



# Formatos de arquivos

Joyce Silva  
joyce.karol@hotmail.com



# Porque tantos formatos?

- Dados são heterogêneos com coberturas e qualidades diferentes



# Porque tantos formatos?

- Dados são heterogêneos com coberturas e qualidades diferentes
- Propósitos, métodos e análises diferentes



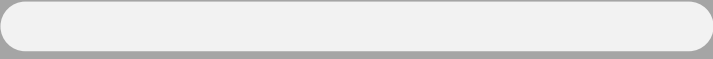

# Porque tantos formatos?

- Dados são heterogêneos com coberturas e qualidades diferentes
- Propósitos, métodos e análises diferentes
- Necessidades mudam



# Leituras brutas

# Sequências



```
>Rotavirus_B_154263.2  
ATCGGATTACACAAGCTA  
GACTAGCCATAGATC
```

- Letrinhas juntas de forma contínua que formam uma string
- String = sequência de caracteres



# Fasta e FastQ

- FASTA e FASTQ são formatos básicos, onipresentes e versáteis de informações para armazenar sequências de nucleotídeos e proteínas.

# FASTA

.fasta ou .fa

1985 > pacote padrão de  
alinhamento de sequências

Sequências de ácidos  
nucléicos, DNA e proteínas.



```
>DQ838640.2 Rotavirus A  
RNA-dependent RNA polymerase  
protein VP1 gene
```

```
GGCTATTAAAGCTGTACAATGGGGGAA  
GTACAATCTAATCTTGTCAGAATATCTAT  
CATTTATATATAATTCACAATCTGCAGTT  
CAAATTCCAATATATTACTCTTCCAACA  
GTGAATTAGAAAATAGATGTATTGAATTT  
CATTCCAAGTGTTTAGAGAACTCAAAG  
AATGGGGTTATCGTTAAGAAAGTTGTTTGT  
TGAATATAATGATGTCATAGAAAATGC  
CACATTACTGTCAATACTATCATATTCTT  
ACGACAAGTATAACGCTGTTGAAAGA  
AAATT
```



# FASTA

.fasta ou .fa

Simple

>cabeçalho

Sequência

- RNA
- DNA
- Aminoácidos



```
>DQ838640.2 Rotavirus A  
RNA-dependent RNA polymerase  
protein VP1 gene
```

```
GGCTATTAAAGCTGTACAATGGGGGAA  
GTACAATCTAATCTTGTCAGAATATCTAT  
CATTTATATATAATTCACAATCTGCAGTT  
CAAATTCCAATATATTACTCTTCCAACA  
GTGAATTAGAAAATAGATGTATTGAATTT  
CATTCCAAGTGTTTAGAGAACTCAAAG  
AATGGGGTTATCGTTAAGAAAGTTGTTTGT  
TGAATATAATGATGTCATAGAAAATGC  
CACATTACTGTCAATACTATCATATTCTT  
ACGACAAGTATAACGCTGTTGAAAGA  
AAATT
```

# FASTA

.fasta ou .fa > .fna > .faa

- Fasta de nucleotídeos .fna
- Fasta de aminoácidos .faa



**>DQ838640.2 Rotavirus A**  
**RNA-dependent RNA polymerase**  
**protein VP1 gene**

```
GGCTATTAAAGCTGTACAATGGGGGAA  
GTACAATCTAATCTTGTCAGAATATCTAT  
CATTTATATATAATTCACAATCTGCAGTT  
CAAATTCCAATATATTACTCTTCCAACA  
GTGAATTAGAAAATAGATGTATTGAATTT  
CATTCCAAGTGTTTAGAGAACTCAAAG  
AATGGGGTTATCGTTAAGAAAGTTGTTTGT  
TGAATATAATGATGTCATAGAAAATGC  
CACATTACTGTCAATACTATCATATTCTT  
ACGACAAGTATAACGCTGTTGAAAGA  
AAATT
```

# Qual

.qual

armazena os valores de  
qualidade



```
>DQ838640.2 Rotavirus A  
RNA-dependent RNA  
polymerase protein VP1 gene
```

```
20 19 10 11 17 18 18 18 19 20  
21 22 06 08 12 12
```

# FASTQ

.fastq ou .fq

Surge em 2010 como  
formato unificado para  
armazenar sequências e  
qualidade



```
@SEQUENCE_ID
```

```
GATTGGGGTTCAAAGCAGTATCGATCAAA  
TAGTAAATCCATTGTCAACTCACAGTTT
```

```
+
```

```
!"*((( (**+))%%%++) (%%%) .1***-+*)"**55CC  
F>>>>>CCCCCCCC65
```

# FASTQ

.fastq ou .fq

@ID\_único

sequência

+

dados de qualidade de  
cada base

@SEQUENCE\_ID

GATTGGGGTTCAAAGCAGTATCGATCAAA  
TAGTAAATCCATTGTCAACTCACAGTTT

+

!"\*((( (\*\*+))%%%++) (%%%) .1\*\*\*-+\*" ) \*\*55CC  
F>>>>>CCCCCCCC65

# Score de qualidade

Tabela ASCII o que permite que apenas um caractere seja usado para cada base

Symbol	ASCII Code	Q-Score	Symbol	ASCII Code	Q-Score
!	33	0	6	54	21
"	34	1	7	55	22
#	35	2	8	56	23
\$	36	3	9	57	24
%	37	4	:	58	25
&	38	5	;	59	26
'	39	6	<	60	27
(	40	7	=	61	28
)	41	8	>	62	29
*	42	9	?	63	30
+	43	10	@	64	31
,	44	11	A	65	32
-	45	12	B	66	33
.	46	13	C	67	34
/	47	14	D	68	35
0	48	15	E	69	36
1	49	16	F	70	37
2	50	17	G	71	38
3	51	18	H	72	39
4	52	19	I	73	40
5	53	20			

