples for which ***pstI*** and ***mseI*** enzymes have been used and a sample specific barcode have been used.



From the output of GBS sequencing to the variant calling file: in command line

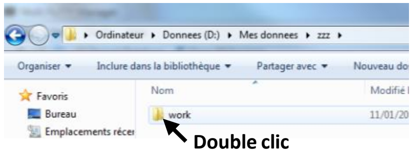
- Obtaining the datasets:
  1. Log onto the cluster

  2. Go to your "work" directory

```
cd
cd work
```

cd ← change directory (with nothing else: it goes to your home - /home/Your_ID)

cd work ← change directory to work

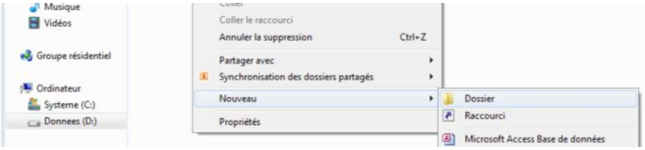


Double clic

  3. Create a folder in which we will be working and go into:

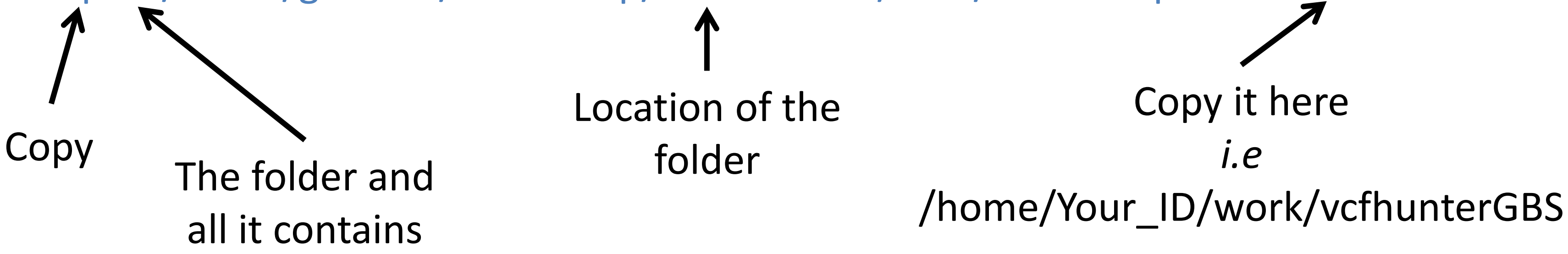
```
mkdir vcfhunterGBS
cd vcfhunterGBS
```

mkdir vcfhunterGBS ← make directory vcfhunterGBS



  4. Copy the folder containing the sequencing information:

```
cp -R /home/gmartin/WorkShop/VCFHUNTER/data/WorkShopDataset .
```



Copy

The folder and all it contains

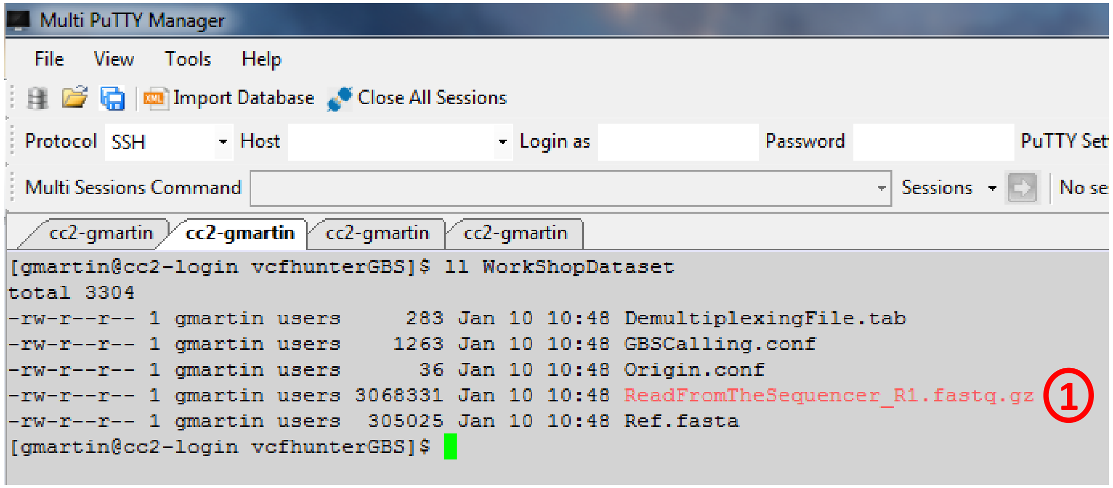
Location of the folder

Copy it here *i.e* /home/Your_ID/work/vcfhunterGBS

From the output of GBS sequencing to the variant calling file: in command line

- Listing the datasets:

  `ll WorkShopDataset`


Multi PuTTY Manager
File View Tools Help
Import Database Close All Sessions
Protocol SSH Host Login as Password PuTTY Set
Multi Sessions Command Sessions No se
cc2-gmartin cc2-gmartin cc2-gmartin cc2-gmartin
1

```
[gmartin@cc2-login vcfhunterGBS]$ ll WorkShopDataset
total 3304
-rw-r--r-- 1 gmartin users     283 Jan 10 10:48 DemultiplexingFile.tab
-rw-r--r-- 1 gmartin users    1263 Jan 10 10:48 GBSCalling.conf
-rw-r--r-- 1 gmartin users      36 Jan 10 10:48 Origin.conf
-rw-r--r-- 1 gmartin users 3068331 Jan 10 10:48 ReadFromTheSequencer_R1.fastq.gz
-rw-r--r-- 1 gmartin users  305025 Jan 10 10:48 Ref.fasta
[gmartin@cc2-login vcfhunterGBS]$
```

1 A compressed file (.gz) containing all reads from all accessions obtained from the sequencer

- To have a look at this file

  `zmore WorkShopDataset/ReadFromTheSequencer_R1.fastq.gz`

  ↑ Reading line by line a zipped file (.gz) ↑ Path to the fastq file

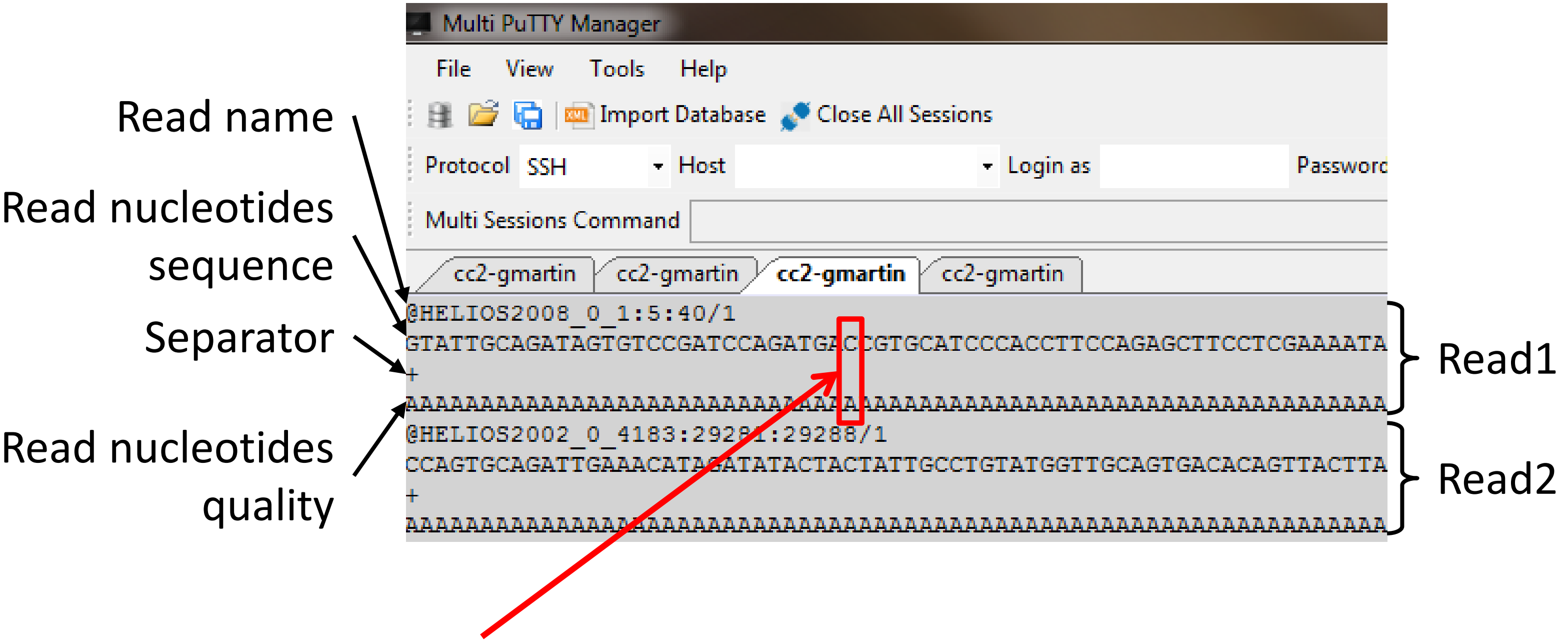
- Because zmore will list the file until its end using the “enter” key, and we do not want that because the file is big, we can “kill” the command with a combination of key:

  `“Ctrl” + “C”`

From the output of GBS sequencing to the variant calling file: in command line

- To have a look at this file
  - `zmore WorkShopDataset/ReadFromTheSequencer_R1.fastq.gz`

Fatsq format (for each read)



Read name

Read nucleotides sequence

Separator

Read nucleotides quality

```
@HELIOS2008_0_1:5:40/1
GTATTGCAGATAGTGTCCGATCCAGATGACCGTGCATCCCACCTTCCAGAGCTTCCTCGAAAATA
+
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
@HELIOS2002_0_4183:29281:29288/1
CCAGTGCAGATTGAAACATAGATATACTACTATTGCCTGTATGGTTGCAGTGACACAGTTACTTA
+
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
```

Read1

Read2

Base quality encoding: for base “C” = A

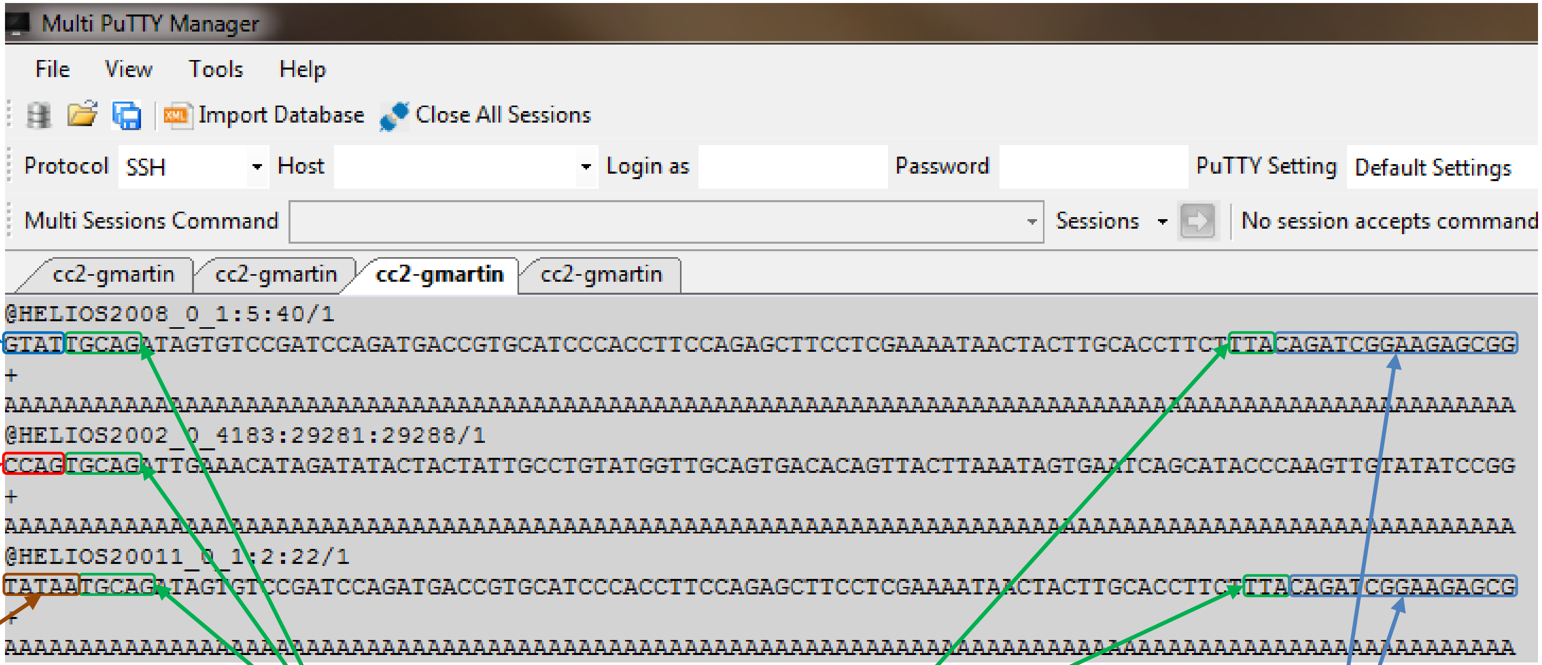
But what does “A” mean?

- Each letter has informatically a numeric value. For example “A” is equal to 65
- We should remove 33 to this value and thus “A” = 65-33 = 32!

```
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJ
 |                         |    |        |
33                        59   64       73
 0.2......................26...31.......41
```

From the output of GBS sequencing to the variant calling file: in command line

- To have a look at this file
  - `zmore WorkShopDataset/ReadFromTheSequencer_R1.fastq.gz`



```
@HELIOS2008_0_1:5:40/1
GTATTGCAGATAGTGTCCGATCCAGATGACCGTGCATCCCACCTTCCAGAGCTTCCTCGAAAATAACTACTTGCACCTTCTTTACAGATCGGAAGAGCGG
+
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
@HELIOS2002_0_4183:29281:29288/1
CCAGTGCAGATTGAAACATAGATATACTACTATTGCCTGTATGGTTGCAGTGACACAGTTACTTAAATAGTGAATCAGCATACCCAAGTTGTATATCCGG
+
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
@HELIOS20011_0_1:2:22/1
TATAATGCAGATAGTGTCCGATCCAGATGACCGTGCATCCCACCTTCCAGAGCTTCCTCGAAAATAACTACTTGCACCTTCTTTACAGATCGGAAGAGCG
+
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
```

Sample8 tag

➔ **Read from sample8**

Sample2 tag

➔ **Read from sample2**

Sample11 tag

➔ **Read from sample11**

pstI restriction site

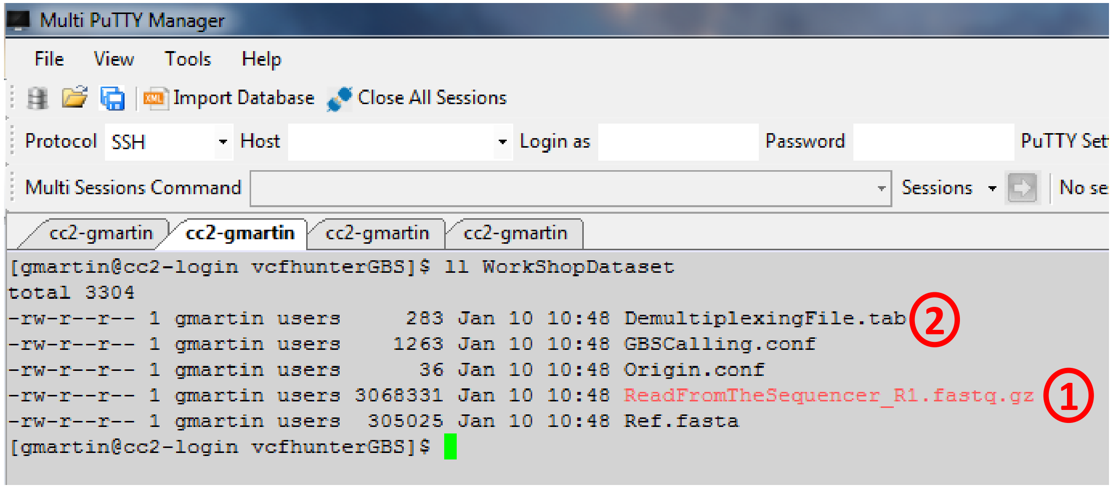
mseI restriction site

Adapter sequence

From the output of GBS sequencing to the variant calling file: in command line

- Listing the datasets:

```
ll WorkShopDataset
```



```
[gmartin@cc2-login vcfhunterGBS]$ ll WorkShopDataset
total 3304
-rw-r--r-- 1 gmartin users     283 Jan 10 10:48 DemultiplexingFile.tab
-rw-r--r-- 1 gmartin users    1263 Jan 10 10:48 GBSCalling.conf
-rw-r--r-- 1 gmartin users      36 Jan 10 10:48 Origin.conf
-rw-r--r-- 1 gmartin users 3068331 Jan 10 10:48 ReadFromTheSequencer_R1.fastq.gz
-rw-r--r-- 1 gmartin users  305025 Jan 10 10:48 Ref.fasta
[gmartin@cc2-login vcfhunterGBS]$
```

① A compressed file (.gz) containing all reads from all accessions obtained from the sequencer

② **A file that will be used to separate reads in distinct file according to the accession they belong**

- To have a look at this file

```
more WorkShopDataset/DemultiplexingFile.tab
```

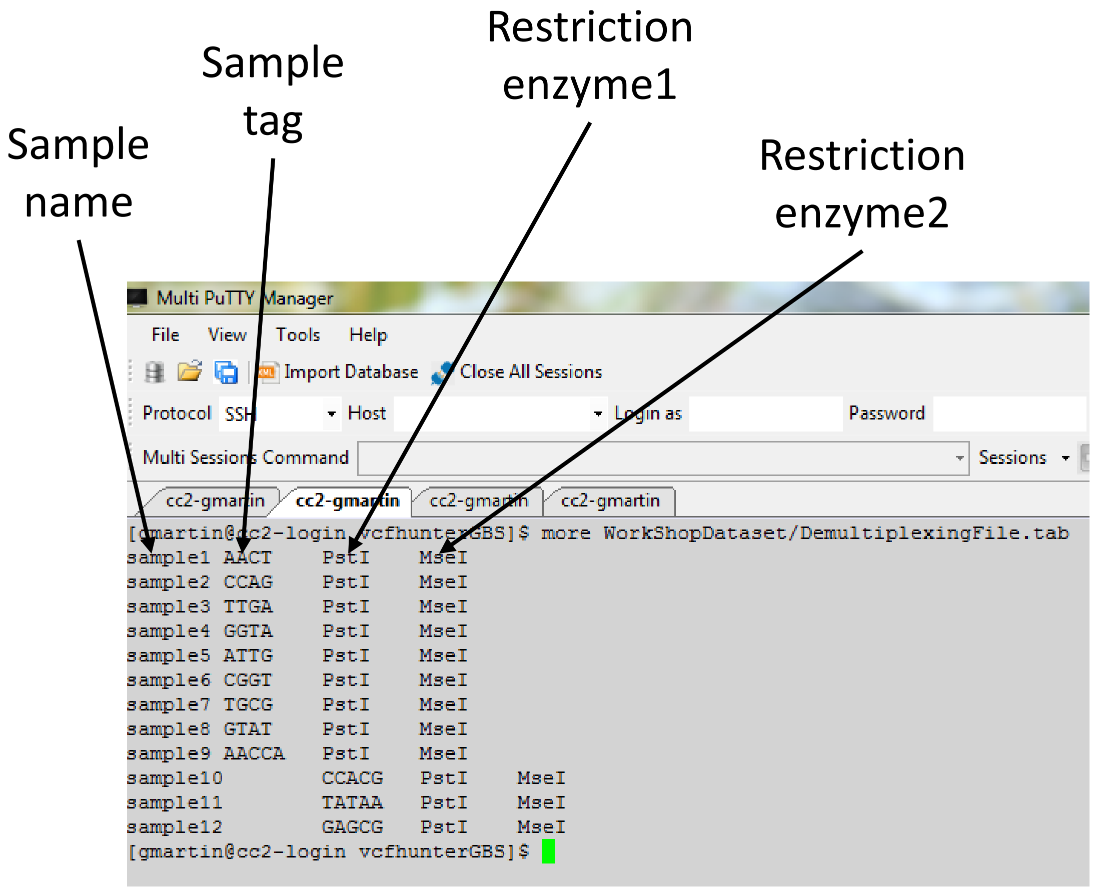
Reading line by line a file

Path to the file

From the output of GBS sequencing to the variant calling file: in command line

- To have a look at this file

```
more WorkShopDataset/DemultiplexingFile.tab
```



Sample name

Sample tag

Restriction enzyme1

Restriction enzyme2

```
[gmartin@cc2-login vcfhunterGBS]$ more WorkShopDataset/DemultiplexingFile.tab
sample1 AACT    PstI    MseI
sample2 CCAG    PstI    MseI
sample3 TTGA    PstI    MseI
sample4 GGTA    PstI    MseI
sample5 ATTG    PstI    MseI
sample6 CGGT    PstI    MseI
sample7 TGCG    PstI    MseI
sample8 GTAT    PstI    MseI
sample9 AACCA   PstI    MseI
sample10        CCACG   PstI    MseI
sample11        TATAA   PstI    MseI
sample12        GAGCG   PstI    MseI
[gmartin@cc2-login vcfhunterGBS]$
```

Sample1	+ AACT
Sample2	+ CCAG
Sample3	+ TTGA
Sample4	+ GGTA
Sample5	+ ATTG
Sample6	+ CGGT
Sample7	+ TGCG
Sample8	+ GTAT
Sample9	+ AACCA
Sample10	+ CCACG
Sample11	+ TATAA
Sample12	+ GAGCG

From the output of GBS sequencing to the variant calling file: in command line

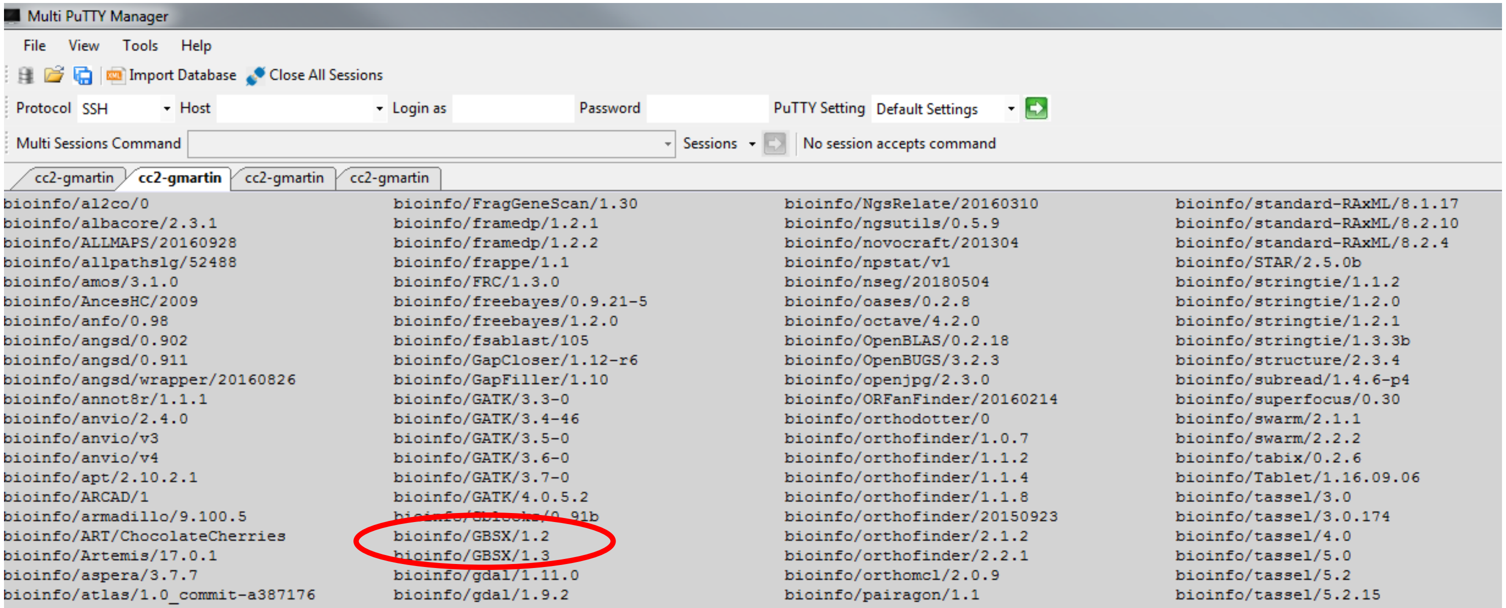
- Now it is time to demultiplex! *i.e.* parse reads in files corresponding to sample.

- For that we will use GBSX (https://github.com/GenomicsCoreLeuven/GBSX, https://doi.org/10.1186/s12859-015-0514-3)

- A small parenthesis: On the AGAP cluster, several modules are already available. To access the list of available modules, use the following command line:

```
module avail
```

A list of modules appears and we can find “GBSX” program in this list!



```
bioinfo/al2co/0                         bioinfo/FragGeneScan/1.30               bioinfo/NgsRelate/20160310              bioinfo/standard-RAxML/8.1.17
bioinfo/albacore/2.3.1                  bioinfo/framedp/1.2.1                   bioinfo/ngsutils/0.5.9                  bioinfo/standard-RAxML/8.2.10
bioinfo/ALLMAPS/20160928                bioinfo/framedp/1.2.2                   bioinfo/novocraft/201304                bioinfo/standard-RAxML/8.2.4
bioinfo/allpathslg/52488                bioinfo/frappe/1.1                      bioinfo/npstat/v1                       bioinfo/STAR/2.5.0b
bioinfo/amos/3.1.0                      bioinfo/FRC/1.3.0                       bioinfo/nseg/20180504                   bioinfo/stringtie/1.1.2
bioinfo/AncesHC/2009                    bioinfo/freebayes/0.9.21-5              bioinfo/oases/0.2.8                     bioinfo/stringtie/1.2.0
bioinfo/anfo/0.98                       bioinfo/freebayes/1.2.0                 bioinfo/octave/4.2.0                    bioinfo/stringtie/1.2.1
bioinfo/angsd/0.902                     bioinfo/fsablast/105                    bioinfo/OpenBLAS/0.2.18                 bioinfo/stringtie/1.3.3b
bioinfo/angsd/0.911                     bioinfo/GapCloser/1.12-r6               bioinfo/OpenBUGS/3.2.3                  bioinfo/structure/2.3.4
bioinfo/angsd/wrapper/20160826          bioinfo/GapFiller/1.10                  bioinfo/openjpg/2.3.0                   bioinfo/subread/1.4.6-p4
bioinfo/annot8r/1.1.1                   bioinfo/GATK/3.3-0                      bioinfo/ORFanFinder/20160214            bioinfo/superfocus/0.30
bioinfo/anvio/2.4.0                     bioinfo/GATK/3.4-46                     bioinfo/orthodotter/0                   bioinfo/swarm/2.1.1
bioinfo/anvio/v3                        bioinfo/GATK/3.5-0                      bioinfo/orthofinder/1.0.7               bioinfo/swarm/2.2.2
bioinfo/anvio/v4                        bioinfo/GATK/3.6-0                      bioinfo/orthofinder/1.1.2               bioinfo/tabix/0.2.6
bioinfo/apt/2.10.2.1                    bioinfo/GATK/3.7-0                      bioinfo/orthofinder/1.1.4               bioinfo/Tablet/1.16.09.06
bioinfo/ARCAD/1                         bioinfo/GATK/4.0.5.2                    bioinfo/orthofinder/1.1.8               bioinfo/tassel/3.0
bioinfo/armadillo/9.100.5               bioinfo/Gblocks/0.91b                   bioinfo/orthofinder/20150923            bioinfo/tassel/3.0.174
bioinfo/ART/ChocolateCherries           bioinfo/GBSX/1.2                        bioinfo/orthofinder/2.1.2               bioinfo/tassel/4.0
bioinfo/Artemis/17.0.1                  bioinfo/GBSX/1.3                        bioinfo/orthofinder/2.2.1               bioinfo/tassel/5.0
bioinfo/aspera/3.7.7                    bioinfo/gdal/1.11.0                     bioinfo/orthomcl/2.0.9                  bioinfo/tassel/5.2
bioinfo/atlas/1.0_commit-a387176        bioinfo/gdal/1.9.2                      bioinfo/pairagon/1.1                    bioinfo/tassel/5.2.15
```

➔ Two versions are available! We will take the 1.2 version

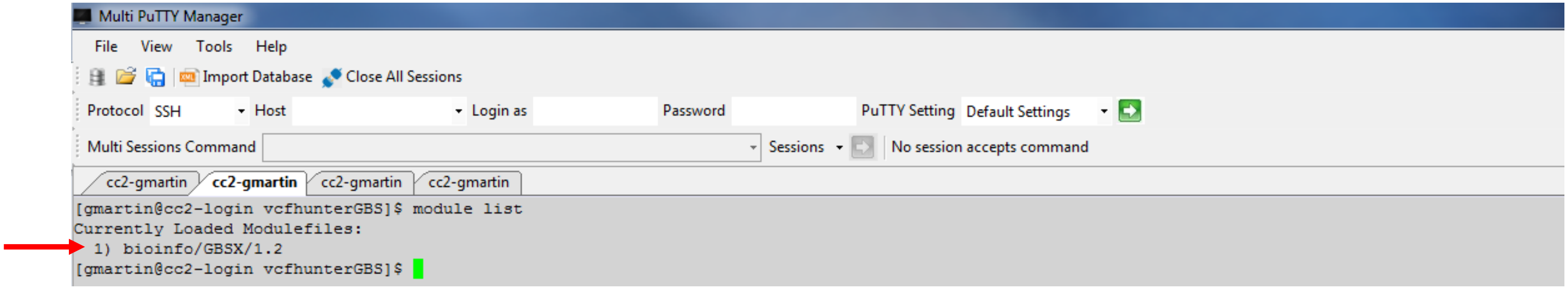
From the output of GBS sequencing to the variant calling file: in command line

- To load this module run the command line:

```
module load bioinfo/GBSX/1.2
```

- The module is now loaded. This can be verified by listing the loaded modules with de following command line:

```
module list
```



Protocol SSH

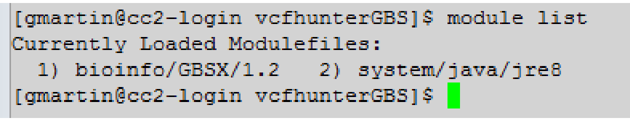
PuTTY Setting Default Settings

```
[gmartin@cc2-login vcfhunterGBS]$ module list
Currently Loaded Modulefiles:
  1) bioinfo/GBSX/1.2
[gmartin@cc2-login vcfhunterGBS]$
```

The GBSX module is loaded. But what you don't know, is that GBSX need another program to be used! This program is JAVA. To load java we will run the command line:

```
module load system/java/jre8
```

You can try again `module list` to verify that java has been loaded



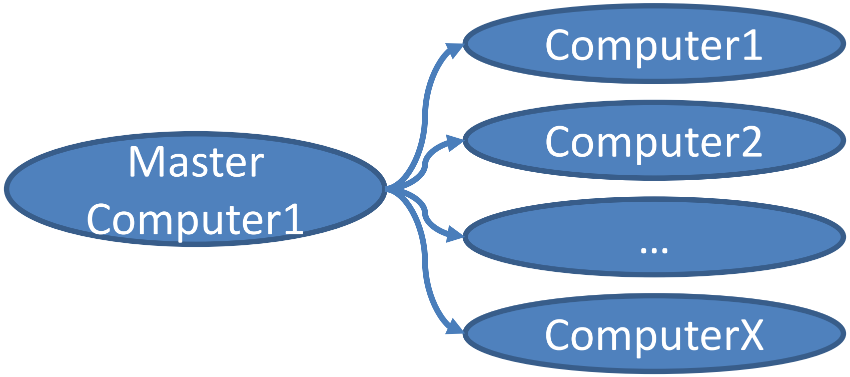
```
[gmartin@cc2-login vcfhunterGBS]$ module list
Currently Loaded Modulefiles:
  1) bioinfo/GBSX/1.2    2) system/java/jre8
[gmartin@cc2-login vcfhunterGBS]$
```

From the output of GBS sequencing to the variant calling file: in command line

- At this point all is ready to demultiplex the fastq file! All we have to do is to run the following command line (in one single line):

```
qsub -q normal.q -l mem_free=12G -b yes -V -N DEMULT java -XX:ParallelGCThreads=1 -Xmx8G
-jar /usr/local/bioinfo/GBSX/1.2/GBSX_v1.1.2.jar --Demultiplexer
-f1 WorkShopDataset/ReadFromTheSequencer_R1.fastq.gz
-i WorkShopDataset/DemultiplexingFile.tab -o Demultiplexed -gzip true -mb 0
```

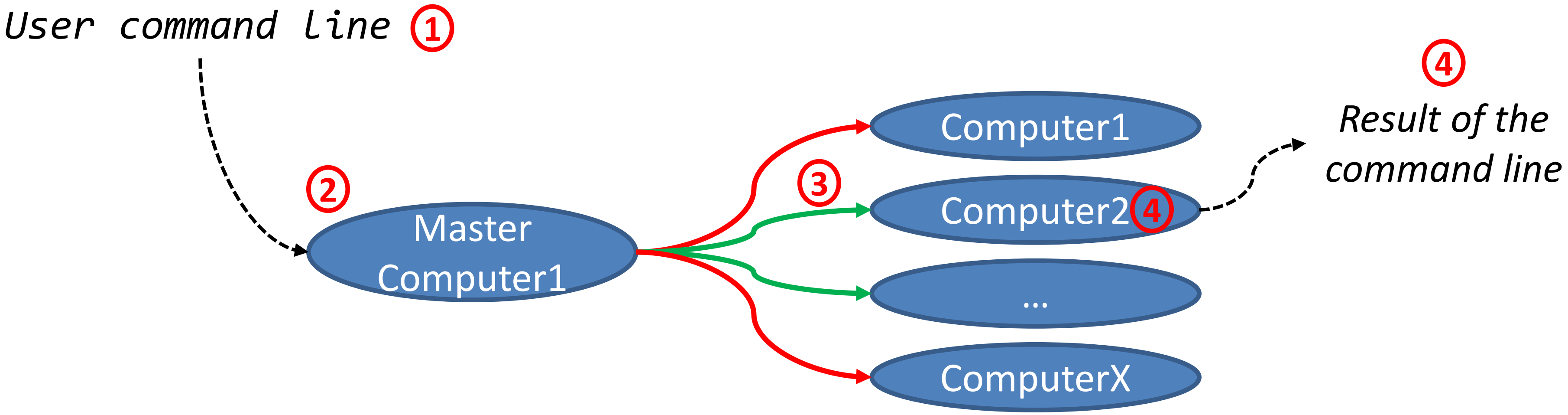
- Now a little piece of explanation:
  - ✓ We are working on a cluster.
    - ❖ This means that we have several computers which are connected so that they can work together.
    - ❖ It also allows that several people can run huge calculation at the same time!
    - ❖ It also means that there is a strict procedure to perform calculation on the cluster and this procedure is associated to the way a cluster work:



A single computer to rule them all

From the output of GBS sequencing to the variant calling file: in command line

How the cluster works?



1. The user tip a command line

2. Which is sent to the master computer

3. Based on this command line, the master computer identify which computer it rules match the command requirements and which of them are available

4. The command line is executed on the chosen computer (in this example **Computer2**)

5. Which returns the result of the command line

From the output of GBS sequencing to the variant calling file: in command line

- Back to the command line:

```
qsub -q normal.q -l mem_free=12G -b yes -V -N DEMULT “java -XX:ParallelGCThreads=1 -Xmx8G
-jar /usr/local/bioinfo/GBSX/1.2/GBSX_v1.1.2.jar --Demultiplexer
-f1 WorkShopDataset/ReadFromTheSequencer_R1.fastq.gz
-i WorkShopDataset/DemultiplexingFile.tab -o Demultiplexed -gzip true -mb 0”
```

- The first part of the command line (in bold) is **used by the master computer**:
  - **qsub**: Means that we will send a command that the master computer needs to analyze to choose the best computer
  - **-q normal.q**: tells the master computer that we will use computer from normal queue. Several queues exist depending on computation requirement:
    - ***normal.q***: access to computers of 48 processors with 192Go shared memory (RAM) and a command line cannot exceed 48hours of running time.
    - ***long.q***: access to computers of 48 processors with 192Go shared memory but there is not running time limit
    - ***bigmem.q***: access to a unique computer of 96 processors with 2,6To shared memory and no time limit
  - **-l mem_free=12G**: precise that the program will use 12G of RAM (so the master computer will check that it is available on the computers). This is a facultative option but necessary when using **java** program to prevent errors…
  - **-b yes**: it is not important, but put it.
  - **-V**: Tell the master computer to load the module previously loaded on the computer it will choose
  - **-N DEMULT**: A name passed to the command line to look at its status (waiting, running or error) on the cluster

From the output of GBS sequencing to the variant calling file: in command line

- Back to the command line:

```
qsub -q normal.q -l mem_free=12G -b yes -V -N DEMULT “java -XX:ParallelGCThreads=1 -Xmx8G
-jar /usr/local/bioinfo/GBSX/1.2/GBSX_v1.1.2.jar --Demultiplexer
-f1 WorkShopDataset/ReadFromTheSequencer_R1.fastq.gz
-i WorkShopDataset/DemultiplexingFile.tab -o Demultiplexed -gzip true -mb 0”
```

- The part of the command line between quotation marks (in bold) is the command line that is executed on the **computer chosen by the master computer**.
  - **/usr/local/bioinfo/GBSX/1.2/GBSX_v1.1.2.jar**: is the program that is used to demultiplex the fastq file. Element in black are options/argument passed to this program to make it work (as a function and its arguments in Excel!).
  - **--Demultiplexer**: Tell the program that we want to demultiplex the fastq
  - **-f1 WorkShopDataset/ReadFromTheSequencer_R1.fastq.gz**: locate the fastq file to demultiplex
  - **-i WorkShopDataset/DemultiplexingFile.tab**: loacte the file containing the multiplexing informations (which tags correspond to which samples and restriction enzymes used)
  - **-o Demultiplexed**: The name of the output folder (this folder will be created by the program).
  - **-gzip true**: Tells the program that output should be compressed to gain space (equivalent to .zip files on Windows)
  - **-mb 0**: Tells the program that 0 mismatch are allowed in the tag to attribute a read to an accession

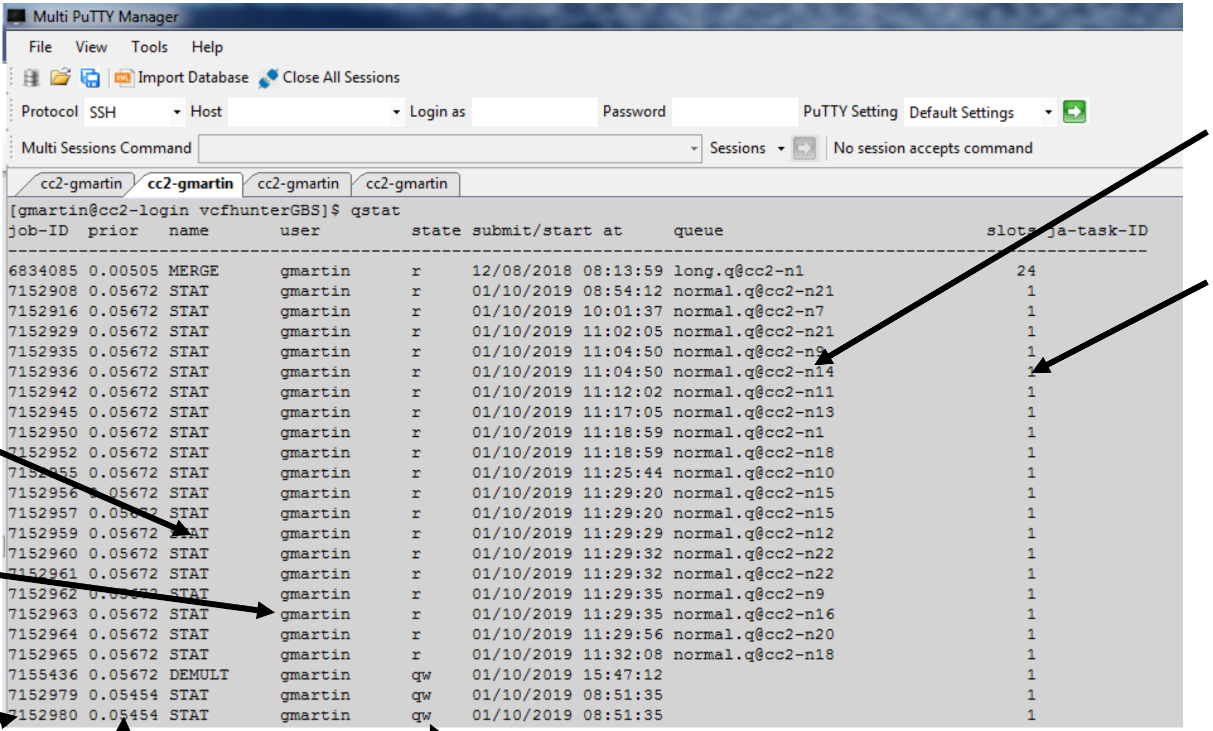
  - java -XX:ParallelGCThreads=1 -Xmx8G -jar: Tells to the computer that the program ***/usr/local/bioinfo/GBSX/1.2/GBSX_v1.1.2.jar*** is written in java language (java), that java should only use one processor (-XX:ParallelGCThreads=1) and that 8G memory are available for java (-Xmx8G -jar). -jar indicate to java that the program is directly after.

