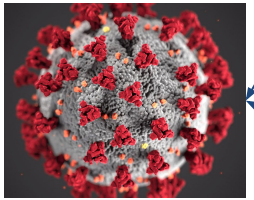
Introduction to viral genome reconstruction using massive sequencing: Sars-cov-2 use case

BU-ISCIII
Bioinformatics Unit, Institute of Health Carlos III
Madrid, Spain

Máster Virología - UCM

November 25 2020

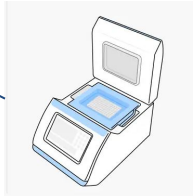
Background



December 2019

SARS-Cov-2

Surveillance:
Genome Sequencing



Detection:
RT-PCR

- RT-PCR Primer & Probe efficiency
- Viral variability behaviour
- Correlate viral lineage & severity
- Help vaccine development
- Future established disease control (flu)
- Coordinated & rapid response
- Understand immune response to infection

Sequencing Approaches

Amplicon-based enrichment

Probe capture enrichment

Wet-lab

ONT long reads

Illumina short reads

Dry-lab

ARTIC network

nf-core/viralrecon

Assembly and intrahost/low-frequency variant calling for viral samples

amplicon assembly covid-19 covid19 illumina metagenomics sars-cov-2 variant-calling viral virus

Data (TB)

