Physalia

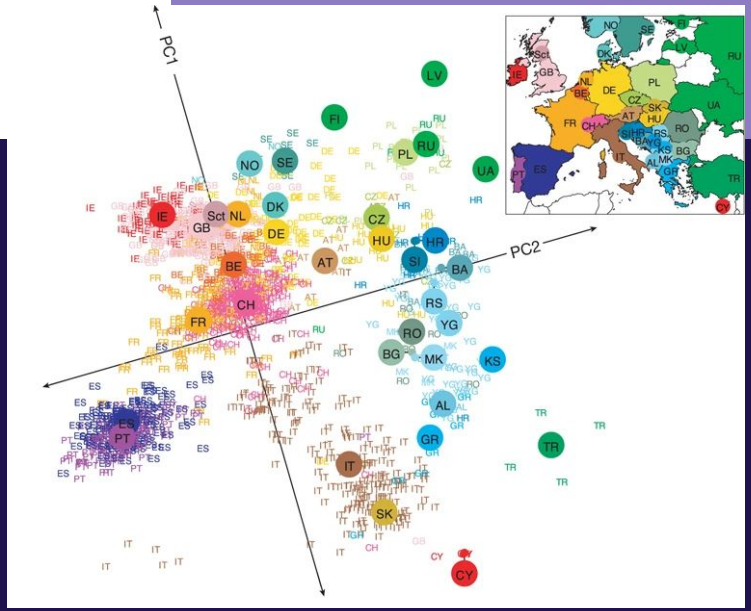
Courses

Introduction to population genomics

25-29/11/2024

Physalia course

Thibault Leroy, Yann Bourgeois



Thibault LEROY
Habilitated, Dr.
Permanent researcher at INRAE
Toulouse, France

Main biological model:
Bees, especially honey bees (*since 2023*)

Previous main biological models:
- *Venturia*, a main fungal pathogen responsible for apple scab disease (PhD, Angers, France)
- Oaks (Postdoc 1, Bordeaux, France)
- Passerine birds (Postdoc 2, Montpellier, France)
- *Populus*/*Tillandsia* (University assistant, Vienna, Austria)
- Roses (Postdoc 3, Angers, France)
- Wheat (Postdoc 4, Clermont-Ferrand, France)

Side projects:
Coho salmon (with Q. Rougemont & L. Bernatchez)
Tropical trees from French Guiana (with S. Schmitt, N. Tysklind, M. Heuertz)
…

Main interests:

- Local adaptation to changing environments

- Gene flow, (adaptive) introgression

- Speciation / Hybrid zones

- Mutation rates and spectra

- Footprints of natural and artificial selection

- Evolution of genomic variation within and between species

- Deleterious mutations /conservation Biology

- Methods in population genomics (demographic inferences, genome scans/GWAS, …)

- Metagenomics

Interests:
'Molecular ecology'
Conservation genomics
Phylodynamics
Transposable elements dynamics

Université de Toulouse

UNI BASEL

جامعة نيويورك ابوظبي
NYU ABU DHABI

UNIVERSITY OF PORTSMOUTH

IRD
Institut de Recherche pour le Développement
FRANCE

2009 2013 2016 2020 2022

Motivations for this Physalia course:

- Introduce population genomics methods
- Adapted to beginners with the idea of following a learning-by-doing strategy

- A lecture of a maximum of 2-hour per day
- Accessible for diverse levels, allowing you to grasp the essentials or explore further

- Practical: autonomous (PDF) + instructors providing support on Slack
- 10-15 minutes discussion all together (the day after?)

- This physalia course has the objective to introduce the topics: you are invited to contact us to ask additional questions after the course, when you needed for your research

Results of the little survey (18 participants)

Your biological models:

- Animals (13/18), including invertebrates (7/18)
- Plants (4/18)
- Protists (1/18)

Your main expectation(s) for this course:

- demographic inferences (10/18)
- population structure (9/18)
- selection (9/18)
- GWAS (6/18)
- basic bioinformatics (5/18)
- landscape genomics (5/18)
- (+ GO term enrichment 1/18)
- (+ methodological choice for sequencing, poolseq vs. individual sequencing etc 1/18)
- (+ summary statistics in population genetics 1/18)

Results of the little survey (18 participants)

More time spent on the introductory courses or on the practical sessions?

- practicals (9/16)
- courses (4/16)
- well balanced (3/16)

Your prior knowledge in population genomics / analyses?

- “Quite experienced” (8/18)
- “None” (7/18)
- “Relatively limited” (3/18)

Prior knowledge in computing and data analysis?

- Experienced (9/18)
- Intermediate (7/18)
- Beginner (2/18)

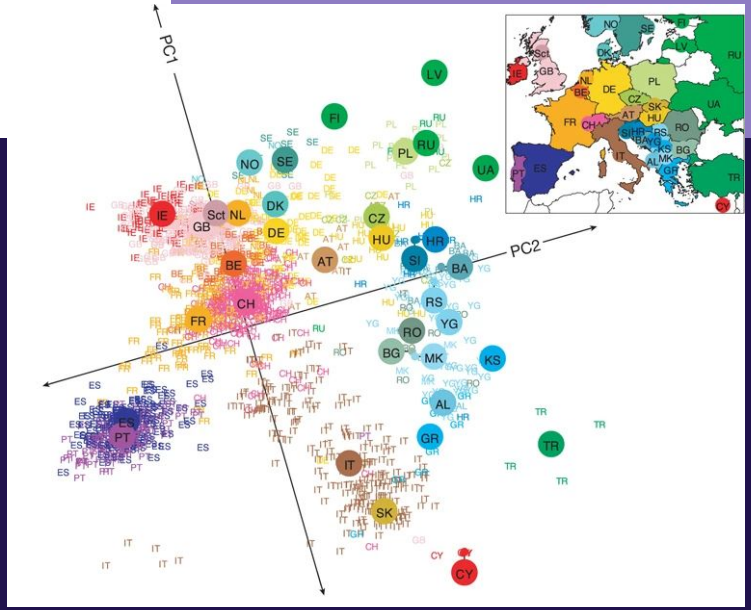
Access to/use of computing clusters ?

- Yes, regularly (12/18)
- Yes, rarely (4/18)
- No (2/18)

Physalia

Courses

Basic bioinformatics



25/11/2024

Physalia course

Thibault Leroy, Yann Bourgeois

Goals for today's lecture

- Pros and cons of using different strategies regarding the sequencing
- Understand the rationale of the analysis
- Describe the common and specific parts of the pipelines depending on the strategies
- Focus on some recent advances in genomics to anticipate future developments (pangenomics, …)

Imagine that you work on the evolution/adaptation of unicorns…

Population 1: unicorns

Population 2: “uni**no**corns”

Imagine that you work on the evolution/adaptation of unicorns…

Population 1: unicorns

Population 2: “uni**no**corns”

<u>You may be interested by many questions:</u>

Are populations of unicorns and uninocorns consistent with a single global population or do they exhibit population structure (*e.g.* due to non-random mating)?

What are the evolutionary history of unicorns /uninocorns?

Are corns adaptive ? deleterious ?

How these populations vary through space?

If you have such questions in mind, you are at the right Physalia course !

Imagine that you work on the evolution/adaptation of unicorns…

Population 1: unicorns

Population 2: “uni**no**corns”

Which sequencing strategy can you use to answer these different questions?

Imagine that you work on the evolution/adaptation of unicorns…

Population 1: unicorns

Population 2: “uni**no**corns”

Which sequencing strategy can you use to answer these different questions?

Individual sequencing, moderate coverage (20-50X / individual)

Individual sequencing, low coverage (< 10X / individual)

Pooling of individual and sequencing (poolseq), moderate to high coverage (>50X)

Short reads (*e.g.* Illumina)
Long reads (*e.g.* PacBio/Nanopore)

Pros and cons of each strategy

Strategy	Individual / moderate coverage	Individual / low coverage	Pool-seq / moderate coverage
Accuracy (individual level)	+++ (Genotype calls)	+ (Genotype likelihood)	- - - (No information at the individual level)
Accuracy (population level) *e.g.* allele frequencies	+++	++ to +++	+ to +++ (depending on the number of individuals <-> pipetting biases)
Affordability Potential sequencing costs (prize are indicative and assume access to an internal facility, *e.g.* in my lab in Toulouse)	+ library 30€/sample (60*30€=1800 €) + Illumina 30X seq eq NovaSeq (1800X*, 5000€) **~7000€**	++ library 30€/sample (60*30€=1800 €) + Illumina seq eq NovaSeq (600X*, 1700€) **~3500€**	+++ libraries 1 per pool (2*60 €=120€) + Illumina seq (eq NovaSeq, 100X*, 350 €) **<500€**

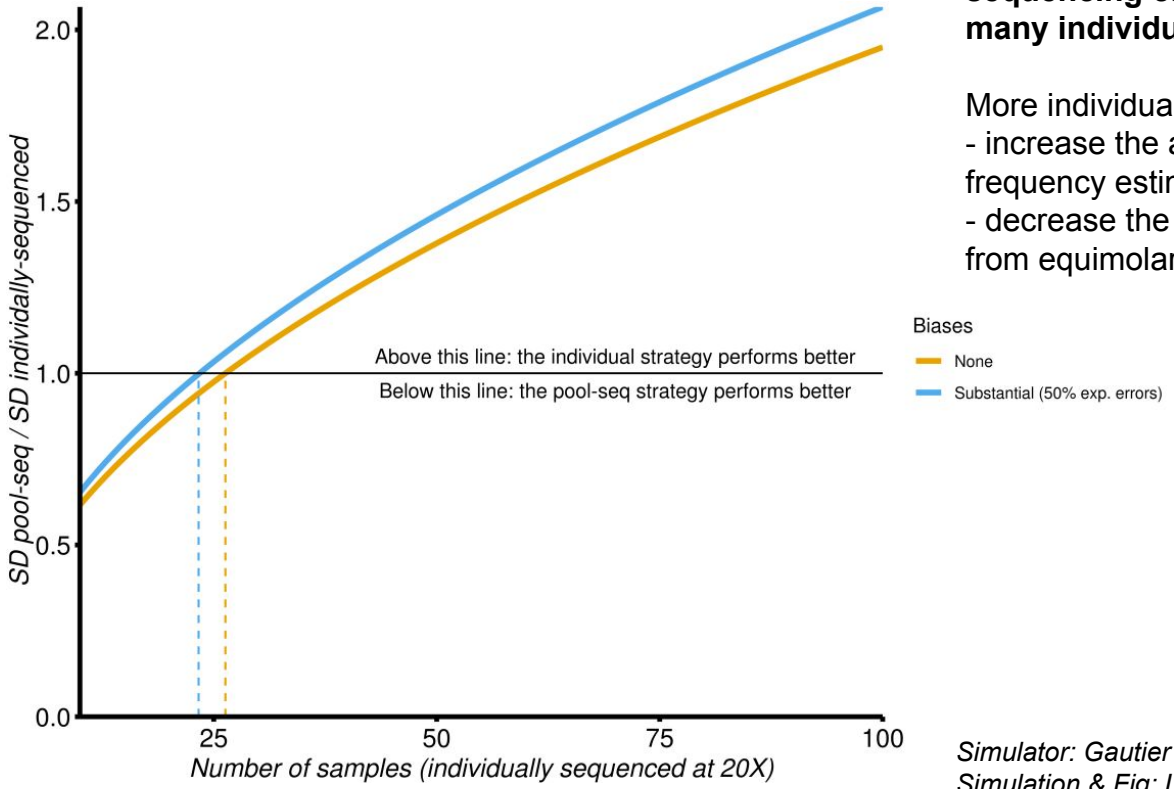
** assuming that unicorns/uninocorns have a genome size of ~250 Mb*

Poolseq can be an interesting strategy for population-level investigation

A pool-seq strategy of 50 individuals sequenced at a mean pool coverage of 100X

vs.

an individual-based genotyping strategy with a growing number of individually sequenced at 20X

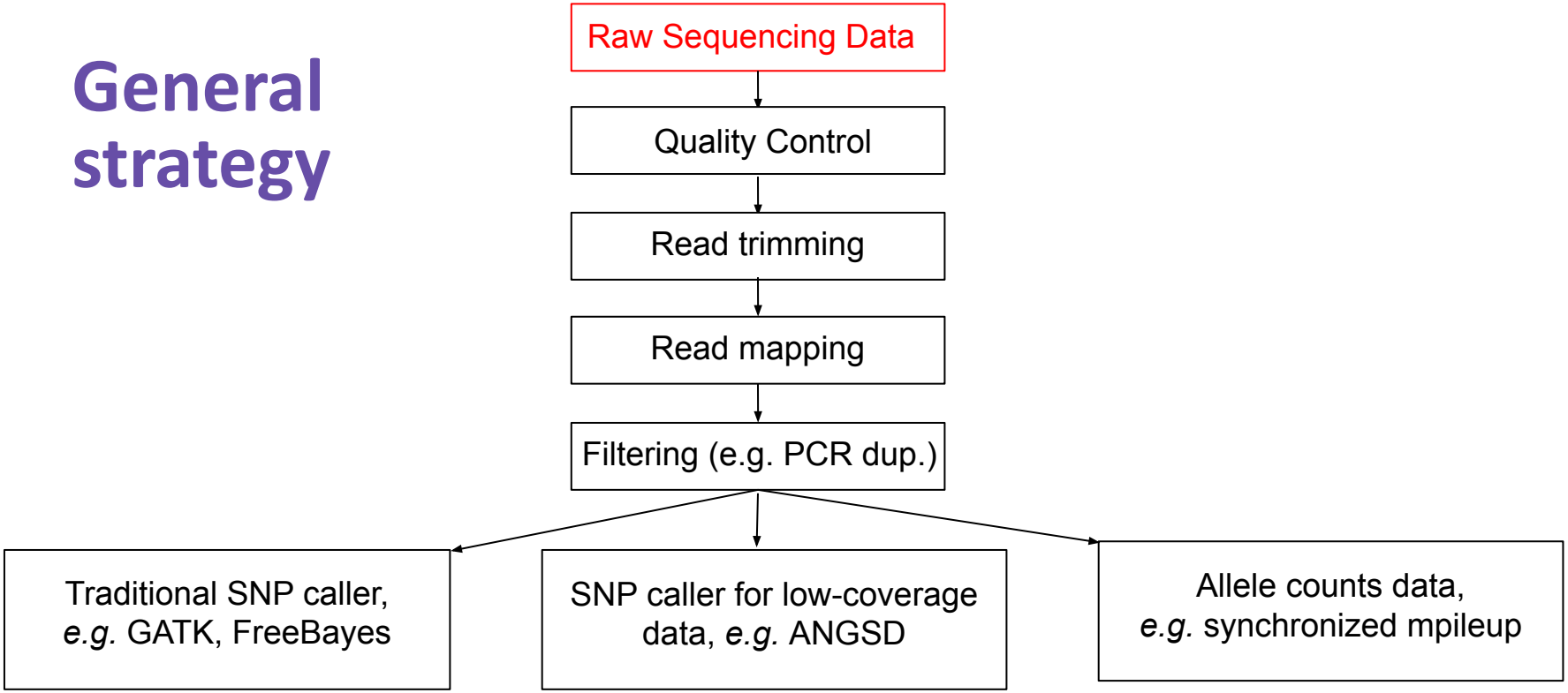


Pool-seq is interesting if sequencing can be performed on many individuals per population!

