MONTAGEM DE GENOMAS DE NOVO E AVALIAÇÃO

GUSTAVO NASCIMENTO, MARIA GABRIELA E KEVIN

ESTRATÉGIAS PARA A MONTAGEM DE GENOMAS DE NOVO

1.

FIRST-GENERATION SEQUENCING

Sanger

2.

SECOND-GENERATION SEQUENCING (NGS)

Illumina

Pirosequenciamento

3.

THIRD-GENERATION SEQUENCING

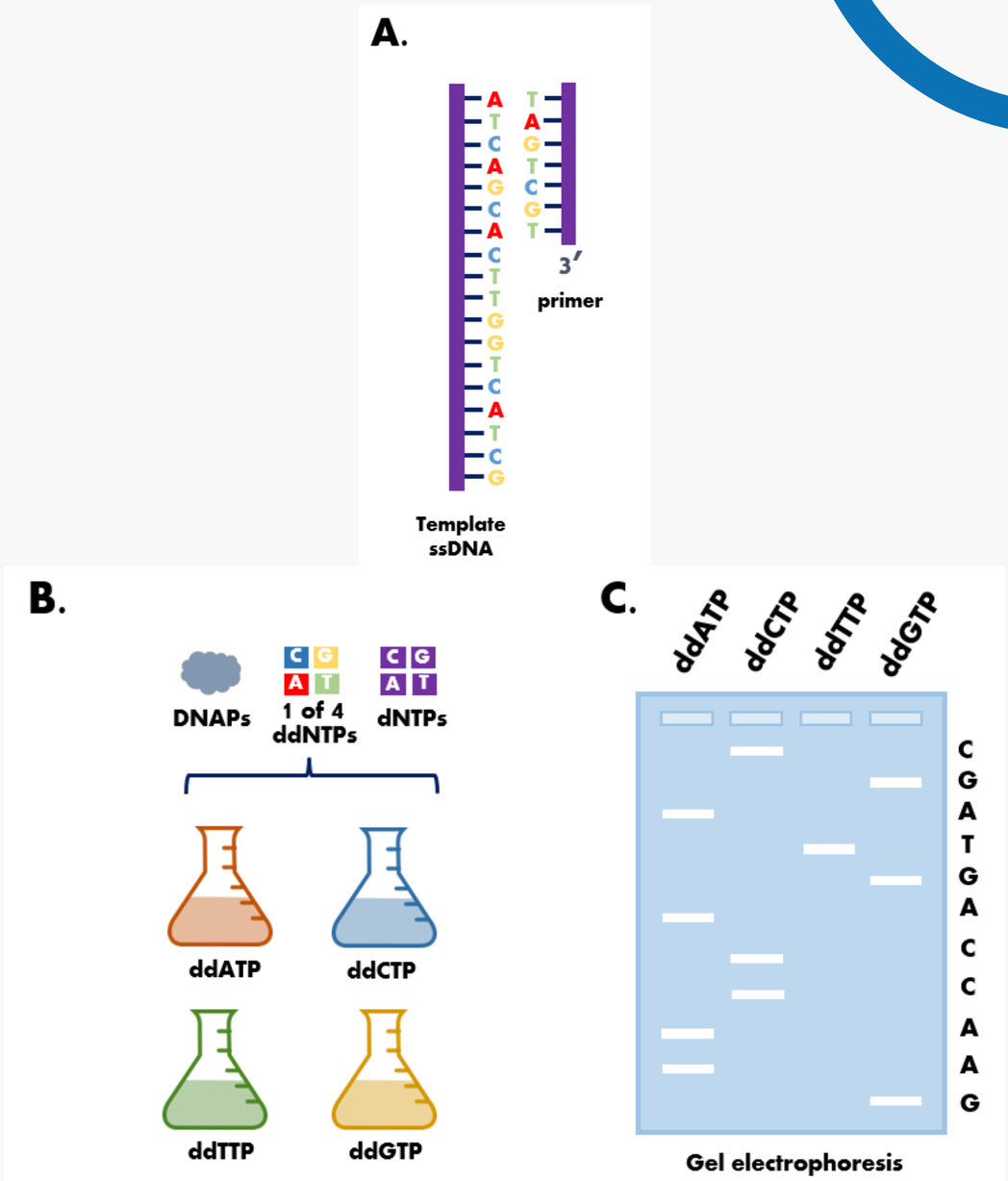
PacBio

ONT

ESTRATÉGIAS PARA A MONTAGEM DE GENOMAS DE NOVO

SANGER

- A primeira geração de sequenciamento
- 1977
- Fragmentos de 500-900bp
- Custo alto por bp



