
QC dati sequenziamento NGS

9 Gennaio 2024

Paolo Uva

Bioinformatica Clinica
IRCCS Istituto Giannina Gaslini



Agenda

- Retrieve datasets for training from History
- Quality control:
 - FASTQ
 - BAM/CRAM, coverage (and recap with UCSC)
 - VCF

How to retrieve datasets

All datasets are available in a Shared history

<https://usegalaxy.eu/u/puva/h/lezione-7---datasets>

The shared history includes 8 files, organized in 3 datasets using Galaxy **tags**.




How to retrieve datasets

All datasets are available in a Shared history

<https://usegalaxy.eu/u/puva/h/lezione-7---datasets>

The shared history includes 8 files, organized in 3 datasets using Galaxy **tags**.

Tags can help you to better organize your history and track datasets.

1. Click on the dataset to expand it
2. Click on **Add Tags** 
3. Add a tag starting with **#**
 - Tags starting with **#** will be automatically propagated to the outputs of tools using this dataset.
4. Press **Enter**
5. Check that the tag appears below the dataset name

Lezione 7 - Datasets

8.08 GB

51: LowQuality_Reads.fastq.gz

#dataset1

50: HighQuality_Reads.fastq.gz

#dataset1

47: Trio_1_Variants.vcf.gz

#dataset3

44: father.dedup.bam

#dataset3

43: mother.dedup.bam

#dataset3

42: proband.dedup.bam

#dataset3

23: Panel_target_regions.bed

#dataset2

22: Panel_alignment.bam

#dataset2

Dataset 1 - FASTQ

Two FASTQ files with:

- high quality reads
- low quality reads

Lezione 7 - Datasets

8.08 GB

51: LowQuality_Reads.fastq.gz

#dataset1

50: HighQuality_Reads.fastq.gz

#dataset1

47: Trio_1_Variants.vcf.gz

#dataset3

44: father.dedup.bam

#dataset3

43: mother.dedup.bam

#dataset3

42: proband.dedup.bam

#dataset3

23: Panel_target_regions.bed

#dataset2

22: Panel_alignment.bam

#dataset2

Dataset 2 - Gene panel

BAM file with reads aligned to the following genes:

ATRX

CDKL5

CNTNAP2

FOXG1

MECP2

MEF2C

NRXN1

SLC9A6

TCF4

UBE3A

ZEB2

Lezione 7 - Datasets

8.08 GB

51: LowQuality_Reads.fastq.gz

#dataset1

50: HighQuality_Reads.fastq.gz

#dataset1

47: Trio_1_Variants.vcf.gz

#dataset3

44: father.dedup.bam

#dataset3

43: mother.dedup.bam

#dataset3

42: proband.dedup.bam

#dataset3

23: Panel_target_regions.bed

#dataset2

22: Panel_alignment.bam

#dataset2

Dataset 3 - Exome (chr8 only)

Trios (mother, father, proband) with:

- BAM file for each individual
- VCF multisample

For this tutorial, BAMs and VCFs only include sequences and variants from a ***region of chr8***