From the output of GBS sequencing to the variant calling file: in command line

- One can check the status of job(s) with the following command line:

```
qstat
```

- Because the job we have sent is a very short one it is likely that it will be finished before you run this command line… Here is an example of the what we can observe:



```
[gmartin@cc2-login vcfhunterGBS]$ qstat
job-ID  prior   name       user         state submit/start at     queue                          slots ja-task-ID
-----------------------------------------------------------------------------------------------------------------
6834085 0.00505 MERGE      gmartin      r     12/08/2018 08:13:59 long.q@cc2-n1                     24
7152908 0.05672 STAT       gmartin      r     01/10/2019 08:54:12 normal.q@cc2-n21                   1
7152916 0.05672 STAT       gmartin      r     01/10/2019 10:01:37 normal.q@cc2-n7                    1
7152929 0.05672 STAT       gmartin      r     01/10/2019 11:02:05 normal.q@cc2-n21                   1
7152935 0.05672 STAT       gmartin      r     01/10/2019 11:04:50 normal.q@cc2-n9                    1
7152936 0.05672 STAT       gmartin      r     01/10/2019 11:04:50 normal.q@cc2-n14                   1
7152942 0.05672 STAT       gmartin      r     01/10/2019 11:12:02 normal.q@cc2-n11                   1
7152945 0.05672 STAT       gmartin      r     01/10/2019 11:17:05 normal.q@cc2-n13                   1
7152950 0.05672 STAT       gmartin      r     01/10/2019 11:18:59 normal.q@cc2-n1                    1
7152952 0.05672 STAT       gmartin      r     01/10/2019 11:18:59 normal.q@cc2-n18                   1
7152955 0.05672 STAT       gmartin      r     01/10/2019 11:25:44 normal.q@cc2-n10                   1
7152956 0.05672 STAT       gmartin      r     01/10/2019 11:29:20 normal.q@cc2-n15                   1
7152957 0.05672 STAT       gmartin      r     01/10/2019 11:29:20 normal.q@cc2-n15                   1
7152959 0.05672 STAT       gmartin      r     01/10/2019 11:29:29 normal.q@cc2-n12                   1
7152960 0.05672 STAT       gmartin      r     01/10/2019 11:29:32 normal.q@cc2-n22                   1
7152961 0.05672 STAT       gmartin      r     01/10/2019 11:29:32 normal.q@cc2-n22                   1
7152962 0.05672 STAT       gmartin      r     01/10/2019 11:29:35 normal.q@cc2-n9                    1
7152963 0.05672 STAT       gmartin      r     01/10/2019 11:29:35 normal.q@cc2-n16                   1
7152964 0.05672 STAT       gmartin      r     01/10/2019 11:29:56 normal.q@cc2-n20                   1
7152965 0.05672 STAT       gmartin      r     01/10/2019 11:32:08 normal.q@cc2-n18                   1
7155436 0.05672 DEMULT     gmartin      qw    01/10/2019 15:47:12                                    1
7152979 0.05454 STAT       gmartin      qw    01/10/2019 08:51:35                                    1
7152980 0.05454 STAT       gmartin      qw    01/10/2019 08:51:35                                    1
```

Name of the job (-N option in the qsub)

Owner of the job

Job ID (unique)

Priority of the job

Computer used

Processor number used

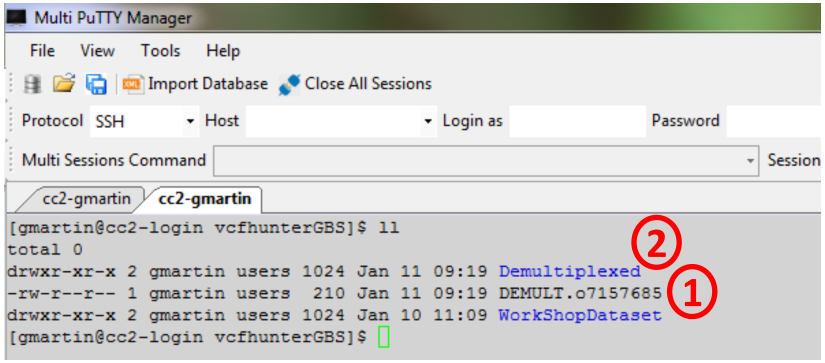
Job status
r = running
qw = waiting to run (no computer available)
other (Eqw, dt, …) = there is a problem

From the output of GBS sequencing to the variant calling file: in command line

- Output of the demultiplexing command line. Listing the current directory:

```
ll
```

- One file and one folder are generated:



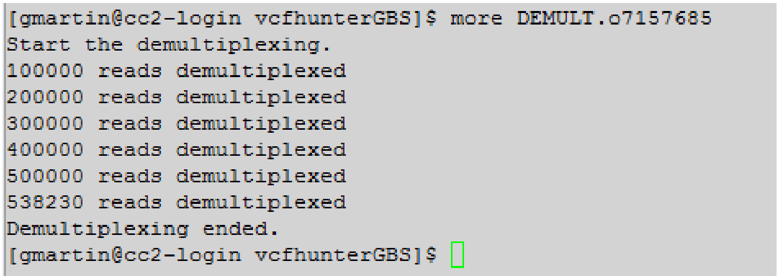
```
[gmartin@cc2-login vcfhunterGBS]$ ll
total 0
drwxr-xr-x 2 gmartin users 1024 Jan 11 09:19 Demultiplexed
-rw-r--r-- 1 gmartin users  210 Jan 11 09:19 DEMULT.o7157685
drwxr-xr-x 2 gmartin users 1024 Jan 10 11:09 WorkShopDataset
[gmartin@cc2-login vcfhunterGBS]$
```

① A file named DEMULT.o7157685

- Correspond to the Name of the job passed to the qsub (`-N DEMULT`) concatenated with the unique job ID attributed by the master computer to the command line (here: **7157685**).

- Because some programs "speak": this file contained what they say. We can have a look at what the program say with the more command:

```
more DEMULT.o7157685
```



```
[gmartin@cc2-login vcfhunterGBS]$ more DEMULT.o7157685
Start the demultiplexing.
100000 reads demultiplexed
200000 reads demultiplexed
300000 reads demultiplexed
400000 reads demultiplexed
500000 reads demultiplexed
538230 reads demultiplexed
Demultiplexing ended.
[gmartin@cc2-login vcfhunterGBS]$
```

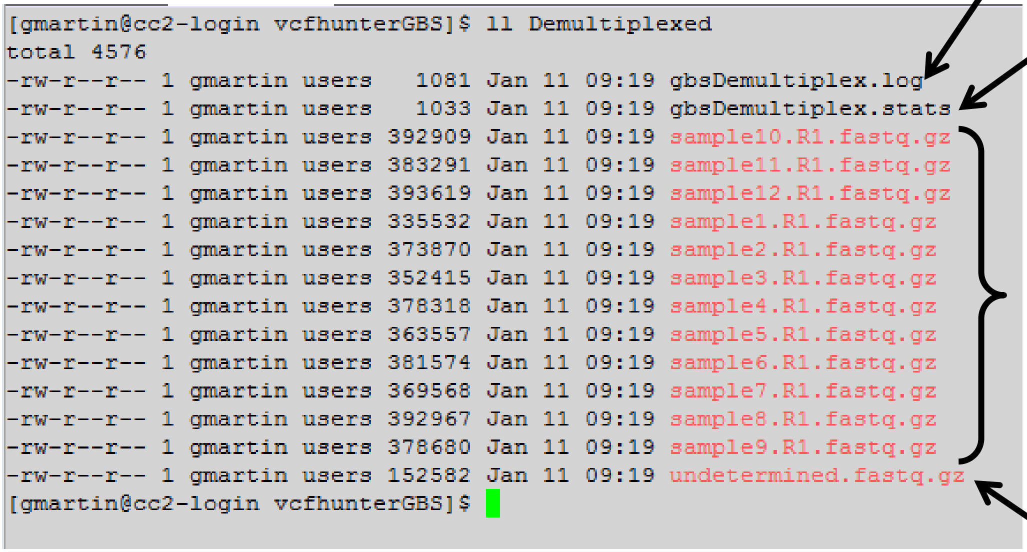
② A folder named Demultiplexed

This folder was created by GBSX as we tell him to do it with the (`-o Demultiplexed`) argument.

From the output of GBS sequencing to the variant calling file: in command line

- Listing the demultiplexed folder:

```
ll Demultiplexed
```



```
[gmartin@cc2-login vcfhunterGBS]$ ll Demultiplexed
total 4576
-rw-r--r-- 1 gmartin users   1081 Jan 11 09:19 gbsDemultiplex.log
-rw-r--r-- 1 gmartin users   1033 Jan 11 09:19 gbsDemultiplex.stats
-rw-r--r-- 1 gmartin users 392909 Jan 11 09:19 sample10.R1.fastq.gz
-rw-r--r-- 1 gmartin users 383291 Jan 11 09:19 sample11.R1.fastq.gz
-rw-r--r-- 1 gmartin users 393619 Jan 11 09:19 sample12.R1.fastq.gz
-rw-r--r-- 1 gmartin users 335532 Jan 11 09:19 sample1.R1.fastq.gz
-rw-r--r-- 1 gmartin users 373870 Jan 11 09:19 sample2.R1.fastq.gz
-rw-r--r-- 1 gmartin users 352415 Jan 11 09:19 sample3.R1.fastq.gz
-rw-r--r-- 1 gmartin users 378318 Jan 11 09:19 sample4.R1.fastq.gz
-rw-r--r-- 1 gmartin users 363557 Jan 11 09:19 sample5.R1.fastq.gz
-rw-r--r-- 1 gmartin users 381574 Jan 11 09:19 sample6.R1.fastq.gz
-rw-r--r-- 1 gmartin users 369568 Jan 11 09:19 sample7.R1.fastq.gz
-rw-r--r-- 1 gmartin users 392967 Jan 11 09:19 sample8.R1.fastq.gz
-rw-r--r-- 1 gmartin users 378680 Jan 11 09:19 sample9.R1.fastq.gz
-rw-r--r-- 1 gmartin users 152582 Jan 11 09:19 undetermined.fastq.gz
[gmartin@cc2-login vcfhunterGBS]$
```

A file summarizing demultiplexing options

A file with demultiplexing statistics

Reads parsed according to the accession they belong to

A file containing reads that could not be attributed to an accession (*i.e.* sequencing error in the tag)

- To have a look at these files:

```
more Demultiplexed/gbsDemultiplex.log
```

(for example)

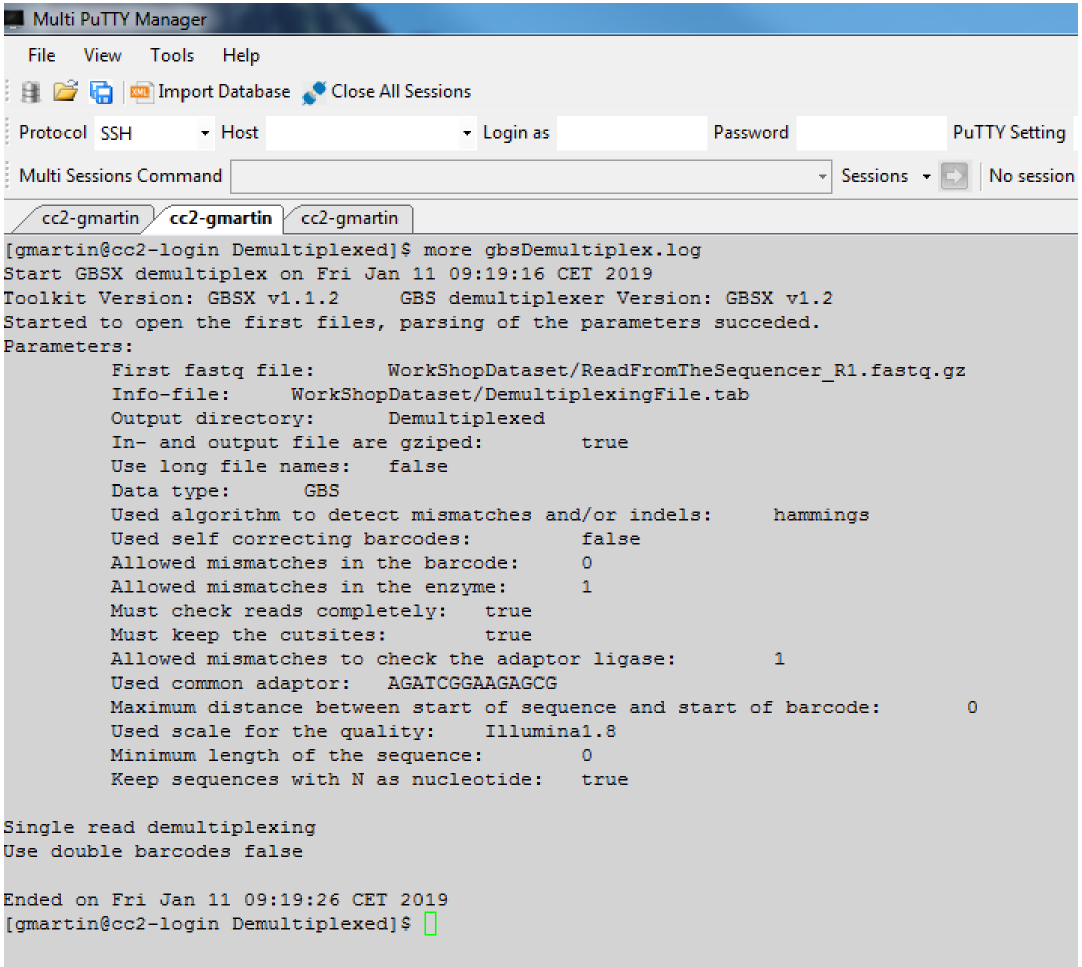
- But because it is boring to always put Demultiplexed/ for all file which are in the directory, we will directly go into this directory:

```
cd Demultiplexed
```

From the output of GBS sequencing to the variant calling file: in command line

- The gbsDemultiplex.log file:

  `more gbsDemultiplex.log`

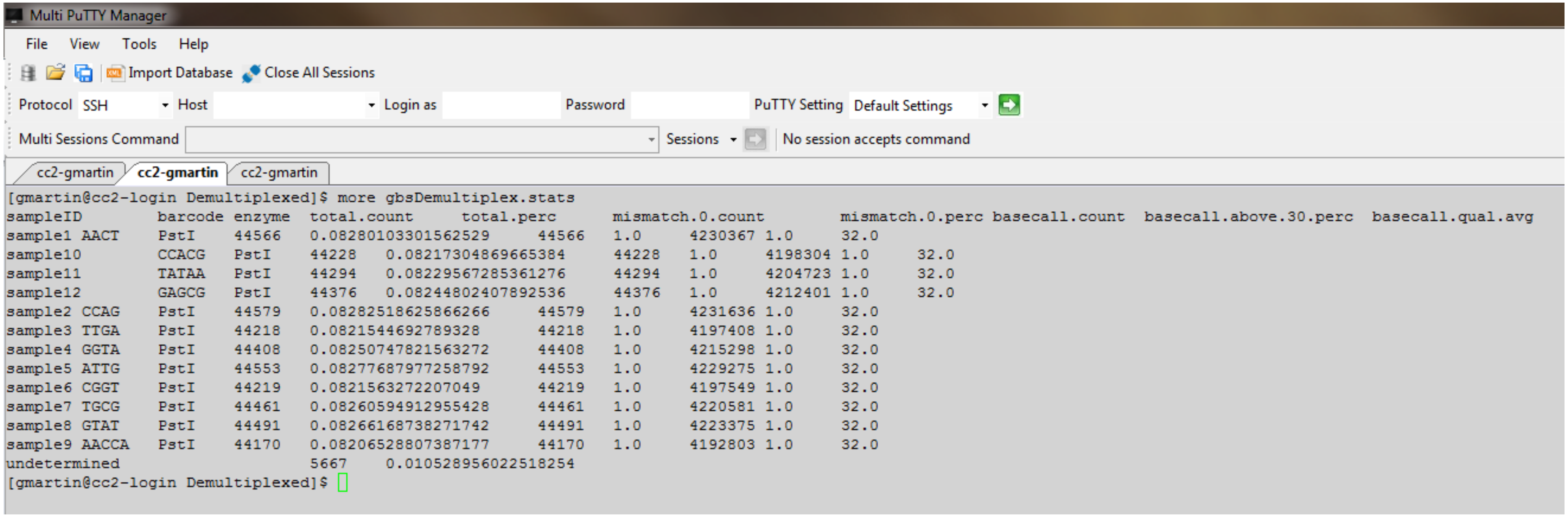


```
[gmartin@cc2-login Demultiplexed]$ more gbsDemultiplex.log
Start GBSX demultiplex on Fri Jan 11 09:19:16 CET 2019
Toolkit Version: GBSX v1.1.2     GBS demultiplexer Version: GBSX v1.2
Started to open the first files, parsing of the parameters succeded.
Parameters:
         First fastq file:      WorkShopDataset/ReadFromTheSequencer_R1.fastq.gz
         Info-file:      WorkShopDataset/DemultiplexingFile.tab
         Output directory:       Demultiplexed
         In- and output file are gziped:        true
         Use long file names:    false
         Data type:      GBS
         Used algorithm to detect mismatches and/or indels:     hammings
         Used self correcting barcodes:         false
         Allowed mismatches in the barcode:     0
         Allowed mismatches in the enzyme:      1
         Must check reads completely:   true
         Must keep the cutsites:        true
         Allowed mismatches to check the adaptor ligase:        1
         Used common adaptor:   AGATCGGAAGAGCG
         Maximum distance between start of sequence and start of barcode:       0
         Used scale for the quality:    Illumina1.8
         Minimum length of the sequence:        0
         Keep sequences with N as nucleotide:   true

Single read demultiplexing
Use double barcodes false

Ended on Fri Jan 11 09:19:26 CET 2019
[gmartin@cc2-login Demultiplexed]$ 
```

From the output of GBS sequencing to the variant calling file: in command line

- The gbsDemultiplex.log file:

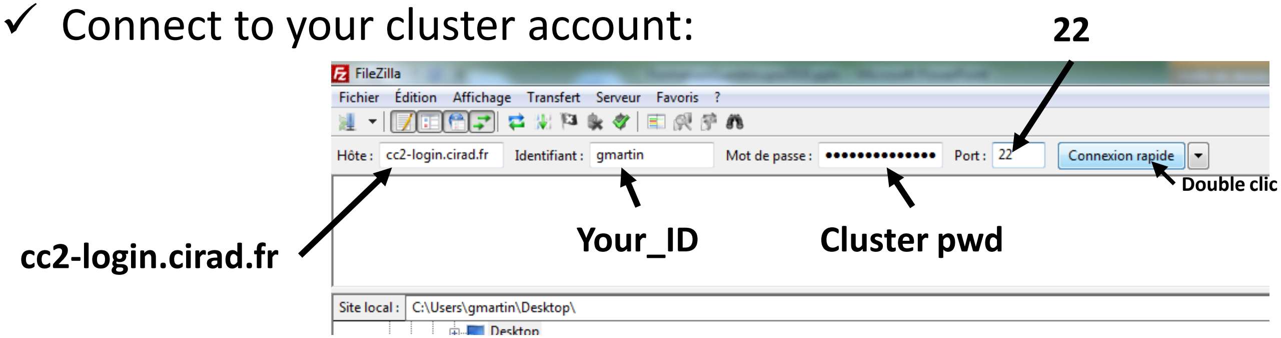
  `more gbsDemultiplex.stats`



Protocol SSH
Host
Login as
Password
PuTTY Setting Default Settings
No session accepts command

```
[gmartin@cc2-login Demultiplexed]$ more gbsDemultiplex.stats
sampleID        barcode enzyme  total.count     total.perc      mismatch.0.count        mismatch.0.perc basecall.count  basecall.above.30.perc  basecall.qual.avg
sample1 AACT    PstI    44566   0.08280103301562529     44566   1.0     4230367 1.0     32.0
sample10        CCACG   PstI    44228   0.08217304869665384     44228   1.0     4198304 1.0     32.0
sample11        TATAA   PstI    44294   0.08229567285361276     44294   1.0     4204723 1.0     32.0
sample12        GAGCG   PstI    44376   0.08244802407892536     44376   1.0     4212401 1.0     32.0
sample2 CCAG    PstI    44579   0.08282518625866266     44579   1.0     4231636 1.0     32.0
sample3 TTGA    PstI    44218   0.0821544692789328      44218   1.0     4197408 1.0     32.0
sample4 GGTA    PstI    44408   0.08250747821563272     44408   1.0     4215298 1.0     32.0
sample5 ATTG    PstI    44553   0.08277687977258792     44553   1.0     4229275 1.0     32.0
sample6 CGGT    PstI    44219   0.0821563272207049      44219   1.0     4197549 1.0     32.0
sample7 TGCG    PstI    44461   0.08260594912955428     44461   1.0     4220581 1.0     32.0
sample8 GTAT    PstI    44491   0.08266168738271742     44491   1.0     4223375 1.0     32.0
sample9 AACCA   PstI    44170   0.08206528807387177     44170   1.0     4192803 1.0     32.0
undetermined                    5667    0.010528956022518254
[gmartin@cc2-login Demultiplexed]$
```

- Not very easy to read… We will load this file on our computer.
  - ✓ For that we need FileZilla: https://filezilla-project.org/
  - ✓ Install it
  - ✓ Connect to your cluster account:



Hôte : cc2-login.cirad.fr
Identifiant : gmartin
Mot de passe : ••••••••••••••
Port : 22
Site local : C:\Users\gmartin\Desktop\

22

Double clic

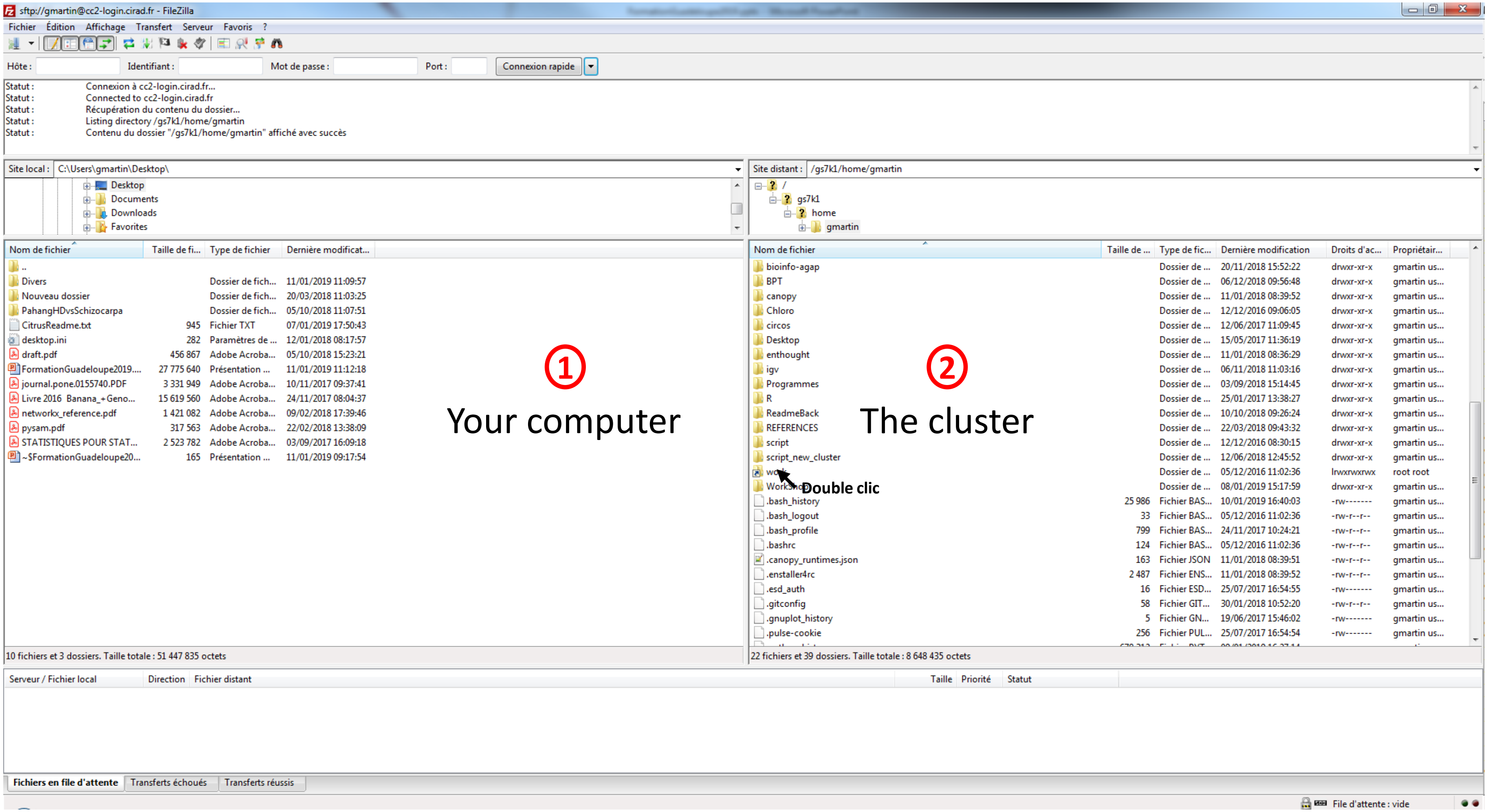
cc2-login.cirad.fr

Your_ID

Cluster pwd

From the output of GBS sequencing to the variant calling file: in command line

- The gbsDemultiplex.log file:



sftp://gmartin@cc2-login.cirad.fr - FileZilla

Fichier Édition Affichage Transfert Serveur Favoris ?

Hôte : Identifiant : Mot de passe : Port : Connexion rapide

```
Statut :	Connexion à cc2-login.cirad.fr...
Statut :	Connected to cc2-login.cirad.fr
Statut :	Récupération du contenu du dossier...
Statut :	Listing directory /gs7k1/home/gmartin
Statut :	Contenu du dossier "/gs7k1/home/gmartin" affiché avec succès
```

Site local : C:\Users\gmartin\Desktop\

- Desktop
- Documents
- Downloads
- Favorites

Nom de fichier	Taille de fi...	Type de fichier	Dernière modificat...
..			
Divers		Dossier de fich...	11/01/2019 11:09:57
Nouveau dossier		Dossier de fich...	20/03/2018 11:03:25
PahangHDvsSchizocarpa		Dossier de fich...	05/10/2018 11:07:51
CitrusReadme.txt	945	Fichier TXT	07/01/2019 17:50:43
desktop.ini	282	Paramètres de ...	12/01/2018 08:17:57
draft.pdf	456 867	Adobe Acroba...	05/10/2018 15:23:21
FormationGuadeloupe2019....	27 775 640	Présentation ...	11/01/2019 11:12:18
journal.pone.0155740.PDF	3 331 949	Adobe Acroba...	10/11/2017 09:37:41
Livre 2016 Banana_+Geno...	15 619 560	Adobe Acroba...	24/11/2017 08:04:37
networkx_reference.pdf	1 421 082	Adobe Acroba...	09/02/2018 17:39:46
pysam.pdf	317 563	Adobe Acroba...	22/02/2018 13:38:09
STATISTIQUES POUR STAT...	2 523 782	Adobe Acroba...	03/09/2017 16:09:18
~$FormationGuadeloupe20...	165	Présentation ...	11/01/2019 09:17:54

10 fichiers et 3 dossiers. Taille totale : 51 447 835 octets

Site distant : /gs7k1/home/gmartin

- /
  - gs7k1
    - home
      - gmartin

Nom de fichier	Taille de ...	Type de fic...	Dernière modification	Droits d'ac...	Propriétair...
bioinfo-agap		Dossier de ...	20/11/2018 15:52:22	drwxr-xr-x	gmartin us...
BPT		Dossier de ...	06/12/2018 09:56:48	drwxr-xr-x	gmartin us...
canopy		Dossier de ...	11/01/2018 08:39:52	drwxr-xr-x	gmartin us...
Chloro		Dossier de ...	12/12/2016 09:06:05	drwxr-xr-x	gmartin us...
circos		Dossier de ...	12/06/2017 11:09:45	drwxr-xr-x	gmartin us...
Desktop		Dossier de ...	15/05/2017 11:36:19	drwxr-xr-x	gmartin us...
enthought		Dossier de ...	11/01/2018 08:36:29	drwxr-xr-x	gmartin us...
igv		Dossier de ...	06/11/2018 11:03:16	drwxr-xr-x	gmartin us...
Programmes		Dossier de ...	03/09/2018 15:14:45	drwxr-xr-x	gmartin us...
R		Dossier de ...	25/01/2017 13:38:27	drwxr-xr-x	gmartin us...
ReadmeBack		Dossier de ...	10/10/2018 09:26:24	drwxr-xr-x	gmartin us...
REFERENCES		Dossier de ...	22/03/2018 09:43:32	drwxr-xr-x	gmartin us...
script		Dossier de ...	12/12/2016 08:30:15	drwxr-xr-x	gmartin us...
script_new_cluster		Dossier de ...	12/06/2018 12:45:52	drwxr-xr-x	gmartin us...
wo		Dossier de ...	05/12/2016 11:02:36	lrwxrwxrwx	root root
Works		Dossier de ...	08/01/2019 15:17:59	drwxr-xr-x	gmartin us...
.bash_history	25 986	Fichier BAS...	10/01/2019 16:40:03	-rw-------	gmartin us...
.bash_logout	33	Fichier BAS...	05/12/2016 11:02:36	-rw-r--r--	gmartin us...
.bash_profile	799	Fichier BAS...	24/11/2017 10:24:21	-rw-r--r--	gmartin us...
.bashrc	124	Fichier BAS...	05/12/2016 11:02:36	-rw-r--r--	gmartin us...
.canopy_runtimes.json	163	Fichier JSON	11/01/2018 08:39:51	-rw-r--r--	gmartin us...
.enstaller4rc	2 487	Fichier ENS...	11/01/2018 08:39:52	-rw-r--r--	gmartin us...
.esd_auth	16	Fichier ESD...	25/07/2017 16:54:55	-rw-------	gmartin us...
.gitconfig	58	Fichier GIT...	30/01/2018 10:52:20	-rw-r--r--	gmartin us...
.gnuplot_history	5	Fichier GN...	19/06/2017 15:46:02	-rw-------	gmartin us...
.pulse-cookie	256	Fichier PUL...	25/07/2017 16:54:54	-rw-------	gmartin us...

22 fichiers et 39 dossiers. Taille totale : 8 648 435 octets

Serveur / Fichier local	Direction	Fichier distant	Taille	Priorité	Statut

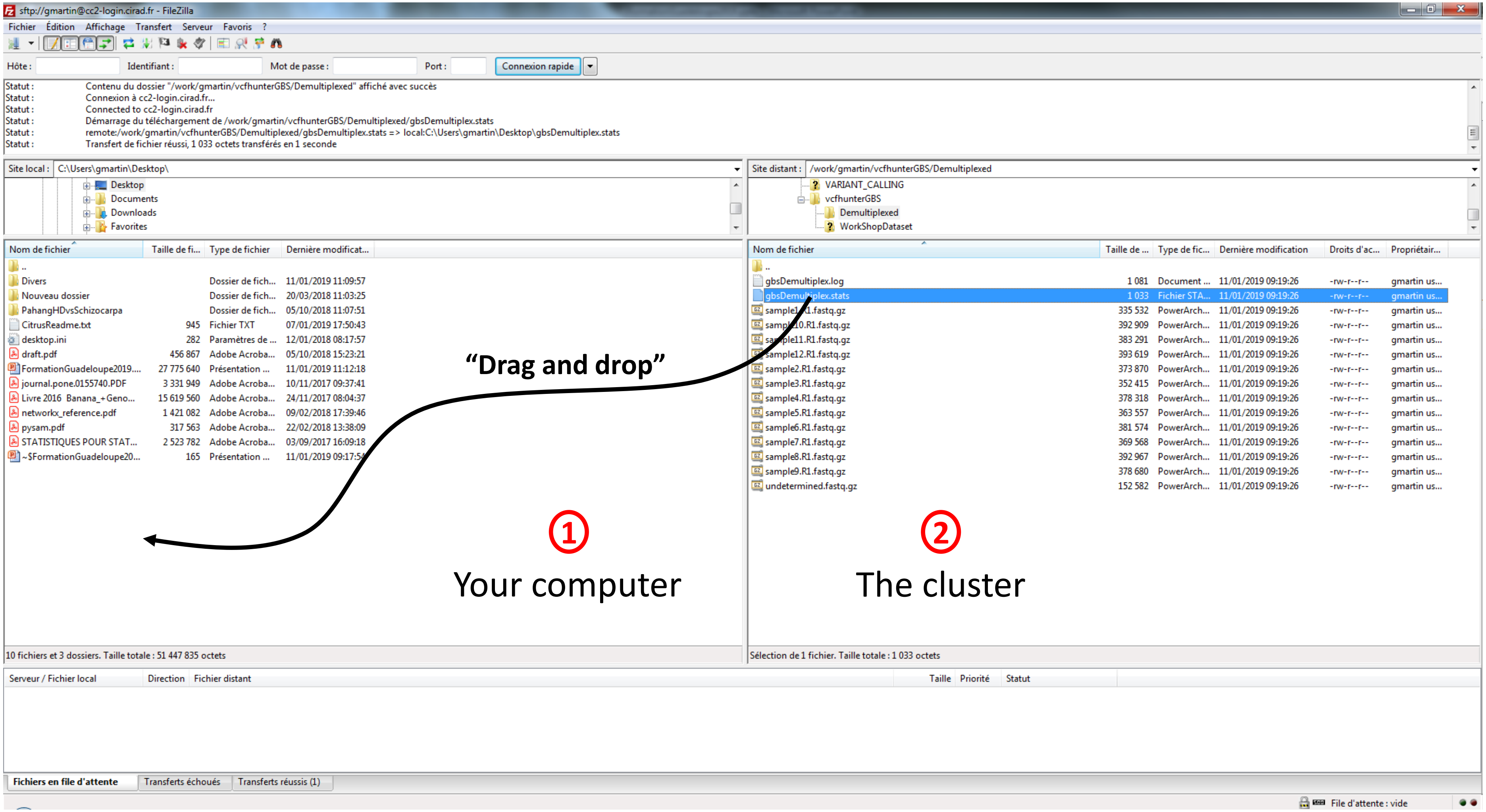
Fichiers en file d'attente | Transferts échoués | Transferts réussis

File d'attente : vide

- Go to the Demultiplexed folder: work ➔ vcfhunterGBS ➔ Demultiplexed

From the output of GBS sequencing to the variant calling file: in command line

- The gbsDemultiplex.log file:


sftp://gmartin@cc2-login.cirad.fr - FileZilla
Fichier Édition Affichage Transfert Serveur Favoris ?
Hôte :
Identifiant :
Mot de passe :
Port :
Connexion rapide

```
Statut :        Contenu du dossier "/work/gmartin/vcfhunterGBS/Demultiplexed" affiché avec succès
Statut :        Connexion à cc2-login.cirad.fr...
Statut :        Connected to cc2-login.cirad.fr
Statut :        Démarrage du téléchargement de /work/gmartin/vcfhunterGBS/Demultiplexed/gbsDemultiplex.stats
Statut :        remote:/work/gmartin/vcfhunterGBS/Demultiplexed/gbsDemultiplex.stats => local:C:\Users\gmartin\Desktop\gbsDemultiplex.stats
Statut :        Transfert de fichier réussi, 1 033 octets transférés en 1 seconde
```

Site local : C:\Users\gmartin\Desktop\

Nom de fichier	Taille de fi...	Type de fichier	Dernière modificat...
..			
Divers		Dossier de fich...	11/01/2019 11:09:57
Nouveau dossier		Dossier de fich...	20/03/2018 11:03:25
PahangHDvsSchizocarpa		Dossier de fich...	05/10/2018 11:07:51
CitrusReadme.txt	945	Fichier TXT	07/01/2019 17:50:43
desktop.ini	282	Paramètres de ...	12/01/2018 08:17:57
draft.pdf	456 867	Adobe Acroba...	05/10/2018 15:23:21
FormationGuadeloupe2019....	27 775 640	Présentation ...	11/01/2019 11:12:18
journal.pone.0155740.PDF	3 331 949	Adobe Acroba...	10/11/2017 09:37:41
Livre 2016 Banana_+Geno...	15 619 560	Adobe Acroba...	24/11/2017 08:04:37
networkx_reference.pdf	1 421 082	Adobe Acroba...	09/02/2018 17:39:46
pysam.pdf	317 563	Adobe Acroba...	22/02/2018 13:38:09
STATISTIQUES POUR STAT...	2 523 782	Adobe Acroba...	03/09/2017 16:09:18
~$FormationGuadeloupe20...	165	Présentation ...	11/01/2019 09:17:54

10 fichiers et 3 dossiers. Taille totale : 51 447 835 octets

Site distant : /work/gmartin/vcfhunterGBS/Demultiplexed

Nom de fichier	Taille de ...	Type de fic...	Dernière modification	Droits d'ac...	Propriétair...
..					
gbsDemultiplex.log	1 081	Document ...	11/01/2019 09:19:26	-rw-r--r--	gmartin us...
gbsDemultiplex.stats	1 033	Fichier STA...	11/01/2019 09:19:26	-rw-r--r--	gmartin us...
sample1.R1.fastq.gz	335 532	PowerArch...	11/01/2019 09:19:26	-rw-r--r--	gmartin us...
sample10.R1.fastq.gz	392 909	PowerArch...	11/01/2019 09:19:26	-rw-r--r--	gmartin us...
sample11.R1.fastq.gz	383 291	PowerArch...	11/01/2019 09:19:26	-rw-r--r--	gmartin us...
sample12.R1.fastq.gz	393 619	PowerArch...	11/01/2019 09:19:26	-rw-r--r--	gmartin us...
sample2.R1.fastq.gz	373 870	PowerArch...	11/01/2019 09:19:26	-rw-r--r--	gmartin us...
sample3.R1.fastq.gz	352 415	PowerArch...	11/01/2019 09:19:26	-rw-r--r--	gmartin us...
sample4.R1.fastq.gz	378 318	PowerArch...	11/01/2019 09:19:26	-rw-r--r--	gmartin us...
sample5.R1.fastq.gz	363 557	PowerArch...	11/01/2019 09:19:26	-rw-r--r--	gmartin us...
sample6.R1.fastq.gz	381 574	PowerArch...	11/01/2019 09:19:26	-rw-r--r--	gmartin us...
sample7.R1.fastq.gz	369 568	PowerArch...	11/01/2019 09:19:26	-rw-r--r--	gmartin us...
sample8.R1.fastq.gz	392 967	PowerArch...	11/01/2019 09:19:26	-rw-r--r--	gmartin us...
sample9.R1.fastq.gz	378 680	PowerArch...	11/01/2019 09:19:26	-rw-r--r--	gmartin us...
undetermined.fastq.gz	152 582	PowerArch...	11/01/2019 09:19:26	-rw-r--r--	gmartin us...