Sanger Sequencing Apollo Institute Disponivel em: https://apollo-institute.org/sanger-sequencing/ Acesso: 16/10/2023

ESTRATÉGIAS PARA A MONTAGEM DE GENOMAS DE NOVO

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

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

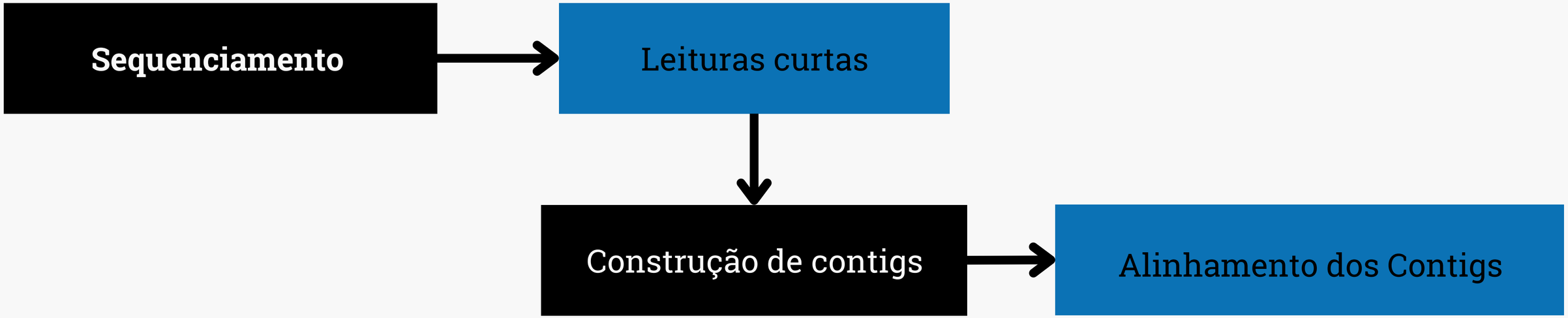
OXFORD NANOPORE

PACIFIC BIOSCIENCES

IBM'S DNA TRANSISTOR