State of the art

January 22 2020 Artic Network protocols

February 21 2020 Galaxy SARS-Cov2 project

https://github.com/BU-ISCIII/SARS_Cov2_consensus-nf

March 18 2020 Our 1st pipeline started

March 30 2020 nf-core collaboration

April 5 2020 1st COVID19 Virtual BioHackathon

nf-core/viralrecon

Assembly and intrahost/low-frequency variant calling for viral samples

June 1 2020 viralrecon released v1.0.0

https://github.com/jaleezyy/covid-19-signal

June 29 2020 covid-19-signal v1.0.0

https://github.com/nodrogluap/nanostripper

September 3 2020 nanostripper v1.0.0

Viralrecon

nf-core/viralrecon

https://github.com/nf-core/viralrecon

nextflow

Multiple compute infrastructures

Portable

Easy to install

Reproducible

Stable code

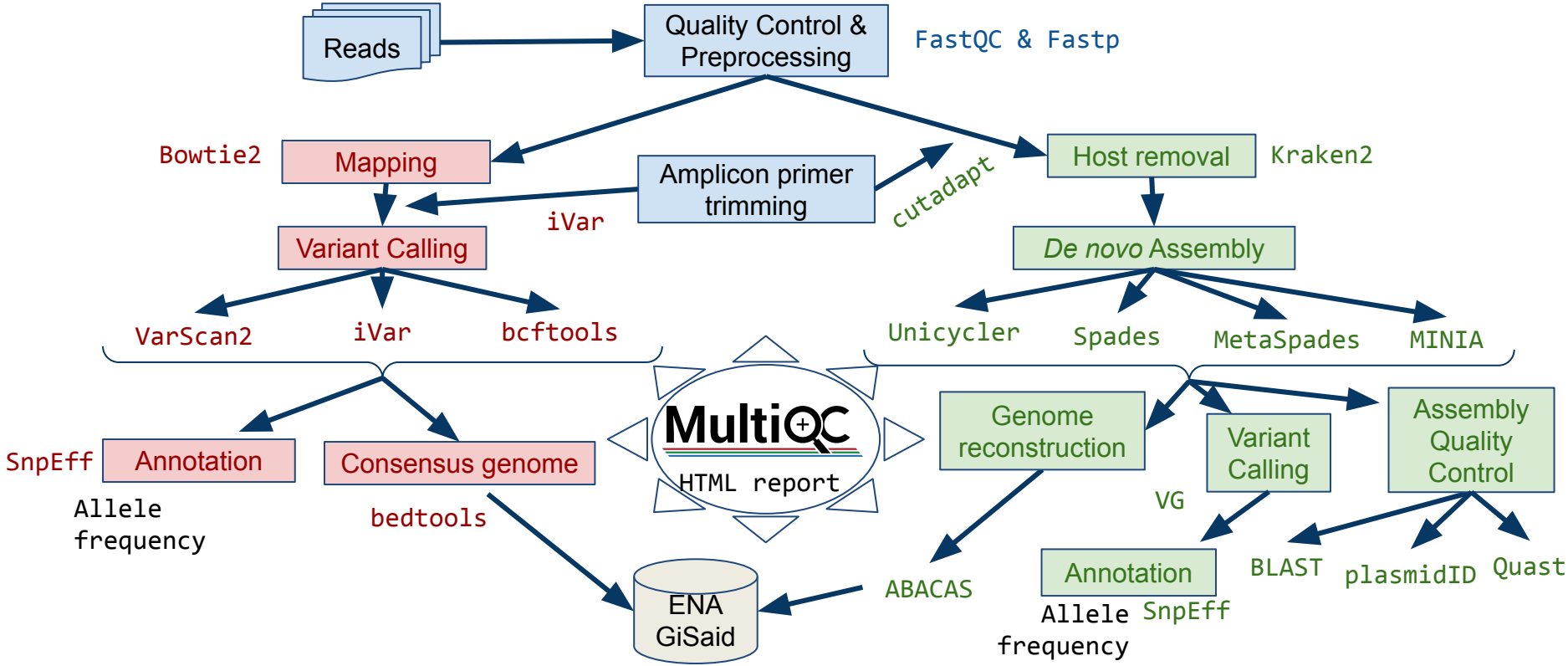
Step in the process

Preprocessing

Mapping

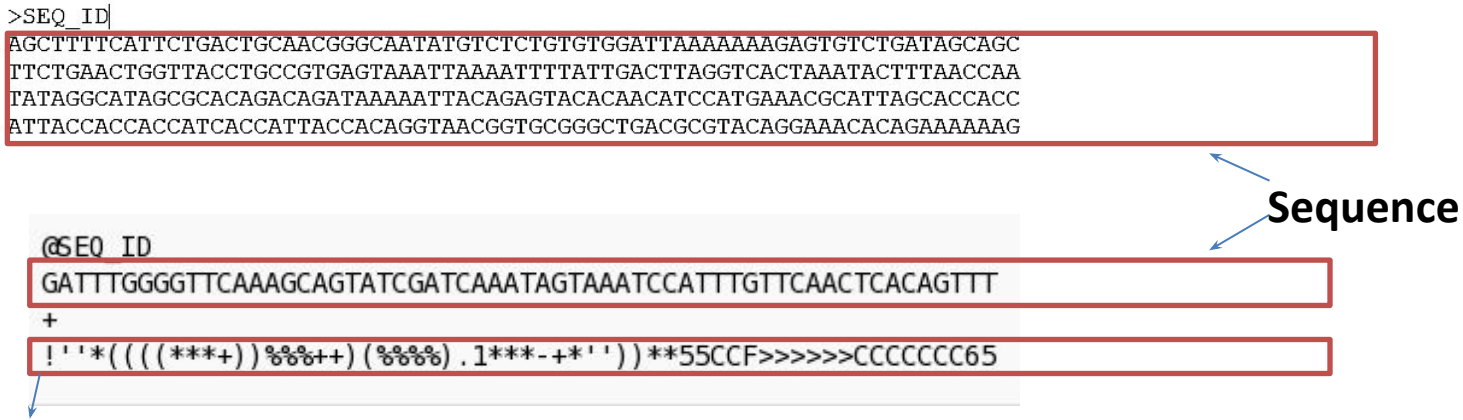
Postprocessing:
Variant calling

Viralrecon


Reads
Quality Control & Preprocessing
FastQC & Fastp
Bowtie2
Mapping
Amplicon primer trimming
cutadapt
Host removal
Kraken2
iVar
Variant Calling
De novo Assembly
VarScan2
iVar
bcftools
Unicycler
Spades
MetaSpades
MINIA
MultiQC
HTML report
SnpEff
Annotation
Consensus genome
Allele frequency
bedtools
Genome reconstruction
Variant Calling
Assembly Quality Control
VG
ENA GiSaid
ABACAS
Annotation
Allele frequency
SnpEff
BLAST
plasmidID
Quast

FASTQ format

- Is a FASTA file with quality information
- Within HTS, FASTA contain genomes y FASTQ reads



Quality: must be 1 bit

Sequencing quality assessment

- To assess quality, software uses **Phred per-base quality** score is used
- Is the **first quality control step** after sequencing. There should be one after every step of the analysis
- After quality assessment user can know how **reliable** are their datasets
- QC will determine the next **filtering** step
- Filtering decisions will **impact** directly in **further analysis**
- Many other steps also use this quality as variable in their **algorithms**