Sélection de 1 fichier. Taille totale : 1 033 octets

“Drag and drop”

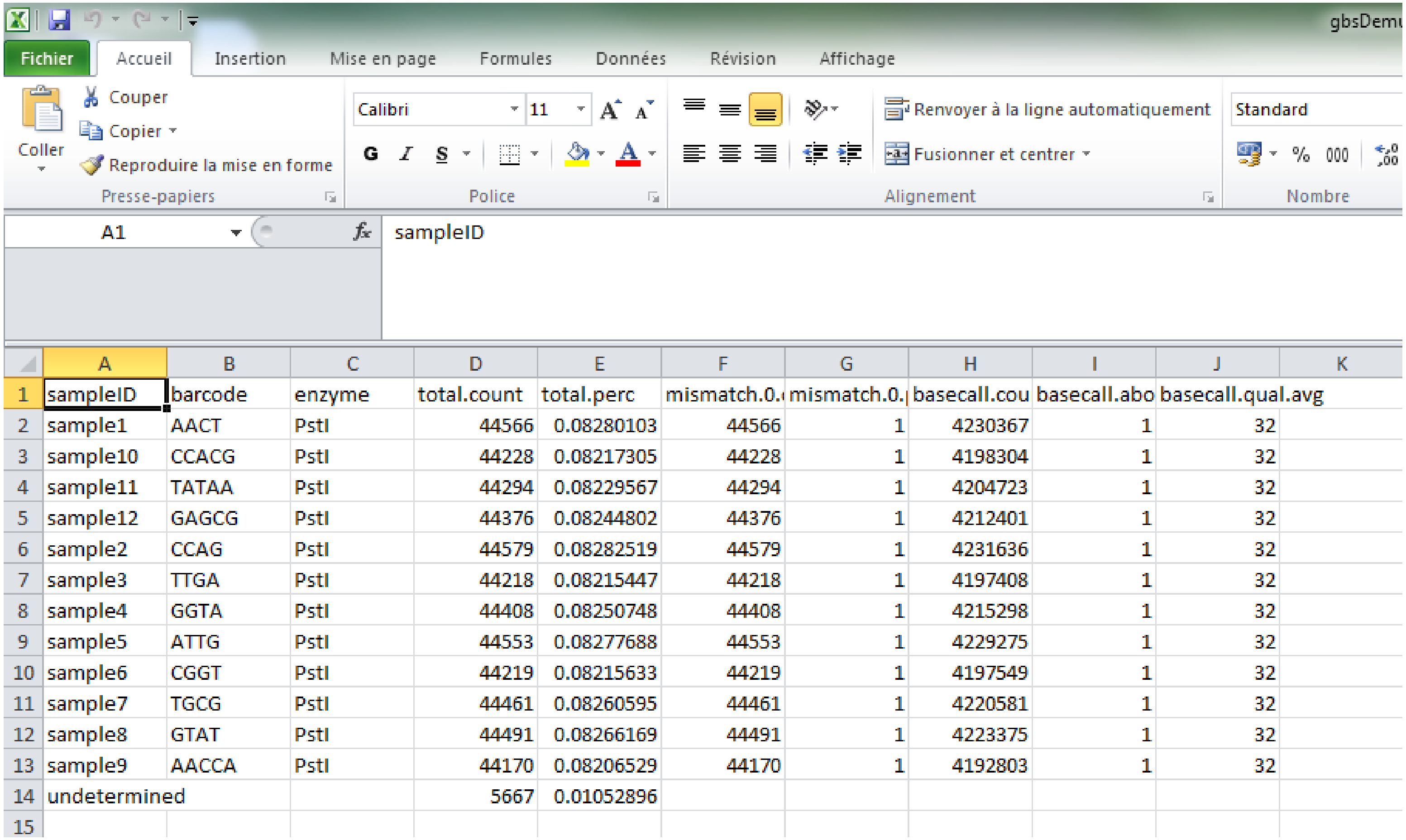
① Your computer

② The cluster

- Your file has been copied to your desktop.
- Open it with Excel!

From the output of GBS sequencing to the variant calling file: in command line

- The gbsDemultiplex.log file:

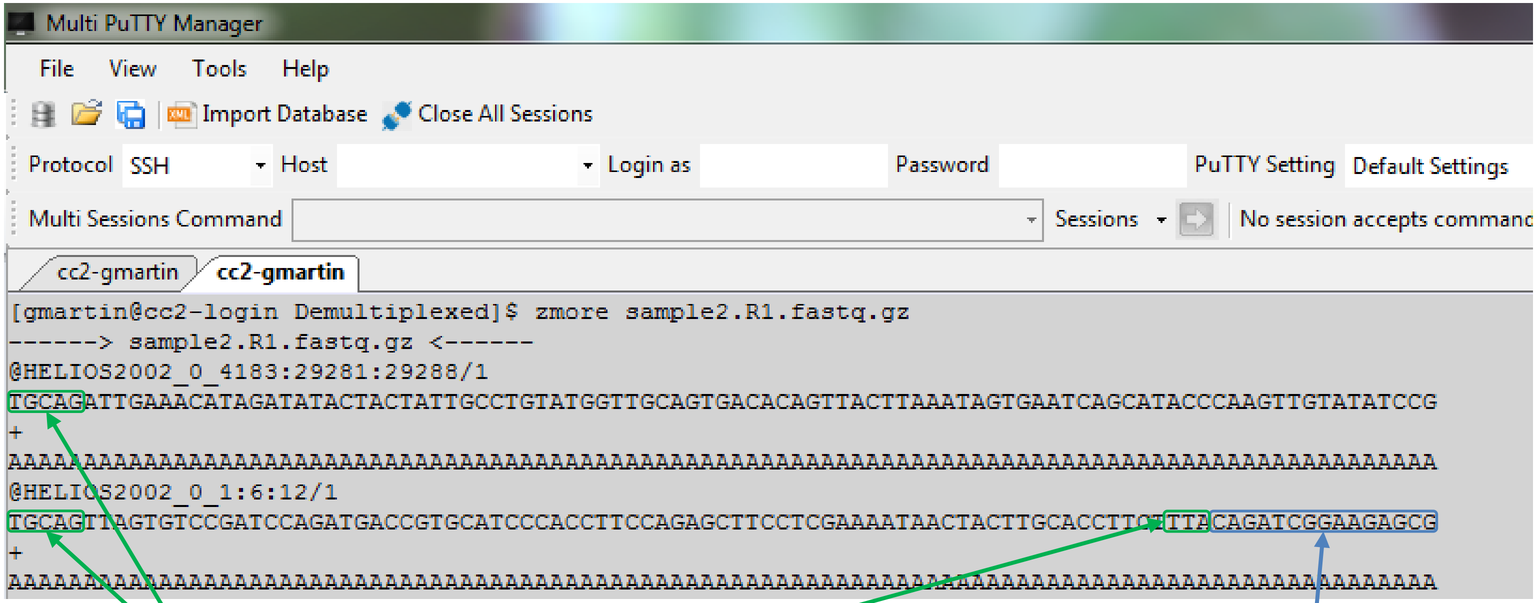


	A	B	C	D	E	F	G	H	I	J	K
1	sampleID	barcode	enzyme	total.count	total.perc	mismatch.0.	mismatch.0.	basecall.cou	basecall.abo	basecall.qual.avg	
2	sample1	AACT	PstI	44566	0.08280103	44566	1	4230367	1	32	
3	sample10	CCACG	PstI	44228	0.08217305	44228	1	4198304	1	32	
4	sample11	TATAA	PstI	44294	0.08229567	44294	1	4204723	1	32	
5	sample12	GAGCG	PstI	44376	0.08244802	44376	1	4212401	1	32	
6	sample2	CCAG	PstI	44579	0.08282519	44579	1	4231636	1	32	
7	sample3	TTGA	PstI	44218	0.08215447	44218	1	4197408	1	32	
8	sample4	GGTA	PstI	44408	0.08250748	44408	1	4215298	1	32	
9	sample5	ATTG	PstI	44553	0.08277688	44553	1	4229275	1	32	
10	sample6	CGGT	PstI	44219	0.08215633	44219	1	4197549	1	32	
11	sample7	TGCG	PstI	44461	0.08260595	44461	1	4220581	1	32	
12	sample8	GTAT	PstI	44491	0.08266169	44491	1	4223375	1	32	
13	sample9	AACCA	PstI	44170	0.08206529	44170	1	4192803	1	32	
14	undetermined			5667	0.01052896						
15											

From the output of GBS sequencing to the variant calling file: in command line

- The sample***X***.R1.fastq.gz files: For example sample2.R1.fastq.gz

```
zmore sample2.R1.fastq.gz
```



```
[gmartin@cc2-login Demultiplexed]$ zmore sample2.R1.fastq.gz
------> sample2.R1.fastq.gz <------
@HELIOS2002_0_4183:29281:29288/1
TGCAGATTGAAACATAGATATACTACTATTGCCTGTATGGTTGCAGTGACACAGTTACTTAAATAGTGAATCAGCATACCCAAGTTGTATATCCG
+
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
@HELIOS2002_0_1:6:12/1
TGCAGTTAGTGTCCGATCCAGATGACCGTGCATCCCACCTTCCAGAGCTTCCTCGAAAATAACTACTTGCACCTTCTTTACAGATCGGAAGAGCG
+
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
```

msel restriction site

pstI restriction site

Adapter sequence

- Sample tags were removed from reads
- Illumina adapters are still present at the end of some read (i.e. when sequenced fragments are shorter than illumina reads) ➔ These adapters should be removed as they do not belong to the sample!

From the output of GBS sequencing to the variant calling file: in command line

- Removing adapters and quality trimming of read.

The quality trimming is not necessary here as this is simulated reads with top quality but in reality as sequencing quality decrease along a read this is necessary.

- For that we will use cutadapt (https://cutadapt.readthedocs.io/en/stable/guide.html, https://doi.org/10.14806/ej.17.1.200)

- To load cutadapt:

```
module purge
module load bioinfo/cutadapt/1.8.1
module load system/python/3.4.3
```

To remove already loaded modules (prevent conflicts)

The cutadapt module

cutadapt also required python module

- To use cutadapt on sample2, run the command line:

```
qsub -q normal.q -b yes -V -N CUTADAPT cutadapt -a CAGATCGGAAGAGCG
-O 10 -q 20,20 -f fastq -m 30 -o sample2.R1.fastq.gz.cut.gz
sample2.R1.fastq.gz
```

From the output of GBS sequencing to the variant calling file: in command line

- Command line explanation

```
qsub -q normal.q -b yes -V -N CUTADAPT cutadapt -a CAGATCGGAAGAGCG -O 10 -q 20,20 -f fastq -m 30 -o
sample2.R1.fastq.gz.cut.gz sample2.R1.fastq.gz
```

- The first part of the command line (in bold) is **used by the master computer** (as previously described):

  - **qsub**: Means that we will send a command that the master computer needs to analyze to choose the best computer

  - **-q normal.q**: tells the master computer that we will use computer from normal queue.

  - **-b yes**: it is not important, but put it.

  - **-V**: Tell the master computer to load the module previously loaded on the computer it will choose to run the program

  - **-N CUTADAPT**: A name passed to the command line to look at its status (waiting, running or error) on the cluster

From the output of GBS sequencing to the variant calling file: in command line

- Command line explanation

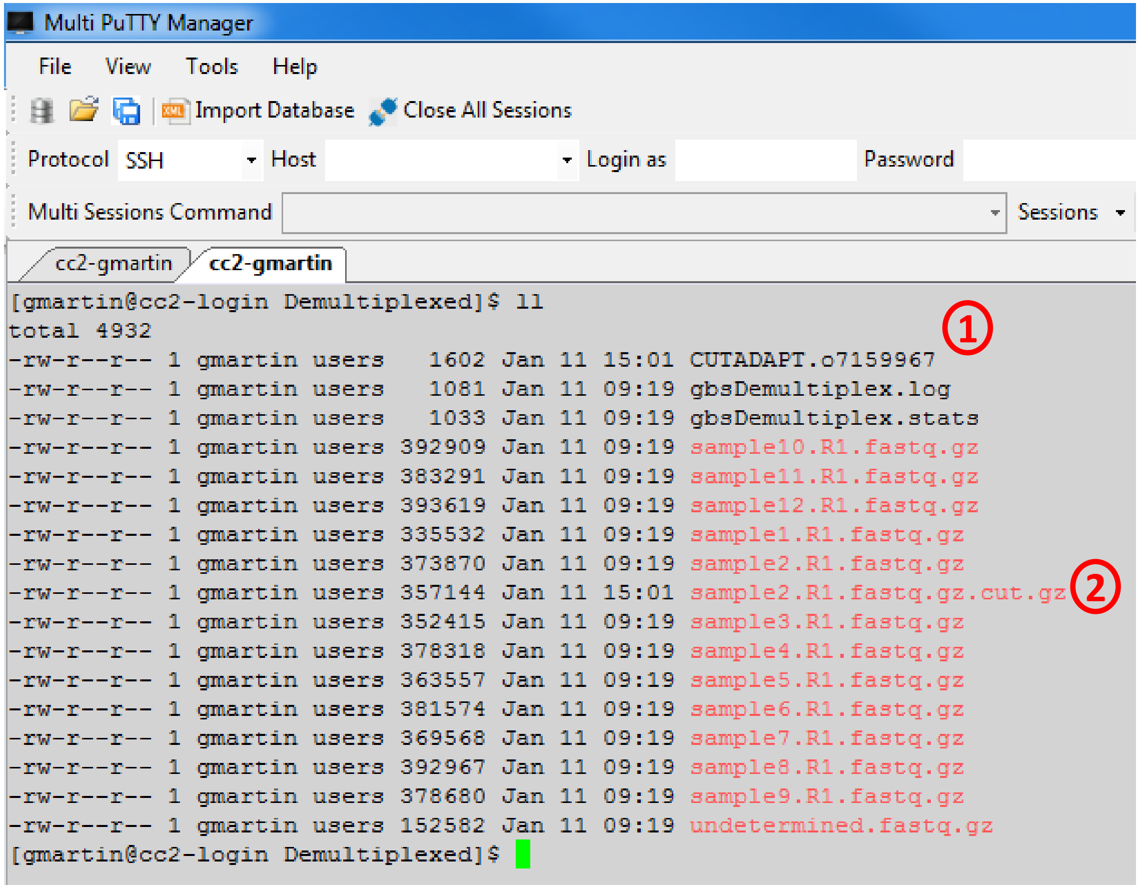
```
qsub -q normal.q -b yes -V -N CUTADAPT cutadapt -a CAGATCGGAAGAGCG -O 10 -q 20,20 -f fastq -m 30 -o
sample2.R1.fastq.gz.cut.gz sample2.R1.fastq.gz
```

- The part of the command line between quotation marks (in bold) is the command line that is executed on the **computer chosen by the master computer**.
  - **cutadapt**: tell that we will be using cutadapt program
  - **-a CAGATCGGAAGAGCG**: tells cutadapt that it should look for adapter sequence at 3’ end and that it should remove this sequence and all that follows.
  - **-O 10**: If the overlap between the read and the adapter is shorter than 10, the read is not modified. This reduces the no. of bases trimmed purely due to short random adapter matches
  - **-q 20,20**: Trim the 5’ and the 3’ until a base quality of 20 is reached
  - **-f fastq** : The input format file is fastq
  - **-m 30** : only read equal or greater than 30 bases will be conserved
  - **-o sample2.R1.fastq.gz.cut.gz**: Name of the output file
  - **sample2.R1.fastq.gz**: Name of the file processed by cutadapt

From the output of GBS sequencing to the variant calling file: in command line

- Outputs: To visualize new file generated, list the files in the repository:

```
ll
```



```
[gmartin@cc2-login Demultiplexed]$ ll
total 4932
-rw-r--r-- 1 gmartin users   1602 Jan 11 15:01 CUTADAPT.o7159967
-rw-r--r-- 1 gmartin users   1081 Jan 11 09:19 gbsDemultiplex.log
-rw-r--r-- 1 gmartin users   1033 Jan 11 09:19 gbsDemultiplex.stats
-rw-r--r-- 1 gmartin users 392909 Jan 11 09:19 sample10.R1.fastq.gz
-rw-r--r-- 1 gmartin users 383291 Jan 11 09:19 sample11.R1.fastq.gz
-rw-r--r-- 1 gmartin users 393619 Jan 11 09:19 sample12.R1.fastq.gz
-rw-r--r-- 1 gmartin users 335532 Jan 11 09:19 sample1.R1.fastq.gz
-rw-r--r-- 1 gmartin users 373870 Jan 11 09:19 sample2.R1.fastq.gz
-rw-r--r-- 1 gmartin users 357144 Jan 11 15:01 sample2.R1.fastq.gz.cut.gz
-rw-r--r-- 1 gmartin users 352415 Jan 11 09:19 sample3.R1.fastq.gz
-rw-r--r-- 1 gmartin users 378318 Jan 11 09:19 sample4.R1.fastq.gz
-rw-r--r-- 1 gmartin users 363557 Jan 11 09:19 sample5.R1.fastq.gz
-rw-r--r-- 1 gmartin users 381574 Jan 11 09:19 sample6.R1.fastq.gz
-rw-r--r-- 1 gmartin users 369568 Jan 11 09:19 sample7.R1.fastq.gz
-rw-r--r-- 1 gmartin users 392967 Jan 11 09:19 sample8.R1.fastq.gz
-rw-r--r-- 1 gmartin users 378680 Jan 11 09:19 sample9.R1.fastq.gz
-rw-r--r-- 1 gmartin users 152582 Jan 11 09:19 undetermined.fastq.gz
[gmartin@cc2-login Demultiplexed]$
```

Two files have been generated:

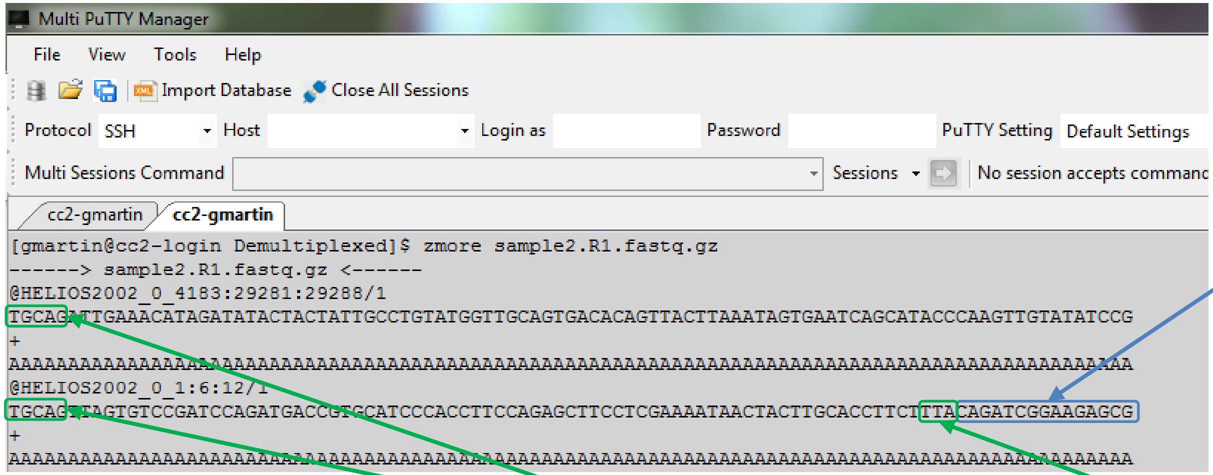
1. The CUTADAPT.oxxxxxxx file containing what cutadapt told us while it was executing

2. The sample2.R1.fastq.gz.cut.gz containing filtered read

From the output of GBS sequencing to the variant calling file: in command line

- The sample2.R1.fastq.gz file before cutadapt:

```
zmore sample2.R1.fastq.gz
```



```
[gmartin@cc2-login Demultiplexed]$ zmore sample2.R1.fastq.gz
------> sample2.R1.fastq.gz <------
@HELIOS2002_0_4183:29281:29288/1
TGCAGATTGAAACATAGATATACTACTATTGCCTGTATGGTTGCAGTGACACAGTTACTTAAATAGTGAATCAGCATACCCAAGTTGTATATCCG
+
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
@HELIOS2002_0_1:6:12/1
TGCAGTTAGTGTCCGATCCAGATGACCGTGCATCCCACCTTCCAGAGCTTCCTCGAAAATAACTACTTGCACCTTCTTTACAGATCGGAAGAGCG
+
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
```

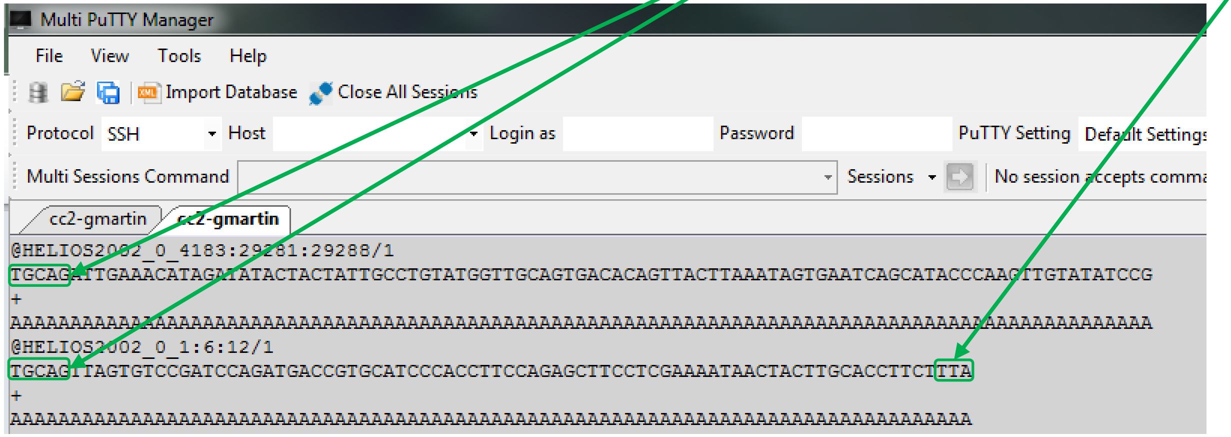
Adapter sequence

- And After cutadapt

```
zmore sample2.R1.fastq.gz.cut.gz
```

pstI restriction site

msel restriction site

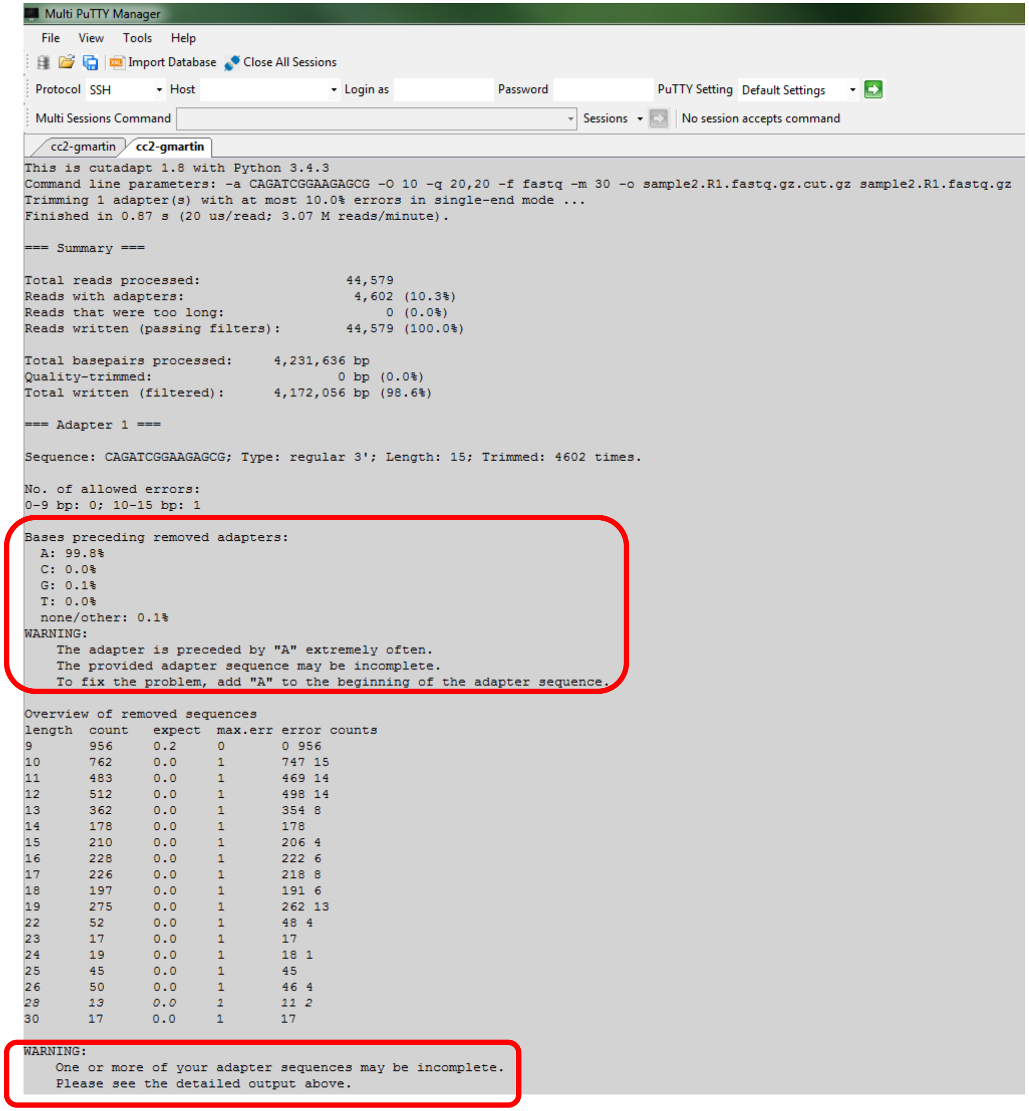


```
@HELIOS2002_0_4183:29281:29288/1
TGCAGATTGAAACATAGATATACTACTATTGCCTGTATGGTTGCAGTGACACAGTTACTTAAATAGTGAATCAGCATACCCAAGTTGTATATCCG
+
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
@HELIOS2002_0_1:6:12/1
TGCAGTTAGTGTCCGATCCAGATGACCGTGCATCCCACCTTCCAGAGCTTCCTCGAAAATAACTACTTGCACCTTCTTTA
+
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
```

From the output of GBS sequencing to the variant calling file: in command line

- The CUTADAPT.oxxxxxxx file:

```
zmore CUTADAPT.oxxxxxxx
```



```
This is cutadapt 1.8 with Python 3.4.3
Command line parameters: -a CAGATCGGAAGAGCG -O 10 -q 20,20 -f fastq -m 30 -o sample2.R1.fastq.gz.cut.gz sample2.R1.fastq.gz
Trimming 1 adapter(s) with at most 10.0% errors in single-end mode ...
Finished in 0.87 s (20 us/read; 3.07 M reads/minute).

=== Summary ===

Total reads processed:                  44,579
Reads with adapters:                     4,602 (10.3%)
Reads that were too long:                    0 (0.0%)
Reads written (passing filters):        44,579 (100.0%)

Total basepairs processed:     4,231,636 bp
Quality-trimmed:                       0 bp (0.0%)
Total written (filtered):      4,172,056 bp (98.6%)

=== Adapter 1 ===

Sequence: CAGATCGGAAGAGCG; Type: regular 3'; Length: 15; Trimmed: 4602 times.

No. of allowed errors:
0-9 bp: 0; 10-15 bp: 1

Bases preceding removed adapters:
  A: 99.8%
  C: 0.0%
  G: 0.1%
  T: 0.0%
  none/other: 0.1%
WARNING:
    The adapter is preceded by "A" extremely often.
    The provided adapter sequence may be incomplete.
    To fix the problem, add "A" to the beginning of the adapter sequence.

Overview of removed sequences
length  count   expect  max.err error counts
9       956     0.2     0       0 956
10      762     0.0     1       747 15
11      483     0.0     1       469 14
12      512     0.0     1       498 14
13      362     0.0     1       354 8
14      178     0.0     1       178
15      210     0.0     1       206 4
16      228     0.0     1       222 6
17      226     0.0     1       218 8
18      197     0.0     1       191 6
19      275     0.0     1       262 13
22      52      0.0     1       48 4
23      17      0.0     1       17
24      19      0.0     1       18 1
25      45      0.0     1       45
26      50      0.0     1       46 4
28      13      0.0     1       11 2
30      17      0.0     1       17

WARNING:
    One or more of your adapter sequences may be incomplete.
    Please see the detailed output above.
```

There is a warning saying that maybe the adapter sequence is incomplete because very often (99.8% of cases), when an adapter is found, the “A” base was found just before…

This is normal because just before the adapter we have our *mseI* restriction site

From the output of GBS sequencing to the variant calling file: in command line

- This command line should be adapted and executed for each sample:

```
qsub -q normal.q -b yes -V -N CUTADAPT cutadapt -a CAGATCGGAAGAGCG -O 10 -q 20,20 -f fastq -m 30 -o sample10.R1.fastq.gz.cut.gz sample10.R1.fastq.gz
qsub -q normal.q -b yes -V -N CUTADAPT cutadapt -a CAGATCGGAAGAGCG -O 10 -q 20,20 -f fastq -m 30 -o sample11.R1.fastq.gz.cut.gz sample11.R1.fastq.gz
qsub -q normal.q -b yes -V -N CUTADAPT cutadapt -a CAGATCGGAAGAGCG -O 10 -q 20,20 -f fastq -m 30 -o sample12.R1.fastq.gz.cut.gz sample12.R1.fastq.gz
qsub -q normal.q -b yes -V -N CUTADAPT cutadapt -a CAGATCGGAAGAGCG -O 10 -q 20,20 -f fastq -m 30 -o sample1.R1.fastq.gz.cut.gz sample1.R1.fastq.gz
qsub -q normal.q -b yes -V -N CUTADAPT cutadapt -a CAGATCGGAAGAGCG -O 10 -q 20,20 -f fastq -m 30 -o sample2.R1.fastq.gz.cut.gz sample2.R1.fastq.gz
qsub -q normal.q -b yes -V -N CUTADAPT cutadapt -a CAGATCGGAAGAGCG -O 10 -q 20,20 -f fastq -m 30 -o sample3.R1.fastq.gz.cut.gz sample3.R1.fastq.gz
qsub -q normal.q -b yes -V -N CUTADAPT cutadapt -a CAGATCGGAAGAGCG -O 10 -q 20,20 -f fastq -m 30 -o sample4.R1.fastq.gz.cut.gz sample4.R1.fastq.gz
qsub -q normal.q -b yes -V -N CUTADAPT cutadapt -a CAGATCGGAAGAGCG -O 10 -q 20,20 -f fastq -m 30 -o sample5.R1.fastq.gz.cut.gz sample5.R1.fastq.gz
qsub -q normal.q -b yes -V -N CUTADAPT cutadapt -a CAGATCGGAAGAGCG -O 10 -q 20,20 -f fastq -m 30 -o sample6.R1.fastq.gz.cut.gz sample6.R1.fastq.gz
qsub -q normal.q -b yes -V -N CUTADAPT cutadapt -a CAGATCGGAAGAGCG -O 10 -q 20,20 -f fastq -m 30 -o sample7.R1.fastq.gz.cut.gz sample7.R1.fastq.gz
qsub -q normal.q -b yes -V -N CUTADAPT cutadapt -a CAGATCGGAAGAGCG -O 10 -q 20,20 -f fastq -m 30 -o sample8.R1.fastq.gz.cut.gz sample8.R1.fastq.gz
qsub -q normal.q -b yes -V -N CUTADAPT cutadapt -a CAGATCGGAAGAGCG -O 10 -q 20,20 -f fastq -m 30 -o sample9.R1.fastq.gz.cut.gz sample9.R1.fastq.gz
```

- This is relatively easy when we have few files but when this should be done on hundreds of files it is a bit annoying… This can be solved with “for” loop in bash programing!

- Here is the command line for our example (advanced programing!):

```
for i in *.fastq.gz

do qsub -q normal.q -b yes -V -N CUTADAPT cutadapt -a
CAGATCGGAAGAGCG -O 10 -q 20,20 -f fastq -m 30 -o $i.cut.gz $i

done
```

From the output of GBS sequencing to the variant calling file: in command line

- This command line should be adapted and executed for each sample:

```
qsub -q normal.q -b yes -V -N CUTADAPT cutadapt -a CAGATCGGAAGAGCG -O 10 -q 20,20 -f fastq -m 30 -o sample10.R1.fastq.gz.cut.gz sample10.R1.fastq.gz
qsub -q normal.q -b yes -V -N CUTADAPT cutadapt -a CAGATCGGAAGAGCG -O 10 -q 20,20 -f fastq -m 30 -o sample11.R1.fastq.gz.cut.gz sample11.R1.fastq.gz
qsub -q normal.q -b yes -V -N CUTADAPT cutadapt -a CAGATCGGAAGAGCG -O 10 -q 20,20 -f fastq -m 30 -o sample12.R1.fastq.gz.cut.gz sample12.R1.fastq.gz
qsub -q normal.q -b yes -V -N CUTADAPT cutadapt -a CAGATCGGAAGAGCG -O 10 -q 20,20 -f fastq -m 30 -o sample1.R1.fastq.gz.cut.gz sample1.R1.fastq.gz
qsub -q normal.q -b yes -V -N CUTADAPT cutadapt -a CAGATCGGAAGAGCG -O 10 -q 20,20 -f fastq -m 30 -o sample2.R1.fastq.gz.cut.gz sample2.R1.fastq.gz
qsub -q normal.q -b yes -V -N CUTADAPT cutadapt -a CAGATCGGAAGAGCG -O 10 -q 20,20 -f fastq -m 30 -o sample3.R1.fastq.gz.cut.gz sample3.R1.fastq.gz
qsub -q normal.q -b yes -V -N CUTADAPT cutadapt -a CAGATCGGAAGAGCG -O 10 -q 20,20 -f fastq -m 30 -o sample4.R1.fastq.gz.cut.gz sample4.R1.fastq.gz
qsub -q normal.q -b yes -V -N CUTADAPT cutadapt -a CAGATCGGAAGAGCG -O 10 -q 20,20 -f fastq -m 30 -o sample5.R1.fastq.gz.cut.gz sample5.R1.fastq.gz
qsub -q normal.q -b yes -V -N CUTADAPT cutadapt -a CAGATCGGAAGAGCG -O 10 -q 20,20 -f fastq -m 30 -o sample6.R1.fastq.gz.cut.gz sample6.R1.fastq.gz
qsub -q normal.q -b yes -V -N CUTADAPT cutadapt -a CAGATCGGAAGAGCG -O 10 -q 20,20 -f fastq -m 30 -o sample7.R1.fastq.gz.cut.gz sample7.R1.fastq.gz
qsub -q normal.q -b yes -V -N CUTADAPT cutadapt -a CAGATCGGAAGAGCG -O 10 -q 20,20 -f fastq -m 30 -o sample8.R1.fastq.gz.cut.gz sample8.R1.fastq.gz
qsub -q normal.q -b yes -V -N CUTADAPT cutadapt -a CAGATCGGAAGAGCG -O 10 -q 20,20 -f fastq -m 30 -o sample9.R1.fastq.gz.cut.gz sample9.R1.fastq.gz
```

- This is relatively easy when we have few files but when this should be done on hundreds of files it is a bit annoying… This can be solved with “for” loop in bash programing!

