

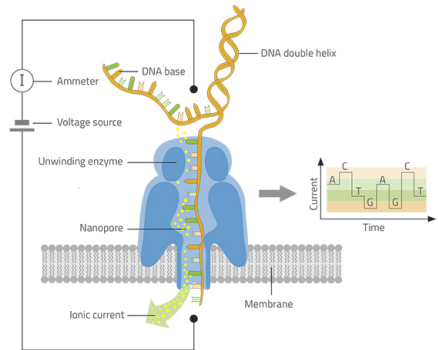
Introdução à bioinformática

Introduction to Bioinformatics

September 28th, 2026

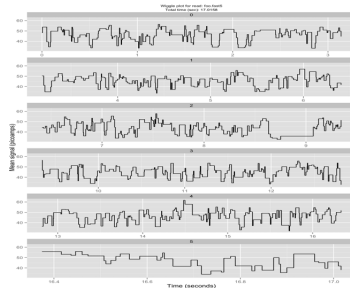


- ❶ Identificação das bases
Basecalling
- ❷ Demultiplexação
Demultiplexing
- ❸ Avaliação da qualidade das leituras
Read quality assessment
- ❹ Construção do consenso
Consensus generation
- ❺ Correção de erros
Error-correcting
- ❻ Identificação da espécie
Species identification



Kerstin Goeprich (www.scienceinschool.org)

- 1 Identificação das bases
Basecalling
- 2 Demultiplexação
Demultiplexing
- 3 Avaliação da qualidade das leituras
Read quality assessment
- 4 Construção do consenso
Consensus generation
- 5 Correção de erros
Error-correcting
- 6 Identificação da espécie
Species identification



```
>read_1  
TTTATAACGCTCACCCCTCTGCATCTC
```

Fastq Format

@HISEQ:119:C42B3ACXX:7:1101:2009:2249 2:N:0:CGACCTG
TCTTGGGGACAGGGAATTTTCATTTCCAAATGAAATCCTCAAAGAACGCCTTTTATTTACAGGAGGCTGTATATCTTAGCCAAAGTGGTAGATCGGAAGA
+
BB<BFFBFB<BFF7BBF<BF<FBB<FFF<FFBFF<BFFFBBF<B<7BB<BFBFB<BFFFF<FFFFFF<BBFB<BB<BBB8BBB7<B<BB<77<BBB77

Quality Score Encoding

```

#####XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....
.....XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....
.....IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII..
.....JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ..
..LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL...
!"%&'()*+,-./0123456789:<=>?@ABCDEFGHIJKLMNPQRSTUUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
33|          59   64       73              104      |
0.....26...31.....40
           -5....0.....9.....40
             0.....9.....40
               3.....9.....40
0.2.....26...31.....41

S - Sanger            Phred+33,    raw reads typically (0, 40)
X - Solexa            Solexa+64,    raw reads typically (-5, 40)
I - Illumina 1.3+     Phred+64,    raw reads typically (0, 40)
J - Illumina 1.5+     Phred+64,    raw reads typically (3, 40)
with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
(Note: See discussion above).
I - Illumina 1.8+     Phred+33,    raw reads typically (0, 41)

```

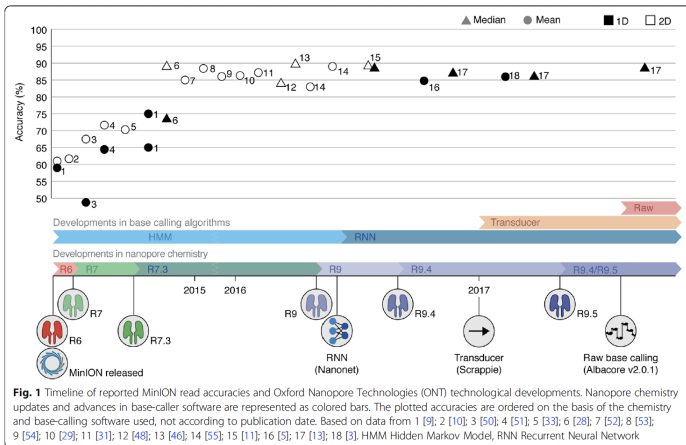
- 1 Identificação das bases
Basecalling
- 2 Demultiplexação
Demultiplexing
- 3 Avaliação da qualidade das leituras
Read quality assessment
- 4 Construção do consenso
Consensus generation
- 5 Correção de erros
Error-correcting
- 6 Identificação da espécie
Species identification



