Conceitos sobre montagem e anotação de variantes virais

wellcome
connecting
science

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz
Instituto Gonçalo Moniz

Túlio Campos, PhD

Instituto Aggeu Magalhães

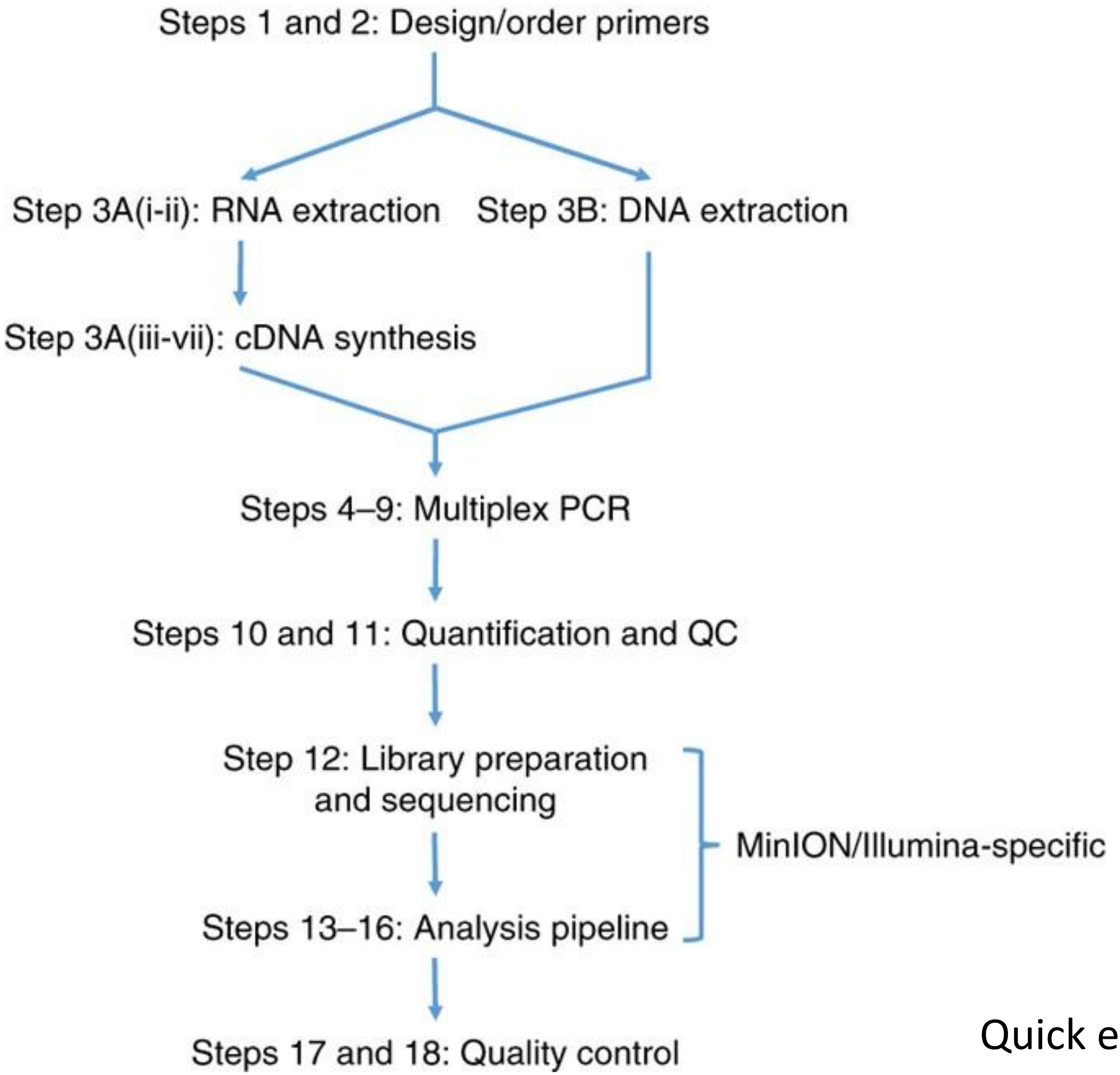
Fiocruz Pernambuco

tulio.campos@fiocruz.br

Informações contidas dos genomas

- De onde surgiu o vírus?
- Como adquiriu a capacidade de infectar humanos?
- Como definir diferentes variantes/cepas?
- Existem variantes/cepas mais infecciosas ou patogênicas?
- Como se espalhou pelo mundo (epidemiologia)?
- Como identificar epítopos para vacinas eficazes?
- Porque é que as vacinas são mais/menos eficazes para algumas variantes/cepas?

Das amostras aos genomas



Steps 1 and 2: Design/order primers

Step 3A(i-ii): RNA extraction

Step 3B: DNA extraction

Step 3A(iii-vii): cDNA synthesis

Steps 4–9: Multiplex PCR

Steps 10 and 11: Quantification and QC

Step 12: Library preparation and sequencing

Steps 13–16: Analysis pipeline

MinION/Illumina-specific

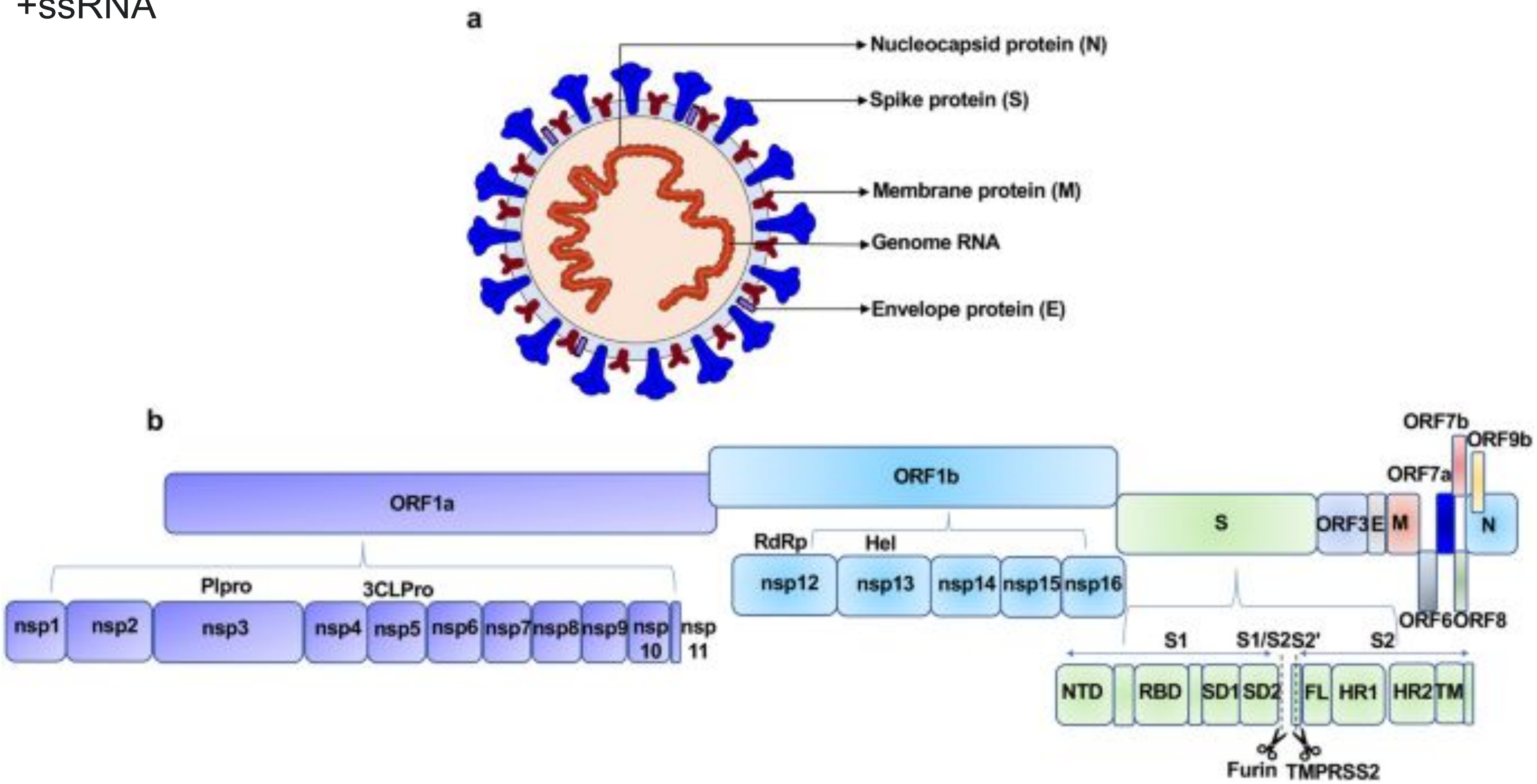
Steps 17 and 18: Quality control

Quick et al. (2017)

SARS-CoV-2 e suas características

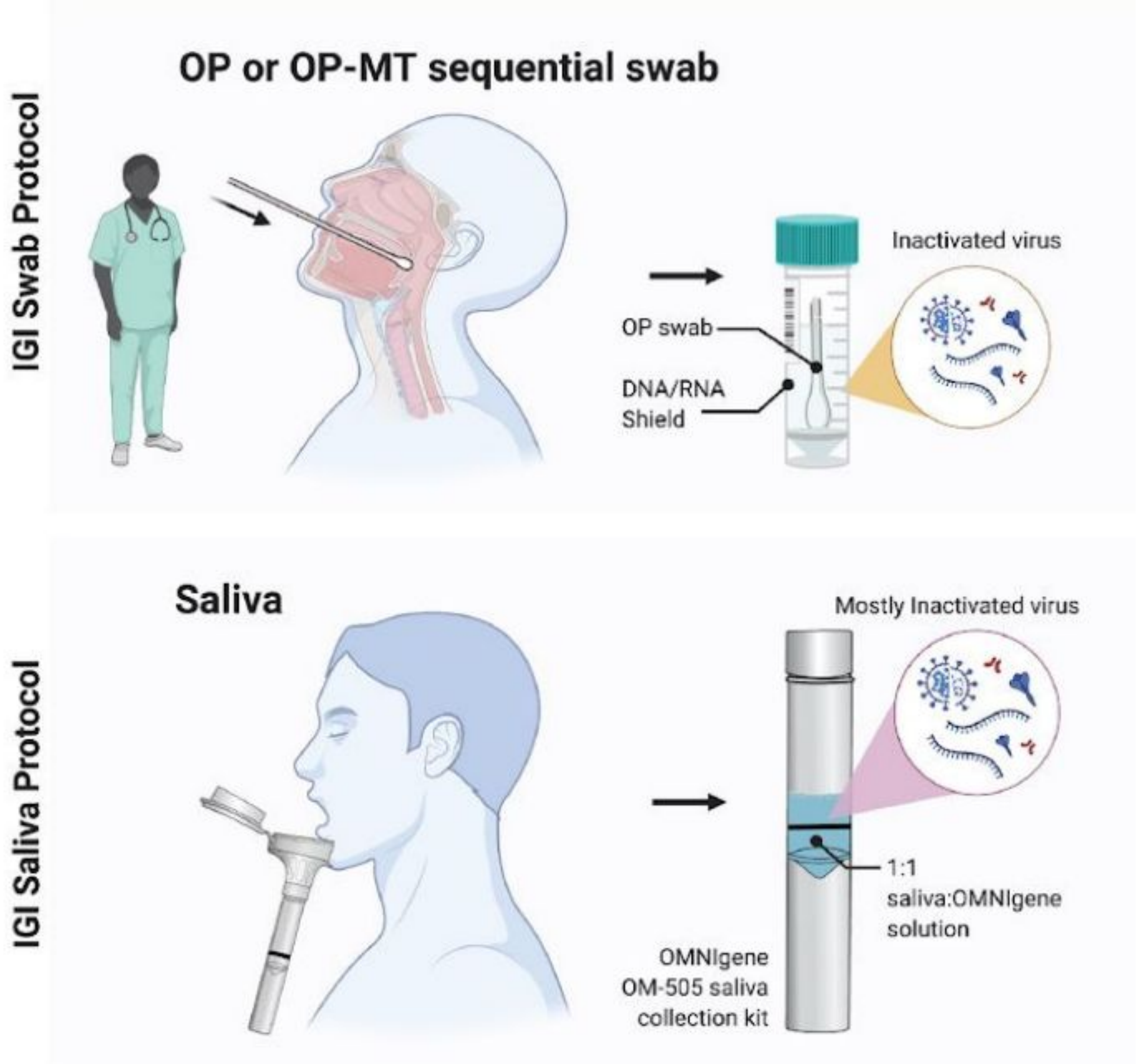
Severe **A**cute **R**espiratory **S**yndrome **Co**rona**V**irus **2**

