

Estratégias para a montagem de genomas, tipos de dados. Avaliação de montagens

Discentes

Danilo Ferreira da Silva

Doutorando em Solos e Nutrição
de Plantas ESALQ/USP

Gabriely Santos de Oliveira

Mestranda em Biologia na Agricultura e no Ambiente

CENA/USP



Estratégias para a montagem de genomas de novo, tipos de dados. Avaliação de montagens

1.

**First-generation
sequencing**
Sanger

2.

**Second-generation
sequencing (NGS)**
Illumina
Pirosequenciamento

3.

**third-generation
sequencing**
PacBio
ONT

Estratégias para a montagem de genomas de novo, tipos de dados. Avaliação de montagens



1 - First-generation sequencing

SANGER

- 1977
 - Fragmentos de 500-900bp
 - Custo alto por bp
-

Estratégias para a montagem de genomas de novo, tipos de dados. Avaliação de montagens



2 - Second-generation Sequencing (NGS)

- Conhecido também como Sequenciamento de Alta Performance
- Fragmentos de 50-300 bp
- Custo baixo por bp
- Automático
- Processamento paralelo massivo de fragmentos de DNA
- Sequenciamento em dias

ILLUMINA

PIROSEQUENCIAMENTO

ION TORRENT



3. Third-generation

Sequencing

- Sequenciamento de fragmentos longos
- Não é preciso fazer PCR da amostra
- ~12kb - ~2Mb
- É necessário pouca amostra
- Baixo - moderado custo por bp
- Sequenciamento em horas

OXFORD NANOPORE

PACIFIC BIOSCIENCES

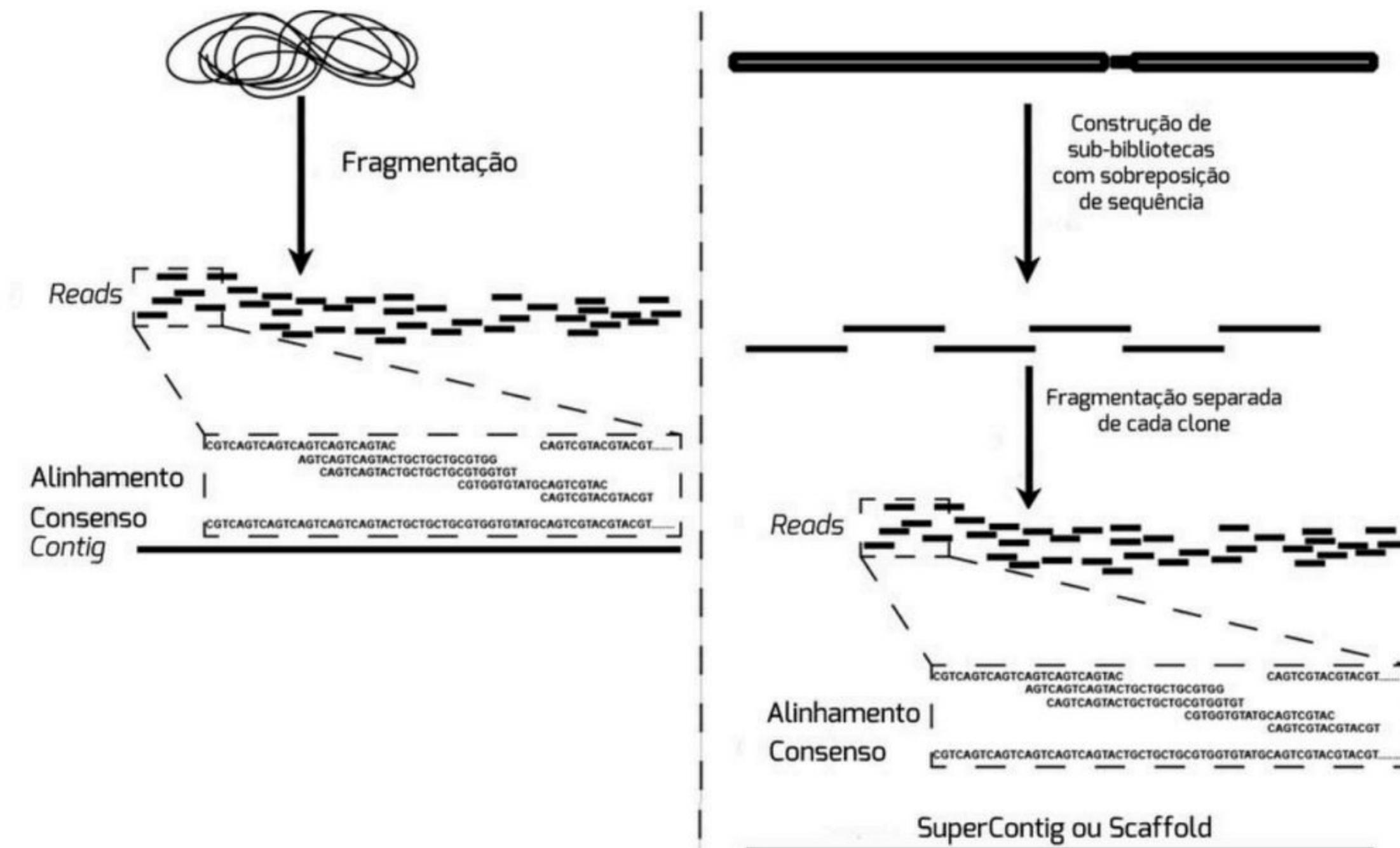
IBM'S DNA TRANSISTOR

Estratégias para a montagem de genomas de novo, tipos de dados. Avaliação de montagens



Estratégias para a montagem de genomas de novo, tipos de dados. Avaliação de montagens

ESTRATEGIAS PARA MONTAGEM COM SHORT READS



Estratégias para a montagem de genomas de novo, tipos de dados. Avaliação de montagens

PROGRAMAS DE MONTAGEM DE SHORT READS



PLATAFORMA ONLINE PARA MONTAGEM DE GENOMAS



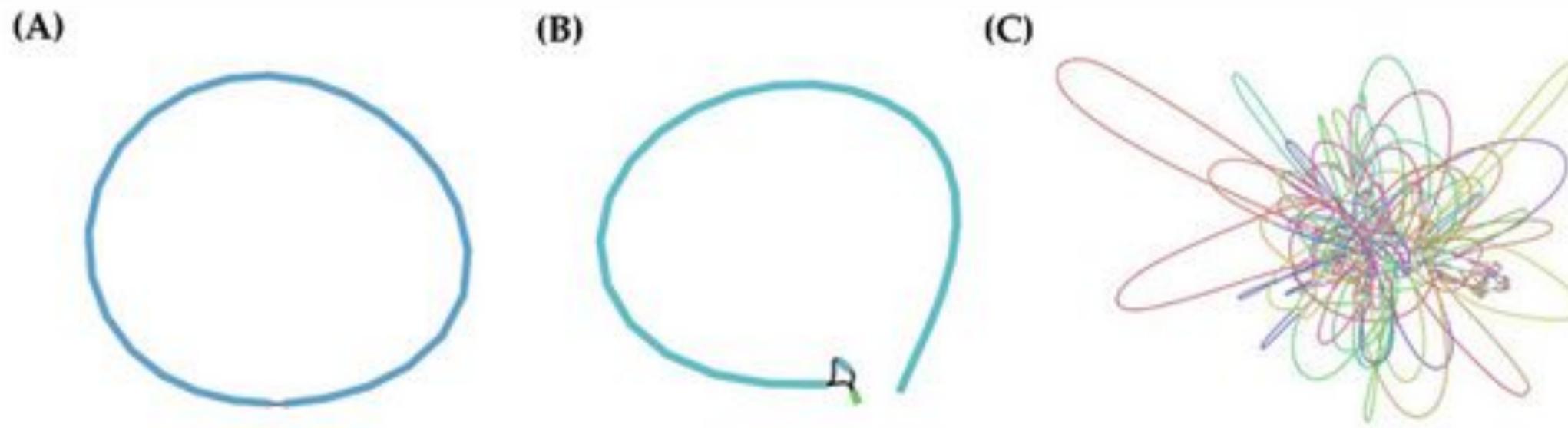
Estratégias para a montagem de genomas de novo, tipos de dados. Avaliação de montagens

ESTRATEGIAS PARA MONTAGEM COM LONG READS



ESTRATEGIAS PARA MONTAGEM COM LONG READS

Figure 2. Comparison of results of independent assembly strategies. (A) Genome assembled with nanopore reads; (B) longest contig assembled with PacBio reads; (C) genome assembled with Illumina reads. Plots were obtained by using Bandage on the "assembly_graph.gfa" output file from SPAdes or the "contig.gfa" output file from Canu. Connections between contigs represent overlaps between contig ends.



Estratégias para a montagem de genomas de novo, tipos de dados. Avaliação de montagens

PROGRAMAS DE MONTAGEM DE LONG READS

COM CORREÇÃO DE ERROS

CANU

FALCON

HGAP

SEM CORREÇÃO DE ERROS

HINGE

MINIMAP

TULIP

Estratégias para a montagem de genomas de novo, tipos de dados. Avaliação de montagens

PROGRAMAS DE MONTAGEM DE LONG READS

PARA MAPEAMENTOS

BWA-MEM

minialign/minimap

BBMap/BBTools

PARA CORREÇÃO DE ERROS

Frame-Pro

MINIMAP

TULIP

PLATAFORMA ONLINE PARA MONTAGEM DE GENOMAS

GALAXY

Estratégias para a montagem de genomas de novo, tipos de dados. Avaliação de montagens

PROGRAMAS DE MONTAGEM DE LONG READS

POLIMENTO DE SEQUENCIAS

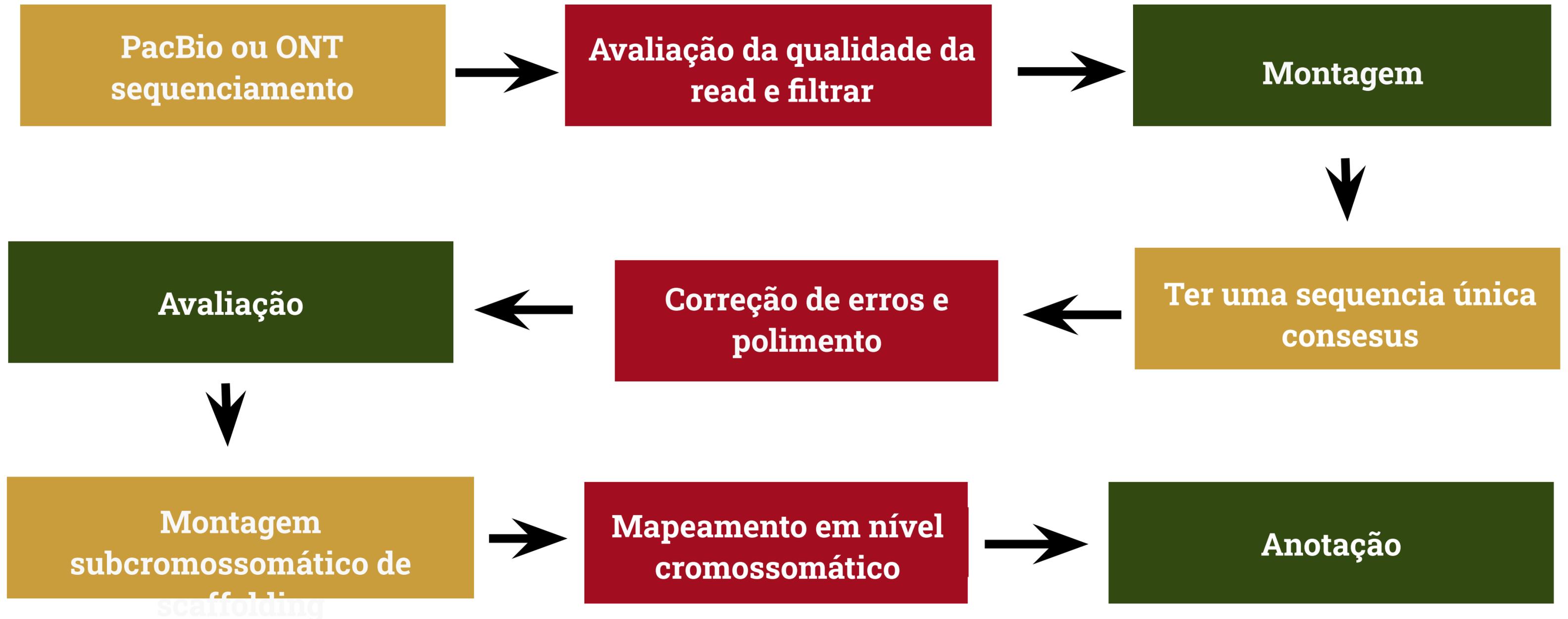
Nanopolish

Racon

Pilon

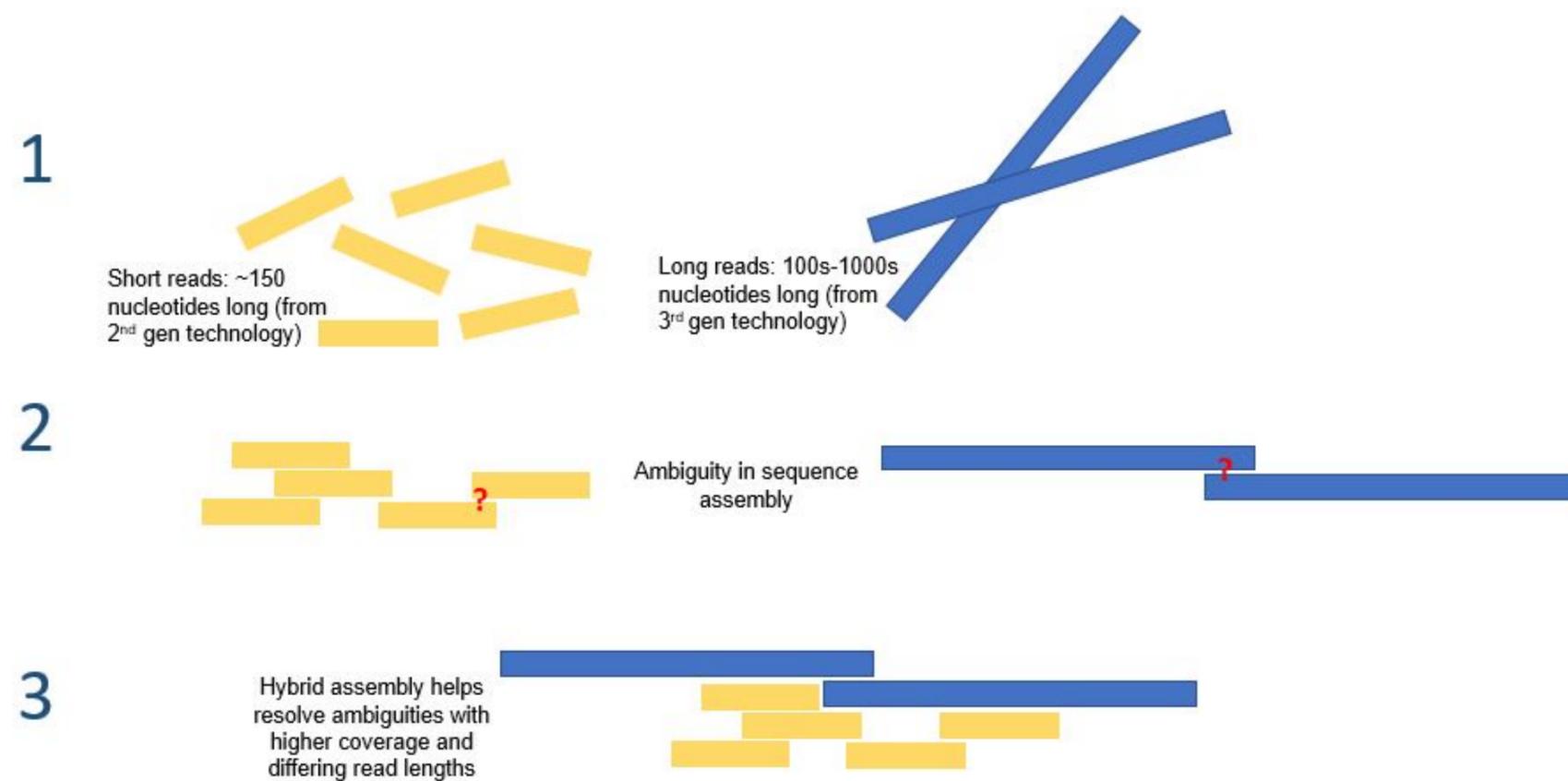
Estratégias para a montagem de genomas de novo, tipos de dados. Avaliação de montagens