Dorado Basecaller

Error rate improvement

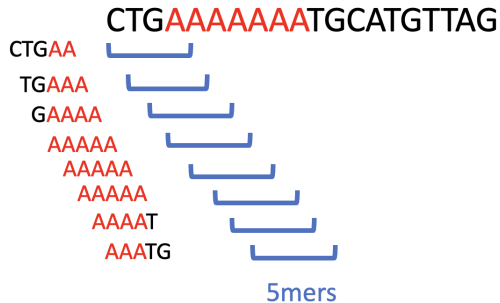
Melhoria da frequência de erros



Rang et al. 2018 *Genome Biology*

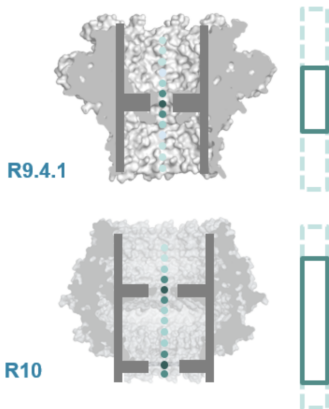
Nanopore errors

Erros da nanopore



Nanopore errors

Erros da nanopore

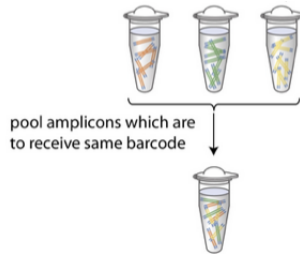


ATCGGAAAAAAAAAATCACGCCACGTCCAAA



Oxford Nanopore Technologies

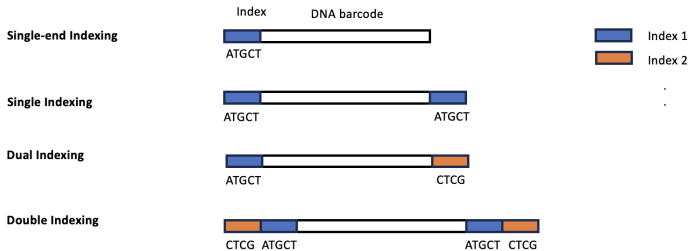
- 1 Identificação das bases
Basecalling
- 2 Demultiplexação
Demultiplexing
- 3 Avaliação da qualidade das leituras
Read quality assessment
- 4 Construção do consenso
Consensus generation
- 5 Correção de erros
Error-correcting
- 6 Identificação da espécie
Species identification



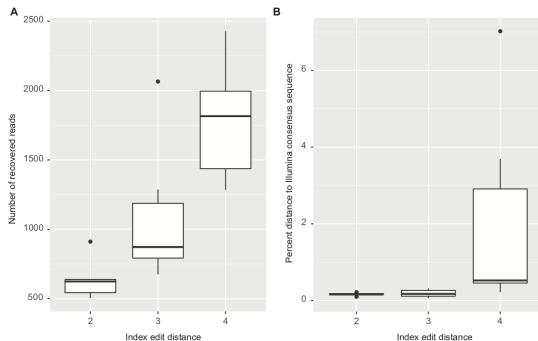
Oxford Nanopore Technologies

Indexing

Indexação

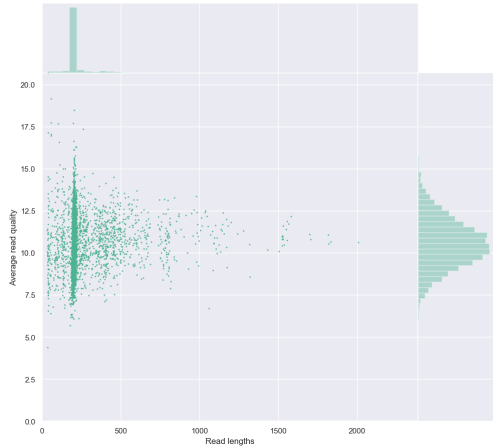


- 1 Identificação das bases
Basecalling
- 2 Demultiplexação
Demultiplexing
- 3 Avaliação da qualidade das leituras
Read quality assessment
- 4 Construção do consenso
Consensus generation
- 5 Correção de erros
Error-correcting
- 6 Identificação da espécie
Species identification