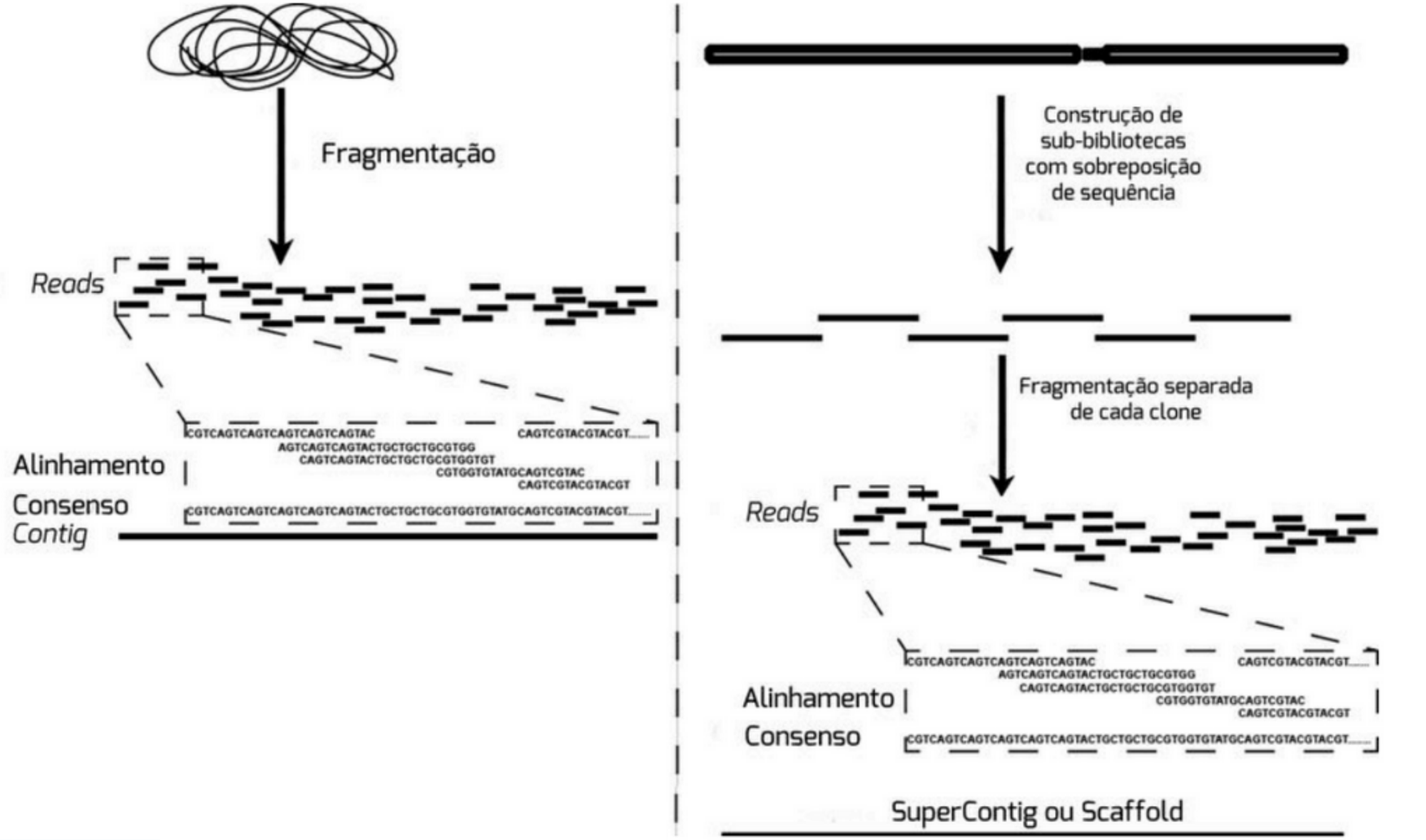
ESTRATÉGIAS PARA A MONTAGEM DE GENOMAS DE NOVO

ESTRATEGIAS PARA MONTAGEM COM SHORT READS



ESTRATÉGIAS PARA A MONTAGEM DE GENOMAS DE NOVO

ESTRATEGIAS PARA MONTAGEM COM SHORT READS



Vergil, Hugo. Bioinformatica da Biologia à flexibilidade Molecular. São Paulo: Sociedade Brasileira de Bioquímica e biologia Molecular - SBBq. 2014

