UNIVERSIDADE DE SÃO PAULO / Centro de Energia Nuclear na Agricultura

CEN5789 - Genômica e Bioinformática

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

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

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

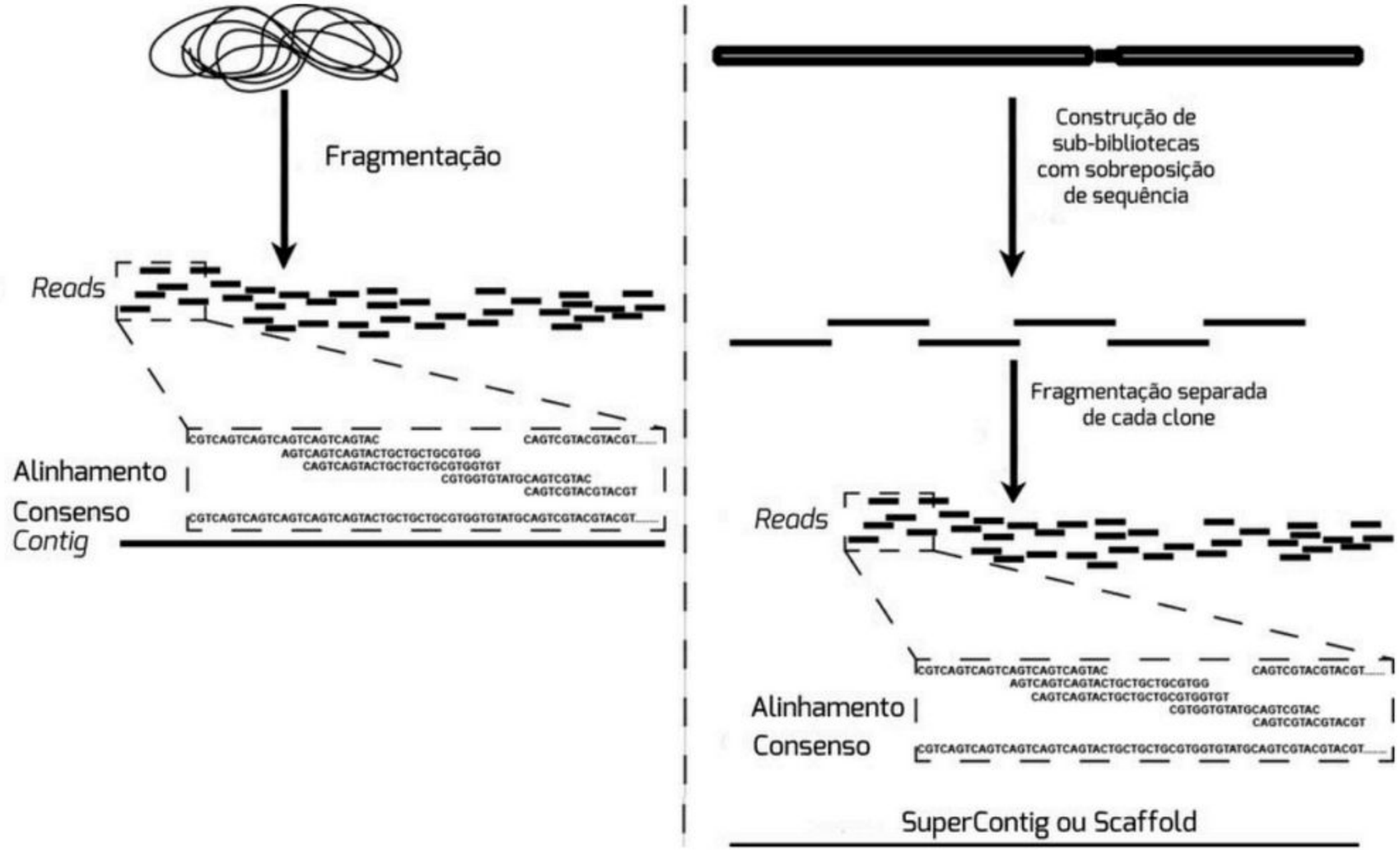
PACIFIC BIOSCIENCES

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



(Hugo., 2014)

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

PROGRAMAS DE MONTAGEM DE SHORT READS

- ABySS
- ALLPATHS
- Edena
- SGA
- SHARCGS
- SHORTY

PLATAFORMA ONLINE PARA MONTAGEM DE GENOMAS

- GALAXY

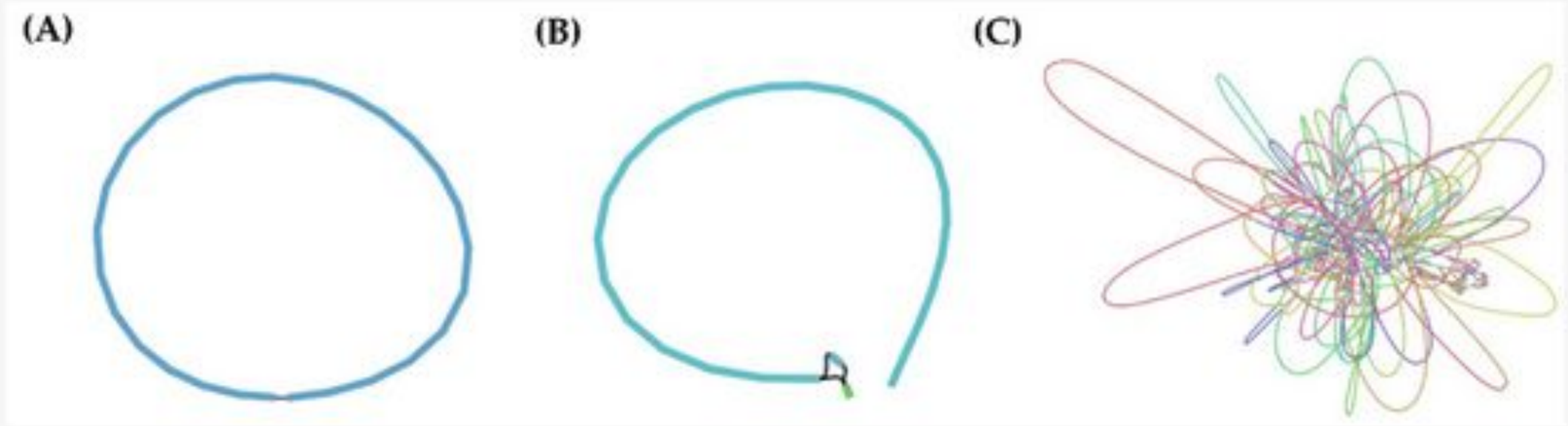
Estratégias para a montagem de genomas de novo, tipos de dados. Avaliação de montagens

ESTRATEGIAS PARA MONTAGEM COM LONG READS

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

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.



Zhang et al.,

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

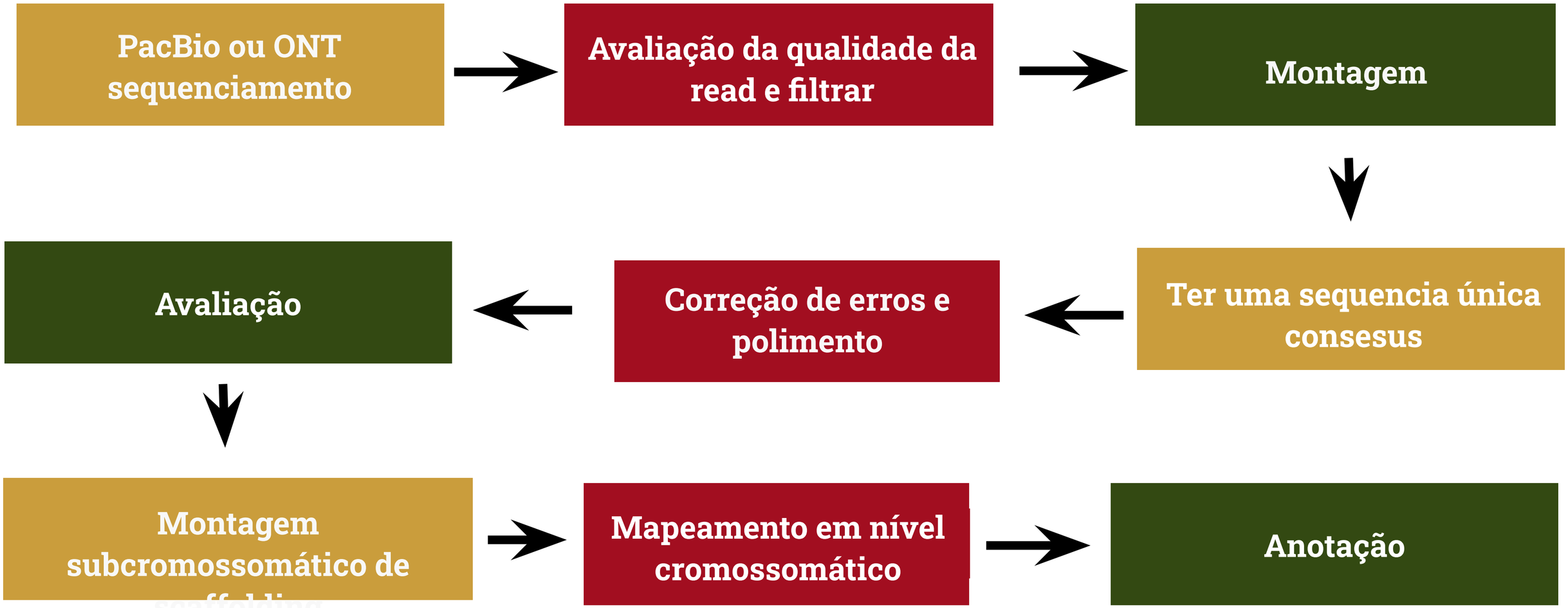
Nanopolish

Racon

Pilon

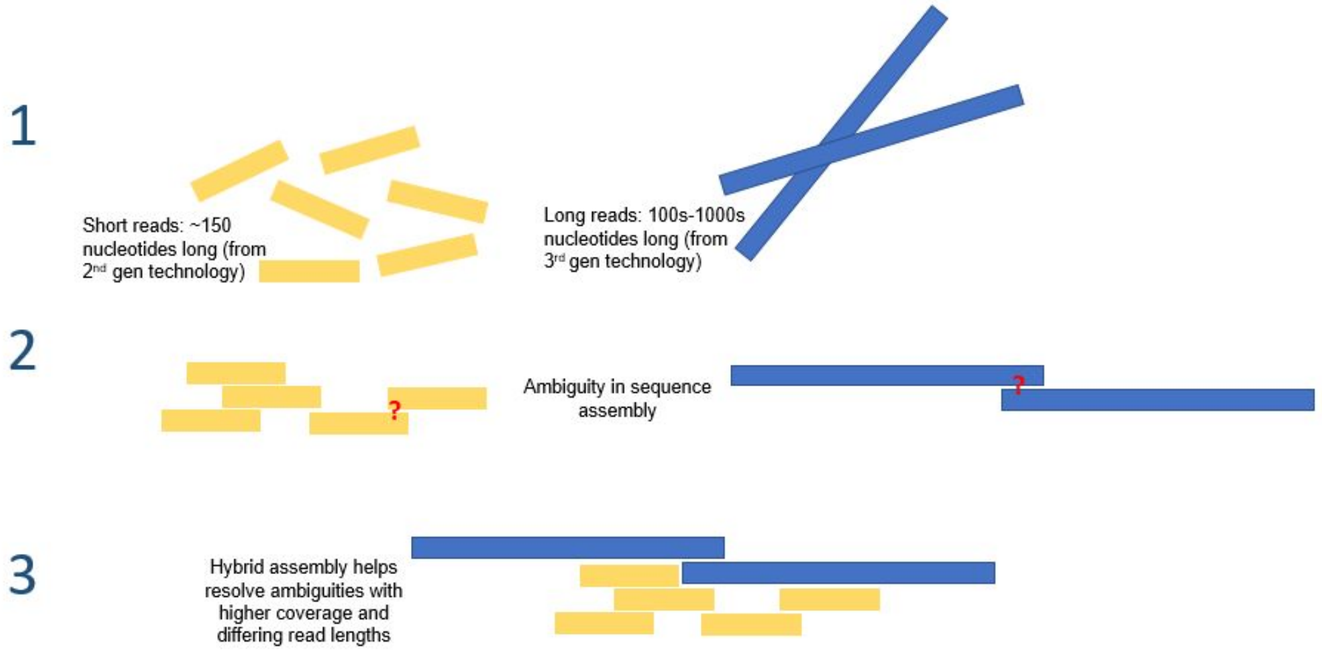
Estratégias para a montagem de genomas de novo, tipos de dados. Avaliação de montagens