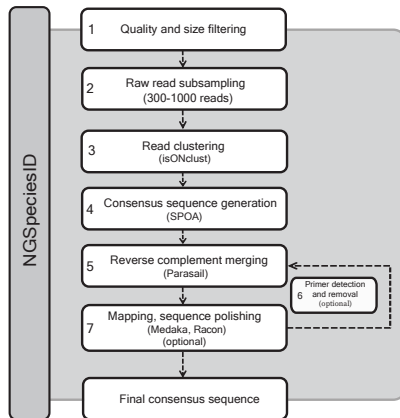
Krehenwinkel et al. 2019

Read lengths vs Average read quality plot



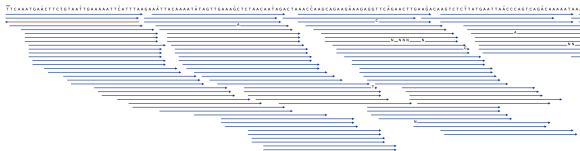
- 1 Identificação das bases
Basecalling
- 2 Demultiplexação
Demultiplexing
- 3 Avaliação da qualidade das leituras
Read quality assessment
- 4 Construção do consenso
Consensus generation
- 5 Correção de erros
Error-correcting
- 6 Identificação da espécie
Species identification

- 1 Identificação das bases
Basecalling
- 2 Demultiplexação
Demultiplexing
- 3 Avaliação da qualidade das leituras
Read quality assessment
- 4 Construção do consenso
Consensus generation
- 5 Correção de erros
Error-correcting
- 6 Identificação da espécie
Species identification

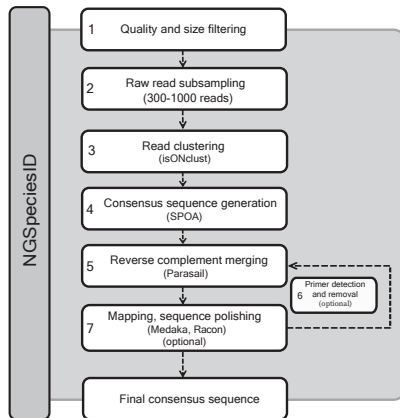


Sahlin, Lim and Prost 2020

Mapeamento à referência vs reconstrução de novo



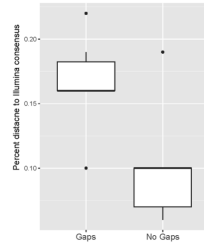
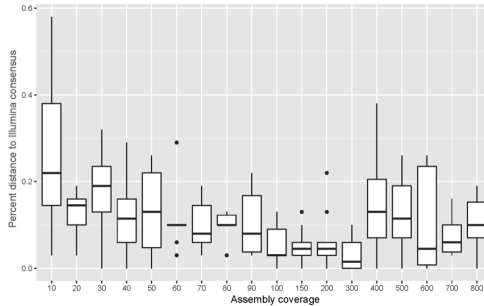
- 1 Identificação das bases
Basecalling
- 2 Demultiplexação
Demultiplexing
- 3 Avaliação da qualidade das leituras
Read quality assessment
- 4 Construção do consenso
Consensus generation
- 5 Correção de erros
Error-correcting
- 6 Identificação da espécie
Species identification