More individuals included in pools:
- increase the accuracy of allele frequency estimation
- decrease the impacts of deviation from equimolarity (*e.g.* pipetting bias)

Simulator: Gautier et al. 2013 Mol Ecol
Simulation & Fig: Leroy & Rougemont 2020

General strategy



VCF = Variant Calling Format

```
Chr1   6   G   A […] GT:DP  0/0:12  0/1:16  0/0:14 […]
Chr1   11  C   G […] GT:DP  0/0:14  1/1:19  0/1:15 […]
…
```

Genotype likelihood (GL)

```
Chr1  42  -2.0,-0.1,-3.0  -1.5,-1.2,-0.2  -0.5,-1.0,-3.5 […]
Chr1  78   1.0,-2.2,-0.3  -0.3,-1.5,-1.8  -3.2,-0.5,-0.7 […]
…
```

Allele counts (popoolation: A:T:C:G:N:*)<

```
Chr1  1  T  0:42:0:0:0:0  0:50:0:0:0:0
Chr1  2  G  7:0:0:35:0:0  0:0:0:50:0:0
Chr1  3  A  42:0:0:0:0:0  50:0:0:0:0:0
Chr1  4  C  0:0:20:22:0:0  0:0:26:24:0:0
….
```

Raw Sequencing Data

fasta vs. fastq

Fasta

> Sequence1
GATGCGGAATGAACTGGGATTCATAACTGCCCCCTGTTAACATTTCGTAAAAGTTGGTCATAAAAC

Fastq

@M07406:112:000000000-L7TK9:1:1101:7541:4763 1:N:0:TTATAACC+GATATCGA
GATGCGGAATGAACTGGGATTCATAACTGCCCCCTGTTAACATTTCGTAAAAGTTGGTCATAAAAC
+
BBBBBBBBBBFFGGFGGGGGGGGHHHHHHHHHHHGGGGGHHHHHHHHHHGGGHHHHFGHHFHHHH

Precise header @SEQ_ID:RUN_ID:FLOWCELL_ID:LANE:SAMPLE:READ:INDEX = High traceability !
Here: @SEQ_ID:RUN_ID:FLOWCELL_ID:LANE:TILE:X:Y:READ:FILTER_FLAG:CTRL_NB:INDEX

The most important information are on lines 2 and 4!
-> The 4th line corresponds to the quality values for the corresponding bases in 2nd line, in the exact same order !

fastq: rationale

Raw Sequencing Data

Fastq

@M07406:112:000000000-L7TK9:1:1101:7541:4763 1:N:0:TTATAACC+GATATCGA

GATGCGGAATGAACTGGGATTCATAACTGCCCCCTGTTAACATTTCGTAAAAGTTGGTCATAAAAC

+

BBBBBBBBBBFFGGFGGGGGGGGHHHHHHHHHHHGGGGGHHHHHHHHHHGGGHHHHFGHHFHHHH

Base quality score (4th line)

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

```
S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 41)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
N - Nanopore      Phred+33,  Duplex reads typically (0, 50)
E - ElemBio AVITI Phred+33,  raw reads typically (0, 55)
P - PacBio        Phred+33,  HiFi reads typically (0, 93)
```

fastq: rationale

Raw Sequencing Data

Fastq

@M07406:112:000000000-L7TK9:1:1101:7541:4763 1:N:0:TTATAACC+GATATCGA
GATGCGGAATGAACTGGGATTCATAACTGCCCCCTGTTAACATTTCGTAAAAGTTGGTCATAAAAC
+
BBBBBBBBBBFFGGFGGGGGGGHHHHHHHHHHGGGGHHHHHHHHHGGGHHHHFGHHFHHHH

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

What does it mean?

Phred Score	Probability of an incorrect base call	Accuracy	Associated character
0	1 in 1	0%	!
10	1 in 10	90%	+
20	1 in 100	99%	5
30	1 in 1000	99.9%	?
40	1 in 10000	99.99%	I

fastq: rationale

Raw Sequencing Data

Fastq

```
@M07406:112:000000000-L7TK9:1:1101:7541:4763 1:N:0:TTATAACC+GATATCGA
GATGCGGAATGAACTGGGATTCATAACTGCCCCCTGTTAACATTTCGTAAAAGTTGGTCATAAAAC
+
BBBBBBBBBBFFGGFGGGGGGGHHHHHHHHHHGGGGHHHHHHHHHGGGHHHHFGHHFHHHH
LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
```

What does it mean?

Associated character	Phred score	Probability of an incorrect base call	Accuracy
B			
F			
G			
H			

fastq: rationale

Raw Sequencing Data

Fastq

@M07406:112:000000000-L7TK9:1:1101:7541:4763 1:N:0:TTATAACC+GATATCGA
GATGCGGAATGAACTGGGATTCATAACTGCCCCCTGTTAACATTTCGTAAAAGTTGGTCATAAAAC
+
BBBBBBBBBBFFGGFGGGGGGGHHHHHHHHHHGGGGHHHHHHHHHHGGGHHHHFGHHFHHHH

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

What does it mean?

Associated character	Phred score	Probability of an incorrect base call	Accuracy
B	33		
F	37		
G	38		
H	39		

fastq: rationale

Raw Sequencing Data

Fastq

@M07406:112:000000000-L7TK9:1:1101:7541:4763 1:N:0:TTATAACC+GATATCGA
GATGCGGAATGAACTGGGATTCATAACTGCCCCCTGTTAACATTTCGTAAAAGTTGGTCATAAAAC
+
BBBBBBBBBBFFGGFGGGGGGGHHHHHHHHHHGGGGGHHHHHHHHHHGGGHHHHFGHHFHHHH

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

What does it mean?

Associated character	Phred score	Probability of an incorrect base call	Accuracy
B	33	0.0005	~99.95%
F	37	0.0002	~99.98%
G	38	0.0002	~99.98%
H	39	0.0001	~99.99%

Raw Sequencing Data

One of the first things we are interested in is:
How many sequences are present in my fastq file?

Fastq.gz files:

There's no need to uncompress the FASTQ files, as most software can directly read `.fastq.gz` files! This helps save space on the computing cluster.

In a command-line Unix terminal:
To read a text file / FASTQ file
more `myfile.fastq`
less `myfile.fastq`

To read a compressed FASTQ file
zmore `myfile.fastq.gz`
zless `myfile.fastq.gz`

Raw Sequencing Data

One of the first things we are interested in is:
How many sequences are present in my fastq file?

Fastq.gz files:

zmore `file.fastq.gz` | grep “@” | wc -l
R1: 696057 (zmore GPS220996_AGTTCAGG-CCAACAGA-L7TK9_L001_R1.fastq.gz | grep "@" | wc -l)
R2: 755300 (zmore GPS220996_AGTTCAGG-CCAACAGA-L7TK9_L001_R2.fastq.gz | grep "@" | wc -l)

zmore `FILE.FASTQ.GZ` | grep “^@” | wc -l
R1: 624811 (zmore GPS220996_AGTTCAGG-CCAACAGA-L7TK9_L001_R1.fastq.gz | grep "^@" | wc -l)
R2: 627027 (zmore GPS220996_AGTTCAGG-CCAACAGA-L7TK9_L001_R2.fastq.gz | grep "^@" | wc -l)

zmore `FILE.FASTQ.GZ` | wc -l -> count the number of lines in a fastq.gz file and then divide this number by 4
zmore `FILE.FASTQ.GZ` | awk '{line++}END{print line/4}' -> single-line command with awk
R1: 624170 (zmore GPS220996_AGTTCAGG-CCAACAGA-L7TK9_L001_R1.fastq.gz | awk '{s++}END{print s/4}')
R2: 624170 (zmore GPS220996_AGTTCAGG-CCAACAGA-L7TK9_L001_R2.fastq.gz | awk '{s++}END{print s/4}')

Raw Sequencing Data

One of the first things we are interested in is:
How many sequences are present in my fastq file?

Fastq.gz files:

zmore `file.fastq.gz` | grep “@” | wc -l
R1: 696057 (zmore GPS220996_AGTTCAGG-CCAACAGA-L7TK9_L00
R2: 755300 (zmore GPS220996_AGTTCAGG-CCAACAGA-L7TK9_L00

zmore `FILE.FASTQ.GZ` | grep “^@” | wc -l
R1: 624811 (zmore GPS220996_AGTTCAGG-CCAACAGA-L7TK9_L00
R2: 627027 (zmore GPS220996_AGTTCAGG-CCAACAGA-L7TK9_L00

zmore `FILE.FASTQ.GZ` | wc -l -> count the number of lines in a fas
zmore `FILE.FASTQ.GZ` | awk '{line++}END{print line/4}'
R1: 624170 (zmore GPS220996_AGTTCAGG-CCAACAGA-L7TK9_L00
R2: 624170 (zmore GPS220996_AGTTCAGG-CCAACAGA-L7TK9_L00

Count the number of “@” using grep

```
@M07406:112:000000000-L7TK9:1:1101:7541:47631:N:0:TTATAACC+GATATCGA
GATGCGGAATGAACTGGGATTCATAACTGCCCCCTGTTAACATTTCGTAAAAGTTGGTCATA
+
A@ABBBBBDDFFGGFGGGGGGGGHHHHHHHHHHGGGGGHHHHHHHHHGGGHHHHFGFEEC@A
```

Count the number of line starting with “@” using grep

```
@M07406:112:000000000-L7TK9:1:1101:7541:4763 1:N:0:TTATAACC+GATATCGA
GATGCGGAATGAACTGGGATTCATAACTGCCCCCTGTTAACATTTCGTAAAAGTTGGTCATA
+
A@ABBBBBDDFFGGFGGGGGGGGHHHHHHHHHHGGGGGHHHHHHHHHGGGHHHHFGFEECCA@
```

Yes, but…

```
@M18478:112:000000000-L7TK9:1:1101:7541:1546 1:N:0:TTATAACC+GATATCGA
TATGCCGAAAAAACTGGGATTCATAACTGCACGCTGCCAACATTTCGTATAAGTTGCCAGAT
+
@@BCBBBBDDFFGGFGGGGGGGGHHHHHHHHHHGGGGGHHHHHHHHHGGGHHHHFGFEECCA@
```

Count the number of blocks of 4 lines in a fastq
The best strategy to use!

QC

FastQC, a convenient tool!

Raw Sequencing Data

↓

Quality Control

FastQC Report

GPS220996_AGTTCAGG-CCAACAGA-L7TK9_L

Summary

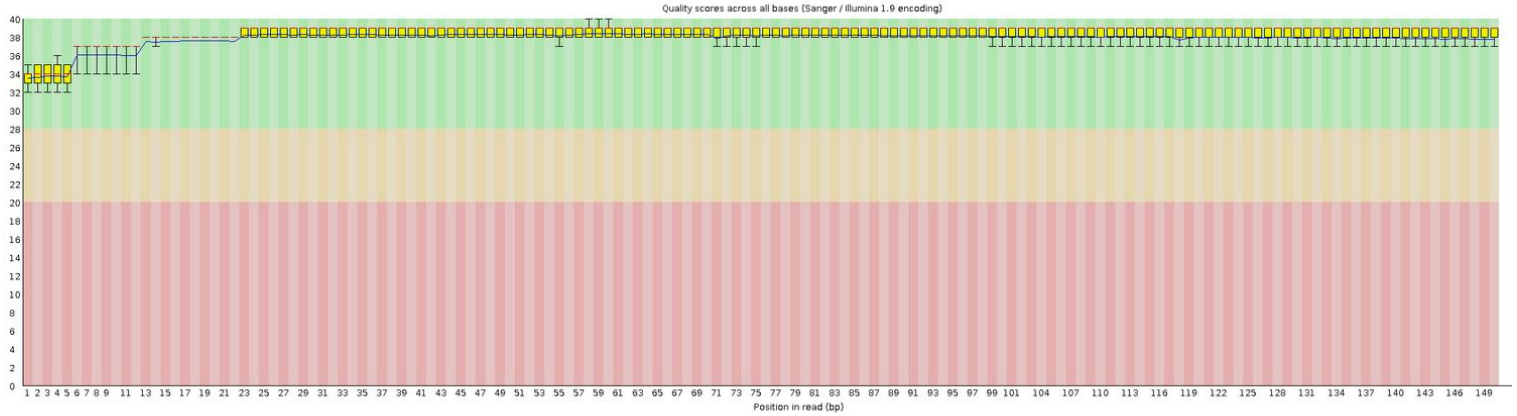
- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

Basic Statistics

Measure	Value
Filename	GPS220996_AGTTCAGG-CCAACAGA-L7TK9_L001_R1_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	624170
Sequences flagged as poor quality	0
Sequence length	150
%GC	36

⟸ Total sequences: Of course, FastQC also provides this information!

Per base sequence quality



QC

FastQC, a convenient tool!

Raw Sequencing Data

↓

Quality Control

FastQC Report

GPS220996_AGTTCAGG-CCAACAGA-L7TK9_L

Summary

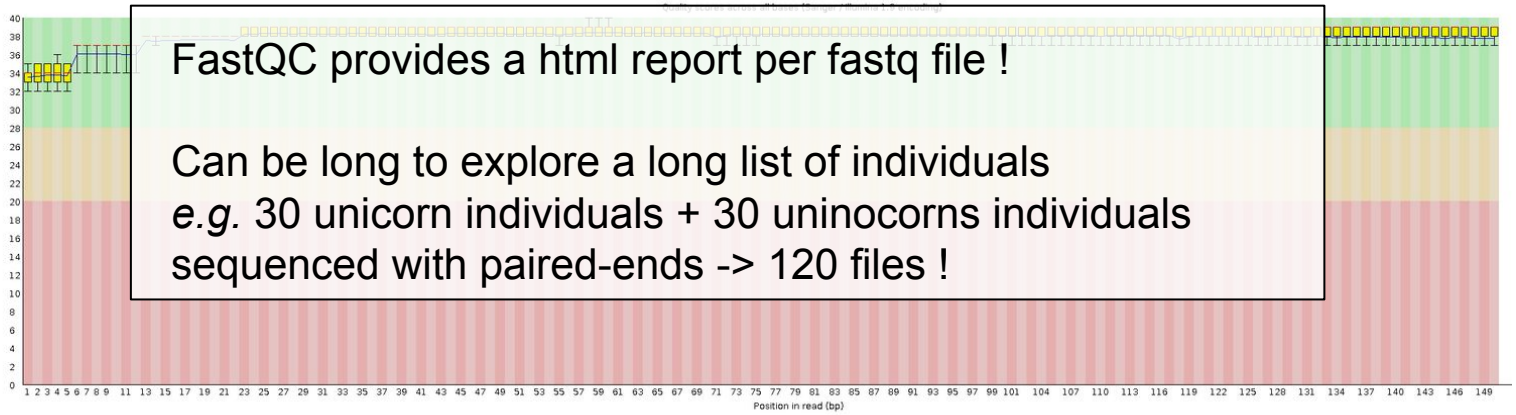
- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

Basic Statistics

Measure	Value
Filename	GPS220996_AGTTCAGG-CCAACAGA-L7TK9_L001_R1_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	624170
Sequences flagged as poor quality	0
Sequence length	150
%GC	36

⟸ Total sequences: Of course, FastQC also provides this information!

Per base sequence quality



FastQC provides a html report per fastq file !

Can be long to explore a long list of individuals
e.g. 30 unicorn individuals + 30 uninocorns individuals
sequenced with paired-ends -> 120 files !