+ssRNA



Zhang et al. (2021)

Passo 1: coleta de amostras

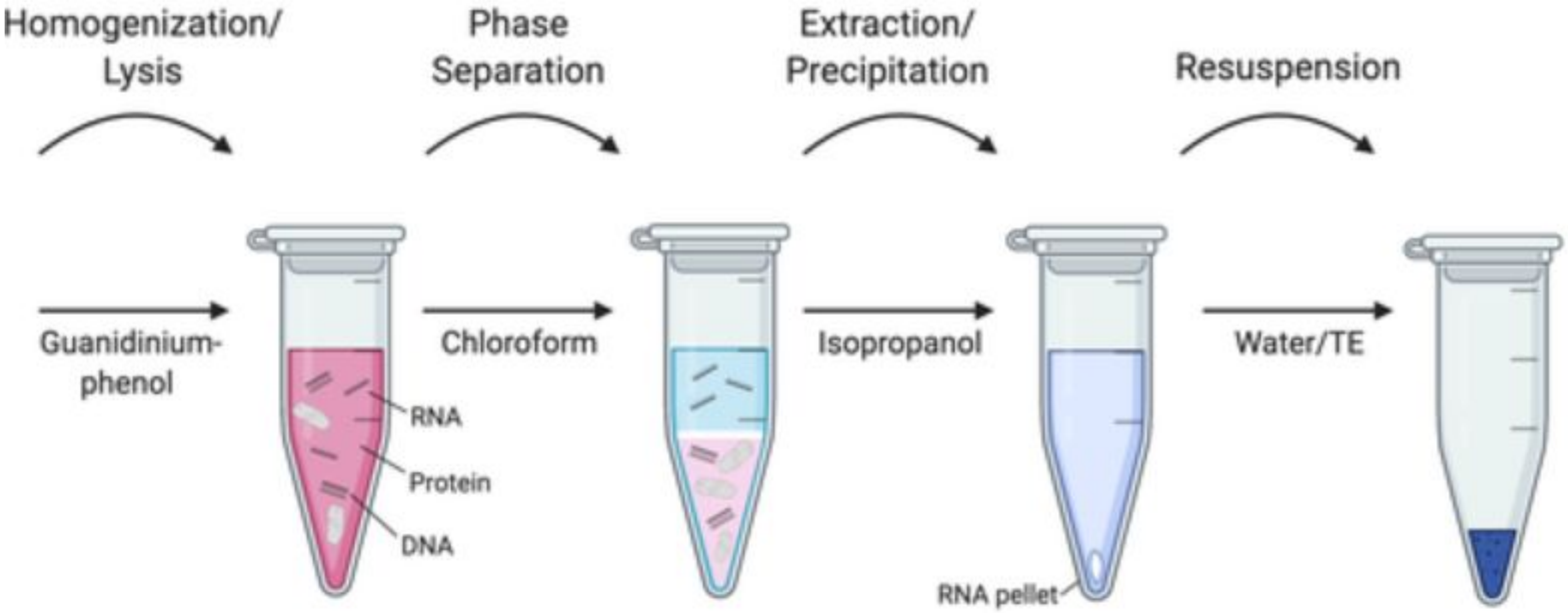


Hamilton et al. (2021)

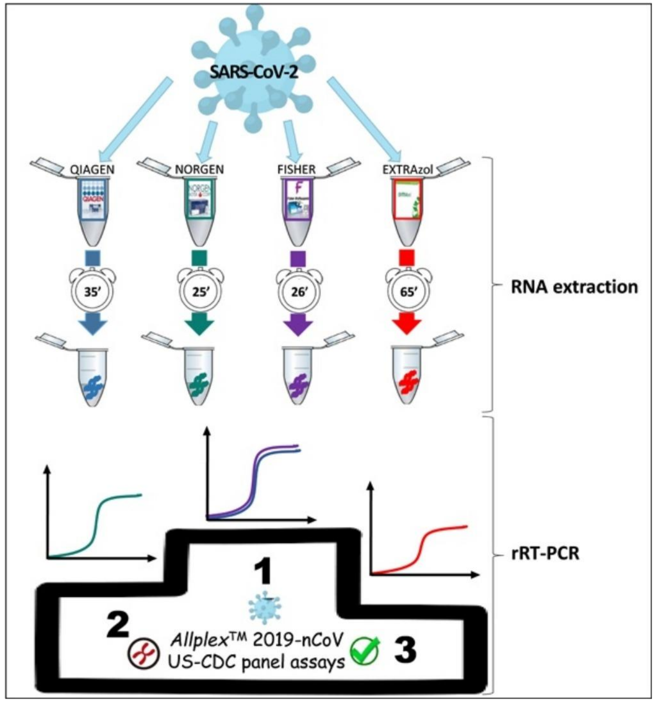
LACEN - Recife/Pernambuco

http://portal.saude.pe.gov.br

Passo 2: Extração de RNA total



Adapted from: https://www.addgene.org/protocols/kit-free-rna-extraction/

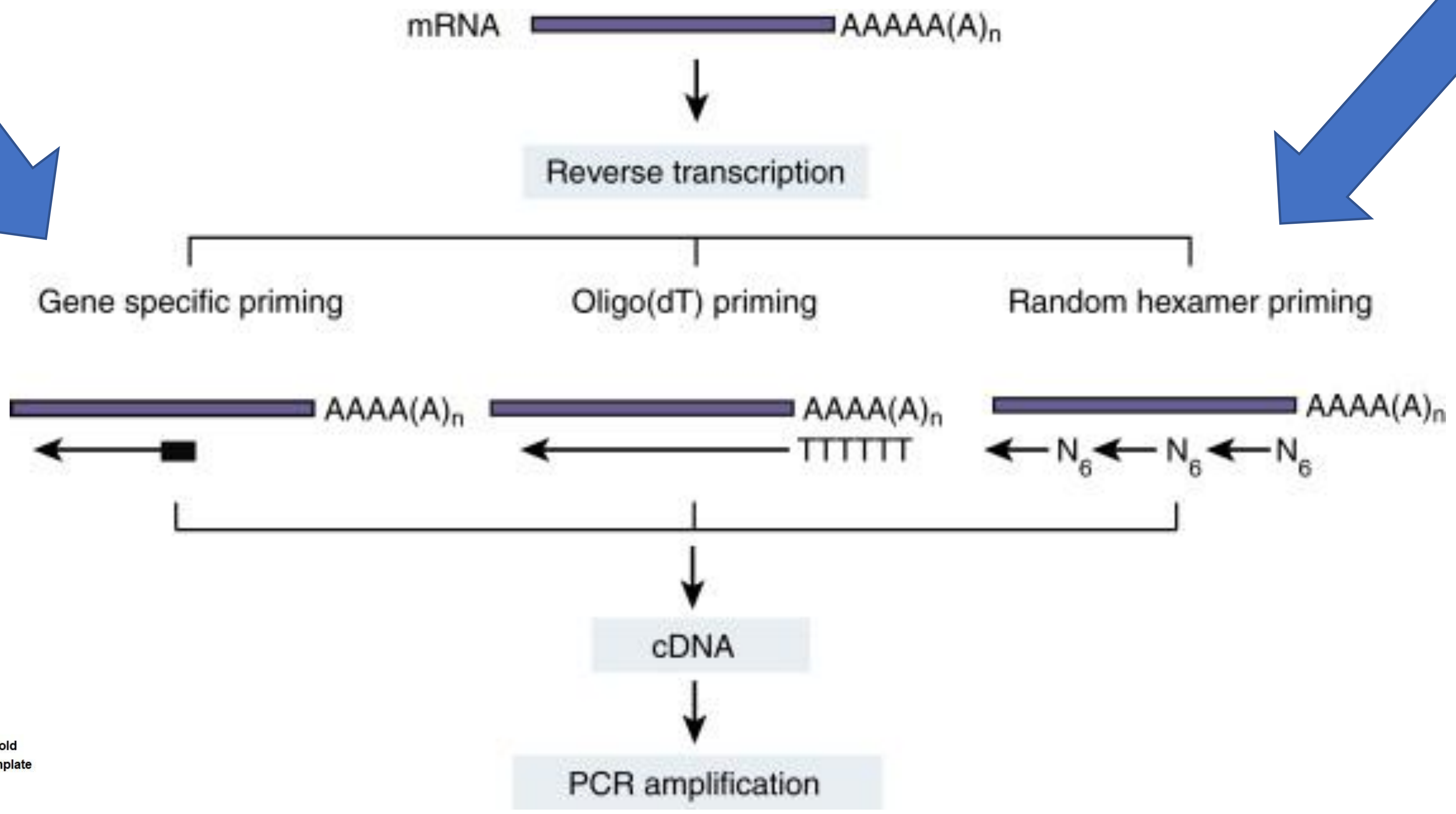


Ambrosi et al., 2021

Passo 3: rRT-PCR (quantificação e cDNA)