Sahlin, Lim and Prost 2020

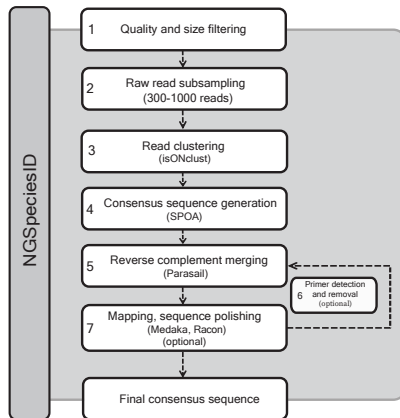
Consensus Accuracy

Precisão do consenso



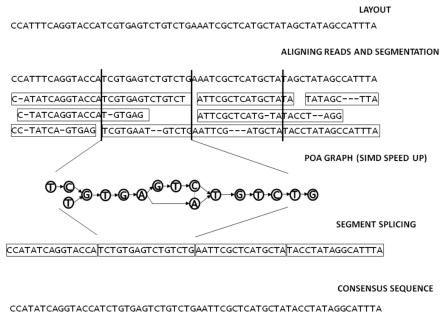
Krehenwinkel et al. 2019 *GigaScience*

- 1 Identificação das bases
Basecalling
- 2 Demultiplexação
Demultiplexing
- 3 Avaliação da qualidade das leituras
Read quality assessment
- 4 Construção do consenso
Consensus generation
- 5 Correção de erros
Error-correcting
- 6 Identificação da espécie
Species identification



Sahlin, Lim and Prost 2020

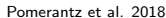
- 1 Identificação das bases
Basecalling
- 2 Demultiplexação
Demultiplexing
- 3 Avaliação da qualidade das leituras
Read quality assessment
- 4 Construção do consenso
Consensus generation
- 5 Correção de erros
Error-correcting
- 6 Identificação da espécie
Species identification



Vaser et al. 2016

A

Bothriechis schlegelii



6 Identificação da espécie Species identification

RID KIKT3V06014 (Expires on 07-23 22:49 pm)

Query ID kciQuery_33635
Description filtered_read1_rac2
Molecule type dna
Query Length 958

Database Name nr
Description Nucleotide collection (nr)
Program BLASTN 2.9.0+ [Help](#)

Other reports: [Search Summary](#) | [Taxonomy reports](#) | [Distance tree of results](#) | [MSA viewer](#)

Graphic Summary

Distribution of the top 100 Blast Hits on 100 subject sequences

Mouse over to see the 150, click to show alignments

Color key for alignment scores

■ <40 ■ 40-50 ■ 50-80 ■ 80-200 ■ >=200

Query 1 150 300 450 600 750 900

Descriptions

Sequences producing significant alignments

Select: [All](#) [None](#) Selected: 0

[Alignments](#) | [Download](#) | [Statistics](#) | [Graphical](#) | [Taxonomy tree of results](#)

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
Alouatta seniculus mitochondrial complete genome	763	763	72%	0.0	86.60%	HQ644333.1
Alouatta palliata claudiana mitochondrial complete genome	747	747	72%	0.0	86.20%	HY202428.1
Alouatta palliata mitochondrial complete genome	730	730	72%	0.0	86.75%	KC797386.1
Alouatta palliata mitochondrial complete genome	730	730	72%	0.0	86.77%	KC797384.1
Alouatta palliata mitochondrial complete genome	730	730	72%	0.0	86.75%	FJ789442.1
Alouatta palliata mitochondrial complete genome	725	725	72%	0.0	86.61%	AB018730.1

- 1 Identificação das bases
Basecalling
- 2 Demultiplexação
Demultiplexing
- 3 Avaliação da qualidade das leituras
Read quality assessment
- 4 Construção do consenso
Consensus generation
- 5 Correção de erros
Error-correcting
- 6 Identificação da espécie
Species identification

- **Bit Score**

- Mede a qualidade do alinhamento
Measures the quality of the alignment.

- **Query Coverage**

- Quantidade de sobreposição das correspondências
How much of the query matches.

- **E-value**

- Probabilidade de uma correspondência ser completamente aleatória (depende do tamanho da base de dados)
Likelihood that a given sequence match is purely by chance (depends on the database size)

- **Percent Identity**

- O quão semelhante é a sequência de interesse e a sequência da base de dados
How similar are the query and the database sequence.