Lezione 7 - Datasets

8.08 GB

51: LowQuality_Reads.fastq.gz

#dataset1

50: HighQuality_Reads.fastq.gz

#dataset1

47: Trio_1_Variants.vcf.gz

#dataset3

44: father.dedup.bam

#dataset3

43: mother.dedup.bam

#dataset3

42: proband.dedup.bam

#dataset3

23: Panel_target_regions.bed

#dataset2

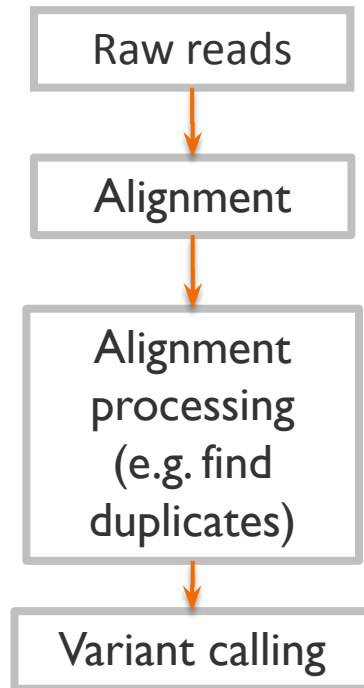
22: Panel_alignment.bam

#dataset2

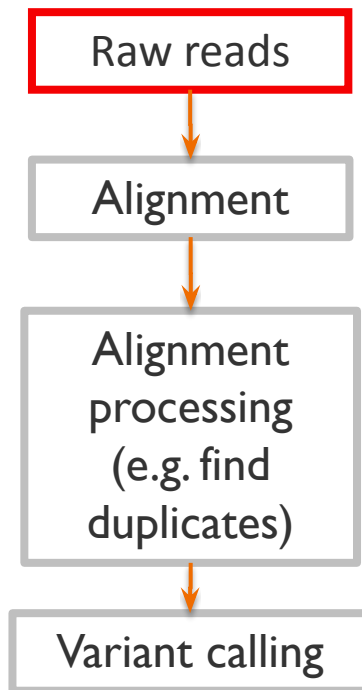
Standard NGS DNA analysis pipeline

Format

Software



FASTQ quality control

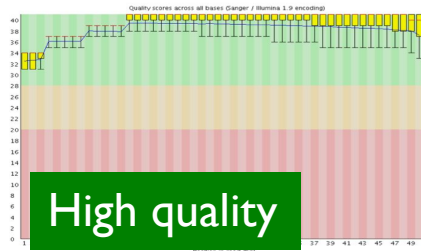


Format
FASTQ

Software
FastQC

Tool	FastQC
Input	FASTQ file (NGS reads)
Output	HTML Report

✓ Per base sequence quality



✗ Per base sequence quality






Exercise 1 - FastQC

- Run **FastQC** on **#dataset 1**
 - Tip: Select **Multiple datasets** option and select both FASTQ files



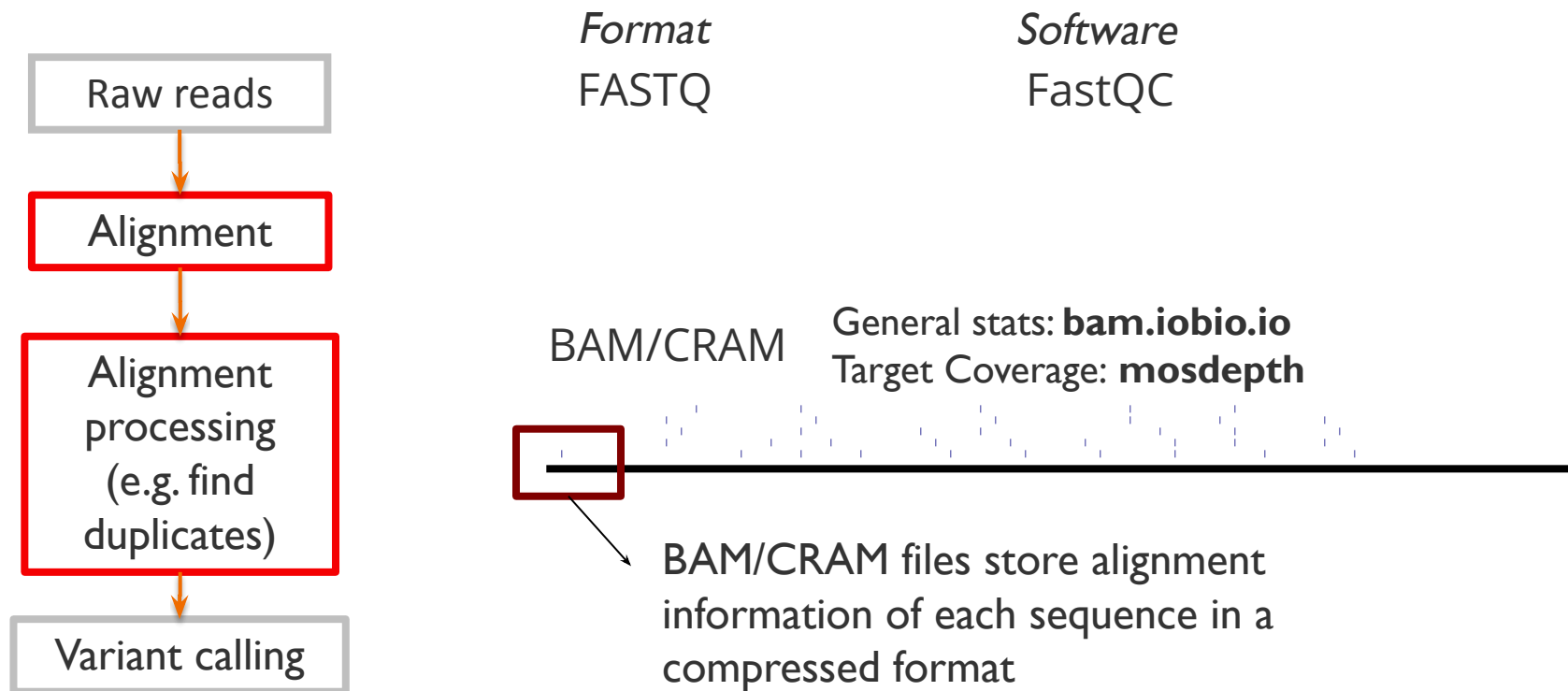
- Combine both **FastQC RawData output** into one report using **MultiQC**