Viral RNA

+



mRNA AAAAA(A)$_n$

Reverse transcription

Gene specific priming

Oligo(dT) priming

Random hexamer priming

AAAA(A)$_n$

AAAA(A)$_n$

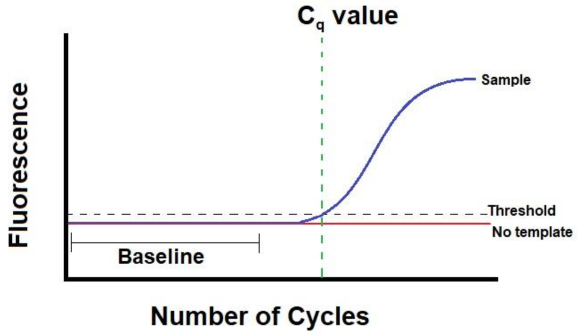
TTTTTT

AAAA(A)$_n$

N_6 N_6 N_6

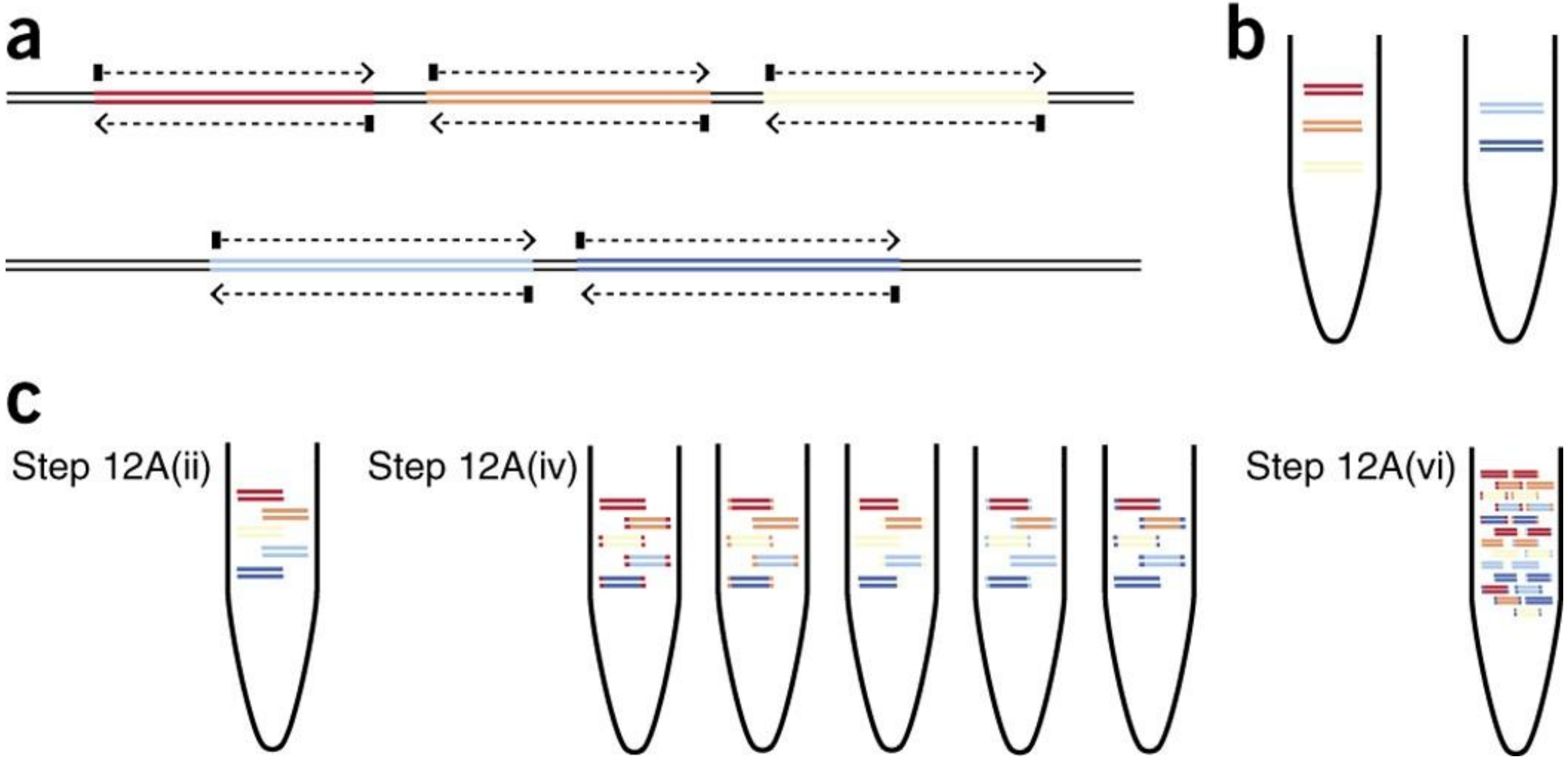
cDNA

PCR amplification



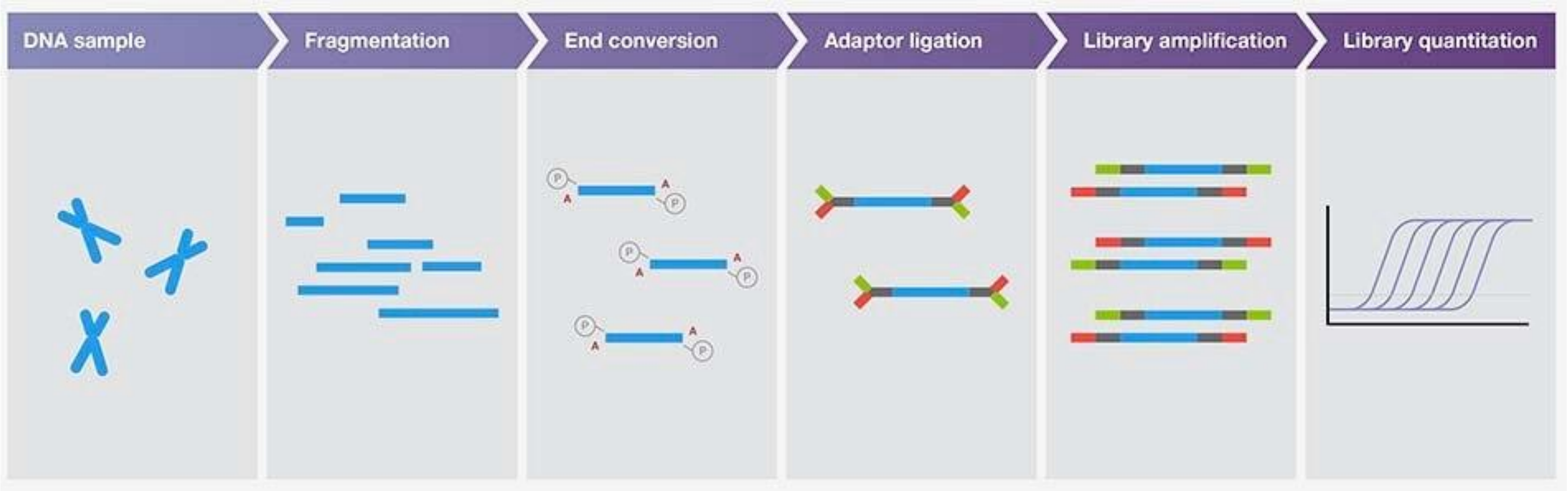
Adaptado de: https://microbeonline.com/rt-pcr-principles-applications/

Passo 4: PCR multiplex ou enriquecimento?



Quick et al. (2017)

Passo 5: preparação de biblioteca Illumina



Fonte: https://www.thermofisher.com/br/en/home/life-science/cloning/cloning-learning-center/invitrogen-school-of-molecular-biology/next-generation-sequencing/dna-sequencing-preparation-illumina.html

Nextera XT vs. COVID-seq

Passo 6: Sequenciamento Illumina

1 Library Preparation

Fragment DNA
Repair ends
Add A overhang
Ligate adapters
Purify

2 Cluster Generation

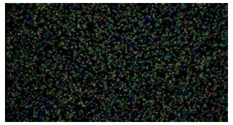
Hybridize to flow cell
Extend hybridized template
Perform bridge amplification
Prepare flow cell for sequencing

Flowcell

3 Sequencing

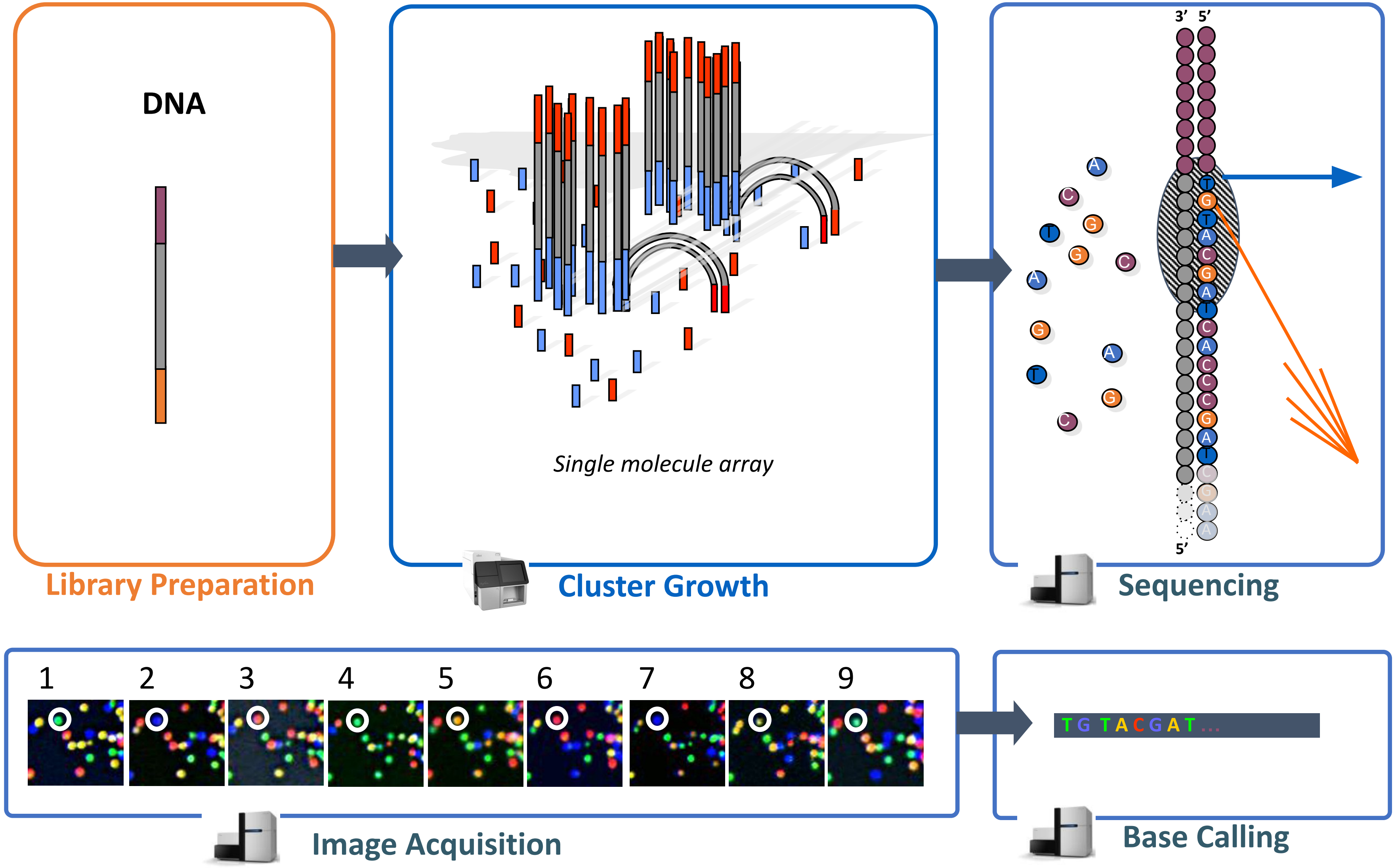
Perform sequencing
Generate base calls

4 Data Analysis



Images
Intensities
Reads
Alignments

Passo 6: Sequenciamento Illumina



DNA

Library Preparation

Single molecule array

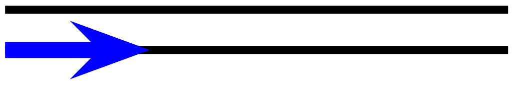
Cluster Growth

Sequencing

Image Acquisition

Base Calling

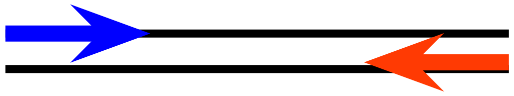
Passo 6: Illumina sequencing



Leituras single-end

Read1

ATGTTCCATAAGC…



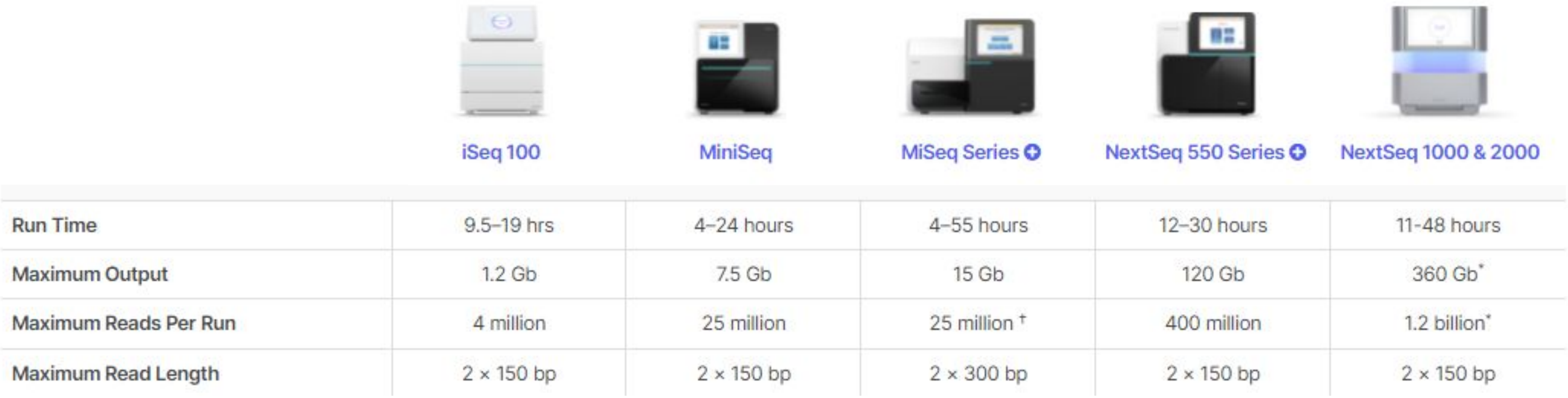
Leituras paired-end

Read1

ATGTTCCATAAGC…

Read2

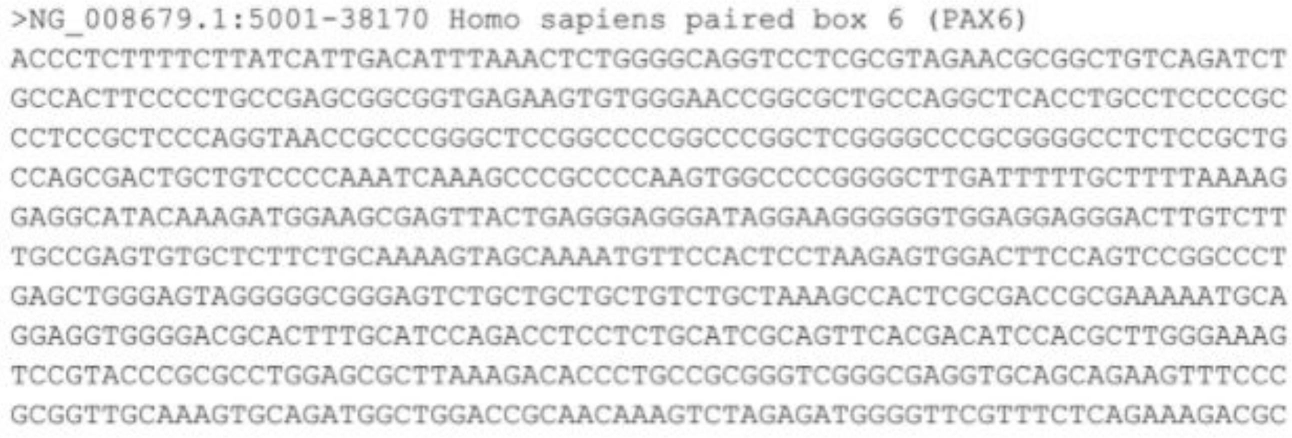
CCGTAATGGCATG…



	iSeq 100	MiniSeq	MiSeq Series ⊕	NextSeq 550 Series ⊕	NextSeq 1000 & 2000
Run Time	9.5–19 hrs	4–24 hours	4–55 hours	12–30 hours	11-48 hours
Maximum Output	1.2 Gb	7.5 Gb	15 Gb	120 Gb	360 Gb*
Maximum Reads Per Run	4 million	25 million	25 million †	400 million	1.2 billion*
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 150 bp