Sequencing quality assessment: Artifacts

HTS methods are bounded by their technical and theoretical limitations and sequencing errors cannot be completely eliminated (Hadigol M, Khiabanian H. 2018)

- **Artifacts in library preparation**
  - Remaining adapters
  - High rate of duplicates
  - GC regions bias
  - Polymerase error rate
  - DNA damage during breakdown
- **Artifacts during sequencing**
  - Low quality in sequence ends(Phasing: cluster loose sync)
  - Complication in certain regions:
    - Repetitions
    - Homopolymers
    - High CG content

FastQC: Basic Statistics

- Self defined overall stats
  - Encoding: Phred33 or Phred64

Basic Statistics

Measure	Value
Filename	bad_sequence.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	395288
Sequences flagged as poor quality	0
Sequence length	40
%GC	47

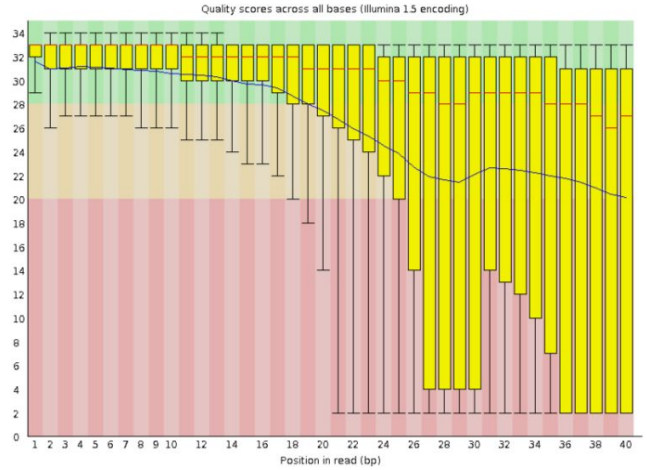
Basic Statistics

Measure	Value
Filename	good_sequence_short.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Sequences flagged as poor quality	0
Sequence length	40
%GC	45

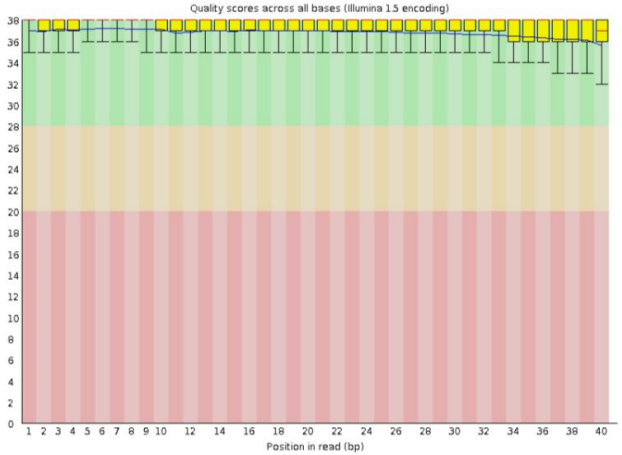
FastQC: Per base sequence quality

- Overview of the range of quality values across all bases at each position in the FastQ file
  - Median, inter-quartile range (25-75%), 10-90% points, mean quality

Per base sequence quality



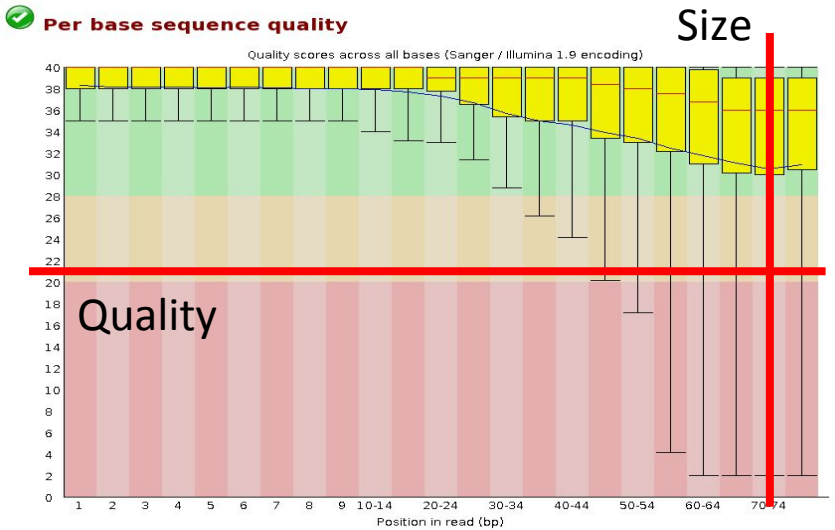
Per base sequence quality



Sequence filtering

- **Remove residual adapters**
  - Depending on used library

- **Filtering parameters**
  - Quality filtering
    - Overall mean quality
    - Local mean quality
      - Sequence end
      - Sliding window
  - Size filtering
    - Overall sequence size
    - Remaining sequence size after filtering



