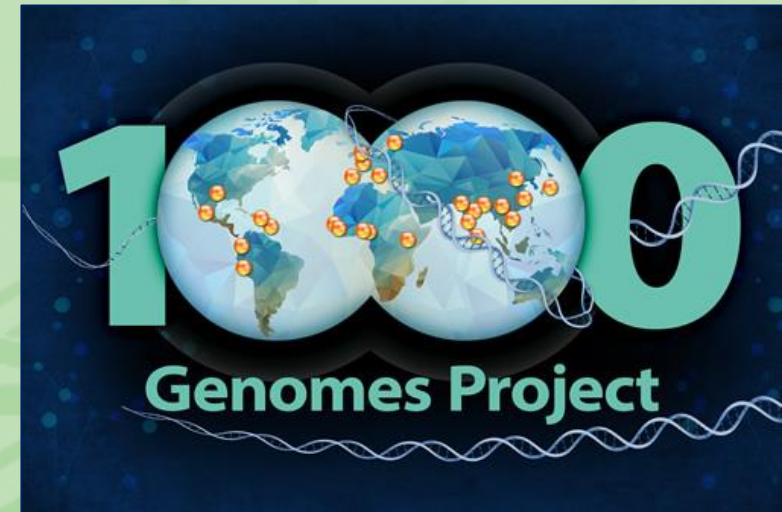
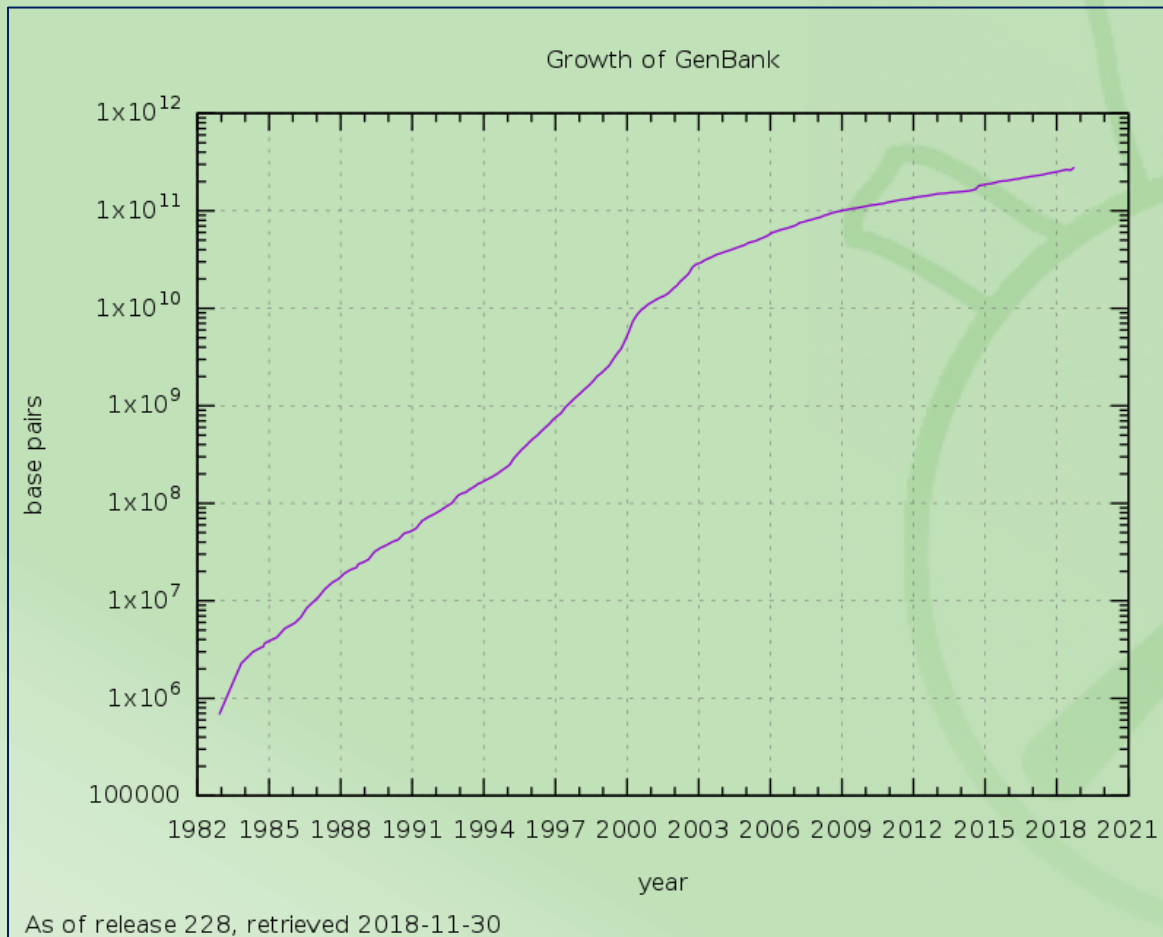



# Noções básicas sobre sequenciamento, montagem e análise de genomas

Dr<sup>a</sup> Desirrê Petters-Vandresen

# Interações patógeno-planta em um contexto genômico

- Grande volume de dados biológicos sendo disponibilizados em bancos de dados, em constante crescimento: inúmeras possibilidades de aplicações em genética funcional



**JGI**  **MycoCosm**  
THE FUNGAL GENOMICS RESOURCE

Home Outreach Video Tutorials About

**1000 Fungal Genomes Project**

[Nominate a genome to sequence](#)

# Desafios

- Como lidar com o crescente volume de dados que surge a partir das mais variadas técnicas e estudos?
- Como acessar, organizar, gerenciar e processar estes grandes conjuntos de dados?
- Como explorar totalmente o potencial dos conjuntos de dados, especialmente em situações de grande demanda computacional?
- Como comparar novos dados com estudos prévios e permitir que esses dados sejam comparados com estudos futuros?

# Desafios

- Como determinar se os dados que estão presentes em bancos de dados atendem às minhas necessidades para análises de genética funcional?
- Como saber uma montagem de genoma ou transcriptoma é confiável?
- Como saber se a metodologia de sequenciamento utilizada é adequada para responder à minha pergunta de estudo?
- Como saber se posso usar dados existentes na literatura ou se precisarei sequenciar um novo genoma?

# Aspectos importantes

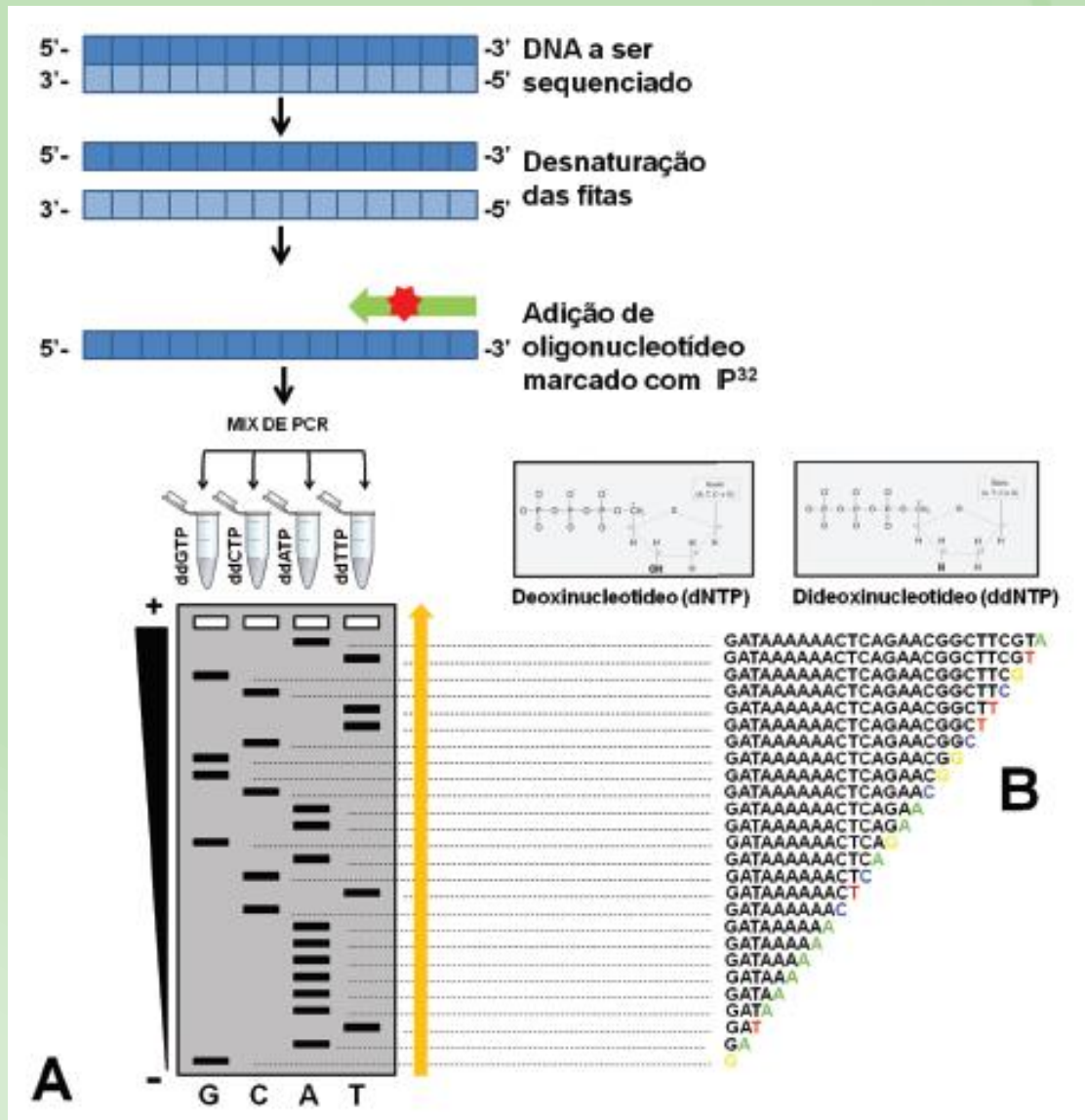
- Noções gerais sobre:
  - Principais metodologias de sequenciamento de genomas e transcriptomas
  - Principais métricas de avaliação de qualidade
  - Formatos de arquivo e bancos de dados biológicos úteis para estudos funcionais