RECOMENDAÇÃO DE PIPELINE



Estratégias para a montagem de genomas de novo, tipos de dados. Avaliação de montagens

ESTRATEGIAS PARA MONTAGEM HIBRIDAS



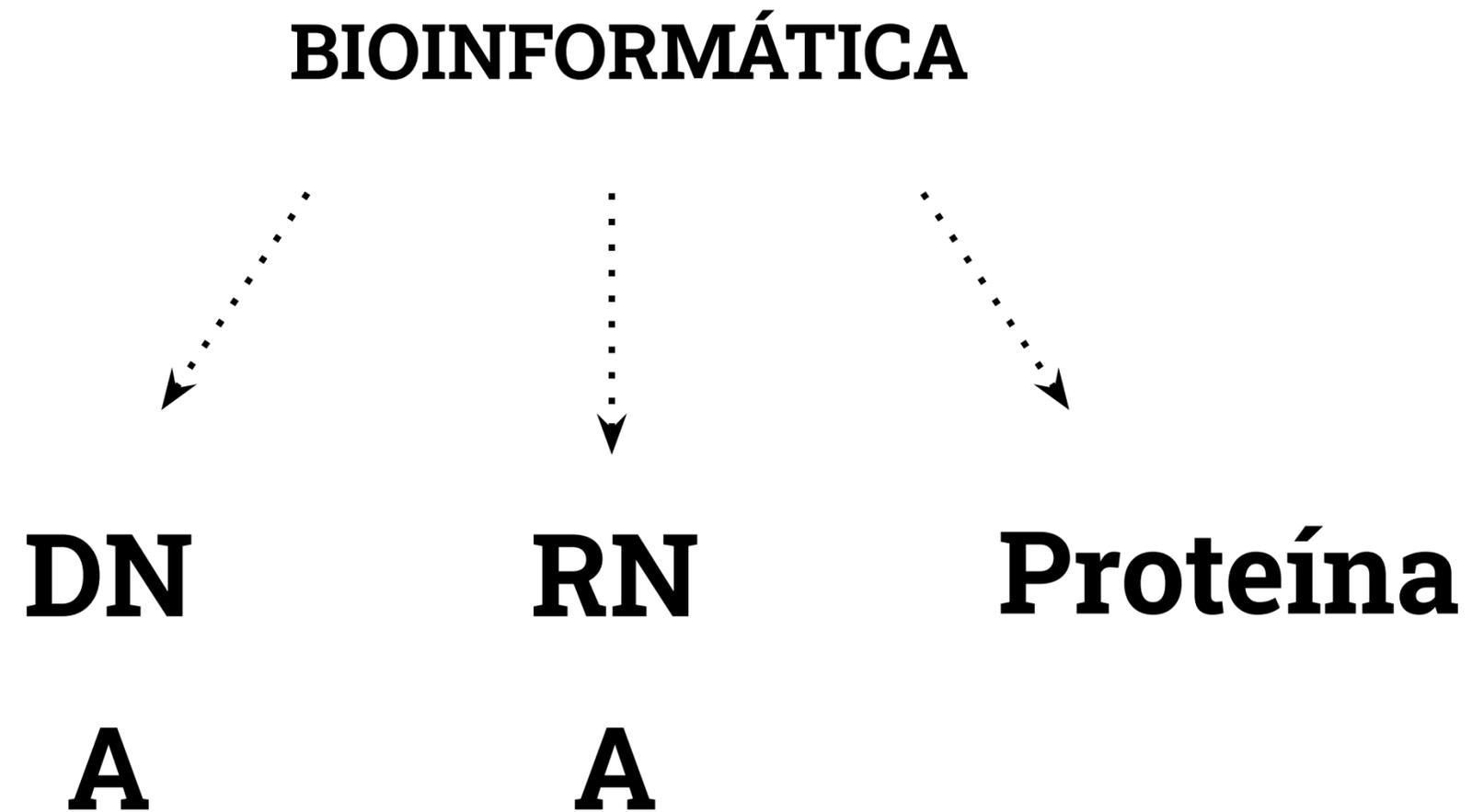
- Cobertura completa e contiguidade
- Illumina + PacBio
- Útil para genomas de plantas poliploide
- Correção de erros
- Melhoria na estrutura do genoma

Wang et al.,
2012

Unicycler

Racon

Estratégias para a montagem de genomas de novo, tipos de dados. Avaliação de montagens



Esses dados são armazenados em formatos digitais

TIPOS DE ARQUIVOS GERADOS

FASTA

- Descrito no final dos anos 80;
 - Com o tempo evoluiu por consenso;
 - Difícil de lidar com linhas longas sem quebra de linha;
 - Formato utilizado em genomas de referência.
-

Estratégias para a montagem de genomas de novo, tipos de dados. Avaliação de montagens

TIPOS DE ARQUIVOS GERADOS

Arquivo

FASTA

```
>NG_008670.1:5001-38170 Homo sapiens paired box 6 (PAX6)
ACCCTCTTTTCTTATCATTGACATTTAAACTCTGGGGCAGGTCCTCGCGTAGAACGCGGGCTGTCAGATCT
GCCACTTCCCCTGCCGAGCGGCGGTGAGAAGTGTGGGAACCGGCGCTGCCAGGCTCACCTGCCTCCCCGC
CCTCCGCTCCCAGGTAACCGCCCCGGGCTCCGGCCCCGGCCCCGGCTCGGGGGCCCGCGGGGGCCTCTCCGCTG
CCAGCGACTGCTGTCCCCAAATCAAAGCCCCGCCCCAAGTGGCCCCGGGGCTTGATTTTTTGCTTTTAAAAG
GAGGCATACAAAGATGGAAGCGAGTTACTGAGGGAGGGATAGGAAGGGGGGTGGAGGAGGGACTTGTCCTT
TGCCGAGTGTGCTCTTCTGCAAAAGTAGCAAAATGTTCCACTCCTAAGAGTGGACTTCCAGTCCGGCCCT
GAGCTGGGAGTAGGGGGCGGGAGTCTGCTGCTGCTGTCTGCTAAAGCCACTCGCGACCGCGAAAAATGCA
GGAGGTGGGGACGCACCTTGCATCCAGACCTCCTCTGCATCGCAGTTCACGACATCCACGCTTGGGAAAG
TCCGTACCCGCGCCTGGAGCGCTTAAAGACACCCTGCCGCGGGTCGGGCGAGGTGCAGCAGAAGTTCCC
CGGGTTGCAAAGTGCAGATGGCTGGACCGCAACAAAGTCTAGAGATGGGGTTCGTTTCTCAGAAAGACGC
```

TIPOS DE ARQUIVOS GERADOS

FASTQ

- Contém sequências + pontuações de qualidade;
 - Formato de dados perpetuado e amplamente aceito;
 1. **Sanger standard: fastq-sanger;**
 2. **Solexa/Illumina: fastq-solexa;**
 3. **Illumina 1.3+: fastq-illumina**
 - Possui variantes:
-

TIPOS DE ARQUIVOS GERADOS

FASTQ

```
@SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGGGGCTTTTTTTGTTTGGAAACCGAAAGG
GTTTGAATTTCAAACCCTTTTCGGTTTCCAACCTTCCAA
AGCAATGCCAATA
+SRR014849.1 EIXKN4201CFU84 length=93
3+&$#"7F@71,'";C?,B;?6B;:EA1EA
1EA5'9B:?:#9EA0D@2EA5':>5?:%A;A8A;?9B;D@
/=<?7=9<2A8==
```

@title and optional description
sequence line(s)
+optional repeat of title line
quality line(s)

- (1) Contém a informação do identificador da sequência iniciando com "@";
- (2) Contém a informação da sequência de bases;
- (3) A linha iniciada por "+" indica o fim da sequência de bases e o início das informações de qualidade na próxima linha.

TIPOS DE ARQUIVOS GERADOS

```
Identifier | @HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
Sequence  | TTAATTGGTAAATAAATCTCCTAATAGCTTAGATNTTACCTTNNNNNNNNNTAGTTTCTTGAGA
+ sign & identifier | +HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
Quality scores | efcffffcfeefffcfffffddf`feed]` ]_Ba_^__[YBBBBBBBBBRTT\]] [] dddd`
```

Base T
phred Quality] = 29

Fastq

- Formato padrão nas pipelines de montagem de genomas *de novo*.
- Usado em análises de bioinformática, como controle de qualidade (FastQC).

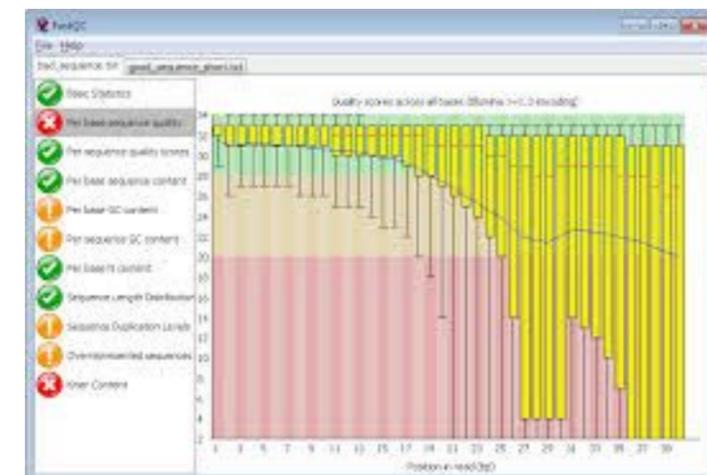
TIPOS DE ARQUIVOS GERADOS

```
Identifier | @HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
Sequence  | TTAATTGGTAAATAAATCTCCTAATAGCTTAGATNTTACCTTNNNNNNNNNTAGTTTCTTGAGA
+ sign & identifier | +HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
Quality scores | efcfffffcfeefffcfffffddf`feed]` ]_Ba_^__[YBBBBBBBBBBRTT\]] [] dddd`
```

Base T
phred Quality] = 29

Fastq

- Formato padrão nas pipelines de montagem de genomas *de novo*.
- Usado em análises de bioinformática, como controle de qualidade (FastQC).



TIPOS DE ARQUIVOS GERADOS



BED

O QUE É?

O arquivo .BED é um formato simples e flexível que representa regiões genômicas.

Delimita intervalos ou segmentos no genoma.

ESTRUTURA DO ARQUIVO BED

Cada linha do arquivo .BED tem, no mínimo, três colunas:

1. Nome do cromossomo (ex.: chr1, chrX).
2. Início da região (posição inicial no cromossomo, 0-based).
3. Fim da região (posição final no cromossomo, 1-based).

TIPOS DE ARQUIVOS GERADOS



BED

USOS:

Utilizado para delimitar regiões de interesse no genoma, como exons, genes, regiões promotoras, hotspots de mutação, etc.

CONTEÚDO DO ARQUIVO:

Contém informações sobre as anotações em dados genômicos, especificamente onde inicia e onde termina uma característica ou uma região específica de DNA.

TIPOS DE ARQUIVOS GERADOS

BED

| 1 | #chrom | start | end | name | score | strand |
|----|--------|-------|-------|---------|-------|--------|
| 2 | chr1 | 1000 | 5000 | geneA | 960 | + |
| 3 | chr1 | 5500 | 9000 | geneB | 850 | - |
| 4 | chr1 | 15000 | 16000 | geneC | 780 | + |
| 5 | chr2 | 100 | 800 | regionX | 1000 | + |
| 6 | chr2 | 1200 | 1500 | regionY | 900 | + |
| 7 | chr2 | 3000 | 5000 | regionZ | 1000 | + |
| 8 | chr3 | 10000 | 11000 | geneD | 970 | - |
| 9 | chr3 | 11500 | 12000 | geneE | 750 | + |
| 10 | chr3 | 2000 | 4000 | regionA | 950 | + |
| 11 | chr4 | 200 | 400 | geneF | 900 | - |
| 12 | chr4 | 500 | 600 | geneG | 600 | + |
| 13 | chr5 | 2500 | 3000 | geneH | 800 | + |
| 14 | chr5 | 7000 | 8000 | geneI | 850 | + |
| 15 | chr5 | 9500 | 10000 | regionB | 920 | - |

- O formato pode conter colunas adicionais que armazenam outras informações, como o nome da região, escore de confiança ou a orientação da sequência (+ ou -).

TIPOS DE ARQUIVOS GERADOS

VCF

O QUE É?

O VCF é um formato de arquivo que armazena informações sobre variantes genéticas identificadas em um organismo, como polimorfismos de nucleotídeo único (SNPs) e outras variantes.

TIPOS DE ARQUIVOS GERADOS

A large, light gray circular graphic on the left side of the slide contains the text 'VCF' in a dark red, bold, sans-serif font. The text is centered within the circle.