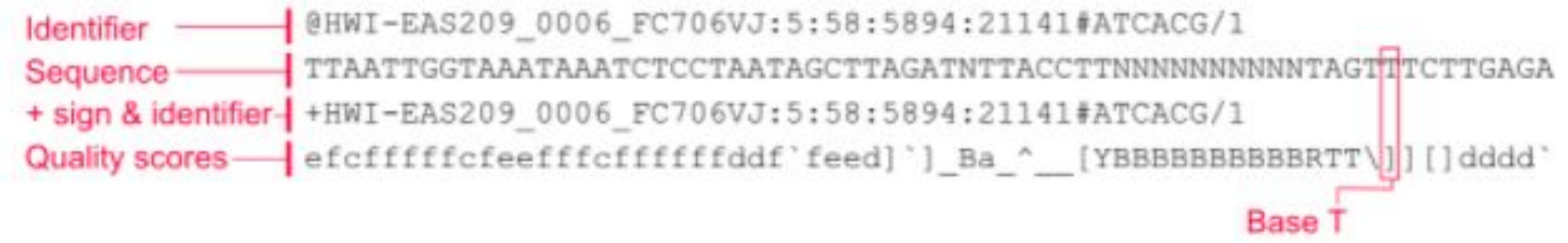
Análise de dados: FASTA/FASTQ

FASTA – genes e genomas



```
>NG_008679.1:5001-38170 Homo sapiens paired box 6 (PAX6)
ACCCTCTTTTCTTATCATTGACATTTAAACTCTGGGCAGGTCCTCGCGTAGAACGCGGCTGTCAGATCT
GCCACTTCCCCTGCCGAGCGGCGGTGAGAAGTGTGGGAACCGGCGCTGCCAGGCTCACCTGCCTCCCCGC
CCTCCGCTCCCAGGTAACCGCCCGGGCTCCGGCCCCGGCCCGGCTCGGGGCCCGCGGGGCCTCTCCGCTG
CCAGCGACTGCTGTCCCCAAATCAAAGCCCGCCCCAAGTGGCCCCGGGGCTTGATTTTTGCTTTTAAAAG
GAGGCATACAAAGATGGAAGCGAGTTACTGAGGGAGGGATAGGAAGGGGGGTGGAGGAGGGACTTGTCTT
TGCCGAGTGTGCTCTTCTGCAAAAGTAGCAAAATGTTCCACTCCTAAGAGTGGACTTCCAGTCCGGCCCT
GAGCTGGGAGTAGGGGGCGGGAGTCTGCTGCTGCTGTCTGCTAAAGCCACTCGCGACCGCGAAAAATGCA
GGAGGTGGGGACGCACTTTGCATCCAGACCTCCTCTGCATCGCAGTTCACGACATCCACGCTTGGGAAAG
TCCGTACCCGCGCCTGGAGCGCTTAAAGACACCCTGCCGCGGGTCGGGCGAGGTGCAGCAGAAGTTTCCC
GCGGTTGCAAAGTGCAGATGGCTGGACCGCAACAAAGTCTAGAGATGGGGTTCGTTTCTCAGAAAGACGC
```

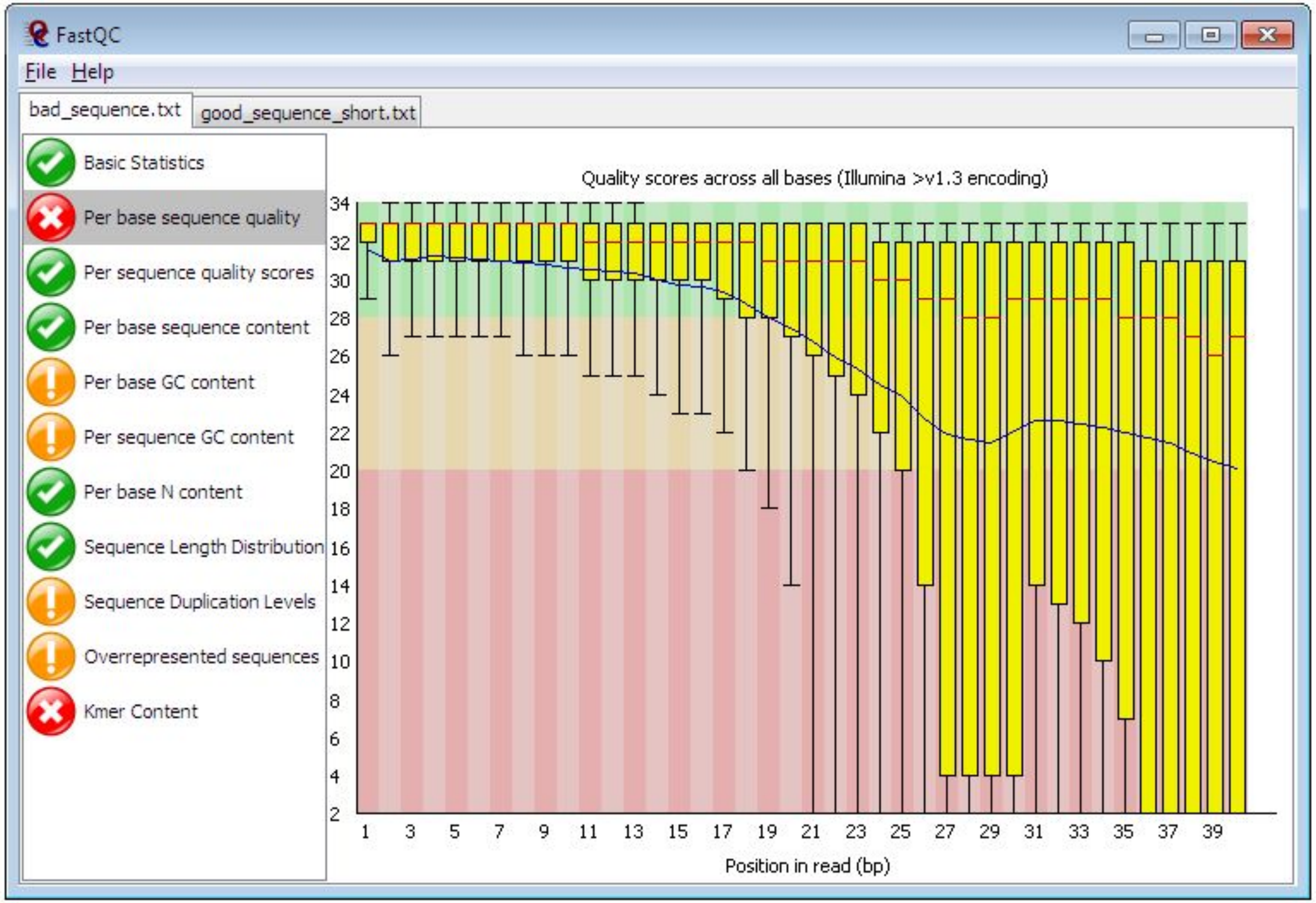
FASTQ – leituras Illumina



```
@HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
TTAATTGGTAAATAAATCTCCTAATAGCTTAGATNTTACCTTNNNNNNNNNNTAGTTTCTTGAGA
+HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
efcfffffcfeefffcffffffddf`feed]`]_Ba_^__[YBBBBBBBBBBRTT\]][]dddd`
```

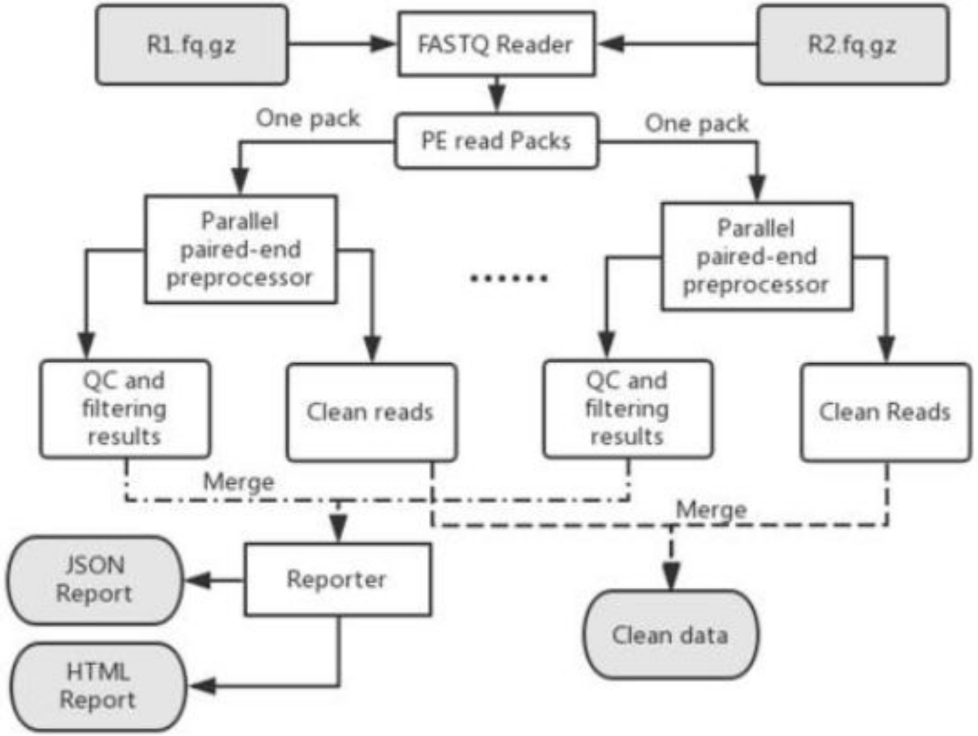
Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Análise de dados: qualidade



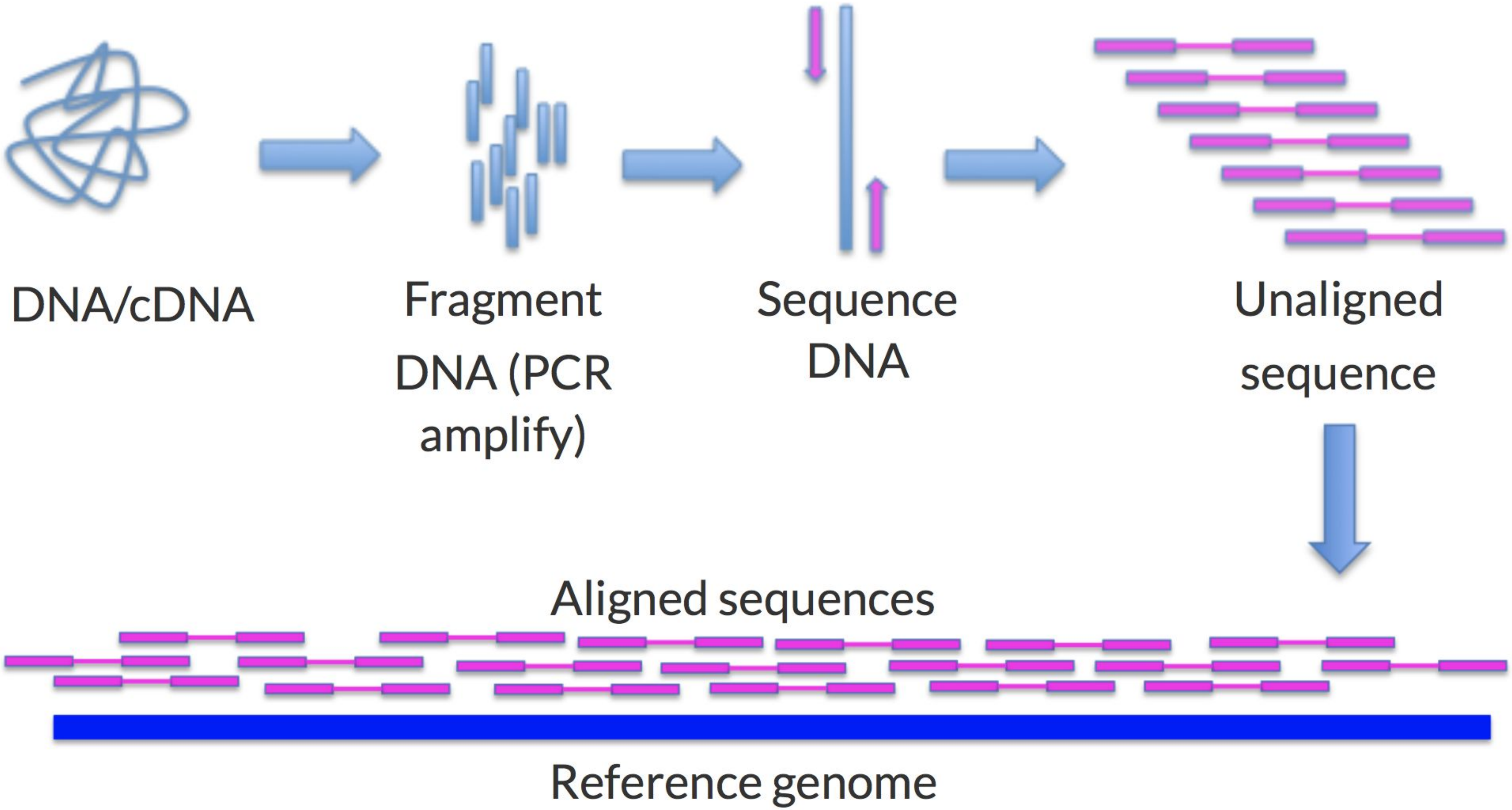
FastQC
File Help
bad_sequence.txt
good_sequence_short.txt
✓ Basic Statistics
✗ Per base sequence quality
✓ Per sequence quality scores
✓ Per base sequence content
! Per base GC content
! Per sequence GC content
✓ Per base N content
✓ Sequence Length Distribution
! Sequence Duplication Levels
! Overrepresented sequences
✗ Kmer Content
Quality scores across all bases (Illumina >v1.3 encoding)
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
Position in read (bp)