# Sequenciamento de genoma ou transcriptoma

- Identificação da sequência de nucleotídeos de uma molécula de DNA ou RNA na sua ordem correta, para conhecer a informação genética presente nesta estrutura
- Além da identidade de cada base, o sequenciamento também fornece informações sobre a confiabilidade de cada uma das bases identificadas
- Avanços na escala de sequenciamento nos últimos 50 anos: do sequenciamento manual ao sequenciamento maciço e paralelo de genomas inteiros em um curto período de tempo

# Sequenciamento de Sanger



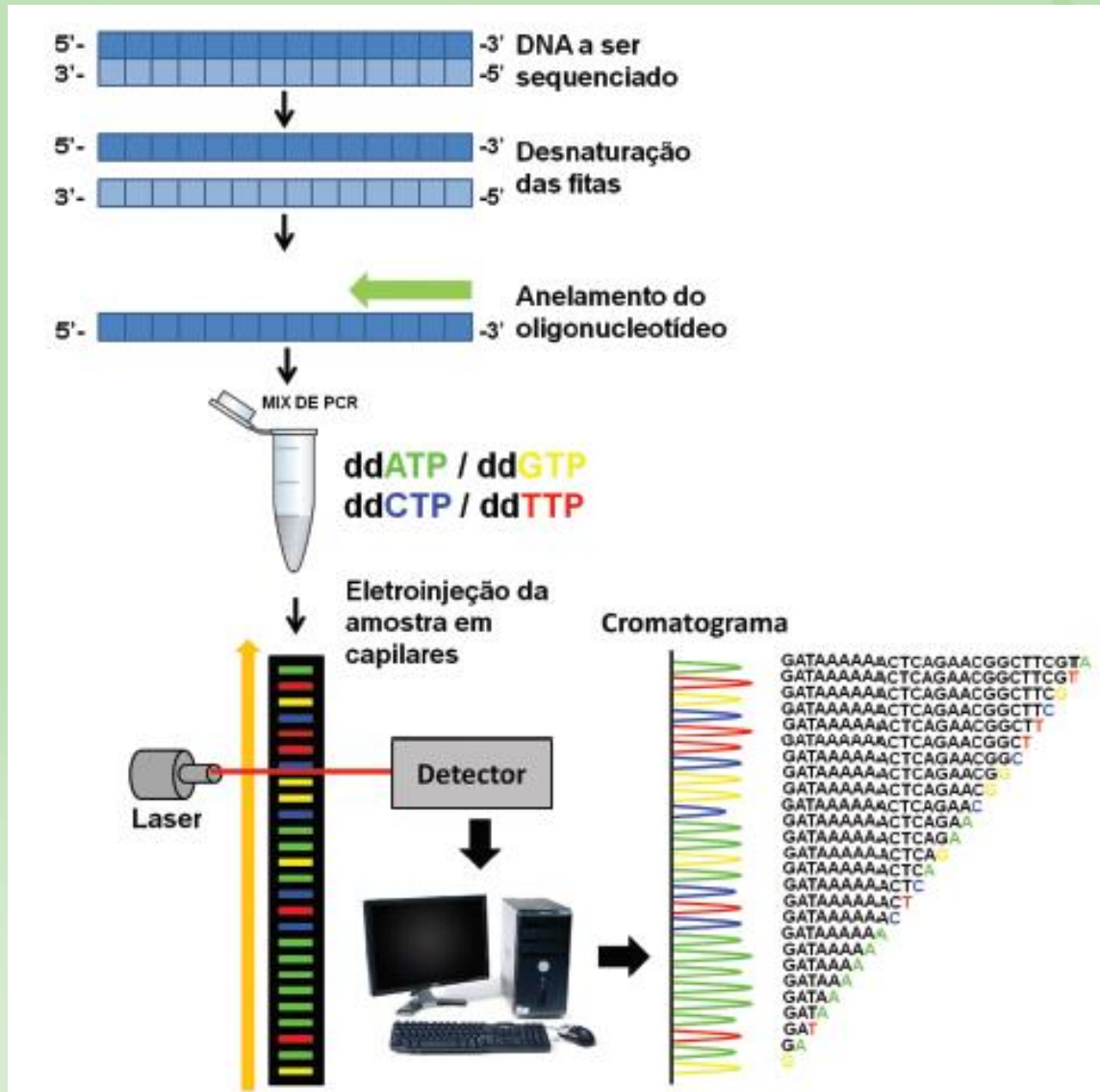
- Reação de PCR com deoxinucleotídeos modificados (dideoxinucleotídeos), marcados com fósforo ( $P^{32}$ ) ou enxofre ( $S^{35}$ ) radioativos
- Incorporação de dideoxinucleotídeos interrompe a síntese da nova molécula de DNA
- Produtos das reações de PCR com dideoxinucleotídeos são submetidos à eletroforese para separação por tamanho, e perfil de bandas é lido de baixo para cima para determinar a sequência

# Automatização do sequenciamento de Sanger

- Substituição dos dideoxinucleotídeos marcados com radiação por dideoxinucleotídeos marcados com fluoróforos:
  - Menor risco à saúde
  - Uso de fluoróforos diferentes para cada uma das bases: emissão de fluorescência em comprimentos de onda distintos e possibilidade de realizar a reação num tubo único
- Geis de eletroforese substituídos por capilares preenchidos com gel
  - Maior quantidade de amostras analisadas no mesmo período de tempo
  - Maior automação e diminuição do trabalho manual do analista



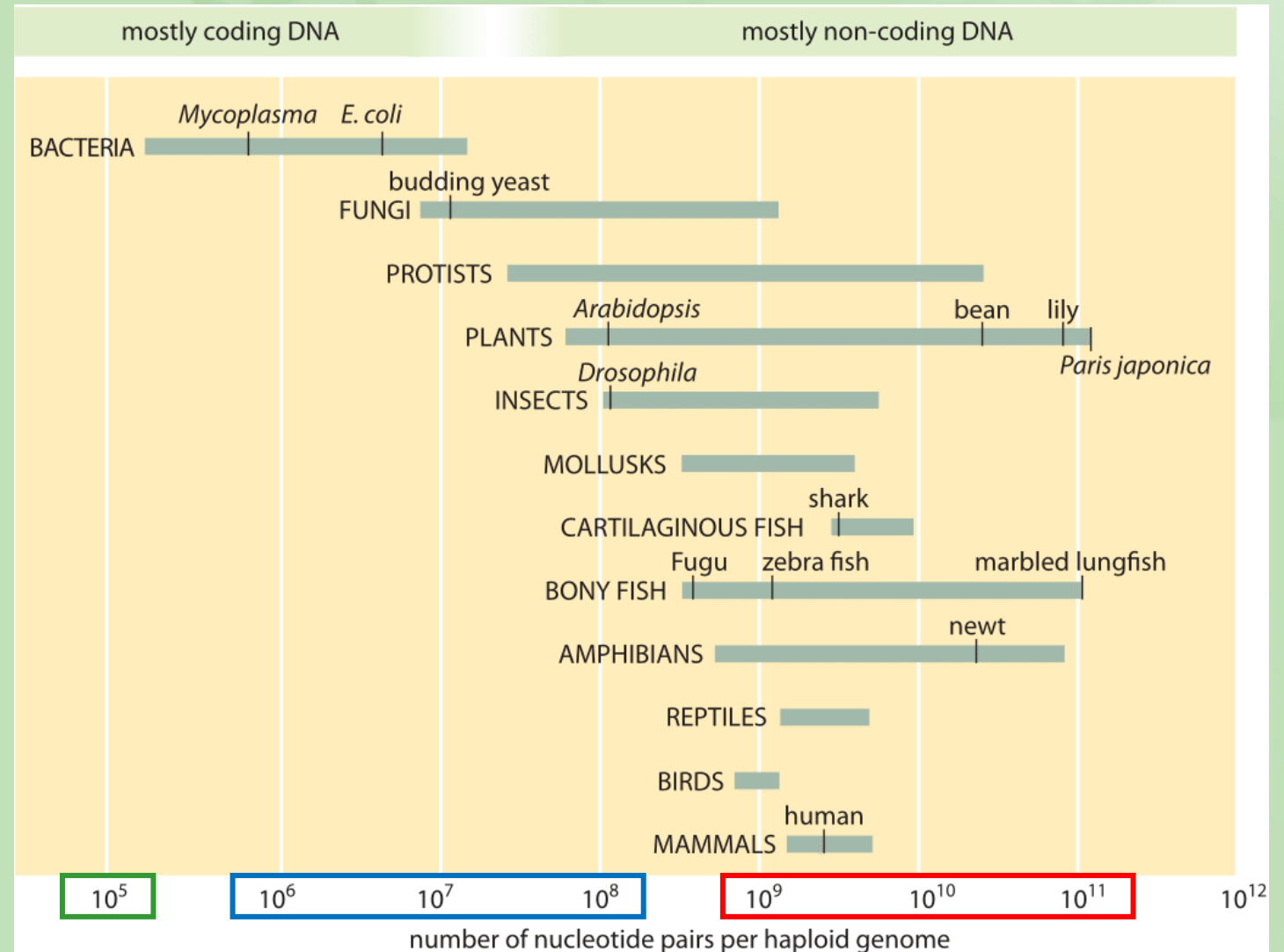
# Automatização do sequenciamento de Sanger