Score de Qualidade



# Arquivo compactado

.gz

O formato de arquivo .gz é um arquivo compactado criado pelo algoritmo de compressão gzip

- .fasta.gz
- .fastq.gz
- .fq.gz
- .faa.gz
- .fna.gz

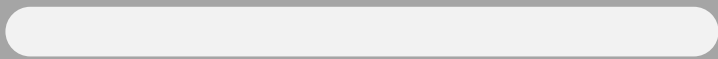




# Leitura de Sequências

- Como eu abro esse arquivo?





>Rotavirus\_B\_154263.2

ATCGGATTACACAAGCTA  
GACTAGCCATAGATC

*Organismo*

Pares de base

*Homo sapiens*

16.310.774.187

*Mus musculus*

9.974.977.889



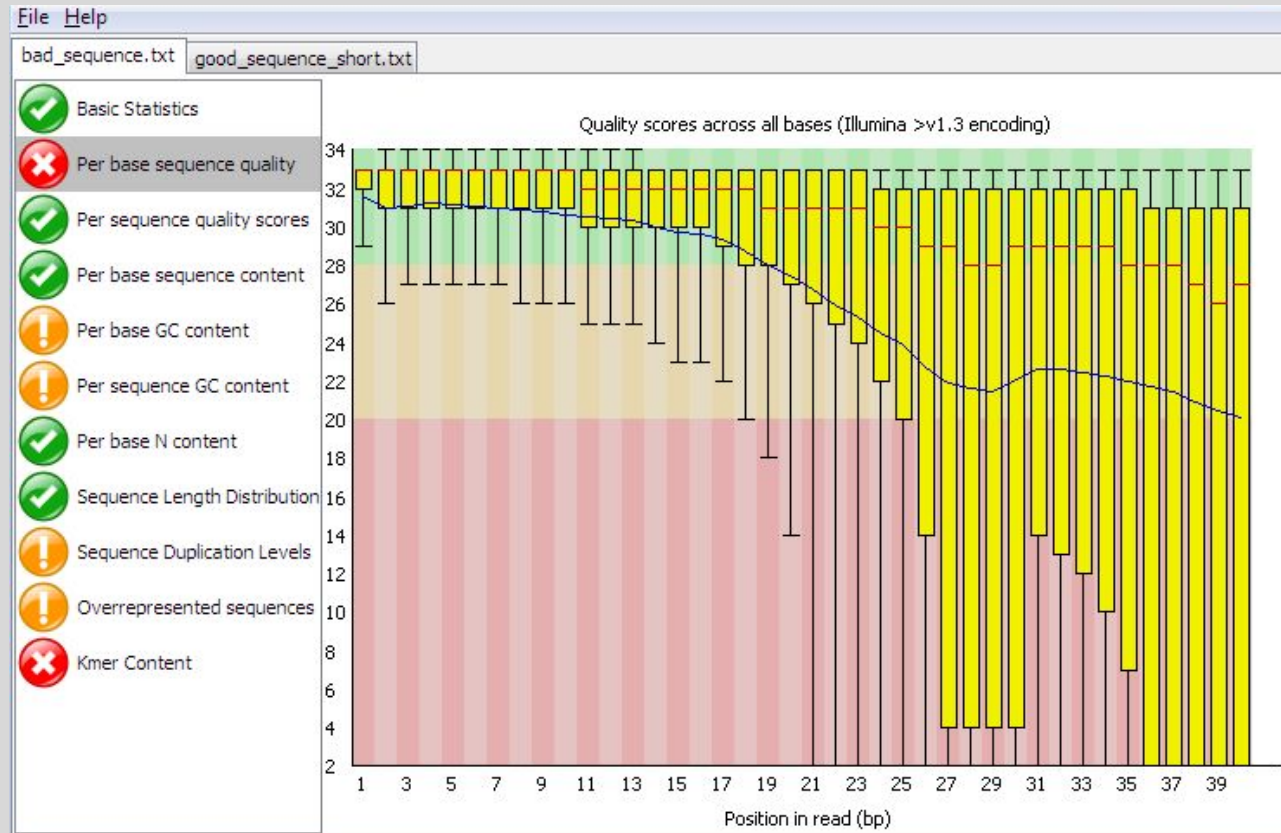
```
>homo_sapiens_  
GRCh38.2
```

```
ATCGGATTACACAAGCTA  
GACTAGCCATAGATC
```



Your PC ran into a problem and needs to restart. We're just collecting some error info, and then we'll restart for you. (0% complete)

If you'd like to know more, you can search online later for this error: HAL\_INITIALIZATION\_FAILED