RECOMENDAÇÃO DE PIPELINE



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

ESTRATEGIAS PARA MONTAGEM HIBRIDAS



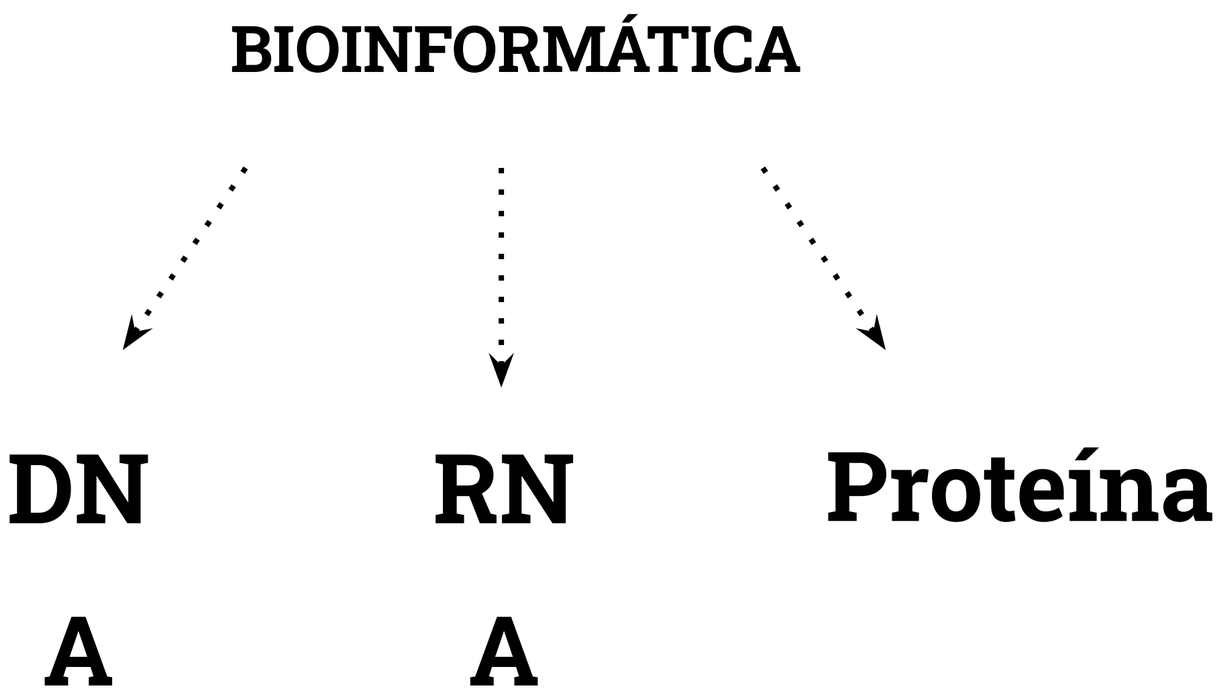
Wang et al., 2012

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

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)
ACCCTCTTTTCTTATCATTGACATTTAAACTCTGGGGCAGGTCCTCGCGTAGAACGCGGCTGTCAGATCT
GCCACTTCCCCTGCCGAGCGGCGGTGAGAAGTGTGGGAACCGGCGCTGCCAGGCTCACCTGCCTCCCCGC
CCTCCGCTCCCAGGTAACCGCCCGGGCTCCGGCCCCGGCCCGGCTCGGGGCCCGCGGGGCCTCTCCGCTG
CCAGCGACTGCTGTCCCCAAATCAAAGCCCGCCCCAAGTGGCCCCGGGGCTTGATTTTTGCTTTTAAAAG
GAGGCATACAAAGATGGAAGCGAGTTACTGAGGGAGGGATAGGAAGGGGGGTGGAGGAGGGACTTGTCTT
TGCCGAGTGTGCTCTTCTGCAAAAGTAGCAAAATGTTCCACTCCTAAGAGTGGACTTCCAGTCCGGCCCT
GAGCTGGGAGTAGGGGGCGGGAGTCTGCTGCTGCTGTCTGCTAAAGCCACTCGCGACCGCGAAAAATGCA
GGAGGTGGGGACGCACTTTGCATCCAGACCTCCTCTGCATCGCAGTTCACGACATCCACGCTTGGGAAAG
TCCGTACCCGCGCCTGGAGCGCTTAAAGACACCCTGCCGCGGGTCGGGCGAGGTGCAGCAGAAGTTTCCC
GCGGTTGCAAAGTGCAGATGGCTGGACCGCAACAAAGTCTAGAGATGGGGTTCGTTTCTCAGAAAGACGC
```

TIPOS DE ARQUIVOS GERADOS

FASTQ

- Contém sequências + pontuações de qualidade;
- Formato de dados perpetuado e amplamente aceito;
- Possui variantes:

1. **Sanger standard: fastq-sanger;**
2. **Solexa/Illumina: fastq-solexa;**
3. **Illumina 1.3+: fastq-illumina**

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

TIPOS DE ARQUIVOS GERADOS

FASTQ

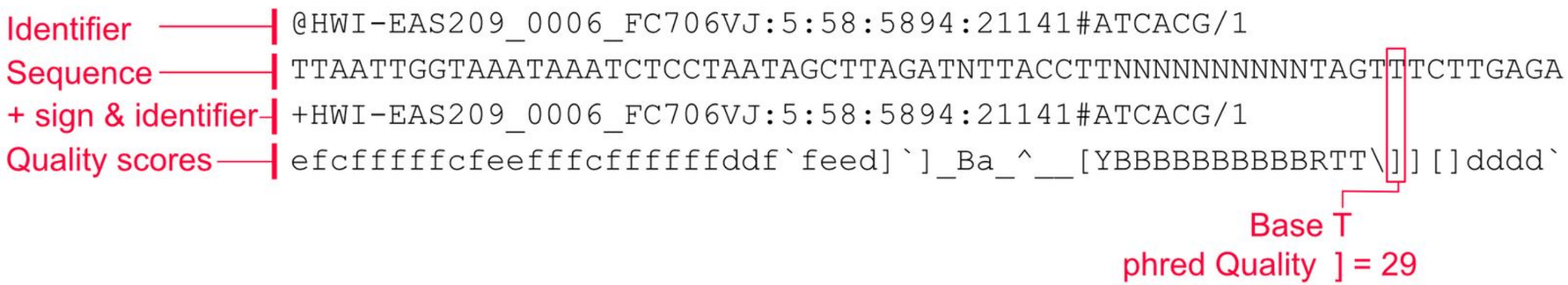
```
@SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGGCTTTTTTTGTTTGGAACCGAAAGG
GTTTTGAATTTCAAACCCTTTTCGGTTTCCAACCTTCCAA
AGCAATGCCAATA
+SRR014849.1 EIXKN4201CFU84 length=93
3+&$#""""""""""""""7F@71,'";C?,B;?6B;:EA1EA
1EA5'9B:?:#9EA0D@2EA5':>5?:%A;A8A;?9B;D@
/=<?7=9<2A8==
```

@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

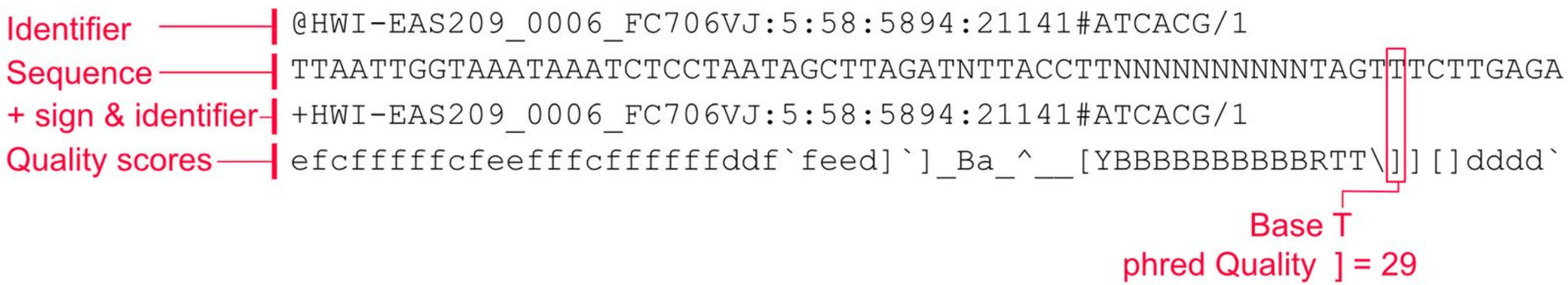
Fastq



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

TIPOS DE ARQUIVOS GERADOS

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

```
#chrom  start   end     name    score   strand
chr1    1000    5000    geneA   960     +
chr1    5500    9000    geneB   850     -
chr1    15000   16000   geneC   780     +
chr2    100     800     regionX 1000    +
chr2    1200    1500    regionY 900     +
chr2    3000    5000    regionZ 1000    +
chr3    10000   11000   geneD   970     -
chr3    11500   12000   geneE   750     +
chr3    2000    4000    regionA 950     +
chr4    200     400     geneF   900     -
chr4    500     600     geneG   600     +
chr5    2500    3000    geneH   800     +
chr5    7000    8000    geneI   850     +
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

VCF

PRINCIPAIS CARACTERISTICAS:

- **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">
##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">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#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