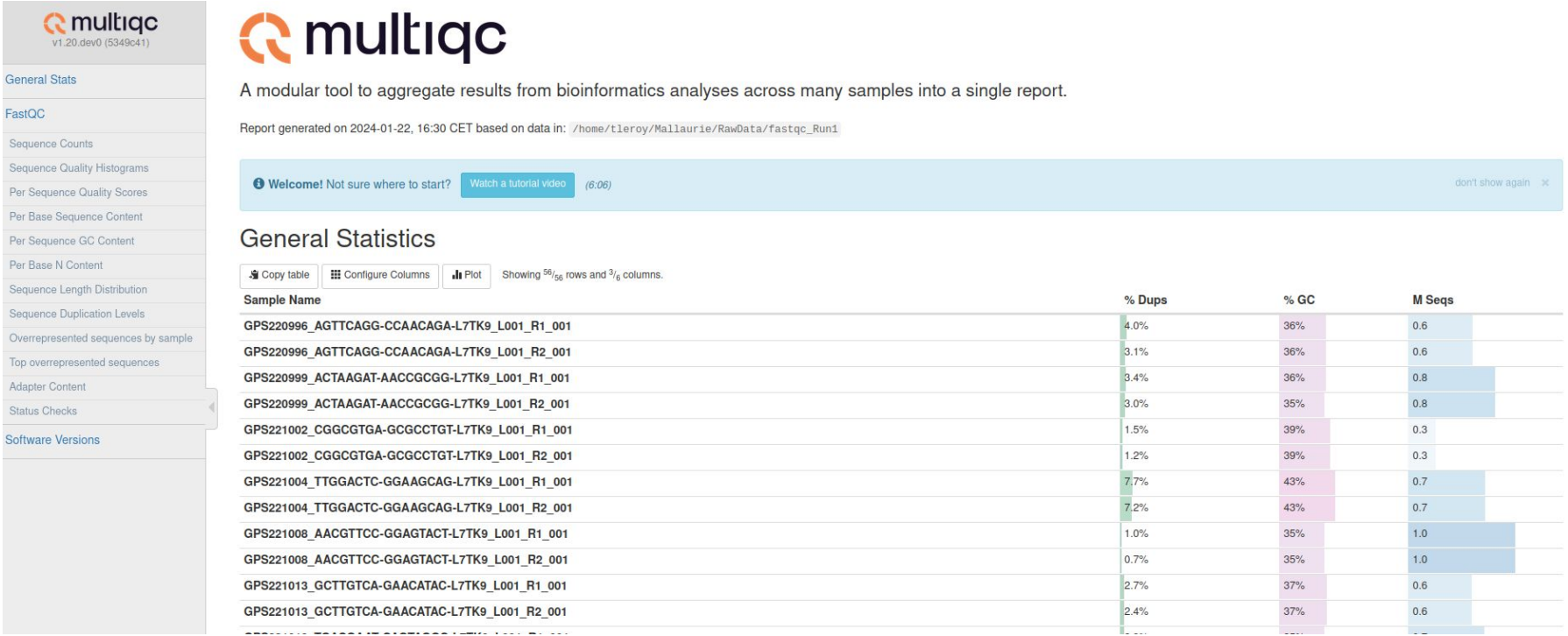
QC

Raw Sequencing Data

↓

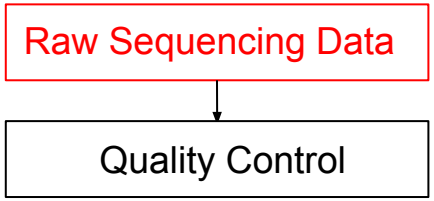
Quality Control

MultiQC, an even more convenient tool!



QC

MultiQC, an even more convenient tool!



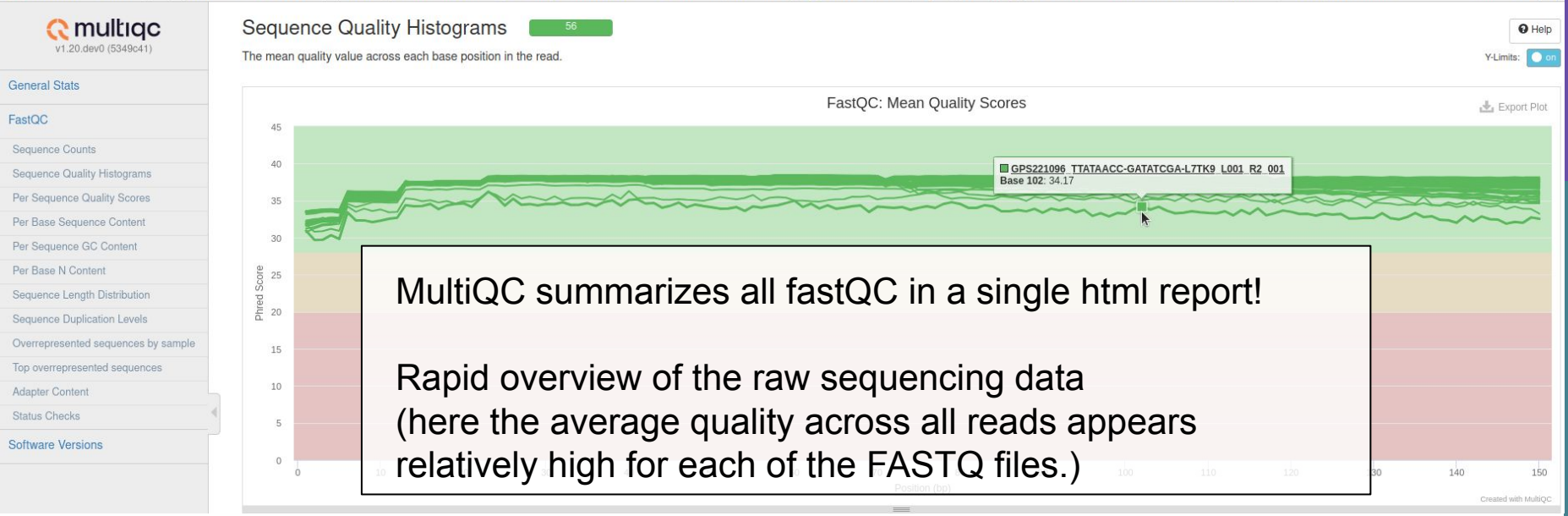
QC

MultiQC, an even more convenient tool!

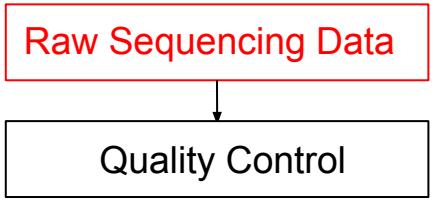
Raw Sequencing Data

↓

Quality Control



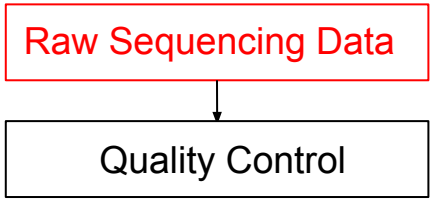
Short reads



Some general information about short read (typically Illumina) sequencing:

- Error probability is relatively low: nowadays ~ 0.1% per base

Short reads

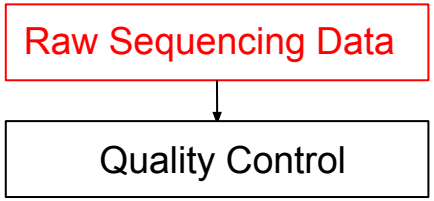


Some general information about short read (typically Illumina) sequencing:

- Error probability is relatively low: nowadays ~ 0.1% per base

Sequencer	**Year of release**	**Read length**	**Accuracy**
Genome Analyzer	2006-2008	36 (GAI) - (2*)75 (GAII)	~98-99%
HiSeq Series	2010	(2*)150	~99.5%
MiSeq	2011	(2*)300	~99.5%
NovaSeq 6000	2017	(2*)250	~99.7%
NovaSeq X	2022	(2*)150	~99.9%

Short reads



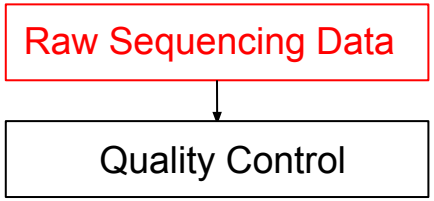
Some general information about short read (typically Illumina) sequencing:

- Error probability is relatively low: nowadays ~ 0.1% per base

Sequencer	Year of release	Read length	Accuracy
Genome Analyzer	2006-2008	36 (GAI) - (2*)75 (GAII)	~98-99%
HiSeq Series	2010	(2*)150	~99.5%
MiSeq	2011	(2*)300	~99.5%
NovaSeq 6000	2017	(2*)250	~99.7%
NovaSeq X	2022	(2*)150	~99.9%

- R2 usually has lower quality scores than R1, especially toward the end of the reads
- Illumina reads are subject to GC bias, extreme GC content (high or low) = lower accuracy (& coverage)

Short reads



Some general information about short read (typically Illumina) sequencing:

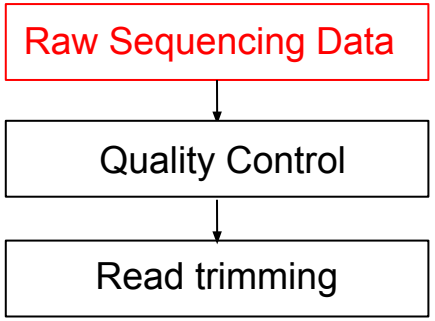
- Error probability is relatively low: nowadays ~ 0.1% per base

Sequencer	Year of release	Read length	Accuracy
Genome Analyzer	2006-2008	36 (GAI) - (2*)75 (GAII)	~98-99%
HiSeq Series	2010	(2*)150	~99.5%
MiSeq	2011	(2*)300	~99.5%
NovaSeq 6000	2017	(2*)250	~99.7%
NovaSeq X	2022	(2*)150	~99.9%

- R2 usually has lower quality scores than R1, especially toward the end of the reads
- Illumina reads are subject to GC bias, extreme GC content (high or low) = lower accuracy (& coverage)

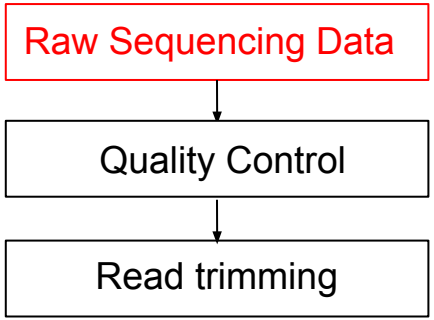
→ **Even with consistently high read quality and progressively lower error rates, read trimming low-quality bases is still recommended to improve overall accuracy and reliability of the data**

Trimming reads



To trim or not to trim, that is the question! → Even with increasing read quality, it remains encouraged for WGS data

Trimming reads



To trim or not to trim, that is the question! → Even with increasing read quality, it remains encouraged for WGS data

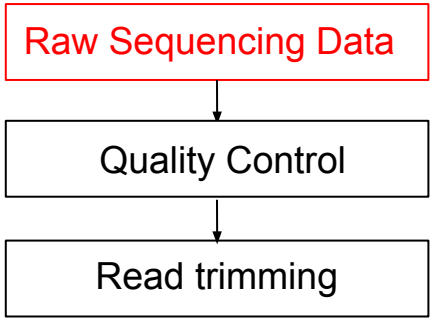
Trimming reads has the objectives:

- Removing adapters

- Removing low quality bases

- Excluding short reads after quality trimming

Trimming reads



To trim or not to trim, that is the question! → Even with increasing read quality, it remains encouraged for WGS data

Trimming reads has the objectives:

- Removing adapters

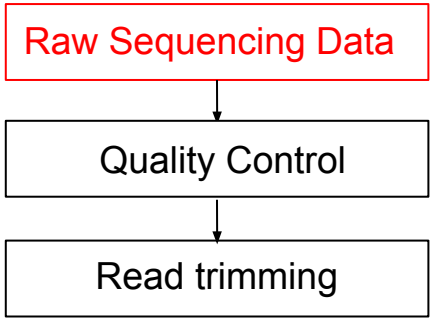
- Removing low quality bases

- Excluding short reads after quality trimming

Trimmomatic is a particularly popular tool, trimming both Single-End (SE) or Paired-End (PE) data

```
trimmomatic PE -threads 4 InfileForward.fastq InfileReverse.fastq  \
        TrimmedOutfileForward_paired.fastq TrimmedOutfileForward_unpaired.fastq \
        TrimmedOutfileReverse_paired.fastq TrimmedOutfileReverse_unpaired.fastq \
        ILLUMINACLIP:Illumina_adapters.fa MINLEN:50 SLIDINGWINDOW:4:20
```

Trimming reads



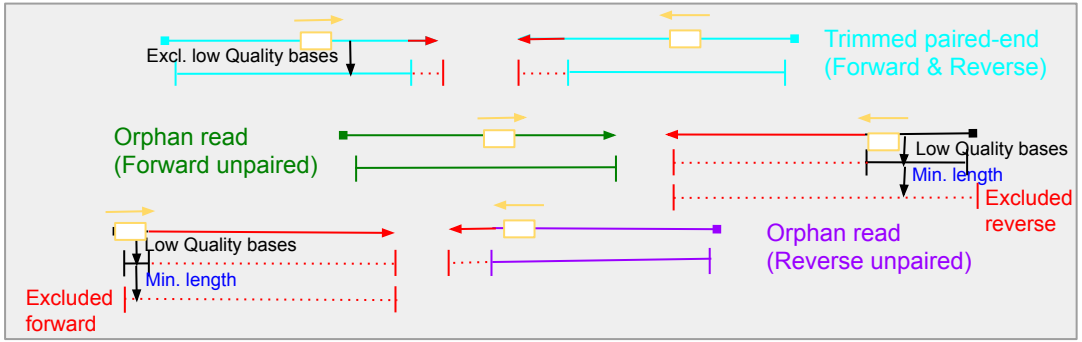
To trim or not to trim, that is the question! → Even with increasing read quality, it remains encouraged for WGS data

Trimming reads has the objectives:

- Removing adapters

- Removing low quality bases

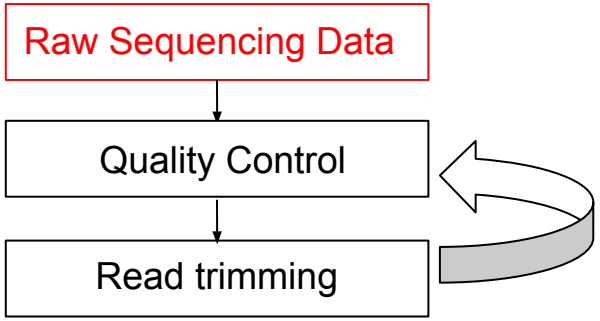
- Excluding short reads after quality trimming



Trimmomatic is a particularly popular tool, trimming both Single-End (SE) or Paired-End (PE) data

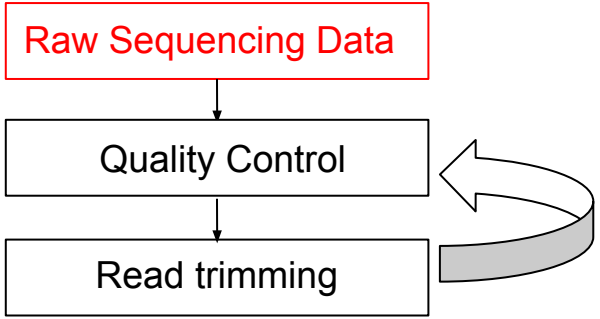
trimmomatic PE -threads 4 ***InfileForward.fastq InfileReverse.fastq*** \
TrimmedOutfileForward_paired.fastq TrimmedOutfileForward_unpaired.fastq \
TrimmedOutfileReverse_paired.fastq TrimmedOutfileReverse_unpaired.fastq \
ILLUMINACLIP:Illumina_adapters.fa MINLEN:50 SLIDINGWINDOW:4:20

Trimming reads



Possibility to control the quality of the trimming by performing another round of QC

Trimming reads



Possibility to control the quality of the trimming by performing another round of QC

FastQC raw data (before)

R1
n=624170 reads

Summary

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

R2
n=624170 reads

Summary

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

FastQC trimmed data (after trimming)

R1 (paired)
n=491646 reads

Summary

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

R2 (paired)
n=491646 reads

Summary

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

R1 (unpaired)
n=131760 reads

Summary

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

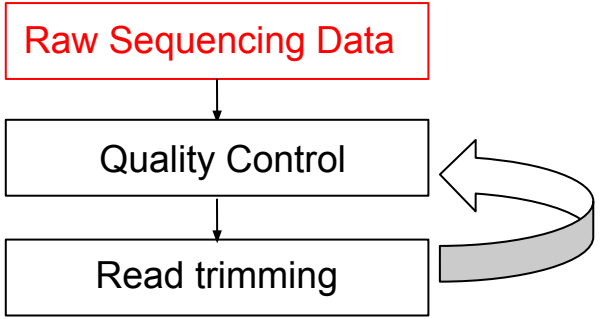
R2 (unpaired)
n=396 reads

Summary

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

We indeed observe the expected improvement of the quality after trimming !