Mapping

Mapping software looks for the best match for each read in the genome.

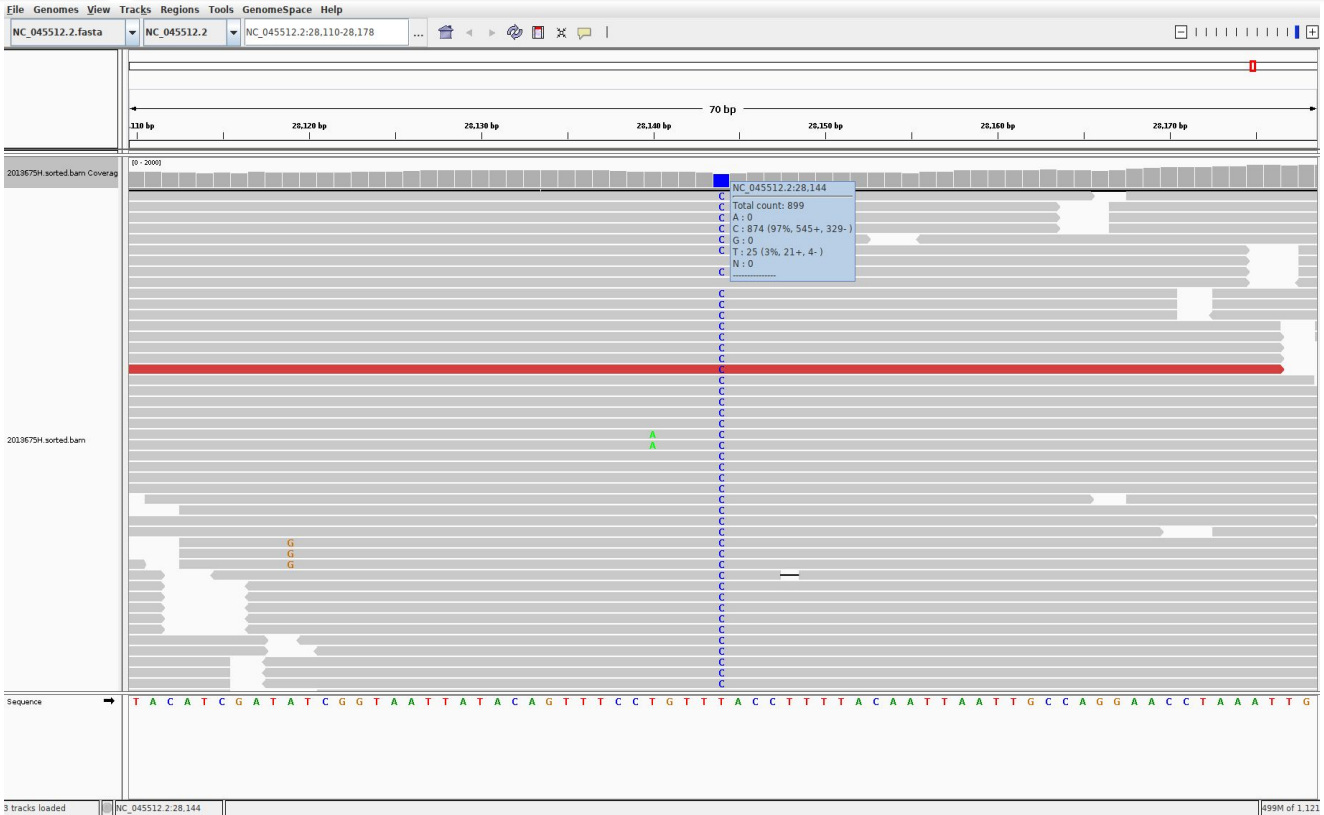
Paired-end reads helps the mapper to find the perfect spot!