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

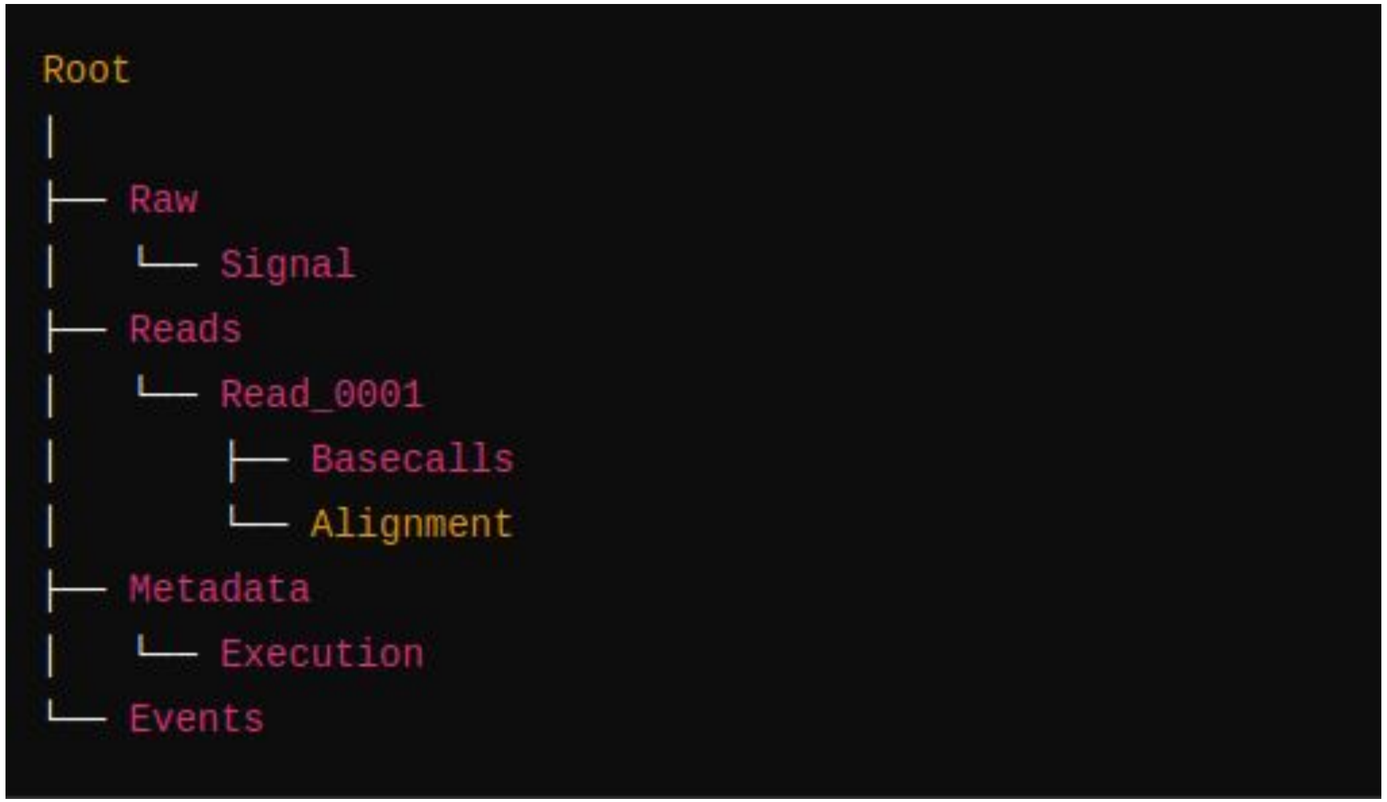
ESTRUTURA DO ARQUIVO VCF:

- **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.

FAST5

TIPOS DE ARQUIVOS GERADOS

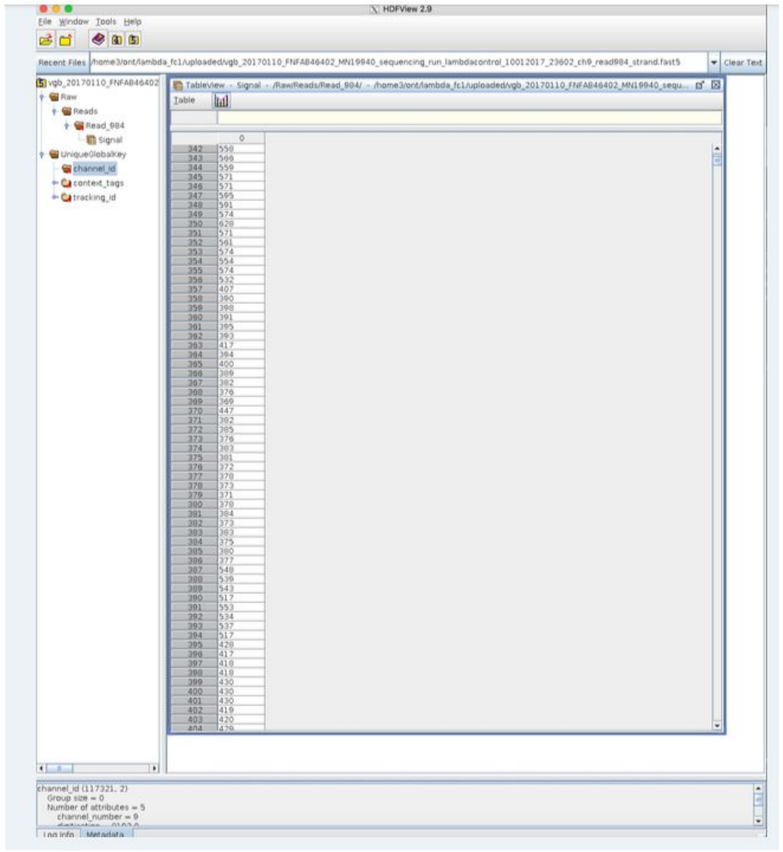
Em um diagrama hierárquico, seria algo assim:



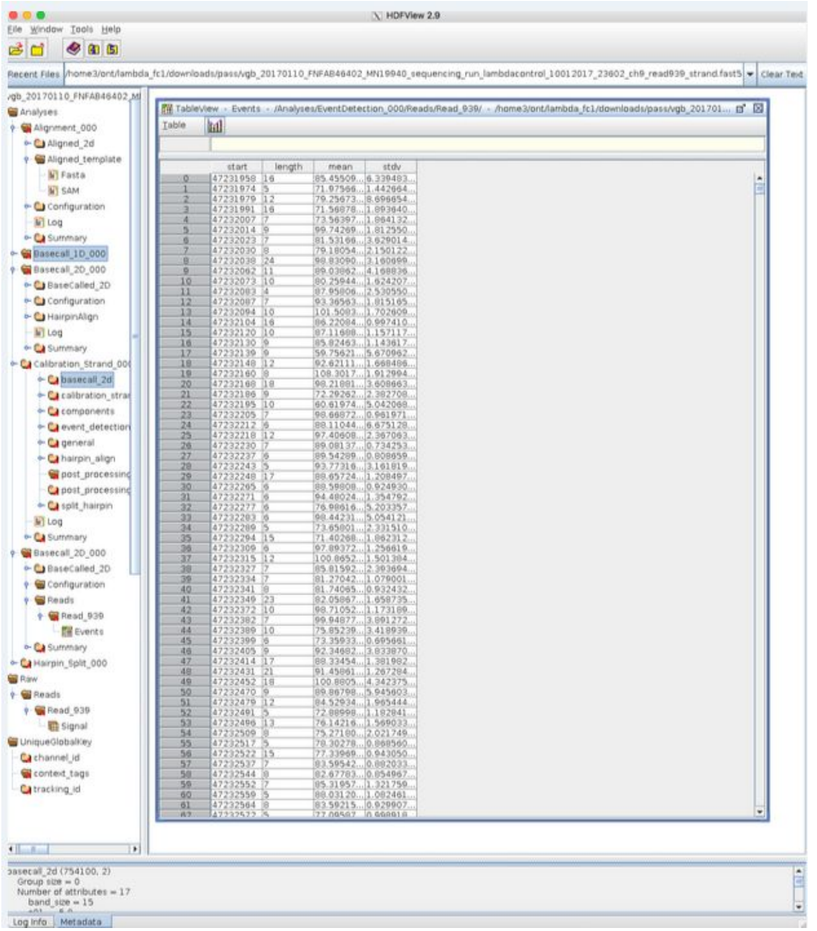
Essa estrutura facilita o armazenamento de diferentes tipos de dados de maneira eficiente, mas também pode ser complexa para processamento, por isso a análise de arquivos FAST5 às vezes pode ser lenta

TIPOS DE ARQUIVOS GERADOS

FAST5



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



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

TIPOS DE ARQUIVOS GERADOS

FAST5

```
hoph01@Alpha:~/Documents/CNT_exp18 h5ls -r /home3/ont/lambda_fc1/uploaded/vgb_20170110_FNFAB46402_MN19940_sequencing_run_lambdacontrol_10012017_23602_ch9_read984_strand.fast5
/                        Group
/Raw                     Group
/Raw/Reads               Group
/Raw/Reads/Read_984      Group
/Raw/Reads/Read_984/Signal Dataset {84995/Inf}
/UniqueGlobalKey         Group
/UniqueGlobalKey/channel_id Group
/UniqueGlobalKey/context_tags Group
/UniqueGlobalKey/tracking_id Group
```

```
HDF5 "/home3/ont/lambda_fc1/download/pass/vgb_20170110_FNFAB46402_MN19940_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/componenta
 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
```

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

FAST5

TIPOS DE ARQUIVOS GERADOS

E obter todos os dados e metadados para um determinado grupo

```
HDF5 "/home3/ont/lambda_fc1/downloads/pass/vgb_20170110_FNFAB46402_MN19940_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, 470, 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

SLOW5

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

SLOW5

PRINCIPAIS CARACTERISTICAS

- **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

SLOW5

```
# 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

BLOW5

PRINCIPAIS CARACTERISTICAS

- **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 CARACTERISTICAS

- **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

SAM

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

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

Alignment section

QNAME (query template name, aka. read ID)

FLAG (indicates alignment information about the read, e.g. paired, aligned, etc.)

RNAME (reference sequence name, e.g. chromosome /transcript id)

POS (1-based position)

MAPQ (mapping quality)

CIGAR (summary of alignment, e.g. insertion, deletion)

RNEXT (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)

PNEXT (Position of the primary alignment of the NEXT read in the template; corresponding to the POS column)

TLEN (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

SEQ (read sequence)

Optional fields in the format of 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 CARACTERISTICAS

- **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.

BAM

TIPOS DE ARQUIVOS GERADOS

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:

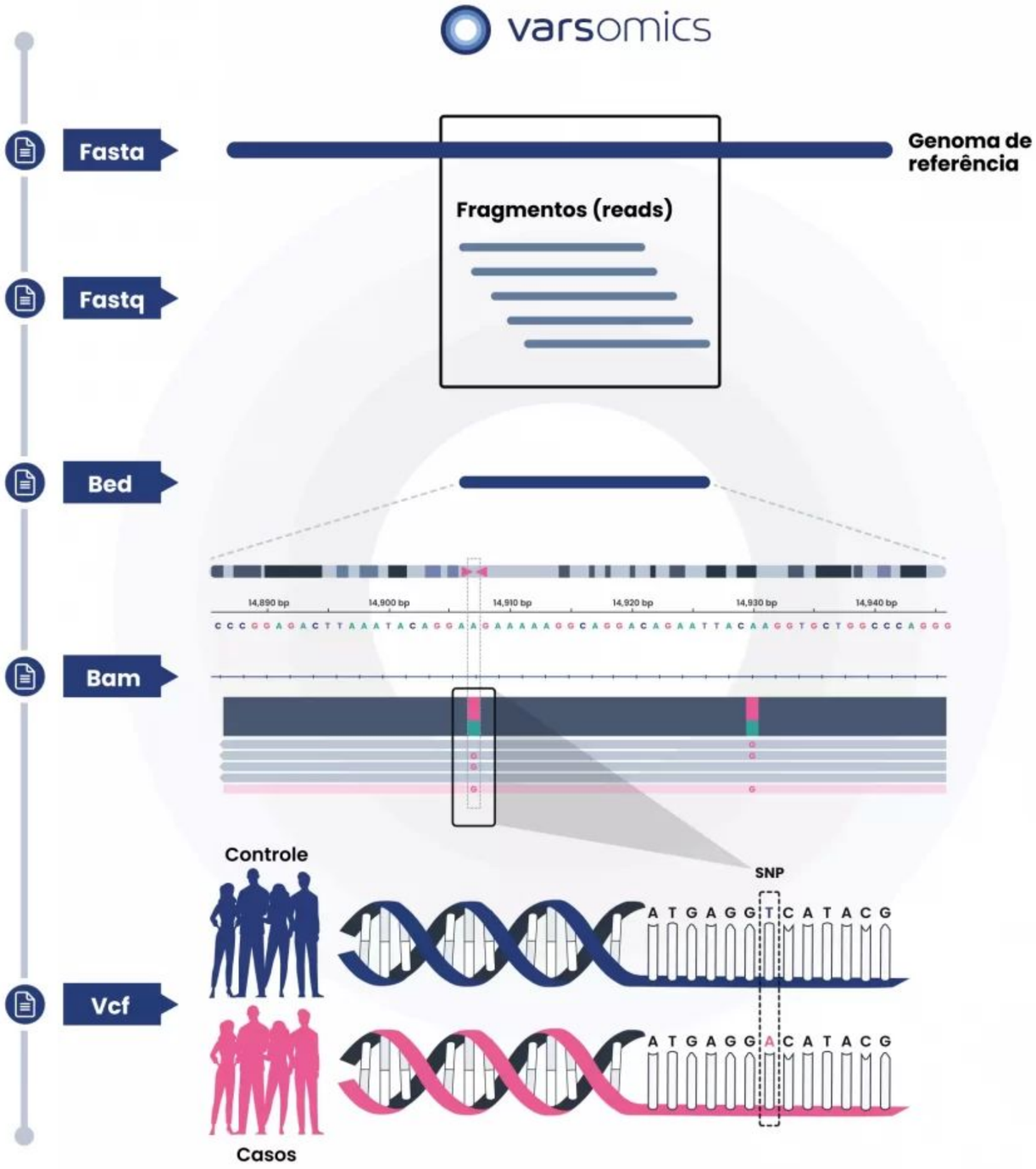
```
samtools faidx ref.fa
samtools view -bt ref.fa.fai -o aln.bam aln.sam
```