Trimming reads



Possibility to control the quality of the trimming by performing another round of QC

FastQC raw data (before)

R1
n=624170 reads

Summary

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

R2
n=624170 reads

Summary

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

FastQC trimmed data (after trimming)

R1 (paired)
n=491646 reads

Summary

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

R2 (paired)
n=491646 reads

Summary

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

R1 (unpaired)
n=131760 reads

Summary

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

R2 (unpaired)
n=396 reads

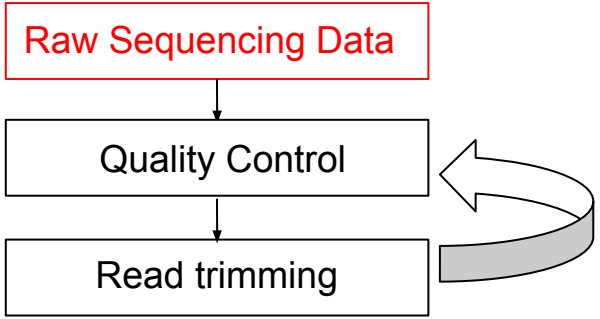
Summary

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

We indeed observe the expected improvement of the quality after trimming !

Note that FastQC tags provide useful indicators and should be treated as warnings. Achieving 'all green' status is not mandatory! Sometimes, warnings are even expected!

Trimming reads



Possibility to control the quality of the trimming by performing another round of QC

F med data (after trimming)

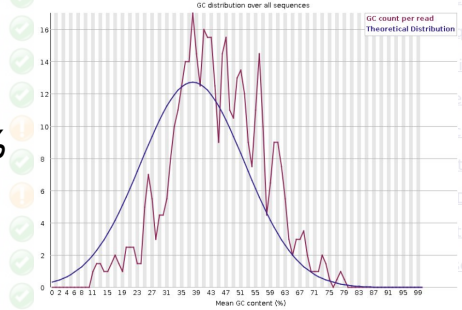
Overall, the quality of R1 reads (first pass) is higher than that of R2 reads (second pass), leading to a higher probability of unpaired R1 compared to unpaired R2 reads

In addition, here the fastq comes for a metagenomic project (ie not a single individual), which has impacts on the warnings

And of course, the number of sequences matters…

Drawing a distribution of GC% based on only 396 reads

=> The last red tag (unpaired R2) is understandable!

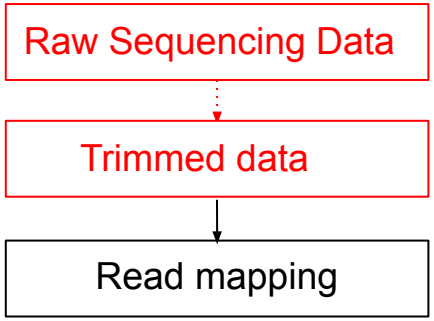


R1 (unpaired) *n=131760 reads*	R2 (unpaired) *n=396 reads*
Summary	**Summary**
Basic Statistics	Basic Statistics
Per base sequence quality	Per base sequence quality
Per tile sequence quality	Per tile sequence quality
Per sequence quality scores	Per sequence quality scores
Per base sequence content	Per base sequence content
Per sequence GC content	Per sequence GC content
Per base N content	Per base N content
Sequence Length Distribution	Sequence Length Distribution
Sequence Duplication Levels	Sequence Duplication Levels
Overrepresented sequences	Overrepresented sequences
Adapter Content	Adapter Content

We indeed observe the expected improvement of the quality after trimming !

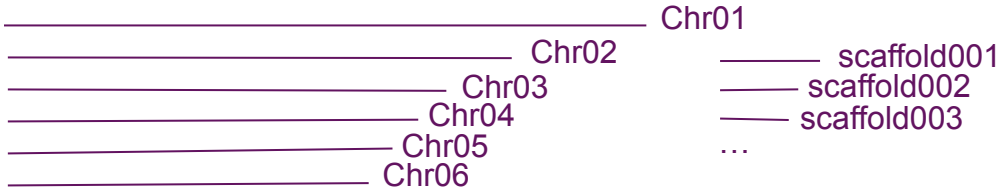
Note that FastQC tags provide useful indicators and should be treated as warnings. Achieving 'all green' status is not mandatory! Sometimes, warnings are even expected!

Mapping reads



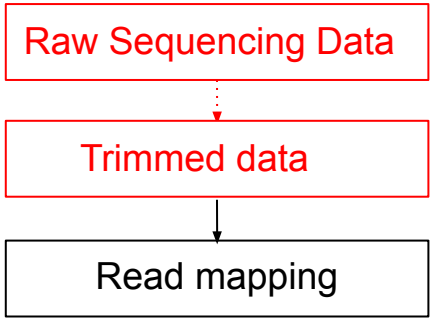
Mapping reads against a reference genome (make sure to choose the right version !) :

Trimmed reads

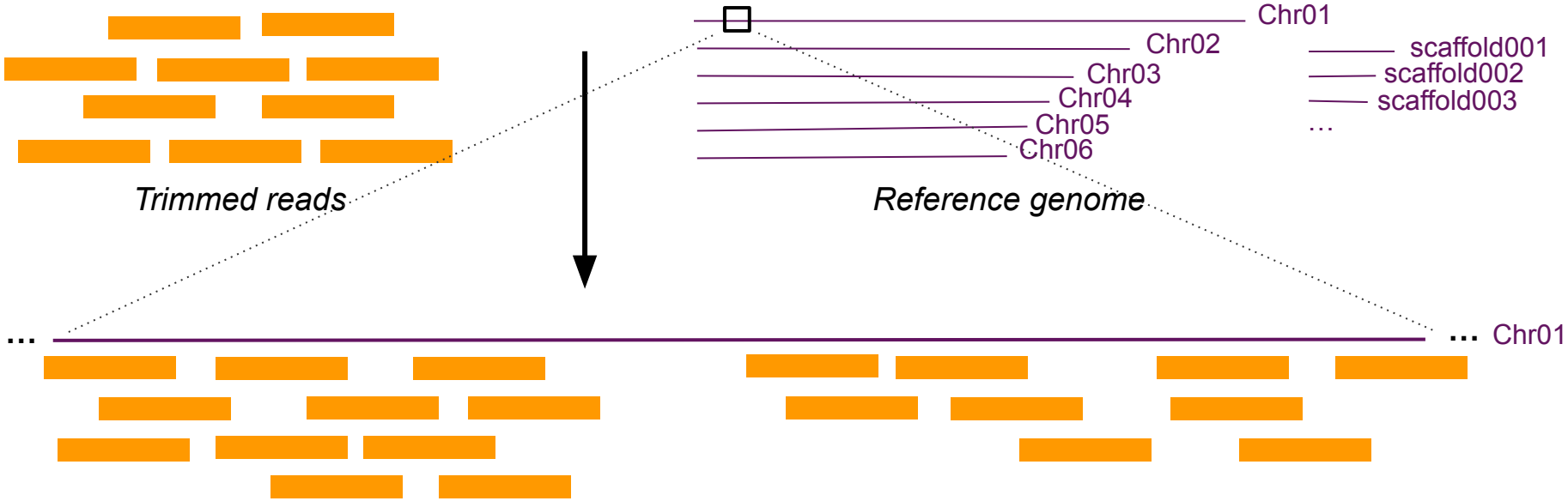


Reference genome

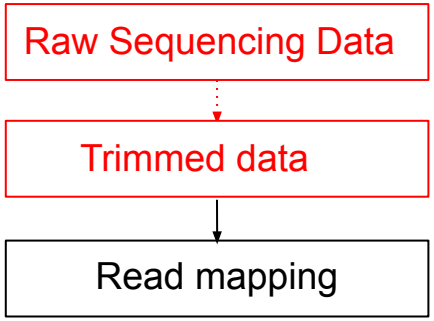
Mapping reads



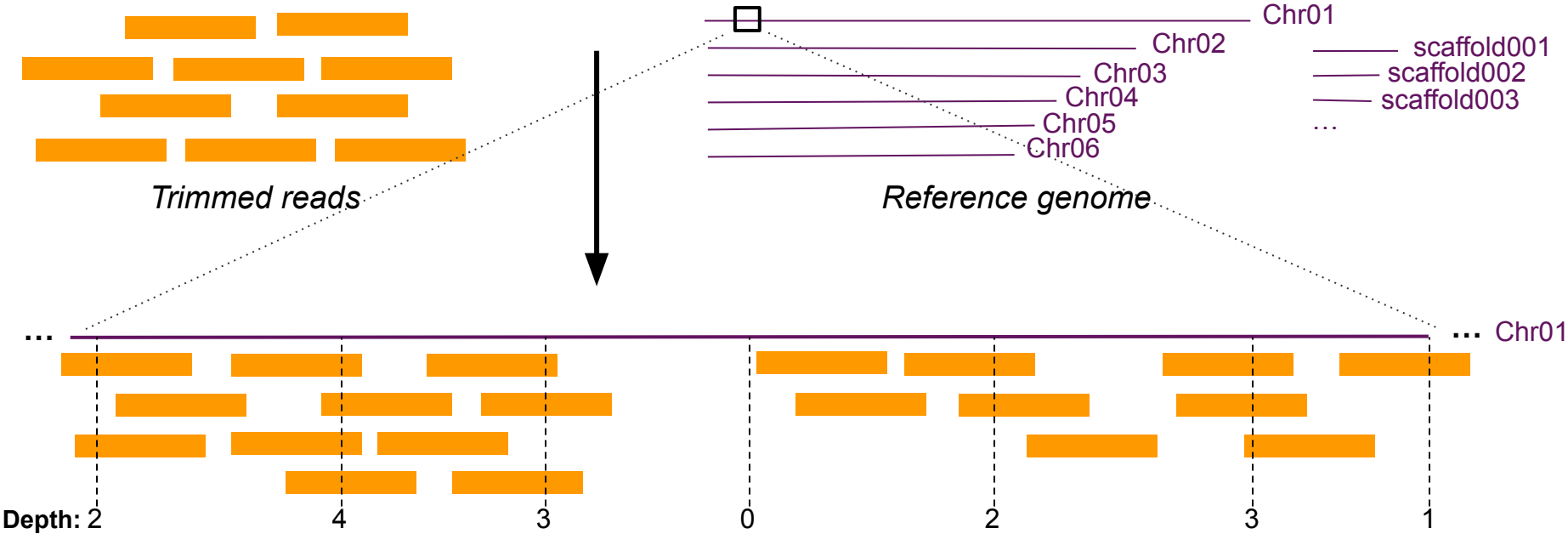
Mapping reads against a reference genome (make sure to choose the right version !) :



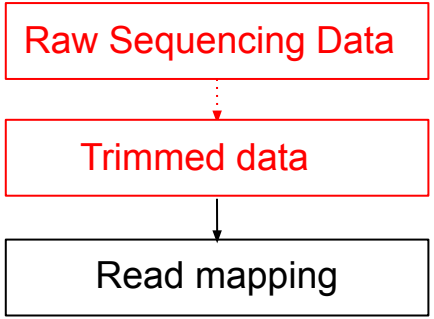
Mapping reads



Mapping reads against a reference genome (make sure to choose the right version !) :



Mapping reads

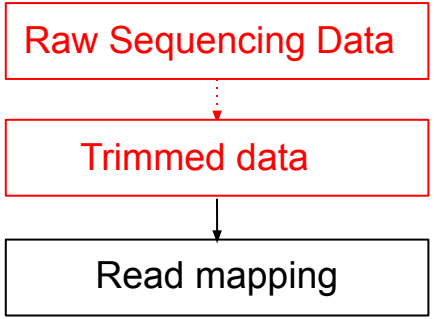


Mappers :

Properly speaking, a mapper is not an aligner!

Read mapping*: locating the approximate position of a read in a reference genome. The goal of read mappers is to identify the general location of reads on the reference genome, without necessarily requiring precise base-by-base alignment.*
Read alignment*: determining the exact sequence correspondence between each base of the read and the reference. This step is more computationally intensive but is expected to improve base-level precision.*

Mapping reads



Mappers :

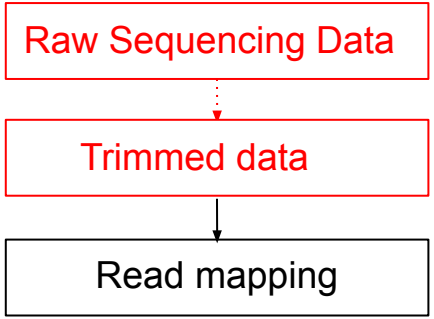
Properly speaking, a mapper is not an aligner!

> ***Read mapping****: locating the approximate position of a read in a reference genome. The goal of read mappers is to identify the general location of reads on the reference genome, without necessarily requiring precise base-by-base alignment.*
> ***Read alignment****: determining the exact sequence correspondence between each base of the read and the reference. This step is more computationally intensive but is expected to improve base-level precision.*

Mapping on a reference requires to have such a reference for the focal species or a (very) closely-related one

Read mapping corresponds to a balance between speed and accuracy: faster algorithms find approximate positions (mapping), while slower, precise methods (alignment) match each base accurately.
Software often allows fine-tuning of detection, *e.g.* --very-fast vs. --very-sensitive modes in Bowtie2

Mapping reads



Mappers :

Properly speaking, a mapper is not an aligner!