SAM format

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,255}`	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0,2^{31}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next read
8	PNEXT	Int	$[0,2^{31}-1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31}+1,2^{31}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001    99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002     0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA     *
r003     0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA        * SA:Z:ref,29,-,6H5M,17,0;
r004     0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC        *
r003  2064 ref 29 17 6H5M       *  0   0 TAGGC              * SA:Z:ref,9,+,5S6M,30,1;
r001   147 ref 37 30 9M         =  7 -39 CAGCGGCAT          * NM:i:1
```

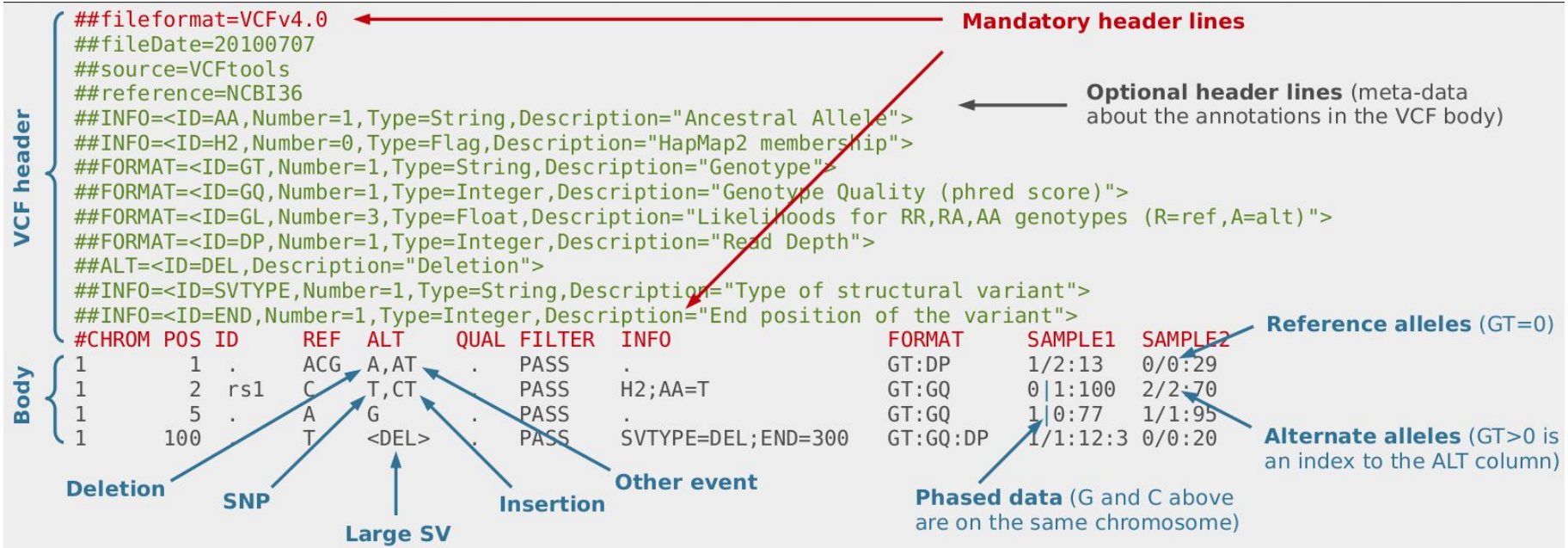
Variant calling



Variant callers walk through each alignment column independently and evaluate if there is a variant.

A variant is called in haploid genomes when it’s supported by at least 90% of the reads.

VCF format


VCF header
Body
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2
1 1 . ACG A,AT . PASS . GT:DP 1/2:13 0/0:29
1 2 rs1 C T,CT . PASS H2;AA=T GT:GQ 0|1:100 2/2:70
1 5 . A G . PASS . GT:GQ 1|0:77 1/1:95
1 100 . T <DEL> . PASS SVTYPE=DEL;END=300 GT:GQ:DP 1/1:12:3 0/0:20
Mandatory header lines
Optional header lines (meta-data about the annotations in the VCF body)
Reference alleles (GT=0)
Alternate alleles (GT>0 is an index to the ALT column)
Deletion
SNP
Large SV
Insertion
Other event
Phased data (G and C above are on the same chromosome)

Depth of coverage vs coverage

Depth of coverage