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

PRINCIPAIS CARACTERISTICAS

- **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

Fornece 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

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

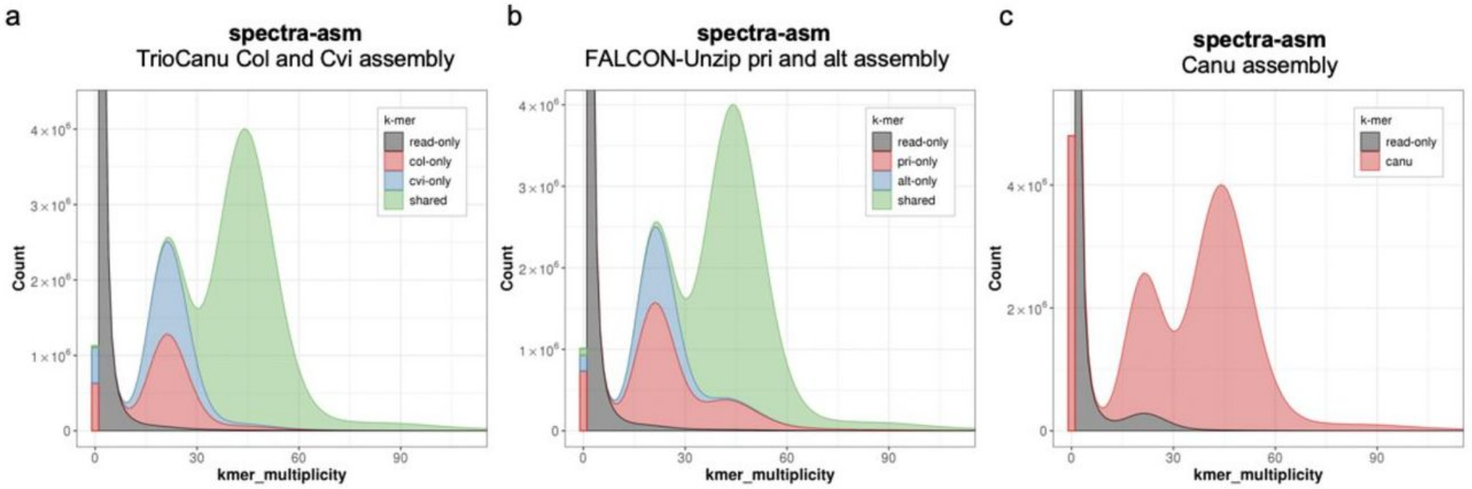
```
All statistics are based on contigs of size >= 500 bp, unless otherwise noted
(e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs)
.

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

Rhie, A., McCarthy, S.A., Fedrigo, O. et al. Towards complete and error-free genome assemblies of all vertebrate species. Nature 592, 737–746 (2021). https://doi.org/10.1038/s41586-021-03451-0

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.

Rhie, A., McCarthy, S.A., Fedrigo, O. et al. Towards complete and error-free genome assemblies of all vertebrate species. Nature 592, 737–746 (2021). https://doi.org/10.1038/s41586-021-03451-0