> ***Read mapping****: locating the approximate position of a read in a reference genome. The goal of read mappers is to identify the general location of reads on the reference genome, without necessarily requiring precise base-by-base alignment.*
> ***Read alignment****: determining the exact sequence correspondence between each base of the read and the reference. This step is more computationally intensive but is expected to improve base-level precision.*

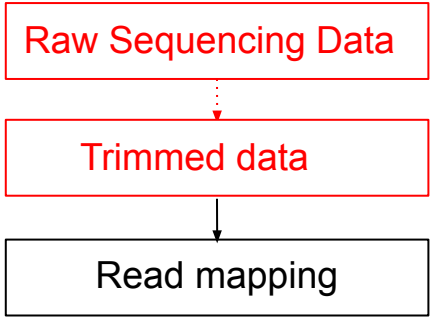
Mapping on a reference requires to have such a reference for the focal species or a (very) closely-related one

Read mapping corresponds to a balance between speed and accuracy: faster algorithms find approximate positions (mapping), while slower, precise methods (alignment) match each base accurately.
Software often allows fine-tuning of detection, *e.g.* --very-fast vs. --very-sensitive modes in Bowtie2

Most popular tools:

- Bowtie and BWA (bowtie2 and bwa-mem2) for WGS data
- STAR and HISAT2 for RNAseq data for RNAseq data (splice-aware algorithms)

Mapping reads



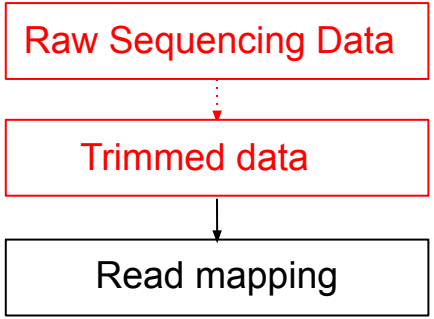
Different format used for the outputs: SAM (text-based) / **BAM (binary)** / CRAM (compressed)

Summarizing the results of the mapping (Samtools flagstat) :

1116143 + 0 in total (QC-passed reads + QC-failed reads)
1115448 + 0 primary → 1115448=trimmed paired-end reads (983292) + unpaired R1 (131760)+ unpaired R2 (396)
[…]
298709 + 0 mapped (26.76% : N/A)
[…]
298014 + 0 primary mapped (26.72% : N/A)
[…]
292198 + 0 properly paired (29.72% : N/A)
[…]

294038 + 0 with itself and mate mapped
[…]
2925 + 0 singletons (0.30% : N/A)
1560 + 0 with mate mapped to a different chr
413 + 0 with mate mapped to a different chr (mapQ>=5)

Mapping reads



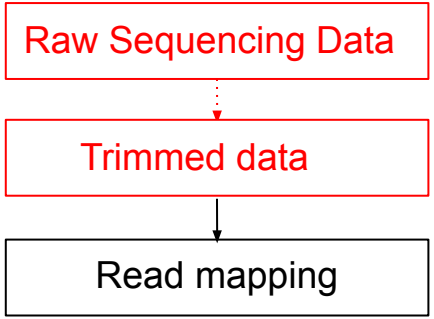
Different format used for the outputs: SAM (text-based) / **BAM (binary)** / CRAM (compressed)

Summarizing the results of the mapping (Samtools flagstat) :

1116143 + 0 in total (QC-passed reads + QC-failed reads)
1115448 + 0 primary → 1115448=trimmed paired-end reads (983292) + unpaired R1 (131760)+ unpaired R2 (396)
[…]
298709 + 0 mapped (26.76% : N/A) → 298709 / 1115448 ~ 26.8%
[…]
298014 + 0 primary mapped (26.72% : N/A) → 298014 / 1115448 ~ 26.7% (only primary, not secondary mapped)
[…]
292198 + 0 properly paired (29.72% : N/A) → 292198 / 983292 ~ 29.7% (PE reads on the same chr, different orientation, relatively short distance between the two PE reads)
[…]

294038 + 0 with itself and mate mapped
[…]
2925 + 0 singletons (0.30% : N/A)
1560 + 0 with mate mapped to a different chr
413 + 0 with mate mapped to a different chr (mapQ>=5)

Mapping reads

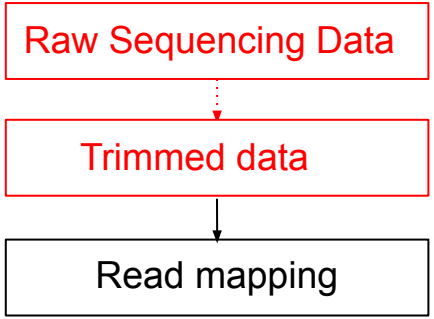


Different format used for the outputs: SAM (text-based) / **BAM (binary)** / CRAM (compressed)

Summarizing the results of the mapping (Samtools flagstat) :

1116143 + 0 in total (QC-passed reads + QC-failed reads)
1115448 + 0 primary → 1115448=trimmed paired-end reads (983292) + unpaired R1 (131760)+ unpaired R2 (396)
[…]
298709 + 0 mapped (26.76% : N/A) → 298709 / 1115448 ~ 26.8%
[…]
298014 + 0 primary mapped (26.72% : N/A) → 298014 / 1115448 ~ 26.7% (only primary, not secondary mapped)
[…]
292198 + 0 properly paired (29.72% : N/A) → 292198 / 983292 ~ 29.7% (PE reads on the same chr, different orientation, relatively short distance between the two PE reads)
[…]
294038 + 0 with itself and mate mapped → 294038 / 983292 (PE reads mapped, even on different chr, distance, …)
[…]
2925 + 0 singletons (0.30% : N/A) → 2925 / 983292 ~ 0.3% (only one of the PE reads is mapped)
1560 + 0 with mate mapped to a different chr → 1560 / 983292 ~ 0.2% (PE reads map on different chr)
413 + 0 with mate mapped to a different chr (mapQ>=5) → 413 / 983292 ~ 0.04% (PE reads map on different chr, each with a “low but decent” mapping quality)

Mapping reads



Mapping qualities (MAPQ)

MAPQ is a score indicating the confidence in the mapping of a read to the genome (expressed in $-10\log_{10}$ probability that the mapping position is wrong).
-> MAPQ is calculated based on the likelihood function, providing values ranging from 0 to 60 (for bwa mem, note that the scale varies from one software to another)

Confidence:
MAPQ10: 90%
MAPQ20: 99%
MAPQ30: 99,9%
MAPQ40: 99,99%

A **low MAPQ** (close to 0) suggests that the read may not be mapped correctly, either because it maps to multiple locations or the mapping is ambiguous.

A **high MAPQ** (close to 60 in bwa mem) suggests that the mapping is very confident and reliable.

Filtering low-confidence reads (e.g. MAPQ <5, <20, …) is generally performed since these mapping are more error-prone, excluding them therefore improve downstream analysis such as SNP calling

Mapping reads

Raw Sequencing Data

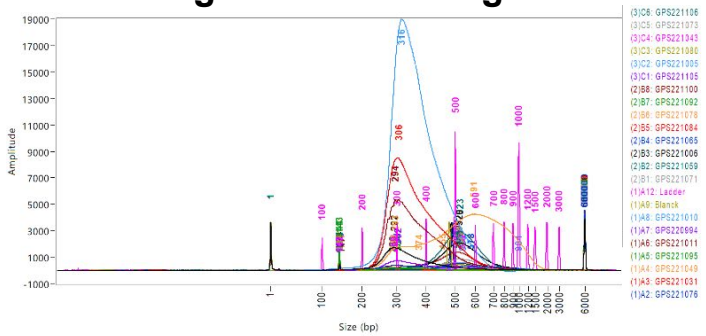
↓

Trimmed data

↓

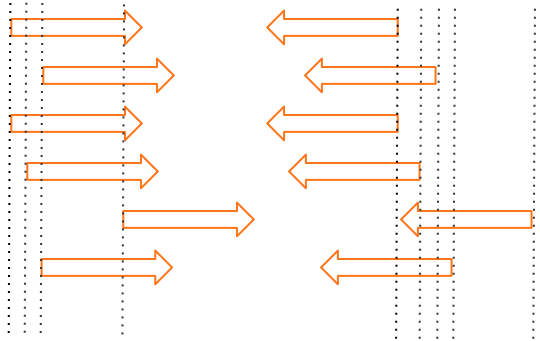
Read mapping

DNA fragmentation using sonication

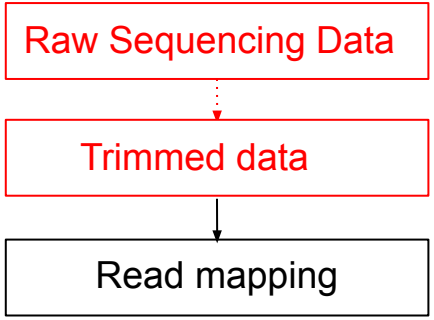


Detecting PCR duplicates (or not)

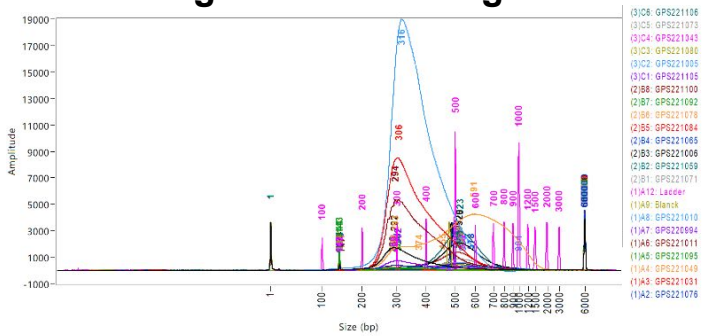
… … Chr01



Mapping reads

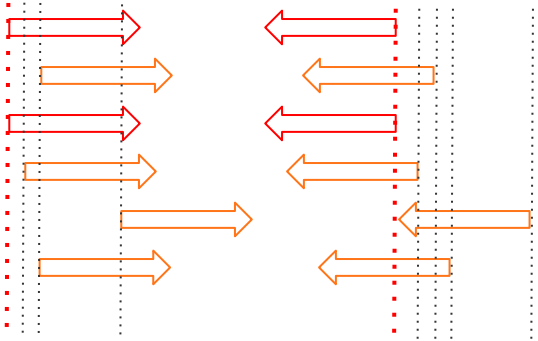


DNA fragmentation using sonication



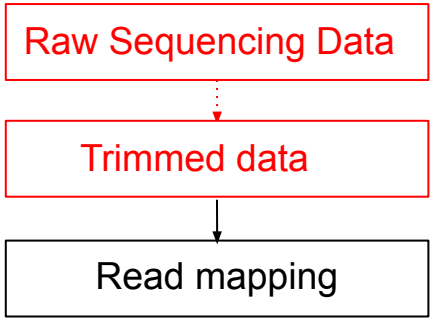
Detecting PCR duplicates (or not)

… … Chr01

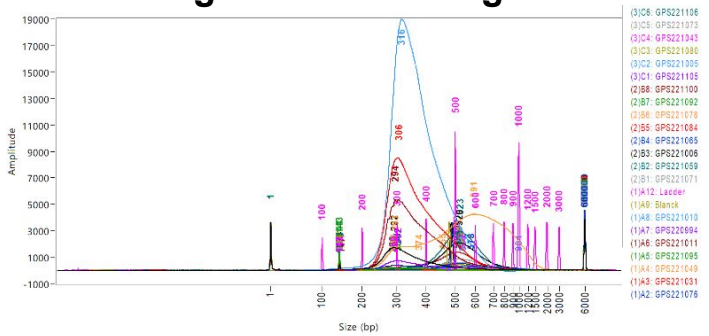


With PE data, it is expected to be quite infrequent to have reads starting and ending at the same locations

Mapping reads

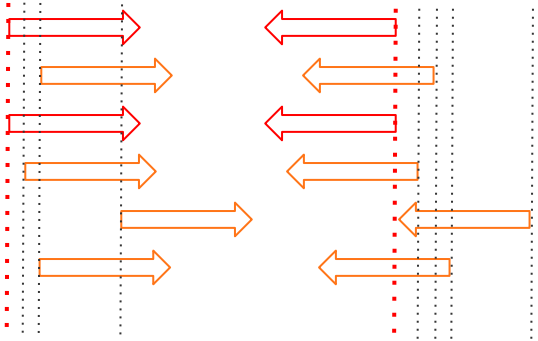


DNA fragmentation using sonication

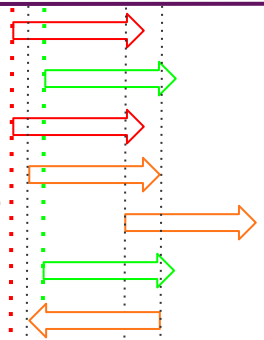


Detecting PCR duplicates (or not)

… … Chr01



With PE data, it is expected to be quite infrequent to have reads starting and ending at the same locations

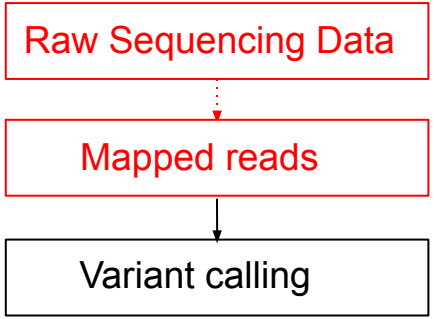


With SE data, it is more challenging to assess duplicates! A risk of overestimating the number of duplicates. A choice to be conservative or not!

With restriction enzymes (e.g. RADseq data), the proportion of false PCR duplicates can be high, as reads originate from the same restriction site. A difference between single-digest vs. double-digest RAD.

Popular tools: Picard (picard Markduplicates) and Samtools (markdup) -> Mark (not remove) duplicates!

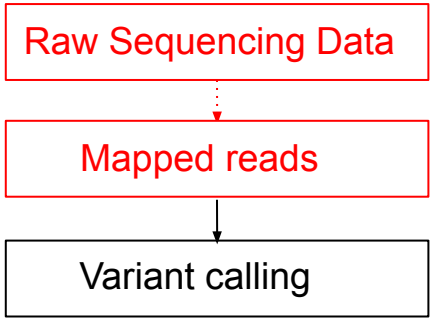
Variant calling



Briefly, variant callers are tools that detect genetic variants such as **single nucleotide polymorphisms (SNPs)**, **insertions**, **deletions**, and **structural variants** from mapped reads

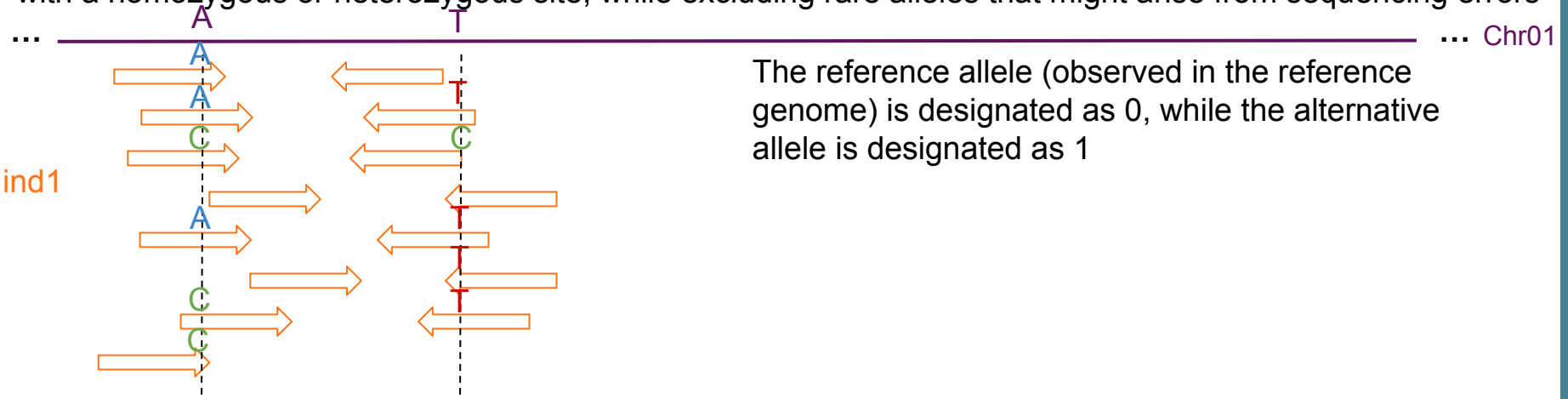
SNP callers aim to identify the most likely genotypes, i.e., determining whether the data are more consistent with a homozygous or heterozygous site, while excluding rare alleles that might arise from sequencing errors

Variant calling



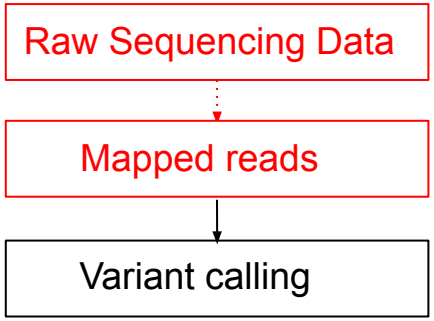
Briefly, variant callers are tools that detect genetic variants such as **single nucleotide polymorphisms (SNPs)**, **insertions**, **deletions**, and **structural variants** from mapped reads

SNP callers aim to identify the most likely genotypes, i.e., determining whether the data are more consistent with a homozygous or heterozygous site, while excluding rare alleles that might arise from sequencing errors



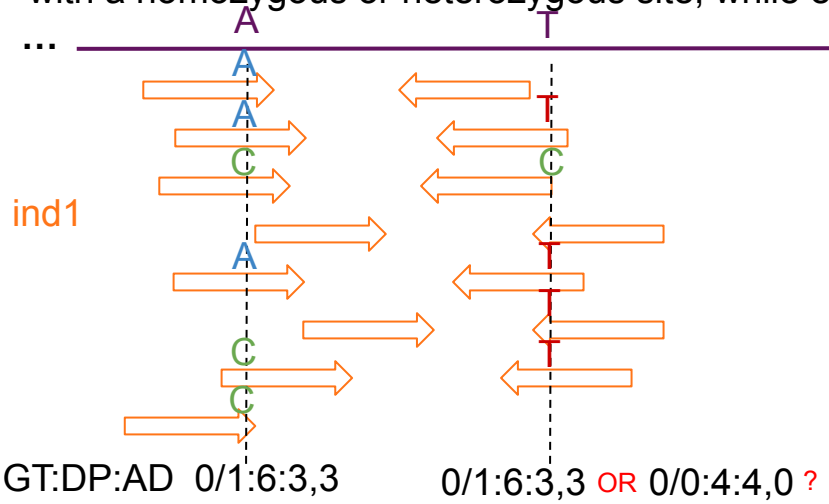
The reference allele (observed in the reference genome) is designated as 0, while the alternative allele is designated as 1