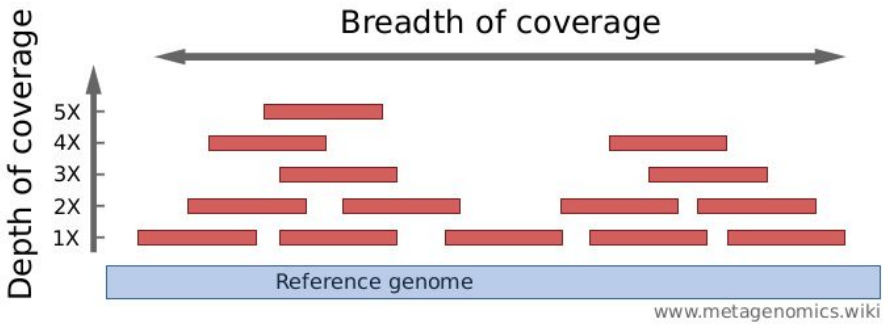
How strong is a genome "covered" by sequenced fragments (short reads)?

Per-base coverage is the average number of times a base of a genome is sequenced. The coverage depth of a genome is calculated as the number of bases of all short reads that match a genome divided by the length of this genome. It is often expressed as 1X, 2X, 3X,... (1, 2, or, 3 times coverage).

Breadth of coverage

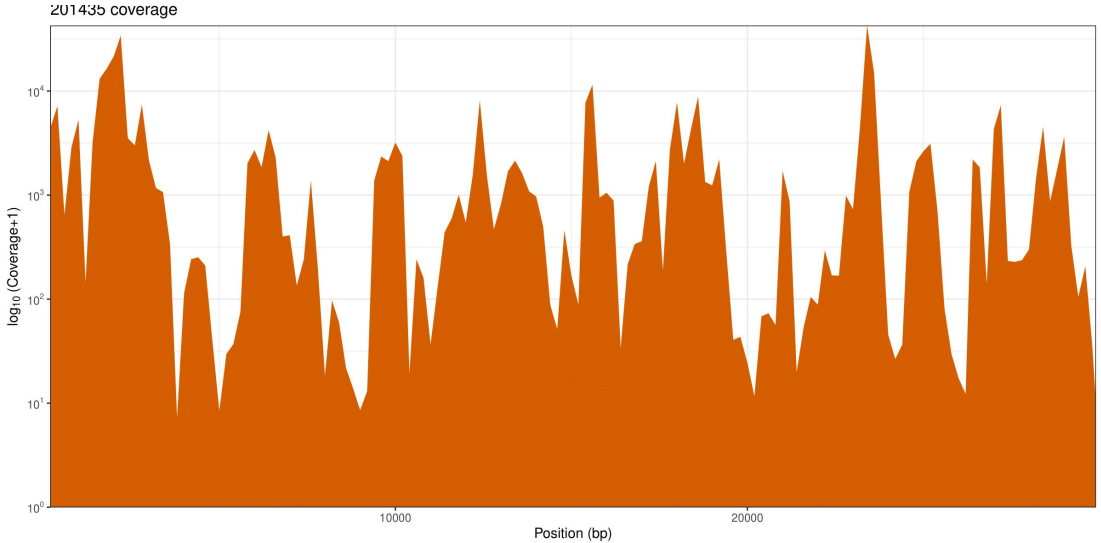
How much of a genome is "covered" by short reads? Are there regions that are not covered, even not by a single read?

Breadth of coverage is the percentage of bases of a reference genome that are covered with a certain depth. For example: 90% of a genome is covered at 1X depth; and still 70% is covered at 5X depth.

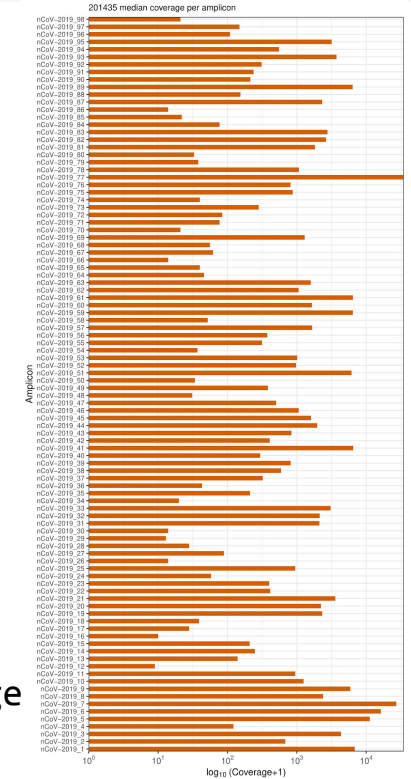


Sars-cov-2 quality control results

Genome coverage



Amplicon coverage



Assembly

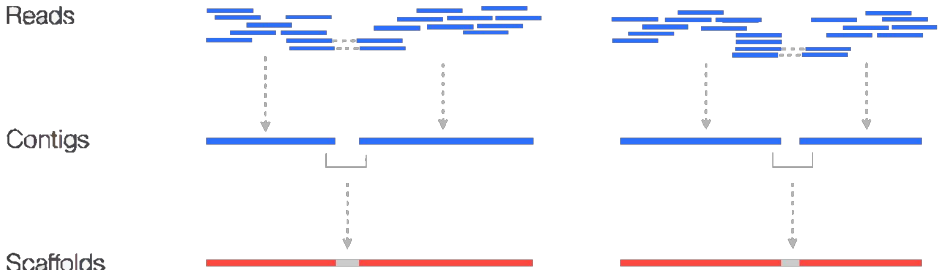
Reconstruct a representation of the original DNA from shorter DNA sequences or small fragments known as reads

- ***De novo:*** with no previous knowledge of the genome to be assembled. It overlap the end of the end of each read in order to créate a longer sequence.

- ***Assembly with reference:*** A similar but not identical genome guides the assembly process. Map reads over supplied genome.

Assembly: contig y scaffold

- **Contig:** continuous sequence made up of overlaping shorter sequences
- **Scaffold:** two or more contigs located and rearranged according to spatial information(pair-end, mate pair, reference)



https://www.biostars.org/p/253222/

And...after all this?