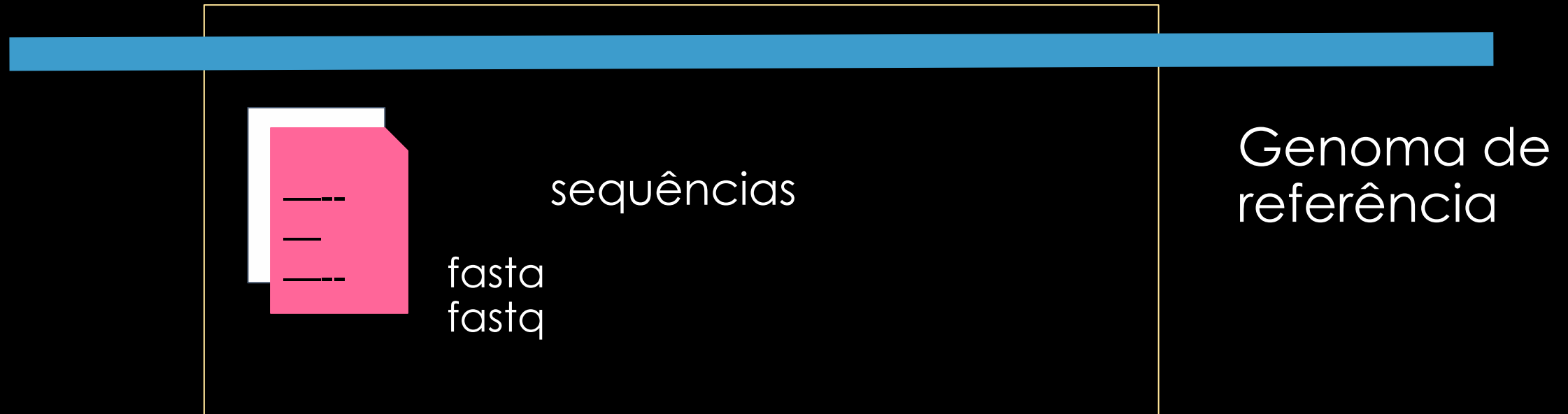
fastQC



# Arquivos de alinhamento

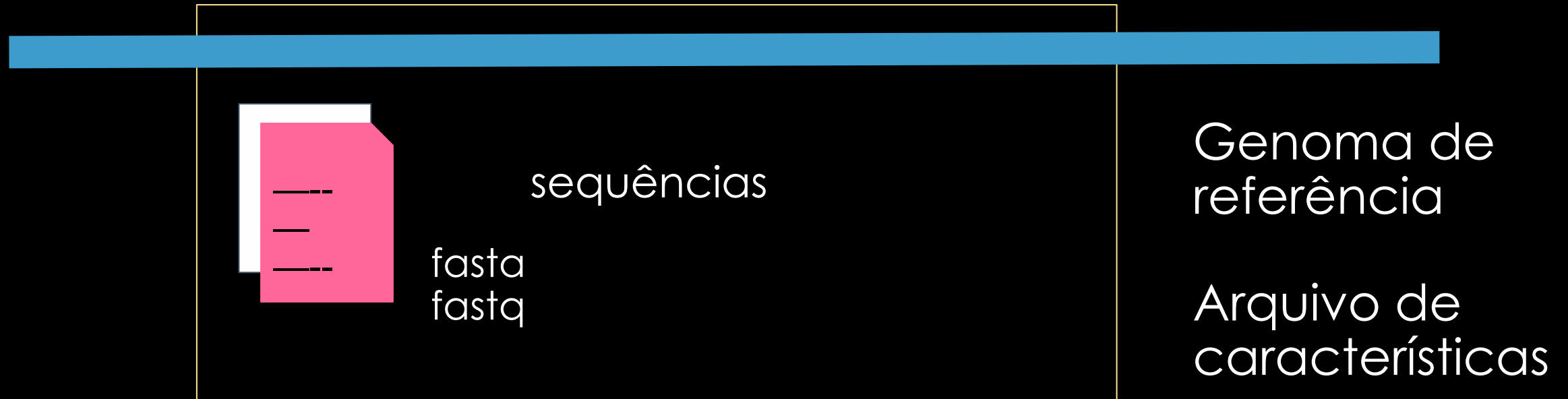


# Arquivo de mapeamento



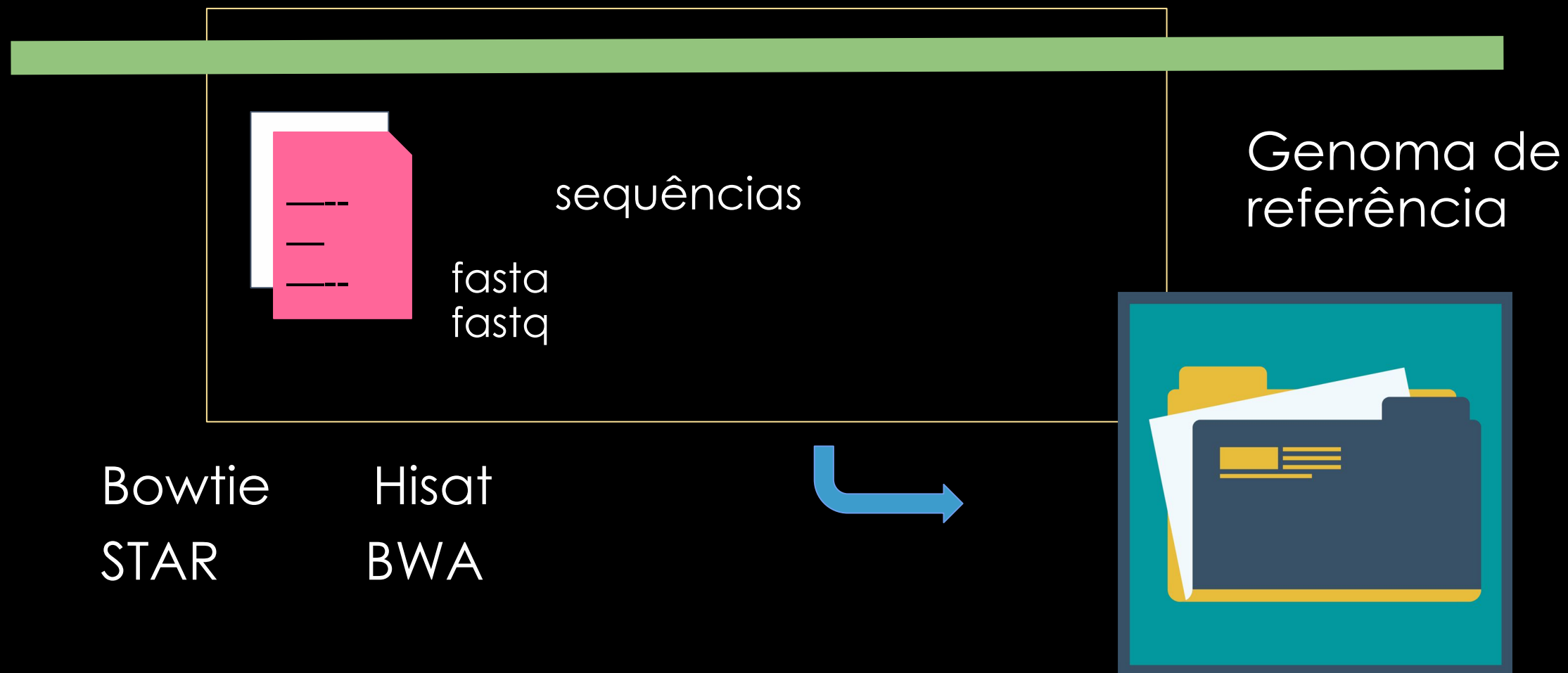


# Arquivo de mapeamento





# Arquivo de mapeamento







# SAM



Formato genérico para armazenar essas informações de alinhamentos de leitura contra sequências de referência >  
Sequence Alignment Map (SAM)

Desenhado para conjuntos de alinhamentos  $10^{11}$   
Ideal para sequências humanas

Arquivos mais pesados



# SAM



Arquivo delimitado por tabulação

Mantém duas linhas de informações válidas armazenadas pelo FASTQ:

- o nome da leitura > @
- código de qualidade



# SAM

Arquivo delimitado por tabulação

Mantém duas linhas de informações válidas armazenadas pelo FASTQ:

- o nome da leitura > @
- código de qualidade

Armazena informações extras vindas do alinhamento

- Cada linha de alinhamento tem 11 campos obrigatórios
- Um número variável de campos opcionais