Variant calling



Briefly, variant callers are tools that detect genetic variants such as **single nucleotide polymorphisms (SNPs)**, **insertions**, **deletions**, and **structural variants** from mapped reads

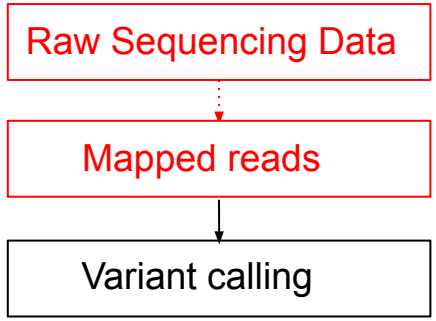
SNP callers aim to identify the most likely genotypes, i.e., determining whether the data are more consistent with a homozygous or heterozygous site, while excluding rare alleles that might arise from sequencing errors

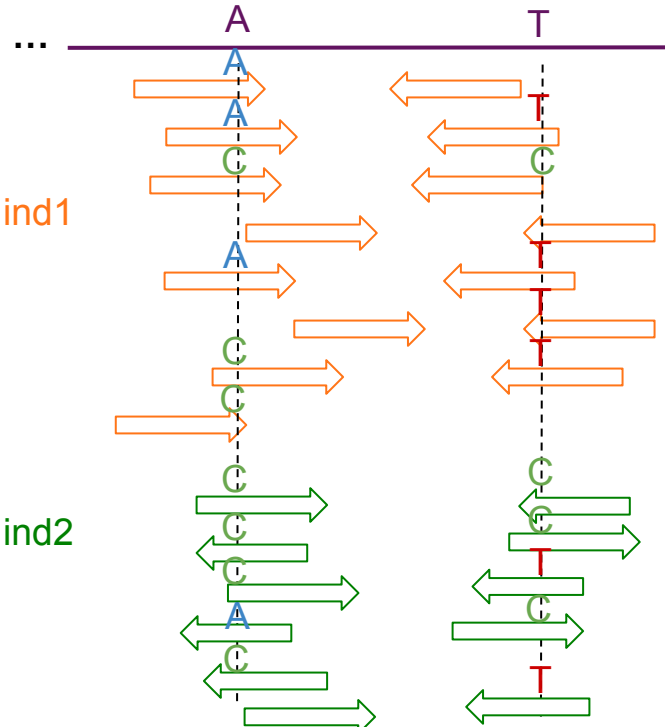


The reference allele (observed in the reference genome) is designated as 0, while the alternative allele is designated as 1

Variant callers assess base quality, mapping quality, and read depth to make accurate predictions of variants.
-> Variant calling is complex, with a difficult balance between identifying true positives and limiting the number of false positives

Variant calling



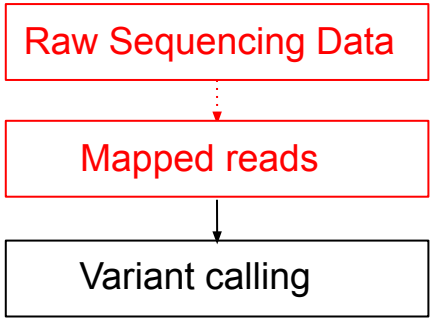


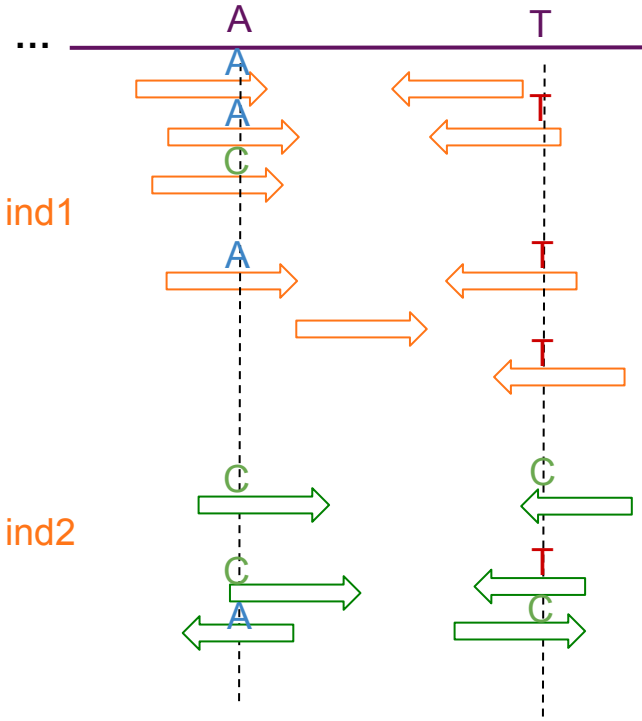
Multi-sample variant calling. By analyzing all individuals together, variant callers can:

- **Increase sensitivity**: Detect low-frequency variants that might be missed in single-sample analyses
- **Improve accuracy**: Use allele frequency data to distinguish true variants from sequencing errors
- **Enable joint genotyping**: Call genotypes across all samples consistently, which helps in downstream analyses like population genetics and association studies

Most popular tools are **GATK**, **FreeBayes** and **Samtools**

Variant calling





Here, the data appears to be associated with low coverage (< 10x)

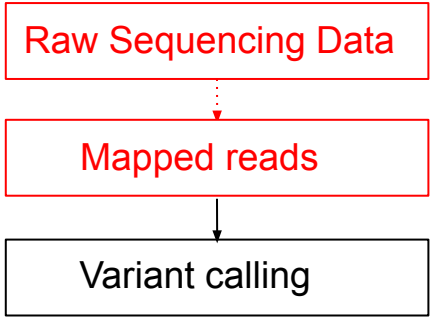
With fewer reads, the probability of accurately determining homozygous or heterozygous genotypes decreases, often leading to more uncertain genotype calls

A more accurate strategy is to compute the probability of each genotype (*i.e.* 0/0, 0/1, 1/1)

PosA	GT:GL	.:-1.5,-0.2,-2.0	.:-0.2,-1.2,-5.1
PosT	GT:GL	.:-1.2,-0.3,-1.2	.:-1.2,-0.3,-1.2

ANGSD is a popular tool, specialized for genotype likelihood calculations, especially in low coverage or population studies

Variant calling



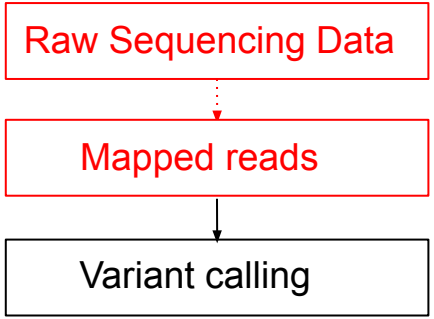
VCF (Variant Calling Format) file

```
##fileformat=VCFv4.2
##FILTER=<ID=PASS,Description="All filters passed">
##ALT=<ID=NON_REF,Description="Represents any possible alternative allele at this location">
##FILTER=<ID=LowQual,Description="Low quality">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=MIN_DP,Number=1,Type=Integer,Description="Minimum DP observed within the GVCF block">
##FORMAT=<ID=PGT,Number=1,Type=String,Description="Physical phasing haplotype information, describing how the alternate alleles are phased in relation to one another">
##FORMAT=<ID=PID,Number=1,Type=String,Description="Physical phasing ID information, where each unique ID within a given sample (but not across samples) connects records within a phasing group">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification">
##FORMAT=<ID=PS,Number=1,Type=Integer,Description="Phasing set (typically the position of the first variant in the set)">
##FORMAT=<ID=RGQ,Number=1,Type=Integer,Description="Unconditional reference genotype confidence, encoded as a phred quality -10*log10 p(genotype call is wrong)">
##FORMAT=<ID=SB,Number=4,Type=Integer,Description="Per-sample component statistics which comprise the Fisher's Exact Test to detect strand bias.">
…
##GATKCommandLine=[...]
#CHROM POS     ID        REF     ALT     QUAL     FILTER   INFO     FORMAT IND01   IND02   IND03 ….
```

The header of a VCF provides extremely important information, explaining how to read the file, the commands used, etc… Read it!

The most important line is the last one starting with a “#”, which provides the column names, including the names of individuals (from column 10)

Variant calling



VCF (Variant Calling Format) file

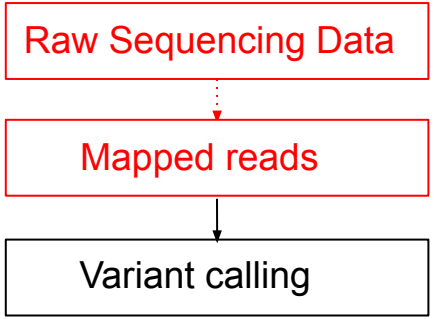
```
##fileformat=VCFv4.2
##FILTER=<ID=PASS,Description="All filters passed">
##ALT=<ID=NON_REF,Description="Represents any possible alternative allele at this location">
##FILTER=<ID=LowQual,Description="Low quality">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=MIN_DP,Number=1,Type=Integer,Description="Minimum DP observed within the GVCF block">
##FORMAT=<ID=PGT,Number=1,Type=String,Description="Physical phasing haplotype information, describing how the alternate alleles are phased in relation to one another">
##FORMAT=<ID=PID,Number=1,Type=String,Description="Physical phasing ID information, where each unique ID within a given sample (but not across samples) connects records within a phasing group">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification">
##FORMAT=<ID=PS,Number=1,Type=Integer,Description="Phasing set (typically the position of the first variant in the set)">
##FORMAT=<ID=RGQ,Number=1,Type=Integer,Description="Unconditional reference genotype confidence, encoded as a phred quality -10*log10 p(genotype call is wrong)">
##FORMAT=<ID=SB,Number=4,Type=Integer,Description="Per-sample component statistics which comprise the Fisher's Exact Test to detect strand bias.">
…
##GATKCommandLine=[...]
```

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	IND01	IND02	IND03 ….
Chr01	5671	.	T	C	332.46	.	AC=2;AF=0.001196;DP=13376	GT:AD:DP:GQ:PL	1/1:1,12:13:2:363,2,0	./.:0,0:0:.:0,0,0	…
Chr01	5698	.	T	C	3633.8	.	AC=8;AF=0.004779;DP=12793	GT:AD:DP:GQ:PL	0/0:9,0:9:24:.:.:0,24,360	0/0:19,0:19:48:0,48,720	…

…

Two first variants are C/T variants, with a T allele in the reference genome (0) and a C as an alternative one (1)

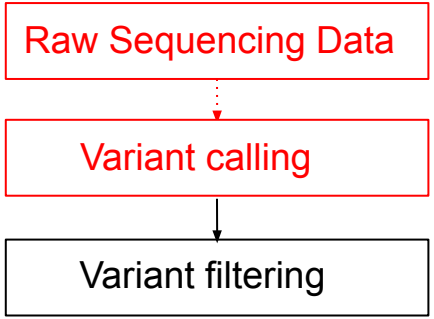
Variant calling



VCF (Variant Calling Format) file

```
##fileformat=VCFv4.2
##FILTER=<ID=PASS,Description="All filters passed">
##ALT=<ID=NON_REF,Description="Represents any possible alternative allele at this location">
##FILTER=<ID=LowQual,Description="Low quality">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=MIN_DP,Number=1,Type=Integer,Description="Minimum DP observed within the GVCF block">
##FORMAT=<ID=PGT,Number=1,Type=String,Description="Physical phasing haplotype information, describing how the alternate alleles are phased in relation to one another">
##FORMAT=<ID=PID,Number=1,Type=String,Description="Physical phasing ID information, where each unique ID within a given sample (but not across samples) connects records within a
phasing group">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification">
##FORMAT=<ID=PS,Number=1,Type=Integer,Description="Phasing set (typically the position of the first variant in the set)">
##FORMAT=<ID=RGQ,Number=1,Type=Integer,Description="Unconditional reference genotype confidence, encoded as a phred quality -10*log10 p(genotype call is wrong)">
##FORMAT=<ID=SB,Number=4,Type=Integer,Description="Per-sample component statistics which comprise the Fisher's Exact Test to detect strand bias.">
…
##GATKCommandLine=[...]
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO            FORMAT IND01        IND02   IND03 ….
Chr01   5671    .       T       C       332.46  .       AC=2;AF=0.001196;DP=13376       GT:AD:DP:GQ:PL  1/1:1,12:13:2:363,2,0        ./.:0,0:0:.:0,0,0  …
Chr01   5698    .       T       C       3633.8  .       AC=8;AF=0.004779;DP=12793       GT:AD:DP:GQ:PL  0/0:9,0:9:24:.:.:0,24,360    0/0:19,0:19:48:0,48,720  …
…
NC_037638.1     41082   .       C       G,T     …       ←  Multiple alt alleles: 0/2 genotype =>C/T; 1/2 genotype =>G/T etc
…
NC_037638.1     48060   .       G       AT      …       ←  INDEL  0/1 genotype = G/AT
…
```

Variant filtering



Selecting most reliable variants

Quality by Depth (QD): ensures that variants have sufficient supporting read depth, filtering out low-quality variants with insufficient read
Mapping Quality (MQ): filtering out variants potentially associated with mapping errors
Base Quality (BQ): average base quality of the reads supporting the variant, filtering out variants with low base quality
Minimum Allele Frequency (AF): filtering out variants with low allele frequency (but could be an issue for some popgen analyses)
Depth of Coverage (DP): ensures that the variant is supported by a large number of reads
Variant type (SNPs *vs.* indels): calling INDELs is more challenging than SNPs. Using SNP variants only can be a good strategy
Missing rate: filtering out variants with excessive missing data is a standard to reduce false positives (true also at the sample level)
…

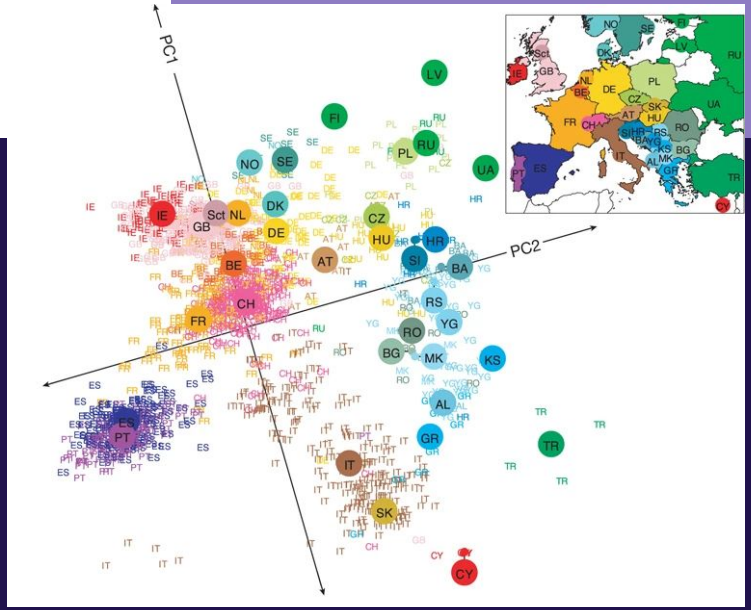
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	IND01	IND02	IND03 ….
Chr01	5671	.	T	C	332.46	.	AC=2;AF=0.001196;DP=13376	GT:AD:DP:GQ:PL	1/1:1,12:13:2:363,2,0		./.:0,0:0::0,0,0 …
Chr01	5698	.	T	C	3633.8	.	AC=8;AF=0.004779;DP=12793	GT:AD:DP:GQ:PL	0/0:9,0:9:24:.:.:0,24,360		0/0:19,0:19:48:0,48,720 …
…											