- Reação de PCR com deoxinucleotídeos marcados com fluoróforos
- Produtos de PCR migram ao longo dos capilares e passam por um feixe de raios laser que excita os fluoróforos, fazendo com que emitam fluorescência
- A intensidade e comprimento de onda da fluorescência é registrada pelo detector e interpretada pelo computador para gerar o cromatograma
- O cromatograma é decodificado na sequência de nucleotídeos do fragmento

# Estratégias de sequenciamento de genoma usando sequenciamento de pequena escala

- Sequenciamento automatizado de pequena escala: fragmentos de ~700 nucleotídeos
- Genomas completos: **milhares**, **milhões** ou até **bilhões** de pares de bases
- Necessidade de fragmentação e posterior montagem para obtenção do genoma completo



# Limitações do sequenciamento de pequena escala nos projetos de genomas e transcriptomas

- Erros ou não detecção das bases iniciais
- Preparação das amostras demorada e laboriosa
- Fragmentos pequenos
- **Necessidade de estratégias mais rápidas, baratas, precisas e com maior capacidade de leitura**

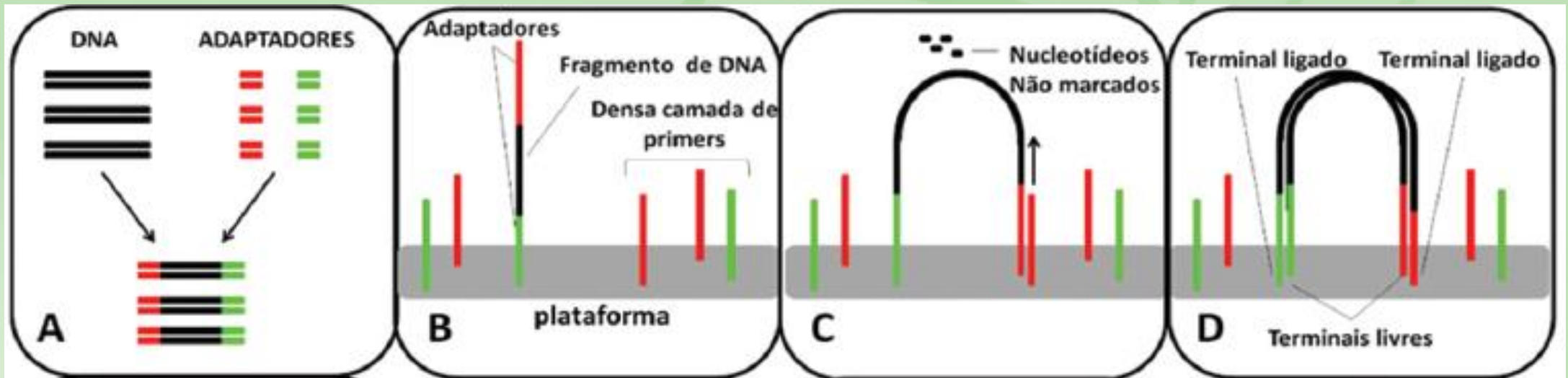
# Sequenciamento de nova geração (larga escala, 2ª geração)

- Reações não são baseadas em eletroforese
- Alta capacidade de geração de uma grande quantidade dados: genomas sequenciados em uma única corrida
- Preparação de bibliotecas independentes de clonagem



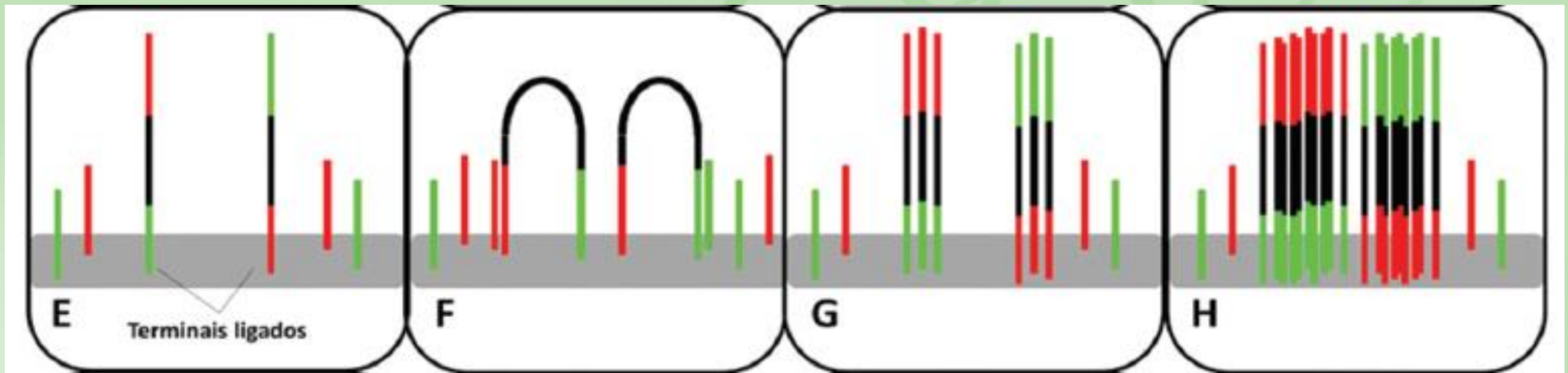
# Illumina

- Preparação: fragmentação do DNA por nebulização, seleção por tamanho e ligação dos adaptadores às extremidades dos fragmentos
- Ligação dos fragmentos à plataforma, que contém uma densa camada de primers. Ocorre a incorporação de nucleotídeos não marcados com fluorescência até que ocorra a amplificação de todo o fragmento
- Há formação da estrutura em ponte (amplificação em ponte), com dois adaptadores presos à placa e dois livres



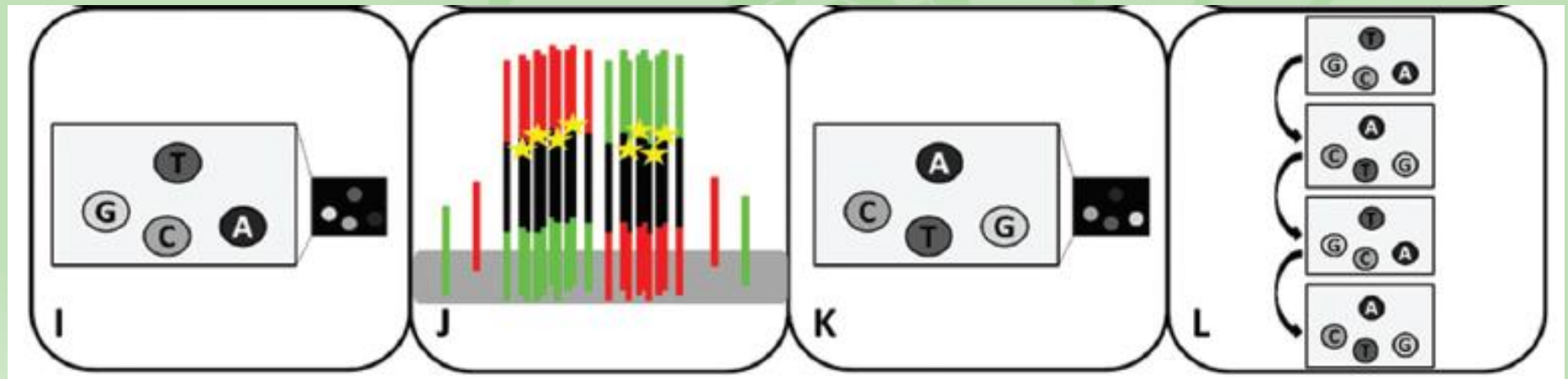
# Illumina

- Novos ciclos de amplificação em ponte ocorrem, até a formação de clusters com mais de um milhão de cópias do mesmo fragmento