- Here is the command line for our example (advanced programing!):

```
for i in *.fastq.gz
```

Initiation of a loop: For all files in the folder finishing by “.fastq.gz”…

Their name is sequentially stored in a variable “i”, and, for each values “i” (each read sample files), the cutadapt command line is executed on the file recorded in the variable i (`$i`) and the output is stored in a file called i+”.cut.gz” (`$i.cut.gz`) .
For example when i = `sample10.R1.fastq.gz : $i.cut.gz = sample10.R1.fastq.gz.cut.gz`

```
do qsub -q normal.q -b yes -V -N CUTADAPT cutadapt -a
CAGATCGGAAGAGCG -O 10 -q 20,20 -f fastq -m 30 -o $i.cut.gz $i
```

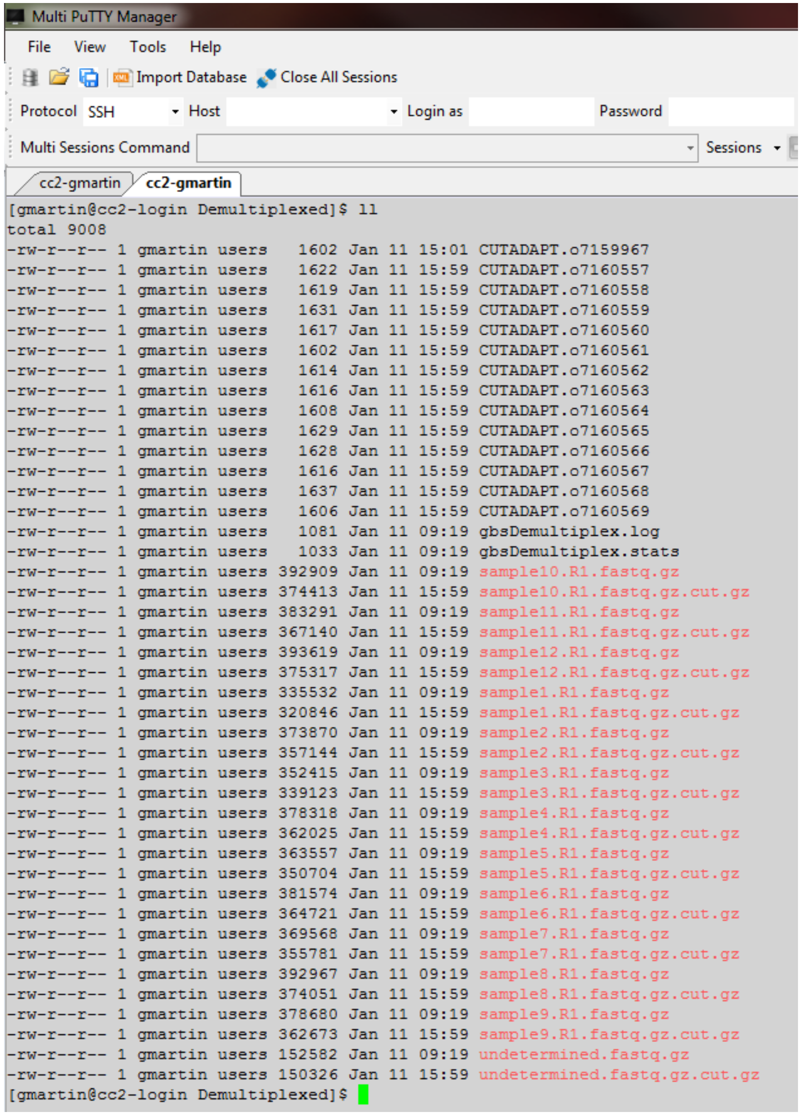
Tell that this is the end of the loop

```
done
```

From the output of GBS sequencing to the variant calling file: in command line

- Listing the files in the folder:

```
ll
```



```
[gmartin@cc2-login Demultiplexed]$ ll
total 9008
-rw-r--r-- 1 gmartin users   1602 Jan 11 15:01 CUTADAPT.o7159967
-rw-r--r-- 1 gmartin users   1622 Jan 11 15:59 CUTADAPT.o7160557
-rw-r--r-- 1 gmartin users   1619 Jan 11 15:59 CUTADAPT.o7160558
-rw-r--r-- 1 gmartin users   1631 Jan 11 15:59 CUTADAPT.o7160559
-rw-r--r-- 1 gmartin users   1617 Jan 11 15:59 CUTADAPT.o7160560
-rw-r--r-- 1 gmartin users   1602 Jan 11 15:59 CUTADAPT.o7160561
-rw-r--r-- 1 gmartin users   1614 Jan 11 15:59 CUTADAPT.o7160562
-rw-r--r-- 1 gmartin users   1616 Jan 11 15:59 CUTADAPT.o7160563
-rw-r--r-- 1 gmartin users   1608 Jan 11 15:59 CUTADAPT.o7160564
-rw-r--r-- 1 gmartin users   1629 Jan 11 15:59 CUTADAPT.o7160565
-rw-r--r-- 1 gmartin users   1628 Jan 11 15:59 CUTADAPT.o7160566
-rw-r--r-- 1 gmartin users   1616 Jan 11 15:59 CUTADAPT.o7160567
-rw-r--r-- 1 gmartin users   1637 Jan 11 15:59 CUTADAPT.o7160568
-rw-r--r-- 1 gmartin users   1606 Jan 11 15:59 CUTADAPT.o7160569
-rw-r--r-- 1 gmartin users   1081 Jan 11 09:19 gbsDemultiplex.log
-rw-r--r-- 1 gmartin users   1033 Jan 11 09:19 gbsDemultiplex.stats
-rw-r--r-- 1 gmartin users 392909 Jan 11 09:19 sample10.R1.fastq.gz
-rw-r--r-- 1 gmartin users 374413 Jan 11 15:59 sample10.R1.fastq.gz.cut.gz
-rw-r--r-- 1 gmartin users 383291 Jan 11 09:19 sample11.R1.fastq.gz
-rw-r--r-- 1 gmartin users 367140 Jan 11 15:59 sample11.R1.fastq.gz.cut.gz
-rw-r--r-- 1 gmartin users 393619 Jan 11 09:19 sample12.R1.fastq.gz
-rw-r--r-- 1 gmartin users 375317 Jan 11 15:59 sample12.R1.fastq.gz.cut.gz
-rw-r--r-- 1 gmartin users 335532 Jan 11 09:19 sample1.R1.fastq.gz
-rw-r--r-- 1 gmartin users 320846 Jan 11 15:59 sample1.R1.fastq.gz.cut.gz
-rw-r--r-- 1 gmartin users 373870 Jan 11 09:19 sample2.R1.fastq.gz
-rw-r--r-- 1 gmartin users 357144 Jan 11 15:59 sample2.R1.fastq.gz.cut.gz
-rw-r--r-- 1 gmartin users 352415 Jan 11 09:19 sample3.R1.fastq.gz
-rw-r--r-- 1 gmartin users 339123 Jan 11 15:59 sample3.R1.fastq.gz.cut.gz
-rw-r--r-- 1 gmartin users 378318 Jan 11 09:19 sample4.R1.fastq.gz
-rw-r--r-- 1 gmartin users 362025 Jan 11 15:59 sample4.R1.fastq.gz.cut.gz
-rw-r--r-- 1 gmartin users 363557 Jan 11 09:19 sample5.R1.fastq.gz
-rw-r--r-- 1 gmartin users 350704 Jan 11 15:59 sample5.R1.fastq.gz.cut.gz
-rw-r--r-- 1 gmartin users 381574 Jan 11 09:19 sample6.R1.fastq.gz
-rw-r--r-- 1 gmartin users 364721 Jan 11 15:59 sample6.R1.fastq.gz.cut.gz
-rw-r--r-- 1 gmartin users 369568 Jan 11 09:19 sample7.R1.fastq.gz
-rw-r--r-- 1 gmartin users 355781 Jan 11 15:59 sample7.R1.fastq.gz.cut.gz
-rw-r--r-- 1 gmartin users 392967 Jan 11 09:19 sample8.R1.fastq.gz
-rw-r--r-- 1 gmartin users 374051 Jan 11 15:59 sample8.R1.fastq.gz.cut.gz
-rw-r--r-- 1 gmartin users 378680 Jan 11 09:19 sample9.R1.fastq.gz
-rw-r--r-- 1 gmartin users 362673 Jan 11 15:59 sample9.R1.fastq.gz.cut.gz
-rw-r--r-- 1 gmartin users 152582 Jan 11 09:19 undetermined.fastq.gz
-rw-r--r-- 1 gmartin users 150326 Jan 11 15:59 undetermined.fastq.gz.cut.gz
[gmartin@cc2-login Demultiplexed]$
```

- A CUTADAPT.oxxxxxxx file has been generated per sample

- A filtered fastq file per sample has been generated per accessions

From the output of GBS sequencing to the variant calling file: in command line

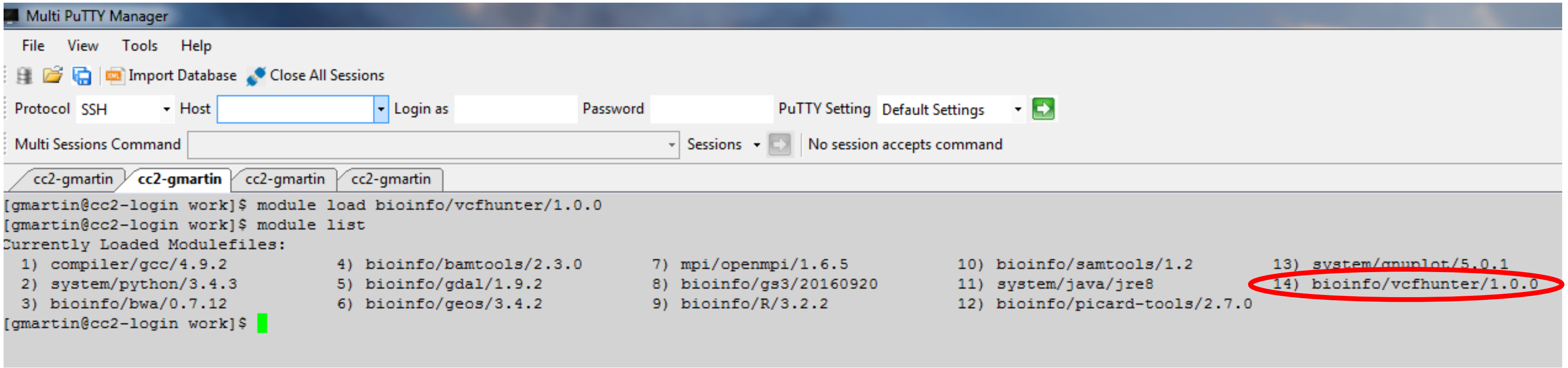
- We will use vcfHunter program which is installed on the AGAP cluster under module *"vcfhunter"*

- To load this module run the command line:

```
module purge
module load bioinfo/vcfhunter/1.0.0
```

- The module is now loaded. This can be verified with de following command line:

```
module list
```



Protocol SSH

PuTTY Setting Default Settings

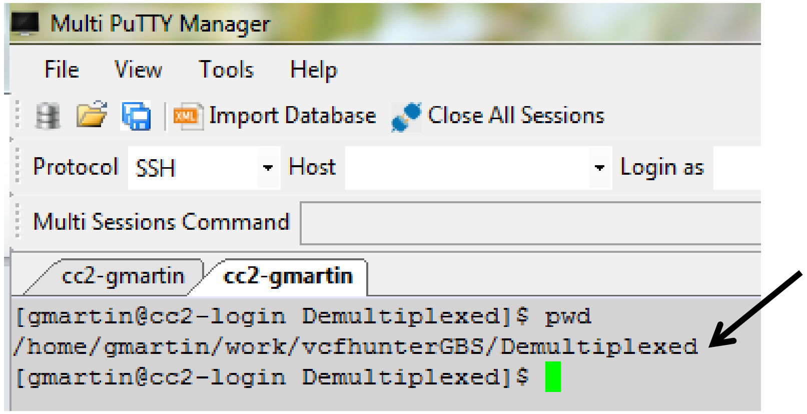
```
[gmartin@cc2-login work]$ module load bioinfo/vcfhunter/1.0.0
[gmartin@cc2-login work]$ module list
Currently Loaded Modulefiles:
  1) compiler/gcc/4.9.2          4) bioinfo/bamtools/2.3.0       7) mpi/openmpi/1.6.5          10) bioinfo/samtools/1.2          13) system/gnuplot/5.0.1
  2) system/python/3.4.3         5) bioinfo/gdal/1.9.2           8) bioinfo/gs3/20160920       11) system/java/jre8              14) bioinfo/vcfhunter/1.0.0
  3) bioinfo/bwa/0.7.12          6) bioinfo/geos/3.4.2           9) bioinfo/R/3.2.2            12) bioinfo/picard-tools/2.7.0
[gmartin@cc2-login work]$
```

- We can see that the vcfhunter module is loaded as well as several other modules which will be used by vcfhunter

From the output of GBS sequencing to the variant calling file: in command line

- We are going to work in a new folder for vcfHunter. This is not necessary but for file ordering, this will be better. But first where are we? To answer this question we use a simple command:

  ```
  pwd
  ```



```
[gmartin@cc2-login Demultiplexed]$ pwd
/home/gmartin/work/vcfhunterGBS/Demultiplexed
[gmartin@cc2-login Demultiplexed]$
```

This locate the path where you are when you execute the pwd command. Instead of "gmartin", you should have your login ID

- From there we want to go back to vcfhunterGBS folder. There are two possibility:

  ```
  cd /home/Your_ID/work/vcfhunterGBS
  ```

  change directory to /home/***Your_ID***/work/vcfhunterGBS

  Or

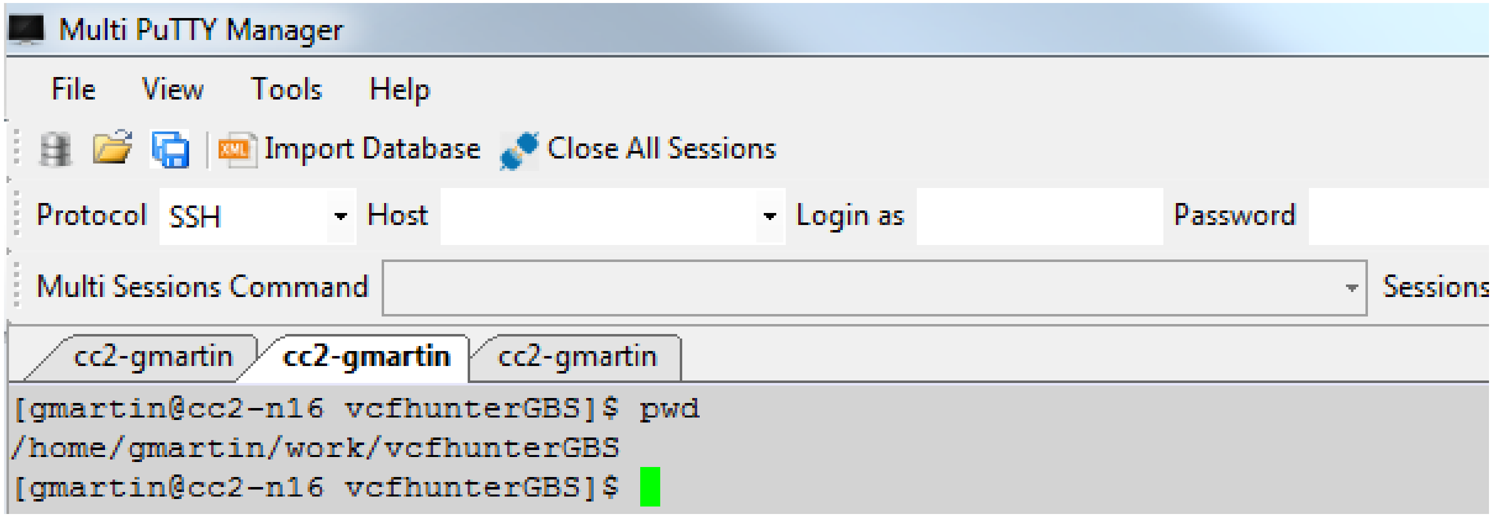
  ```
  cd ..
  ```

  change directory to one folder before. And one folder before there is vcfhunterGBS

From the output of GBS sequencing to the variant calling file: in command line

- Where are we now?

pwd



```
[gmartin@cc2-n16 vcfhunterGBS]$ pwd
/home/gmartin/work/vcfhunterGBS
[gmartin@cc2-n16 vcfhunterGBS]$
```

- Now we create the new folder

```
mkdir Mapping
```

- And we go into this folder

```
cd Mapping
```

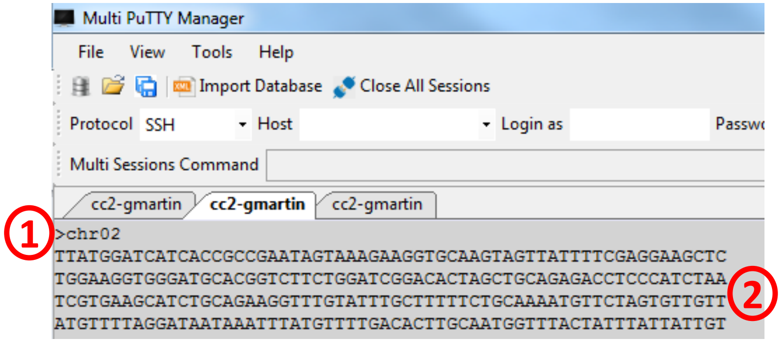
From the output of GBS sequencing to the variant calling file: in command line

- At this stage, we have 12 fastq files:
  - ✓ One for each samples, which comprised cleaned/filtered reads.
  - ✓ These files are located in a folder named `Demultiplexed`, located `/home/Your_ID/work/vcfhunterGBS`

- To run vcfHunter program, we also need an additional file which contained the reference sequence (in fasta format), on which we will align the reads. This file is already present in the `WorkShopDataset` folder located here: `/home/Your_ID/work/vcfhunterGBS/WorkShopDataset`. This file is named `Ref.fasta` (It is the folder you copied at the beginning of this exercise).

- Because at this time we are in the `Mapping` folder loacted `/home/Your_ID/work/vcfhunterGBS/Mapping`, to have a look at this file we should go back from one folder (`..`) to enter the `WorkShopDataset` folder and then access to `Ref.fasta` file. Thus, to have a look at this file:

```
more ../WorkShopDataset/Ref.fasta
```



```
>chr02
TTATGGATCATCACCGCCGAATAGTAAAGAAGGTGCAAGTAGTTATTTTCGAGGAAGCTC
TGGAAGGTGGGATGCACGGTCTTCTGGATCGGACACTAGCTGCAGAGACCTCCCATCTAA
TCGTGAAGCATCTGCAGAAGGTTTGTATTTGCTTTTTCTGCAAAATGTTCTAGTGTTGTT
ATGTTTTAGGATAATAAATTTATGTTTTGACACTTGCAATGGTTTACTATTTATTATTGT
```

Standard fasta format with each sequences beginning with a “>”+sequence name①, followed by DNA sequence②.

From the output of GBS sequencing to the variant calling file: in command line

- The sample fastq read file and reference fasta files should be passed recorded in a configuration unique file which will be given to ***vcfHunter*** program.

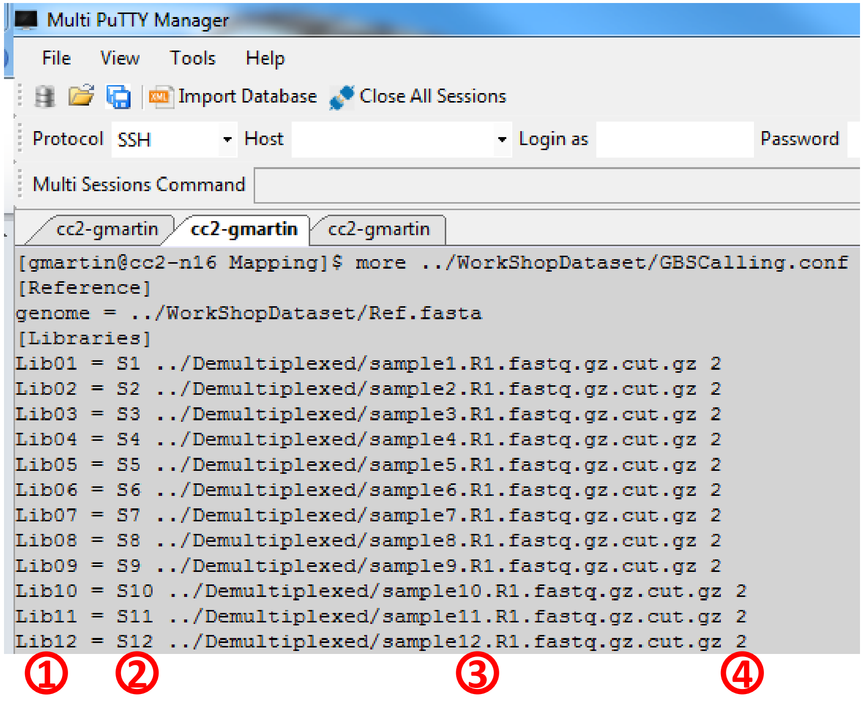
- For this example, the configuration file (`GBSCalling.conf`) has already been created can be found here: `/home/Your_ID/work/vcfhunterGBS/WorkShopDataset`. To have a look at this file and because we are in the Mapping folder we just created:

```
more ../WorkShopDataset/GBSCalling.conf
```

A [Reference] section locating how to access reference fasta.

A [Libraries] section locating how to access sample fastq reads files and additional information to sample:

① Unique ID for each fastq
② Sample Name (Name that will appear in the vcf)
③ How to access to the fastq read file
④ Accession ploidy



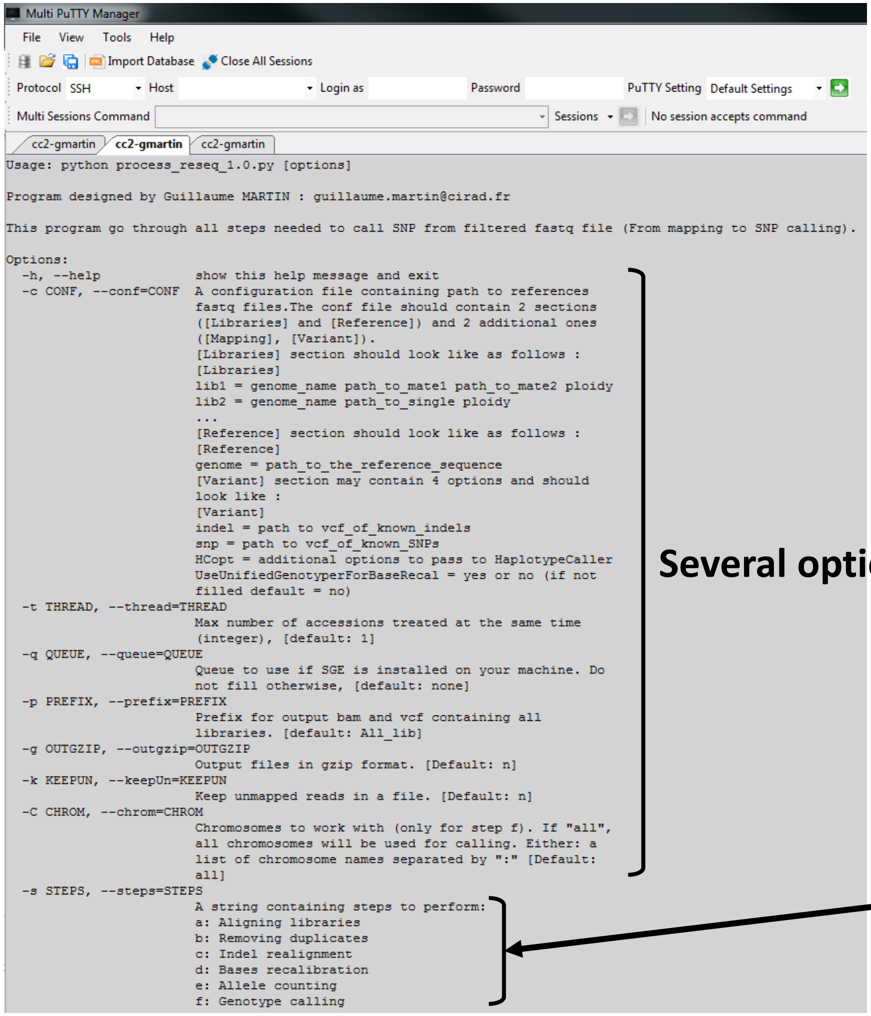
```
[gmartin@cc2-n16 Mapping]$ more ../WorkShopDataset/GBSCalling.conf
[Reference]
genome = ../WorkShopDataset/Ref.fasta
[Libraries]
Lib01 = S1 ../Demultiplexed/sample1.R1.fastq.gz.cut.gz 2
Lib02 = S2 ../Demultiplexed/sample2.R1.fastq.gz.cut.gz 2
Lib03 = S3 ../Demultiplexed/sample3.R1.fastq.gz.cut.gz 2
Lib04 = S4 ../Demultiplexed/sample4.R1.fastq.gz.cut.gz 2
Lib05 = S5 ../Demultiplexed/sample5.R1.fastq.gz.cut.gz 2
Lib06 = S6 ../Demultiplexed/sample6.R1.fastq.gz.cut.gz 2
Lib07 = S7 ../Demultiplexed/sample7.R1.fastq.gz.cut.gz 2
Lib08 = S8 ../Demultiplexed/sample8.R1.fastq.gz.cut.gz 2
Lib09 = S9 ../Demultiplexed/sample9.R1.fastq.gz.cut.gz 2
Lib10 = S10 ../Demultiplexed/sample10.R1.fastq.gz.cut.gz 2
Lib11 = S11 ../Demultiplexed/sample11.R1.fastq.gz.cut.gz 2
Lib12 = S12 ../Demultiplexed/sample12.R1.fastq.gz.cut.gz 2
```

- Possible to generate this file with a loop for Those who want to try!

From the output of GBS sequencing to the variant calling file: in command line

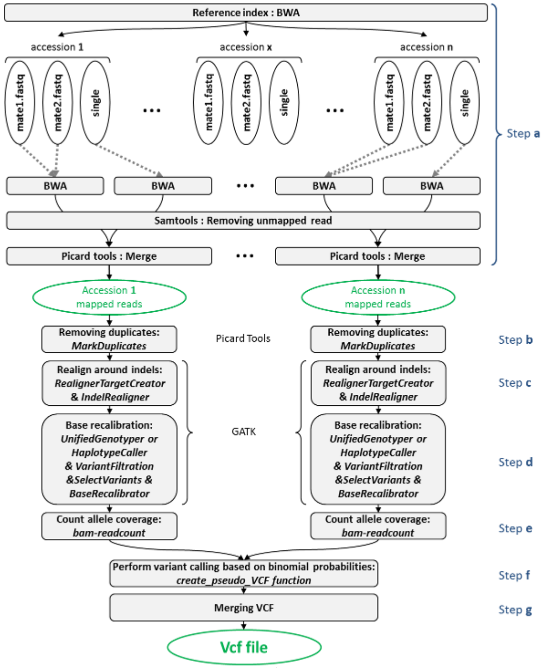
- One last thing before using vcfHunter module: This program has several programs, we will use process_reseq_1.0.py program which have several options, to have access to a description of these options, you can try the following command line:

```
process_reseq_1.0.py -h
```



```
Usage: python process_reseq_1.0.py [options]

Program designed by Guillaume MARTIN : guillaume.martin@cirad.fr

This program go through all steps needed to call SNP from filtered fastq file (From mapping to SNP calling).

Options:
  -h, --help            show this help message and exit
  -c CONF, --conf=CONF  A configuration file containing path to references
                        fastq files.The conf file should contain 2 sections
                        ([Libraries] and [Reference]) and 2 additional ones
                        ([Mapping], [Variant]).
                        [Libraries] section should look like as follows :
                        [Libraries]
                        lib1 = genome_name path_to_mate1 path_to_mate2 ploidy
                        lib2 = genome_name path_to_single ploidy
                        ...
                        [Reference] section should look like as follows :
                        [Reference]
                        genome = path_to_the_reference_sequence
                        [Variant] section may contain 4 options and should
                        look like :
                        [Variant]
                        indel = path to vcf_of_known_indels
                        snp = path to vcf_of_known_SNPs
                        HCopt = additional options to pass to HaplotypeCaller
                        UseUnifiedGenotyperForBaseRecal = yes or no (if not
                        filled default = no)
  -t THREAD, --thread=THREAD
                        Max number of accessions treated at the same time
                        (integer), [default: 1]
  -q QUEUE, --queue=QUEUE
                        Queue to use if SGE is installed on your machine. Do
                        not fill otherwise, [default: none]
  -p PREFIX, --prefix=PREFIX
                        Prefix for output bam and vcf containing all
                        libraries. [default: All_lib]
  -g OUTGZIP, --outgzip=OUTGZIP
                        Output files in gzip format. [Default: n]
  -k KEEPUN, --keepUn=KEEPUN
                        Keep unmapped reads in a file. [Default: n]
  -C CHROM, --chrom=CHROM
                        Chromosomes to work with (only for step f). If "all",
                        all chromosomes will be used for calling. Either: a
                        list of chromosome names separated by ":" [Default:
                        all]
  -s STEPS, --steps=STEPS
                        A string containing steps to perform:
                        a: Aligning libraries
                        b: Removing duplicates
                        c: Indel realignment
                        d: Bases recalibration
                        e: Allele counting
                        f: Genotype calling
```

Several options



Reference index : BWA
accession 1
accession x
accession n
mate1.fastq
mate2.fastq
single
BWA
Samtools : Removing unmapped read
Picard tools : Merge
Accession 1 mapped reads
Accession n mapped reads
Removing duplicates: MarkDuplicates
Picard Tools
Realign around indels: RealignerTargetCreator & IndelRealigner
GATK
Base recalibration: UnifiedGenotyper or HaplotypeCaller & VariantFiltration &SelectVariants & BaseRecalibrator
Count allele coverage: bam-readcount
Perform variant calling based on binomial probabilities: create_pseudo_VCF function
Merging VCF
Vcf file
Step a
Step b
Step c
Step d
Step e
Step f
Step g

Distinct steps: which will be performed sequentially for better explanation

From the output of GBS sequencing to the variant calling file: in command line

- Running read mapping process

```
qsub -q normal.q -l mem_free=12G -b yes -V -N GBSa "process_reseq_1.0.py -c
../WorkShopDataset/GBSCalling.conf -p GBSset -s a -t 1"
```

- The first part of the command line (in bold) is **used by the master computer** (as previously described):

  - **qsub**: Means that we will send a command that the master computer needs to analyze to choose the best computer

  - **-q normal.q**: tells the master computer that we will use computer from normal queue.

  - **-l mem_free=12G**: precise that the program will use 12G of RAM (so the master computer will check that it is available on the computers). This is necessary because this step will use **java** program and this will prevent errors…

  - **-b yes**: it is not important, but put it.

  - **-V**: Tell the master computer to load the module previously loaded on the computer it will choose

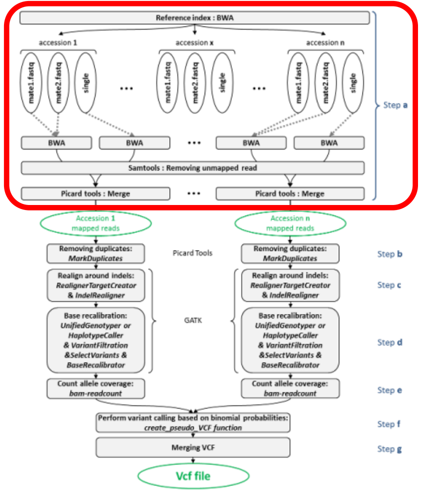
  - **-N GBSa**: A name passed to the command line to look at its status (waiting, running or error) on the cluster

From the output of GBS sequencing to the variant calling file: in command line

- Running read mapping process

```
qsub -q normal.q -l mem_free=12G -b yes -V -N GBSa "process_reseq_1.0.py -c
../WorkShopDataset/GBSCalling.conf -p GBSset -s a -t 1"
```

- The part of the command line between quotation marks (in bold) is the command line that is executed on the **computer chosen by the master computer**:
  - **process_reseq_1.0.py**: We will use process_reseq_1.0.py program
  - **-c ../WorkShopDataset/GBSCalling.conf**: Locates the configuration file
  - **-p GBSset**: A prefix for final output file
  - **-s a**: Tell the program that we will perform step “a” of the workflow

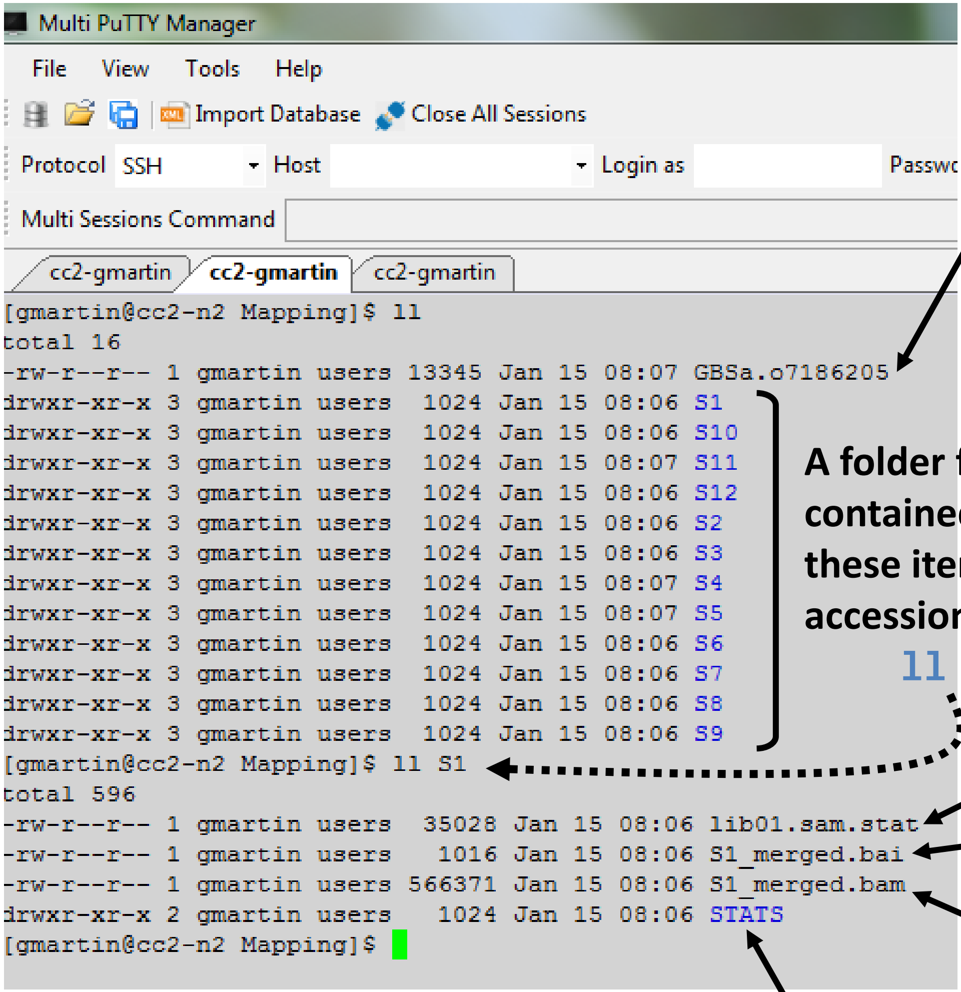

Reference index : BWA
accession 1
accession x
accession n
mate1.fastq
mate2.fastq
single
...
BWA
Samtools : Removing unmapped read
Picard tools : Merge
Step a
Accession 1 mapped reads
Accession n mapped reads
Removing duplicates: MarkDuplicates
Picard Tools
Step b
Realign around indels: RealignerTargetCreator & IndelRealigner
Step c
GATK
Base recalibration: UnifiedGenotyper or HaplotypeCaller & VariantFiltration &SelectVariants & BaseRecalibrator
Step d
Count allele coverage: bam-readcount
Step e
Perform variant calling based on binomial probabilities: create_pseudo_VCF function
Step f
Merging VCF
Step g
Vcf file

  - **-t 1**: Tell the program that only one processor is available. This means that each accessions will be treated sequencially

From the output of GBS sequencing to the variant calling file: in command line

- Listing the files generated:

  ll



```
[gmartin@cc2-n2 Mapping]$ ll
total 16
-rw-r--r-- 1 gmartin users 13345 Jan 15 08:07 GBSa.o7186205
drwxr-xr-x 3 gmartin users  1024 Jan 15 08:06 S1
drwxr-xr-x 3 gmartin users  1024 Jan 15 08:06 S10
drwxr-xr-x 3 gmartin users  1024 Jan 15 08:07 S11
drwxr-xr-x 3 gmartin users  1024 Jan 15 08:06 S12
drwxr-xr-x 3 gmartin users  1024 Jan 15 08:06 S2
drwxr-xr-x 3 gmartin users  1024 Jan 15 08:06 S3
drwxr-xr-x 3 gmartin users  1024 Jan 15 08:07 S4
drwxr-xr-x 3 gmartin users  1024 Jan 15 08:07 S5
drwxr-xr-x 3 gmartin users  1024 Jan 15 08:06 S6
drwxr-xr-x 3 gmartin users  1024 Jan 15 08:06 S7
drwxr-xr-x 3 gmartin users  1024 Jan 15 08:06 S8
drwxr-xr-x 3 gmartin users  1024 Jan 15 08:06 S9
[gmartin@cc2-n2 Mapping]$ ll S1
total 596
-rw-r--r-- 1 gmartin users  35028 Jan 15 08:06 lib01.sam.stat
-rw-r--r-- 1 gmartin users   1016 Jan 15 08:06 S1_merged.bai
-rw-r--r-- 1 gmartin users 566371 Jan 15 08:06 S1_merged.bam
drwxr-xr-x 2 gmartin users   1024 Jan 15 08:06 STATS
[gmartin@cc2-n2 Mapping]$
```

The GBSa.oxxxxxxx file containing what process_reseq_1.0.py told us while it was executing

A folder for each accession. Which contained several items. To have a look at these items, for example for S1 accession:

ll S1

Read mapping statistics

A .bai file: is an index of a bam file for computation performance

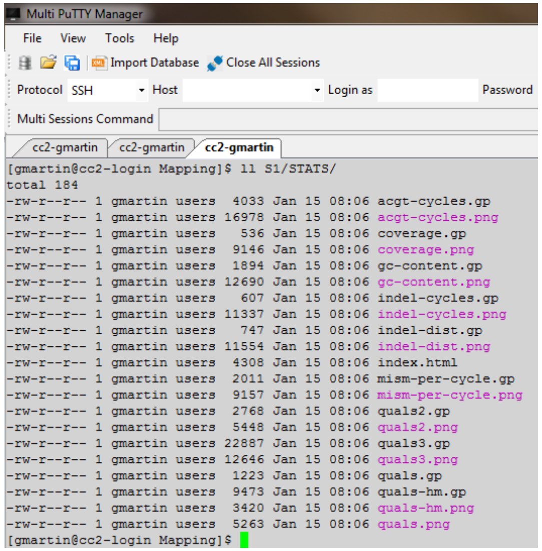
A .bam file: contained sample 1 reads aligned onto the reference