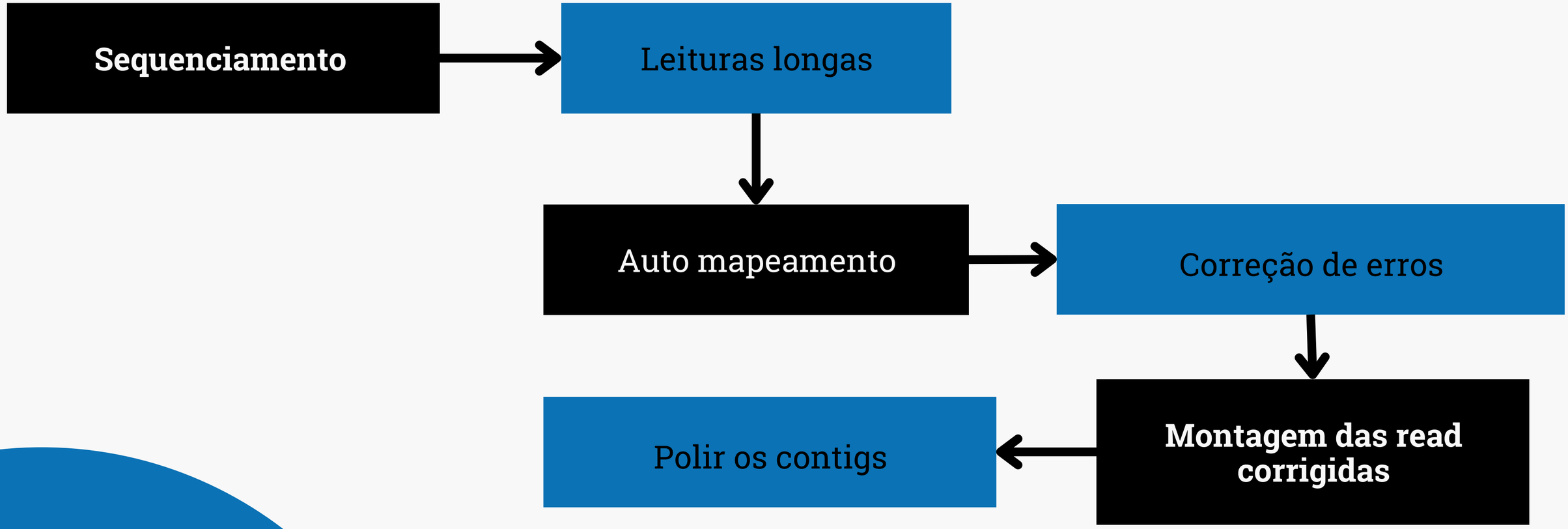
ESTRATÉGIAS PARA A MONTAGEM DE GENOMAS DE NOVO

PROGRAMAS DE MONTAGEM DE SHORT READS

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

ESTRATÉGIAS PARA A MONTAGEM DE GENOMAS DE NOVO

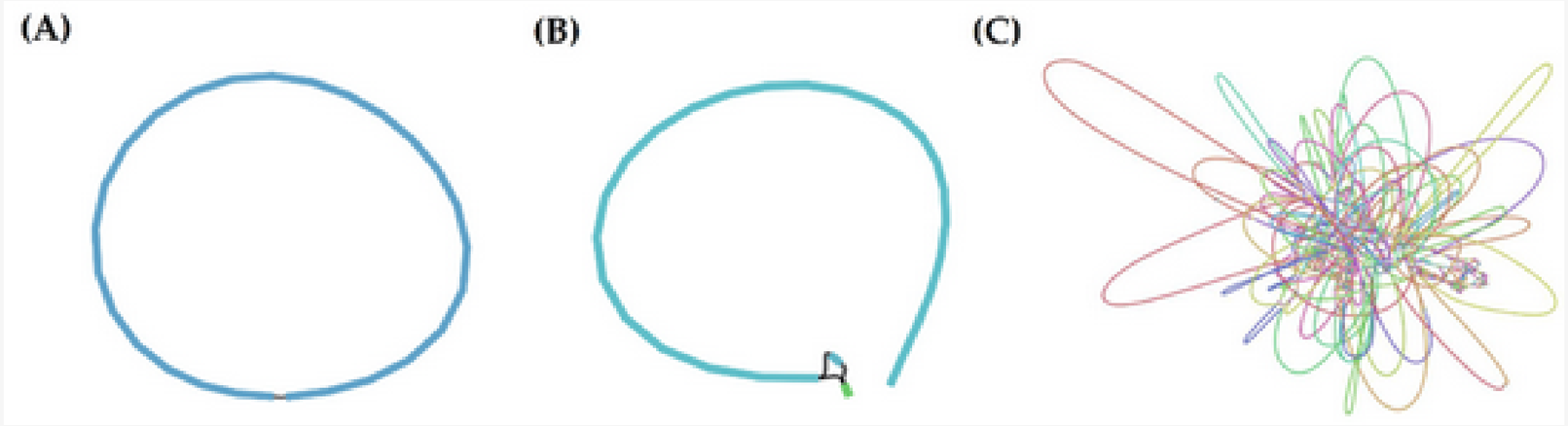
ESTRATEGIAS PARA MONTAGEM COM LONG READS



ESTRATÉGIAS PARA A MONTAGEM DE GENOMAS DE NOVO

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, P.; Jiang, D.; Wang, Y.; Yao, X.; Luo, Y.; Yang, Z. Comparison of De Novo Assembly Strategies for Bacterial Genomes. Int. J. Mol. Sci. 2021, 22, 7668. https://doi.org/10.3390/ijms22147668