VCF

PRINCIPAIS CARACTERÍSTICAS:

- **Variantes SNPs:** O VCF é frequentemente utilizado para representar variantes SNP, mas também pode incluir indels (inserções e deleções) e outros tipos de variantes genéticas;
- **Versionado:** O formato VCF é versionado, permitindo a inclusão de metadados e definições que especificam a versão do formato e os parâmetros utilizados na chamada de variantes;
- **Comparação ao Genoma de Referência:** Cada variante no arquivo é geralmente apresentada em relação a um genoma de referência, o que permite identificar a posição e o tipo da variante em comparação ao genoma padrão;

TIPOS DE ARQUIVOS GERADOS

VCF

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
1 ##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
2 ##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
3 #CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA00001
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ 0|0:48:1:51,51
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0|0:49:3:58,50
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T GT:GQ:DP:HQ 0|0:54:7:56,60
20 1234567 microsat1 GTC G,GTCT 50 PASS NS=3;DP=9;AA=G GT:GQ:DP 0/1:35:4
```

(1) Cabeçalho de metadados, (2) Cabeçalho e (3) Linhas contendo os dados e suas devidas anotações.

O VCF é amplamente utilizado em estudos genômicos e pesquisas de associação genômica (GWAS), facilitando a análise e comparação de variantes entre indivíduos e populações.

TIPOS DE ARQUIVOS GERADOS

A large, light gray circular graphic on the left side of the slide contains the text 'FAST5' in a bold, dark red font. The text is centered within the circle, which is outlined with a thin red border.

FAST5

O QUE É?

O FAST5 é um formato de arquivo hierárquico usado para armazenar dados gerados por sequenciadores de nanopore, como os da Oxford Nanopore Technologies.

TIPOS DE ARQUIVOS GERADOS



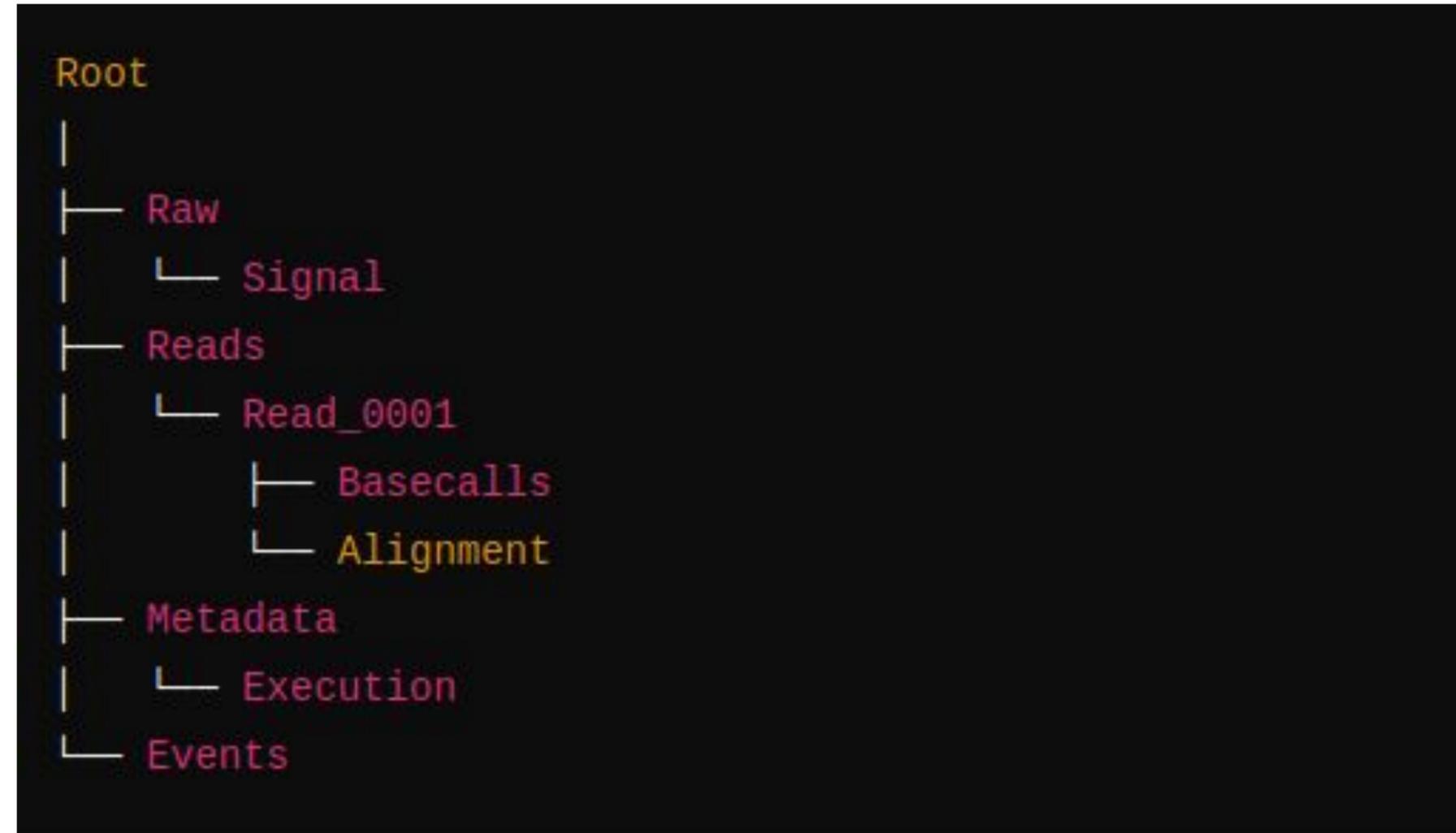
FAST5

ESTRUTURA DO ARQUIVO VCF:

- **Hierárquico:** Permitindo a organização eficiente de dados em diferentes níveis, como informações de sequência, qualidade e metadados associados.
- **Esquema Específico:** Baseado no formato HDF5 (Hierarchical Data Format version 5), que é otimizado para o armazenamento e gerenciamento de grandes volumes de dados.
- **Armazenamento de Grandes Dados:** É projetado para lidar com grandes conjuntos de dados gerados durante o sequenciamento, incluindo informações detalhadas sobre as leituras e os sinais elétricos coletados pelo sequenciador.
- **Biblioteca Exclusiva:** O formato FAST5 é exclusivo para Nanopore e, por isso, não é amplamente utilizado fora desse contexto.

TIPOS DE ARQUIVOS GERADOS

Em um diagrama hierárquico, seria algo assim:



Essa estrutura facilita o armazenamento de diferentes tipos de dados de maneira eficiente, mas também pode ser complexa para processamento, por isso a análise de arquivos FAST5 às vezes pode ser lenta

FAST5

TIPOS DE ARQUIVOS GERADOS

FAST5

The screenshot shows the HDFView 2.9 interface. The main window displays a table with two columns: '0' and '1'. The data consists of pairs of integers, such as (342, 559), (343, 569), (344, 559), (345, 571), (346, 571), (347, 595), (348, 591), (349, 574), (350, 628), (351, 571), (352, 591), (353, 574), (354, 554), (355, 574), (356, 532), (357, 407), (358, 390), (359, 398), (360, 391), (361, 395), (362, 393), (363, 417), (364, 384), (365, 400), (366, 389), (367, 382), (368, 376), (369, 399), (370, 447), (371, 382), (372, 395), (373, 376), (374, 383), (375, 381), (376, 372), (377, 378), (378, 373), (379, 371), (380, 370), (381, 384), (382, 373), (383, 383), (384, 375), (385, 386), (386, 377), (387, 548), (388, 539), (389, 543), (390, 517), (391, 553), (392, 534), (393, 537), (394, 517), (395, 428), (396, 417), (397, 418), (398, 418), (399, 430), (400, 430), (401, 430), (402, 418), (403, 420), (404, 426).

Nesta fase, o arquivo FAST5 tem apenas um conjunto de dados que é o conjunto de dados "Sinal".

The screenshot shows the HDFView 2.9 interface with a table of analysis events. The table has columns for 'start', 'length', 'mean', and 'stdv'. The data includes various analysis results such as 'Basecall_ID_000', 'Basecall_2D_000', 'BaseCalled_2D', 'HairpinAlign', 'Calibration_Strand_000', 'basecall_2d', 'calibration_stra', 'components', 'event_detection', 'general', 'hairpin_align', 'post_processing', 'split_hairpin', 'Summary', 'Basecall_2D_000', 'BaseCalled_2D', 'Configuration', 'Reads', 'Read_939', 'Events', 'Summary', 'Hairpin_Split_000', 'Raw', 'Reads', 'Read_939', 'Signal', 'UniqueGlobalkey', 'channel_id', 'context_tags', and 'tracking_id'. The table contains 63 rows of data.

Vários arquivos de log para as diferentes análises e ainda contém o conjunto de dados de sinal bruto.

TIPOS DE ARQUIVOS GERADOS

```
hugh@138Alpha:~/Documents/ONT_exp118 $ ls -l ~/home3/ont/lambda_fc1/uploads/vqb_20170118_FRFAS44402_MSI9940_sequencing_run_lambdacontrol_10012017_23602_ch9_read984_strand.fast5
/
/Raw Group
/Raw/Beads Group
/Raw/Beads/Read_984 Group
/Raw/Beads/Read_984/Signal Dataset (84995/Inf)
/UniqueGlobalKey Group
/UniqueGlobalKey/channel_id Group
/UniqueGlobalKey/context_tags Group
/UniqueGlobalKey/tracking_id Group
```

```
RPFS ~/home3/ont/lambda_fc1/downloads/pass/vqb_20170118_FRFAS44402_MSI9940_sequencing_run_lambdacontrol_10012017_23602_ch9_read939_strand.fast5 {
FILE_CONTENTS {
group /
group /Analyses
group /Analyses/Alignment_000
group /Analyses/Alignment_000/Aligned_2d
dataset /Analyses/Alignment_000/Aligned_2d/Fasta
dataset /Analyses/Alignment_000/Aligned_2d/SAM
group /Analyses/Alignment_000/Aligned_template
dataset /Analyses/Alignment_000/Aligned_template/Fasta
dataset /Analyses/Alignment_000/Aligned_template/SAM
group /Analyses/Alignment_000/Configuration
group /Analyses/Alignment_000/Configuration/aggregator
group /Analyses/Alignment_000/Configuration/basecall_1d
group /Analyses/Alignment_000/Configuration/basecall_2d
group /Analyses/Alignment_000/Configuration/calibration_strand
group /Analyses/Alignment_000/Configuration/components
group /Analyses/Alignment_000/Configuration/general
group /Analyses/Alignment_000/Configuration/genome_mapping
group /Analyses/Alignment_000/Configuration/hairpin_align
group /Analyses/Alignment_000/Configuration/post_processing_3000Hz
group /Analyses/Alignment_000/Configuration/split_hairpin
dataset /Analyses/Alignment_000/Log
group /Analyses/Alignment_000/Summary
group /Analyses/Alignment_000/Summary/genome_mapping_2d
```

FAST5

A partir do sinal bruto, é possível listar todos os grupos recursivamente.

TIPOS DE ARQUIVOS GERADOS

E obter todos os dados e metadados para um determinado grupo

FAST5

```
EDF5 "/home3/ont/lambda_fc1/downloads/pass/vgb_20170110_FNFAB46402_MH19940_sequencing_run_lambdacontrol_10012017_23602_ch9_read939_strand.fast5" {
GROUP "/Raw/Reads/Read_939" {
  ATTRIBUTE "duration" {
    DATATYPE H5T_STD_U32LE
    DATASPACE SCALAR
    DATA {
      (0): 142677
    }
  }
  ATTRIBUTE "median_before" {
    DATATYPE H5T_IEEE_F64LE
    DATASPACE SCALAR
    DATA {
      (0): 225.326
    }
  }
  ATTRIBUTE "read_id" {
    DATATYPE H5T_STRING {
      STRSIZE 37;
      STRPAD H5T_STR_NULLTERM;
      CSET H5T_CSET_ASCII;
      CTYPE H5T_C_S1;
    }
    DATASPACE SCALAR
    DATA {
      (0): "9260274d-d570-4c5d-bdc1-95a9f365295f"
    }
  }
  ATTRIBUTE "read_number" {
    DATATYPE H5T_STD_U32LE
    DATASPACE SCALAR
    DATA {
      (0): 939
    }
  }
  ATTRIBUTE "start_mux" {
    DATATYPE H5T_STD_I32LE
    DATASPACE SCALAR
    DATA {
      (0): 3
    }
  }
  ATTRIBUTE "start_time" {
    DATATYPE H5T_STD_U64LE
    DATASPACE SCALAR
    DATA {
      (0): 47230594
    }
  }
  DATASET "Signal" {
    DATATYPE H5T_STD_I16LE
    DATASPACE SIMPLE ( ( 142677 ) / ( H5S_UNLIMITED ) )
    DATA {
      (0): 1216, 653, 494, 487, 468, 478, 510, 535, 506, 454, 476, 483, 475,
      (13): 488, 472, 505, 474, 474, 488, 485, 480, 493, 481, 479, 485, 481,
      (26): 472, 491, 493, 480, 480, 487, 477, 500, 484, 488, 486, 493, 458,
      (39): 480, 491, 487, 477, 489, 478, 485, 476, 489, 486, 488, 490, 480,
      (52): 480, 484, 493, 475, 486, 477, 478, 489, 481, 482, 492, 480, 474,
      (65): 486, 426, 483, 508, 486, 487, 479, 476, 486, 473, 485, 487, 484,
      (78): 456, 485, 484, 466, 466, 483, 484, 484, 474, 480, 498, 481, 484,
      (91): 483, 477, 479, 473, 488, 482, 480, 478, 496, 479, 490, 489, 483,
      (104): 487, 473, 477, 479, 478, 480, 474, 475, 472, 475, 486, 498, 503,
      (117): 481, 493, 485, 475, 489, 478, 487, 479, 480, 488, 491, 490, 487,
      (130): 481, 483, 479, 484, 483, 484, 480, 499, 488, 486, 474, 485, 428,
      (143): 474, 474, 486, 466, 486, 481, 479, 476, 485, 489, 482, 484, 491,
      (156): 481, 491, 473, 485, 488, 486, 479, 494, 486, 491, 477, 464, 488,
      (169): 489, 484, 495, 478, 493, 476, 494, 475, 477, 477, 487, 480, 482,
      (182): 474, 478, 477, 481, 480, 483, 481, 481, 469, 482, 478, 479, 445,
      (195): 482, 484, 477, 493, 489, 484, 481, 470, 481, 491, 482, 471, 487,
```

TIPOS DE ARQUIVOS GERADOS



O QUE É?

O SLOW5 é um formato desenvolvido para ser uma alternativa otimizada ao formato FAST5, criado para resolver alguns dos problemas de desempenho e armazenamento encontrados no FAST5.



TIPOS DE ARQUIVOS GERADOS

PRINCIPAIS CARACTERÍSTICAS

- **Baseado no FAST5:** Criado a partir do formato FAST5, codificando todas as informações presentes no FAST5, como os sinais brutos e metadados gerados pelo sequenciamento de nanopore.
- **Desempenho e Eficiência:** Foi projetado para ser mais rápido e eficiente em termos de espaço e desempenho de leitura/escrita, com objetivo de processar grandes volumes de dados de forma mais ágil.
- **Não Depende de Biblioteca Exclusiva:** Diferente do FAST5, que depende de bibliotecas específicas (como HDF5) para leitura e escrita de dados, o SLOW5 não depende de uma biblioteca única, o que o torna mais acessível e fácil de integrar em diferentes pipelines de bioinformática.

TIPOS DE ARQUIVOS GERADOS



```
# Cabeçalho com as informações das colunas
read_id  channel_number  start_time  signal

# Dados da primeira leitura
read_001  144              159684563   [300, 295, 290, 305, ...]
read_002  146              159684578   [298, 299, 300, 297, ...]
read_003  148              159684590   [305, 310, 308, 299, ...]