# BAM



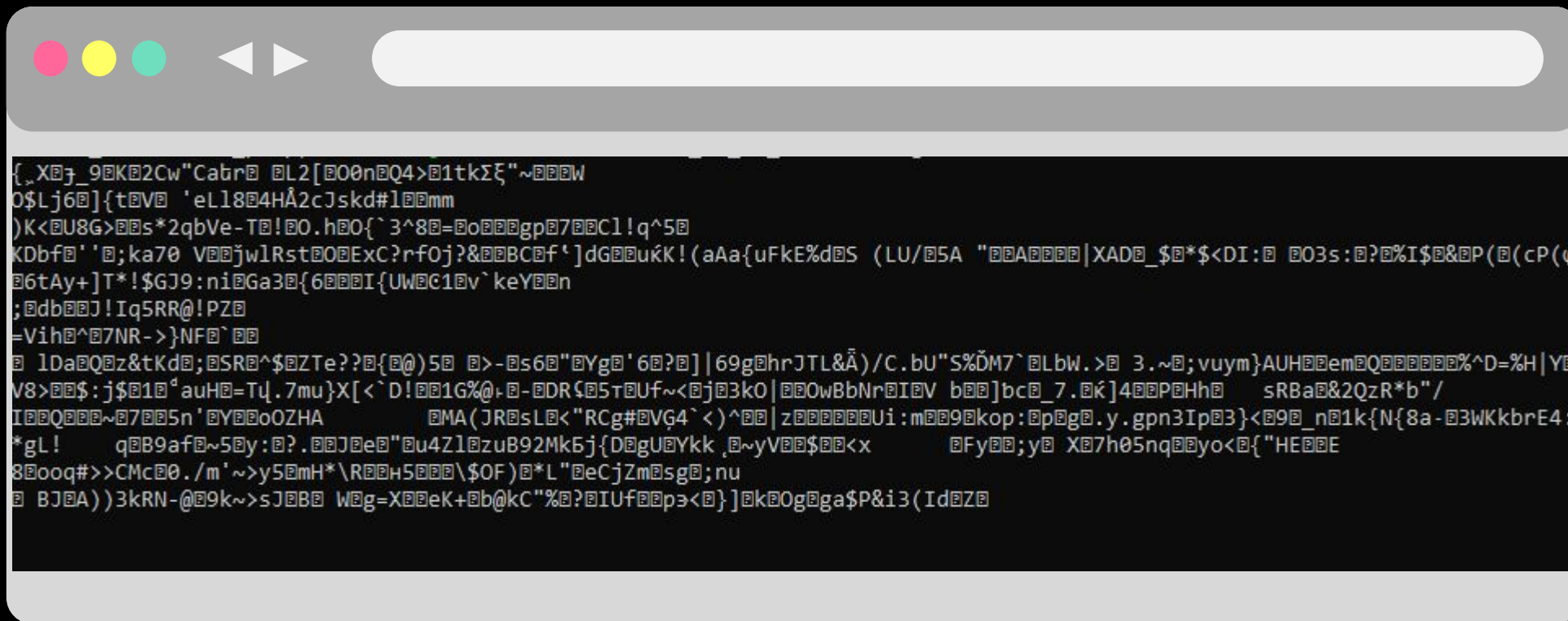
Versão binária do SAM

Costuma ser mais usado por ocupar menor espaço de armazenamento

Mantém as mesmas informações que o arquivo de alinhamento SAM



# BAM





# BAM



```
Load BAM file ... Done

#=====
#All numbers are READ count
#=====

Total records:                               47263034

QC failed:                                   0
Optical/PCR duplicate:                       0
Non primary hits                             14551135
Unmapped reads:                              0
mapq < mapq_cut (non-unique):                4289754

mapq >= mapq_cut (unique):                   28422145
Read-1:                                       14251021
Read-2:                                       14171124
Reads map to '+':                           14210087
Reads map to '-':                           14212058
Non-splice reads:                           25382616
Splice reads:                               3039529
Reads mapped in proper pairs:                28331724
Proper-paired reads map to different chrom:0
```