# Illumina

- Após a formação dos clusters, dideoxinucleotídeos são adicionados antes da próxima amplificação
- Após a incorporação dos dideoxinucleotídeos durante a amplificação, um feixe de raios laser excita os fluoróforos e a luz emitida é registrada pelo detector, fazendo a leitura de bases naquela posição
- Ocorre uma lavagem para remoção dos grupos bloqueadores presentes nas extremidades 3' dos dideoxinucleotídeos para que a reação possa continuar
- Esse processo é repetido sucessivamente até que toda a extensão do fragmento de DNA seja polimerizada e o fragmento seja sequenciado
- Por fim, as leituras são decodificadas para determinar a sequência de bases dos fragmentos



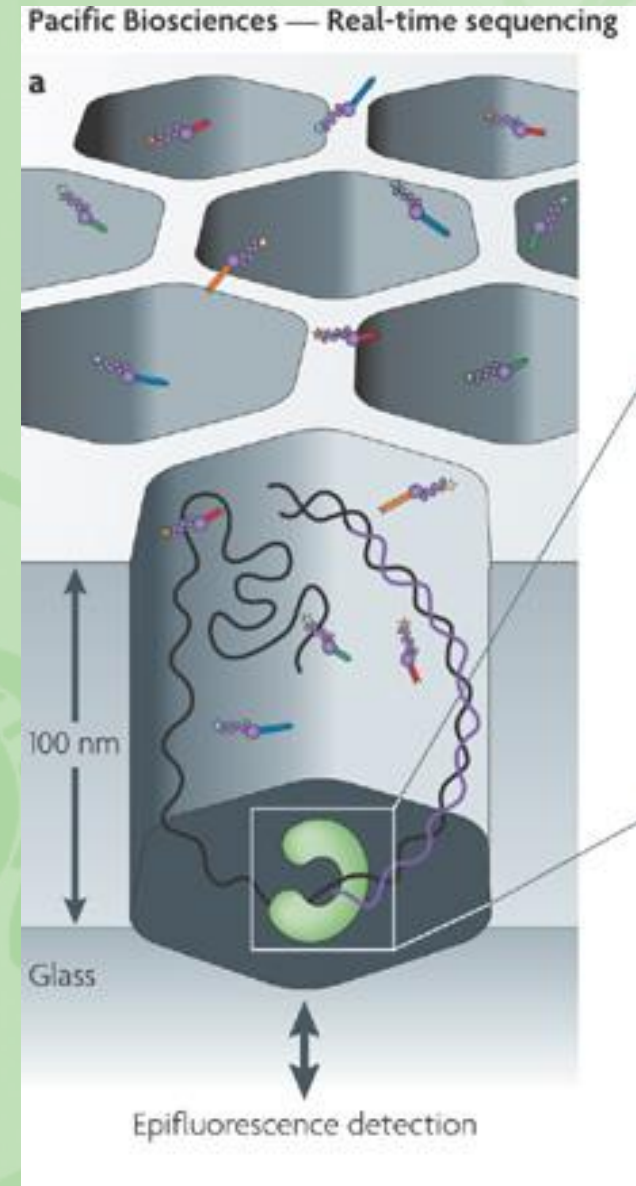
# Sequenciamento de nova geração (larga escala, 3ª geração)

- Capacidade de sequenciamento de uma única molécula de DNA
- Sem necessidade de amplificação
- Altíssima capacidade de geração de um grande volume de dados em curto período de tempo: um genoma humano pode ser sequenciado em uma única corrida



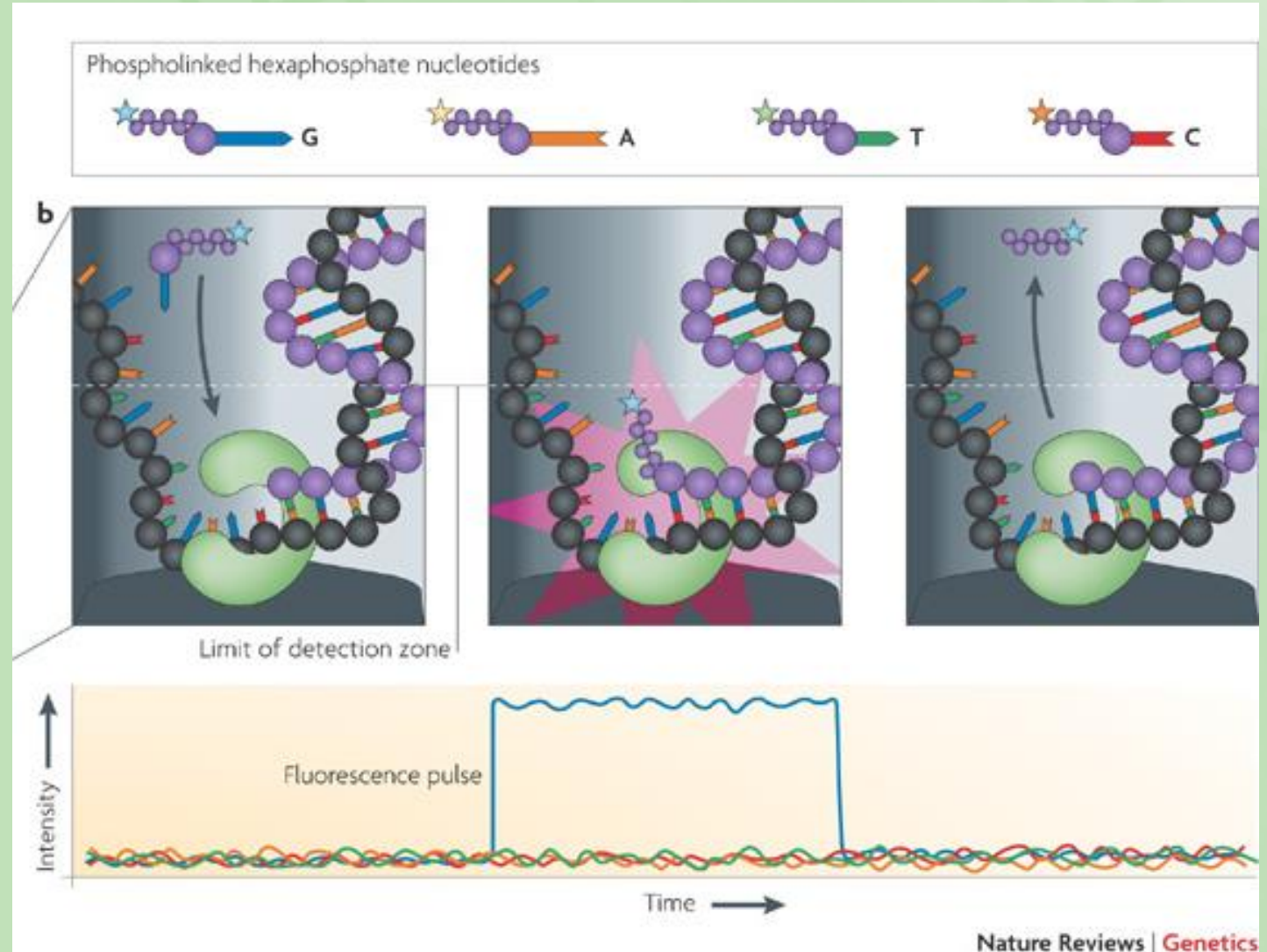
# Pacific Biosciences SMRT Sequencing (*Single Molecule Real Time Sequencing*)

- Janela de observação em nano-escala (ZMW, *zero-mode waveguide*), com um volume extremamente reduzido, suficiente para visualizar a incorporação de um único nucleotídeo pela DNA polimerase
- Uma única DNA polimerase contendo uma única molécula de DNA molde é fixada no fundo da ZMW



# Pacific Biosciences SMRT Sequencing (Single Molecule Real Time Sequencing)

- Nucleotídeos com fluoróforos são utilizados, e quando um nucleotídeo é incorporado a marcação fluorescente é clivada, emitindo luz, que é detectada e transformada em dado de sequência