# Metadados podem incluir: device_id, run_id, duration, etc.
```

o SLOW5 se organiza de forma tabular e plana, sendo mais rápido para leitura e processamento.

| read_id | channel_number | start_time | signal |
|----------|----------------|------------|-----------------|
| read_001 | 144 | 159684563 | [300, 295, ...] |
| read_002 | 146 | 159684578 | [298, 299, ...] |
| read_003 | 148 | 159684590 | [305, 310, ...] |

Se fosse exibido em forma de tabela, os dados do SLOW5 poderiam ser organizados dessa forma.

TIPOS DE ARQUIVOS GERADOS



BLOW5

O QUE É?

O BLOW5 é um formato de dados binário que evolui do SLOW5, projetado para otimizar o armazenamento e o processamento de dados de sequenciamento de alta densidade, especialmente para sequenciamento de nanoporos.

TIPOS DE ARQUIVOS GERADOS



PRINCIPAIS CARACTERÍSTICAS

- **Codificação Binária:** Armazena dados em formato binário, semelhante ao SLOW5, proporcionando eficiência em espaço e desempenho.
- **Alocação de Espaço Eficiente:** Otimiza o uso de espaço ao reduzir a redundância de metadados, facilitando o acesso e o processamento de dados em grandes volumes.
- **Desempenho Aprimorado:** Projetado para alta performance, especialmente em ambientes de análise paralela, melhorando a velocidade e a eficiência das análises.

TIPOS DE ARQUIVOS GERADOS



SAM

O QUE É?

O SAM é um formato de arquivo amplamente utilizado para armazenar informações de alinhamento de leituras de sequenciamento contra uma sequência de referência genômica.



SAM

TIPOS DE ARQUIVOS GERADOS

PRINCIPAIS CARACTERÍSTICAS

- **Mapeamento de Leituras:** Realiza um mapeamento eficiente de leituras curtas ou longas contra grandes sequências de referência, como o genoma de um organismo.
- **Alinhamentos:** Armazena alinhamentos de leitura, incluindo informações detalhadas sobre a posição de cada leitura no genoma de referência, se ela se alinha perfeitamente ou contém indels (inserções ou deleções).
- **Escala de Conjuntos Grandes:** É projetado para lidar com grandes conjuntos de alinhamentos, com capacidade de armazenar dados de sequências de 1.011 pares de bases (ou mais).

TIPOS DE ARQUIVOS GERADOS

Um exemplo anotado do formato SAM

```

A
      10      20      30      40
Coor   12345678901234 5678901234567890123456789012345
ref    AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1      TTAGATAAAGGATA*CTG
+r002      aaaAGATAA*GGATA
+r003      gcctaAGCTAA
+r004      ATAGCT.....TCAGC
-r003      ttagctTAGGC
-r001/2      CAGCGGCAT
    
```

O resultado do alinhamento representado no formato SAM.

B

```

@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
    
```

Header section (points to @HD and @SQ lines)

Alignment section (points to the alignment lines)

QUAL (read quality; * meaning such information is not available)

| QNAME | FLAG | RNAME | POS | MAPQ | CIGAR | RNEXT | PNEXT | TLEN | SEQ | Optional fields |
|-------------------------------------|--|---|--------------------|-------------------|--|---|---|--|-----------------|------------------|
| (query template name, aka. read ID) | (indicates alignment information about the read, e.g. paired, aligned, etc.) | (reference sequence name, e.g. chromosome /transcript id) | (1-based position) | (mapping quality) | (summary of alignment, e.g. insertion, deletion) | (reference sequence name of the primary alignment of the NEXT read; for paired-end sequencing, NEXT read is the paired read; corresponding to the RNAME column) | (Position of the primary alignment of the NEXT read in the template; corresponding to the POS column) | (the number of bases covered by the reads from the same fragment. In this particular case, it's 45 - 7 + 1 = 39 as highlighted in Panel A). Sign: plus for leftmost read, and minus for rightmost read | (read sequence) | (TAG:TYPE:VALUE) |

É um formato de texto simples, projetado para ser eficiente em termos de espaço e de fácil leitura, facilitando a análise de grandes volumes de dados genômicos.



TIPOS DE ARQUIVOS GERADOS



BAM

O QUE É?

O BAM é o formato binário do SAM (Sequence Alignment/Map), criado para representar de forma mais eficiente os dados de alinhamento. Ele armazena exatamente as mesmas informações que o SAM, mas em um formato comprimido e otimizado.

TIPOS DE ARQUIVOS GERADOS



BAM

PRINCIPAIS CARACTERÍSTICAS

- **Representação Binária:** O BAM converte o conteúdo de um arquivo SAM, permitindo economizar espaço de armazenamento e aumentar a eficiência durante a leitura e gravação de dados.
- **Mesmas Informações que o SAM:** Todas as informações, como os nomes das leituras, posições no genoma, mapeamentos e pontuações de qualidade, são mantidas exatamente no formato BAM.
- **Compactação:** Com o formato BAM é possível uma redução significativa no tamanho do arquivo, especialmente para grandes conjuntos de dados de sequenciamento.

TIPOS DE ARQUIVOS GERADOS

BAM

samtools/samtools

Tools (written in C using htslib) for manipulating next-generation sequencing data



- Import SAM to BAM when @SQ lines are present in the header:

```
samtools view -bo aln.bam aln.sam
```

If @SQ lines are absent:

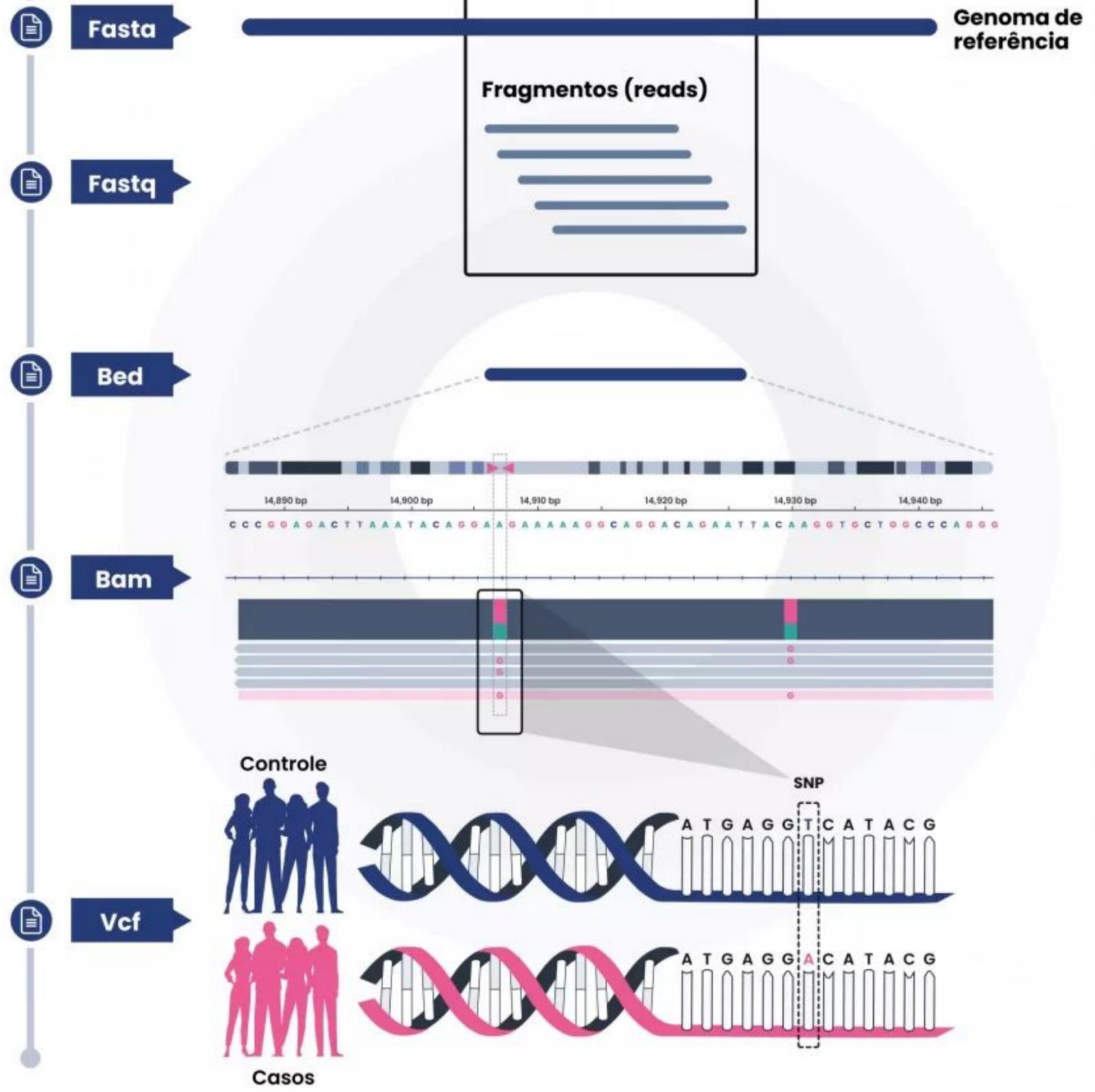
```
samtools faidx ref.fa  
samtools view -bt ref.fa.fai -o aln.bam aln.sam
```

where *ref.fa.fai* is generated automatically by the **faidx** command.

PRINCIPAIS CARACTERÍSTICAS

- **Representação Binária:** O BAM converte o conteúdo de um arquivo SAM, permitindo economizar espaço de armazenamento e aumentar a eficiência durante a leitura e gravação de dados.
- **Mesmas Informações que o SAM:** Todas as informações, como os nomes das leituras, posições no genoma, mapeamentos e pontuações de qualidade, são mantidas exatamente no formato BAM.
- **Compactação:** Com o formato BAM é possível uma redução significativa no tamanho do arquivo, especialmente para grandes conjuntos de dados de sequenciamento.

TIPOS DE ARQUIVOS GERADOS



Após o mapeamento dos reads (FASTQ);

Contra o genoma de referência (FASTA);

E identificação das regiões específicas (BED);

No mapeamento (SAM/BAM),

É possível obter as variantes nestas regiões (VCF).

AVALIAÇÃO DE MONTAGENS

Qualidade

Precisão

Identificar falhas

Garantindo que os dados sejam adequados para análises posteriores, como anotações funcionais e estudos evolutivos.

AVALIAÇÃO DE MONTAGENS

BASE ACCURACY



Refere-se à precisão da sequência em nível de nucleotídeos, ou seja, a capacidade de um método de sequenciamento de identificar corretamente as bases (A, T, C, G) em uma montagem genômica.

AVALIAÇÃO DE MONTAGENS

BASE ACCURACY



Refere-se à precisão da sequência em nível de nucleotídeos, ou seja, a capacidade de um método de sequenciamento de identificar corretamente as bases (A, T, C, G) em uma montagem genômica.

CONTINUITY



Avalia se as sequências montadas estão conectadas de maneira adequada, ou seja, se as contigs (fragmentos contíguos de DNA) ou scaffolds (conjuntos de contigs interligados) formam uma sequência contínua sem lacunas significativas.

AVALIAÇÃO DE MONTAGENS

QUAST

Quality Assessment Tool for Genome Assemblies by [CAB](#)

Fornecer métricas como tamanho total da montagem, tamanho do contig/scaffold maior, número de contigs/scaffolds, é N50 entre outros.

```
All statistics are based on contigs of size >= 500 bp, unless otherwise noted
(e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs)
.
Assembly                HFHC_hap02
# contigs (>= 0 bp)      3234
# contigs (>= 1000 bp)   3234
# contigs (>= 10000 bp)  3234
# contigs (>= 100000 bp) 2340
# contigs (>= 1000000 bp) 956
# contigs (>= 10000000 bp) 162
Total length (>= 0 bp)   6587868591
Total length (>= 1000 bp) 6587868591
Total length (>= 10000 bp) 6587868591
Total length (>= 100000 bp) 6536227698
Total length (>= 1000000 bp) 6061590638
Total length (>= 10000000 bp) 3077019577
# contigs                3234
Largest contig           77421221
Total length             6587868591
GC (%)                   44.68
N50                      9173517
N75                       4128966
L50                       185
L75                       452
# N's per 100 kbp        0.00
~
~
```

QUAST v.5.0.2

AVALIAÇÃO DE MONTAGENS

QUAST

Quality Assessment Tool for Genome Assemblies by CAB

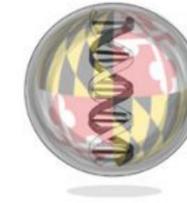
Fornecer métricas como tamanho total da montagem, tamanho do contig/scaffold maior, número de contigs/scaffolds, é N50 entre outros.