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	IND01	IND02	IND03 ….
Chr01	5671	.	T	C	332.46	PASS	AC=2;AF=0.001196;DP=13376	GT:AD:DP:GQ:PL	1/1:1,12:13:2:363,2,0		./.:0,0:0::0,0,0 …
Chr01	5698	.	T	C	3633.8	LowQual	AC=8;AF=0.004779;DP=12793	GT:AD:DP:GQ:PL	0/0:9,0:9:24:.:.:0,24,360		0/0:19,0:19:48:0,48,720 …
…											

Physalia

Courses

Basic bioinformatics



25/11/2024

Physalia course

Thibault Leroy, Yann Bourgeois

Physalia

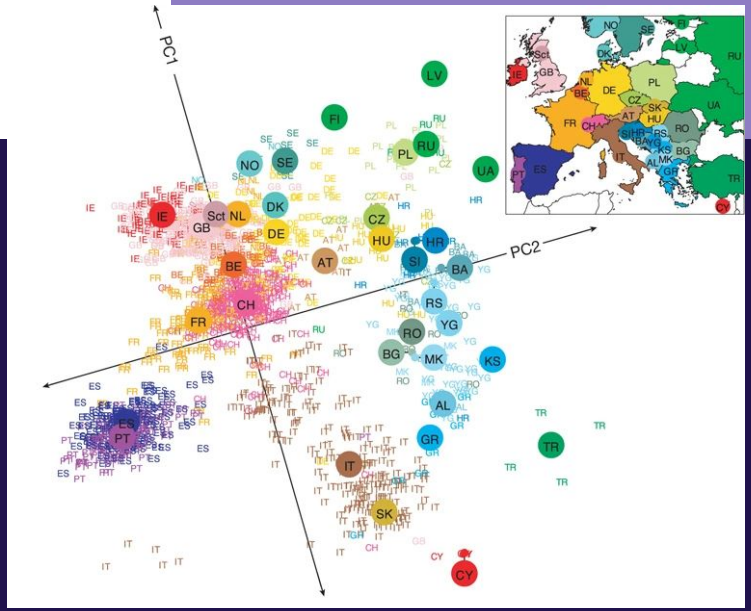
Courses

Basic bioinformatics Tutorial: Wrap-up

25/11/2024

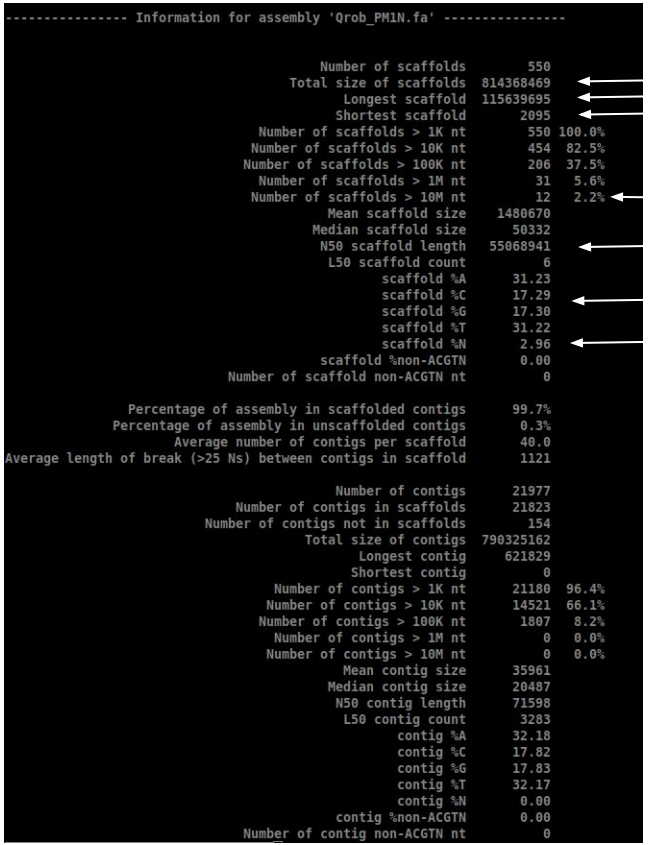
Physalia course

Thibault Leroy, Yann Bourgeois



Step 1: Evaluating the reference genome

/home/ubuntu/Share/software/assemblathon2-analysis/assemblathon_stats.pl Qrob_PM1N.fa > Qrob_PM1N.fa.assemblathon.txt



```
---------- Information for assembly 'Qrob_PM1N.fa' ----------

                                     Number of scaffolds        550
                                 Total size of scaffolds  814368469
                                       Longest scaffold  115639695
                                      Shortest scaffold       2095
                           Number of scaffolds > 1K nt        550 100.0%
                          Number of scaffolds > 10K nt        454  82.5%
                         Number of scaffolds > 100K nt        206  37.5%
                           Number of scaffolds > 1M nt         31   5.6%
                          Number of scaffolds > 10M nt         12   2.2%
                                 Mean scaffold size    1480670
                               Median scaffold size      50332
                               N50 scaffold length   55068941
                                L50 scaffold count          6
                                      scaffold %A       31.23
                                      scaffold %C       17.29
                                      scaffold %G       17.30
                                      scaffold %T       31.22
                                      scaffold %N        2.96
                            scaffold %non-ACGTN        0.00
                   Number of scaffold non-ACGTN nt           0

                Percentage of assembly in scaffolded contigs      99.7%
              Percentage of assembly in unscaffolded contigs       0.3%
                    Average number of contigs per scaffold        40.0
Average length of break (>25 Ns) between contigs in scaffold       1121

                                     Number of contigs       21977
                          Number of contigs in scaffolds       21823
                      Number of contigs not in scaffolds         154
                                Total size of contigs  790325162
                                      Longest contig     621829
                                     Shortest contig           0
                           Number of contigs > 1K nt       21180  96.4%
                          Number of contigs > 10K nt       14521  66.1%
                         Number of contigs > 100K nt        1807   8.2%
                           Number of contigs > 1M nt           0   0.0%
                          Number of contigs > 10M nt           0   0.0%
                                 Mean contig size       35961
                               Median contig size       20487
                               N50 contig length       71598
                                L50 contig count        3283
                                      contig %A       32.18
                                      contig %C       17.82
                                      contig %G       17.83
                                      contig %T       32.17
                                      contig %N        0.00
                            contig %non-ACGTN        0.00
                   Number of contig non-ACGTN nt           0
```

Assembly size 814 Mb (a bit more than the expected size of ~750 Mb)
longest scaffold: 115 Mb
2095 scaffolds

12 scaffolds of > 10 Mb (= 12 chromosomes)

N50: 55 Mb (6 longest scaffolds to reach half of the genome ~407 Mb)

GC content ~ 34.6%

N content ~ 3.0%

Based on the Assemblathon results, how long is the assembled genome?

-> 814.4 Mb

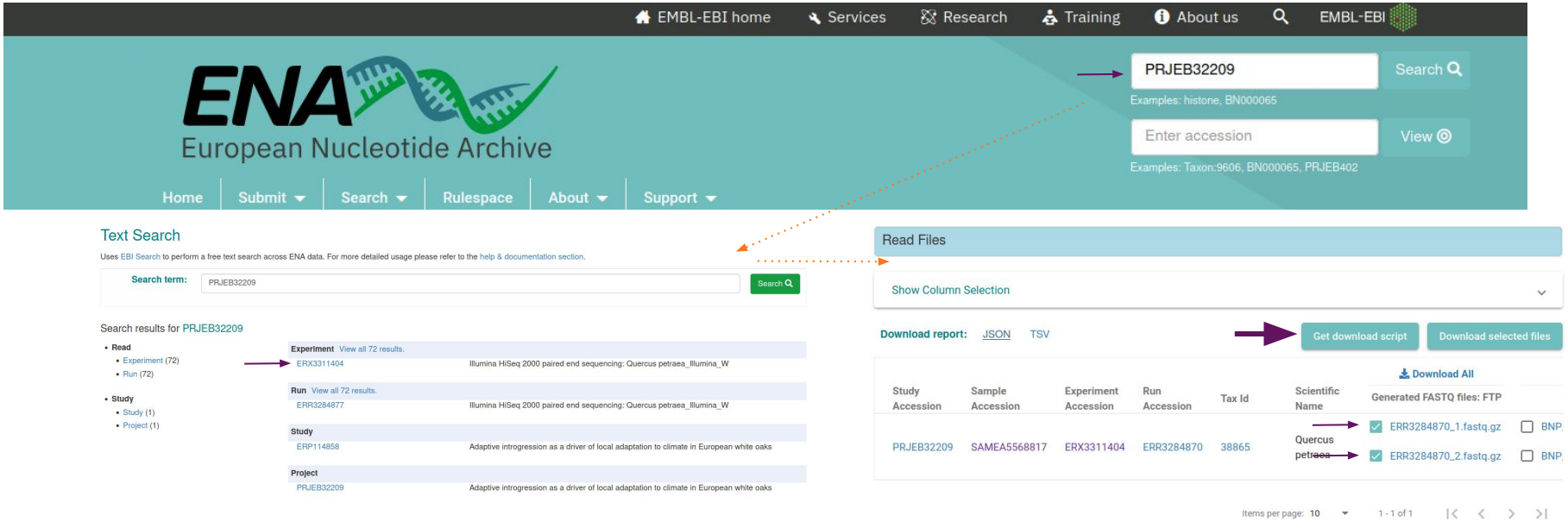
Does its total length exceed the expected genome size for this species (0.75 Gb)?

-> **Yes**, some duplicated regions of the genome in the assembly?

Additionally, is this assembly resolved at the chromosome level?

-> **Yes**, almost (12 long scaffolds), but still fragmented

Step 2: Downloading sequencing data from public repositories like the SRA (Sequence Read Archive)



Direct access to the right ftp link to download the data (direct from a computing cluster)

wget -nc ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR328/003/ERR3284873/ERR3284873_1.fastq.gz
wget -nc ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR328/003/ERR3284873/ERR3284873_2.fastq.gz

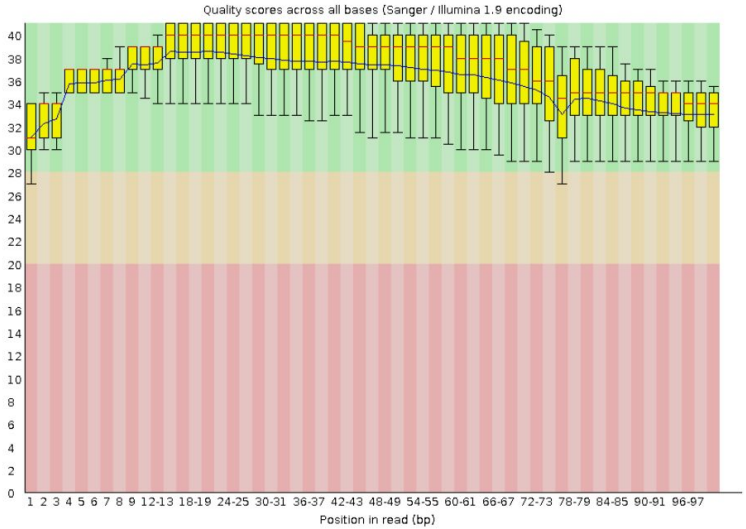
Step 3: Quality control of raw sequencing data

fastqc

Basic Statistics

Measure	Value
Filename	ERR3284869_1.subset.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	19306
Total Bases	1.8 Mbp
Sequences flagged as poor quality	0
Sequence length	30-101
%GC	33

Per base sequence quality



multiqc

multiqc

A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

Report generated on 2024-11-18, 07:35 CET based on data in:

- `/work/genphyse/cytogen/Thibault/beegenomics_2023disk/beegenomics/PuceOak/Puce_oak/analysis_physalia/ERR3284869_1.fastq.gz.subset4kreads`

Welcome! Not sure where to start? Watch a tutorial video *(6:06)* don't show again

General Statistics

Copy table | Configure columns | Scatter plot | Violin plot | Showing 6/6 rows and 3/6 columns. | Export as CSV

High percentage of duplicates (Kmer)

Sample Name	% Dups	% GC	M Seqs
ERR3284869_1.subset	37.7%	33%	0.0 M
ERR3284869_2.subset	36.7%	33%	0.0 M
ERR3284873_1.subset	34.7%	33%	0.0 M
ERR3284873_2.subset	34.7%	33%	0.0 M
ERR3284898_1.subset	29.2%	33%	0.0 M
ERR3284898_2.subset	29.0%	33%	0.0 M

FastQC

Step 3: Quality control of raw sequencing data

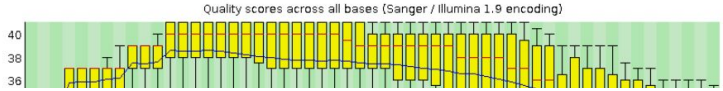
multiqc *subset* (multiqc .)
multiqc subset
"*" transformed in " " in the pdf :(

fastqc

Basic Statistics

Measure	Value
Filename	ERR3284869_1.subset.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	19306
Total Bases	1.8 Mbp
Sequences flagged as poor quality	0
Sequence length	30-101
%GC	33

Per base sequence quality



multiqc

multiqc

A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

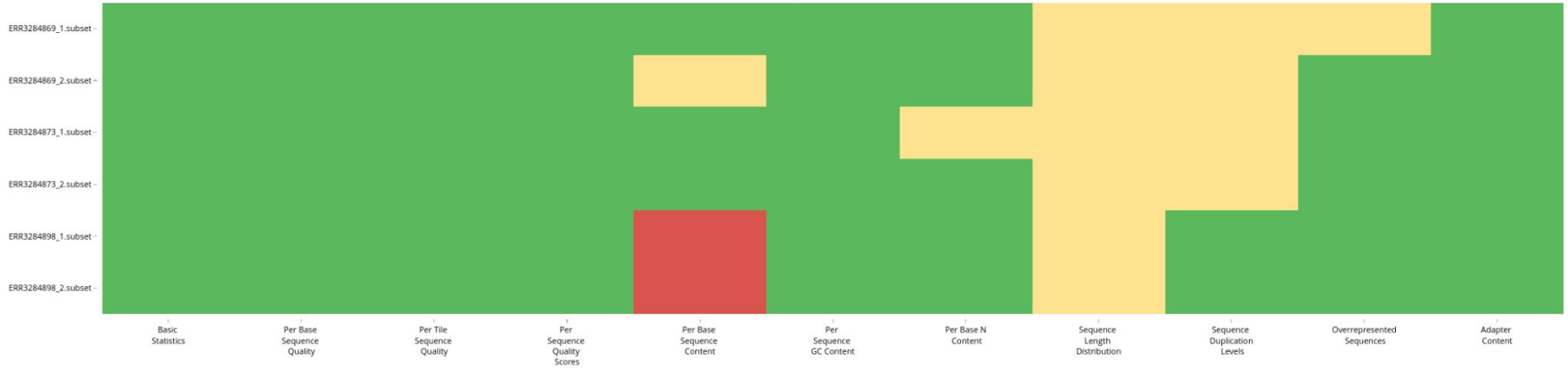
Report generated on 2024-11-18, 07:35 CET based on data in:

- /work/genphyse/cytogen/Thibault/beegenomics_2023disk/beegenomics/PuceOak/Puce_oak/analysis_physalia/ERR3284869_1.fastq.gz.subset4kreads