```
>NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuh
complete genome
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAA
CGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAAC
TAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTG
TTGCAGCCGATCATCAGCACATCTAGGTTTCGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTC
CCTGGTTTCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTAC
GTGGCTTTGGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGG
CTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAAACGTTCGGAT
GCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAAGGCATTCAGTACGGTC
GTAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGGCGAAATACCAGTGGCTTACCGCAAGGTTCT
TCTTCGTAAGAACGGTAATAAAGGAGCTGGTGGCCATAGTTACGGCGCCGATCTAAAGTCATTTGACTTA
GGCGACGAGCTTGGCACTGATCCTTATGAAGATTTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTG
TTACCCGTGAACTCATGCGTGAGCTTAACGGAGGGGCATACACTCGCTATGTCGATAACAACTTCTGTGG
CCCTGATGGCTACCCTCTTGAGTGCATTAAAGACCTTCTAGCACGTGCTGGTAAAGCTTCATGCACTTTG
TCCGAACAACTGGACTTTATTGACACTAAGAGGGGTGTATACTGCTGCCGTGAACATGAGCATGAAATTG
CTTGGTACACGGAACGTTCTGAAAAGAGCTATGAATTGCAGACACCTTTTGAAATTAAATTGGCAAAGAA
ATTTGACACCTTCAATGGGGAATGTCCAAATTTTGTATTTCCCTTAAATTCCATAATCAAGACTATTCAA
CCAAGGGTTGAAAAGAAAAAGCTTGATGGCTTTATGGGTAGAATTCGATCTGTCTATCCAGTTGCGTCAC
```

.fasta consensus genome

DB

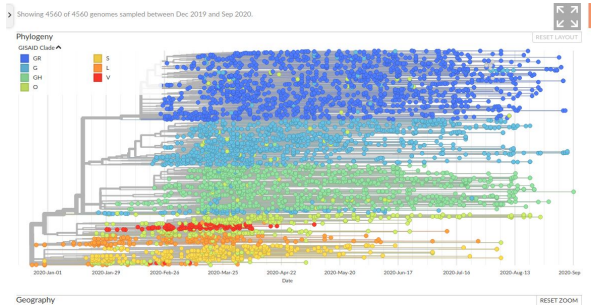
ENA

European Nucleotide Archive

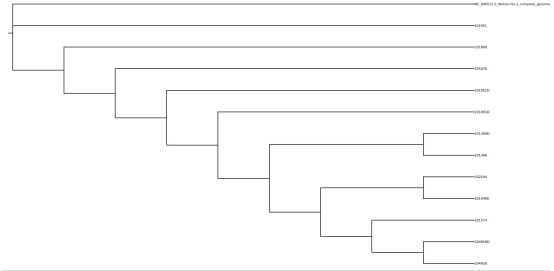
GISAID

2008 - 2020

Clade classification



Outbreak detection

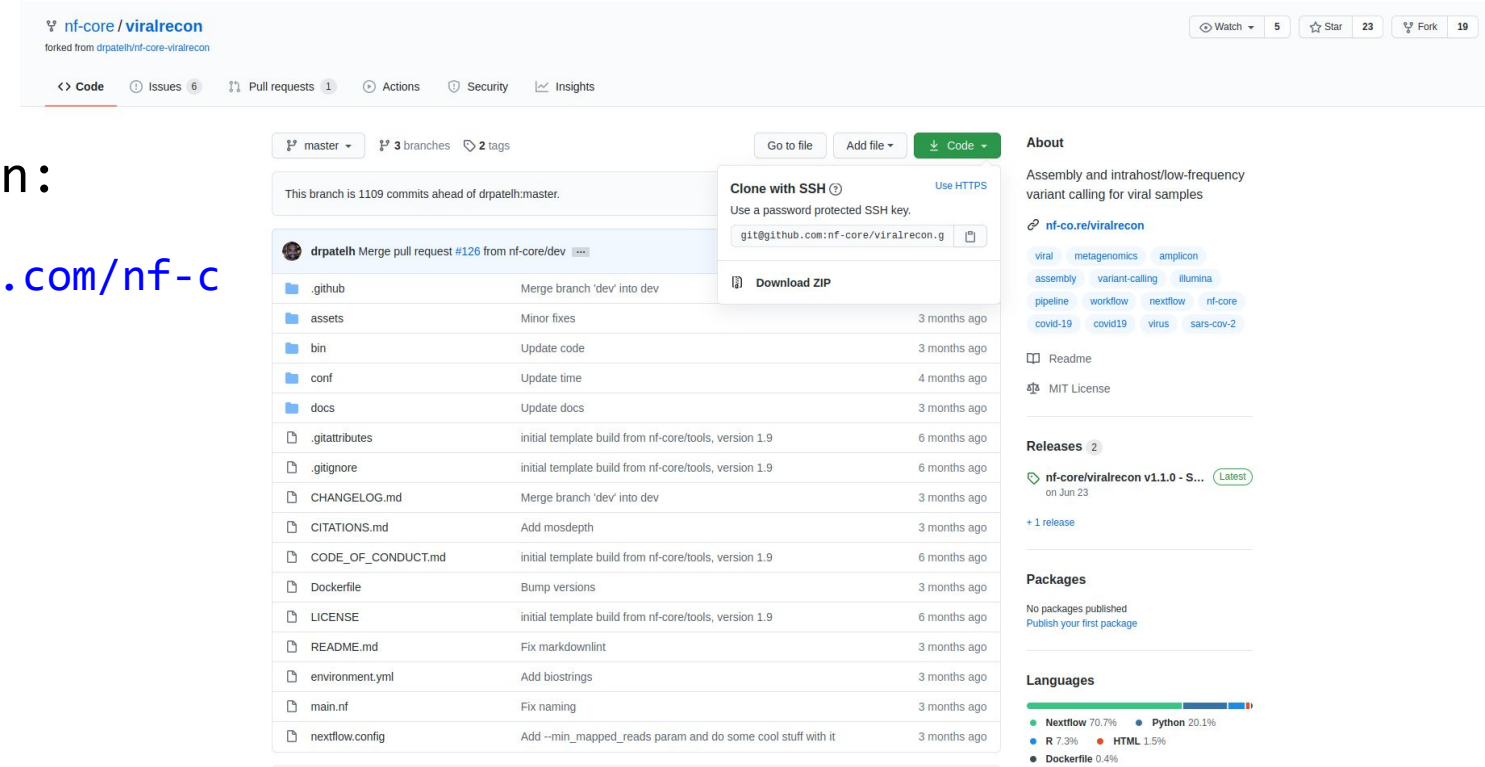


Other analyses

Conclusions

You can find viralrecon in:

https://github.com/nf-core/viralrecon



Thanks for your attention!

(Find us in https://github.com/BU-ISCIII)

Special mention to the nf-core community (https://nf-co.re/)