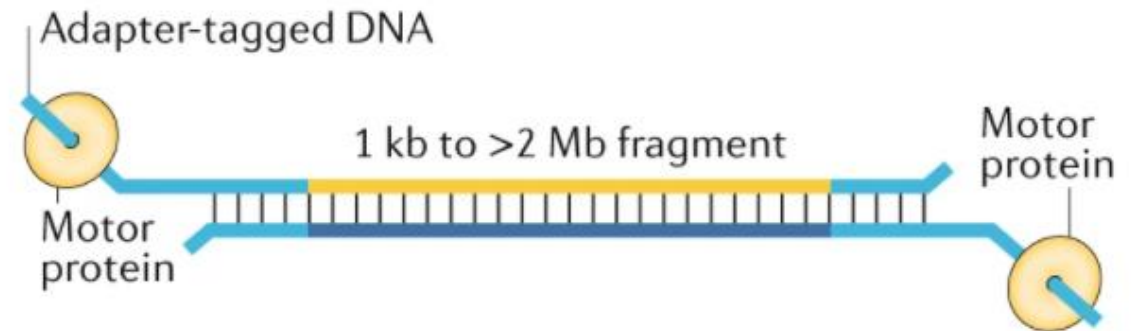


# Oxford Nanopore Technologies

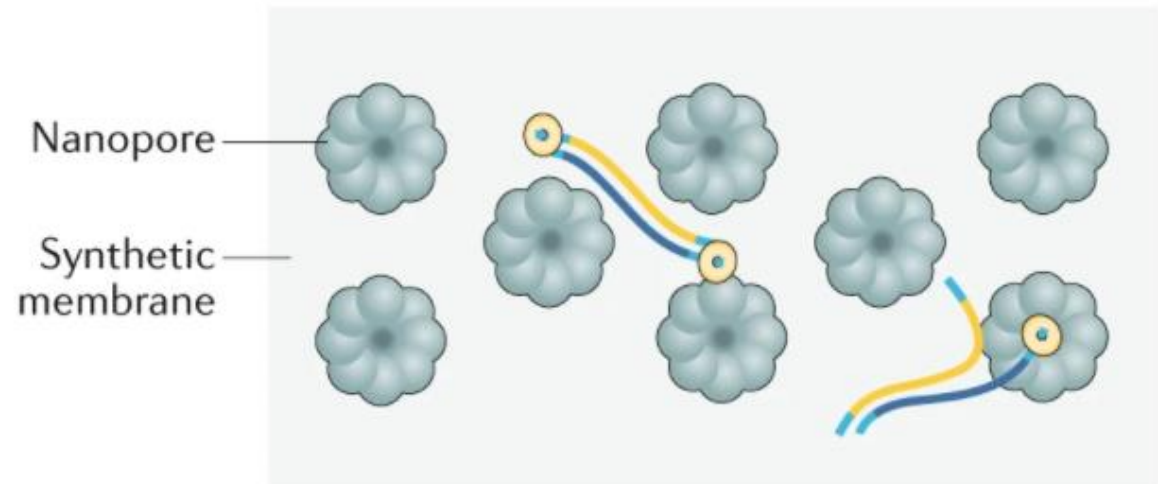
- DNA é marcado com adaptadores com proteínas motoras em uma ou ambas as extremidades e é combinado à proteínas carregadoras, que o direcionam aos nanoporos
- A plataforma de sequenciamento contém milhares de nanoporos de proteicos associados à uma membrana sintética

## b ONT sequencing

### Template topology



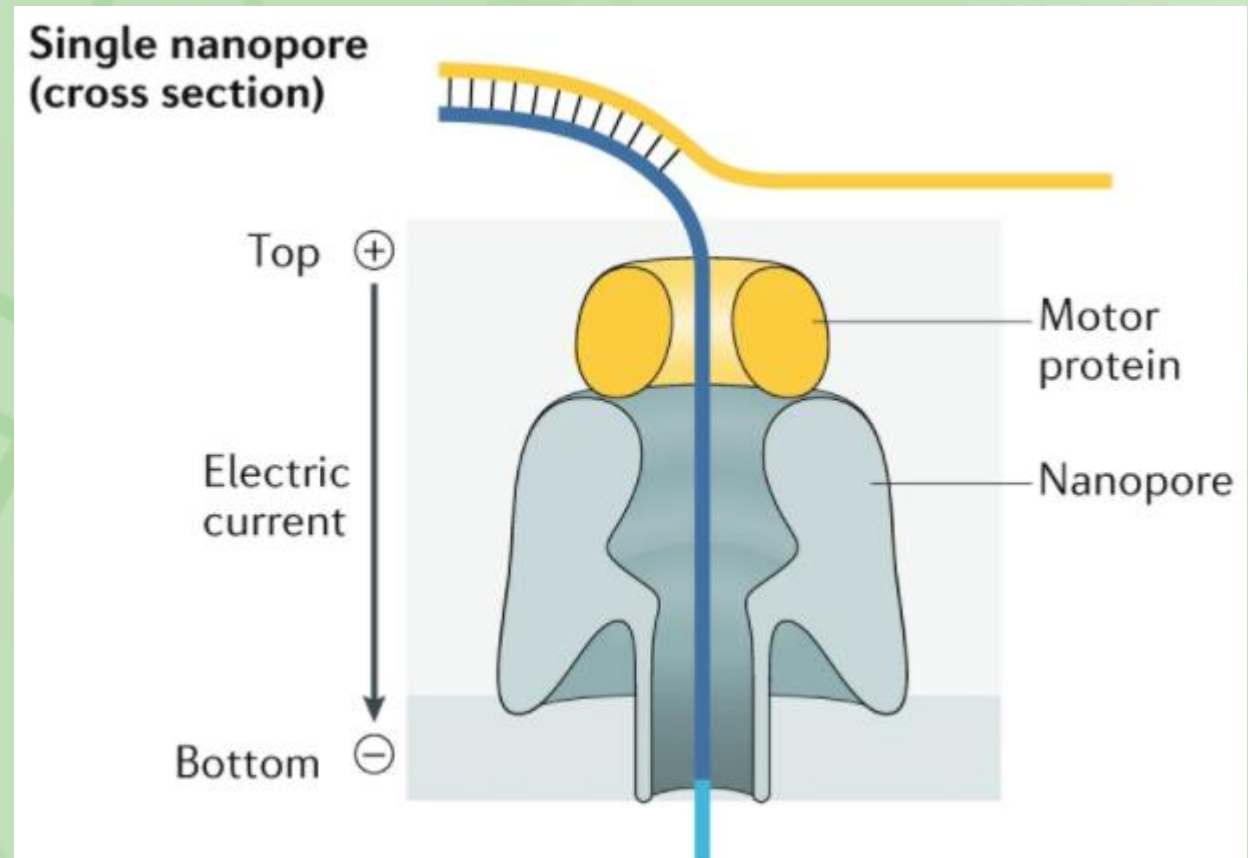
### Flow cell (top view)





# Oxford Nanopore Technologies

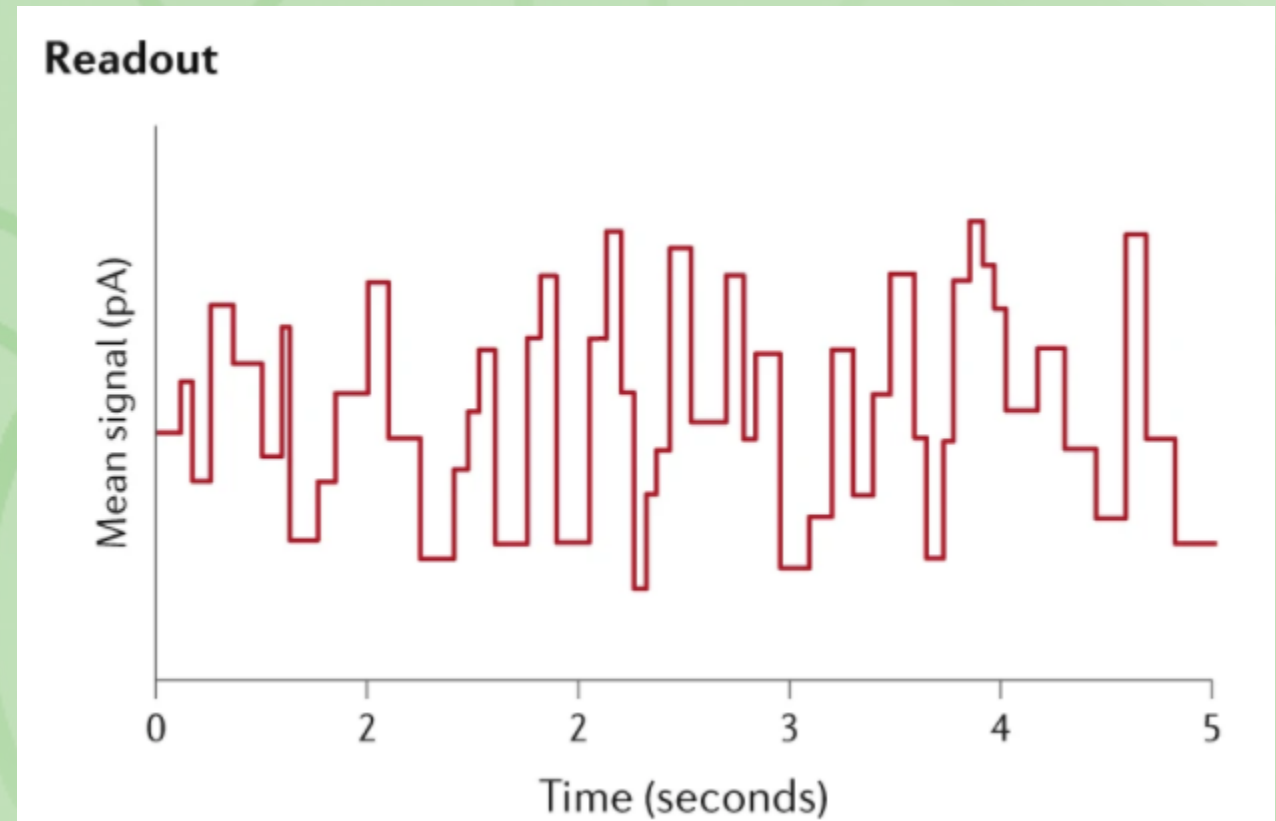
- O adaptador se insere na abertura do nanoporo, e a proteína motora começa a separar as fitas do DNA
- Uma corrente elétrica é aplicada e, em conjunto com a proteína motora, conduz o DNA carregado negativamente através do poro numa velocidade de ~450 bases por segundo