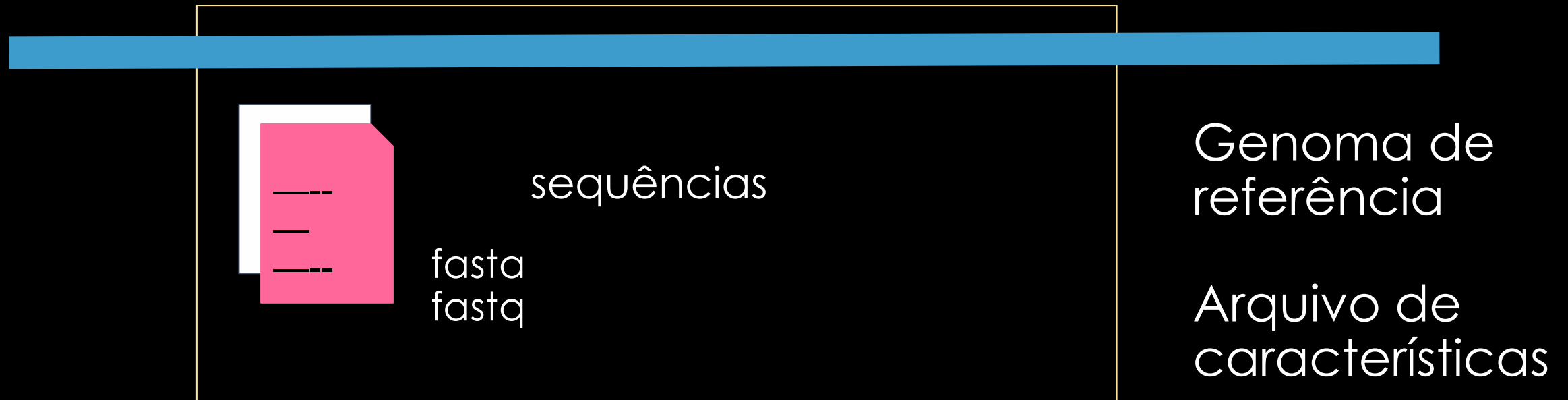
## RSeQC



# Arquivos de anotação



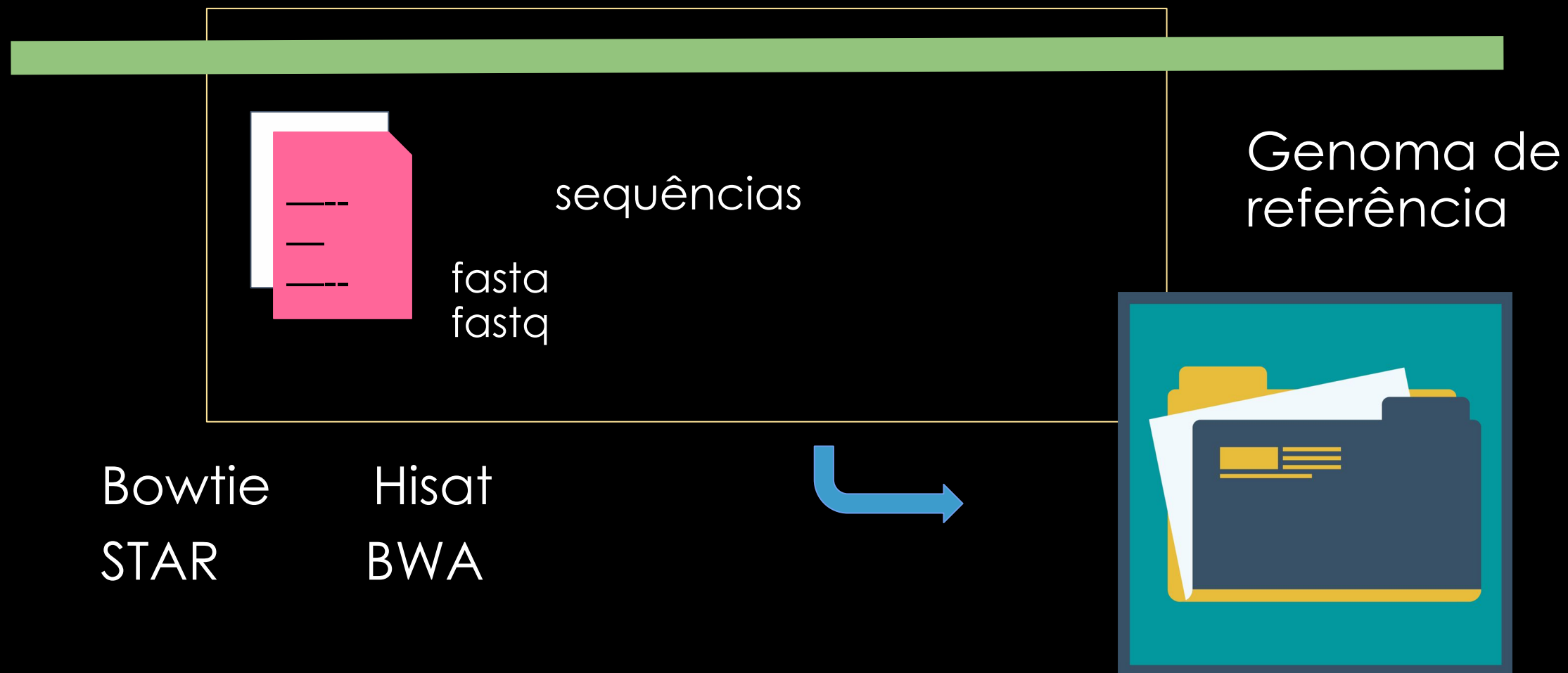
# Arquivo de mapeamento







# Arquivo de mapeamento



# GFF

- General Feature Format
- Formato de características gerais
- Usado para anotar genes e outras características de sequências de DNA, RNA e proteína de genomas

# GFF

- São delimitados por tabulações com 9 campos por linha
- A mesma estrutura para os primeiros 8 campos
  - conteúdo
  - formato do 9º campo

# Campos do GFF

<div><div><div></div><div></div><div></div></div><div><div></div><div></div></div><div></div></div>									
Posição	1	2	3	4	5	6	7	8	9
Nome	SeqID (sequência e ID)	Source	Tipo	Começo	Final	Score	Strand	Fase	Atributos
Descrição	Nome da sequência onde fica a feature	De onde veio a feature (software ou banco de dados)	Gene ou exon	Início da feature	Fim da feature	Confiança da fonte de anotação	Sentido 5'-3' (+) ou 3'-5' (-)	Indica onde o recurso começa com referência na leitura	Informações adicionais

# Versões

## **GFF2 e GTF2**

General feature format version 2

- Obsoleto
- Suporta duas hierarquias

gene → transcrito

Gene Transfer Format 2.2

- Derivação usada pelo Ensembl

## **GFF3**

Generic Feature Format Version  
3

- Formato preferencial
- Suporta 3+ hierarquias:  
gene → transcrito → exon
- Hierarquia pode ser arbitrária
- Dá significado a certas informações

# Versões

## GFF2 e GTF2

General feature format version 2