ESTRATÉGIAS PARA A MONTAGEM DE GENOMAS DE NOVO

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

PROGRAMAS DE MONTAGEM DE LONG READS

MAPEAMENTO

- BWA-MEM
- minialign/minimap
- BBMap/BBTools

CORREÇÃO DE ERROS

- Frame-Pro
- MINIMAP
- TULIP

ESTRATÉGIAS PARA A MONTAGEM DE GENOMAS DE NOVO

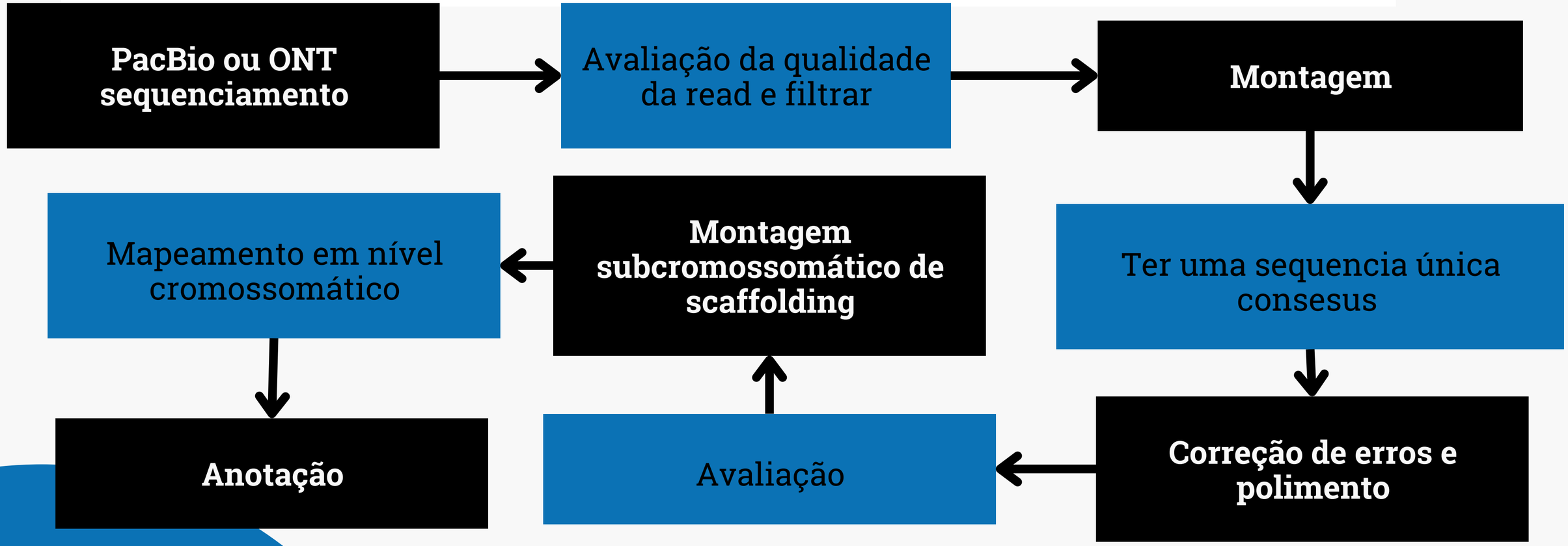
PROGRAMAS DE MONTAGEM DE LONG READS

POLIMENTO DE SEQUENCIAS

- Nanopolish
- Racon
- pilon

ESTRATÉGIAS PARA A MONTAGEM DE GENOMAS DE NOVO

RECOMENDAÇÃO DE PIPELINE



ESTRATÉGIAS PARA A MONTAGEM DE GENOMAS DE NOVO

ESTRATEGIAS PARA MONTAGEM HIBRIDAS

- Compensa pelos problemas de casa estrategia
- Illumina + PacBio
- Útil para genomas de plantas poliploide
- Consegue maior sequencias e contiguidade de scaffolds.

Unicycler

MaSuRCA

BIOINFORMÁTICA

DNA

RNA

Proteína

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.