General Statistics

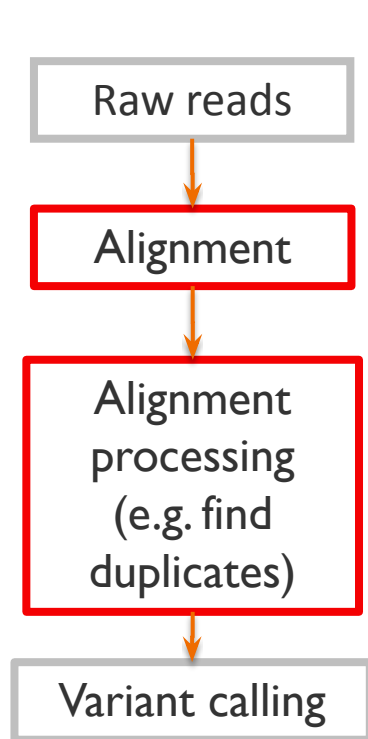
 Copy table	 Configure Columns	 Plot	Showing 2/2 rows and 3/5 columns.		
Sample Name	% Dups	% GC	M Seqs		
HighQuality_Reads_fastq_gz	23.3%	44%	1.8		
LowQuality_Reads_fastq_gz	0.0%	41%	0.0		



BAM/CRAM quality control



BAM/CRAM quality control



Format
FASTQ

Software
FastQC

BAM/CRAM General stats: **bam.iobio.io**
Target Coverage: **mosdepth**

Tool	bam.iobio.io
Input	BAM
Output	Interactive plot

Tool	mosdepth
Input	BAM, BED with target regions, thresholds
Output	Tabular files

bam.iobio.io

44: father.dedup.bam

43: mother.dedup.bam

42: proband.dedup.bam

Add Tags

402.3 MB

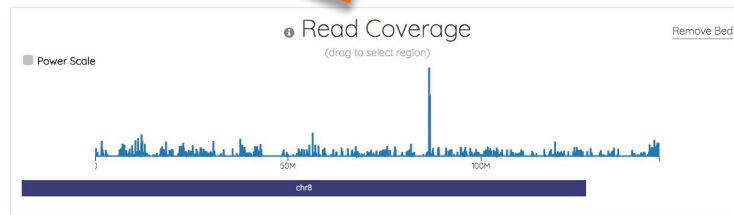
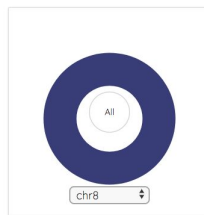
format **bam**, database **hg19**