A folder containing read and alignment statistics

From the output of GBS sequencing to the variant calling file: in command line

- Listing one of the stat folder:

```
ll S1/STATS/
```


Multi PuTTY Manager
File View Tools Help
Import Database Close All Sessions
Protocol SSH Host Login as Password
Multi Sessions Command
cc2-gmartin cc2-gmartin cc2-gmartin

```
[gmartin@cc2-login Mapping]$ ll S1/STATS/
total 184
-rw-r--r-- 1 gmartin users  4033 Jan 15 08:06 acgt-cycles.gp
-rw-r--r-- 1 gmartin users 16978 Jan 15 08:06 acgt-cycles.png
-rw-r--r-- 1 gmartin users   536 Jan 15 08:06 coverage.gp
-rw-r--r-- 1 gmartin users  9146 Jan 15 08:06 coverage.png
-rw-r--r-- 1 gmartin users  1894 Jan 15 08:06 gc-content.gp
-rw-r--r-- 1 gmartin users 12690 Jan 15 08:06 gc-content.png
-rw-r--r-- 1 gmartin users   607 Jan 15 08:06 indel-cycles.gp
-rw-r--r-- 1 gmartin users 11337 Jan 15 08:06 indel-cycles.png
-rw-r--r-- 1 gmartin users   747 Jan 15 08:06 indel-dist.gp
-rw-r--r-- 1 gmartin users 11554 Jan 15 08:06 indel-dist.png
-rw-r--r-- 1 gmartin users  4308 Jan 15 08:06 index.html
-rw-r--r-- 1 gmartin users  2011 Jan 15 08:06 mism-per-cycle.gp
-rw-r--r-- 1 gmartin users  9157 Jan 15 08:06 mism-per-cycle.png
-rw-r--r-- 1 gmartin users  2768 Jan 15 08:06 quals2.gp
-rw-r--r-- 1 gmartin users  5448 Jan 15 08:06 quals2.png
-rw-r--r-- 1 gmartin users 22887 Jan 15 08:06 quals3.gp
-rw-r--r-- 1 gmartin users 12646 Jan 15 08:06 quals3.png
-rw-r--r-- 1 gmartin users  1223 Jan 15 08:06 quals.gp
-rw-r--r-- 1 gmartin users  9473 Jan 15 08:06 quals-hm.gp
-rw-r--r-- 1 gmartin users  3420 Jan 15 08:06 quals-hm.png
-rw-r--r-- 1 gmartin users  5263 Jan 15 08:06 quals.png
[gmartin@cc2-login Mapping]$
```

- Several files are generated but one summarize all of them: the one named: index.html
- This is an html file readable by firefox. To have a look at this file:

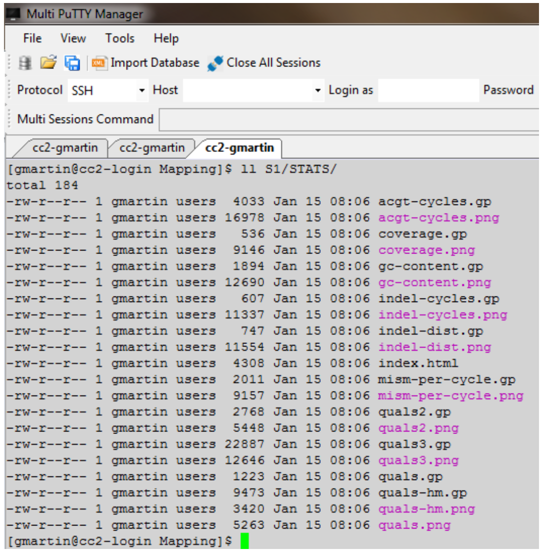
```
firefox S1/STATS/index.html
```

- This command open a firefox window:

From the output of GBS sequencing to the variant calling file: in command line

- Listing one of the stat folder:

```
ll S1/STATS/
```

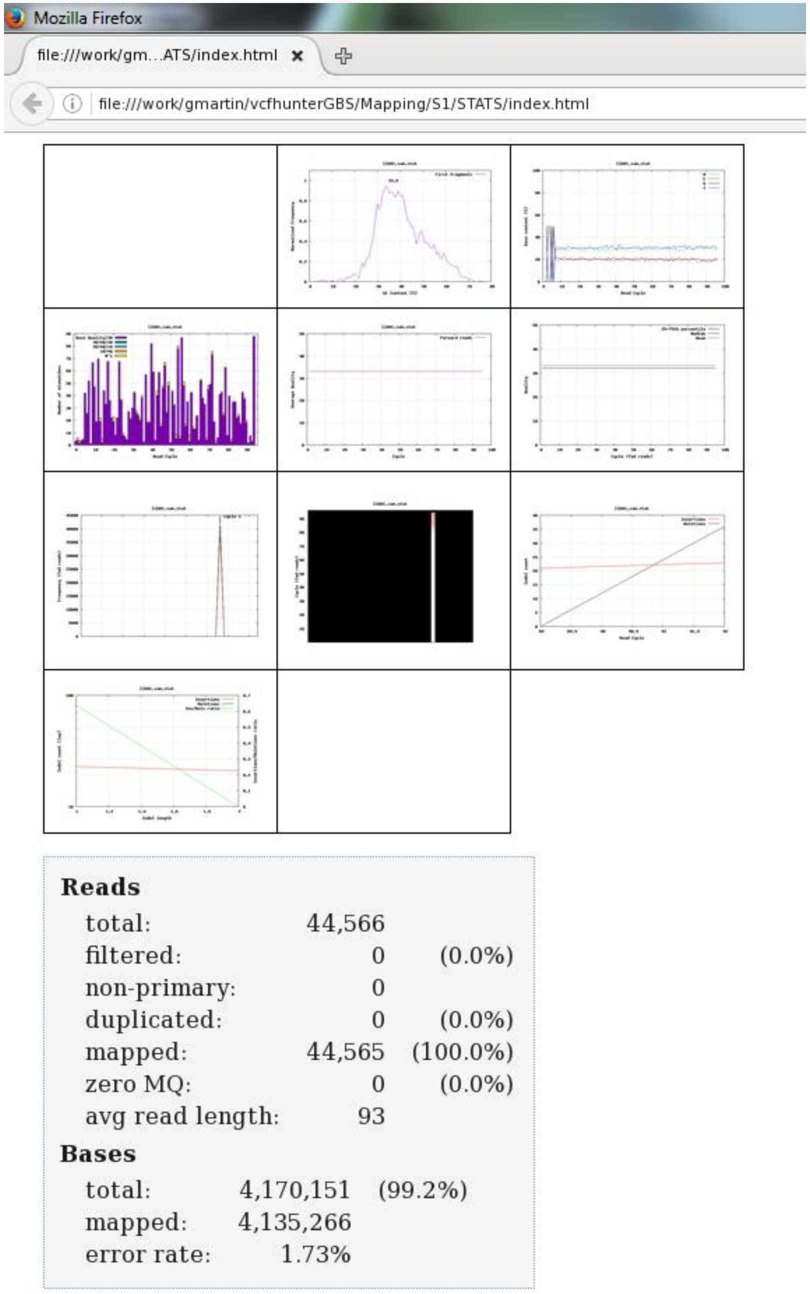


```
[gmartin@cc2-login Mapping]$ ll S1/STATS/
total 184
-rw-r--r-- 1 gmartin users  4033 Jan 15 08:06 acgt-cycles.gp
-rw-r--r-- 1 gmartin users 16978 Jan 15 08:06 acgt-cycles.png
-rw-r--r-- 1 gmartin users   536 Jan 15 08:06 coverage.gp
-rw-r--r-- 1 gmartin users  9146 Jan 15 08:06 coverage.png
-rw-r--r-- 1 gmartin users  1894 Jan 15 08:06 gc-content.gp
-rw-r--r-- 1 gmartin users 12690 Jan 15 08:06 gc-content.png
-rw-r--r-- 1 gmartin users   607 Jan 15 08:06 indel-cycles.gp
-rw-r--r-- 1 gmartin users 11337 Jan 15 08:06 indel-cycles.png
-rw-r--r-- 1 gmartin users   747 Jan 15 08:06 indel-dist.gp
-rw-r--r-- 1 gmartin users 11554 Jan 15 08:06 indel-dist.png
-rw-r--r-- 1 gmartin users  4308 Jan 15 08:06 index.html
-rw-r--r-- 1 gmartin users  2011 Jan 15 08:06 mism-per-cycle.gp
-rw-r--r-- 1 gmartin users  9157 Jan 15 08:06 mism-per-cycle.png
-rw-r--r-- 1 gmartin users  2768 Jan 15 08:06 quals2.gp
-rw-r--r-- 1 gmartin users  5448 Jan 15 08:06 quals2.png
-rw-r--r-- 1 gmartin users 22887 Jan 15 08:06 quals3.gp
-rw-r--r-- 1 gmartin users 12646 Jan 15 08:06 quals3.png
-rw-r--r-- 1 gmartin users  1223 Jan 15 08:06 quals.gp
-rw-r--r-- 1 gmartin users  9473 Jan 15 08:06 quals-hm.gp
-rw-r--r-- 1 gmartin users  3420 Jan 15 08:06 quals-hm.png
-rw-r--r-- 1 gmartin users  5263 Jan 15 08:06 quals.png
[gmartin@cc2-login Mapping]$
```

- Several files are generated but one summarize all of them: the one named: index.html
- This is an html file readable by firefox. To have a look at this file:

```
firefox S1/STATS/index.html
```

- This command open a firefox window:


Mozilla Firefox
file:///work/gm...ATS/index.html
file:///work/gmartin/vcfhunterGBS/Mapping/S1/STATS/index.html

```
Reads
  total:              44,566
  filtered:                0     (0.0%)
  non-primary:             0
  duplicated:              0     (0.0%)
  mapped:             44,565   (100.0%)
  zero MQ:                 0     (0.0%)
  avg read length:        93
Bases
  total:          4,170,151   (99.2%)
  mapped:         4,135,266
  error rate:         1.73%
```

From the output of GBS sequencing to the variant calling file: in command line

- The alignment file (bam format): These file are compressed binary files (easier to use by programs) but not directly readable for human… These file can still be observed with the ***samtools*** program with the command line:

```
samtools view -h S1/S1_merged.bam | more
```

Convert the bam is sam format (readable by human)

Read this converted file line by line

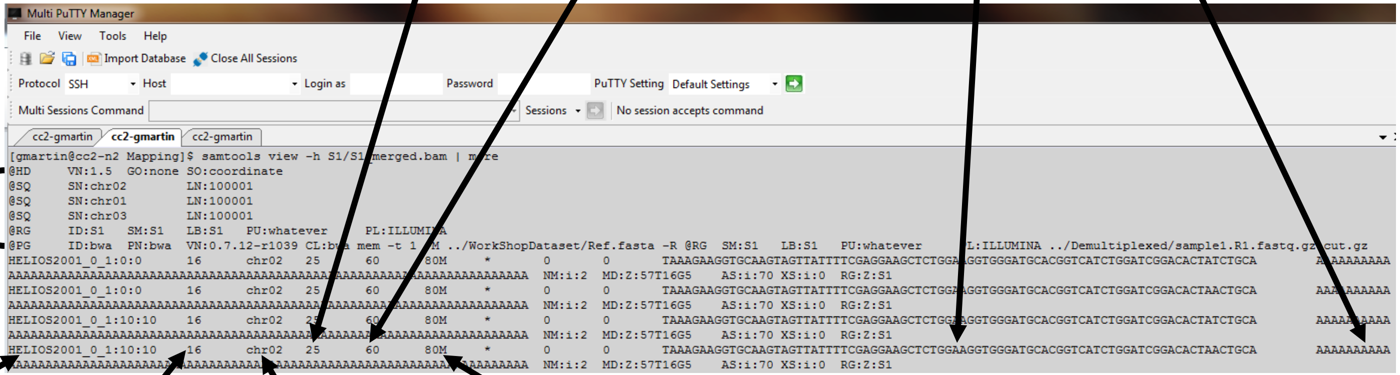
Mapping quality **Read sequence**

Read position

Read quality

Header containing information on:

- **reference sequences**
- **Aligner used**



```
[gmartin@cc2-n2 Mapping]$ samtools view -h S1/S1_merged.bam | more
@HD     VN:1.5  GO:none SO:coordinate
@SQ     SN:chr02        LN:100001
@SQ     SN:chr01        LN:100001
@SQ     SN:chr03        LN:100001
@RG     ID:S1   SM:S1   LB:S1   PU:whatever     PL:ILLUMINA
@PG     ID:bwa  PN:bwa  VN:0.7.12-r1039 CL:bwa mem -t 1 -M ../WorkShopDataset/Ref.fasta -R @RG  SM:S1   LB:S1   PU:whatever     PL:ILLUMINA ../Demultiplexed/sample1.R1.fastq.gz_cut.gz
HELIOS2001_0_1:0:0      16      chr02   25      60      80M     *       0       0       TAAAGAAGGTGCAAGTAGTTATTTTCGAGGAAGCTCTGGAAGGTGGGATGCACGGTCATCTGGATCGGACACTATCTGCA        AAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA  NM:i:2  MD:Z:57T16G5    AS:i:70 XS:i:0  RG:Z:S1
HELIOS2001_0_1:0:0      16      chr02   25      60      80M     *       0       0       TAAAGAAGGTGCAAGTAGTTATTTTCGAGGAAGCTCTGGAAGGTGGGATGCACGGTCATCTGGATCGGACACTAACTGCA        AAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA  NM:i:2  MD:Z:57T16G5    AS:i:70 XS:i:0  RG:Z:S1
HELIOS2001_0_1:10:10    16      chr02   25      60      80M     *       0       0       TAAAGAAGGTGCAAGTAGTTATTTTCGAGGAAGCTCTGGAAGGTGGGATGCACGGTCATCTGGATCGGACACTATCTGCA        AAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA  NM:i:2  MD:Z:57T16G5    AS:i:70 XS:i:0  RG:Z:S1
HELIOS2001_0_1:10:10    16      chr02   25      60      80M     *       0       0       TAAAGAAGGTGCAAGTAGTTATTTTCGAGGAAGCTCTGGAAGGTGGGATGCACGGTCATCTGGATCGGACACTAACTGCA        AAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA  NM:i:2  MD:Z:57T16G5    AS:i:70 XS:i:0  RG:Z:S1
```

Read 1
Read 2
Read 3
Read 4

Read name

Tag regarding read mapping information
https://broadinstitute.github.io/picard/explain-flags.html

CIGAR (80M = 80 Match)

Read mapping chromosome

+ Other informations (see link for more information) https://samtools.github.io/hts-specs/SAMv1.pdf

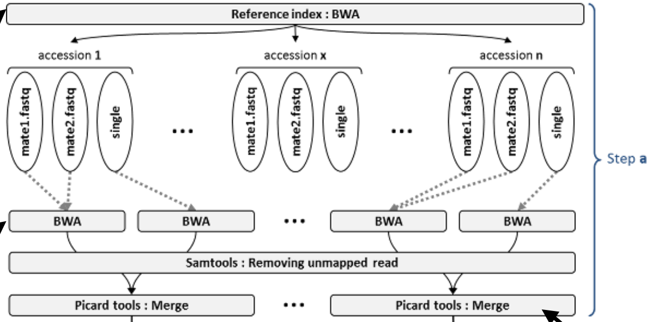
- **To quit:** “Ctrl” + “C”

From the output of GBS sequencing to the variant calling file: in command line

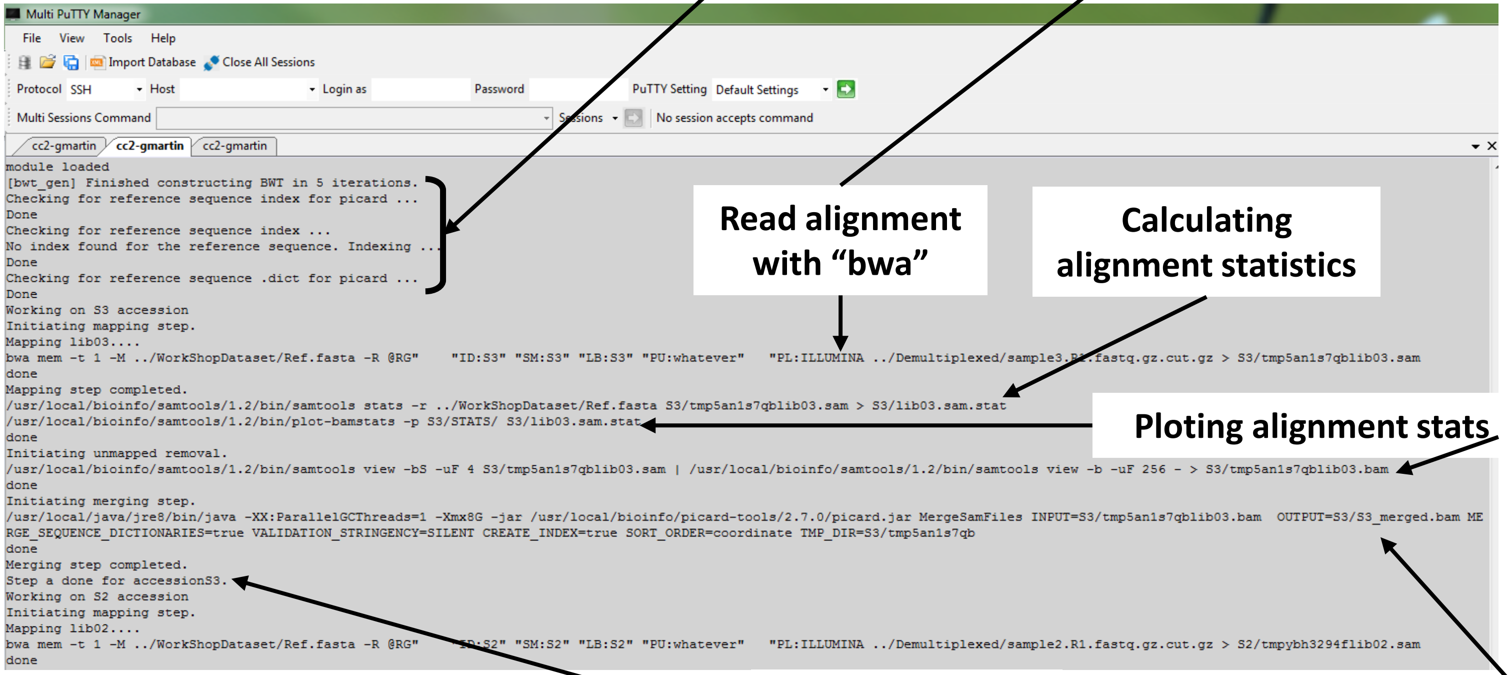
- The GBSa.oxxxxxxx file:

  `more GBSa.oxxxxxxxx`

- List steps performed during step “a”



Reference indexation



```
module loaded
[bwt_gen] Finished constructing BWT in 5 iterations.
Checking for reference sequence index for picard ...
Done
Checking for reference sequence index ...
No index found for the reference sequence. Indexing ...
Done
Checking for reference sequence .dict for picard ...
Done
Working on S3 accession
Initiating mapping step.
Mapping lib03....
bwa mem -t 1 -M ../WorkShopDataset/Ref.fasta -R @RG"    "ID:S3" "SM:S3" "LB:S3" "PU:whatever"   "PL:ILLUMINA ../Demultiplexed/sample3.R1.fastq.gz.cut.gz > S3/tmp5an1s7qblib03.sam
done
Mapping step completed.
/usr/local/bioinfo/samtools/1.2/bin/samtools stats -r ../WorkShopDataset/Ref.fasta S3/tmp5an1s7qblib03.sam > S3/lib03.sam.stat
/usr/local/bioinfo/samtools/1.2/bin/plot-bamstats -p S3/STATS/ S3/lib03.sam.stat
done
Initiating unmapped removal.
/usr/local/bioinfo/samtools/1.2/bin/samtools view -bS -uF 4 S3/tmp5an1s7qblib03.sam | /usr/local/bioinfo/samtools/1.2/bin/samtools view -b -uF 256 - > S3/tmp5an1s7qblib03.bam
done
Initiating merging step.
/usr/local/java/jre8/bin/java -XX:ParallelGCThreads=1 -Xmx8G -jar /usr/local/bioinfo/picard-tools/2.7.0/picard.jar MergeSamFiles INPUT=S3/tmp5an1s7qblib03.bam  OUTPUT=S3/S3_merged.bam ME
RGE_SEQUENCE_DICTIONARIES=true VALIDATION_STRINGENCY=SILENT CREATE_INDEX=true SORT_ORDER=coordinate TMP_DIR=S3/tmp5an1s7qb
done
Merging step completed.
Step a done for accessionS3.
Working on S2 accession
Initiating mapping step.
Mapping lib02....
bwa mem -t 1 -M ../WorkShopDataset/Ref.fasta -R @RG"    "ID:S2" "SM:S2" "LB:S2" "PU:whatever"   "PL:ILLUMINA ../Demultiplexed/sample2.R1.fastq.gz.cut.gz > S2/tmpybh3294flib02.sam
done
```

Read alignment with “bwa”

Calculating alignment statistics

Ploting alignment stats

Removing unmapped read and secondary alignment

Concatenating reads from the same accessions but from several libraries (not necessary here but performed anyway)

Starting a new accession

- **To quit:** `“Ctrl” + “C”` **or** `“enter”` **until the end of file**

From the output of GBS sequencing to the variant calling file: in command line

- Running read indel realignment:

```
qsub -q normal.q -l mem_free=12G -b yes -V -N GBSc "process_reseq_1.0.py -c
../WorkShopDataset/GBSCalling.conf -p GBSset -s c -t 1"
```

- The first part of the command line (in bold) is **used by the master computer** (as previously described):

  - **qsub**: Means that we will send a command that the master computer needs to analyze to choose the best computer

  - **-q normal.q**: tells the master computer that we will use computer from normal queue.

  - **-l mem_free=12G**: precise that the program will use 12G of RAM (so the master computer will check that it is available on the computers). This is necessary because this step will use **java** program and this will prevent errors…

  - **-b yes**: it is not important, but put it.

  - **-V**: Tell the master computer to load the module previously loaded on the computer it will choose

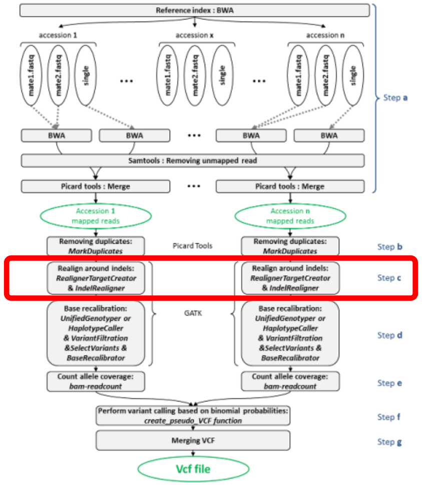
  - **-N GBSc**: A name passed to the command line to look at its status (waiting, running or error) on the cluster

From the output of GBS sequencing to the variant calling file: in command line

- Running read indel realignment:

```
qsub -q normal.q -l mem_free=12G -b yes -V -N GBSc "process_reseq_1.0.py -c
../WorkShopDataset/GBSCalling.conf -p GBSset -s c -t 1"
```

- The part of the command line between quotation marks (in bold) is the command line that is executed on the **computer chosen by the master computer**:
  - ❖ **process_reseq_1.0.py**: We will use process_reseq_1.0.py program
  - ❖ **-c ../WorkShopDataset/GBSCalling.conf**: Locates the configuration file
  - ❖ **-p GBSset**: A prefix for final output file
  - ❖ **-s c**: Tell the program that we will perform step "c" of the workflow

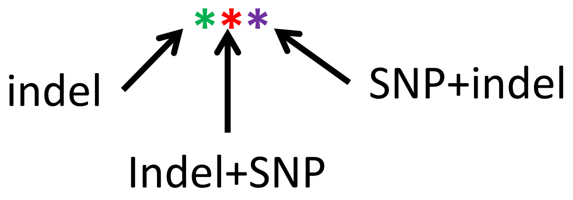

Reference index : BWA
accession 1
accession x
accession n
mate1.fastq
mate2.fastq
single
...
BWA
Samtools : Removing unmapped read
Picard tools : Merge
Accession 1 mapped reads
Accession n mapped reads
Step a
Removing duplicates: MarkDuplicates
Picard Tools
Step b
Realign around indels: RealignerTargetCreator & IndelRealigner
Step c
Base recalibration: UnifiedGenotyper or HaplotypeCaller & VariantFiltration &SelectVariants & BaseRecalibrator
GATK
Step d
Count allele coverage: bam-readcount
Step e
Perform variant calling based on binomial probabilities: create_pseudo_VCF function
Step f
Merging VCF
Step g
Vcf file

  - ❖ **-t 1**: Tell the program that only one processor is available. This means that each accessions will be treated sequencially

From the output of GBS sequencing to the variant calling file: in command line

- Why performing indel realalignment?
  - ✓ Because the alignment around indel can be problematic…

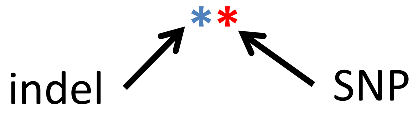
```
Reference  GCAACAAGGGTTACAGATCGGAAAAGAGCGGTTCAGCAGGAATGCCG
               CAAGGGTTACAGATCGGAAA-TAGCGGTTCAGCA
                  GGGTTACAGATCGGAAAT-AGCGGTTCAGCAGGAATGCCG
                 AGGGTTACAGATCGGAA-ATAGCGGTTCAGCAGGAATGCCG
```



  - ✓ ➔ several polimorphism with the same sequence!

➔ Realignment around indel:

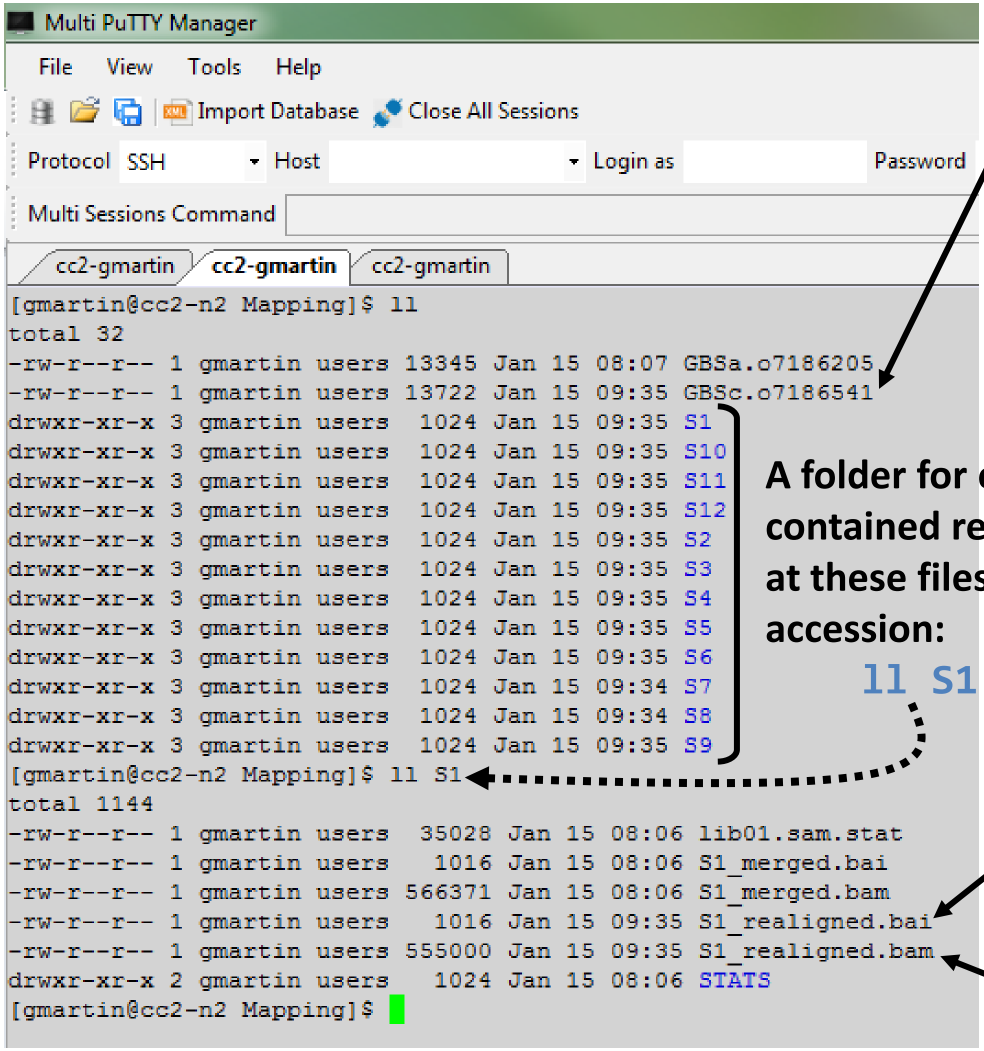
```
Reference  GCAACAAGGGTTACAGATCGGAAAAGAGCGGTTCAGCAGGAATGCCG
               CAAGGGTTACAGATCGGAAA-TAGCGGTTCAGCA
                  GGGTTACAGATCGGAAA-TAGCGGTTCAGCAGGAATGCCG
                 AGGGTTACAGATCGGAAA-TAGCGGTTCAGCAGGAATGCCG
```



From the output of GBS sequencing to the variant calling file: in command line

- Listing the files generated:

```
ll
```



```
[gmartin@cc2-n2 Mapping]$ ll
total 32
-rw-r--r-- 1 gmartin users 13345 Jan 15 08:07 GBSa.o7186205
-rw-r--r-- 1 gmartin users 13722 Jan 15 09:35 GBSc.o7186541
drwxr-xr-x 3 gmartin users  1024 Jan 15 09:35 S1
drwxr-xr-x 3 gmartin users  1024 Jan 15 09:35 S10
drwxr-xr-x 3 gmartin users  1024 Jan 15 09:35 S11
drwxr-xr-x 3 gmartin users  1024 Jan 15 09:35 S12
drwxr-xr-x 3 gmartin users  1024 Jan 15 09:35 S2
drwxr-xr-x 3 gmartin users  1024 Jan 15 09:35 S3
drwxr-xr-x 3 gmartin users  1024 Jan 15 09:35 S4
drwxr-xr-x 3 gmartin users  1024 Jan 15 09:35 S5
drwxr-xr-x 3 gmartin users  1024 Jan 15 09:35 S6
drwxr-xr-x 3 gmartin users  1024 Jan 15 09:34 S7
drwxr-xr-x 3 gmartin users  1024 Jan 15 09:34 S8
drwxr-xr-x 3 gmartin users  1024 Jan 15 09:35 S9
[gmartin@cc2-n2 Mapping]$ ll S1
total 1144
-rw-r--r-- 1 gmartin users  35028 Jan 15 08:06 lib01.sam.stat
-rw-r--r-- 1 gmartin users   1016 Jan 15 08:06 S1_merged.bai
-rw-r--r-- 1 gmartin users 566371 Jan 15 08:06 S1_merged.bam
-rw-r--r-- 1 gmartin users   1016 Jan 15 09:35 S1_realigned.bai
-rw-r--r-- 1 gmartin users 555000 Jan 15 09:35 S1_realigned.bam
drwxr-xr-x 2 gmartin users   1024 Jan 15 08:06 STATS
[gmartin@cc2-n2 Mapping]$
```

The GBSc.oxxxxxxx file containing what process_reseq_1.0.py told us while it was executing

A folder for each accession which contained realigned reads. To have a look at these files, for example for S1 accession:

```
ll S1
```

A .bai file: is an index of the realigned bam file for computation performance

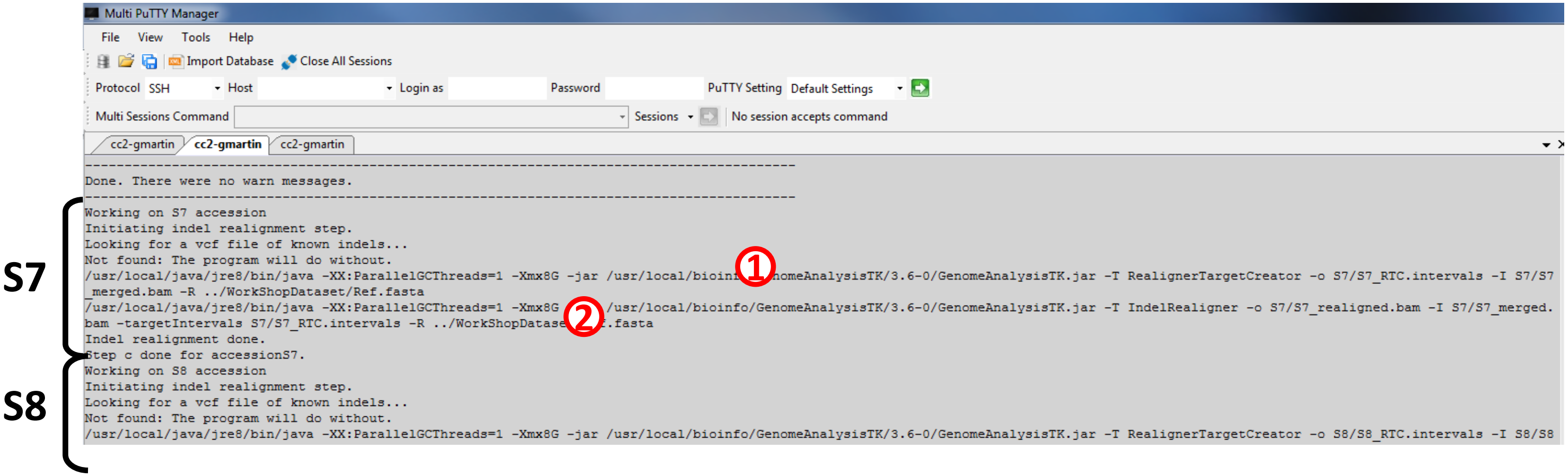
A realigned.bam file: contained sample 1 reads realigned around indels

From the output of GBS sequencing to the variant calling file: in command line

- The GBSa.oxxxxxxx file:

  `more GBSc.oxxxxxxx`

- List steps performed during step “c”



```
------------------------------------------------------------------------------------------
Done. There were no warn messages.
------------------------------------------------------------------------------------------
Working on S7 accession
Initiating indel realignment step.
Looking for a vcf file of known indels...
Not found: The program will do without.
/usr/local/java/jre8/bin/java -XX:ParallelGCThreads=1 -Xmx8G -jar /usr/local/bioinfo/GenomeAnalysisTK/3.6-0/GenomeAnalysisTK.jar -T RealignerTargetCreator -o S7/S7_RTC.intervals -I S7/S7
_merged.bam -R ../WorkShopDataset/Ref.fasta
/usr/local/java/jre8/bin/java -XX:ParallelGCThreads=1 -Xmx8G -jar /usr/local/bioinfo/GenomeAnalysisTK/3.6-0/GenomeAnalysisTK.jar -T IndelRealigner -o S7/S7_realigned.bam -I S7/S7_merged.
bam -targetIntervals S7/S7_RTC.intervals -R ../WorkShopDataset/Ref.fasta
Indel realignment done.
Step c done for accessionS7.
Working on S8 accession
Initiating indel realignment step.
Looking for a vcf file of known indels...
Not found: The program will do without.
/usr/local/java/jre8/bin/java -XX:ParallelGCThreads=1 -Xmx8G -jar /usr/local/bioinfo/GenomeAnalysisTK/3.6-0/GenomeAnalysisTK.jar -T RealignerTargetCreator -o S8/S8_RTC.intervals -I S8/S8
```

- Indel realignment was performed using GATK (https://software.broadinstitute.org/gatk/) in two steps (https://software.broadinstitute.org/gatk/documentation/tooldocs/3.8-0/org_broadinstitute_gatk_tools_walkers_indels_RealignerTargetCreator.php):

① “Determining (small) suspicious intervals which are likely in need of realignment”

② “Running the realigner over those intervals”

SouthGreen bioinformatics platform GenomeHarvest diversity, organization and dynamics

From the output of GBS sequencing to the variant calling file: in command line

- Running allele count:

```
qsub -q normal.q -b yes -V -N GBSe "process_reseq_1.0.py -c
../WorkShopDataset/GBSCalling.conf -p GBSset -s e -t 1"
```

- The first part of the command line (in bold) is **used by the master computer** (as previously described):

  - **qsub**: Means that we will send a command that the master computer needs to analyze to choose the best computer

  - **-q normal.q**: tells the master computer that we will use computer from normal queue.

  - **-b yes**: it is not important, but put it.