Arquivo FASTA

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

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**

```
@SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGGCTTTTTTTGTTTGGAACCGAAAGG
GTTTTGAATTTCAAACCCTTTTCGGTTTCCAACCTTCCAA
AGCAATGCCAATA
+SRR014849.1 EIXKN4201CFU84 length=93
3+&$#""""""""""""""7F@71,'";C?,B;?6B;:EA1EA
1EA5'9B:?:#9EAOD@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.

BED

- Representa regiões genômicas de forma simples e flexível;
- Contém informações sobre as anotações do dado genômico, onde inicia e onde termina alguma característica em um segmento de DNA;
- Usado para delimitar regiões de interesse no genoma.

VCF

- Armazena as informações das variantes genéticas identificadas em um organismo;
- Variantes SNPs;
- Versionado;
- Compara ao genoma de referência.

ARQUIVO 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.

FAST5

Oxford Nanopore

- Hierárquico;
- Formato de dados com um esquema específico;
- Armazenar grandes dados;
- Biblioteca exclusiva;
- Pode causar lentidão durante a análise.

SLOW5

Oxford Nanopore

- Criado a parti do FAST5;
- Codifica todas as informações encontradas no FAST5;
- Não depende de biblioteca única.

TIPOS DE DADOS PARA ALINHAMENTO

SAM

Sequence Alignment Map

- Realiza mapeamento de leitura eficiente contra grandes sequências de referência;
- Armazena alinhamentos de leitura para uma sequência de referência;
- Escala conjuntos de alinhamento de 1.011 ou mais pb.
- Uma seção de cabeçalho e uma seção de alinhamento.