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:

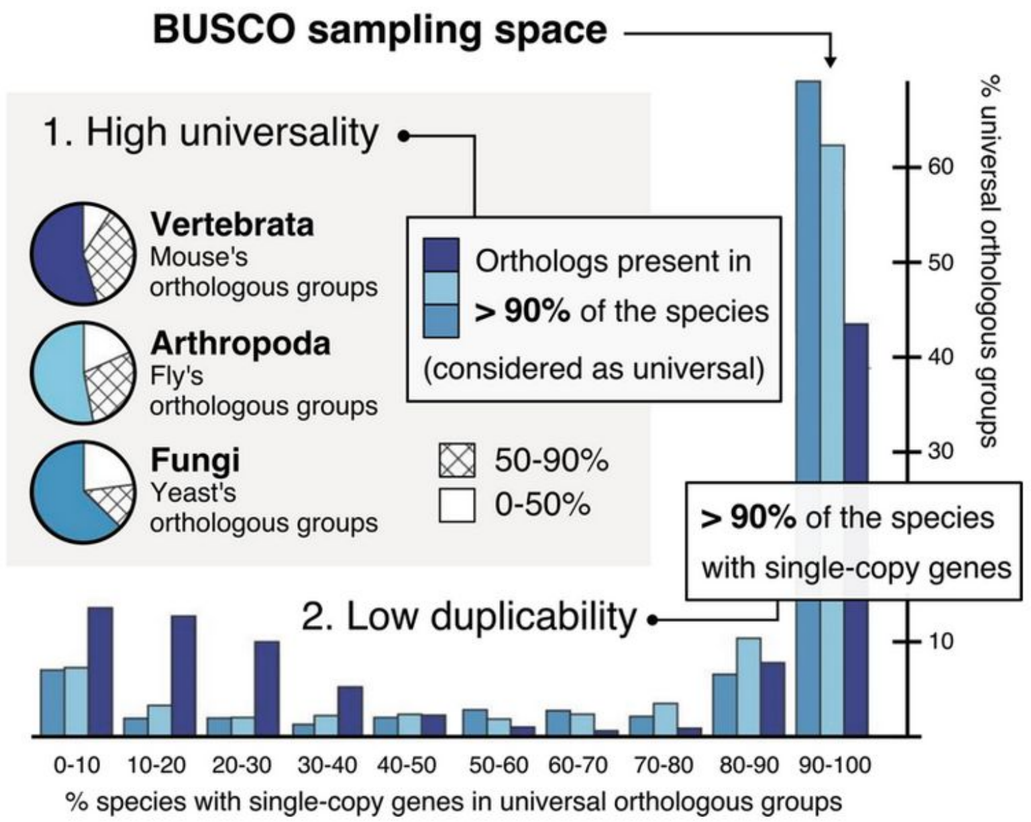
1. IDENTIFICAÇÃO DE GENES;
2. COMPARAÇÃO COM GENOMAS DE REFERÊNCIA;

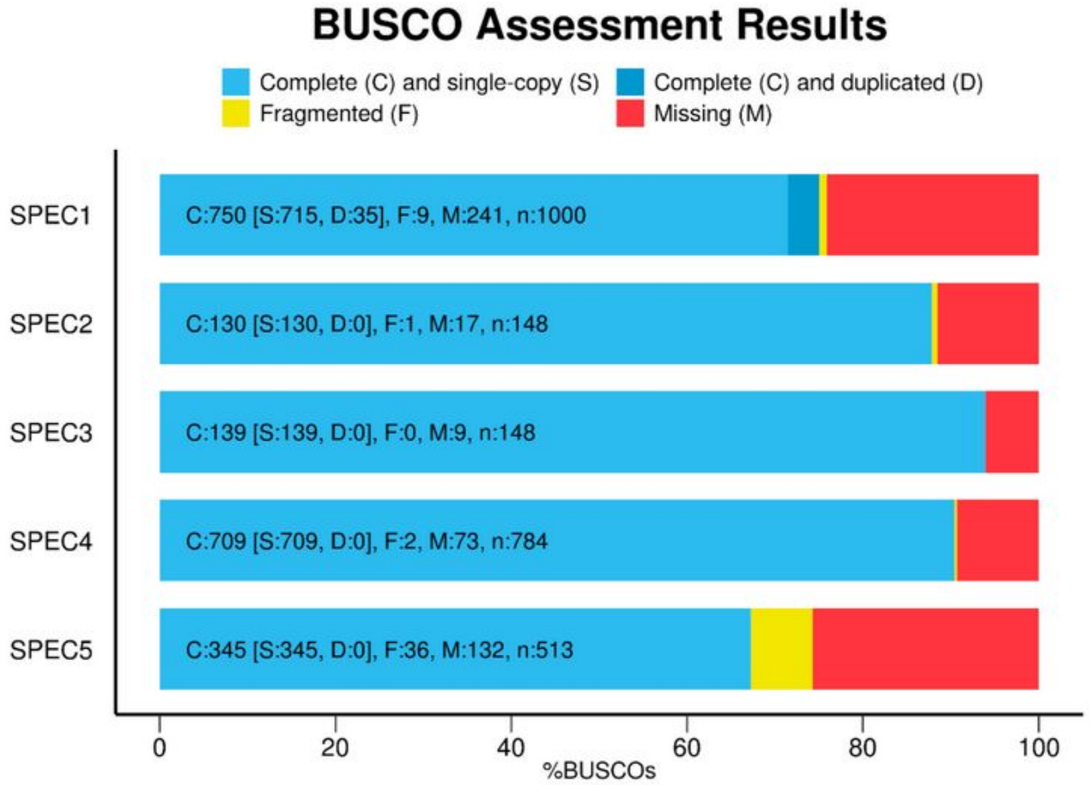
BUSCO

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





AVALIAÇÃO DE MONTAGENS

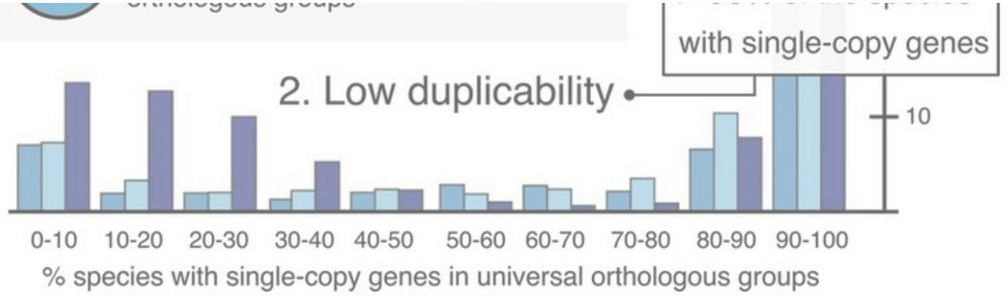
BUSCO

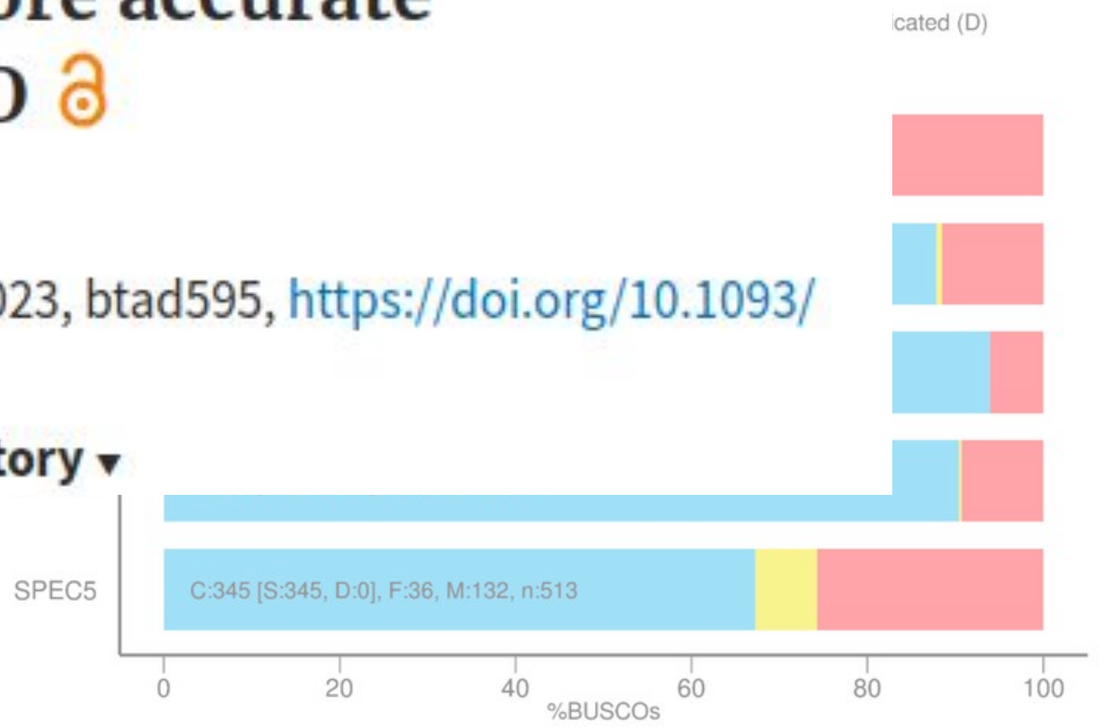
compleasm: a faster and more accurate reimplementation of BUSCO

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

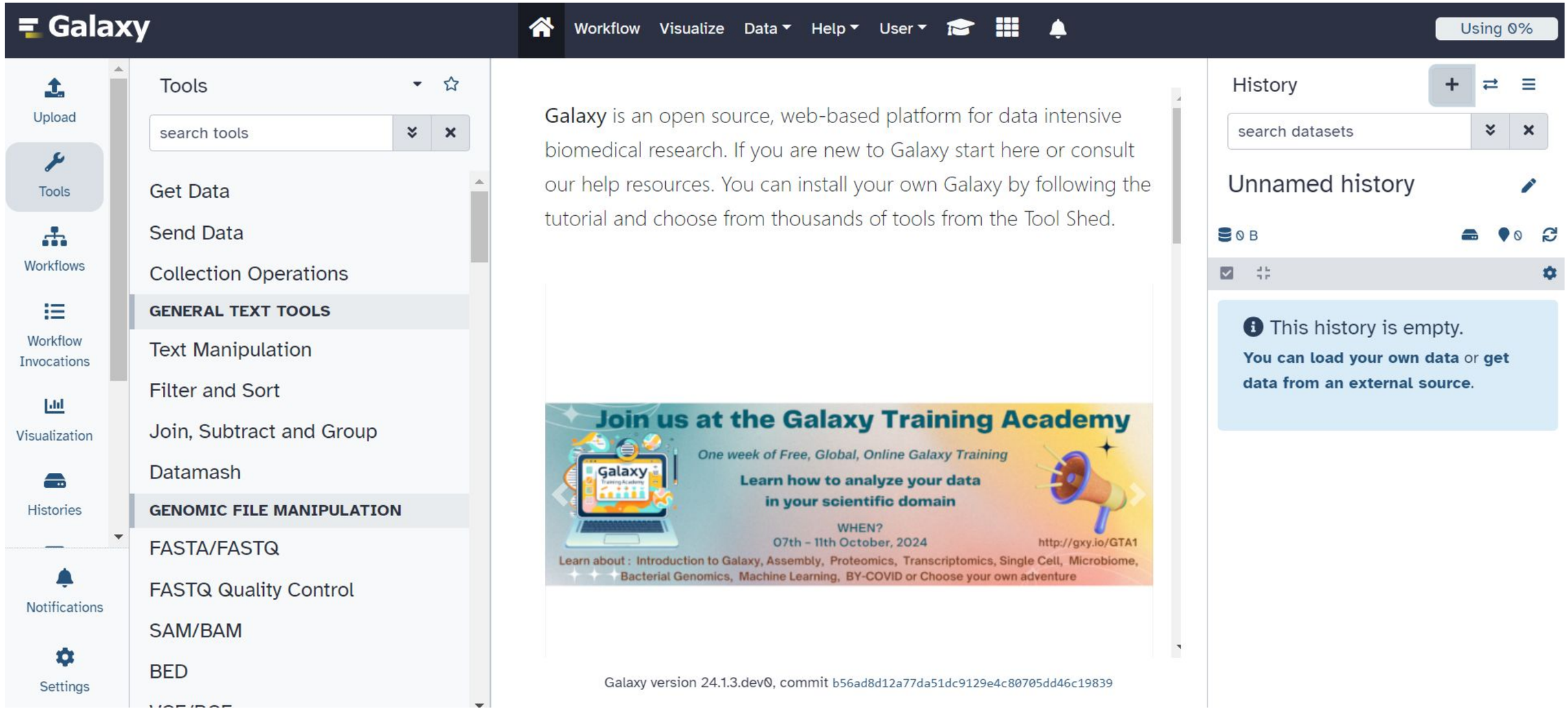
Esses genes BUSCO são categorizados nas seguintes classes:

- **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

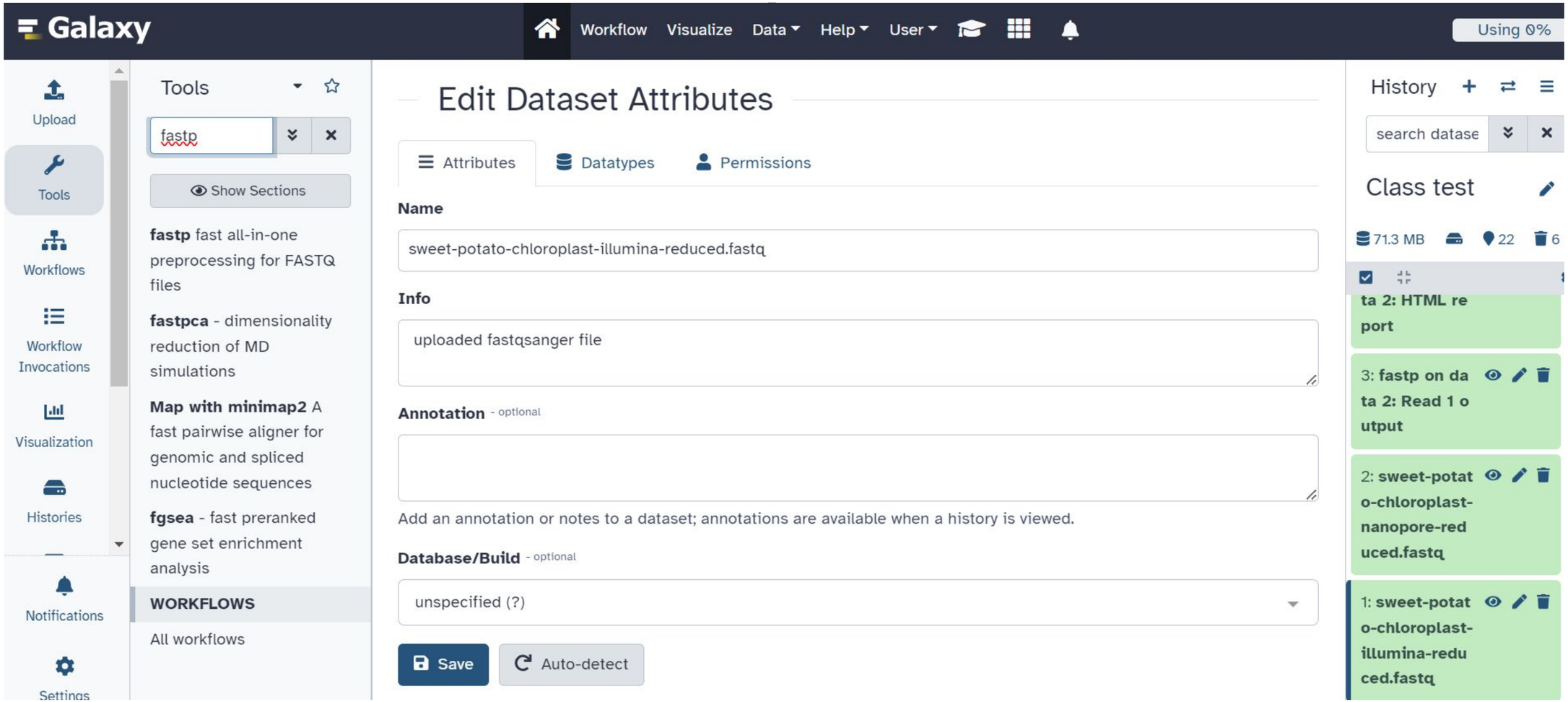
Estratégias para a montagem de genomas de novo, tipos de dados. Avaliação de montagens

Upload dos arquivos na Plataforma Galaxy



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

Controle de qualidade das sequências na Plataforma