Análise de dados: filtragem/trimagem



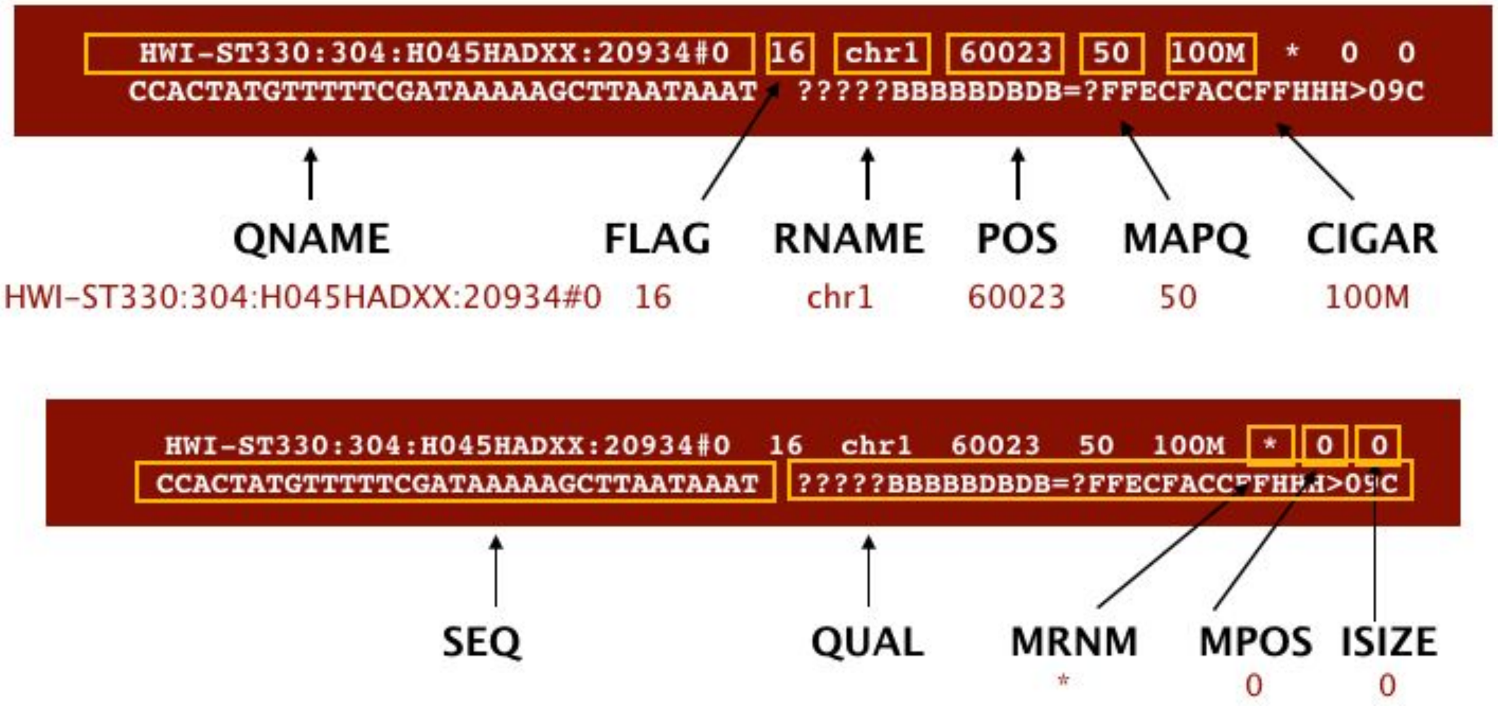
Example: fastp – Chen et al. (2018)

Análise de dados: alinhamento



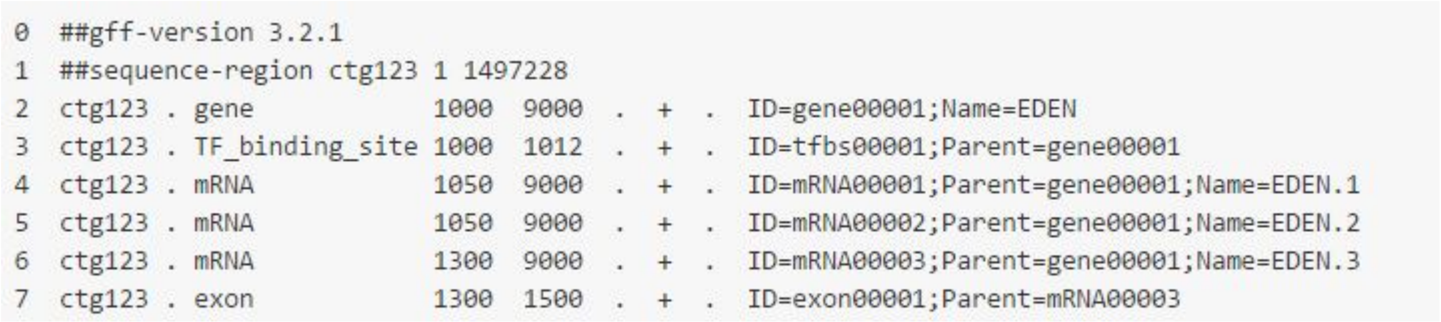
Análise de dados: SAM/BAM/GFF

SAM/BAM



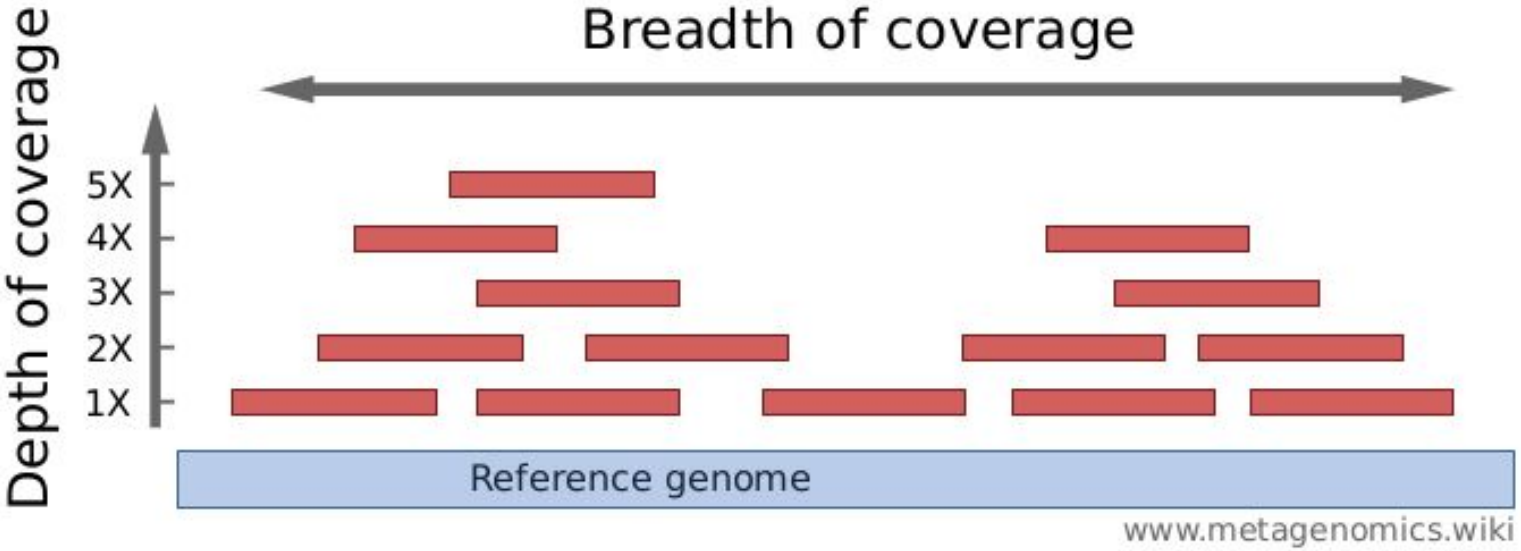
```
HWI-ST330:304:H045HADXX:20934#0 16 chr1 60023 50 100M * 0 0
CCACTATGTTTTTCGATAAAAAGCTTAATAAAT ?????BBBBBDBDB=?FFECFACCFFHHH>09C
```

GFF



```
##gff-version 3.2.1
##sequence-region ctg123 1 1497228
ctg123 . gene             1000  9000  .  +  .  ID=gene00001;Name=EDEN
ctg123 . TF_binding_site 1000  1012  .  +  .  ID=tfbs00001;Parent=gene00001
ctg123 . mRNA             1050  9000  .  +  .  ID=mRNA00001;Parent=gene00001;Name=EDEN.1
ctg123 . mRNA             1050  9000  .  +  .  ID=mRNA00002;Parent=gene00001;Name=EDEN.2
ctg123 . mRNA             1300  9000  .  +  .  ID=mRNA00003;Parent=gene00001;Name=EDEN.3
ctg123 . exon             1300  1500  .  +  .  ID=exon00001;Parent=mRNA00003
```

Análise de dados: profundidade vs. cobertura

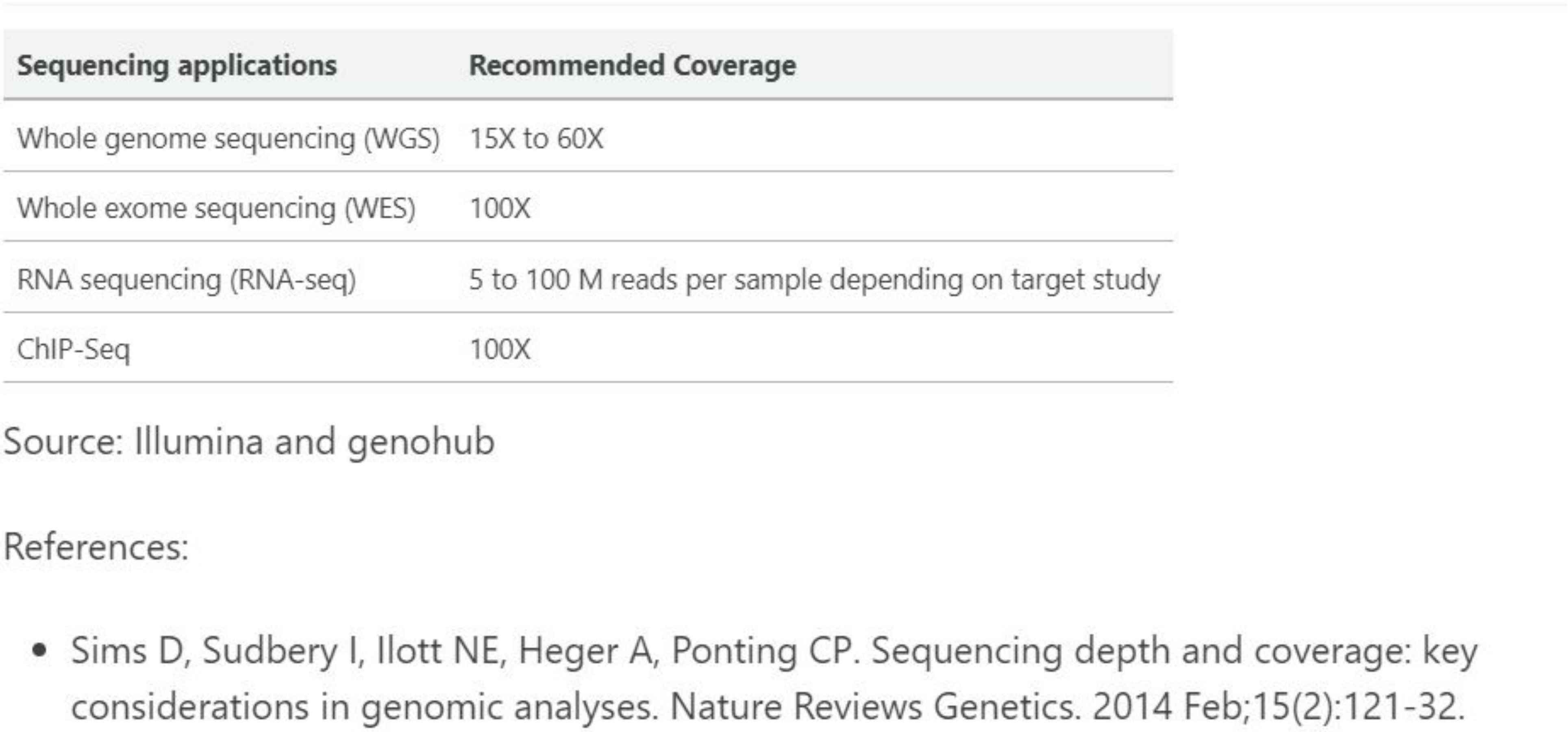


Exemplo:
Genoma: 10 Mbp
Sequenciamento: 5 milhões de leituras x 100bp = 50Mbp (sequenciados totais)