# Oxford Nanopore Technologies

- À medida que o DNA se move pelo poro, causa perturbações à corrente elétrica, as quais são específicas para cada um dos nucleotídeos
- O perfil de mudanças na corrente elétrica pode ser utilizado para identificar a sequência de bases da molécula



# Comparativo entre metodologias

Método	Sanger	454	Illumina	PacBio	Nanopore
Comprimento dos reads	400 - 900 pb	700 bp	100 – 300 pb	10 – 100 kb	Variável (até 1000 kb)
Taxa de erro	0.01 %	0.1 %	0.1%	5 – 15%*	5 – 20%*
Eficiência (bases por corrida)	1.9 - 84 Kb	1 Mb	200 – 600 Gb	10 – 20 Gb	5 – 10 Gb
Tempo de corrida	20 min – 3 horas	24 horas	1 – 3 dias	~ 30 horas	1 minuto até 72 horas
Prós	Alta confiabilidade	Velocidade	Alta confiabilidade e custo baixo	Reads longos, velocidade e alta eficiência	Reads longos, velocidade e alta eficiência
Contras	Baixa eficiência	Baixa eficiência e alto custo	Reads curtos, velocidade	Taxa de erro elevada e alto custo	Taxa de erro elevada e alto custo

Adaptado de:

Basal & Boucher et al. 2019 **iScience**. DOI: [10.1016/j.isci.2019.06.035](https://doi.org/10.1016/j.isci.2019.06.035)  
 Liu et al. 2012 **BioMed Research International**. DOI: [10.1155/2012/251364](https://doi.org/10.1155/2012/251364)  
<https://www.pacb.com/products-and-services/sequel-system/>

\* Em baixa cobertura

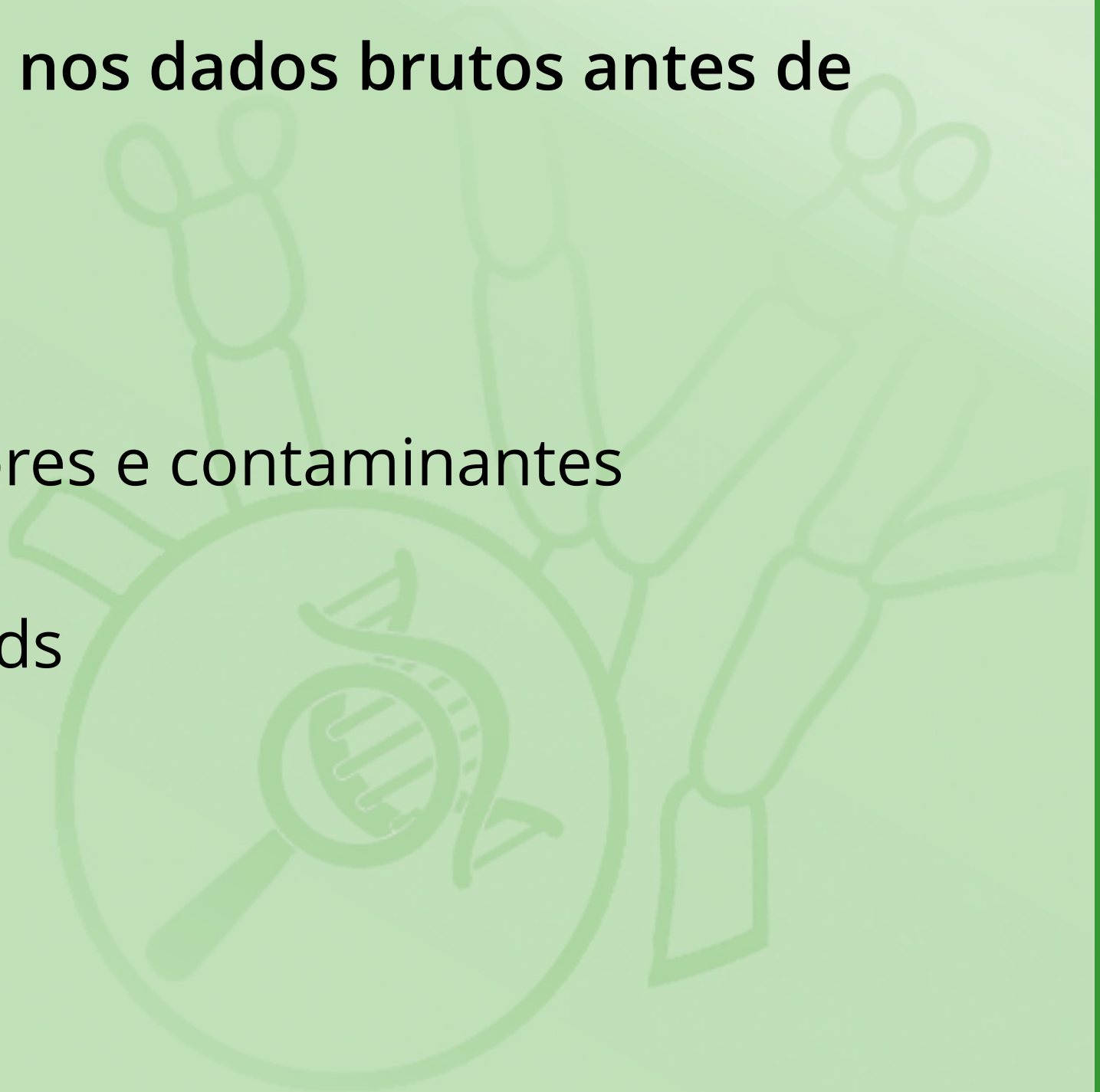
# Sequenciei um genoma ou obtive dados já sequenciados: e agora?

- Avaliar os dados brutos **com muita atenção**

- **Utilizar reads brutos que saem do equipamento ou genoma/transcriptoma obtidos sem qualquer tipo de controle de qualidade não é uma boa ideia!**
- Sequências erradas podem ter um grande impacto negativo nos resultados e conclusões de um estudo

# O que devemos avaliar nos dados brutos antes de utilizar?

- Qualidade das bases
- Presença de adaptadores e contaminantes
- Comprimento dos reads
- Quantidade de reads