- Obsoleto
- Suporta duas hierarquias  
gene → transcrito

Gene Transfer Format 2.2

- Derivação usada pelo Ensembl

## GFF3

- canonical genes non-coding transcripts
- parent (part-of) relationships
- alignments
- ontology association and database cross references
- single exon genes
- polycistronic transcripts
- genes containing inteins
- trans-spliced transcripts
- programmed frameshifts
- operons

Organismo	Pares de base
<i>Homo sapiens</i>	16.310.774.187
<i>Mus musculus</i>	9.974.977.889
<i>Rattus norvegicus</i>	6.521.253.272
<i>Bos taurus</i>	5.386.258.455
<i>Zea mays</i>	5.062.731.057
<i>Sus scrofa</i>	4.887.861.860
<i>Danio rerio</i>	3.120.857.462
<i>Strongylocentrotus purpuratus</i>	1.435.236.534
<i>Macaca mulatta</i>	1.256.203.101
<i>Oryza sativa Japonica Group</i>	1.255.686.573
<i>Nicotiana tabacum</i>	1.197.357.811
<i>Xenopus (Silurana) tropicalis</i>	1.249.938.611
<i>Drosophila melanogaster</i>	1.119.965.220
<i>Pan troglodytes</i>	1.008.323.292
<i>Arabidopsis thaliana</i>	1.144.226.616
<i>Canis lupus familiaris</i>	951.238.343
<i>Vitis vinifera</i>	999.010.073
<i>Gallus gallus</i>	899.631.338
<i>Glycine max</i>	906.638.854
<i>Triticum aestivum</i>	898.689.329

# GenBank

- GenBank é um banco de dados de anotações de sequências de nucleotídeos, suas traduções de proteínas.
- Esse banco de dados é produzido e mantido pelo National Center for Biotechnology Information (NCBI).