BAM

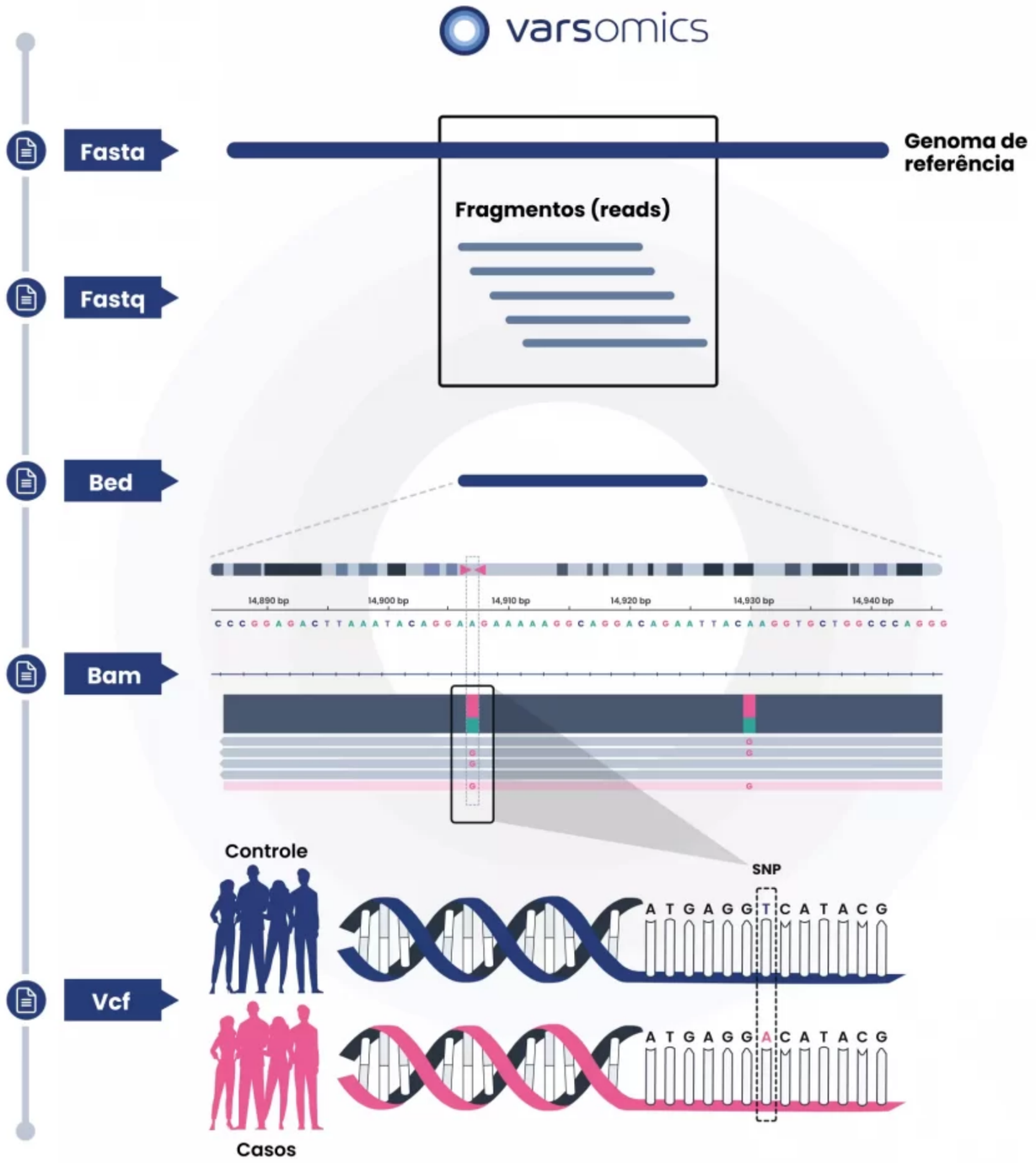
- Formato complementar Binary Alignment/Map;
- Representação binária do SAM e mantém exatamente as mesmas informações;
- O objetivo é reduzir o espaço que ocupa no disco.

BLOW5

Oxford Nanopore

- Formato SLOW5 codificado em formato binário;
- Análogo ao formato SAM/BAM para armazenar alinhamentos de sequência;
- Alocação de espaço mais simples e à redução da redundância de metadados.

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).



GENOME ASSEMBLY VALIDATION

THEMES

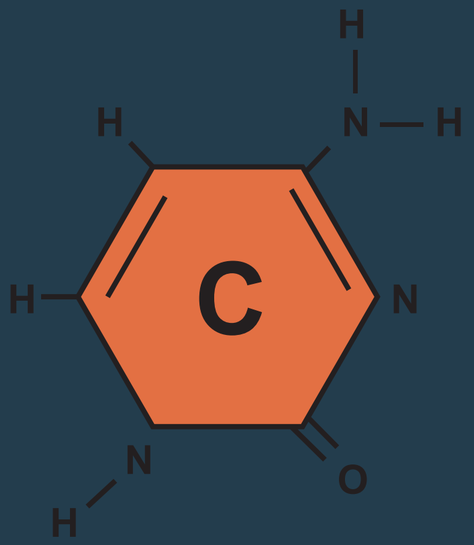
- Continuity
- Structural accuracy
- REAPR
- Base accuracy
- Functional completeness
- BUSCO
- Chromosome status

CONTINUITY

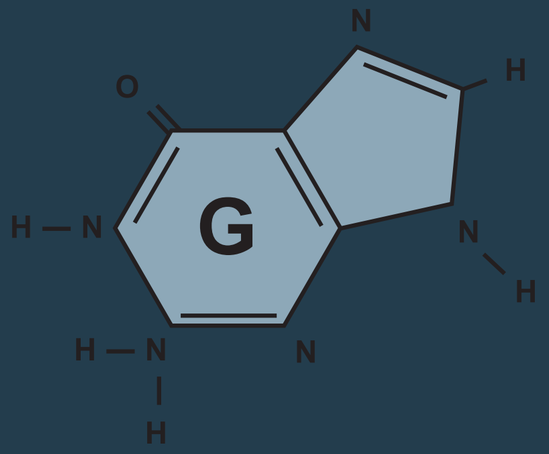
Quality category	Metric	Finished	VGP-2020	VGP-2016	B10k-2014
Notation	***x.y.P.Q.C***	c.c.Pc.Q60.C100	7.c.P6.Q50.C95	6.7.P5.Q40.C90	4.5.Q30
Continuity	**Contig NG50 (*x*)**	= Chr. NG50	>10 Mb	>1 Mb	>10 kb
	Scaffolds NG50 (*y*)	= Chr. NG50	= Chr. NG50	>10 Mb	>100 kb
	Gaps per Gb	No gaps	<200	<1,000	<10,000

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

BLOBTOOLS



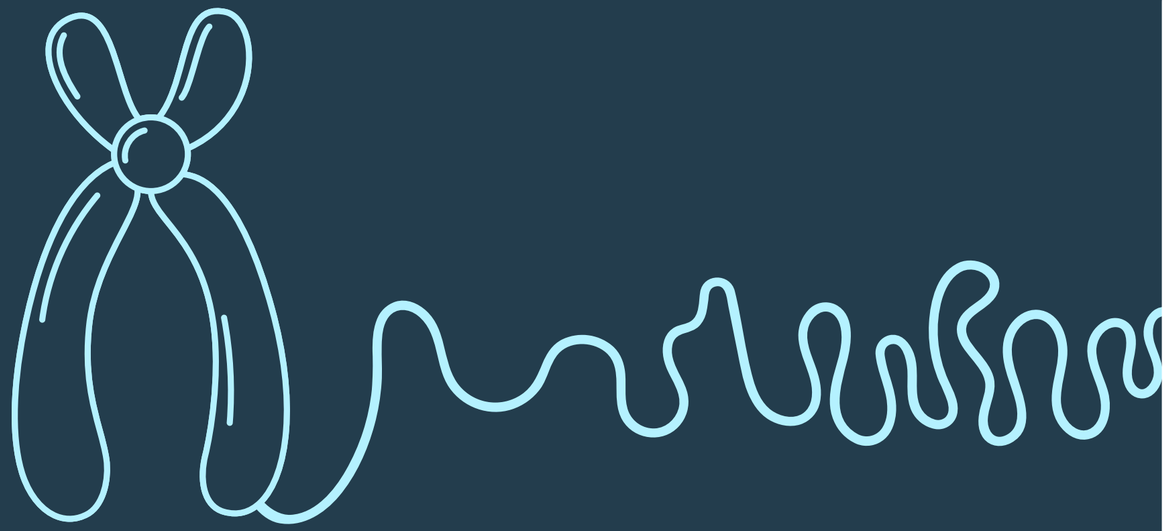
BlobTools is a bioinformatics tool used for the visualization, quality control, and taxonomic partitioning of genome datasets. It can assist in primary partitioning of data, leading to improved assemblies, and screening of final assemblies for potential contaminants



STRUCTURAL ACCURACY

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