```
All statistics are based on contigs of size >= 500 bp, unless otherwise noted
(e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs)
.
Assembly HFHC_hap02
# contigs (>= 0 bp) 3234
# contigs (>= 1000 bp) 3234
# contigs (>= 10000 bp) 3234
# contigs (>= 100000 bp) 2340
# contigs (>= 1000000 bp) 956
# contigs (>= 10000000 bp) 162
Total length (>= 0 bp) 6587868591
Total length (>= 1000 bp) 6587868591
Total length (>= 10000 bp) 6587868591
Total length (>= 100000 bp) 6536227698
Total length (>= 1000000 bp) 6061590638
Total length (>= 10000000 bp) 3077019577
# contigs 3234
Largest contig 77421221
Total length 6587868591
GC (%) 44.68
N50 9173517
N75 4128966
L50 185
L75 452
# N's per 100 kbp 0.00
~
~
```

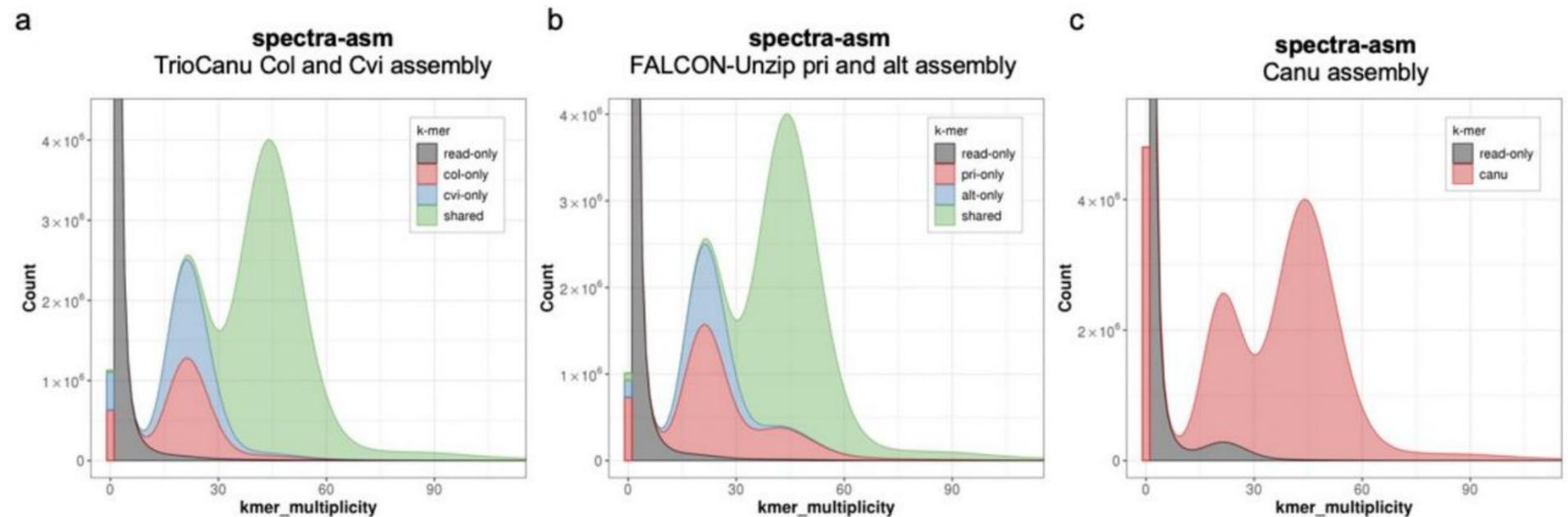
QUAST v.5.0.2

marbl/merqury

k-mer based assembly evaluation



Realiza a avaliação de completude e estatísticas de espectro de kmers e faseamento;



Merqury v.5.0.2

Gráficos de espectro de montagem Merqury para avaliar a completude k-mer.

AVALIAÇÃO DE MONTAGENS

PRECISÃO ESTRUTURAL



Avalia se as sequências montadas estão conectadas de maneira adequada, ou seja, se as contigs (fragmentos contíguos de DNA) ou scaffolds (conjuntos de contigs interligados) formam uma sequência contínua sem lacunas significativas.

AVALIAÇÃO DE MONTAGENS

PRECISÃO ESTRUTURAL



Avalia se as sequências montadas estão conectadas de maneira adequada, ou seja, se as contigs (fragmentos contíguos de DNA) ou scaffolds (conjuntos de contigs interligados) formam uma sequência contínua sem lacunas significativas.

| Quality category | Metric | Finished | VGP-2020 | VGP-2016 |
|---------------------|--------------------|--------------------|----------|----------|
| Structural accuracy | Reliable blocks | = Chr. NG50 | >10 Mb | >1 Mb |
| | False duplications | 0% | <1% | <5% |
| | Curation | Conflicts resolved | Manual | Manual |

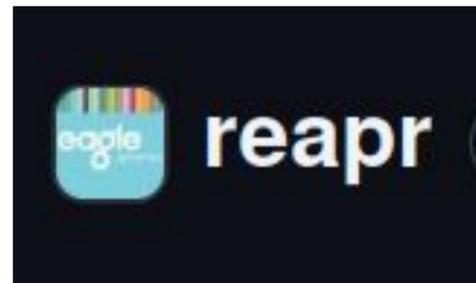
AVALIAÇÃO DE MONTAGENS

PRECISÃO ESTRUTURAL



Avalia se as sequências montadas estão conectadas de maneira adequada, ou seja, se as contigs (fragmentos contíguos de DNA) ou scaffolds (conjuntos de contigs interligados) formam uma sequência contínua sem lacunas significativas.

| Quality category | Metric | Finished | VGP-2020 | VGP-2016 |
|---------------------|--------------------|--------------------|----------|----------|
| Structural accuracy | Reliable blocks | = Chr. NG50 | >10 Mb | >1 Mb |
| | False duplications | 0% | <1% | <5% |
| | Curation | Conflicts resolved | Manual | Manual |



Avalia a precisão de uma montagem do genoma usando leituras de extremidade mapeadas, sem o uso de um genoma de referência para comparação.

AVALIAÇÃO DE MONTAGENS

COMPLETUDE FUNCIONAL



Avaliação da presença e funcionalidade de genes e elementos essenciais na montagem do genoma.

AVALIAÇÃO DE MONTAGENS

COMPLETUDE FUNCIONAL



Avaliação da presença e funcionalidade de genes e elementos essenciais na montagem do genoma.

PRINCIPAIS ASPECTOS:

1. IDENTIFICAÇÃO DE GENES;

2. COMPARAÇÃO COM GENOMAS
DE REFERÊNCIA;

AVALIAÇÃO DE MONTAGENS

COMPLETUDE FUNCIONAL



Avaliação da presença e funcionalidade de genes e elementos essenciais na montagem do genoma.

PRINCIPAIS ASPECTOS:

1. IDENTIFICAÇÃO DE GENES;

2. COMPARAÇÃO COM GENOMAS
DE REFERÊNCIA;

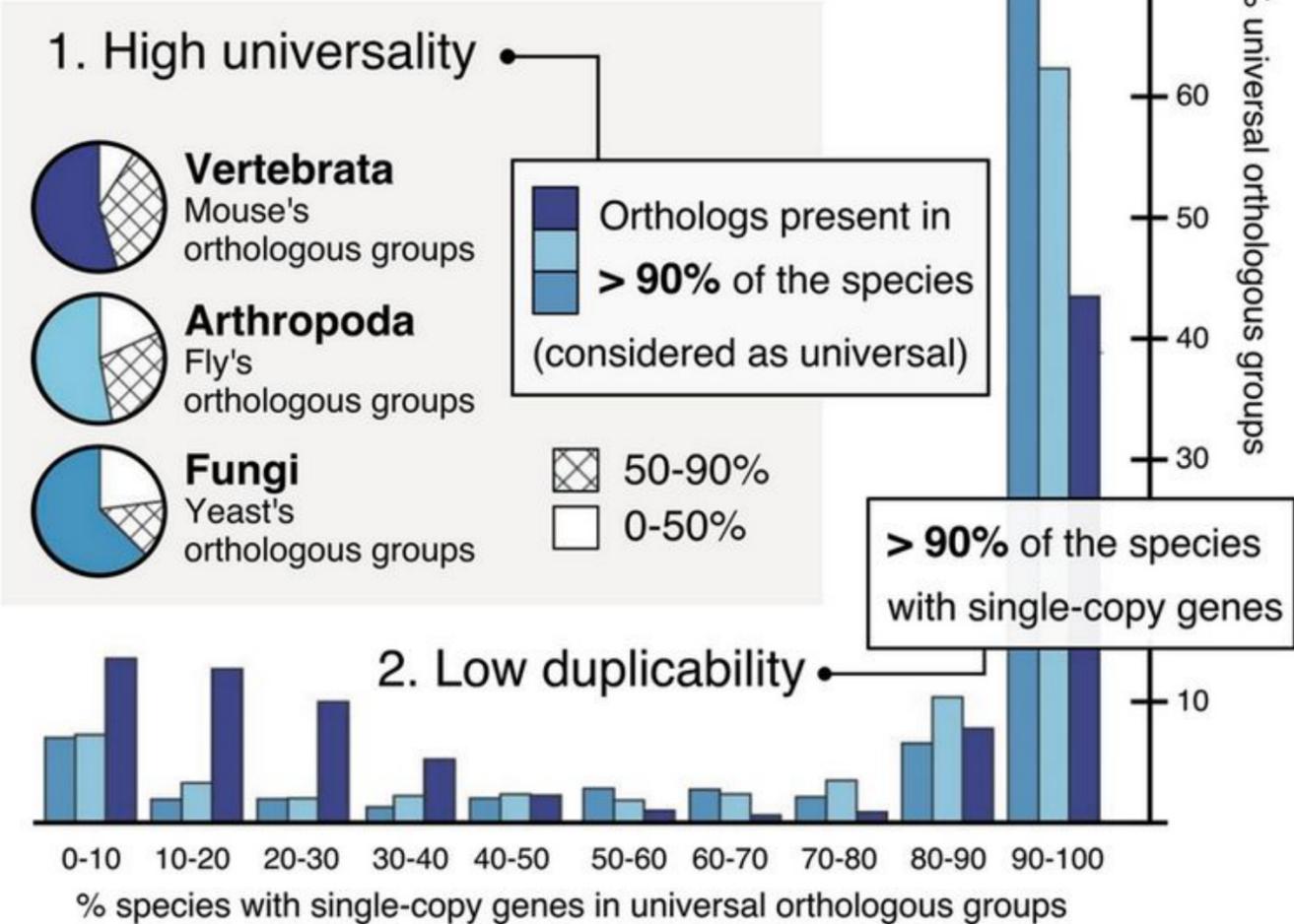
BUSCO

Avalia a qualidade das anotações do genoma e a integridade das montagens do genoma e fornece medidas quantitativas para avaliar a integridade do gene em termos de genes esperados encontrados em um conjunto de dados.