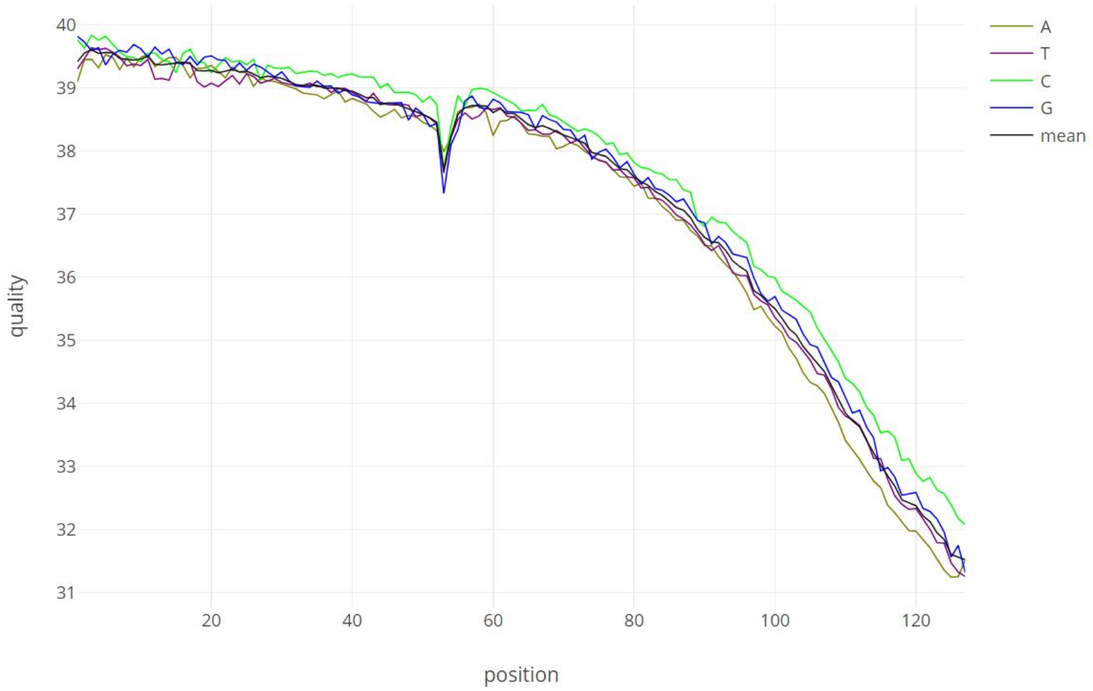


Estratégias para a montagem de genomas de novo, tipos de dados. Avaliação de montagens

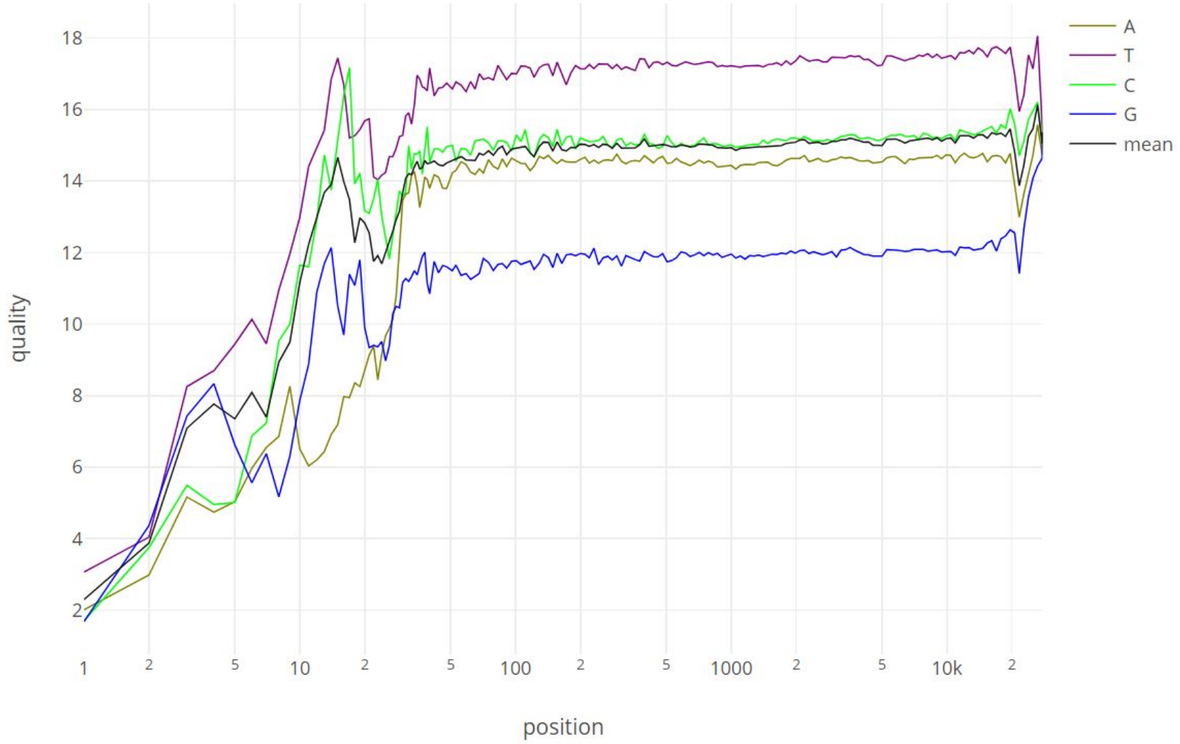
Controle de qualidade das sequências na Plataforma

Galaxy

Short reads



Long reads



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

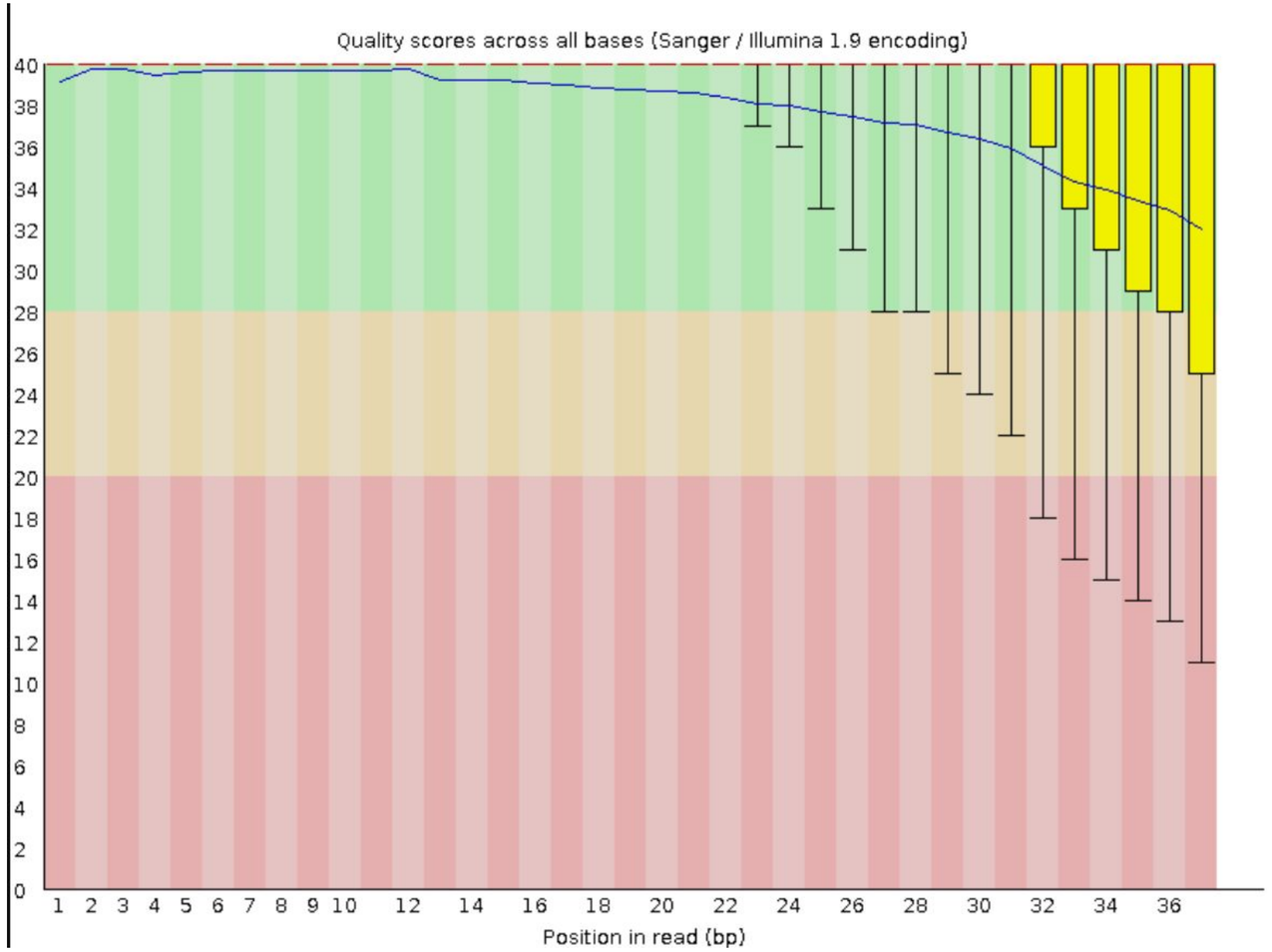
Controle de qualidade das sequências na Plataforma

Galaxy

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

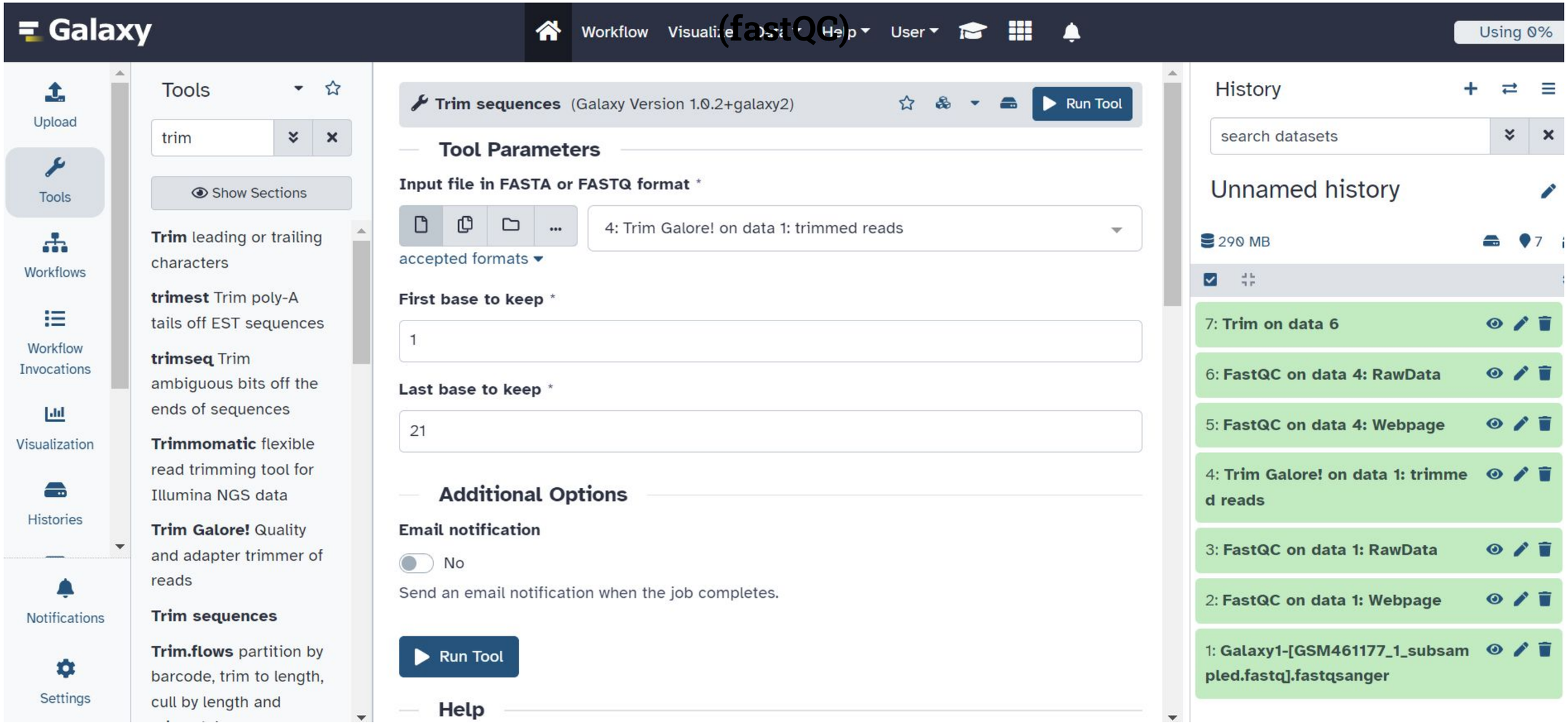
Controle de qualidade das sequências na Plataforma Galaxy

(fastQC)



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

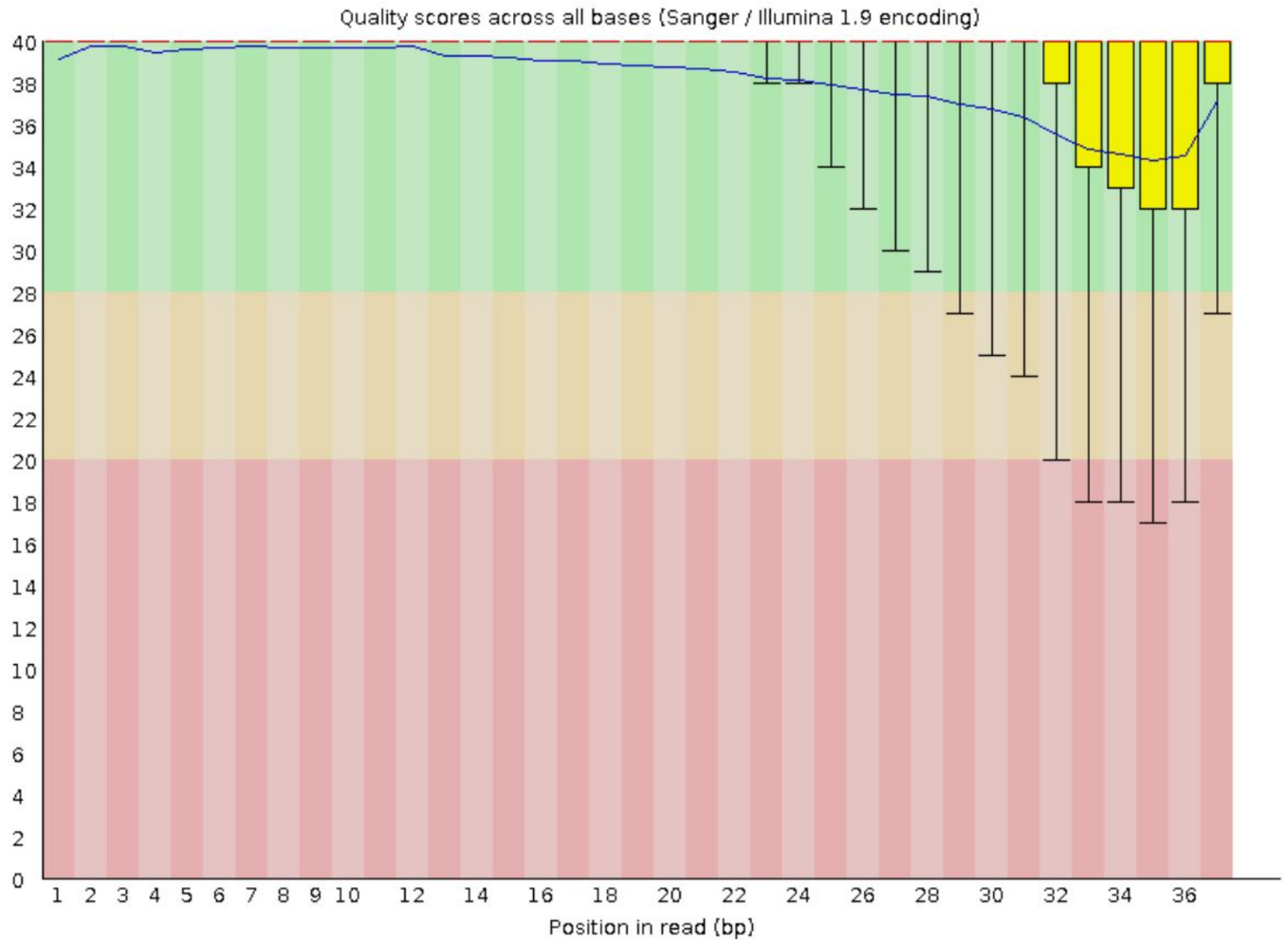
Controle de qualidade das sequências na Plataforma Galaxy (fastQC)



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

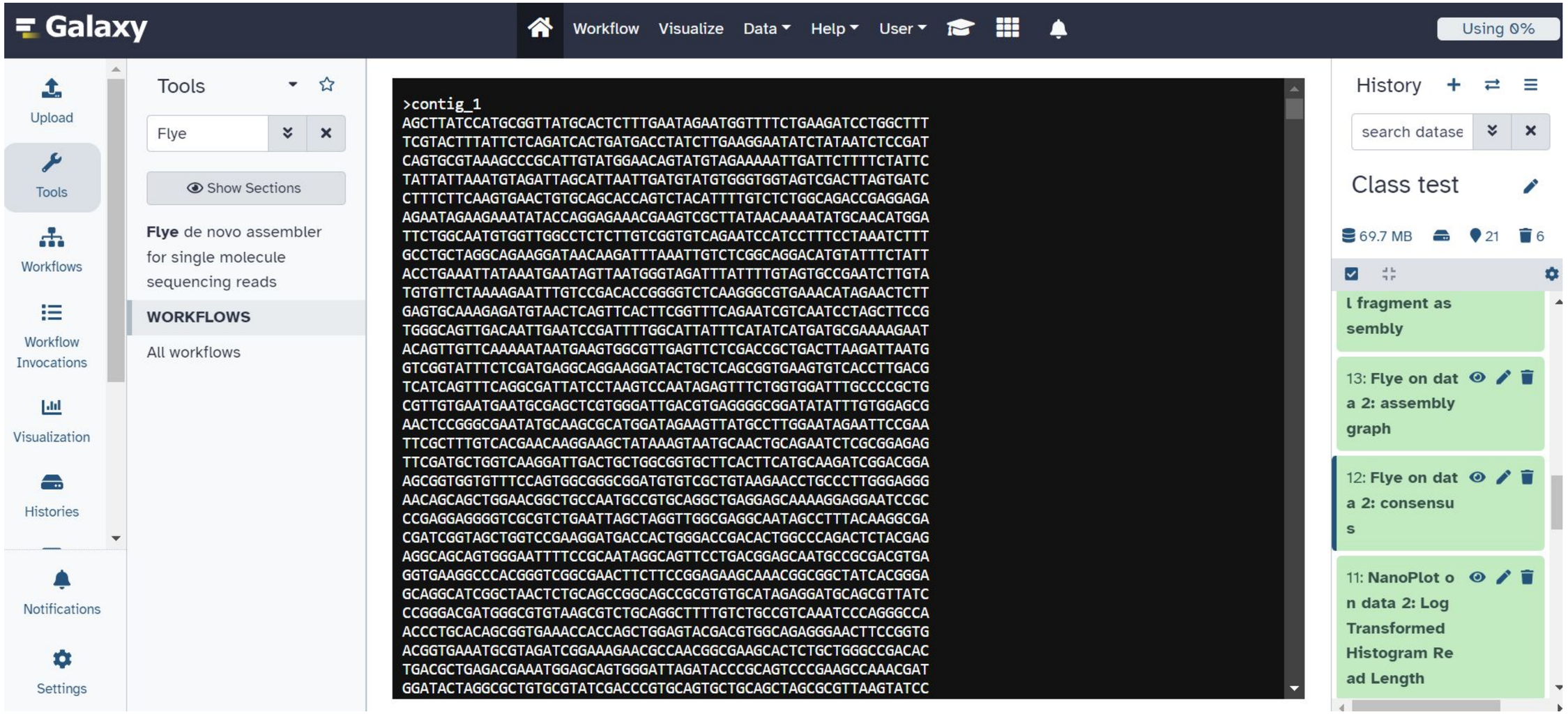
Controle de qualidade das sequências na Plataforma Galaxy

(fastQC)



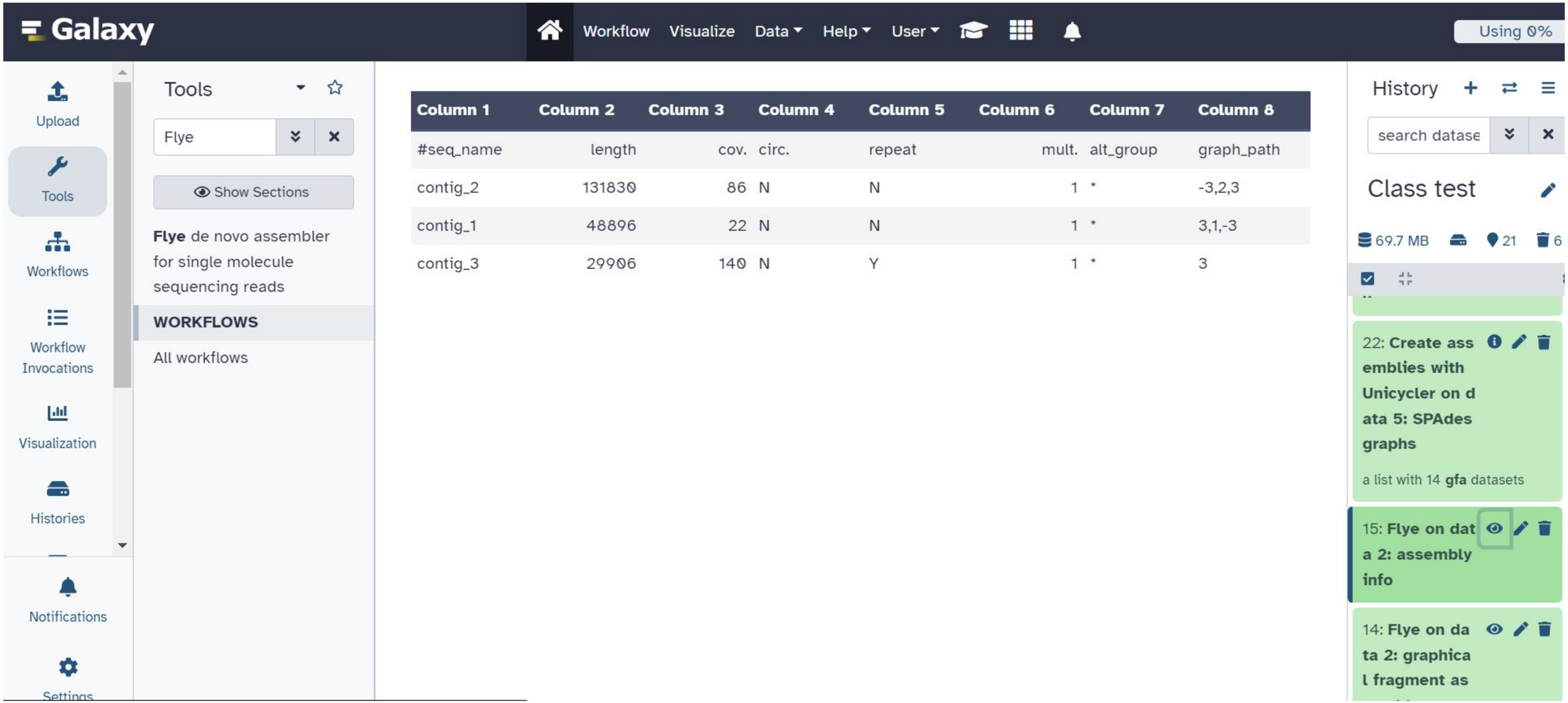