  - **-V**: Tell the master computer to load the module previously loaded on the computer it will choose

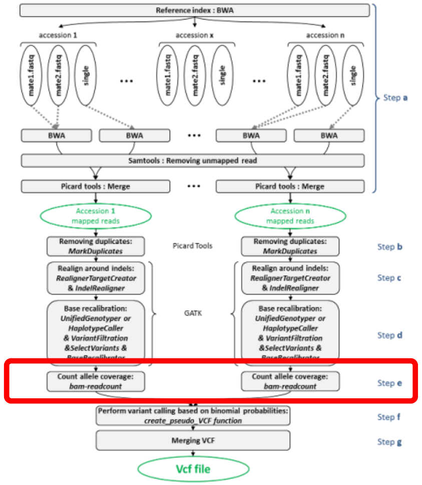
  - **-N GBSe**: A name passed to the command line to look at its status (waiting, running or error) on the cluster

From the output of GBS sequencing to the variant calling file: in command line

- Running allele count:

```
qsub -q normal.q -b yes -V -N GBSe "process_reseq_1.0.py -c
../WorkShopDataset/GBSCalling.conf -p GBSset -s e -t 1"
```

- The part of the command line between quotation marks (in bold) is the command line that is executed on the **computer chosen by the master computer**:
  - **process_reseq_1.0.py**: We will use process_reseq_1.0.py program
  - **-c ../WorkShopDataset/GBSCalling.conf**: Locates the configuration file
  - **-p GBSset**: A prefix for final output file
  - **-s e**: Tell the program that we will perform step “e” of the workflow



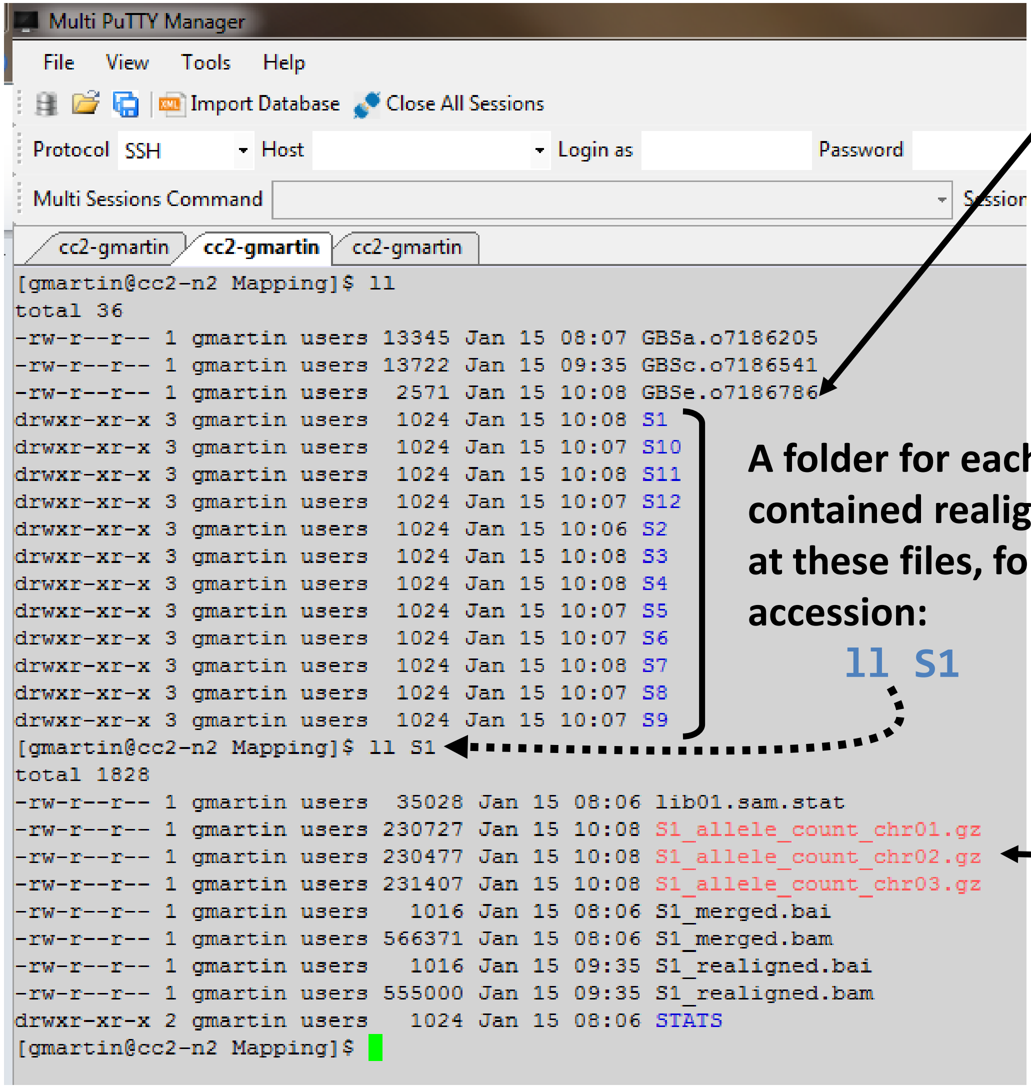
Reference index : BWA
accession 1
accession x
accession n
mate1.fastq
mate2.fastq
single
...
BWA
Samtools : Removing unmapped read
Picard tools : Merge
Step a
Accession 1 mapped reads
Accession n mapped reads
Removing duplicates: MarkDuplicates
Picard Tools
Step b
Realign around indels: RealignerTargetCreator & IndelRealigner
Step c
GATK
Base recalibration: UnifiedGenotyper or HaplotypeCaller & VariantFiltration &SelectVariants & BaseRecalibrator
Step d
Count allele coverage: bam-readcount
Step e
Perform variant calling based on binomial probabilities: create_pseudo_VCF function
Step f
Merging VCF
Step g
Vcf file

  - **-t 1**: Tell the program that only one processor is available. This means that each accessions will be treated sequencially

From the output of GBS sequencing to the variant calling file: in command line

- Listing the files generated:

```
ll
```



```
[gmartin@cc2-n2 Mapping]$ ll
total 36
-rw-r--r-- 1 gmartin users 13345 Jan 15 08:07 GBSa.o7186205
-rw-r--r-- 1 gmartin users 13722 Jan 15 09:35 GBSc.o7186541
-rw-r--r-- 1 gmartin users  2571 Jan 15 10:08 GBSe.o7186786
drwxr-xr-x 3 gmartin users  1024 Jan 15 10:08 S1
drwxr-xr-x 3 gmartin users  1024 Jan 15 10:07 S10
drwxr-xr-x 3 gmartin users  1024 Jan 15 10:08 S11
drwxr-xr-x 3 gmartin users  1024 Jan 15 10:07 S12
drwxr-xr-x 3 gmartin users  1024 Jan 15 10:06 S2
drwxr-xr-x 3 gmartin users  1024 Jan 15 10:08 S3
drwxr-xr-x 3 gmartin users  1024 Jan 15 10:08 S4
drwxr-xr-x 3 gmartin users  1024 Jan 15 10:07 S5
drwxr-xr-x 3 gmartin users  1024 Jan 15 10:07 S6
drwxr-xr-x 3 gmartin users  1024 Jan 15 10:08 S7
drwxr-xr-x 3 gmartin users  1024 Jan 15 10:07 S8
drwxr-xr-x 3 gmartin users  1024 Jan 15 10:07 S9
[gmartin@cc2-n2 Mapping]$ ll S1
total 1828
-rw-r--r-- 1 gmartin users  35028 Jan 15 08:06 lib01.sam.stat
-rw-r--r-- 1 gmartin users 230727 Jan 15 10:08 S1_allele_count_chr01.gz
-rw-r--r-- 1 gmartin users 230477 Jan 15 10:08 S1_allele_count_chr02.gz
-rw-r--r-- 1 gmartin users 231407 Jan 15 10:08 S1_allele_count_chr03.gz
-rw-r--r-- 1 gmartin users   1016 Jan 15 08:06 S1_merged.bai
-rw-r--r-- 1 gmartin users 566371 Jan 15 08:06 S1_merged.bam
-rw-r--r-- 1 gmartin users   1016 Jan 15 09:35 S1_realigned.bai
-rw-r--r-- 1 gmartin users 555000 Jan 15 09:35 S1_realigned.bam
drwxr-xr-x 2 gmartin users   1024 Jan 15 08:06 STATS
[gmartin@cc2-n2 Mapping]$
```

The GBSe.oxxxxxxx file containing what process_reseq_1.0.py told us while it was executing

A folder for each accession which contained realigned reads. To have a look at these files, for example for S1 accession:

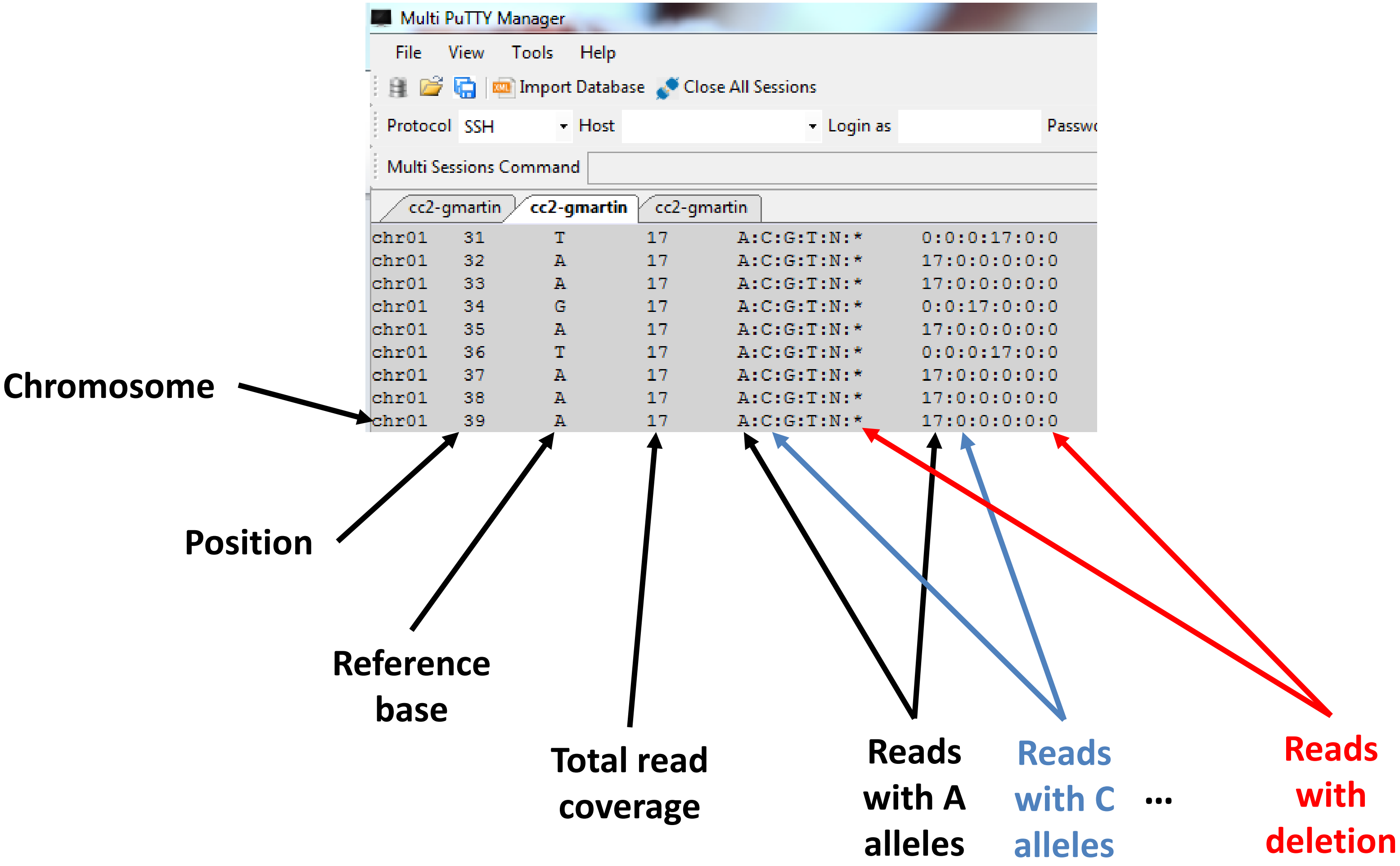
```
ll S1
```

Three files (one for each chromosomes) which count for each covered position by reads, the number of read supporting each possible alleles

From the output of GBS sequencing to the variant calling file: in command line

- Example of ***S1_allele_count_chr01.gz*** file:

  `zmore S1/S1_allele_count_chr01.gz`



```
chr01    31       T        17       A:C:G:T:N:*      0:0:0:17:0:0
chr01    32       A        17       A:C:G:T:N:*      17:0:0:0:0:0
chr01    33       A        17       A:C:G:T:N:*      17:0:0:0:0:0
chr01    34       G        17       A:C:G:T:N:*      0:0:17:0:0:0
chr01    35       A        17       A:C:G:T:N:*      17:0:0:0:0:0
chr01    36       T        17       A:C:G:T:N:*      0:0:0:17:0:0
chr01    37       A        17       A:C:G:T:N:*      17:0:0:0:0:0
chr01    38       A        17       A:C:G:T:N:*      17:0:0:0:0:0
chr01    39       A        17       A:C:G:T:N:*      17:0:0:0:0:0
```

Chromosome

Position

Reference base

Total read coverage

Reads with A alleles

Reads with C alleles

…

Reads with deletion

- **To quit: “`Ctrl`” + “`C`” or “`enter`” until the end of file**

From the output of GBS sequencing to the variant calling file: in command line

- Creating the variant calling file (VCF):

```
qsub -q normal.q -pe parallel_smp 3 -b yes -V -N GBSf "process_reseq_1.0.py
-c ../WorkShopDataset/GBSCalling.conf -p GBSset -s f -t 3"
```

- The first part of the command line (in bold) is **used by the master computer** (as previously described):

  - ❖ **qsub**: Means that we will send a command that the master computer needs to analyze to choose the best computer

  - ❖ **-q normal.q**: tells the master computer that we will use computer from normal queue.

  - ❖ **-pe parallel_smp 3**: tells the master computer that we need 3 processor (this can be used to gain speed in computation time if the program allowed it)

  - ❖ **-b yes**: it is not important, but put it.

  - ❖ **-V**: Tell the master computer to load the module previously loaded on the computer it will choose

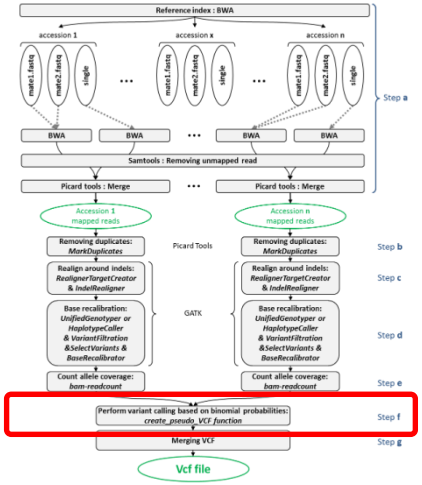
  - ❖ **-N GBSf**: A name passed to the command line to look at its status (waiting, running or error) on the cluster

From the output of GBS sequencing to the variant calling file: in command line

- Creating the variant calling file (VCF):

```
qsub -q normal.q -pe parallel_smp 3 -b yes -V -N GBSf "process_reseq_1.0.py
-c ../WorkShopDataset/GBSCalling.conf -p GBSset -s f -t 3"
```

- The part of the command line between quotation marks (in bold) is the command line that is executed on the **computer chosen by the master computer**:
  - ❖ **process_reseq_1.0.py**: We will use process_reseq_1.0.py program
  - ❖ **-c ../WorkShopDataset/GBSCalling.conf**: Locates the configuration file
  - ❖ **-p GBSset**: A prefix for final output file
  - ❖ **-s f**: Tell the program that we will perform step "f" of the workflow

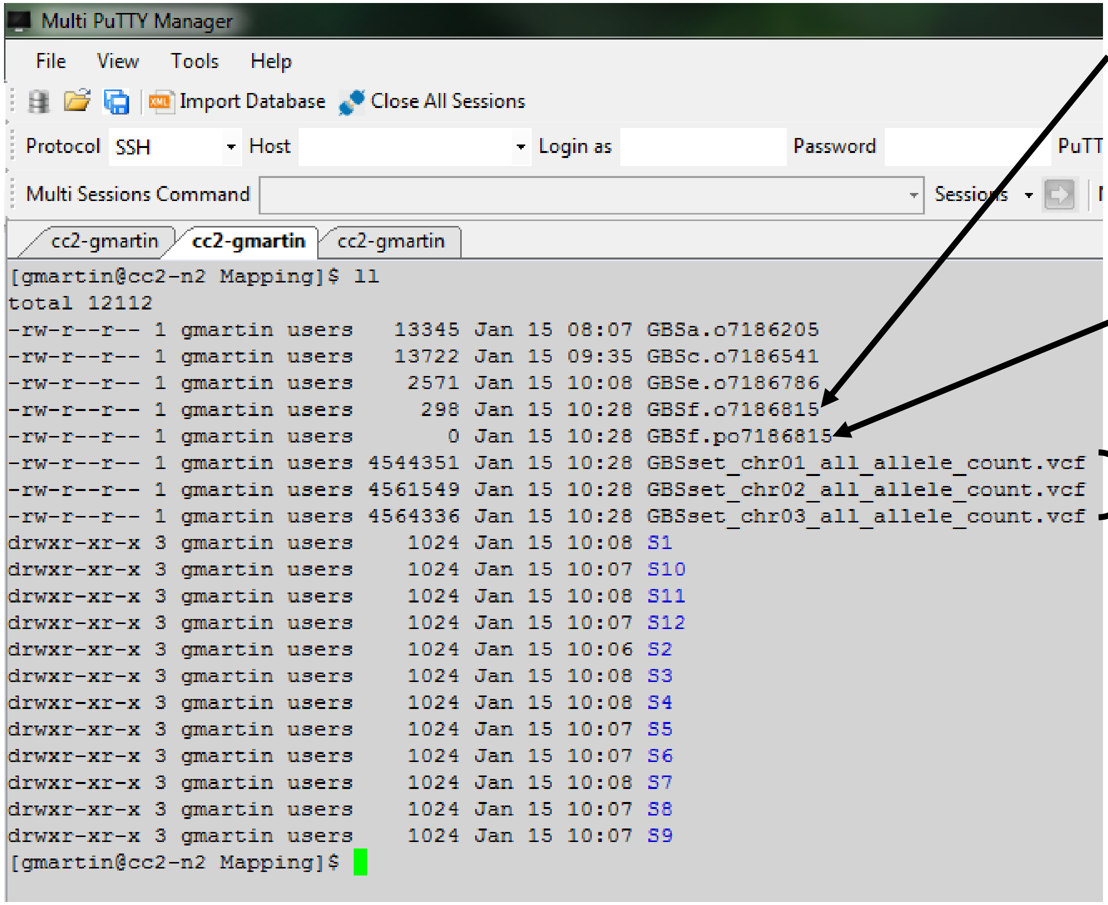

Reference index : BWA
accession 1
accession x
accession n
mate1.fastq
mate2.fastq
single
...
BWA
Samtools : Removing unmapped read
Picard tools : Merge
Accession 1 mapped reads
Accession n mapped reads
Removing duplicates: MarkDuplicates
Picard Tools
Realign around indels: RealignerTargetCreator & IndelRealigner
GATK
Base recalibration: UnifiedGenotyper or HaplotypeCaller & VariantFiltration &SelectVariants & BaseRecalibrator
Count allele coverage: bam-readcount
Perform variant calling based on binomial probabilities: create_pseudo_VCF function
Merging VCF
Vcf file
Step a
Step b
Step c
Step d
Step e
Step f
Step g

  - ❖ **-t 3**: Tell the program that only three processors are available (allowed by -pe parallel_smp 3). With this option, all three chromosomes will be treated independently at the same time by one processor each. This allowed to gain computation time

From the output of GBS sequencing to the variant calling file: in command line

- Listing the files generated:

```
ll
```



```
[gmartin@cc2-n2 Mapping]$ ll
total 12112
-rw-r--r-- 1 gmartin users   13345 Jan 15 08:07 GBSa.o7186205
-rw-r--r-- 1 gmartin users   13722 Jan 15 09:35 GBSc.o7186541
-rw-r--r-- 1 gmartin users    2571 Jan 15 10:08 GBSe.o7186786
-rw-r--r-- 1 gmartin users     298 Jan 15 10:28 GBSf.o7186815
-rw-r--r-- 1 gmartin users       0 Jan 15 10:28 GBSf.po7186815
-rw-r--r-- 1 gmartin users 4544351 Jan 15 10:28 GBSset_chr01_all_allele_count.vcf
-rw-r--r-- 1 gmartin users 4561549 Jan 15 10:28 GBSset_chr02_all_allele_count.vcf
-rw-r--r-- 1 gmartin users 4564336 Jan 15 10:28 GBSset_chr03_all_allele_count.vcf
drwxr-xr-x 3 gmartin users    1024 Jan 15 10:08 S1
drwxr-xr-x 3 gmartin users    1024 Jan 15 10:07 S10
drwxr-xr-x 3 gmartin users    1024 Jan 15 10:08 S11
drwxr-xr-x 3 gmartin users    1024 Jan 15 10:07 S12
drwxr-xr-x 3 gmartin users    1024 Jan 15 10:06 S2
drwxr-xr-x 3 gmartin users    1024 Jan 15 10:08 S3
drwxr-xr-x 3 gmartin users    1024 Jan 15 10:08 S4
drwxr-xr-x 3 gmartin users    1024 Jan 15 10:07 S5
drwxr-xr-x 3 gmartin users    1024 Jan 15 10:07 S6
drwxr-xr-x 3 gmartin users    1024 Jan 15 10:08 S7
drwxr-xr-x 3 gmartin users    1024 Jan 15 10:07 S8
drwxr-xr-x 3 gmartin users    1024 Jan 15 10:07 S9
[gmartin@cc2-n2 Mapping]$
```

The GBSf.oxxxxxxx file containing what process_reseq_1.0.py told us while it was executing

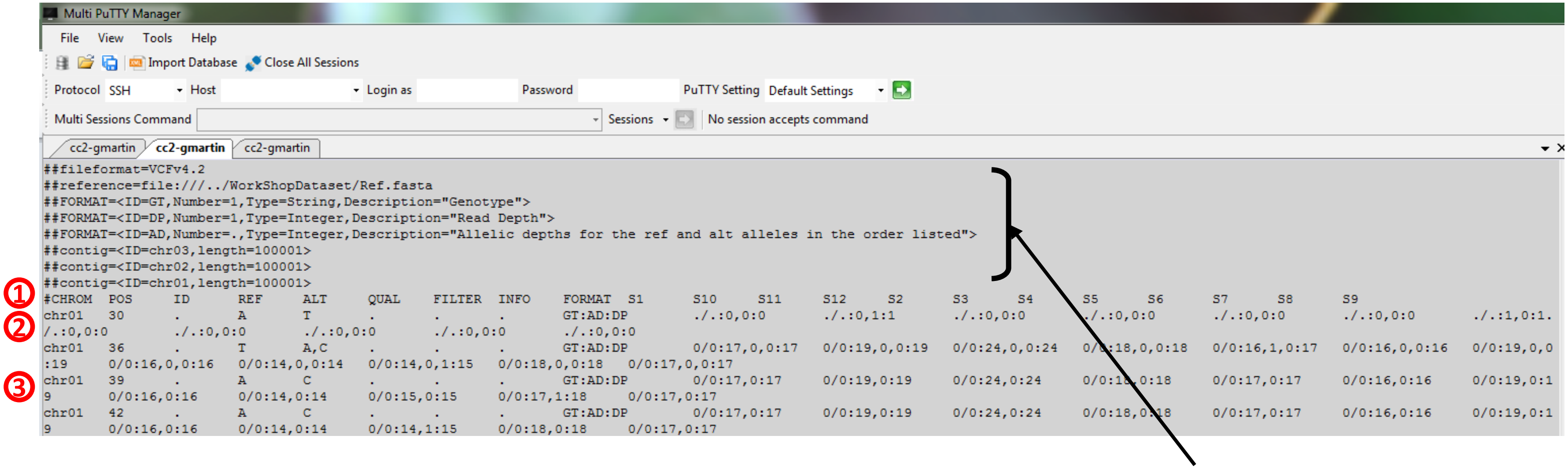
An always empty file associated to -pe parallel_smp 3 options

Three vcf files containing genotyping informations, one for each chromosomes

From the output of GBS sequencing to the variant calling file: in command line

- What can be found in a vcf format:

```
more GBSset_chr01_all_allele_count.vcf
```



Protocol SSH Host Login as Password PuTTY Setting Default Settings

Multi Sessions Command Sessions No session accepts command

```
##fileformat=VCFv4.2
##reference=file:///../WorkShopDataset/Ref.fasta
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##contig=<ID=chr03,length=100001>
##contig=<ID=chr02,length=100001>
##contig=<ID=chr01,length=100001>
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  S1      S10     S11     S12     S2      S3      S4      S5      S6      S7      S8      S9
chr01   30      .       A       T       .       .       .       GT:AD:DP        ./.:0,0:0       ./.:0,1:1       ./.:0,0:0       ./.:0,0:0       ./.:0,0:0       ./.:0,0:0       ./.:1,0:1.
/.:0,0:0        ./.:0,0:0       ./.:0,0:0       ./.:0,0:0       ./.:0,0:0
chr01   36      .       T       A,C     .       .       .       GT:AD:DP        0/0:17,0,0:17   0/0:19,0,0:19   0/0:24,0,0:24   0/0:18,0,0:18   0/0:16,1,0:17   0/0:16,0,0:16   0/0:19,0,0
:19     0/0:16,0,0:16   0/0:14,0,0:14   0/0:14,0,1:15   0/0:18,0,0:18   0/0:17,0,0:17
chr01   39      .       A       C       .       .       .       GT:AD:DP        0/0:17,0:17     0/0:19,0:19     0/0:24,0:24     0/0:18,0:18     0/0:17,0:17     0/0:16,0:16     0/0:19,0:1
9       0/0:16,0:16     0/0:14,0:14     0/0:15,0:15     0/0:17,1:18     0/0:17,0:17
chr01   42      .       A       C       .       .       .       GT:AD:DP        0/0:17,0:17     0/0:19,0:19     0/0:24,0:24     0/0:18,0:18     0/0:17,0:17     0/0:16,0:16     0/0:19,0:1
9       0/0:16,0:16     0/0:14,0:14     0/0:14,1:15     0/0:18,0:18     0/0:17,0:17
```

Header of the vcf file containing information about:

- ✓ **Reference file location**
- ✓ **Genotype format description**
- ✓ **Reference sequence name and size**

① **Real header of variant calling file**

② **Variant line 1**

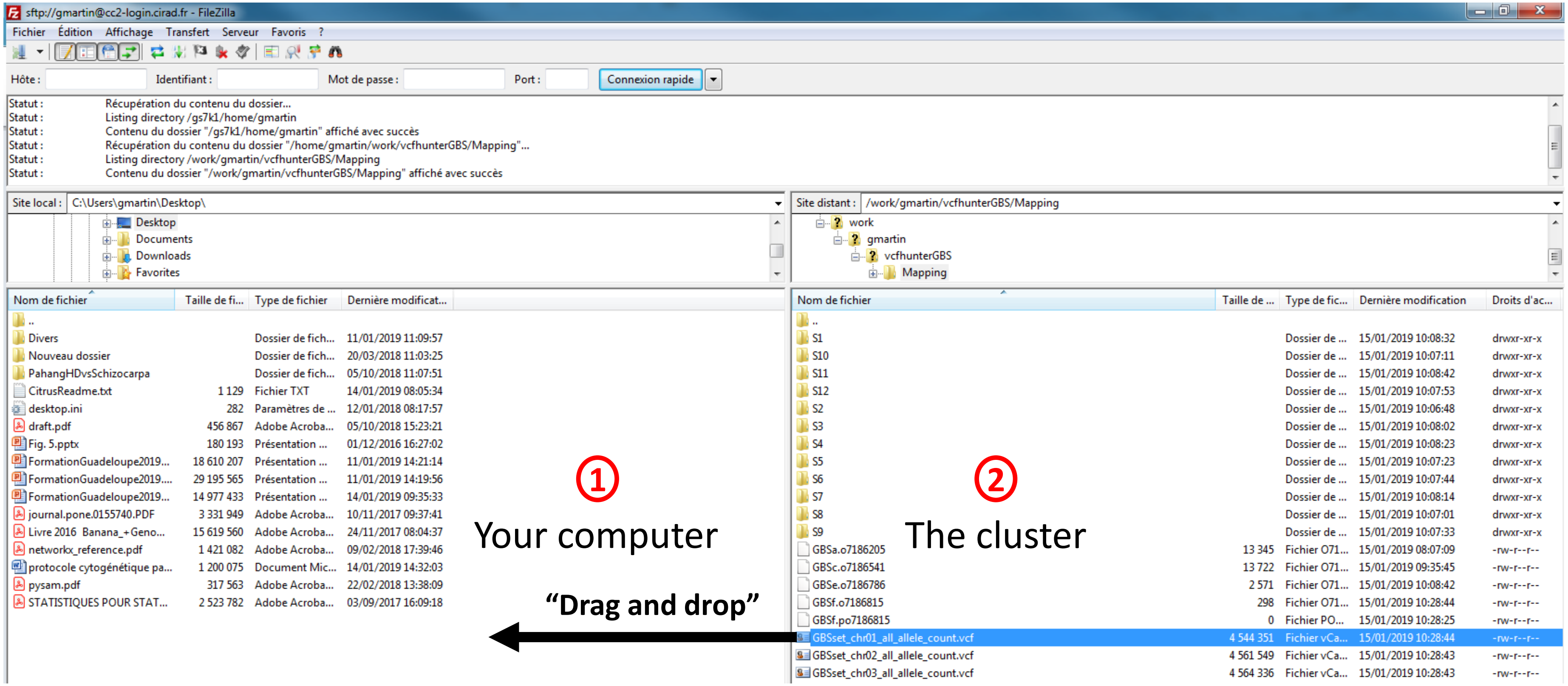
③ **Variant line 3**

- **To quit:** `“Ctrl” + “C”` **or** `“enter”` **until the end of file**

From the output of GBS sequencing to the variant calling file: in command line

- Looking at the vcf file with excel because it is easier (Not to do on real dataset!):

- Using filezilla to get the data on our computer:


sftp://gmartin@cc2-login.cirad.fr - FileZilla
Fichier Édition Affichage Transfert Serveur Favoris ?
Hôte :
Identifiant :
Mot de passe :
Port :
Connexion rapide
Site local : C:\Users\gmartin\Desktop\
Desktop
Documents
Downloads
Favorites
Site distant : /work/gmartin/vcfhunterGBS/Mapping
work
gmartin
vcfhunterGBS
Mapping

```
Statut :        Récupération du contenu du dossier...
Statut :        Listing directory /gs7k1/home/gmartin
Statut :        Contenu du dossier "/gs7k1/home/gmartin" affiché avec succès
Statut :        Récupération du contenu du dossier "/home/gmartin/work/vcfhunterGBS/Mapping"...
Statut :        Listing directory /work/gmartin/vcfhunterGBS/Mapping
Statut :        Contenu du dossier "/work/gmartin/vcfhunterGBS/Mapping" affiché avec succès
```

Nom de fichier	Taille de fi...	Type de fichier	Dernière modificat...
..			
Divers		Dossier de fich...	11/01/2019 11:09:57
Nouveau dossier		Dossier de fich...	20/03/2018 11:03:25
PahangHDvsSchizocarpa		Dossier de fich...	05/10/2018 11:07:51
CitrusReadme.txt	1 129	Fichier TXT	14/01/2019 08:05:34
desktop.ini	282	Paramètres de ...	12/01/2018 08:17:57
draft.pdf	456 867	Adobe Acroba...	05/10/2018 15:23:21
Fig. 5.pptx	180 193	Présentation ...	01/12/2016 16:27:02
FormationGuadeloupe2019...	18 610 207	Présentation ...	11/01/2019 14:21:14
FormationGuadeloupe2019....	29 195 565	Présentation ...	11/01/2019 14:19:56
FormationGuadeloupe2019...	14 977 433	Présentation ...	14/01/2019 09:35:33
journal.pone.0155740.PDF	3 331 949	Adobe Acroba...	10/11/2017 09:37:41
Livre 2016 Banana_+Geno...	15 619 560	Adobe Acroba...	24/11/2017 08:04:37
networkx_reference.pdf	1 421 082	Adobe Acroba...	09/02/2018 17:39:46
protocole cytogénétique pa...	1 200 075	Document Mic...	14/01/2019 14:32:03
pysam.pdf	317 563	Adobe Acroba...	22/02/2018 13:38:09
STATISTIQUES POUR STAT...	2 523 782	Adobe Acroba...	03/09/2017 16:09:18

Nom de fichier	Taille de ...	Type de fic...	Dernière modification	Droits d'ac...
..				
S1		Dossier de ...	15/01/2019 10:08:32	drwxr-xr-x
S10		Dossier de ...	15/01/2019 10:07:11	drwxr-xr-x
S11		Dossier de ...	15/01/2019 10:08:42	drwxr-xr-x
S12		Dossier de ...	15/01/2019 10:07:53	drwxr-xr-x
S2		Dossier de ...	15/01/2019 10:06:48	drwxr-xr-x
S3		Dossier de ...	15/01/2019 10:08:02	drwxr-xr-x
S4		Dossier de ...	15/01/2019 10:08:23	drwxr-xr-x
S5		Dossier de ...	15/01/2019 10:07:23	drwxr-xr-x
S6		Dossier de ...	15/01/2019 10:07:44	drwxr-xr-x
S7		Dossier de ...	15/01/2019 10:08:14	drwxr-xr-x
S8		Dossier de ...	15/01/2019 10:07:01	drwxr-xr-x
S9		Dossier de ...	15/01/2019 10:07:33	drwxr-xr-x
GBSa.o7186205	13 345	Fichier O71...	15/01/2019 08:07:09	-rw-r--r--
GBSc.o7186541	13 722	Fichier O71...	15/01/2019 09:35:45	-rw-r--r--
GBSe.o7186786	2 571	Fichier O71...	15/01/2019 10:08:42	-rw-r--r--
GBSf.o7186815	298	Fichier O71...	15/01/2019 10:28:44	-rw-r--r--
GBSf.po7186815	0	Fichier PO...	15/01/2019 10:28:25	-rw-r--r--
GBSset_chr01_all_allele_count.vcf	4 544 351	Fichier vCa...	15/01/2019 10:28:44	-rw-r--r--
GBSset_chr02_all_allele_count.vcf	4 561 549	Fichier vCa...	15/01/2019 10:28:43	-rw-r--r--
GBSset_chr03_all_allele_count.vcf	4 564 336	Fichier vCa...	15/01/2019 10:28:43	-rw-r--r--

① Your computer

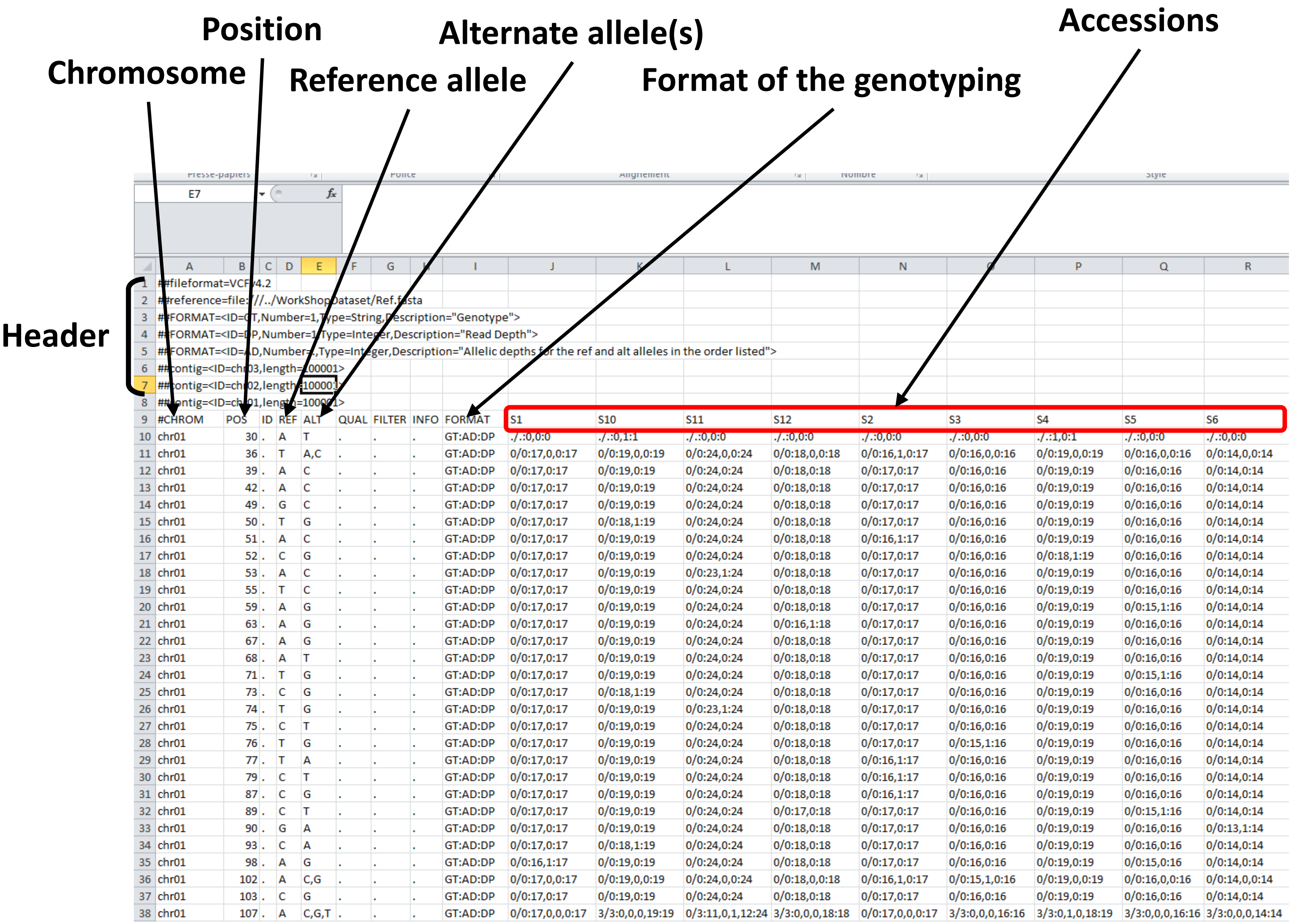
② The cluster

"Drag and drop"

- Open it with Excel!