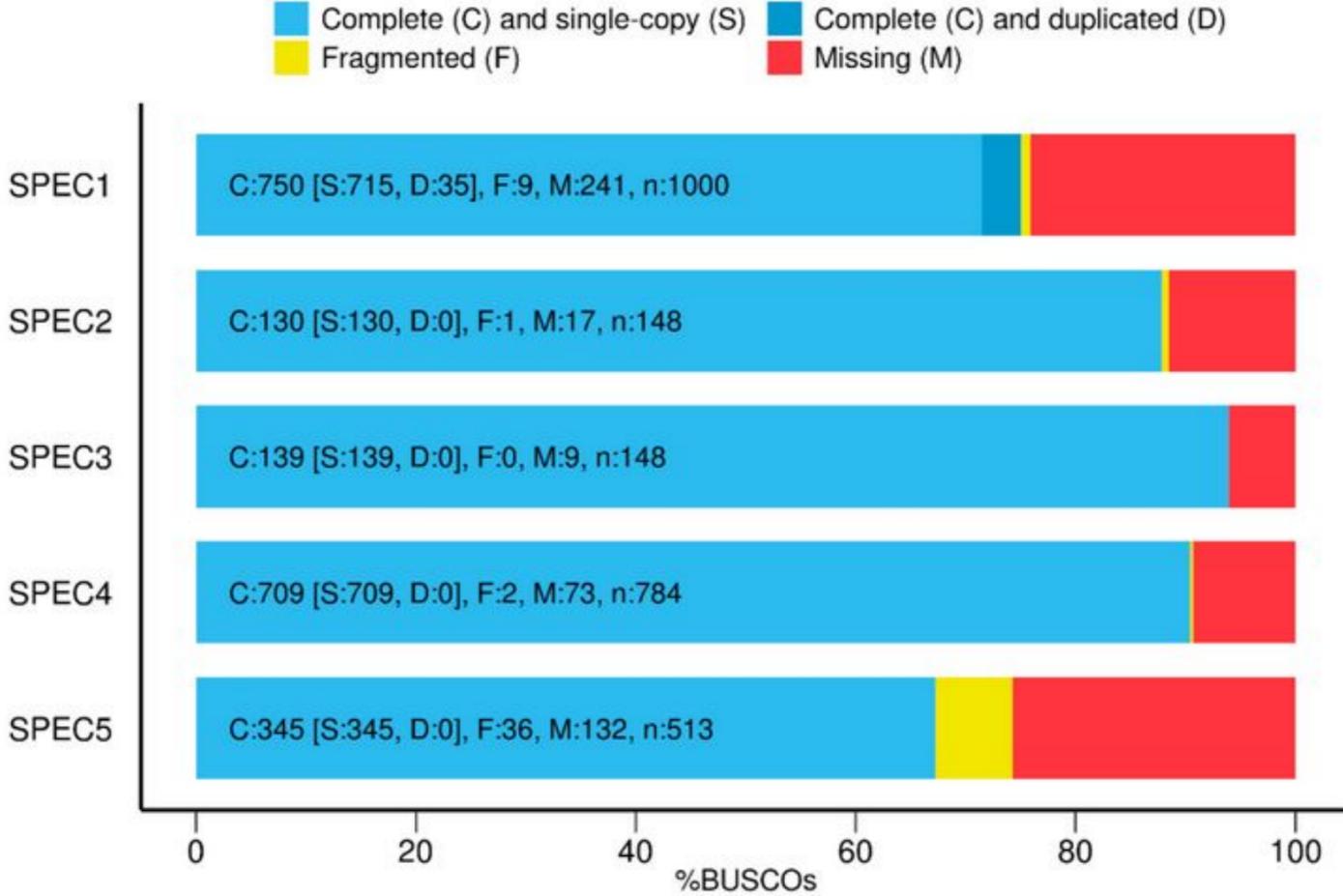
AVALIAÇÃO DE MONTAGENS

BUSCO

BUSCO sampling space



BUSCO Assessment Results



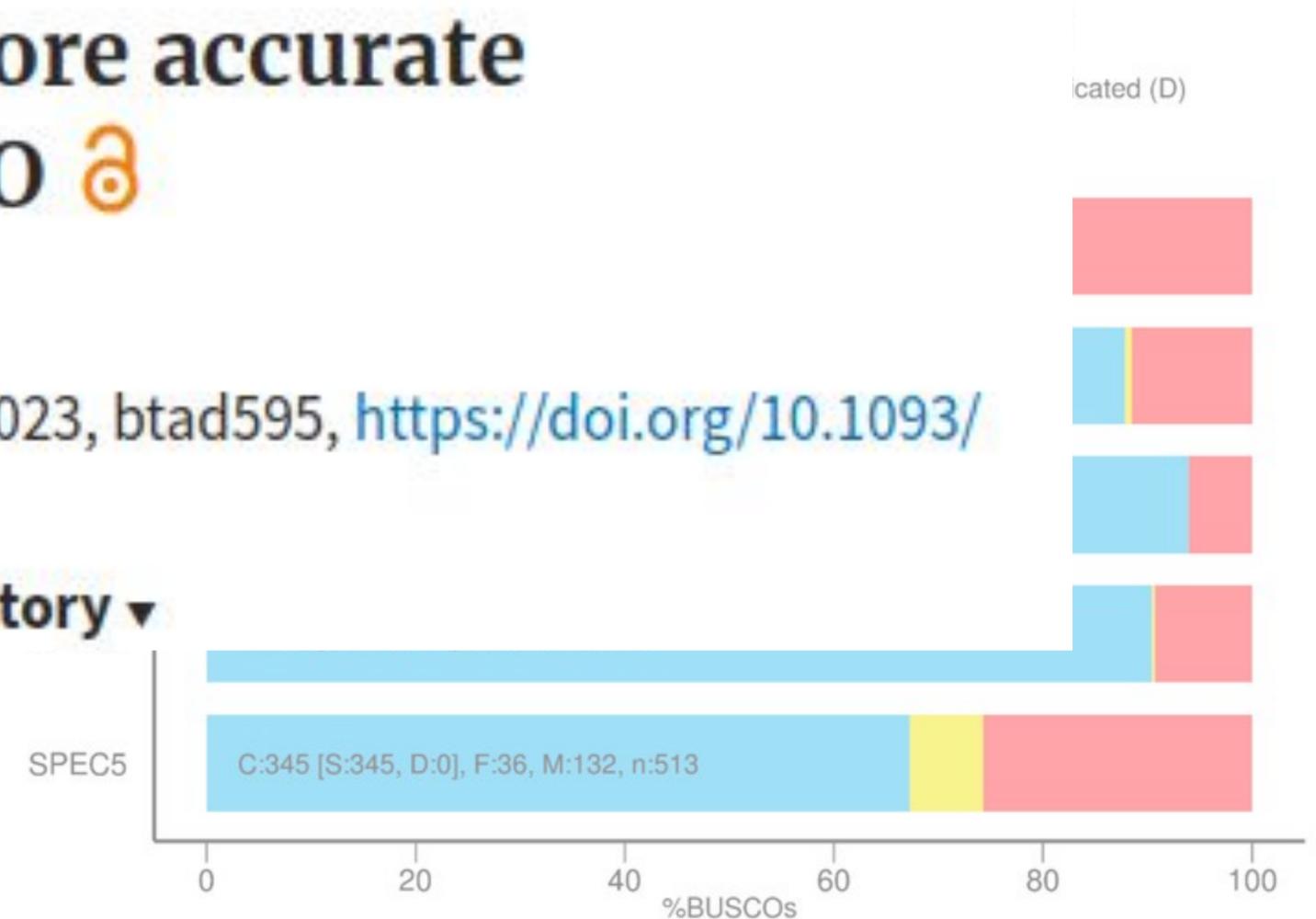
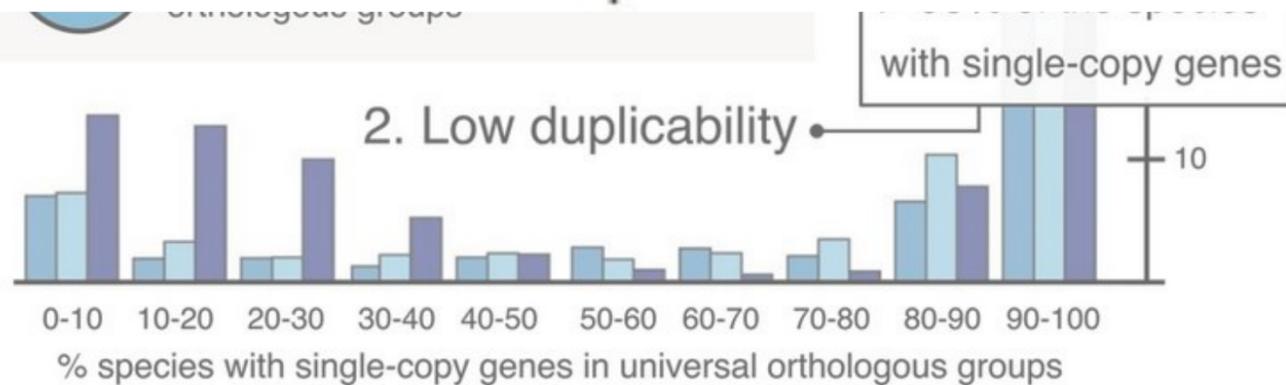
BUSCO

compleasm: a faster and more accurate reimplementation of BUSCO

Neng Huang, Heng Li

Bioinformatics, Volume 39, Issue 10, October 2023, btad595, <https://doi.org/10.1093/bioinformatics/btad595>

Published: 27 September 2023 Article history



AVALIAÇÃO DE MONTAGENS

Compleasm v0.2.6

Esses genes BUSCO são categorizados nas seguintes classes:

- **S (Single Copy Complete Genes)**: Os genes BUSCO que podem estar totalmente alinhados na montagem, com apenas uma cópia presente.
- **D (Douplementos Completos)**: Os genes BUSCO que podem estar completamente alinhados na montagem, com mais de uma cópia presente.
- **F (Genes fragmentados, subclasse 1)**: Os genes BUSCO que apenas uma parte do gene está presente na montagem, e o resto do gene não pode ser alinhado.
- **I (Genes fragmentados, subclasse 2)**: Os genes BUSCO em que uma seção do gene se alinha a uma posição na montagem, enquanto a parte restante se alinha com outra posição.
- **M (Genes Missing)**: Os genes BUSCO sem alinhamento presente na montagem.

**MONTAGEM DE GENOMA
USANDO A
PLATAFORMA GALAXY**



Estratégias para a montagem de genomas de novo, tipos de dados. Avaliação de montagens

Upload dos arquivos na Plataforma Galaxy

The screenshot displays the Galaxy web interface. At the top, the Galaxy logo is on the left, and navigation links for Workflow, Visualize, Data, Help, and User are in the center. A 'Using 0%' indicator is on the right. The left sidebar contains icons for Upload, Tools, Workflows, Workflow Invocations, Visualization, Histories, Notifications, and Settings. The main 'Tools' panel is open, showing a search bar and categories like 'GENERAL TEXT TOOLS' and 'GENOMIC FILE MANIPULATION'. The central workspace contains a text block and a banner for the 'Galaxy Training Academy'. The right sidebar shows the 'History' panel, which is currently empty.

Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources. You can install your own Galaxy by following the tutorial and choose from thousands of tools from the Tool Shed.

Join us at the Galaxy Training Academy
One week of Free, Global, Online Galaxy Training
Learn how to analyze your data in your scientific domain
WHEN?
07th - 11th October, 2024
<http://gxy.io/GTA1>
Learn about : Introduction to Galaxy, Assembly, Proteomics, Transcriptomics, Single Cell, Microbiome, Bacterial Genomics, Machine Learning, BY-COVID or Choose your own adventure

Galaxy version 24.1.3.dev0, commit b56ad8d12a77da51dc9129e4c80705dd46c19839

Estratégias para a montagem de genomas de novo, tipos de dados. Avaliação de montagens

Controle de qualidade das sequências na Plataforma

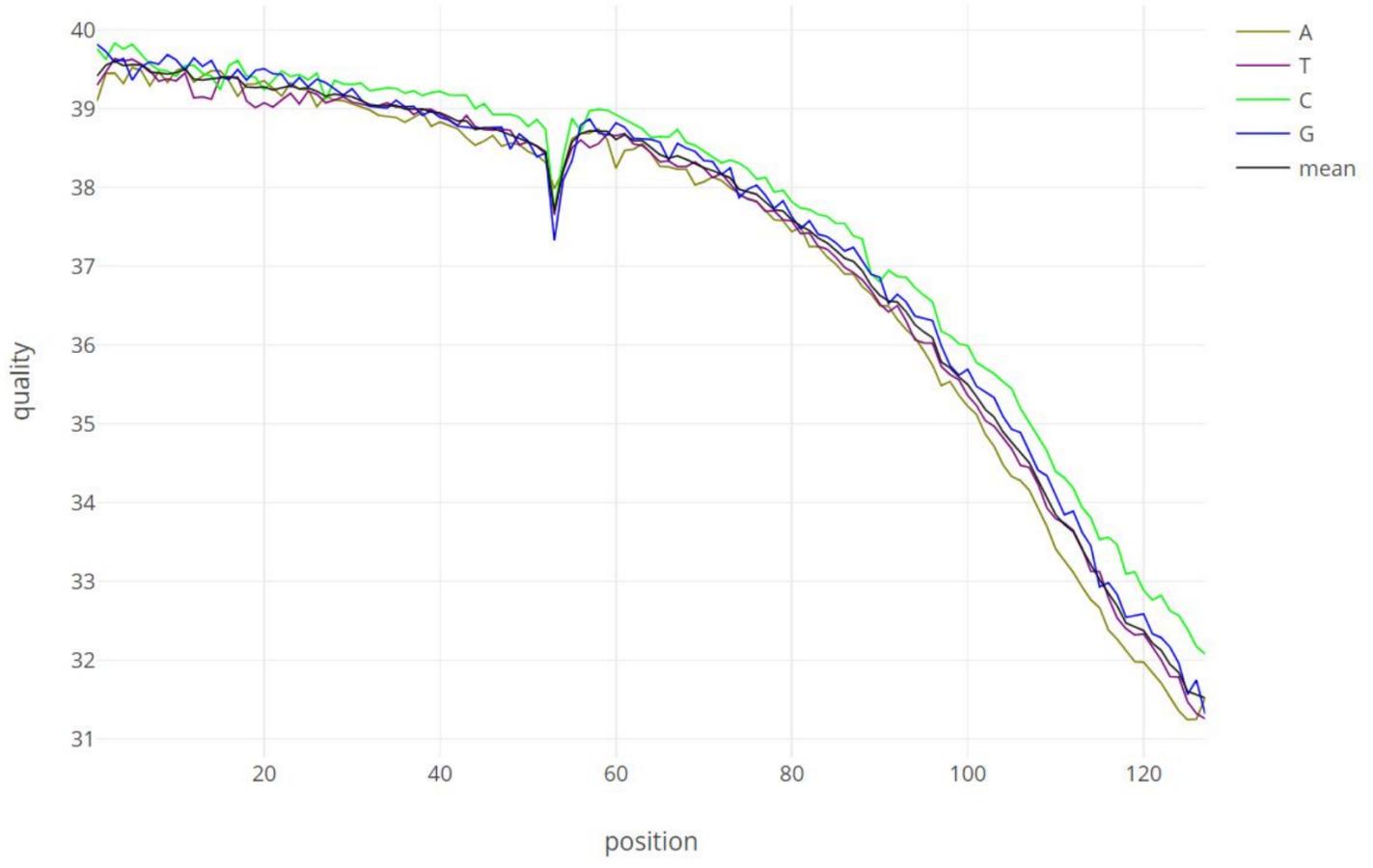
The screenshot displays the Galaxy web interface. At the top, the 'Galaxy' logo is on the left, and navigation links for 'Workflow', 'Visualize', 'Data', 'Help', and 'User' are in the center. A 'Using 0%' indicator is on the right. The left sidebar contains icons for 'Upload', 'Tools', 'Workflows', 'Workflow Invocations', 'Visualization', 'Histories', 'Notifications', and 'Settings'. The 'Tools' panel is active, showing a search for 'fastp' and a list of tools including 'fastp', 'fastpca', 'Map with minimap2', and 'fgsea'. The main content area is titled 'Edit Dataset Attributes' and has three tabs: 'Attributes', 'Datatypes', and 'Permissions'. The 'Attributes' tab is selected, showing fields for 'Name' (sweet-potato-chloroplast-illumina-reduced.fastq), 'Info' (uploaded fastqsanger file), 'Annotation' (optional), and 'Database/Build' (optional, unspecified (?)). 