Estratégias para a montagem de genomas de novo, tipos de dados. Avaliação de montagens

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



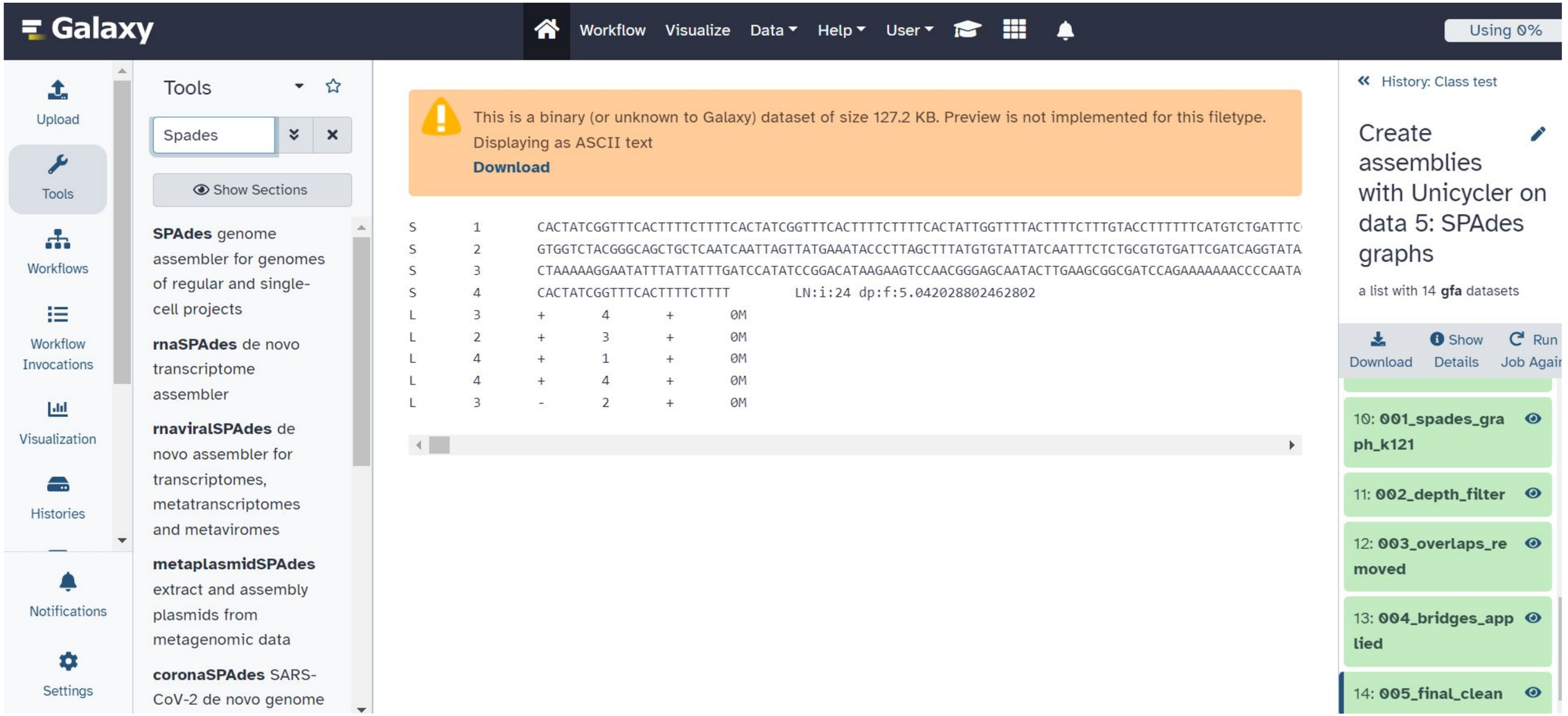
Estratégias para a montagem de genomas de novo, tipos de dados. Avaliação de montagens

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



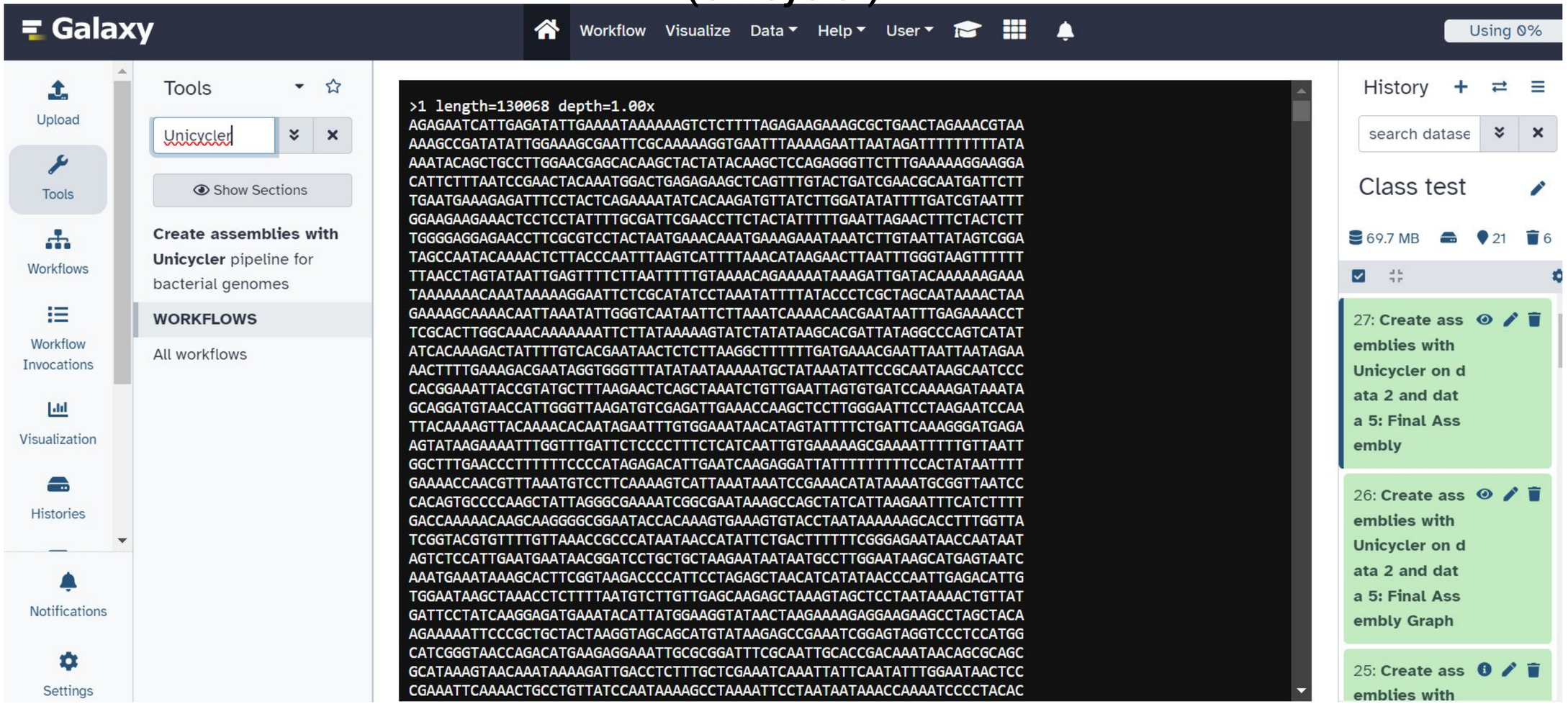
Estratégias para a montagem de genomas de novo, tipos de dados. Avaliação de montagens

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



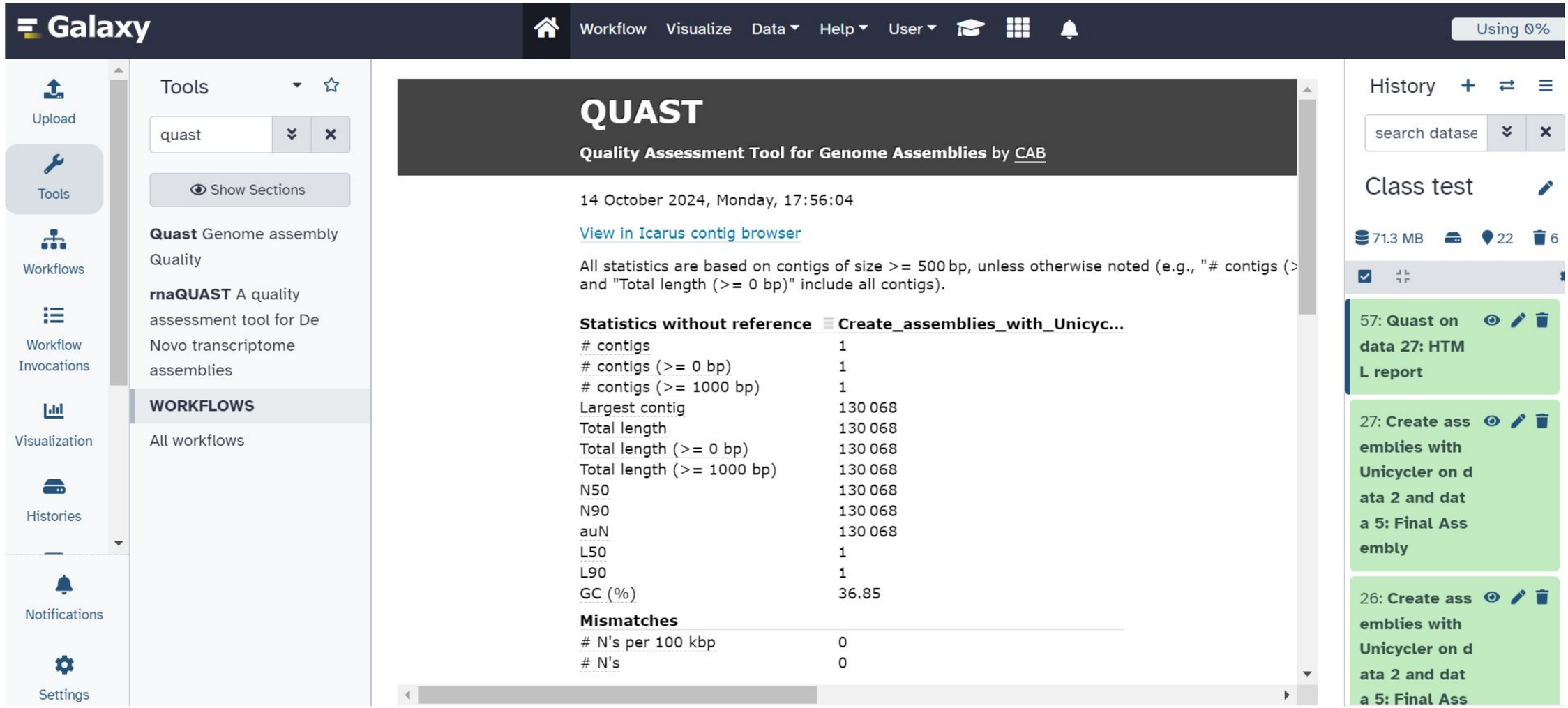
Estratégias para a montagem de genomas de novo, tipos de dados. Avaliação de montagens

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



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

Qualidade e completude da montagem no Galaxy (QUAST)



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 Galoree 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