REAPR

REAPR is a tool that evaluates the accuracy of a genome assembly using mapped paired end reads, without the use of a reference genome for comparison

BASE ACCURACY

Quality category	Metric	Finished	VGP-2020	VGP-2016
Base accuracy	**Base pair QV (*Q*)**	>60	>50	>40
	***k*-mer completeness**	100% complete	>95%	>90%

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

FUNCTIONAL COMPLETENESS

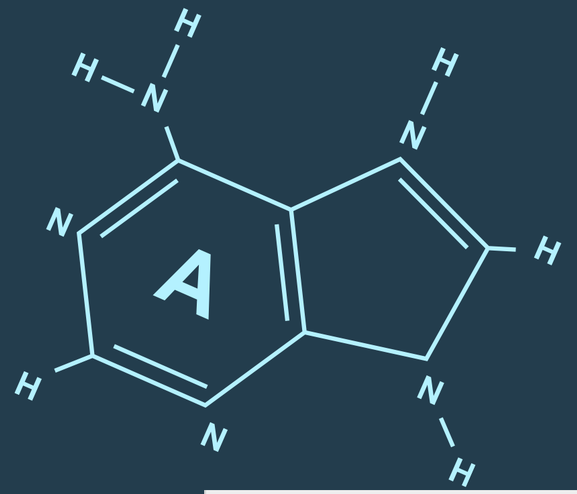
Quality category	Metric	Finished	VGP-2020	VGP-2016	B10k-2014
Functional completeness	**Genes**	>98% complete	>95% complete	>90%	>80%
	Transcript mappability	>98%	>90%	>80%	>70%

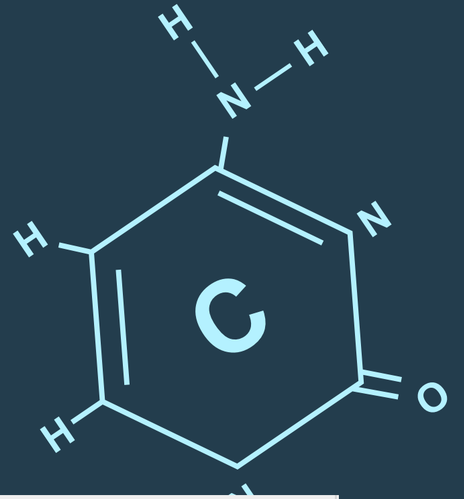
Rhie, A., McCarthy, S.A., Fedrigo, O. et al. Towards complete and error-free genome assemblies of all vertebrate species. Nature 592, 737-746 (2021). https://doi.org/10.1038/s41586-021-03451-0

BUSCO

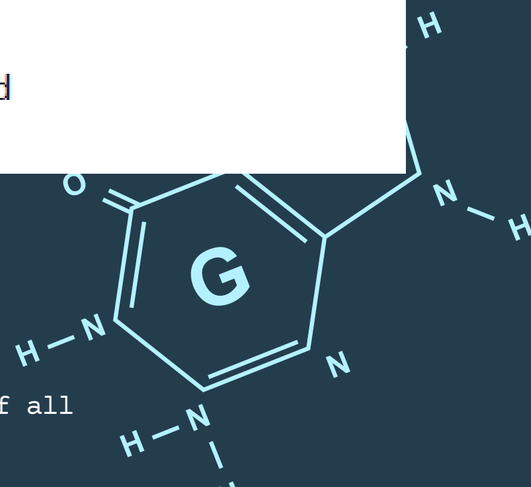
BUSCO is a bioinformatics tool used to assess the quality of genome annotations and the completeness of genome assemblies. BUSCO provides quantitative measures for assessing gene completeness in terms of expected genes found in a dataset.

CHROMOSOME STATUS





Quality category	Metric	Finished	VGP-2020	VGP-2016
Chromosome status	**Assigned (C)**	>100%	>95%	>90%
	Sex chromosomes	Right order, no gaps	Localized homo pairs	At least one shared (for example, X or Z)
	Organelles (for example, MT)	One complete allele	One complete allele	Fragmented



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