uploaded bam file

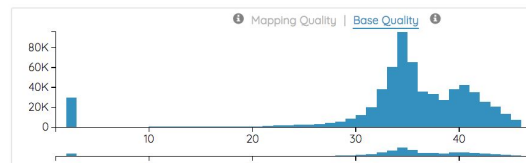
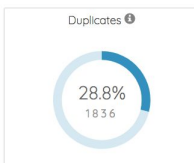
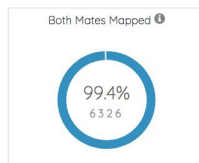
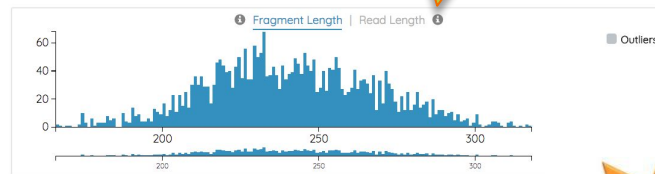
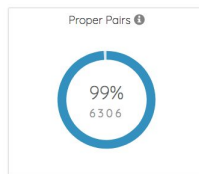
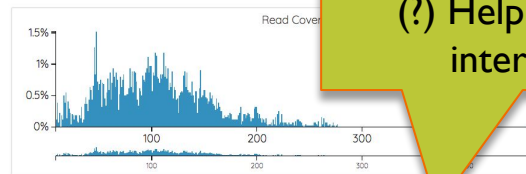
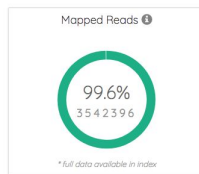
Binary bam alignments file

Select exonic regions
(GrcH37)

Click to increase the number
of sampled reads



Reads
Sampled
6
thousand



(?) Help for plot/stat
interpretation

In Galaxy, click on
Visualize
than
display at bam.iobio

LET'S
PRACTICE

#dataset 3

mosdepth compute coverage of target regions

Tool	mosdepth
Input	BAM, BED with target regions, thresholds
Output	Tabular files

This tool can be used to identify exons with low coverage

Example of output file:

#chrom	start	end	region	20X	30X
chr18	52895455	52895592	NM_001083962.2_cds_18_0_chr18_52895456_r	137	137
chr18	52896077	52896307	NM_001083962.2_cds_17_0_chr18_52896078_r	230	230

Number of bases covered = or > threshold

If we compute the fraction, e.g. $c5/(c3-c2)$:

start	end	region	20X	30X	frac20	frac30
52895455	52895592	NM_001083962.2_cds_18_0_chr18_52895456_r	137	137	1.0	1.0
52896077	52896307	NM_001083962.2_cds_17_0_chr18_52896078_r	230	230	1.0	1.0

Fraction of region = or > threshold

mosdepth compute coverage of target regions

- Input BAM/CRAM: **Panel_alignment.bam**
- Compute depth by region: **Compute depth in regions specified by a BED file**
- BED file specifying regions: **BED file with exons of your gene [*]**
- Advanced options:
 - Specify thresholds for output when using region output: **20,30**

The number of bases in each region covered = or > the thresholds is in ***mosdepth thresholds BED*** output file. To convert the absolute number of bases into fractions:

- run **Compute on rows** using the expressions
 - $c5/(c3-c2)$
 - $c6/(c3-c2)$



[*] BED file can be obtained from UCSC Table Browser (next slide)

How to retrieve exon regions in Galaxy

- Get data -> UCSC Main Table Browser
 - assembly: *the same used for Panel_alignment.bam*
 - group: **Genes and Gene Predictions**
 - track: **NCBI RefSeq**
 - table: **RefSeq Select and MANE**
 - region: **genome**
 - identifiers: *enter a gene of the panel [*]*
 - output format: **BED**
 - Send output to: **Galaxy**
 - Create one BED record per: **Coding Exons**