Portanto, a **profundidade** média esperada em cada posição é 5x. No entanto, isso deve ser calculado em cada posição.

Já a **cobertura** refere-se à porcentagem do genoma suportada por leituras de sequenciamento

Análise de dados: profundidade vs. cobertura



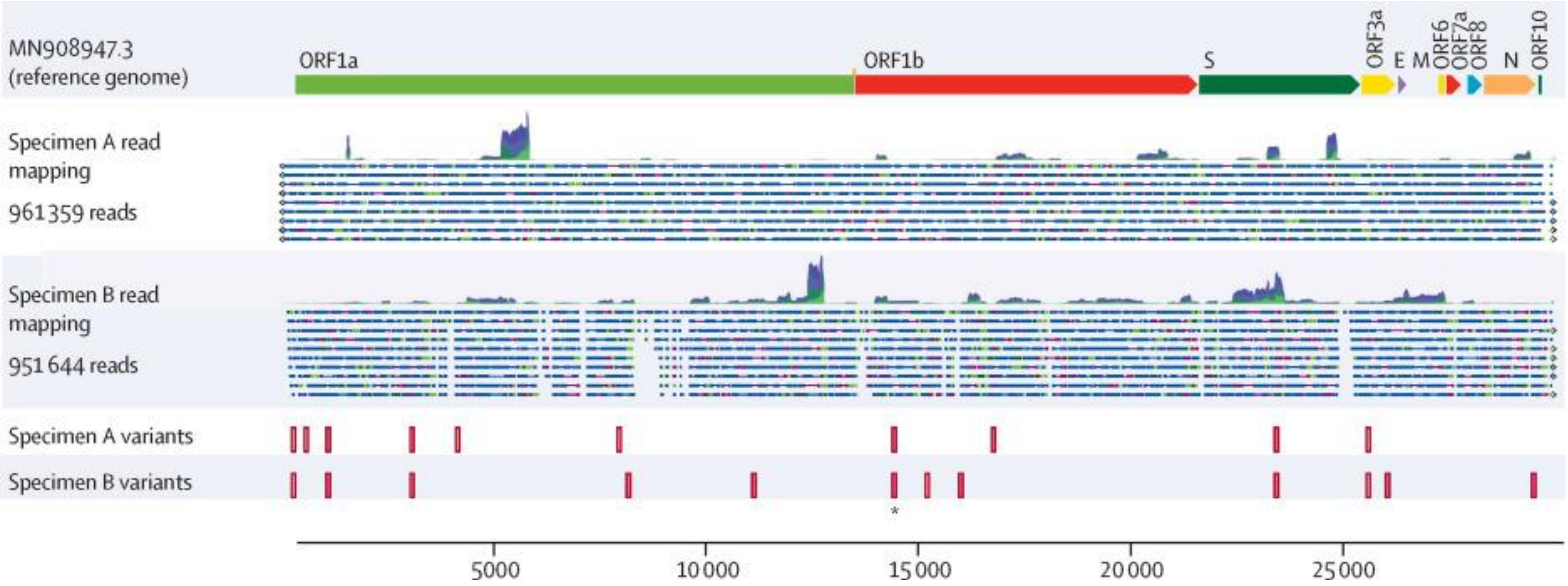
Sequencing applications	Recommended Coverage
Whole genome sequencing (WGS)	15X to 60X
Whole exome sequencing (WES)	100X
RNA sequencing (RNA-seq)	5 to 100 M reads per sample depending on target study
ChIP-Seq	100X

Source: Illumina and genohub

References:

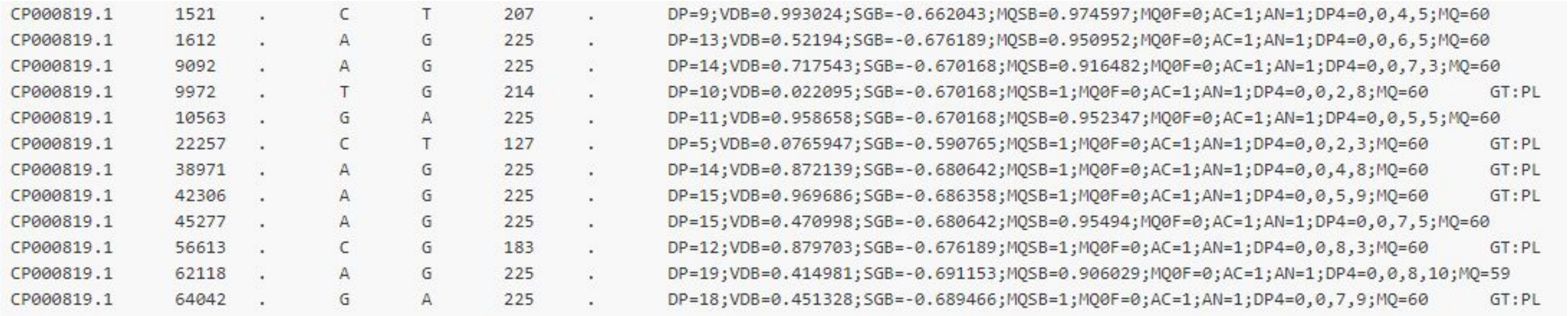
- Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nature Reviews Genetics. 2014 Feb;15(2):121-32.

Análise de dados: alinhamento

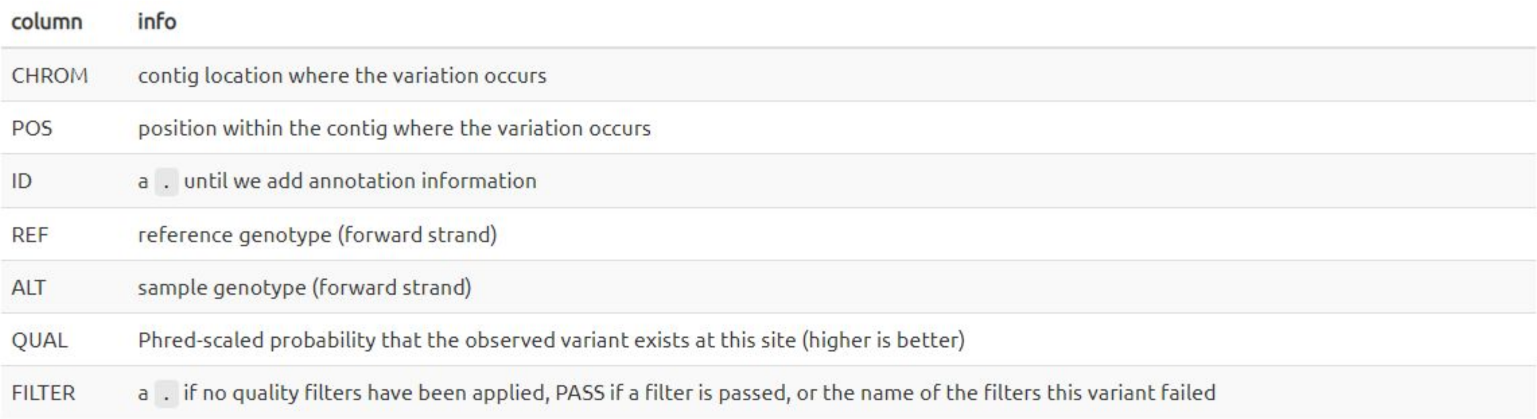


Tillett et al. (2021)

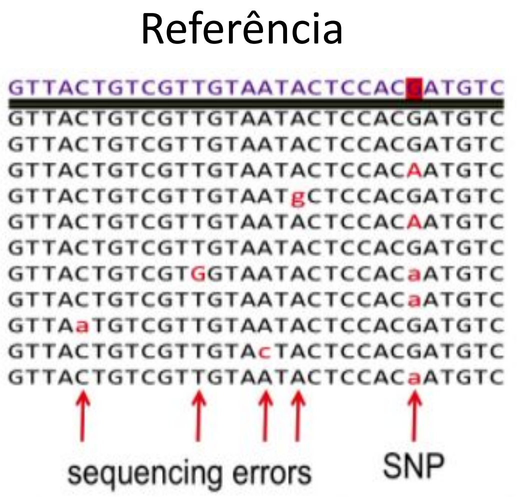
Análise de dados: chamada de variantes (VCF)