Welcome! Not sure where to start? Watch a tutorial video (6:06) don't show again

General Statistics

FastQC: Status Checks
11 samples



Step 4: Read trimming

" automatically transformed in “ :(

"

```
for j in ERR3284869 ERR3284873 ERR3284898; do
        acc=$(echo "./$j")
        outacc=$(echo "../Trimming/$j")
        trimmomatic PE -threads 1 -phred33 "$acc"_1.subset.fastq.gz "$acc"_2.subset.fastq.gz
        "$outacc"_1.cleaned.subset.fastq.gz "$outacc"_1.cleaned_unpaired.subset.fastq.gz
        "$outacc"_2.cleaned.subset.fastq.gz "$outacc"_2.cleaned_unpaired.subset.fastq.gz
        ILLUMINACLIP:/home/ubuntu/src/conda/envs/Workshop_TL_YB_Calling2/share/trimmomatic-0.39-2/adaptersTruSeq3-PE-2.fa:2:30:10
        LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:50
done
```

Remaining PE reads
Remaining SE reads R1 (R2 lost)
Remaining SE reads R2 (R1 lost)

General Statistics

Copy table | Configure columns | Scatter plot | Violin plot | Showing 12/12 rows and 3/6 columns. | Export as CSV

Sample Name	% Dups	% GC	M Seqs
ERR3284869_1_cleaned.subset	36.9%	33%	0.0 M
ERR3284869_1_cleaned_unpaired.subset	41.3%	34%	0.0 M
ERR3284869_2_cleaned.subset	36.2%	33%	0.0 M
ERR3284869_2_cleaned_unpaired.subset	35.1%	32%	0.0 M
ERR3284873_1_cleaned.subset	34.7%	33%	0.0 M
ERR3284873_1_cleaned_unpaired.subset	33.9%	33%	0.0 M
ERR3284873_2_cleaned.subset	34.7%	33%	0.0 M
ERR3284873_2_cleaned_unpaired.subset	32.0%	33%	0.0 M
ERR3284898_1_cleaned.subset	29.2%	33%	0.0 M
ERR3284898_1_cleaned_unpaired.subset	24.1%	35%	0.0 M
ERR3284898_2_cleaned.subset	29.1%	33%	0.0 M
ERR3284898_2_cleaned_unpaired.subset	27.6%	33%	0.0 M

Step 4: Read trimming

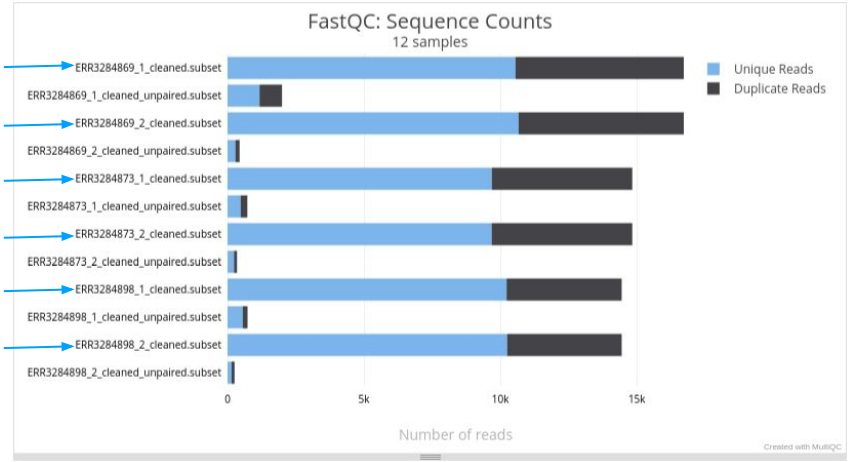
```
for j in ERR3284869 ERR3284873 ERR3284898; do
        acc=$(echo "./$j")
        outacc=$(echo "../Trimming/$j")
        trimmomatic PE -threads 1 -phred33 "$acc"_1.subset.fastq.gz "$acc"_2.subset.fastq.gz
        "$outacc"_1.cleaned.subset.fastq.gz "$outacc"_1.cleaned_unpaired.subset.fastq.gz
        "$outacc"_2.cleaned.subset.fastq.gz "$outacc"_2.cleaned_unpaired.subset.fastq.gz
        ILLUMINACLIP:/home/ubuntu/src/conda/envs/Workshop_TL_YB_Calling2/share/trimmomatic-0.39-2/adaptersTruSeq3-PE-2.fa:2:30:10
        LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:50
done
```

Remaining PE reads
Remaining SE reads R1 (R2 lost)
Remaining SE reads R2 (R1 lost)

pool1
- Remaining PE reads R1
- Remaining PE reads R2

pool2
- Remaining PE reads R1
- Remaining PE reads R2

pool3
- Remaining PE reads R1
- Remaining PE reads R2



Step 4: Read trimming

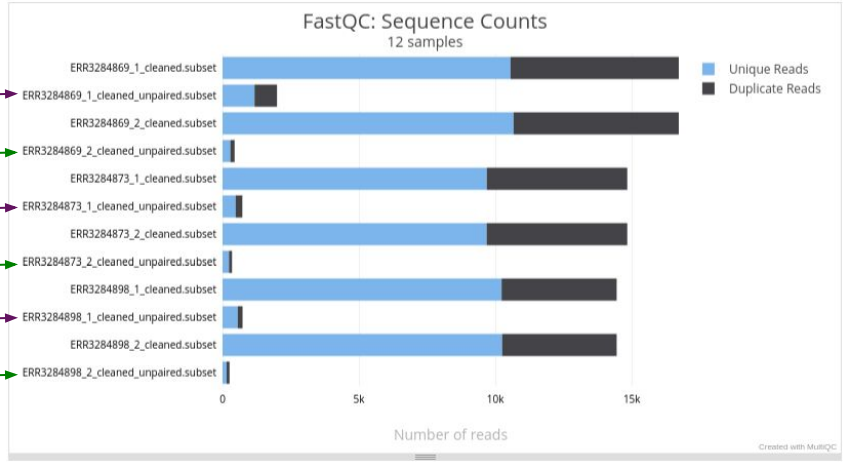
```
for j in ERR3284869 ERR3284873 ERR3284898; do
        acc=$(echo "./$j")
        outacc=$(echo "../Trimming/$j")
        trimmomatic PE -threads 1 -phred33 "$acc"_1.subset.fastq.gz "$acc"_2.subset.fastq.gz
        "$outacc"_1.cleaned.subset.fastq.gz "$outacc"_1.cleaned_unpaired.subset.fastq.gz
        "$outacc"_2.cleaned.subset.fastq.gz "$outacc"_2.cleaned_unpaired.subset.fastq.gz
        ILLUMINACLIP:/home/ubuntu/src/conda/envs/Workshop_TL_YB_Calling2/share/trimmomatic-0.39-2/adaptersTruSeq3-PE-2.fa:2:30:10
        LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:50
done
```

Remaining PE reads
Remaining SE reads R1 (R2 lost)
Remaining SE reads R2 (R1 lost)

pool1
- Remaining SE reads R1 (R2 lost)
- Remaining SE reads R2 (R1 lost)

pool2
- Remaining SE reads R1 (R2 lost)
- Remaining SE reads R2 (R1 lost)

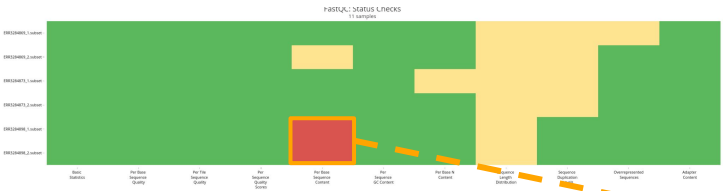
pool3
- Remaining SE reads R1 (R2 lost)
- Remaining SE reads R2 (R1 lost)

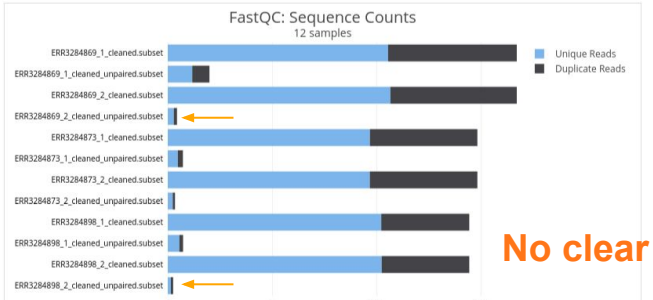


Step 4: Read trimming

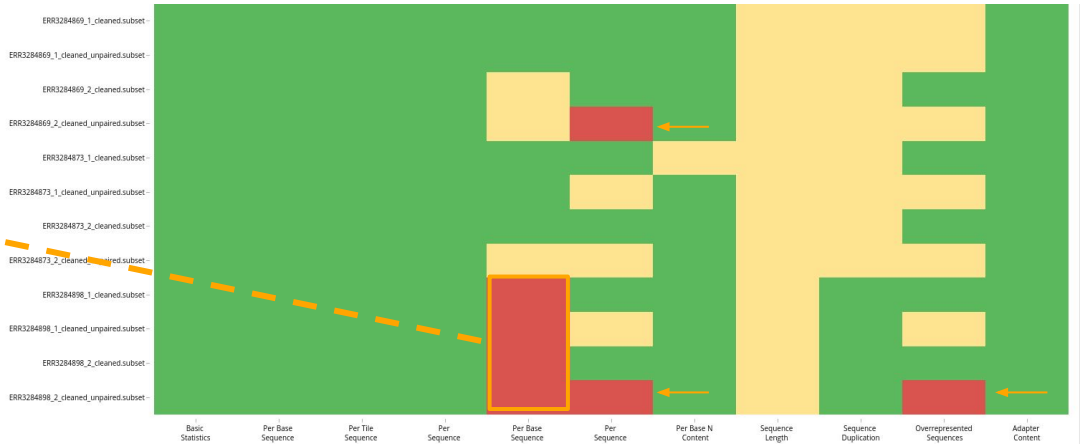
```
for j in ERR3284869 ERR3284873 ERR3284898; do
        acc=$(echo "./$j")
        outacc=$(echo "../Trimming/$j")
        trimmomatic PE -threads 1 -phred33 "$acc"_1.subset.fastq.gz "$acc"_2.subset.fastq.gz
        "$outacc"_1.cleaned.subset.fastq.gz "$outacc"_1.cleaned_unpaired.subset.fastq.gz
        "$outacc"_2.cleaned.subset.fastq.gz "$outacc"_2.cleaned_unpaired.subset.fastq.gz
        ILLUMINACLIP:/home/ubuntu/src/conda/envs/Workshop_TL_YB_Calling2/share/trimmomatic-0.39-2/adaptersTruSeq3-PE-2.fa:2:30:10
        LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:50
done
```

multiqc before trimming





multiqc after trimming



No clear improvements, some additional red tags, for some unpaired R2!

Step 5: Mapping reads against a reference genome

```
cd ~/Day1_bioinfo/data-oak/Trimming
for j in ERR3284869 ERR3284873 ERR3284898; do
        acc=$(echo “$j”)
        outacc=$(echo “../Mapping/$j””_subset_trimmedPE”)
        bwa-mem2 mem ../$REFERENCE_GENOME “$acc”_1.cleaned.subset.fastq.gz
        “$acc”_2.cleaned.subset.fastq.gz | samtools view -Sb - > $outacc.bam
done
```

Step 6: Mapping reads against a reference genome

ERR3284869 (PE reads after trimming)

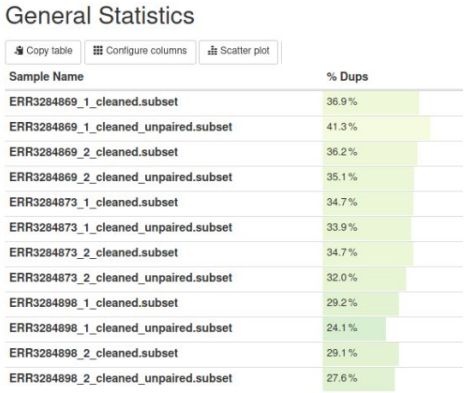
34162 + 0 in total (QC-passed reads + QC-failed reads)
33438 + 0 primary
0 + 0 secondary
724 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
33760 + 0 mapped (98.82% : N/A)
33036 + 0 primary mapped (98.80% : N/A)
33438 + 0 paired in sequencing
16719 + 0 read1
16719 + 0 read2
29108 + 0 properly paired (87.05% : N/A)
32730 + 0 with itself and mate mapped
306 + 0 singletons (0.92% : N/A)
2740 + 0 with mate mapped to a different chr
1413 + 0 with mate mapped to a different chr (mapQ>=5)

Step 7: Removing PCR duplicates

```
for filemetrics in *.duplication_metrics.txt; do
        duplicates=$(grep "Library" $filemetrics)
        echo "$filemetrics        $duplicates"
done
```

filename	percentage duplicates
ERR3284869_withtrimming.trimmedPE.bam.duplication_metrics.txt	**0.072401**
ERR3284869_withtrimming.trimmedunpaired1.bam.duplication_metrics.txt	0.102675
ERR3284869_withtrimming.trimmedunpaired2.bam.duplication_metrics.txt	0.086877
ERR3284873_withtrimming.trimmedPE.bam.duplication_metrics.txt	**0.074513**
ERR3284873_withtrimming.trimmedunpaired1.bam.duplication_metrics.txt	0.069848
ERR3284873_withtrimming.trimmedunpaired2.bam.duplication_metrics.txt	0.07938
ERR3284898_withtrimming.trimmedPE.bam.duplication_metrics.txt	**0.084082**
ERR3284898_withtrimming.trimmedunpaired1.bam.duplication_metrics.txt	0.062445
ERR3284898_withtrimming.trimmedunpaired2.bam.duplication_metrics.txt	0.071926

MultiQC
(simpler approach, potentially totally different values)



Step 8: Allele counts & mpileup files for pool-seq data

mpileup

Qrob_Chr01	1	T	0	*	*	1	^8.	C	0	*	*
Qrob_Chr01	2	C	0	*	*	1	.	C	0	*	*
Qrob_Chr01	3	T	0	*	*	1	.	C	0	*	*
Qrob_Chr01	4	G	0	*	*	2	.^9.	F@	0	*	*
Qrob_Chr01	5	A	0	*	*	2	..	F@	0	*	*
Qrob_Chr01	6	A	0	*	*	3	G.^8.	F@C	1	^7.	@
Qrob_Chr01	7	G	0	*	*	3	A..	FFC	3	…	FF@
Qrob_Chr01	8	T	0	*	*	4	...^8.	FD@@	3	…	EFB
Qrob_Chr01	9	A	0	*	*	3	...	HEF	3	…	FHD
Qrob_Chr01	10	T	0	*	*	4	....	HFF@	3	…	HHF

Synchronized mpileup

Qrob_Chr01	1	T	0:0:0:0:0:0	0:1:0:0:0:0	0:0:0:0:0:0	
Qrob_Chr01	2	C	0:0:0:0:0:0	0:0:1:0:0:0	0:0:0:0:0:0	
Qrob_Chr01	3	T	0:0:0:0:0:0	0:1:0:0:0:0	0:0:0:0:0:0	
Qrob_Chr01	4	G	0:0:0:0:0:0	0:0:0:2:0:0	0:0:0:0:0:0	
Qrob_Chr01	5	A	0:0:0:0:0:0	2:0:0:0:0:0	0:0:0:0:0:0	
Qrob_Chr01	6	A	0:0:0:0:0:0	2:0:0:1:0:0	1:0:0:0:0:0	
Qrob_Chr01	7	G	0:0:0:0:0:0	1:0:0:2:0:0	0:0:0:3:0:0	
Qrob_Chr01	8	T	0:0:0:0:0:0	0:4:0:0:0:0	0:3:0:0:0:0	→ data that will be imported for Friday's proposal
Qrob_Chr01	9	A	0:0:0:0:0:0	3:0:0:0:0:0	3:0:0:0:0:0	
Qrob_Chr01	10	T	0:0:0:0:0:0	0:4:0:0:0:0	0:3:0:0:0:0	

Note: For each pool, the format is A-count:T-count:C-count:G-count:N-count:-count, which represents the number of reads supporting the alleles A, T, C, G, ambigious allele (N) and indels (*).*