[*] ATRX, CDKL5, CNTNAP2, FOXG1, MECP2, MEF2C, NRXN1, SLC9A6, TCF4, UBE3A, ZEB2



mosdepth compute coverage of target regions

- Input BAM/CRAM: **Panel_alignment.bam**
- Compute depth by region: **Compute depth in regions specified by a BED file**
- BED file specifying regions: **BED file with exons of your gene**
- Advanced options:
 - Specify thresholds for output when using region output: **20,30**

The number of bases in each region covered = or > the thresholds is in **mosdepth thresholds BED** output file. To convert the absolute number of bases into fractions:

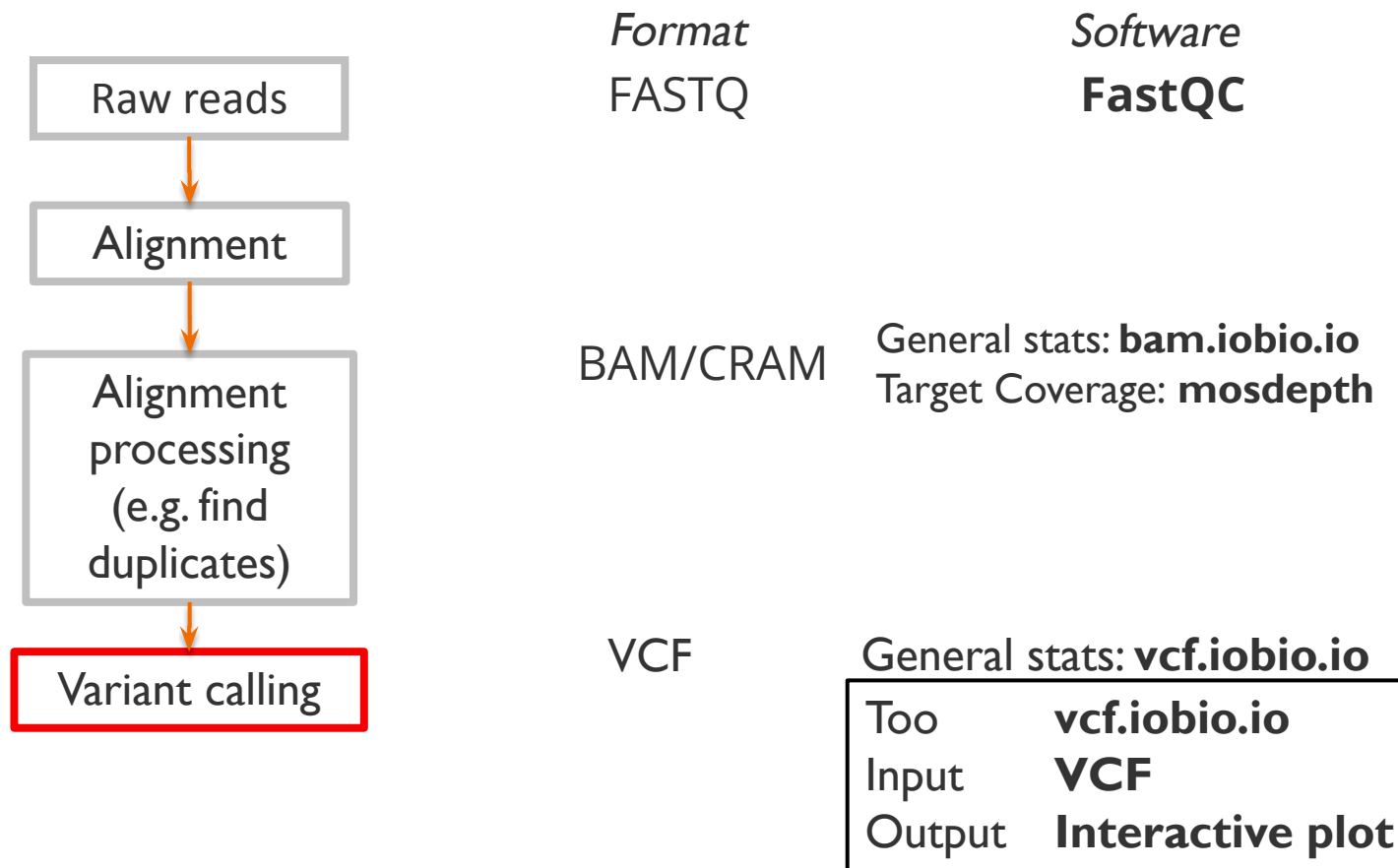
- run **Compute on rows** using the expressions
 - $c5/(c3-c2)$
 - $c6/(c3-c2)$