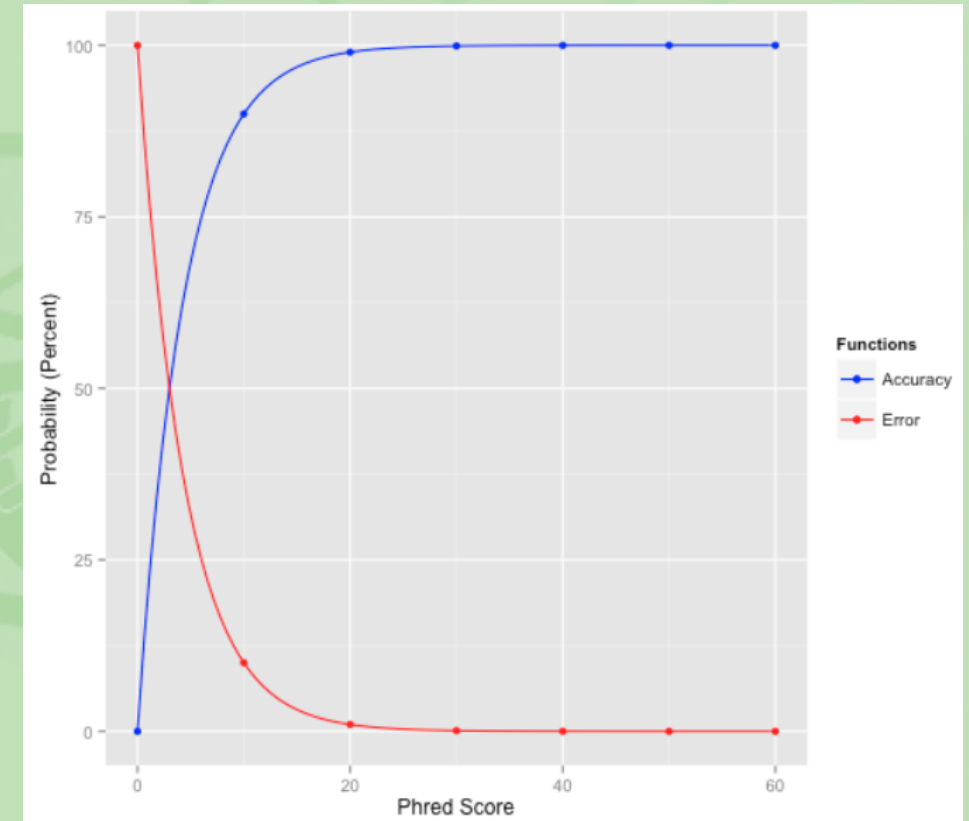




# Indicador de qualidade Q (*Phred quality score*)

- Baseado na probabilidade de erro ( $E$ ) na identificação de uma base em determinada posição do read
- Define a acurácia de uma base
  - 90%: um erro em cada 10 leituras (0.1),  $Q = 10$
  - 99%: um erro em cada 100 leituras (0.01),  $Q = 20$
  - 99,9%: um erro em cada 1.000 leituras (0.001),  $Q = 30$
  - 99,99%: um erro em cada 10.000 (0.0001),  $Q = 40$
- $Q < 20$ , a perda de confiabilidade é muito alta e rápida
- $Q > 20$ , o aumento na confiabilidade não é tão significativo
- **20 ou 25 como valores de corte em muitos casos**

$$Q = -10 \log E$$



# Formato FASTQ

- Formato de armazenamento de sequências biológicas e scores de qualidade correspondentes às bases

```
1 @A00178:149:H7K7YDSXY:4:1101:1506:1000 1:N:0:GCACTCAT+ATGAGTGC
2 CNGCTCTGGTCATCCGTCTCGGCTCGCGAGATTCAAGCGTTGCCGTCAACCTTGGCAATGTAGACAAGGA
  GGTCGAGGACACGGCGGGAGTAGCCCCACCTCGTTGTCGTACCAGGAGACGAGCTTGACGAAGTTCTCGTT
  GAGCGAGATAC
3 +
4 F#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:
  FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:
  FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:
  FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:
  FFFF:FF:FF:
```

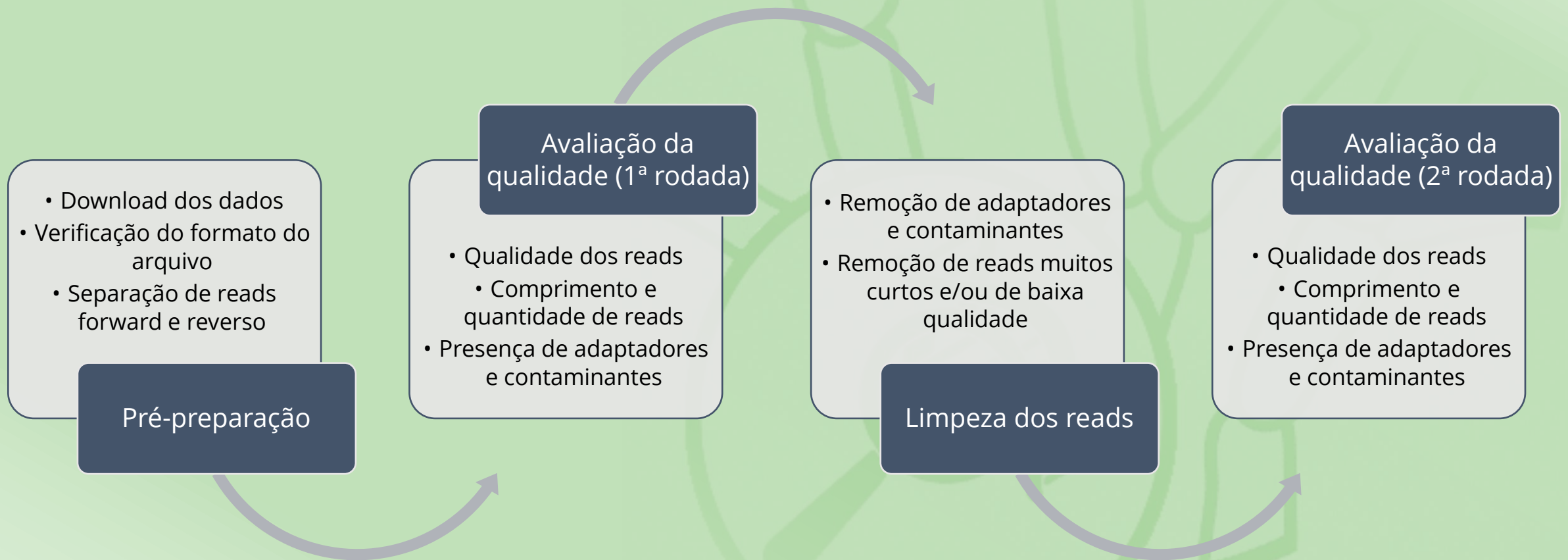
**Linha 01:** começa com um @ e contém o identificador da sequência (similar à primeira linha do formato FASTA)

**Linha 02:** sequência em nucleotídeos

**Linha 03:** começa com um + e pode conter o identificar da sequência novamente

**Linha 04:** contém os valores de qualidade para a sequência na linha 02. Mesmo número de caracteres que a linha 02 (cada símbolo é correspondente à uma letra)

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger											
Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 \$	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41 )	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			





# E se for utilizar um genoma/transcriptoma já pronto?

- Observar seção de materiais e métodos na publicação associada ao genoma ou transcriptoma e confirmar se o controle de qualidade foi realizado antes da montagem

For *P. capitalensis* LGMF01 and *P. citricarpa* LGMF06, libraries of the paired-end reads were processed with NxTrim (O'Connell et al., 2015) to remove Nextera adapters and generate mate-pair, paired-end, single-end and unknown libraries. The script “deinterleave\_fastq.sh” (<https://gist.github.com/nathanhaigh/3521724>) was used to separate the reads from the mate-pair and paired-end libraries in “forward” and “reverse” files. For quality filtering, Trimmomatic v 0.38 (Bolger et al., 2014) was used to (1) trim bases at the start and end of the reads below a quality threshold of 25, (2) trim low quality segments using a 4 bp sliding window and a quality threshold of 15, and (3) discard reads shorter than 50 bp. For *P. citribraziliensis* LGMF08, the reads were filtered in Trimmomatic to remove Illumina adapters and trim for quality using the same parameters described before but discarding reads shorter than 90 bp.

FastQC v 0.11.8 (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) was employed to check the quality of reads after processing in Trimmomatic. De novo genome assemblies were generated with SPAdes v 3.13 (Bankevich et al., 2012), using default parameters. Contigs smaller than 500 bp were filtered and removed from the final assemblies, which were then evaluated with QUAST v 4.6.3 (Gurevich et al., 2013). Library and assembly statistics for the new assemblies are summarized in Table S1 and assemblies are available at Zenodo (<https://doi.org/10.5281/zenodo.3750350>).