'Save' and 'Auto-detect' buttons are at the bottom. The right sidebar shows a 'History' panel with a search bar and a list of workflow steps, including 'Class test', 'ta 2: HTML report', '3: fastp on data 2: Read 1 output', '2: sweet-potato-chloroplast-nanopore-reduced.fastq', and '1: sweet-potato-chloroplast-illumina-reduced.fastq'.

Estratégias para a montagem de genomas de novo, tipos de dados. Avaliação de montagens

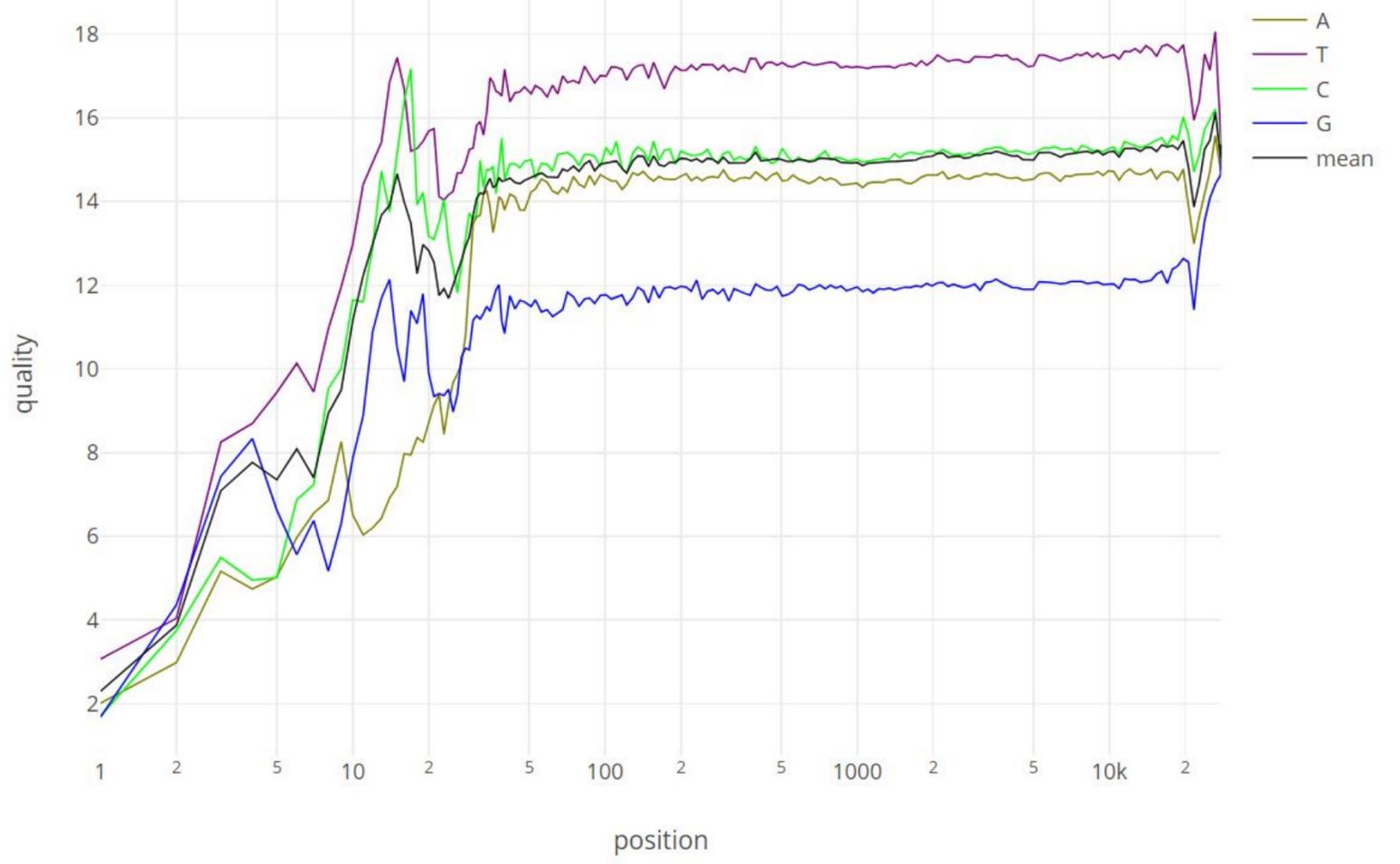
Controle de qualidade das sequências na Plataforma

Galaxy

Short reads



Long reads

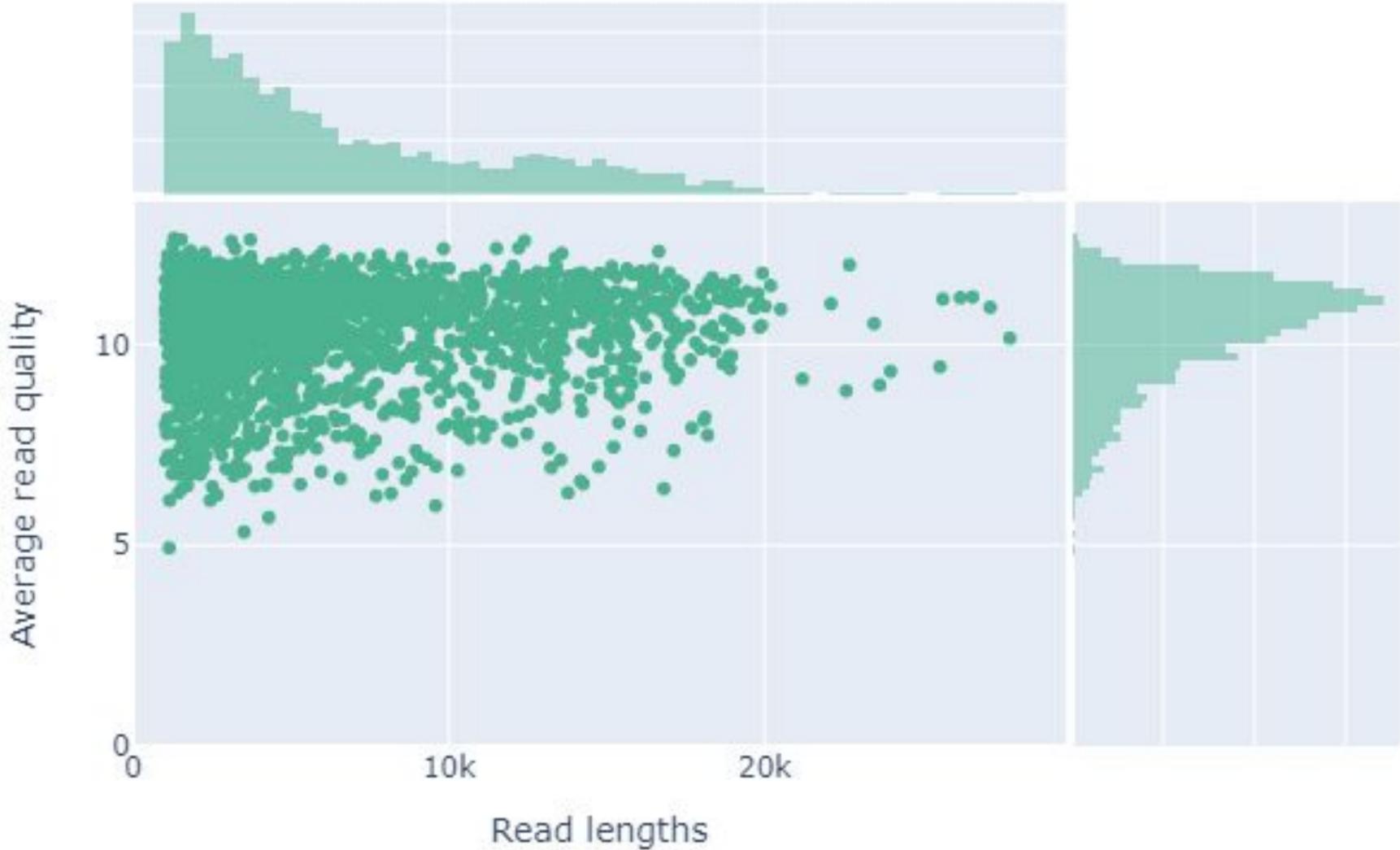


Estratégias para a montagem de genomas de novo, tipos de dados. Avaliação de montagens

Controle de qualidade das sequências na Plataforma

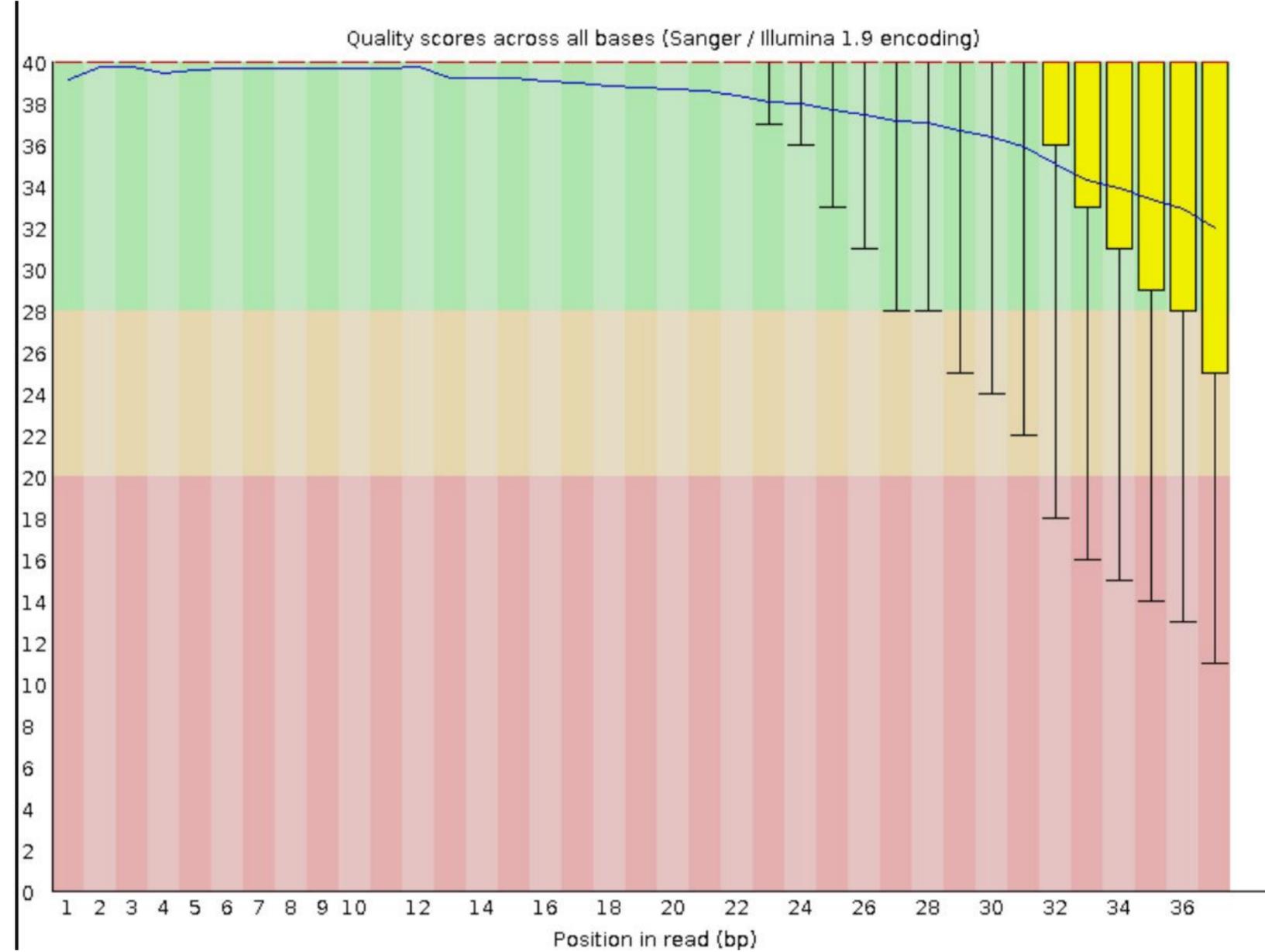
Galaxy

Read lengths vs Average read quality plot using dots



Estratégias para a montagem de genomas de novo, tipos de dados. Avaliação de montagens

Controle de qualidade das sequências na Plataforma Galaxy (fastQC)



Estratégias para a montagem de genomas de novo, tipos de dados. Avaliação de montagens

Controle de qualidade das sequências na Plataforma Galaxy

Galaxy Workflow Visualize (fastQC) Help User Using 0%

Tools trim

Trim sequences (Galaxy Version 1.0.2+galaxy2) Run Tool

Tool Parameters

Input file in FASTA or FASTQ format *

4: Trim Galore! on data 1: trimmed reads

accepted formats

First base to keep *

1

Last base to keep *

21

Additional Options

Email notification

No

Send an email notification when the job completes.

Run Tool

History

search datasets

Unnamed history

290 MB 7

7: Trim on data 6

6: FastQC on data 4: RawData

5: FastQC on data 4: Webpage

4: Trim Galore! on data 1: trimmed reads

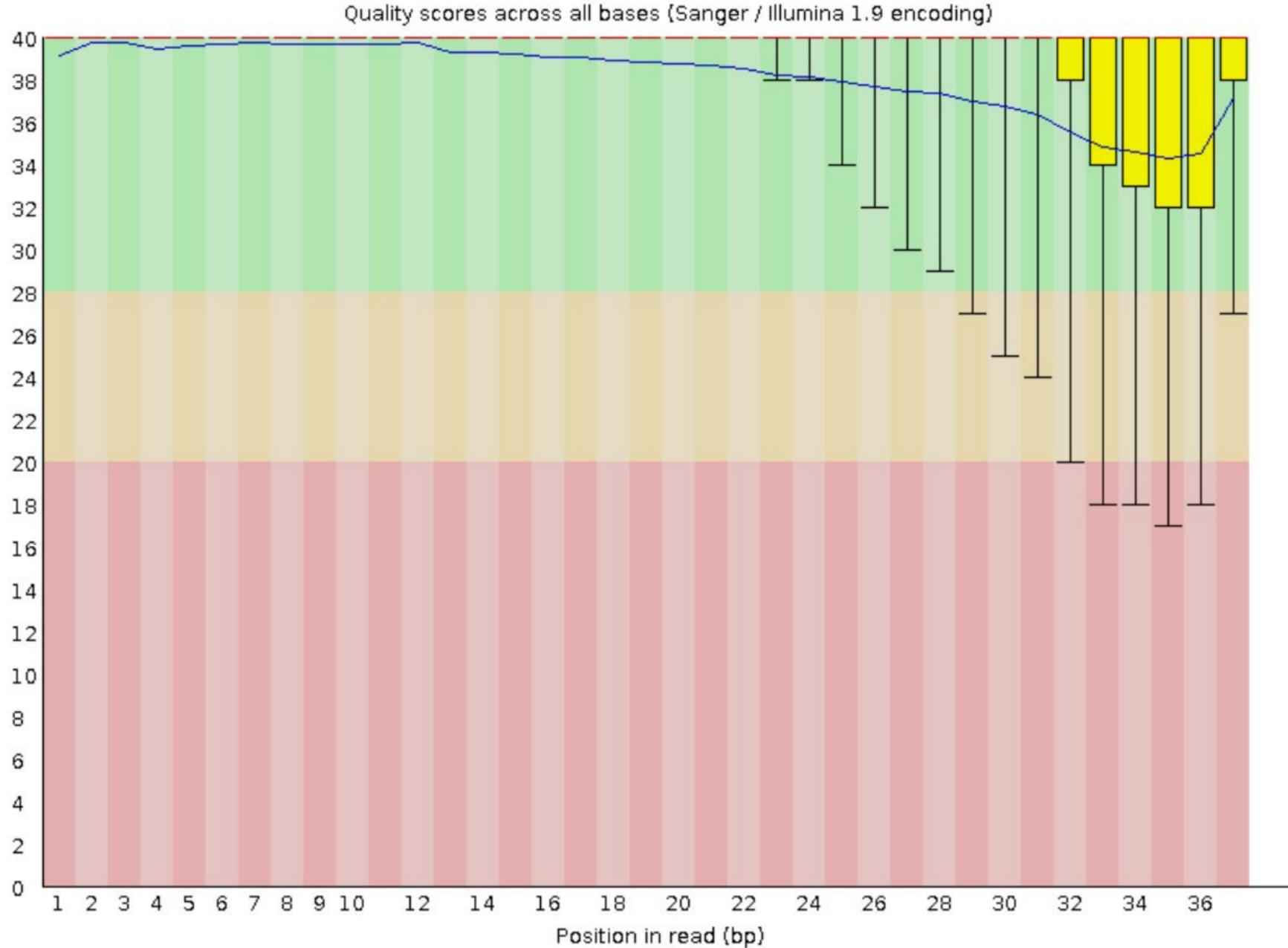
3: FastQC on data 1: RawData

2: FastQC on data 1: Webpage

1: Galaxy1-[GSM461177_1_subsampled.fastq].fastqsanger

Estratégias para a montagem de genomas de novo, tipos de dados. Avaliação de montagens

Controle de qualidade das sequências na Plataforma Galaxy (fastQC)



Estratégias para a montagem de genomas de novo, tipos de dados. Avaliação de montagens

Montagem de genomas com dados de long reads no Galaxy (Flye)

The screenshot displays the Galaxy web interface. At the top, the 'Galaxy' logo is on the left, and navigation links for 'Workflow', 'Visualize', 'Data', 'Help', and 'User' are in the center. A 'Using 0%' indicator is on the right. The left sidebar contains icons for 'Upload', 'Tools', 'Workflows', 'Workflow Invocations', 'Visualization', 'Histories', 'Notifications', and 'Settings'. The main area is divided into three sections: 'Tools', 'WORKFLOWS', and 'History'. The 'Tools' section shows 'Flye' selected, with a 'Show Sections' button. The 'WORKFLOWS' section lists 'All workflows'. The 'History' section shows a list of workflow steps, including 'l fragment as ssembly', '13: Flye on dat a 2: assembly graph', '12: Flye on dat a 2: consensu s', and '11: NanoPlot o n data 2: Log Transformed Histogram Re ad Length'. The central panel displays the output of the Flye workflow, showing a contig labeled '>contig_1' followed by a long sequence of DNA bases.