From the output of GBS sequencing to the variant calling file: in command line

- The vcf file format:



Position

Chromosome

Alternate allele(s)

Reference allele

Format of the genotyping

Accessions

Header

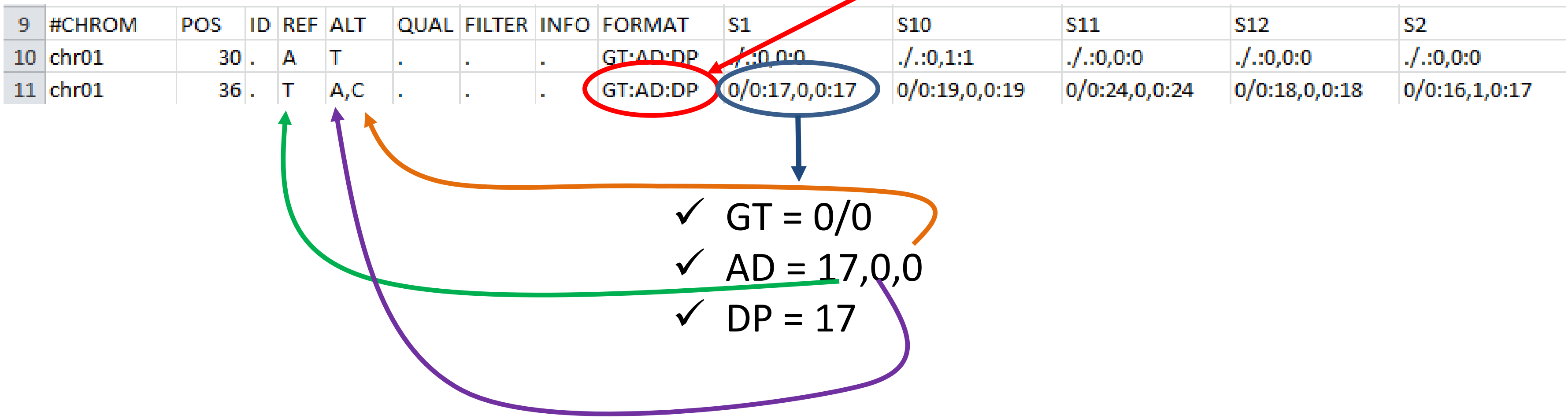
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	##fileformat=VCFv4.2																	
2	##reference=file:///../WorkShopDataset/Ref.fasta																	
3	##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">																	
4	##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">																	
5	##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">																	
6	##contig=<ID=chr03,length=100001>																	
7	##contig=<ID=chr02,length=100001>																	
8	##contig=<ID=chr01,length=100001>																	
9	#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	S1	S10	S11	S12	S2	S3	S4	S5	S6
10	chr01	30	.	A	T	.	.	.	GT:AD:DP	./.:0,0:0	./.:0,1:1	./.:0,0:0	./.:0,0:0	./.:0,0:0	./.:0,0:0	./.:1,0:1	./.:0,0:0	./.:0,0:0
11	chr01	36	.	T	A,C	.	.	.	GT:AD:DP	0/0:17,0,0:17	0/0:19,0,0:19	0/0:24,0,0:24	0/0:18,0,0:18	0/0:16,1,0:17	0/0:16,0,0:16	0/0:19,0,0:19	0/0:16,0,0:16	0/0:14,0,0:14
12	chr01	39	.	A	C	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14
13	chr01	42	.	A	C	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14
14	chr01	49	.	G	C	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14
15	chr01	50	.	T	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:18,1:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14
16	chr01	51	.	A	C	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:16,1:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14
17	chr01	52	.	C	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:18,1:19	0/0:16,0:16	0/0:14,0:14
18	chr01	53	.	A	C	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:23,1:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14
19	chr01	55	.	T	C	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14
20	chr01	59	.	A	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:15,1:16	0/0:14,0:14
21	chr01	63	.	A	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:16,1:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14
22	chr01	67	.	A	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14
23	chr01	68	.	A	T	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14
24	chr01	71	.	T	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:15,1:16	0/0:14,0:14
25	chr01	73	.	C	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:18,1:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14
26	chr01	74	.	T	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:23,1:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14
27	chr01	75	.	C	T	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14
28	chr01	76	.	T	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:15,1:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14
29	chr01	77	.	T	A	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:16,1:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14
30	chr01	79	.	C	T	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:16,1:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14
31	chr01	87	.	C	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:16,1:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14
32	chr01	89	.	C	T	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:17,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:15,1:16	0/0:14,0:14
33	chr01	90	.	G	A	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:13,1:14
34	chr01	93	.	C	A	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:18,1:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14
35	chr01	98	.	A	G	.	.	.	GT:AD:DP	0/0:16,1:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:15,0:16	0/0:14,0:14
36	chr01	102	.	A	C,G	.	.	.	GT:AD:DP	0/0:17,0,0:17	0/0:19,0,0:19	0/0:24,0,0:24	0/0:18,0,0:18	0/0:16,1,0:17	0/0:15,1,0:16	0/0:19,0,0:19	0/0:16,0,0:16	0/0:14,0,0:14
37	chr01	103	.	C	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14
38	chr01	107	.	A	C,G,T	.	.	.	GT:AD:DP	0/0:17,0,0,0:17	3/3:0,0,0,19:19	0/3:11,0,1,12:24	3/3:0,0,0,18:18	0/0:17,0,0,0:17	3/3:0,0,0,16:16	3/3:0,1,0,18:19	3/3:0,0,0,16:16	3/3:0,0,0,14:14

From the output of GBS sequencing to the variant calling file: in command line

- The vcf file format:

Describe the way the genotype is formatted for each accessions:

- ✓ GT = genotype
- ✓ AD = allele depth
- ✓ DP = depth



9	#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	S1	S10	S11	S12	S2
10	chr01	30	.	A	T	.	.	.	GT:AD:DP	./.:0,0:0	./.:0,1:1	./.:0,0:0	./.:0,0:0	./.:0,0:0
11	chr01	36	.	T	A,C	.	.	.	GT:AD:DP	0/0:17,0,0:17	0/0:19,0,0:19	0/0:24,0,0:24	0/0:18,0,0:18	0/0:16,1,0:17

- ✓ GT = 0/0
- ✓ AD = 17,0,0
- ✓ DP = 17

Based on these allelic depths, calculation of the likelihood of each haplotypes:

- ✓ 0/0 = T/T
- ✓ 0/1 = T/A
- ✓ 0/2 = T/C
- ✓ 1/2 = A/C
- ✓ 1/1 = A/A
- ✓ 2/2 = C/C

From the output of GBS sequencing to the variant calling file: in command line

- Because it is sometime easier to have only one file for all chromosomes, this unique file can be produced with this last command line:

```
qsub -q normal.q -b yes -V -N GBSg "process_reseq_1.0.py -c
../WorkShopDataset/GBSCalling.conf -p GBSset -s g -t 1"
```

- The first part of the command line (in bold) is **used by the master computer** (as previously described):

  - **qsub**: Means that we will send a command that the master computer needs to analyze to choose the best computer

  - **-q normal.q**: tells the master computer that we will use computer from normal queue.

  - **-b yes**: it is not important, but put it.

  - **-V**: Tell the master computer to load the module previously loaded on the computer it will choose

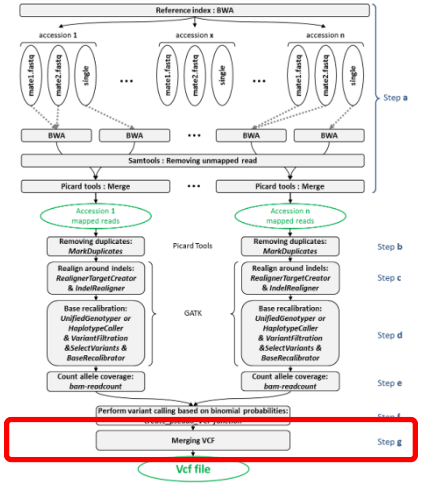
  - **-N GBSf**: A name passed to the command line to look at its status (waiting, running or error) on the cluster

From the output of GBS sequencing to the variant calling file: in command line

- Because it is sometime easier to have only one file for all chromosomes, this unique file can be produced with this last command line:

```
qsub -q normal.q -b yes -V -N GBSg "process_reseq_1.0.py -c
../WorkShopDataset/GBSCalling.conf -p GBSset -s g -t 1"
```

- The part of the command line between quotation marks (in bold) is the command line that is executed on the **computer chosen by the master computer**:
  - **process_reseq_1.0.py**: We will use process_reseq_1.0.py program
  - **-c ../WorkShopDataset/GBSCalling.conf**: Locates the configuration file
  - **-p GBSset**: A prefix for final output file
  - **-s g**: Tell the program that we will perform step “g” of the workflow



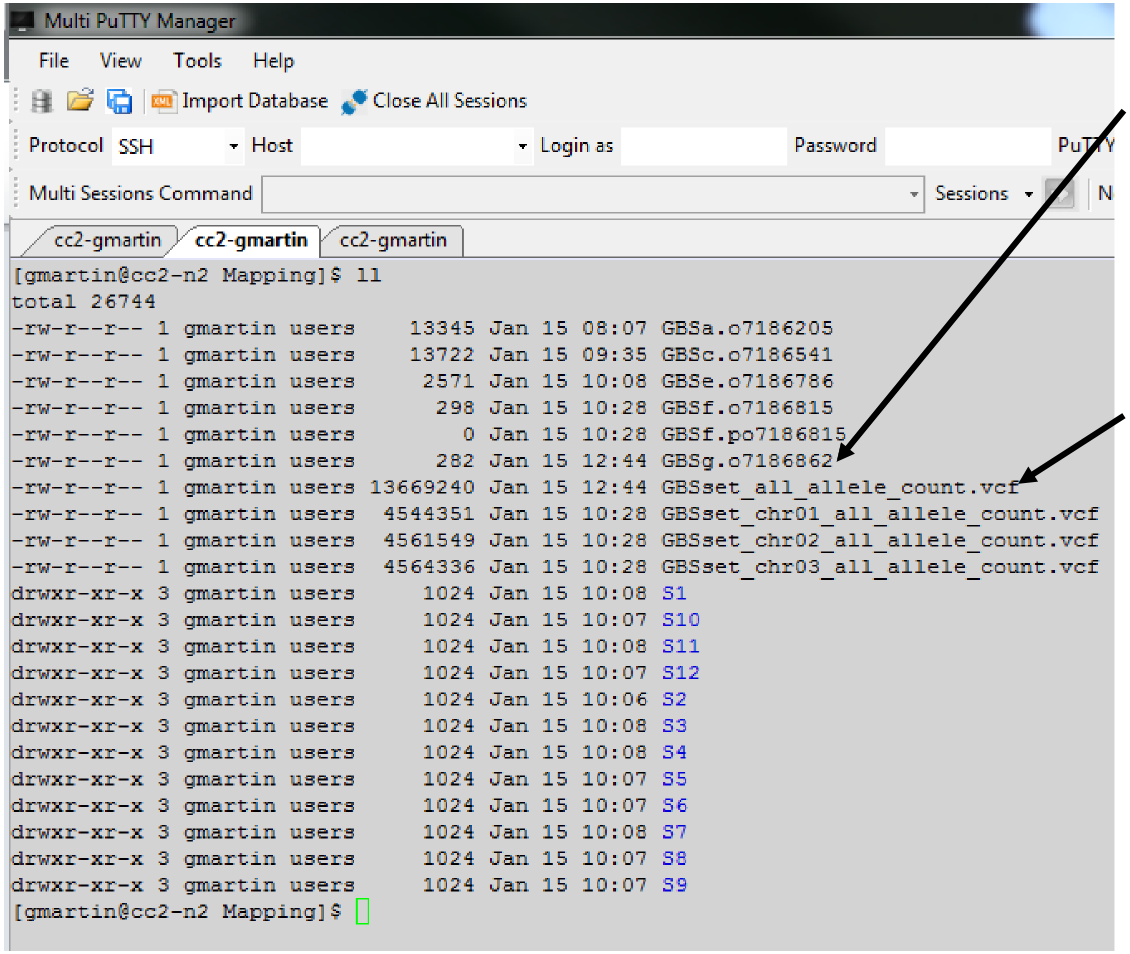
Reference index : BWA
accession 1
accession x
accession n
mate1.fastq
mate2.fastq
single
...
BWA
Samtools : Removing unmapped read
Picard tools : Merge
Step a
Accession 1 mapped reads
Accession n mapped reads
Removing duplicates: MarkDuplicates
Picard Tools
Step b
Realign around indels: RealignerTargetCreator & IndelRealigner
Step c
GATK
Base recalibration: UnifiedGenotyper or HaplotypeCaller & VariantFiltration &SelectVariants & BaseRecalibrator
Step d
Count allele coverage: bam-readcount
Step e
Perform variant calling based on binomial probabilities:
Step f
Merging VCF
Step g
Vcf file

  - **-t 1**: Tell the program that only one processor is available.

From the output of GBS sequencing to the variant calling file: in command line

- Listing the file generated:

  ll


Multi PuTTY Manager
File View Tools Help
Import Database Close All Sessions
Protocol SSH Host Login as Password PuTTY
Multi Sessions Command Sessions N
cc2-gmartin cc2-gmartin cc2-gmartin

```
[gmartin@cc2-n2 Mapping]$ ll
total 26744
-rw-r--r-- 1 gmartin users    13345 Jan 15 08:07 GBSa.o7186205
-rw-r--r-- 1 gmartin users    13722 Jan 15 09:35 GBSc.o7186541
-rw-r--r-- 1 gmartin users     2571 Jan 15 10:08 GBSe.o7186786
-rw-r--r-- 1 gmartin users      298 Jan 15 10:28 GBSf.o7186815
-rw-r--r-- 1 gmartin users        0 Jan 15 10:28 GBSf.po7186815
-rw-r--r-- 1 gmartin users      282 Jan 15 12:44 GBSg.o7186862
-rw-r--r-- 1 gmartin users 13669240 Jan 15 12:44 GBSset_all_allele_count.vcf
-rw-r--r-- 1 gmartin users  4544351 Jan 15 10:28 GBSset_chr01_all_allele_count.vcf
-rw-r--r-- 1 gmartin users  4561549 Jan 15 10:28 GBSset_chr02_all_allele_count.vcf
-rw-r--r-- 1 gmartin users  4564336 Jan 15 10:28 GBSset_chr03_all_allele_count.vcf
drwxr-xr-x 3 gmartin users     1024 Jan 15 10:08 S1
drwxr-xr-x 3 gmartin users     1024 Jan 15 10:07 S10
drwxr-xr-x 3 gmartin users     1024 Jan 15 10:08 S11
drwxr-xr-x 3 gmartin users     1024 Jan 15 10:07 S12
drwxr-xr-x 3 gmartin users     1024 Jan 15 10:06 S2
drwxr-xr-x 3 gmartin users     1024 Jan 15 10:08 S3
drwxr-xr-x 3 gmartin users     1024 Jan 15 10:08 S4
drwxr-xr-x 3 gmartin users     1024 Jan 15 10:07 S5
drwxr-xr-x 3 gmartin users     1024 Jan 15 10:07 S6
drwxr-xr-x 3 gmartin users     1024 Jan 15 10:08 S7
drwxr-xr-x 3 gmartin users     1024 Jan 15 10:07 S8
drwxr-xr-x 3 gmartin users     1024 Jan 15 10:07 S9
[gmartin@cc2-n2 Mapping]$
```

The GBSg.oxxxxxxx file containing what process_reseq_1.0.py told us while it was executing

A vcf file containing all chromosomes

From the output of GBS sequencing to the variant calling file: in command line

- To discriminate between sequencing errors and true variant site we developed an additional program which allowed to select true polymorphous SNP according to selected parameters. This program is called ***VcfPreFilter.1.0.py*** and can be executed with the following command line:

```
qsub -q normal.q -b yes -V -N PREFLTR "VcfPreFilter.1.0.py -v
GBSset_all_allele_count.vcf -m 10 -M 10000 -f 0.05 -c 3 -o
GBSset_prefiltered.vcf -d y"
```

- The first part of the command line (in bold) is **used by the master computer** (as previously described):

  - **qsub**: Means that we will send a command that the master computer needs to analyze to choose the best computer
  - **-q normal.q**: tells the master computer that we will use computer from normal queue.
  - **-b yes**: it is not important, but put it.
  - **-V**: Tell the master computer to load the module previously loaded on the computer it will choose
  - **-N PREFLTR**: A name passed to the command line to look at its status (waiting, running or error) on the cluster

From the output of GBS sequencing to the variant calling file: in command line

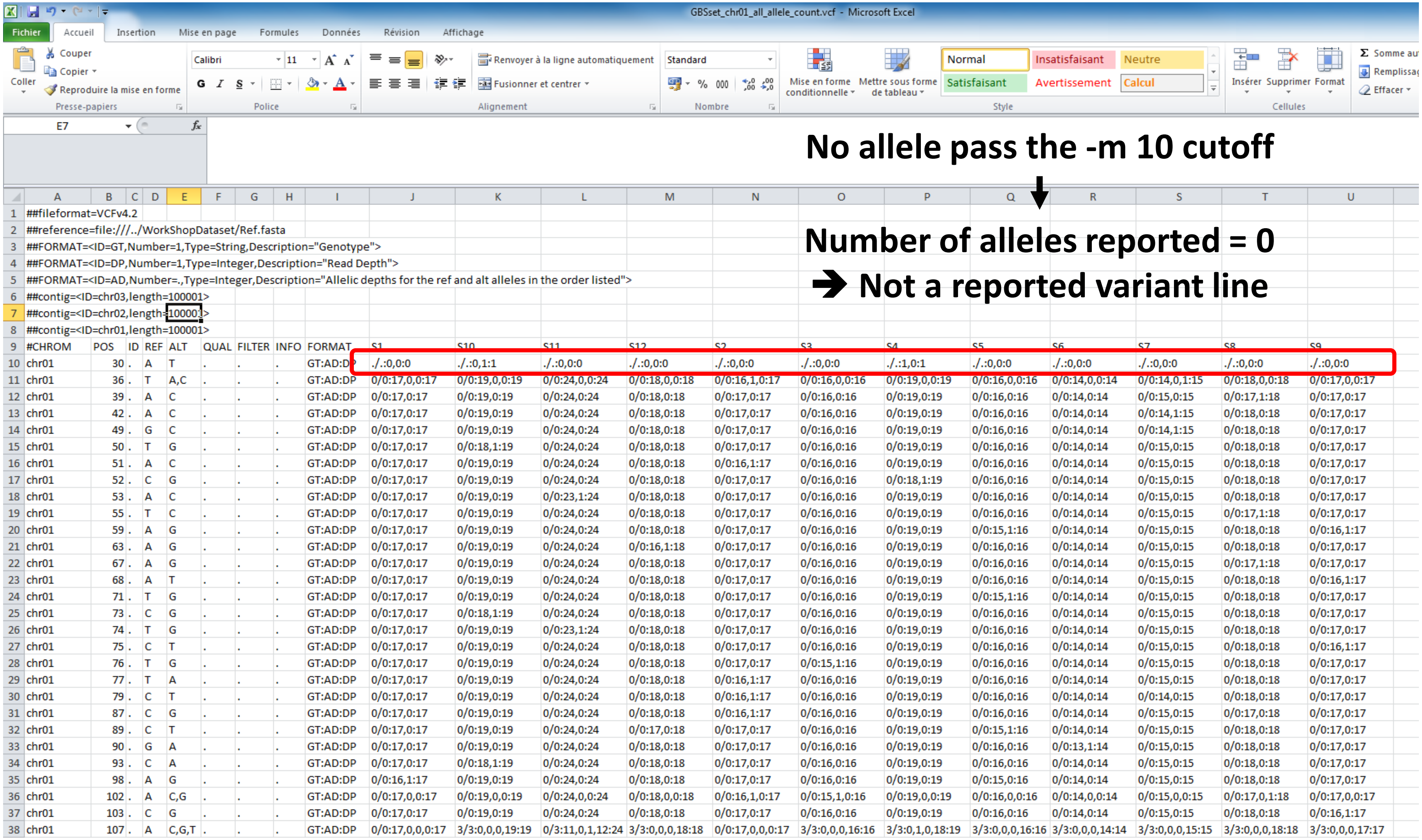
```
qsub -q normal.q -b yes -V -N PREFLTR "VcfPreFilter.1.0.py -v
GBSset_all_allele_count.vcf -m 10 -M 10000 -f 0.05 -c 3 -o
GBSset_prefiltered.vcf -d y"
```

- The part of the command line between quotation marks (in bold) is the command line that is executed on the **computer chosen by the master computer**:
  - **VcfPreFilter.1.0.py**: We will use VcfPreFilter.1.0.py program
  - **-v GBSset_all_allele_count.vcf**: Locates the vcf file to filter
  - **-m 10** : Only datapoint with coverage supported by more than 10 reads will be considered
  - **-M 10000**: Only datapoint with coverage supported by less than 10000 reads will be considered (to manage very high repeats)
  - **-f 0.05**: An allele is kept if it is present in at least this proportion in at least one accession.
  - **-c 3**: An allele is kept if it is supported by at least **3** reads in at least one accession.
  - **-o GBSset_prefiltered.vcf**: Name of the output file
  - **-d y**: Perform only diallelic calling (i.e. for triploid accessions, A/C/G genotype is not possible because only two alleles are allowed in a genotype: A/A/C or A/G/G, or … genotype are tested).

- According to -m, -M, -f and -c parameters the number of possible alleles is counted (including the reference sequence allele, and if this number is strictly greater than 1, the line is identified as a polymorphous line that should be reported)

From the output of GBS sequencing to the variant calling file: in command line

- Prefiltering example: `-m 10 -M 10000 -f 0.05 -c 3`



No allele pass the -m 10 cutoff

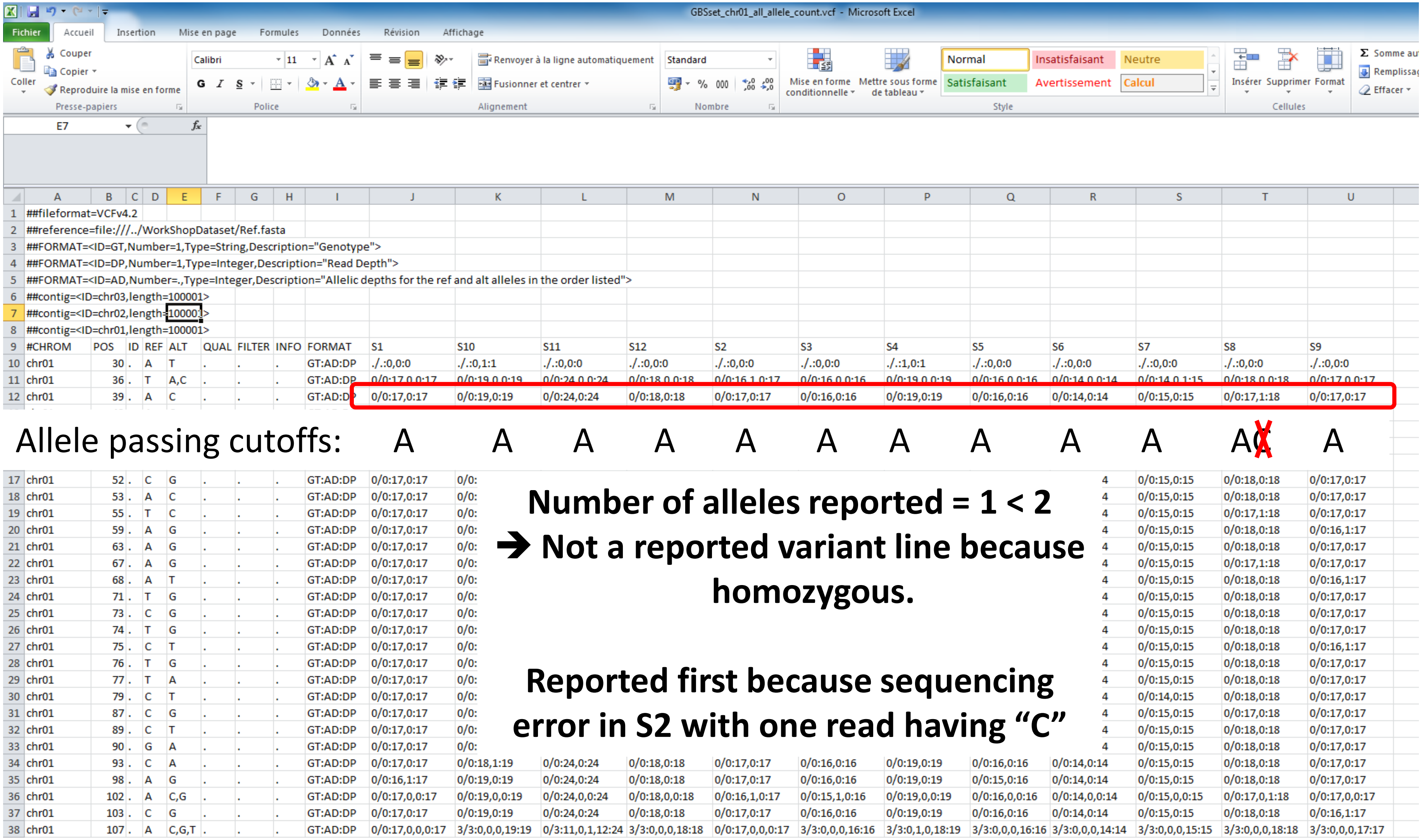
Number of alleles reported = 0

➔ Not a reported variant line

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	##fileformat=VCFv4.2																				
2	##reference=file:///../WorkShopDataset/Ref.fasta																				
3	##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">																				
4	##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">																				
5	##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">																				
6	##contig=<ID=chr03,length=100001>																				
7	##contig=<ID=chr02,length=100001>																				
8	##contig=<ID=chr01,length=100001>																				
9	#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	S1	S10	S11	S12	S2	S3	S4	S5	S6	S7	S8	S9
10	chr01	30	.	A	T	.	.	.	GT:AD:DP	./.:0,0:0	./.:0,1:1	./.:0,0:0	./.:0,0:0	./.:0,0:0	./.:0,0:0	./.:1,0:1	./.:0,0:0	./.:0,0:0	./.:0,0:0	./.:0,0:0	./.:0,0:0
11	chr01	36	.	T	A,C	.	.	.	GT:AD:DP	0/0:17,0,0:17	0/0:19,0,0:19	0/0:24,0,0:24	0/0:18,0,0:18	0/0:16,1,0:17	0/0:16,0,0:16	0/0:19,0,0:19	0/0:16,0,0:16	0/0:14,0,0:14	0/0:14,0,1:15	0/0:18,0,0:18	0/0:17,0,0:17
12	chr01	39	.	A	C	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14	0/0:15,0:15	0/0:17,1:18	0/0:17,0:17
13	chr01	42	.	A	C	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14	0/0:14,1:15	0/0:18,0:18	0/0:17,0:17
14	chr01	49	.	G	C	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14	0/0:14,1:15	0/0:18,0:18	0/0:17,0:17
15	chr01	50	.	T	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:18,1:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14	0/0:15,0:15	0/0:18,0:18	0/0:17,0:17
16	chr01	51	.	A	C	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:16,1:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14	0/0:15,0:15	0/0:18,0:18	0/0:17,0:17
17	chr01	52	.	C	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:18,1:19	0/0:16,0:16	0/0:14,0:14	0/0:15,0:15	0/0:18,0:18	0/0:17,0:17
18	chr01	53	.	A	C	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:23,1:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14	0/0:15,0:15	0/0:18,0:18	0/0:17,0:17
19	chr01	55	.	T	C	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14	0/0:15,0:15	0/0:17,1:18	0/0:17,0:17
20	chr01	59	.	A	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:15,1:16	0/0:14,0:14	0/0:15,0:15	0/0:18,0:18	0/0:16,1:17
21	chr01	63	.	A	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:16,1:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14	0/0:15,0:15	0/0:18,0:18	0/0:17,0:17
22	chr01	67	.	A	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14	0/0:15,0:15	0/0:17,1:18	0/0:17,0:17
23	chr01	68	.	A	T	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14	0/0:15,0:15	0/0:18,0:18	0/0:16,1:17
24	chr01	71	.	T	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:15,1:16	0/0:14,0:14	0/0:15,0:15	0/0:18,0:18	0/0:17,0:17
25	chr01	73	.	C	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:18,1:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14	0/0:15,0:15	0/0:18,0:18	0/0:17,0:17
26	chr01	74	.	T	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:23,1:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14	0/0:15,0:15	0/0:18,0:18	0/0:17,0:17
27	chr01	75	.	C	T	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14	0/0:15,0:15	0/0:18,0:18	0/0:16,1:17
28	chr01	76	.	T	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:15,1:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14	0/0:15,0:15	0/0:18,0:18	0/0:17,0:17
29	chr01	77	.	T	A	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:16,1:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14	0/0:15,0:15	0/0:18,0:18	0/0:17,0:17
30	chr01	79	.	C	T	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:16,1:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14	0/0:14,0:15	0/0:18,0:18	0/0:17,0:17
31	chr01	87	.	C	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:16,1:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14	0/0:15,0:15	0/0:17,0:18	0/0:17,0:17
32	chr01	89	.	C	T	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:17,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:15,1:16	0/0:14,0:14	0/0:15,0:15	0/0:18,0:18	0/0:17,0:17
33	chr01	90	.	G	A	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:13,1:14	0/0:15,0:15	0/0:18,0:18	0/0:17,0:17
34	chr01	93	.	C	A	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:18,1:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14	0/0:15,0:15	0/0:18,0:18	0/0:17,0:17
35	chr01	98	.	A	G	.	.	.	GT:AD:DP	0/0:16,1:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:15,0:16	0/0:14,0:14	0/0:15,0:15	0/0:18,0:18	0/0:17,0:17
36	chr01	102	.	A	C,G	.	.	.	GT:AD:DP	0/0:17,0,0:17	0/0:19,0,0:19	0/0:24,0,0:24	0/0:18,0,0:18	0/0:16,1,0:17	0/0:15,1,0:16	0/0:19,0,0:19	0/0:16,0,0:16	0/0:14,0,0:14	0/0:15,0,0:15	0/0:17,0,1:18	0/0:17,0,0:17
37	chr01	103	.	C	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14	0/0:15,0:15	0/0:18,0:18	0/0:16,1:17
38	chr01	107	.	A	C,G,T	.	.	.	GT:AD:DP	0/0:17,0,0,0:17	3/3:0,0,0,19:19	0/3:11,0,1,12:24	3/3:0,0,0,18:18	0/0:17,0,0,0:17	3/3:0,0,0,16:16	3/3:0,1,0,18:19	3/3:0,0,0,16:16	3/3:0,0,0,14:14	3/3:0,0,0,15:15	3/3:0,0,0,18:18	3/3:0,0,0,17:17

From the output of GBS sequencing to the variant calling file: in command line

- Prefiltering example: `-m 10 -M 10000 -f 0.05 -c 3`



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	##fileformat=VCFv4.2																				
2	##reference=file:///../WorkShopDataset/Ref.fasta																				
3	##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">																				
4	##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">																				
5	##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">																				
6	##contig=<ID=chr03,length=100001>																				
7	##contig=<ID=chr02,length=100001>																				
8	##contig=<ID=chr01,length=100001>																				
9	#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	S1	S10	S11	S12	S2	S3	S4	S5	S6	S7	S8	S9
10	chr01	30	.	A	T	.	.	.	GT:AD:DP	./.:0,0:0	./.:0,1:1	./.:0,0:0	./.:0,0:0	./.:0,0:0	./.:0,0:0	./.:1,0:1	./.:0,0:0	./.:0,0:0	./.:0,0:0	./.:0,0:0	./.:0,0:0
11	chr01	36	.	T	A,C	.	.	.	GT:AD:DP	0/0:17,0,0:17	0/0:19,0,0:19	0/0:24,0,0:24	0/0:18,0,0:18	0/0:16,1,0:17	0/0:16,0,0:16	0/0:19,0,0:19	0/0:16,0,0:16	0/0:14,0,0:14	0/0:14,0,1:15	0/0:18,0,0:18	0/0:17,0,0:17
12	chr01	39	.	A	C	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14	0/0:15,0:15	0/0:17,1:18	0/0:17,0:17

Allele passing cutoffs: A A A A A A A A A A A~~C~~ A

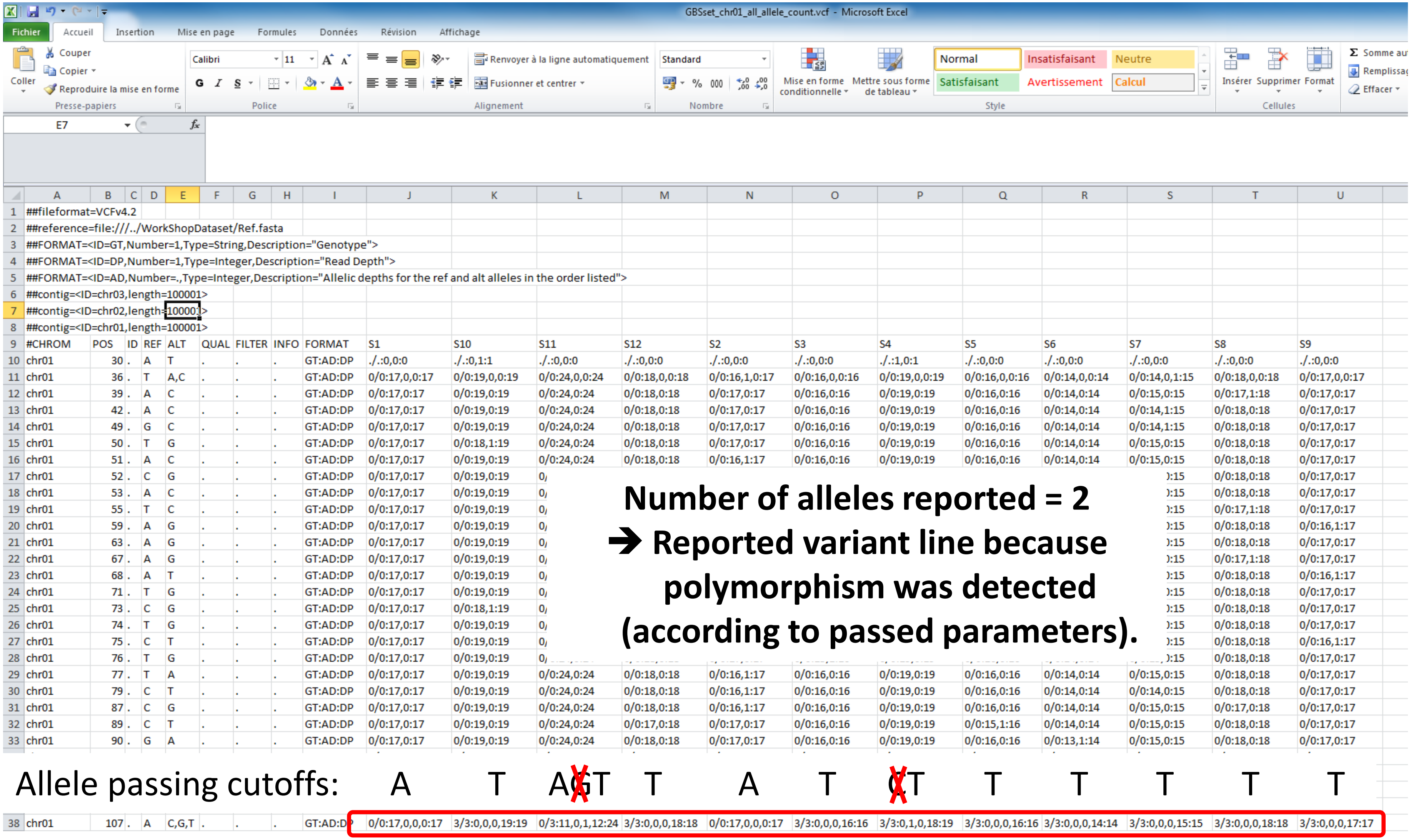
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
17	chr01	52	.	C	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:							4	0/0:15,0:15	0/0:18,0:18	0/0:17,0:17
18	chr01	53	.	A	C	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:							4	0/0:15,0:15	0/0:18,0:18	0/0:17,0:17
19	chr01	55	.	T	C	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:							4	0/0:15,0:15	0/0:17,1:18	0/0:17,0:17
20	chr01	59	.	A	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:							4	0/0:15,0:15	0/0:18,0:18	0/0:16,1:17
21	chr01	63	.	A	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:							4	0/0:15,0:15	0/0:18,0:18	0/0:17,0:17
22	chr01	67	.	A	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:							4	0/0:15,0:15	0/0:17,1:18	0/0:17,0:17
23	chr01	68	.	A	T	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:							4	0/0:15,0:15	0/0:18,0:18	0/0:16,1:17
24	chr01	71	.	T	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:							4	0/0:15,0:15	0/0:18,0:18	0/0:17,0:17
25	chr01	73	.	C	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:							4	0/0:15,0:15	0/0:18,0:18	0/0:17,0:17
26	chr01	74	.	T	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:							4	0/0:15,0:15	0/0:18,0:18	0/0:17,0:17
27	chr01	75	.	C	T	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:							4	0/0:15,0:15	0/0:18,0:18	0/0:16,1:17
28	chr01	76	.	T	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:							4	0/0:15,0:15	0/0:18,0:18	0/0:17,0:17
29	chr01	77	.	T	A	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:							4	0/0:15,0:15	0/0:18,0:18	0/0:17,0:17
30	chr01	79	.	C	T	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:							4	0/0:14,0:15	0/0:18,0:18	0/0:17,0:17
31	chr01	87	.	C	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:							4	0/0:15,0:15	0/0:17,0:18	0/0:17,0:17
32	chr01	89	.	C	T	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:							4	0/0:15,0:15	0/0:18,0:18	0/0:17,0:17
33	chr01	90	.	G	A	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:							4	0/0:15,0:15	0/0:18,0:18	0/0:17,0:17
34	chr01	93	.	C	A	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:18,1:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14	0/0:15,0:15	0/0:18,0:18	0/0:17,0:17
35	chr01	98	.	A	G	.	.	.	GT:AD:DP	0/0:16,1:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:15,0:16	0/0:14,0:14	0/0:15,0:15	0/0:18,0:18	0/0:17,0:17
36	chr01	102	.	A	C,G	.	.	.	GT:AD:DP	0/0:17,0,0:17	0/0:19,0,0:19	0/0:24,0,0:24	0/0:18,0,0:18	0/0:16,1,0:17	0/0:15,1,0:16	0/0:19,0,0:19	0/0:16,0,0:16	0/0:14,0,0:14	0/0:15,0,0:15	0/0:17,0,1:18	0/0:17,0,0:17
37	chr01	103	.	C	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14	0/0:15,0:15	0/0:18,0:18	0/0:16,1:17
38	chr01	107	.	A	C,G,T	.	.	.	GT:AD:DP	0/0:17,0,0,0:17	3/3:0,0,0,19:19	0/3:11,0,1,12:24	3/3:0,0,0,18:18	0/0:17,0,0,0:17	3/3:0,0,0,16:16	3/3:0,1,0,18:19	3/3:0,0,0,16:16	3/3:0,0,0,14:14	3/3:0,0,0,15:15	3/3:0,0,0,18:18	3/3:0,0,0,17:17

Number of alleles reported = 1 < 2
➔ Not a reported variant line because homozygous.