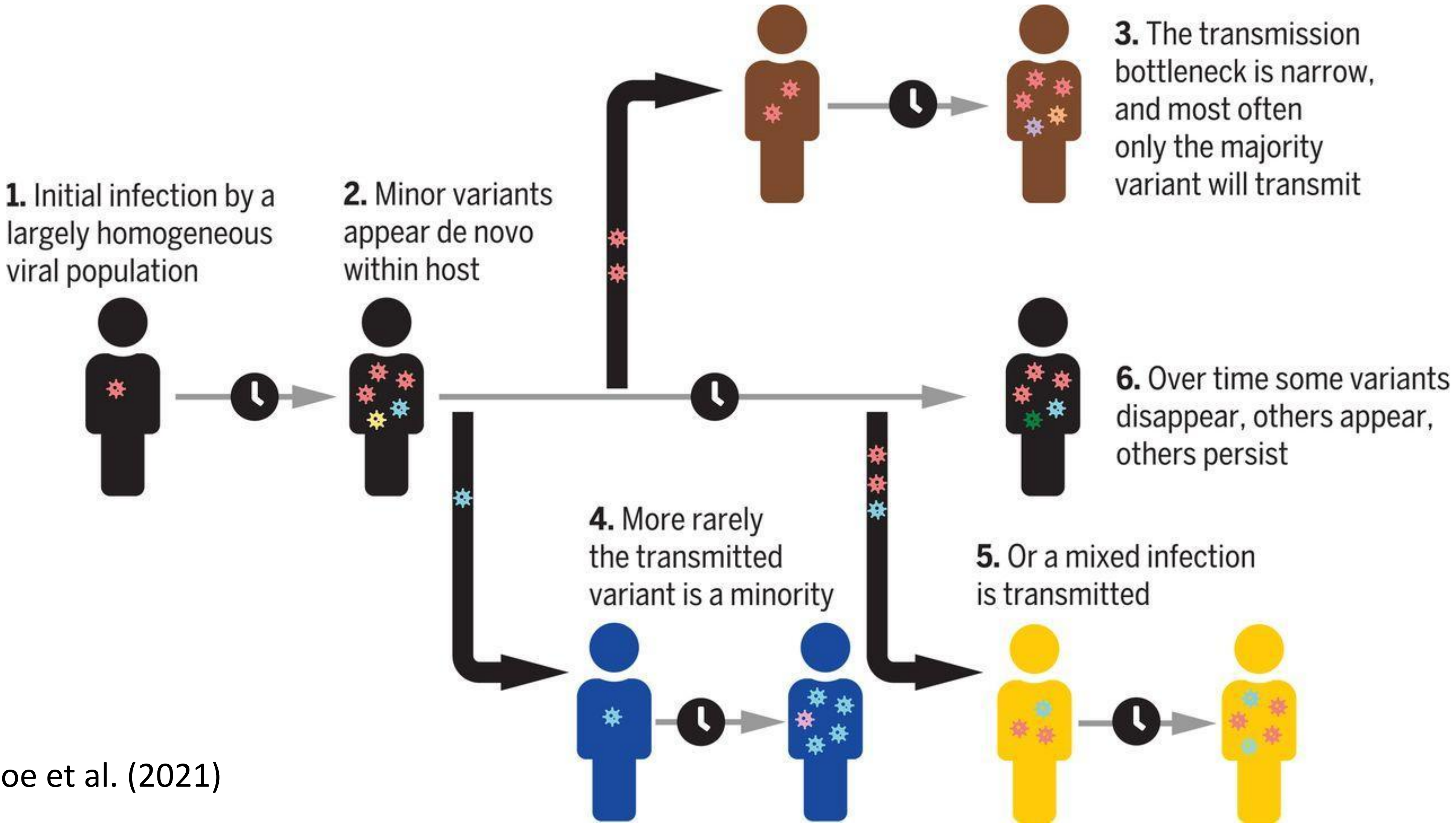
```
CP000819.1      1521    .       C       T       207     .       DP=9;VDB=0.993024;SGB=-0.662043;MQSB=0.974597;MQ0F=0;AC=1;AN=1;DP4=0,0,4,5;MQ=60
CP000819.1      1612    .       A       G       225     .       DP=13;VDB=0.52194;SGB=-0.676189;MQSB=0.950952;MQ0F=0;AC=1;AN=1;DP4=0,0,6,5;MQ=60
CP000819.1      9092    .       A       G       225     .       DP=14;VDB=0.717543;SGB=-0.670168;MQSB=0.916482;MQ0F=0;AC=1;AN=1;DP4=0,0,7,3;MQ=60
CP000819.1      9972    .       T       G       214     .       DP=10;VDB=0.022095;SGB=-0.670168;MQSB=1;MQ0F=0;AC=1;AN=1;DP4=0,0,2,8;MQ=60      GT:PL
CP000819.1      10563   .       G       A       225     .       DP=11;VDB=0.958658;SGB=-0.670168;MQSB=0.952347;MQ0F=0;AC=1;AN=1;DP4=0,0,5,5;MQ=60
CP000819.1      22257   .       C       T       127     .       DP=5;VDB=0.0765947;SGB=-0.590765;MQSB=1;MQ0F=0;AC=1;AN=1;DP4=0,0,2,3;MQ=60      GT:PL
CP000819.1      38971   .       A       G       225     .       DP=14;VDB=0.872139;SGB=-0.680642;MQSB=1;MQ0F=0;AC=1;AN=1;DP4=0,0,4,8;MQ=60      GT:PL
CP000819.1      42306   .       A       G       225     .       DP=15;VDB=0.969686;SGB=-0.686358;MQSB=1;MQ0F=0;AC=1;AN=1;DP4=0,0,5,9;MQ=60      GT:PL
CP000819.1      45277   .       A       G       225     .       DP=15;VDB=0.470998;SGB=-0.680642;MQSB=0.95494;MQ0F=0;AC=1;AN=1;DP4=0,0,7,5;MQ=60
CP000819.1      56613   .       C       G       183     .       DP=12;VDB=0.879703;SGB=-0.676189;MQSB=1;MQ0F=0;AC=1;AN=1;DP4=0,0,8,3;MQ=60      GT:PL
CP000819.1      62118   .       A       G       225     .       DP=19;VDB=0.414981;SGB=-0.691153;MQSB=0.906029;MQ0F=0;AC=1;AN=1;DP4=0,0,8,10;MQ=59
CP000819.1      64042   .       G       A       225     .       DP=18;VDB=0.451328;SGB=-0.689466;MQSB=1;MQ0F=0;AC=1;AN=1;DP4=0,0,7,9;MQ=60      GT:PL
```



column	info
CHROM	contig location where the variation occurs
POS	position within the contig where the variation occurs
ID	a `.` until we add annotation information
REF	reference genotype (forward strand)
ALT	sample genotype (forward strand)
QUAL	Phred-scaled probability that the observed variant exists at this site (higher is better)
FILTER	a `.` if no quality filters have been applied, PASS if a filter is passed, or the name of the filters this variant failed

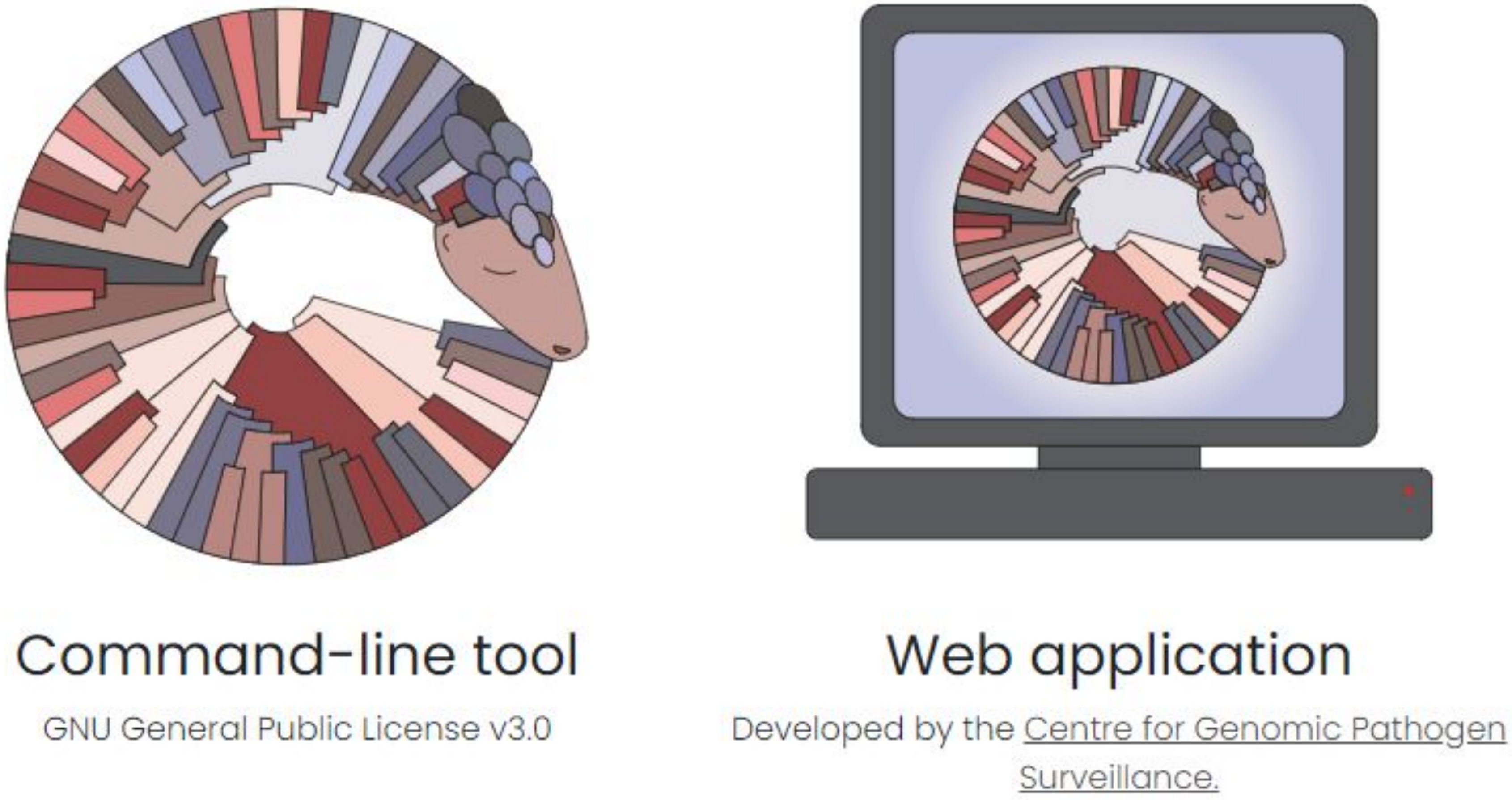

Referência
GTTACTGTCGTTGTAATACTCCACGATGTC
GTTACTGTCGTTGTAATACTCCACGATGTC
GTTACTGTCGTTGTAATACTCCACGATGTC
GTTACTGTCGTTGTAATACTCCACAATGTC
GTTACTGTCGTTGTAATgCTCCACGATGTC
GTTACTGTCGTTGTAATACTCCACAATGTC
GTTACTGTCGTTGTAATACTCCACGATGTC
GTTACTGTCGTGGTAATACTCCACaATGTC
GTTACTGTCGTTGTAATACTCCACaATGTC
GTTAaTGTCGTTGTAATACTCCACGATGTC
GTTACTGTCGTTGTAcTACTCCACGATGTC
GTTACTGTCGTTGTAATACTCCACaATGTC
sequencing errors
SNP

Análise de dados: consenso e variantes


1. Initial infection by a largely homogeneous viral population
2. Minor variants appear de novo within host
3. The transmission bottleneck is narrow, and most often only the majority variant will transmit
4. More rarely the transmitted variant is a minority
5. Or a mixed infection is transmitted
6. Over time some variants disappear, others appear, others persist

Lythgoe et al. (2021)

Análise de dados: definição de linhagem (Pangolin/Nextclade)



Command-line tool

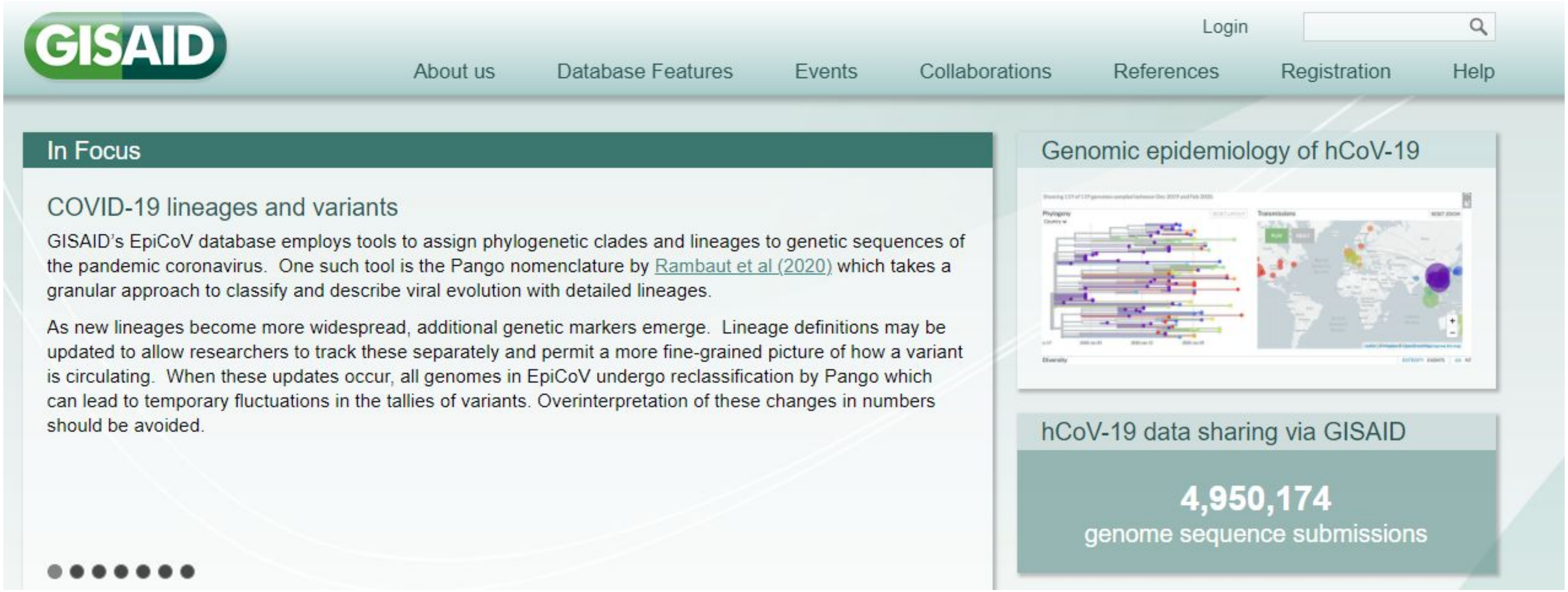
GNU General Public License v3.0

Web application

Developed by the Centre for Genomic Pathogen Surveillance.

Rambaut et al. (2020)

Publicação de genomas: GISAID



In Focus

COVID-19 lineages and variants

GISAID's EpiCoV database employs tools to assign phylogenetic clades and lineages to genetic sequences of the pandemic coronavirus. One such tool is the Pango nomenclature by Rambaut et al (2020) which takes a granular approach to classify and describe viral evolution with detailed lineages.