```
>contig_1
AGCTTATCCATGCGGTTATGCACTCTTTGAATAGAATGGTTTTCTGAAGATCCTGGCTTT
TCGTACTTTATTCTCAGATCACTGATGACCTATCTTGAAGGAATATCTATAATCTCCGAT
CAGTGCGTAAAGCCCGCATTGTATGGAACAGTATGTAGAAAAATTGATTCTTTTCTATTC
TATTATTAATGTAGATTAGCATTAAATTGATGTATGTGGGTGGTAGTCGACTTAGTGATC
CTTTCTTCAAGTGAAGTGTGCAAGCAGTCTACATTTTGTCTCTGGCAGACCGAGGAGA
AGAATAGAAGAAATATACCAGGAGAAACGAAGTCGCTTATAACAAAATATGCAACATGGA
TTCTGGCAATGTGGTTGGCCTCTCTTGTGCGGTGTCAGAATCCATCCTTTCTAAATCTTT
GCCTGCTAGGCAGAAGGATAACAAGATTTAAATTGTCTCGGCAGGACATGATTTCTATT
ACCTGAAATTATAAATGAATAGTTAATGGGTAGATTTATTTTGTAGTGCCGAATCTTGTA
TGTGTTCTAAAAGAATTTGTCCGACACCGGGTCTCAAGGGCGTGAACATAGAATCTT
GAGTGCAAAGAGATGTAACCTCAGTTCACCTCGGTTTTCAGAATCGTCAATCCTAGCTTCCG
TGGGCAGTTGACAATTGAATCCGATTTTGGCATTATTTTCATATCATGATGCGAAAAGAAT
ACAGTTGTTCAAAAATAATGAAGTGGCGTTGAGTTCTCGACCGCTGACTTAAGATTAATG
GTCGGTATTTCTCGATGAGGCAGGAAGGATACTGCTCAGCGGTGAAGTGTACCTTGACG
TCATCAGTTTTCAGGCGATTATCCTAAGTCCAATAGAGTTTCTGGTGGATTTGCCCGCTG
CGTTGTGAATGAATGCGAGCTCGTGGGATTGACGTGAGGGGCGGATATATTTGTGGAGCG
AACTCCGGGCGAATATGCAAGCGCATGGATAGAAGTTATGCCTTGAATAGAATTCGGAA
TTCGCTTTGTACGAACAAGGAAGCTATAAAGTAATGCAACTGCAGAATCTCGCGGAGAG
TTCGATGCTGGTCAAGGATTGACTGCTGGCGGTGCTTCACTTCATGCAAGATCGGACGGA
AGCGGTGGTGTTCAGTGGCGGGCGGATGTGTCGCTGTAAGAACCTGCCCTTGGGAGGG
AACAGCAGCTGGAACGGCTGCCAATGCCGTGCAGGCTGAGGAGCAAAGGAGGAATCCGC
CCGAGGAGGGGTCGCGTCTGAATTAGCTAGGTTGGCGAGGCAATAGCCTTTACAAGGCGA
CGATCGGTAGCTGGTCCGAAGGATGACCACCTGGGACCGACACTGGCCAGACTCTACGAG
AGGCAGCAGTGGGAATTTTCCGCAATAGGCAGTTCTTCTGACGGAGCAATGCCGCGACGTGA
GGTGAAGGCCACGGGTGGCGAACTTCTTCCGGAGAAGCAAACGGCGGCTATCACGGGA
GCAGGCATCGGCTAACTCTGCAGCCGGCAGCCGCGTGTGCATAGAGGATGCAGCGTTATC
CCGGGACGATGGGCGTGTAAAGCTGTGCAGGCTTTTGTCTGCCGTCAAATCCCAGGGCCA
ACCCTGCACAGCGGTGAAACCACCAGCTGGAGTACGACGTGGCAGAGGGAATTCGGGTG
ACGGTGAAATGCGTAGATCGGAAAGAACGCCAACGGCGAAGCACTCTGCTGGGCCGACAC
TGACGCTGAGACGAAATGGAGCAGTGGGATTAGATACCCGAGTCCCAGGCAACGAT
GGATACTAGGCGCTGTGCGTATCGACCCGTGCAGTGTGCAGCTAGCGCGTTAAGTATCC
```

Estratégias para a montagem de genomas de novo, tipos de dados. Avaliação de montagens

Montagem de genomas com dados de long reads no Galaxy (Flye)

The screenshot shows the Galaxy web interface. On the left is a navigation sidebar with icons for Upload, Tools, Workflows, Workflow Invocations, Visualization, Histories, Notifications, and Settings. The main area is divided into three sections: Tools, a table of assembly contigs, and History.

Tools Section: The 'Flye' tool is selected. Below the tool name is a 'Show Sections' button. The description reads: 'Flye de novo assembler for single molecule sequencing reads'.

Table of Assembly Contigs:

| Column 1 | Column 2 | Column 3 | Column 4 | Column 5 | Column 6 | Column 7 | Column 8 |
|-----------|----------|----------|----------|----------|----------|-----------|------------|
| #seq_name | length | cov. | circ. | repeat | mult. | alt_group | graph_path |
| contig_2 | 131830 | 86 | N | N | 1 * | | -3,2,3 |
| contig_1 | 48896 | 22 | N | N | 1 * | | 3,1,-3 |
| contig_3 | 29906 | 140 | N | Y | 1 * | | 3 |

History Section: Shows a list of workflow steps. The current step is '22: Create assemblies with Unicycler on data 5: SPAdes graphs'. Below it are '15: Flye on data 2: assembly info' and '14: Flye on data 2: graphical fragment as'.

Estratégias para a montagem de genomas de novo, tipos de dados. Avaliação de montagens

Montagem de genomas com dados de short reads no Galaxy (Spades)

The screenshot displays the Galaxy web interface. On the left is a navigation sidebar with icons for Upload, Tools, Workflows, Workflow Invocations, Visualization, Histories, Notifications, and Settings. The main area is divided into three sections:

- Tools:** A search bar contains 'Spades'. Below it, a 'Show Sections' button is visible. A list of tools includes:
 - SPAdes genome assembler** for genomes of regular and single-cell projects
 - rnaSPAdes** de novo transcriptome assembler
 - rnaviralSPAdes** de novo assembler for transcriptomes, metatranscriptomes and metaviromes
 - metaplasmidSPAdes** extract and assembly plasmids from metagenomic data
 - coronaSPAdes** SARS-CoV-2 de novo genome
- Dataset Preview:** An orange warning box states: 'This is a binary (or unknown to Galaxy) dataset of size 127.2 KB. Preview is not implemented for this filetype. Displaying as ASCII text'. Below this, a 'Download' button is present. The dataset content is shown as a table of sequence lines:

| | | |
|---|---|---|
| S | 1 | CACTATCGGTTTCACTTTTCTTTTCACTATCGGTTTCACTTTTCTTTTCACTATTGGTTTTACTTTTCTTTGTACCTTTTTTCATGTCTGATTTTC |
| S | 2 | GTGGTCTACGGGCAGCTGCTCAATCAATTAGTTATGAAATACCCCTTAGCTTTATGTGTATTATCAATTTCTCTGCGTGTGATTTCGATCAGGTATA |
| S | 3 | CTAAAAAGGAATATTTATTATTTGATCCATATCCGGACATAAGAAGTCCAACGGGAGCAATACTTGAAGCGGCGATCCAGAAAAAACCCCAATA |
| S | 4 | CACTATCGGTTTCACTTTTCTTTT LN:i:24 dp:f:5.042028802462802 |
| L | 3 | + 4 + 0M |
| L | 2 | + 3 + 0M |
| L | 4 | + 1 + 0M |
| L | 4 | + 4 + 0M |
| L | 3 | - 2 + 0M |
- History:** A panel on the right titled 'History: Class test' shows a workflow for 'Create assemblies with Unicycler on data 5: SPAdes graphs'. It lists 14 gfa datasets. Below the list are buttons for 'Download', 'Show Details', and 'Run Job Again'. The steps in the history are:
 - 10: 001_spades_gra ph_k121
 - 11: 002_depth_filter
 - 12: 003_overlaps_re moved
 - 13: 004_bridges_app lied
 - 14: 005_final_clean

Estratégias para a montagem de genomas de novo, tipos de dados. Avaliação de montagens

Montagem de genomas com dados híbridos (short and long reads) no Galaxy

(Unicycler)

The screenshot displays the Galaxy web interface. On the left is a navigation sidebar with icons for Upload, Tools, Workflows, Workflow Invocations, Visualization, Histories, Notifications, and Settings. The main area is divided into three sections: Tools, Workflows, and a large text output window. The Tools section shows 'Unicycler' selected in a search box, with a 'Show Sections' button below it. The Workflows section is titled 'Create assemblies with Unicycler pipeline for bacterial genomes' and lists 'All workflows'. The central text window displays the output of the Unicycler pipeline, starting with a header: `>1 length=130068 depth=1.00x`. The output consists of a long sequence of DNA bases. On the right side, there is a 'History' panel showing a search bar and a list of workflow steps. The top step is '27: Create assemblies with Unicycler on data 2 and data 5: Final Assembly', followed by '26: Create assemblies with Unicycler on data 2 and data 5: Final Assembly Graph', and '25: Create assemblies with...'. The top right corner of the interface shows 'Using 0%'.

Estratégias para a montagem de genomas de novo, tipos de dados. Avaliação de montagens

Qualidade e completude da montagem no Galaxy (QUAST)

The screenshot displays the Galaxy web interface. The top navigation bar includes the Galaxy logo, a home icon, and menu items for Workflow, Visualize, Data, Help, and User. A 'Using 0%' indicator is visible in the top right. The left sidebar contains navigation options: Upload, Tools, Workflows, Workflow Invocations, Visualization, Histories, Notifications, and Settings. The 'Tools' section is active, showing a search for 'quast' and a 'Show Sections' button. Below this, two tool descriptions are visible: 'Quast Genome assembly Quality' and 'rnaQUAST A quality assessment tool for De Novo transcriptome assemblies'. The 'WORKFLOWS' section shows 'All workflows'. The main content area displays the 'QUAST' tool results for a genome assembly. The title is 'QUAST Quality Assessment Tool for Genome Assemblies by CAB'. The execution time is '14 October 2024, Monday, 17:56:04'. A link to 'View in Icarus contig browser' is provided. A note states: 'All statistics are based on contigs of size ≥ 500 bp, unless otherwise noted (e.g., "# contigs (> 500 bp)" and "Total length (≥ 0 bp)" include all contigs)'. A table of statistics is shown, with a header 'Statistics without reference' and a sub-header 'Create_assemblies_with_Unicyc...'. The table lists various metrics and their values. The 'Mismatches' section shows 0 N's per 100 kbp and 0 N's. The right sidebar shows the 'History' section with a search bar and a list of jobs. The top job is '57: Quast on data 27: HTM L report'. Below it is '27: Create ass embly with Unicycler on d ata 2 and dat a 5: Final Ass embly'. The bottom job is '26: Create ass embly with Unicycler on d ata 2 and dat a 5: Final Ass embly'. The 'Class test' section shows '71.3 MB', '22', and '6'.

QUAST
Quality Assessment Tool for Genome Assemblies by CAB

14 October 2024, Monday, 17:56:04

[View in Icarus contig browser](#)

All statistics are based on contigs of size ≥ 500 bp, unless otherwise noted (e.g., "# contigs (> 500 bp)" and "Total length (≥ 0 bp)" include all contigs).

| Statistics without reference | | Create_assemblies_with_Unicyc... |
|--------------------------------|---------|----------------------------------|
| # contigs | 1 | |
| # contigs (≥ 0 bp) | 1 | |
| # contigs (≥ 1000 bp) | 1 | |
| Largest contig | 130 068 | |
| Total length | 130 068 | |
| Total length (≥ 0 bp) | 130 068 | |
| Total length (≥ 1000 bp) | 130 068 | |
| N50 | 130 068 | |
| N90 | 130 068 | |
| auN | 130 068 | |
| L50 | 1 | |
| L90 | 1 | |
| GC (%) | 36.85 | |
| Mismatches | | |
| # N's per 100 kbp | 0 | |
| # N's | 0 | |

Estratégias para a montagem de genomas de novo, tipos de dados. Avaliação de montagens

- **Montagem de genomas com short Reads (Ex: Illumina)**

1. FastQC.
2. Trim Galore
3. SPAdes:
4. Velvet:
5. QUAST:
6. BUSCO:

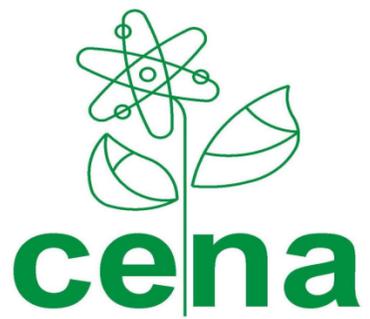
- **Montagem de Genomas com Long Reads (Ex: PacBio, Oxford Nanopore)**

1. NanoPlot:
2. Filtlong:
3. Canu ou Flye:
4. Pilon:
5. QUAST e BUSCO

- **Montagem híbrida (Short + Long Reads)**

1. FastQC (para short reads) e NanoPlot (para long reads)
2. Trim Galore e Filtlong
3. SPAdes (Modo Híbrido):
4. Unicycler:
5. Pilon
6. QUAST e BUSCO.

| TIPO DE LEITURA | Pré-processamento | Montagem Principal | Correção de Erros | Avaliação |
|------------------|--|--------------------------------------|-------------------|--------------|
| Short Reads | FastQC, Trim Galore! | SPAdes, Velvet | - | QUAST, BUSCO |
| Long Reads | NanoPlot, Filtlong | Canu, Flye | Pilon | QUAST, BUSCO |
| Montagem Híbrida | FastQC, NanoPlot, Trim Galore!, Filtlong | SPAdes (Híbrido), Unicycler, MaSuRCA | Pilon | QUAST, BUSCO |



Obrigado!

Danilo Ferreira da Silva

Doutorando em Solos e Nutrição
de Plantas ESALQ/USP

Gabriely Santos de Oliveira

Mestranda em Biologia na Agricultura e no Ambiente

CENA/USP