# GenBank

- International Nucleotide Sequence Database Collaboration (INSDC)
  - DNA DataBank of Japan (DDBJ)
  - European Nucleotide Archive (ENA)
- .gbk ou .gb



◀

▶

https://www.ncbi.nlm.nih.gov/genbank/

NIH

National Library of Medicine

National Center for Biotechnology Information

Log in

GenBank

Nucleotide

Search

GenBank

Submit

Genomes

WGS

Metagenomes

TPA

TSA

INSDC

Documentation

Other

GenBank Overview

What is GenBank?

GenBank<sup>®</sup> is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences ([Nucleic Acids Research, 2013 Jan 41\(D1\):D36-42](#)). GenBank is part of the [International Nucleotide Sequence Database Collaboration](#), which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.

A GenBank release occurs every two months and is available from the [ftp site](#). The [release notes](#) for the current version of GenBank provide detailed information about the release and notifications of upcoming changes to GenBank. Release notes for [previous GenBank releases](#) are also available. GenBank growth [statistics](#) for both the traditional GenBank divisions and the WGS division are available from each release.

An [annotated sample GenBank record](#) for a *Saccharomyces cerevisiae* gene demonstrates many of the features of the GenBank flat file format.

Access to GenBank

There are several ways to search and retrieve data from GenBank.

- Search GenBank for sequence identifiers and annotations with [Entrez Nucleotide](#).
- Search and align GenBank sequences to a query sequence using [BLAST](#) (Basic Local Alignment Search Tool). See [BLAST info](#) for more information about the numerous BLAST databases.
- Search, link, and download sequences programatically using [NCBI e-utilities](#).
- The ASN.1 and flatfile formats are available at NCBI's anonymous FTP server: <ftp://ftp.ncbi.nlm.nih.gov/ncbi-asn1> and <ftp://ftp.ncbi.nlm.nih.gov/genbank>.

GenBank Data Usage

The GenBank database is designed to provide and encourage access within the scientific community to the most up-to-date and comprehensive DNA sequence information. Therefore, NCBI places no restrictions on the use or distribution of the GenBank data. However, some submitters may claim patent, copyright, or other intellectual property rights in all or a portion of the data they have submitted. NCBI is not in a position to assess the validity of such claims, and therefore cannot provide comment or unrestricted permission concerning the use, copying, or distribution of the information contained in GenBank.

GenBank Resources

[GenBank Home](#)

[Submission Types](#)

[Submission Tools](#)

[Search GenBank](#)

[Update GenBank Records](#)

GenBank

# Estrutura do Arquivo.gbk

- CAMPOS – com a explicação no espaço ao lado
- O campo LOCUS define o ID da sequência e alguns meta-dados:
  - Pares de base
  - DNA, RNA ou proteína, topologia
  - Tipo de organismo
  - Data de submissão

ACCESSION - código de acesso principal e alternativos

SOURCE - informações de taxonomia

# Estrutura do Arquivo.gbk

NIH

National Library of Medicine  
National Center for Biotechnology Information

Log in

Nucleotide

Nucleotide

coronavirus

Search

Create alert Advanced

Help

Species

Animals (4,596)

Plants (57)

Fungi (4)

Protists (2)

Bacteria (1,165)

Archaea (1)

Viruses (8,360,279)

Customize ...

Molecule types

genomic DNA/RNA (8,435,293)

mRNA (4,297)

Customize ...

Source databases

INSDC (GenBank) (8,473,422)

RefSeq (3,378)

Customize ...

Sequence Type

Nucleotide (8,477,043)

EST (3)

Genetic compartments

Mitochondrion (357)

Sequence length

Custom range...

Release date

Custom range...

Revision date

Custom range...

Clear all

Show additional filters

Summary

20 per page

Sort by Default order

Send to:

Filters: [Manage Filters](#)

TAXONOMY

Was this helpful?

Betacoronavirus - viruses, genus

[Search \(txid694002\)](#)

Gammacoronavirus - viruses, genus

[Search \(txid694013\)](#)

Alphacoronavirus - viruses, genus

[Search \(txid693996\)](#)

Results by taxon

Top Organisms [\[Tree\]](#)

Severe acute respiratory syndrome-related coronavirus (8307291)

synthetic construct (23449)

Avian coronavirus (15568)

Porcine epidemic diarrhea virus (7399)

Alphacoronavirus 1 (5756)

All other taxa (117583)

More...

Find related data

Database: [Select](#)

Find items

Search details

"Alphacoronavirus"[Organism] OR

"Betacoronavirus"[Organism] OR

"Gammacoronavirus"[Organism] OR

coronavirus[All Fields]

Search

See more...