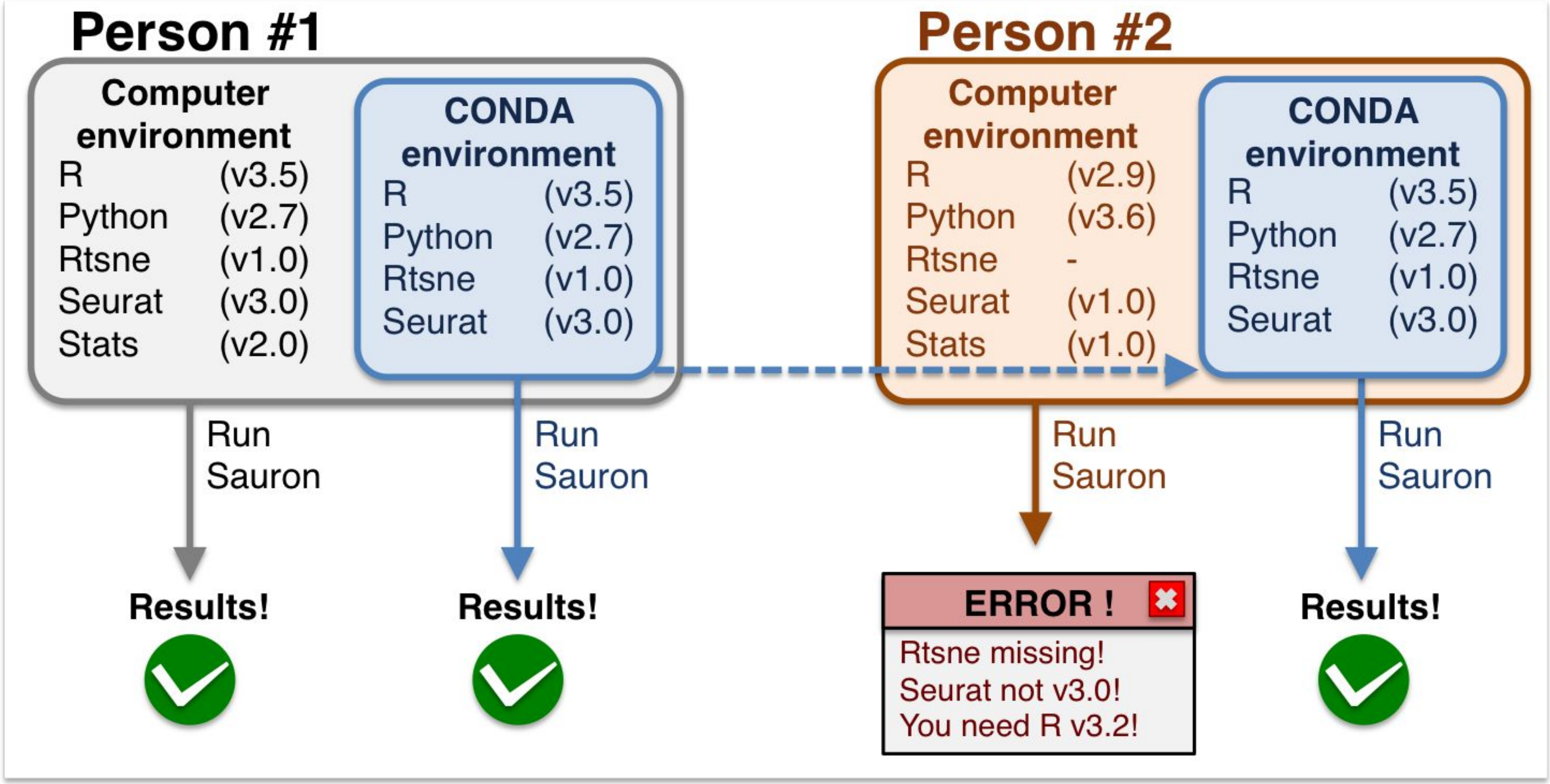
As new lineages become more widespread, additional genetic markers emerge. Lineage definitions may be updated to allow researchers to track these separately and permit a more fine-grained picture of how a variant is circulating. When these updates occur, all genomes in EpiCoV undergo reclassification by Pango which can lead to temporary fluctuations in the tallies of variants. Overinterpretation of these changes in numbers should be avoided.

Genomic epidemiology of hCoV-19

hCoV-19 data sharing via GISAID

4,950,174
genome sequence submissions

Conda



Person #1

Computer environment

R	(v3.5)
Python	(v2.7)
Rtsne	(v1.0)
Seurat	(v3.0)
Stats	(v2.0)

CONDA environment

R	(v3.5)
Python	(v2.7)
Rtsne	(v1.0)
Seurat	(v3.0)

Run Sauron

Results!

Run Sauron

Results!

Person #2

Computer environment

R	(v2.9)
Python	(v3.6)
Rtsne	-
Seurat	(v1.0)
Stats	(v1.0)

CONDA environment

R	(v3.5)
Python	(v2.7)
Rtsne	(v1.0)
Seurat	(v3.0)

Run Sauron

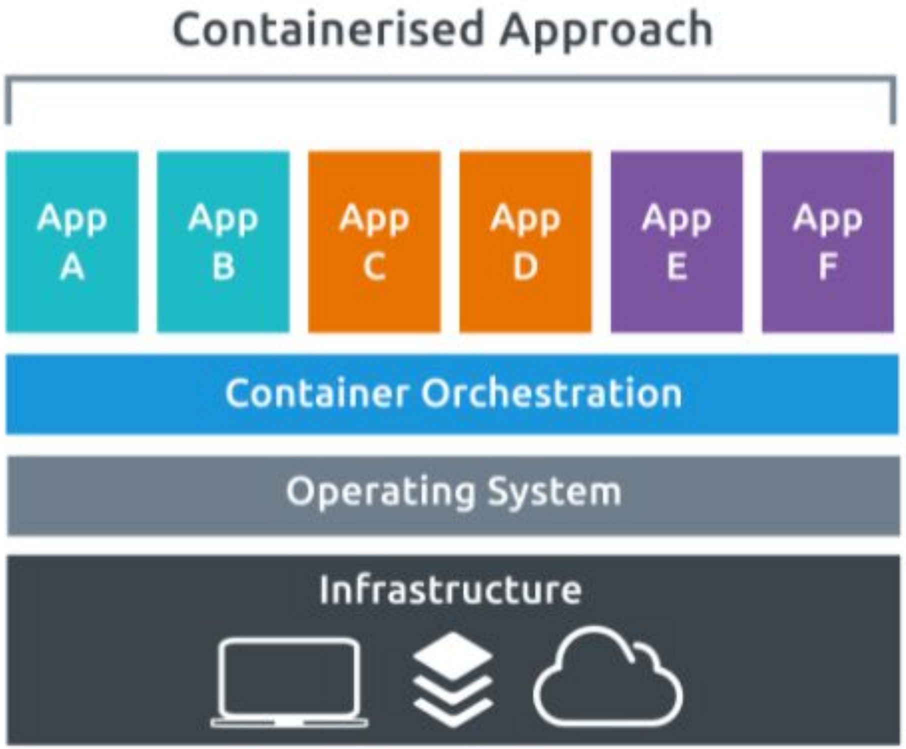
ERROR !

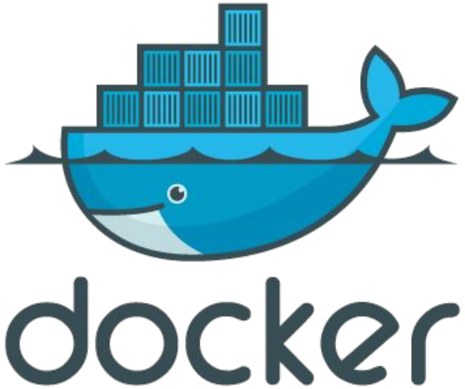
Rtsne missing!
Seurat not v3.0!
You need R v3.2!

Run Sauron

Results!

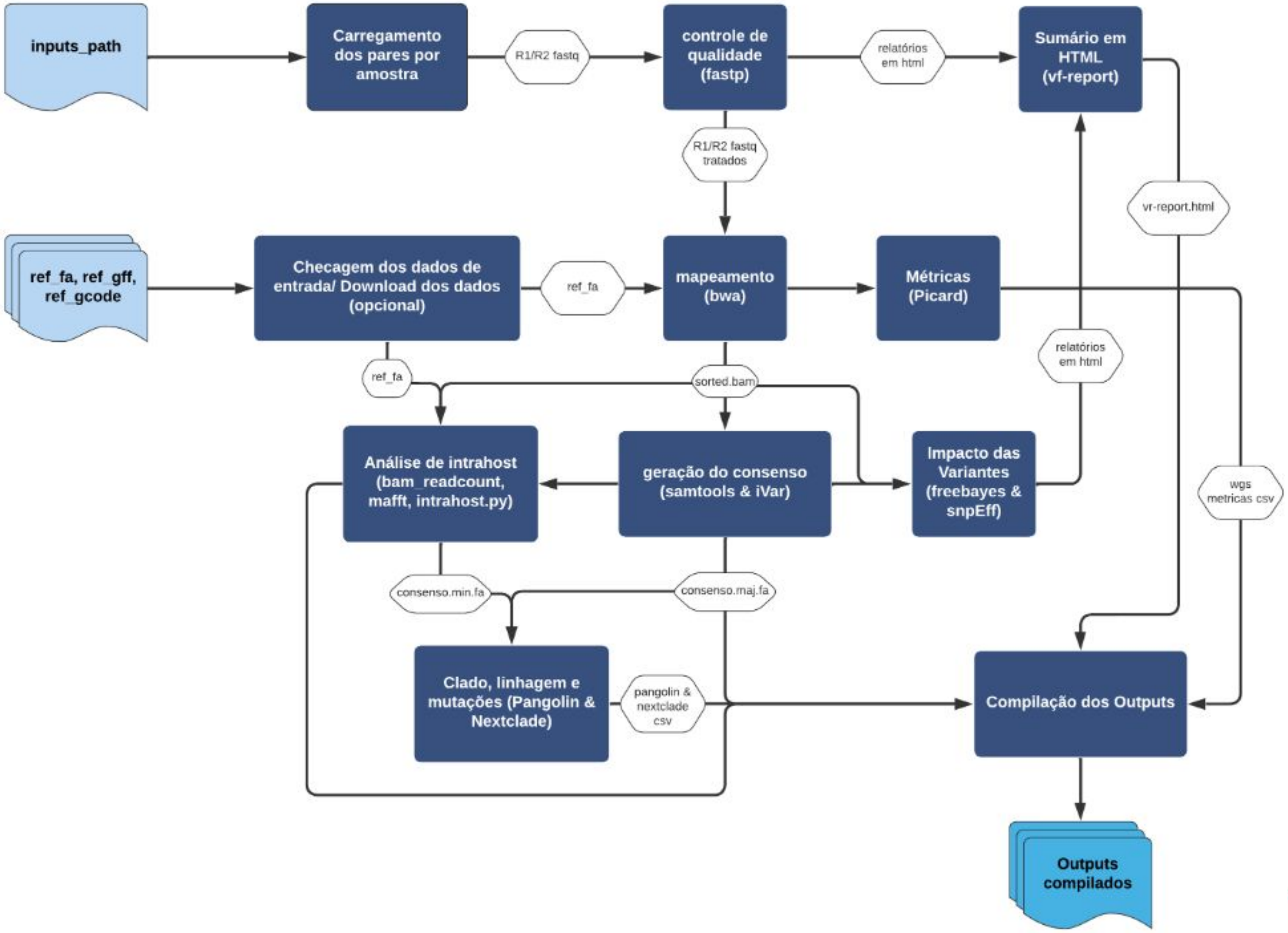
Contêineres





https://cloudhelix.io/blog/post/introduction-containerisation-enterprises

Workflow: ViralFlow



inputs_path

Carregamento dos pares por amostra

R1/R2 fastq

controle de qualidade (fastp)

relatórios em html

Sumário em HTML (vf-report)

R1/R2 fastq tratados

vr-report.html

ref_fa, ref_gff, ref_gcode

Checagem dos dados de entrada/ Download dos dados (opcional)

ref_fa

mapeamento (bwa)

Métricas (Picard)

relatórios em html

ref_fa

sorted.bam

Análise de intrahost (bam_readcount, mafft, intrahost.py)

geração do consenso (samtools & iVar)

Impacto das Variantes (freebayes & snpEff)

wgs metricas csv

consenso.min.fa

consenso.maj.fa

Clado, linhagem e mutações (Pangolin & Nextclade)

pangolin & nextclade csv

Compilação dos Outputs

Outputs compilados

ViralFlow

ACUUUCGAUCUCUG

nextflow