Reported first because sequencing error in S2 with one read having “C”

From the output of GBS sequencing to the variant calling file: in command line

- Prefiltering example: `-m 10 -M 10000 -f 0.05 -c 3`



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	##fileformat=VCFv4.2																				
2	##reference=file:///../WorkShopDataset/Ref.fasta																				
3	##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">																				
4	##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">																				
5	##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">																				
6	##contig=<ID=chr03,length=100001>																				
7	##contig=<ID=chr02,length=100001>																				
8	##contig=<ID=chr01,length=100001>																				
9	#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	S1	S10	S11	S12	S2	S3	S4	S5	S6	S7	S8	S9
10	chr01	30	.	A	T	.	.	.	GT:AD:DP	./.:0,0:0	./.:0,1:1	./.:0,0:0	./.:0,0:0	./.:0,0:0	./.:0,0:0	./.:1,0:1	./.:0,0:0	./.:0,0:0	./.:0,0:0	./.:0,0:0	./.:0,0:0
11	chr01	36	.	T	A,C	.	.	.	GT:AD:DP	0/0:17,0,0:17	0/0:19,0,0:19	0/0:24,0,0:24	0/0:18,0,0:18	0/0:16,1,0:17	0/0:16,0,0:16	0/0:19,0,0:19	0/0:16,0,0:16	0/0:14,0,0:14	0/0:14,0,1:15	0/0:18,0,0:18	0/0:17,0,0:17
12	chr01	39	.	A	C	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14	0/0:15,0:15	0/0:17,1:18	0/0:17,0:17
13	chr01	42	.	A	C	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14	0/0:14,1:15	0/0:18,0:18	0/0:17,0:17
14	chr01	49	.	G	C	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14	0/0:14,1:15	0/0:18,0:18	0/0:17,0:17
15	chr01	50	.	T	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:18,1:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14	0/0:15,0:15	0/0:18,0:18	0/0:17,0:17
16	chr01	51	.	A	C	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:16,1:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14	0/0:15,0:15	0/0:18,0:18	0/0:17,0:17
17	chr01	52	.	C	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/							:15	0/0:18,0:18	0/0:17,0:17
18	chr01	53	.	A	C	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/							:15	0/0:18,0:18	0/0:17,0:17
19	chr01	55	.	T	C	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/							:15	0/0:17,1:18	0/0:17,0:17
20	chr01	59	.	A	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/							:15	0/0:18,0:18	0/0:16,1:17
21	chr01	63	.	A	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/							:15	0/0:18,0:18	0/0:17,0:17
22	chr01	67	.	A	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/							:15	0/0:17,1:18	0/0:17,0:17
23	chr01	68	.	A	T	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/							:15	0/0:18,0:18	0/0:16,1:17
24	chr01	71	.	T	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/							:15	0/0:18,0:18	0/0:17,0:17
25	chr01	73	.	C	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:18,1:19	0/							:15	0/0:18,0:18	0/0:17,0:17
26	chr01	74	.	T	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/							:15	0/0:18,0:18	0/0:17,0:17
27	chr01	75	.	C	T	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/							:15	0/0:18,0:18	0/0:16,1:17
28	chr01	76	.	T	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/							:15	0/0:18,0:18	0/0:17,0:17
29	chr01	77	.	T	A	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:16,1:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14	0/0:15,0:15	0/0:18,0:18	0/0:17,0:17
30	chr01	79	.	C	T	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:16,1:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14	0/0:14,0:15	0/0:18,0:18	0/0:17,0:17
31	chr01	87	.	C	G	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:16,1:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:14,0:14	0/0:15,0:15	0/0:17,0:18	0/0:17,0:17
32	chr01	89	.	C	T	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:17,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:15,1:16	0/0:14,0:14	0/0:15,0:15	0/0:18,0:18	0/0:17,0:17
33	chr01	90	.	G	A	.	.	.	GT:AD:DP	0/0:17,0:17	0/0:19,0:19	0/0:24,0:24	0/0:18,0:18	0/0:17,0:17	0/0:16,0:16	0/0:19,0:19	0/0:16,0:16	0/0:13,1:14	0/0:15,0:15	0/0:18,0:18	0/0:17,0:17
38	chr01	107	.	A	C,G,T	.	.	.	GT:AD:DP	0/0:17,0,0,0:17	3/3:0,0,0,19:19	0/3:11,0,1,12:24	3/3:0,0,0,18:18	0/0:17,0,0,0:17	3/3:0,0,0,16:16	3/3:0,1,0,18:19	3/3:0,0,0,16:16	3/3:0,0,0,14:14	3/3:0,0,0,15:15	3/3:0,0,0,18:18	3/3:0,0,0,17:17

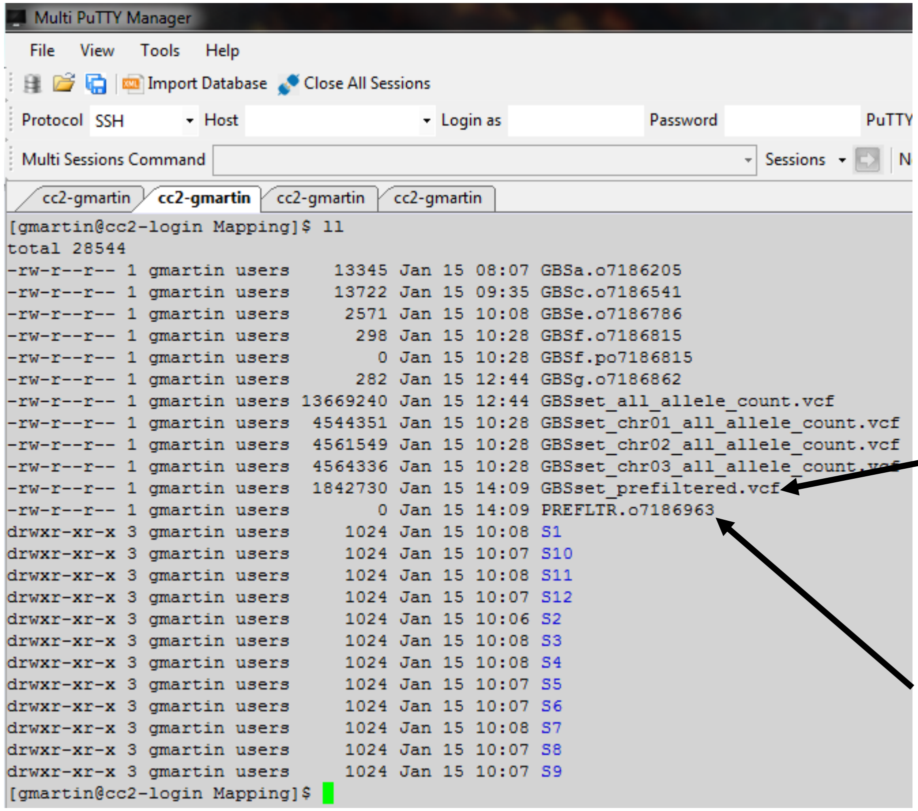
Number of alleles reported = 2
➔ Reported variant line because polymorphism was detected (according to passed parameters).

Allele passing cutoffs: A T A~~G~~T T A T ~~C~~T T T T T T

From the output of GBS sequencing to the variant calling file: in command line

- Listing the files generated:

  ```
  ll
  ```



Protocol SSH

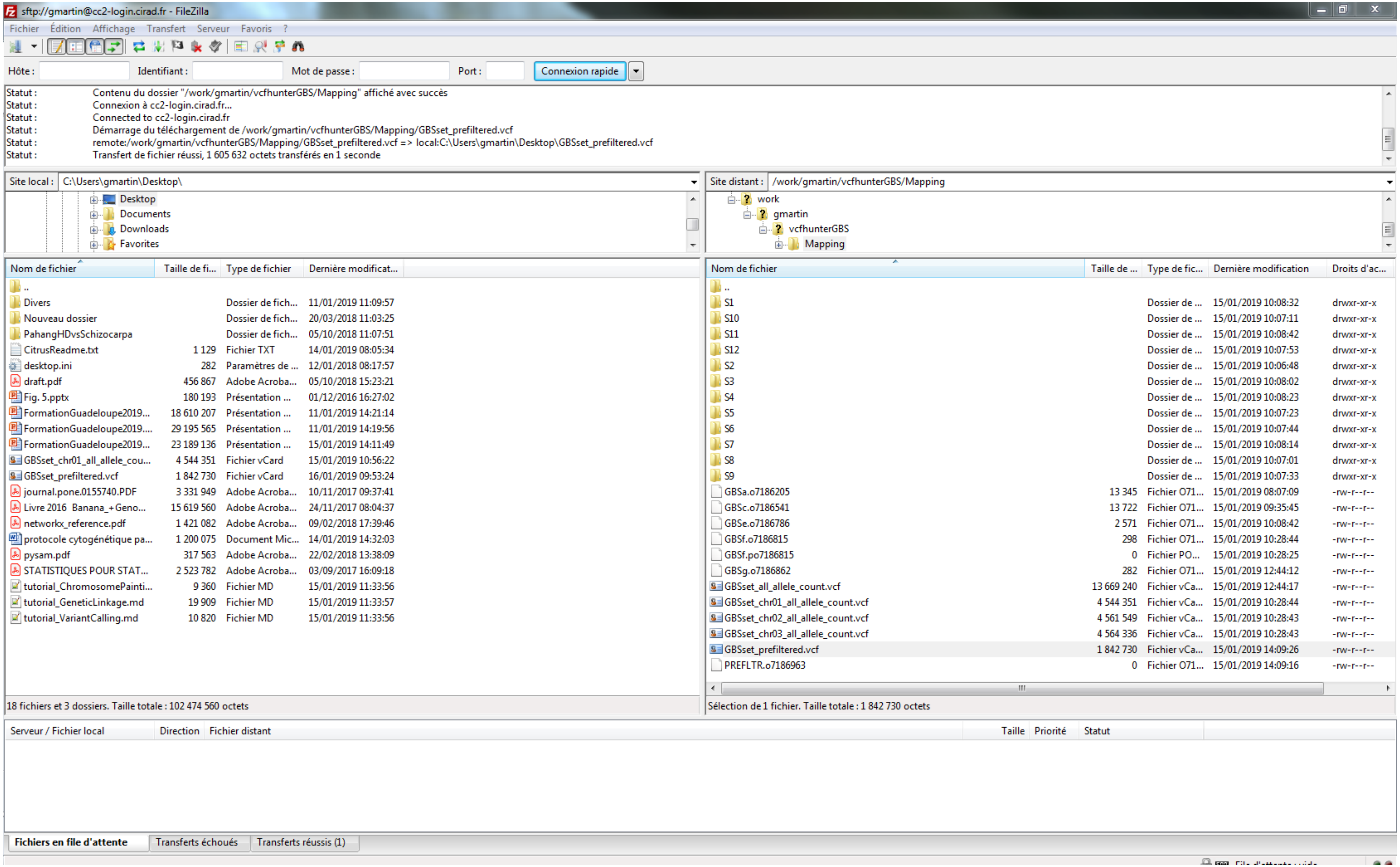
```
[gmartin@cc2-login Mapping]$ ll
total 28544
-rw-r--r-- 1 gmartin users    13345 Jan 15 08:07 GBSa.o7186205
-rw-r--r-- 1 gmartin users    13722 Jan 15 09:35 GBSc.o7186541
-rw-r--r-- 1 gmartin users     2571 Jan 15 10:08 GBSe.o7186786
-rw-r--r-- 1 gmartin users      298 Jan 15 10:28 GBSf.o7186815
-rw-r--r-- 1 gmartin users        0 Jan 15 10:28 GBSf.po7186815
-rw-r--r-- 1 gmartin users      282 Jan 15 12:44 GBSg.o7186862
-rw-r--r-- 1 gmartin users 13669240 Jan 15 12:44 GBSset_all_allele_count.vcf
-rw-r--r-- 1 gmartin users  4544351 Jan 15 10:28 GBSset_chr01_all_allele_count.vcf
-rw-r--r-- 1 gmartin users  4561549 Jan 15 10:28 GBSset_chr02_all_allele_count.vcf
-rw-r--r-- 1 gmartin users  4564336 Jan 15 10:28 GBSset_chr03_all_allele_count.vcf
-rw-r--r-- 1 gmartin users  1842730 Jan 15 14:09 GBSset_prefiltered.vcf
-rw-r--r-- 1 gmartin users        0 Jan 15 14:09 PREFLTR.o7186963
drwxr-xr-x 3 gmartin users     1024 Jan 15 10:08 S1
drwxr-xr-x 3 gmartin users     1024 Jan 15 10:07 S10
drwxr-xr-x 3 gmartin users     1024 Jan 15 10:08 S11
drwxr-xr-x 3 gmartin users     1024 Jan 15 10:07 S12
drwxr-xr-x 3 gmartin users     1024 Jan 15 10:06 S2
drwxr-xr-x 3 gmartin users     1024 Jan 15 10:08 S3
drwxr-xr-x 3 gmartin users     1024 Jan 15 10:08 S4
drwxr-xr-x 3 gmartin users     1024 Jan 15 10:07 S5
drwxr-xr-x 3 gmartin users     1024 Jan 15 10:07 S6
drwxr-xr-x 3 gmartin users     1024 Jan 15 10:08 S7
drwxr-xr-x 3 gmartin users     1024 Jan 15 10:07 S8
drwxr-xr-x 3 gmartin users     1024 Jan 15 10:07 S9
[gmartin@cc2-login Mapping]$
```

A vcf file prefiltered

PREFLTR.oxxxxxxx file containing what VcfPreFilter.1.0.py told us while it was executing

From the output of GBS sequencing to the variant calling file: in command line

- Download this file with filezilla:



sftp://gmartin@cc2-login.cirad.fr - FileZilla

Hôte : Identifiant : Mot de passe : Port : Connexion rapide

```
Statut :	Contenu du dossier "/work/gmartin/vcfhunterGBS/Mapping" affiché avec succès
Statut :	Connexion à cc2-login.cirad.fr...
Statut :	Connected to cc2-login.cirad.fr
Statut :	Démarrage du téléchargement de /work/gmartin/vcfhunterGBS/Mapping/GBSset_prefiltered.vcf
Statut :	remote:/work/gmartin/vcfhunterGBS/Mapping/GBSset_prefiltered.vcf => local:C:\Users\gmartin\Desktop\GBSset_prefiltered.vcf
Statut :	Transfert de fichier réussi, 1 605 632 octets transférés en 1 seconde
```

Site local : C:\Users\gmartin\Desktop\

- Desktop
- Documents
- Downloads
- Favorites

Nom de fichier	Taille de fi...	Type de fichier	Dernière modificat...
..			
Divers		Dossier de fich...	11/01/2019 11:09:57
Nouveau dossier		Dossier de fich...	20/03/2018 11:03:25
PahangHDvsSchizocarpa		Dossier de fich...	05/10/2018 11:07:51
CitrusReadme.txt	1 129	Fichier TXT	14/01/2019 08:05:34
desktop.ini	282	Paramètres de ...	12/01/2018 08:17:57
draft.pdf	456 867	Adobe Acroba...	05/10/2018 15:23:21
Fig. 5.pptx	180 193	Présentation ...	01/12/2016 16:27:02
FormationGuadeloupe2019...	18 610 207	Présentation ...	11/01/2019 14:21:14
FormationGuadeloupe2019....	29 195 565	Présentation ...	11/01/2019 14:19:56
FormationGuadeloupe2019...	23 189 136	Présentation ...	15/01/2019 14:11:49
GBSset_chr01_all_allele_cou...	4 544 351	Fichier vCard	15/01/2019 10:56:22
GBSset_prefiltered.vcf	1 842 730	Fichier vCard	16/01/2019 09:53:24
journal.pone.0155740.PDF	3 331 949	Adobe Acroba...	10/11/2017 09:37:41
Livre 2016 Banana_+Geno...	15 619 560	Adobe Acroba...	24/11/2017 08:04:37
networkx_reference.pdf	1 421 082	Adobe Acroba...	09/02/2018 17:39:46
protocole cytogénétique pa...	1 200 075	Document Mic...	14/01/2019 14:32:03
pysam.pdf	317 563	Adobe Acroba...	22/02/2018 13:38:09
STATISTIQUES POUR STAT...	2 523 782	Adobe Acroba...	03/09/2017 16:09:18
tutorial_ChromosomePainti...	9 360	Fichier MD	15/01/2019 11:33:56
tutorial_GeneticLinkage.md	19 909	Fichier MD	15/01/2019 11:33:57
tutorial_VariantCalling.md	10 820	Fichier MD	15/01/2019 11:33:56

18 fichiers et 3 dossiers. Taille totale : 102 474 560 octets

Site distant : /work/gmartin/vcfhunterGBS/Mapping

- work
  - gmartin
    - vcfhunterGBS
      - Mapping

Nom de fichier	Taille de ...	Type de fic...	Dernière modification	Droits d'ac...
..				
S1		Dossier de ...	15/01/2019 10:08:32	drwxr-xr-x
S10		Dossier de ...	15/01/2019 10:07:11	drwxr-xr-x
S11		Dossier de ...	15/01/2019 10:08:42	drwxr-xr-x
S12		Dossier de ...	15/01/2019 10:07:53	drwxr-xr-x
S2		Dossier de ...	15/01/2019 10:06:48	drwxr-xr-x
S3		Dossier de ...	15/01/2019 10:08:02	drwxr-xr-x
S4		Dossier de ...	15/01/2019 10:08:23	drwxr-xr-x
S5		Dossier de ...	15/01/2019 10:07:23	drwxr-xr-x
S6		Dossier de ...	15/01/2019 10:07:44	drwxr-xr-x
S7		Dossier de ...	15/01/2019 10:08:14	drwxr-xr-x
S8		Dossier de ...	15/01/2019 10:07:01	drwxr-xr-x
S9		Dossier de ...	15/01/2019 10:07:33	drwxr-xr-x
GBSa.o7186205	13 345	Fichier O71...	15/01/2019 08:07:09	-rw-r--r--
GBSc.o7186541	13 722	Fichier O71...	15/01/2019 09:35:45	-rw-r--r--
GBSe.o7186786	2 571	Fichier O71...	15/01/2019 10:08:42	-rw-r--r--
GBSf.o7186815	298	Fichier O71...	15/01/2019 10:28:44	-rw-r--r--
GBSf.po7186815	0	Fichier PO...	15/01/2019 10:28:25	-rw-r--r--
GBSg.o7186862	282	Fichier O71...	15/01/2019 12:44:12	-rw-r--r--
GBSset_all_allele_count.vcf	13 669 240	Fichier vCa...	15/01/2019 12:44:17	-rw-r--r--
GBSset_chr01_all_allele_count.vcf	4 544 351	Fichier vCa...	15/01/2019 10:28:44	-rw-r--r--
GBSset_chr02_all_allele_count.vcf	4 561 549	Fichier vCa...	15/01/2019 10:28:43	-rw-r--r--
GBSset_chr03_all_allele_count.vcf	4 564 336	Fichier vCa...	15/01/2019 10:28:43	-rw-r--r--
GBSset_prefiltered.vcf	1 842 730	Fichier vCa...	15/01/2019 14:09:26	-rw-r--r--
PREFLTR.o7186963	0	Fichier O71...	15/01/2019 14:09:16	-rw-r--r--

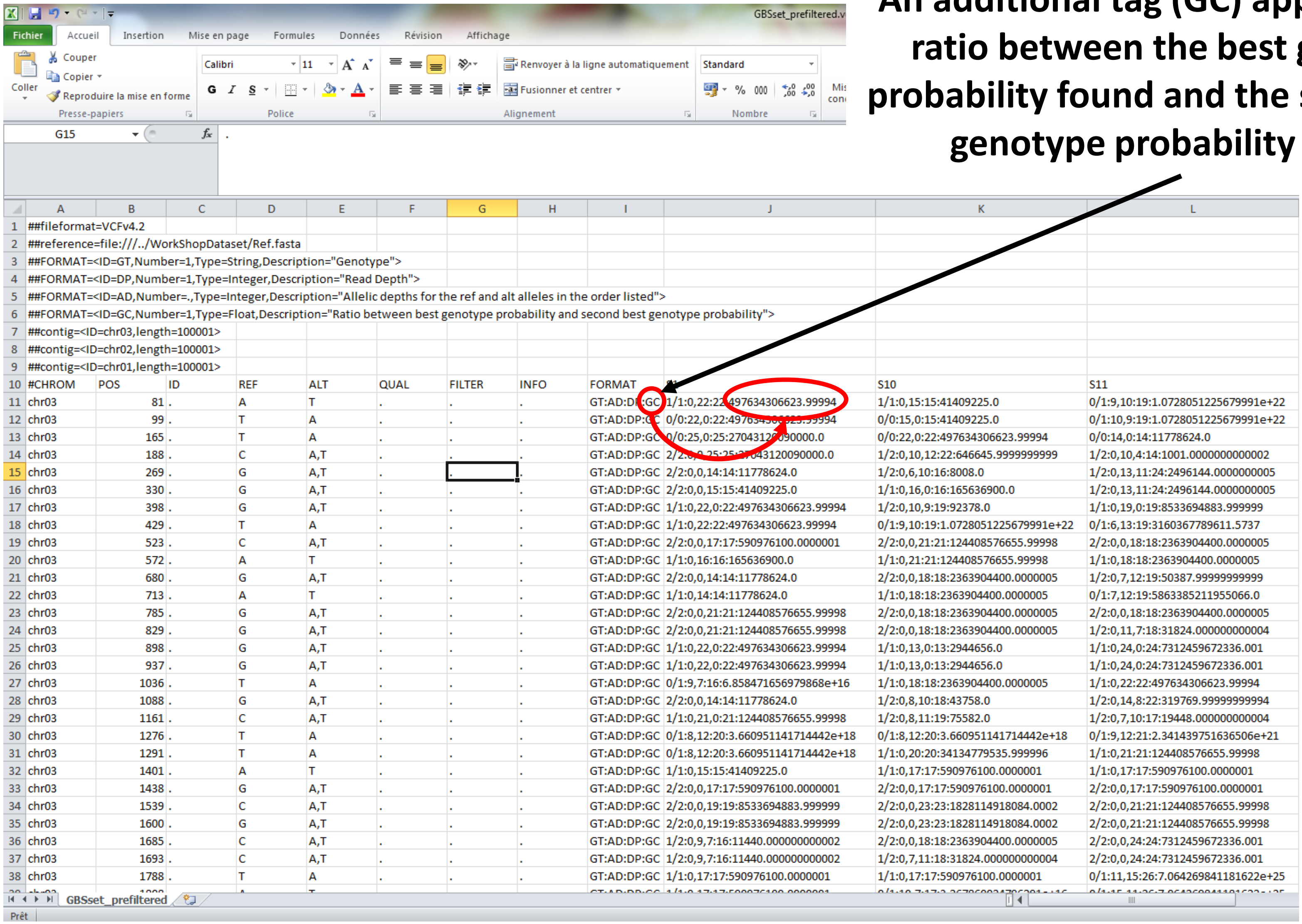
Sélection de 1 fichier. Taille totale : 1 842 730 octets

Serveur / Fichier local	Direction	Fichier distant	Taille	Priorité	Statut

From the output of GBS sequencing to the variant calling file: in command line

- Open the vcf with excel: (less missing data)

An additional tag (GC) appeared: the ratio between the best genotype probability found and the second best genotype probability found.



	A	B	C	D	E	F	G	H	I	J	K	L
1	##fileformat=VCFv4.2											
2	##reference=file:///../WorkShopDataset/Ref.fasta											
3	##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">											
4	##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">											
5	##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">											
6	##FORMAT=<ID=GC,Number=1,Type=Float,Description="Ratio between best genotype probability and second best genotype probability">											
7	##contig=<ID=chr03,length=100001>											
8	##contig=<ID=chr02,length=100001>											
9	##contig=<ID=chr01,length=100001>											
10	#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	S1	S10	S11
11	chr03	81	.	A	T	.	.	.	GT:AD:DP:GC	1/1:0,22:22:497634306623.99994	1/1:0,15:15:41409225.0	0/1:9,10:19:1.0728051225679991e+22
12	chr03	99	.	T	A	.	.	.	GT:AD:DP:GC	0/0:22,0:22:497634306623.99994	0/0:15,0:15:41409225.0	0/1:10,9:19:1.0728051225679991e+22
13	chr03	165	.	T	A	.	.	.	GT:AD:DP:GC	0/0:25,0:25:27043120090000.0	0/0:22,0:22:497634306623.99994	0/0:14,0:14:11778624.0
14	chr03	188	.	C	A,T	.	.	.	GT:AD:DP:GC	2/2:0,0,25:25:27043120090000.0	1/2:0,10,12:22:646645.9999999999	1/2:0,10,4:14:1001.0000000000002
15	chr03	269	.	G	A,T	.	.	.	GT:AD:DP:GC	2/2:0,0,14:14:11778624.0	1/2:0,6,10:16:8008.0	1/2:0,13,11:24:2496144.0000000005
16	chr03	330	.	G	A,T	.	.	.	GT:AD:DP:GC	2/2:0,0,15:15:41409225.0	1/1:0,16,0:16:165636900.0	1/2:0,13,11:24:2496144.0000000005
17	chr03	398	.	G	A,T	.	.	.	GT:AD:DP:GC	1/1:0,22,0:22:497634306623.99994	1/2:0,10,9:19:92378.0	1/1:0,19,0:19:8533694883.999999
18	chr03	429	.	T	A	.	.	.	GT:AD:DP:GC	1/1:0,22:22:497634306623.99994	0/1:9,10:19:1.0728051225679991e+22	0/1:6,13:19:3160367789611.5737
19	chr03	523	.	C	A,T	.	.	.	GT:AD:DP:GC	2/2:0,0,17:17:590976100.0000001	2/2:0,0,21:21:124408576655.99998	2/2:0,0,18:18:2363904400.0000005
20	chr03	572	.	A	T	.	.	.	GT:AD:DP:GC	1/1:0,16:16:165636900.0	1/1:0,21:21:124408576655.99998	1/1:0,18:18:2363904400.0000005
21	chr03	680	.	G	A,T	.	.	.	GT:AD:DP:GC	2/2:0,0,14:14:11778624.0	2/2:0,0,18:18:2363904400.0000005	1/2:0,7,12:19:50387.99999999999
22	chr03	713	.	A	T	.	.	.	GT:AD:DP:GC	1/1:0,14:14:11778624.0	1/1:0,18:18:2363904400.0000005	0/1:7,12:19:5863385211955066.0
23	chr03	785	.	G	A,T	.	.	.	GT:AD:DP:GC	2/2:0,0,21:21:124408576655.99998	2/2:0,0,18:18:2363904400.0000005	2/2:0,0,18:18:2363904400.0000005
24	chr03	829	.	G	A,T	.	.	.	GT:AD:DP:GC	2/2:0,0,21:21:124408576655.99998	2/2:0,0,18:18:2363904400.0000005	1/2:0,11,7:18:31824.000000000004
25	chr03	898	.	G	A,T	.	.	.	GT:AD:DP:GC	1/1:0,22,0:22:497634306623.99994	1/1:0,13,0:13:2944656.0	1/1:0,24,0:24:7312459672336.001
26	chr03	937	.	G	A,T	.	.	.	GT:AD:DP:GC	1/1:0,22,0:22:497634306623.99994	1/1:0,13,0:13:2944656.0	1/1:0,24,0:24:7312459672336.001
27	chr03	1036	.	T	A	.	.	.	GT:AD:DP:GC	0/1:9,7:16:6.858471656979868e+16	1/1:0,18:18:2363904400.0000005	1/1:0,22:22:497634306623.99994
28	chr03	1088	.	G	A,T	.	.	.	GT:AD:DP:GC	2/2:0,0,14:14:11778624.0	1/2:0,8,10:18:43758.0	1/2:0,14,8:22:319769.99999999994
29	chr03	1161	.	C	A,T	.	.	.	GT:AD:DP:GC	1/1:0,21,0:21:124408576655.99998	1/2:0,8,11:19:75582.0	1/2:0,7,10:17:19448.000000000004
30	chr03	1276	.	T	A	.	.	.	GT:AD:DP:GC	0/1:8,12:20:3.660951141714442e+18	0/1:8,12:20:3.660951141714442e+18	0/1:9,12:21:2.341439751636506e+21
31	chr03	1291	.	T	A	.	.	.	GT:AD:DP:GC	0/1:8,12:20:3.660951141714442e+18	1/1:0,20:20:34134779535.999996	1/1:0,21:21:124408576655.99998
32	chr03	1401	.	A	T	.	.	.	GT:AD:DP:GC	1/1:0,15:15:41409225.0	1/1:0,17:17:590976100.0000001	1/1:0,17:17:590976100.0000001
33	chr03	1438	.	G	A,T	.	.	.	GT:AD:DP:GC	2/2:0,0,17:17:590976100.0000001	2/2:0,0,17:17:590976100.0000001	2/2:0,0,17:17:590976100.0000001
34	chr03	1539	.	C	A,T	.	.	.	GT:AD:DP:GC	2/2:0,0,19:19:8533694883.999999	2/2:0,0,23:23:1828114918084.0002	2/2:0,0,21:21:124408576655.99998
35	chr03	1600	.	G	A,T	.	.	.	GT:AD:DP:GC	2/2:0,0,19:19:8533694883.999999	2/2:0,0,23:23:1828114918084.0002	2/2:0,0,21:21:124408576655.99998
36	chr03	1685	.	C	A,T	.	.	.	GT:AD:DP:GC	1/2:0,9,7:16:11440.000000000002	2/2:0,0,18:18:2363904400.0000005	2/2:0,0,24:24:7312459672336.001
37	chr03	1693	.	C	A,T	.	.	.	GT:AD:DP:GC	1/2:0,9,7:16:11440.000000000002	1/2:0,7,11:18:31824.000000000004	2/2:0,0,24:24:7312459672336.001
38	chr03	1788	.	T	A	.	.	.	GT:AD:DP:GC	1/1:0,17:17:590976100.0000001	1/1:0,17:17:590976100.0000001	0/1:11,15:26:7.064269841181622e+25

From the output of GBS sequencing to the variant calling file: in command line

- This prefiltering step was designed to discriminate between variant lines resulting from sequencing errors and true variant line.

- However, one can want to apply additional filters such as reporting only diallelic polymorphous SNP, minimal coverage confidence to call a variant, missing data proportion, etc…

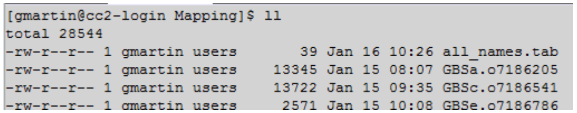
- For that we first need to generate a file containing a list of accessions we want to apply filter on. If we want to apply this filter on all accessions of the vcf, this file can be generated by a “simple” command line that will work on any vcf files you have!

```
head -n 10000 GBSset_prefiltered.vcf | grep "#CHROM" | sed 's/\t/\n/g' |
tail -n +10 > all_names.tab
```

- We take the first 10000 lines of the vcf: `head -n 10000 GBSset_prefiltered.vcf`
- Of these 10000 lines, we get the line with the accessions names which also contained “#CHROM”: `grep "#CHROM"`
- Of this line we convert tabulation into carriage return: `sed 's/\t/\n/g'`
- And we take all lines from the result, but only from line number 10 to the end: `tail -n +10`
- The selected lines are written to a file named: `all_names.tab`

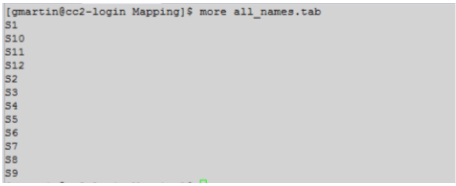
From the output of GBS sequencing to the variant calling file: in command line

- Once the name file as been created: this can be verified with `ll` command:



```
[gmartin@cc2-login Mapping]$ ll
total 28544
-rw-r--r-- 1 gmartin users       39 Jan 16 10:26 all_names.tab
-rw-r--r-- 1 gmartin users    13345 Jan 15 08:07 GBSa.o7186205
-rw-r--r-- 1 gmartin users    13722 Jan 15 09:35 GBSc.o7186541
-rw-r--r-- 1 gmartin users     2571 Jan 15 10:08 GBSe.o7186786
```

- This file contained accession names: `more all_names.tab`



```
[gmartin@cc2-login Mapping]$ more all_names.tab
S1
S10
S11
S12
S2
S3
S4
S5
S6
S7
S8
S9
```

- A third script, called vcfFilter.1.0.py as been designed to filter the vcf (GBSset_prefiltered.vcf). For example, we may want to:

(1) convert to missing data:

- ✓ all datapoints which are not supported by at least **15** reads (no sufficient coverage to call good genotype)
- ✓ all datapoints which are not supported by more than **300** reads (probably repeat sequences)
- ✓ all datapoints for which each alleles is not supported by **3** read and a minimal read proportion of 0.2

(2) remove all line which contained missing data,

(3) remove mono, tri and tetra allelic sites,

(4) write the output in a file which prefix is **GBSset_Filtered.**

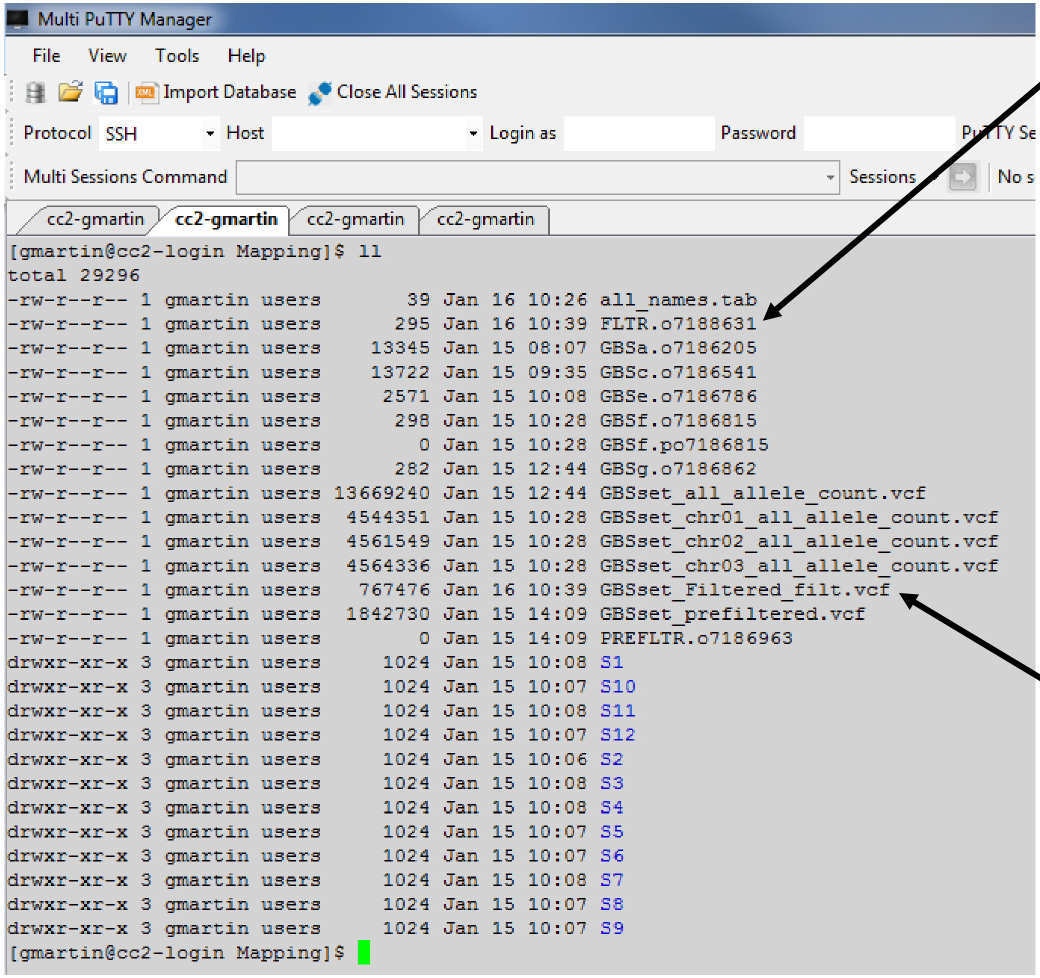
To apply these filters do not try the command following command line:

```
qsub -q normal.q -b yes -V -N FLTR "vcfFilter.1.0.py --vcf
GBSset_prefiltered.vcf --names all_names.tab --MinCov 15 --MaxCov 300 --
MinAl 3 --MinFreq 0.2 --nMiss 0 --RmAlAlt 1:3:4 --prefix GBSset_Filtered"
```

From the output of GBS sequencing to the variant calling file: in command line

- Listing the files generated:

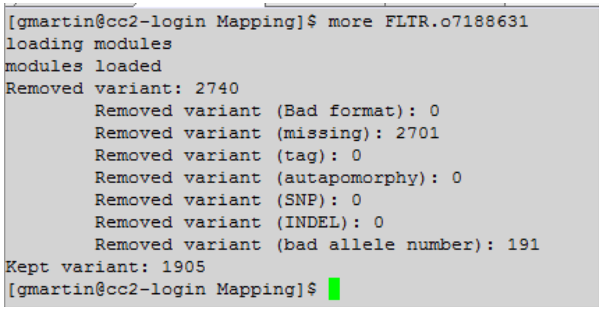
  `ll`



```
[gmartin@cc2-login Mapping]$ ll
total 29296
-rw-r--r-- 1 gmartin users       39 Jan 16 10:26 all_names.tab
-rw-r--r-- 1 gmartin users      295 Jan 16 10:39 FLTR.o7188631
-rw-r--r-- 1 gmartin users    13345 Jan 15 08:07 GBSa.o7186205
-rw-r--r-- 1 gmartin users    13722 Jan 15 09:35 GBSc.o7186541
-rw-r--r-- 1 gmartin users     2571 Jan 15 10:08 GBSe.o7186786
-rw-r--r-- 1 gmartin users      298 Jan 15 10:28 GBSf.o7186815
-rw-r--r-- 1 gmartin users        0 Jan 15 10:28 GBSf.po7186815
-rw-r--r-- 1 gmartin users      282 Jan 15 12:44 GBSg.o7186862
-rw-r--r-- 1 gmartin users 13669240 Jan 15 12:44 GBSset_all_allele_count.vcf
-rw-r--r-- 1 gmartin users  4544351 Jan 15 10:28 GBSset_chr01_all_allele_count.vcf
-rw-r--r-- 1 gmartin users  4561549 Jan 15 10:28 GBSset_chr02_all_allele_count.vcf
-rw-r--r-- 1 gmartin users  4564336 Jan 15 10:28 GBSset_chr03_all_allele_count.vcf
-rw-r--r-- 1 gmartin users   767476 Jan 16 10:39 GBSset_Filtered_filt.vcf
-rw-r--r-- 1 gmartin users  1842730 Jan 15 14:09 GBSset_prefiltered.vcf
-rw-r--r-- 1 gmartin users        0 Jan 15 14:09 PREFLTR.o7186963
drwxr-xr-x 3 gmartin users     1024 Jan 15 10:08 S1
drwxr-xr-x 3 gmartin users     1024 Jan 15 10:07 S10
drwxr-xr-x 3 gmartin users     1024 Jan 15 10:08 S11
drwxr-xr-x 3 gmartin users     1024 Jan 15 10:07 S12
drwxr-xr-x 3 gmartin users     1024 Jan 15 10:06 S2
drwxr-xr-x 3 gmartin users     1024 Jan 15 10:08 S3
drwxr-xr-x 3 gmartin users     1024 Jan 15 10:08 S4
drwxr-xr-x 3 gmartin users     1024 Jan 15 10:07 S5
drwxr-xr-x 3 gmartin users     1024 Jan 15 10:07 S6
drwxr-xr-x 3 gmartin users     1024 Jan 15 10:08 S7
drwxr-xr-x 3 gmartin users     1024 Jan 15 10:07 S8
drwxr-xr-x 3 gmartin users     1024 Jan 15 10:07 S9
[gmartin@cc2-login Mapping]$
```

FLTR.oxxxxxxx file containing what vcfFilter.1.0.py told us while it was executing

`more FLTR.oxxxxxxx`



```
[gmartin@cc2-login Mapping]$ more FLTR.o7188631
loading modules
modules loaded
Removed variant: 2740
        Removed variant (Bad format): 0
        Removed variant (missing): 2701
        Removed variant (tag): 0
        Removed variant (autapomorphy): 0
        Removed variant (SNP): 0
        Removed variant (INDEL): 0
        Removed variant (bad allele number): 191
Kept variant: 1905
[gmartin@cc2-login Mapping]$
```

The filtered vcf file

From the output of GBS sequencing to the variant calling file: in command line

- A tutorial for variant calling of WGS data is also available here:

  https://github.com/SouthGreenPlatform/VcfHunter/blob/master/tutorial_VariantCalling.md

- Vcfhunter module contained additional tools for genetic mapping analysis and chromosome painting described and available here:

  https://github.com/SouthGreenPlatform/VcfHunter