Recent activity

coronavirus (8477046)

Nucleotide

txid10942[Organism:exp] (422)

Nucleotide

Escherichia phage HK629

Genome

Genome for Nucleotide (Select 428782011) (1)

Genome

Enterobacteria phage HK629, complete genome

Nucleotide

Items: 1 to 20 of 8477046

<< First < Prev Page 1 of 423853 Next > Last >>

☐ Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/CA-CDC-VSX-A356888/2023 ORF1ab polyprotein (ORF1ab), ORF1a polyprotein (ORF1ab), surface glycoprotein (S), ORF3a protein (ORF3a), envelope protein (E), membrane glycoprotein (M), ORF6 protein (ORF6), ORF7a protein (ORF7a), and ORF7b (ORF7b) genes, complete cds; ORF8 gene, complete sequence; and nucleocapsid phosphoprotein (N) and ORF10 protein (ORF10) genes, complete cds

29,813 bp linear RNA

Accession: OR592427.1 GI: 2581254236

[BioProject](#) [BioSample](#) [Protein](#) [Taxonomy](#)

[GenBank](#) [FASTA](#) [Graphics](#)

☐ Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/CA-CDC-VSX-A356892/2023 ORF1ab polyprotein (ORF1ab), ORF1a polyprotein (ORF1ab), surface glycoprotein (S), ORF3a protein (ORF3a), envelope protein (E), membrane glycoprotein (M), ORF6 protein (ORF6), ORF7a protein (ORF7a), and ORF7b (ORF7b) genes, complete cds; ORF8 gene, complete sequence; and nucleocapsid phosphoprotein (N) and ORF10 protein (ORF10) genes, complete cds

29,807 bp linear RNA

# Estrutura do Arquivo.gbk

GenBank

Send to: ▾

Change region shown ▾

Customize view ▾

Analyze this sequence ▴

Run BLAST

Pick Primers

Highlight Sequence Features

Find in this Sequence

NCBI Virus ▴

Retrieve, view, and download SARS-CoV-2 coronavirus genomic and protein sequences.

Related information ▴

BioProject

BioSample

Protein

Taxonomy

Recent activity ▴

Turn Off Clear

Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV- Nucleotide

coronavirus (8477046) Nucleotide

txid10942[Organism:exp] (422) Nucleotide

Escherichia phage HK629 Genome

Genome for Nucleotide (Select 428782011) (1) Genome

See more...

## Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/CA-CDC-VSX-A356888/2023 ORF1ab polyprotein (ORF1ab), ORF1a polyprotein (ORF1ab), surface glycoprotein (S), ORF3a protein (ORF3a), envelope protein (E), membrane glycoprotein (M), ORF6...

GenBank: OR592427.1

[FASTA](#) [Graphics](#)

[Go to:](#) ▾

LOCUS OR592427 29813 bp RNA linear VRL 24-SEP-2023  
DEFINITION Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/CA-CDC-VSX-A356888/2023 ORF1ab polyprotein (ORF1ab), ORF1a polyprotein (ORF1ab), surface glycoprotein (S), ORF3a protein (ORF3a), envelope protein (E), membrane glycoprotein (M), ORF6 protein (ORF6), ORF7a protein (ORF7a), and ORF7b (ORF7b) genes, complete cds; ORF8 gene, complete sequence; and nucleocapsid phosphoprotein (N) and ORF10 protein (ORF10) genes, complete cds.  
ACCESSION OR592427  
VERSION OR592427.1  
DBLINK BioProject: [PRJNA720050](#)  
BioSample: [SAMN37523022](#)  
KEYWORDS purposeofsampling:baselinesurveillance.  
SOURCE Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)  
ORGANISM [Severe acute respiratory syndrome coronavirus 2](#)  
Viruses; Riboviria; Orthornavirae; Pisuviricota; Pisoniviricetes; Nidovirales; Cornidovirineae; Coronaviridae; Orthocoronavirinae; Betacoronavirus; Sarbecovirus; Severe acute respiratory syndrome-related coronavirus.  
REFERENCE 1 (bases 1 to 29813)  
AUTHORS Howard,D., Batra,D., Cook,P.W., Caravas,J., Rambo-Martin,B., Unoarumhi,Y., Schmerer,M., Lacey,K.A., Ca,H., Morrison,S., Gulvik,C., Sula,E., Paden,C.R., Mandal,P., Bajwa,M., Thornburg,N., Chau,R., Mandal,P., Momin,N. and MacCannell,D.  
TITLE CDC Sars CoV2 Sequencing Baseline Constellation  
JOURNAL Unpublished  
REFERENCE 2 (bases 1 to 29813)  
AUTHORS Howard,D., Batra,D., Cook,P.W., Caravas,J., Rambo-Martin,B., Unoarumhi,Y., Schmerer,M., Lacey,K.A., Ca,H., Morrison,S., Gulvik,C., Sula,E., Paden,C.R., Mandal,P., Bajwa,M., Thornburg,N., Chau,R., Mandal,P., Momin,N. and MacCannell,D.  
TITLE Direct Submission  
JOURNAL Submitted (24-SEP-2023) Respiratory Viruses Branch, Division of Viral Diseases, Centers for Disease Control and Prevention, 1600 Clifton Rd, Atlanta, GA 30329, USA  
COMMENT ##Assembly-Data-START##  
Assembly Method :: Helix klados-fastagenerator-6.0.1  
Sequencing Technology :: Illumina NovaSeq  
##Assembly-Data-END##  
FEATURES  
source Location/Qualifiers  
1..29813  
/organism="Severe acute respiratory syndrome coronavirus 2"



# .gb e .gbk

O formato .gbk é uma versão mais recente do formato .gb  
Uma extensão usada para armazenar informações de  
sequência de DNA e proteína



# .gb e .gbk

Podem ser usados em programas de bioinformática

Metadados como a fonte da amostra, uma descrição e informações do autor



# obrigada!

contato:

joyce.karol@hotmail.com