#dataset 2

Question: *do you have poorly covered exons in your gene? Are we at risk of missing Patho/Likely Patho variants (e.g. in NRXN1) according to ClinVar?*

VCF quality control



vcf.iobio.io

In Galaxy, run
vcf.iobio

vcf.iobio.io

http://sigu-training.cineca.it/display_application/2fd1d72d379aa820/iobio_vcf/vcf_iobio/0aa34fac8daa4bd8/data/galaxy_2fd1d72d379aa820.vcf.gz

GRCh37 All References

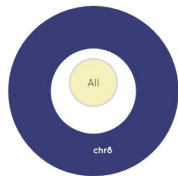
Filter samples

father mother proband

Click to increase the
number of sampled reads

34 thousand [↑]
variants sampled

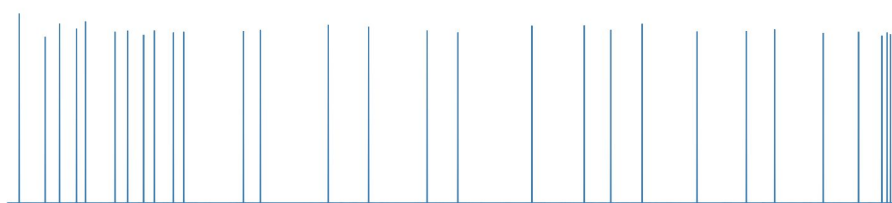
References ⓘ



Variant Density ⓘ

(click bottom chart to select a reference)

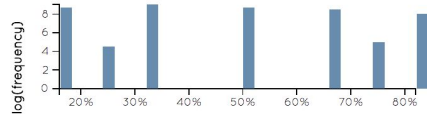
☒ GRCh37 exonic regions



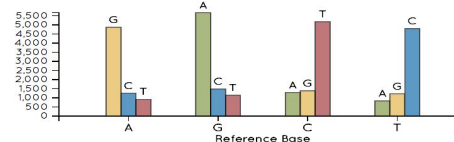
Ts/Tv Ratio ⓘ



Allele Frequency Spectrum ⓘ



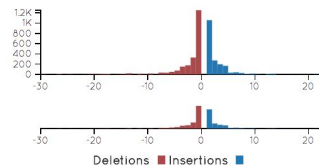
Base Changes ⓘ



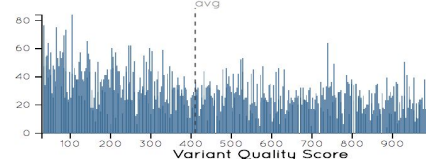
Variant Types ⓘ



Insertion & Deletion Lengths ⓘ ☒ Outliers



Variant Quality ⓘ



LET'S
PRACTICE

Commonly used thresholds

FASTQ

- Most bases above Q30 (FastQC)

BAM

- % un-mapped reads < 2% (bam.iobio.io)
- % duplicated reads ~15% (bam.iobio.io)

VCF

- Ts/Tv ratio > 2 (vcf.iobio.io)

Check if our samples pass these thresholds



Summary

- In-depth quality control of raw sequencing reads (issues with DNA quality, library construction, pooling): **FastQC**
- Un-mapped reads (contamination), duplication (library construction), target coverage (test accuracy): **bam.iobio**, **mosdepth**
- Quality of variant calls (false positives): **vcf.iobio**

Prossimo argomento

16/01/2024

UCSC: Visualizzazione dati, utilizzo di custom track, database, table browser

Marta Rusmini