# Reads confiáveis vs. montagem de genoma confiável

- Boa qualidade e confiabilidade de reads não está diretamente relacionada à boa qualidade do genoma final
- Necessidade de garantir que o processo de montagem foi adequado
- Diferentes abordagens de montagem para atender à diferentes necessidades

# Por que é necessário “montar” um genoma?

- Cenário ideal: sequenciar o genoma inteiro ou o maior tamanho de fragmento possível
- Condições reais: mesmo técnicas mais recentes como PacBio e ONT que sequenciam reads longos não são capazes de sequenciar cromossomos grandes inteiros
- Necessidade de utilizar os fragmentos obtidos para obter o genoma completo

# Abordagens para montagem de genomas

- ***De novo***

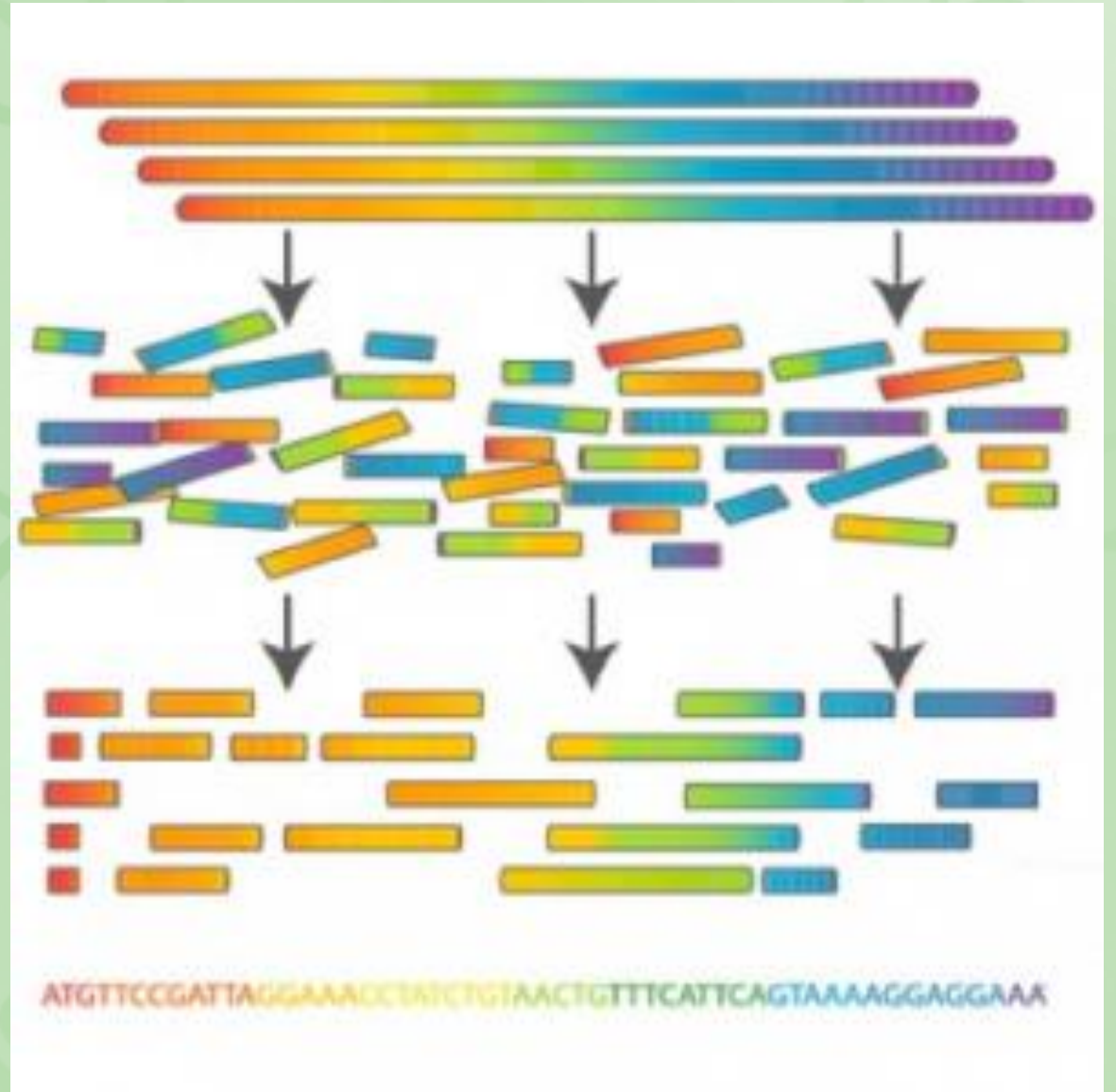
- Reconstruir a sequência completa “do zero”, sem utilizar outro genoma como referência

- **Baseado em referência**

- Alinhamento dos reads à um genoma de referência já montado, e partir dos alinhamentos construir os contigs
- Processo mais simples que uma montagem de novo
- Possibilidade de detecção de alguns tipos de variantes, porém pode mascarar grandes rearranjos estruturais

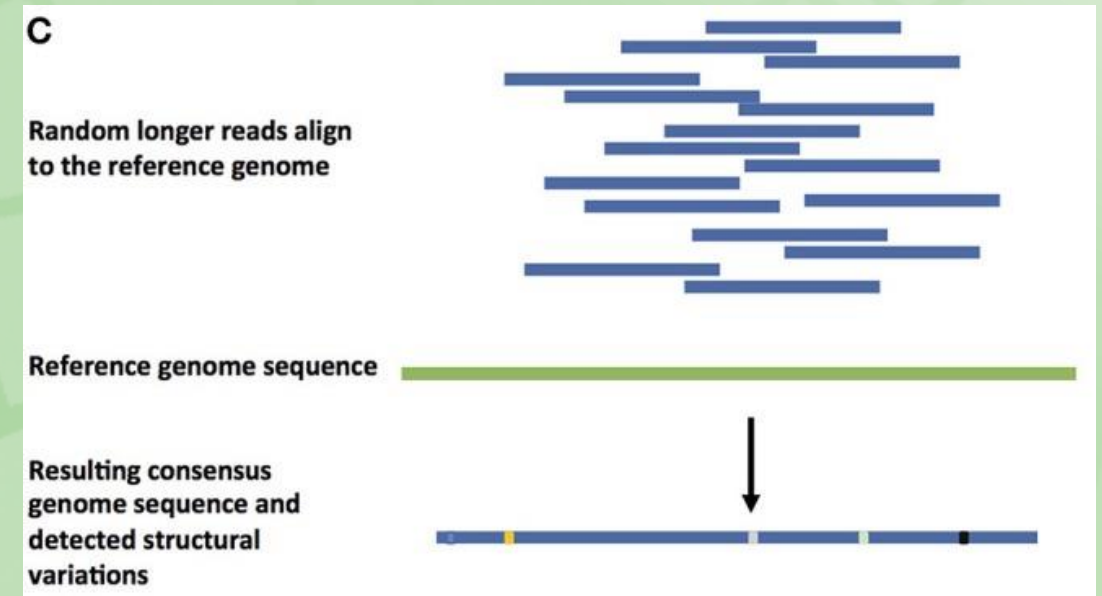
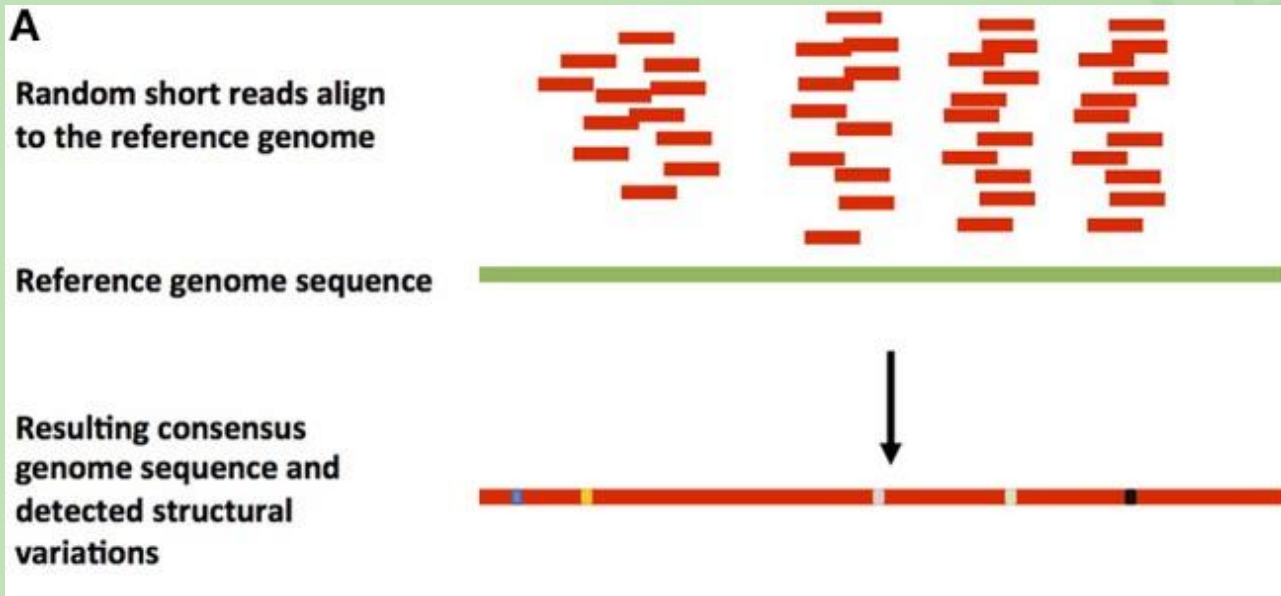
# Montagem de novo

- Utilização dos reads (fragmentos) e informações sobre regiões de sobreposição para produzir sequências únicas e contínuas (contigs)





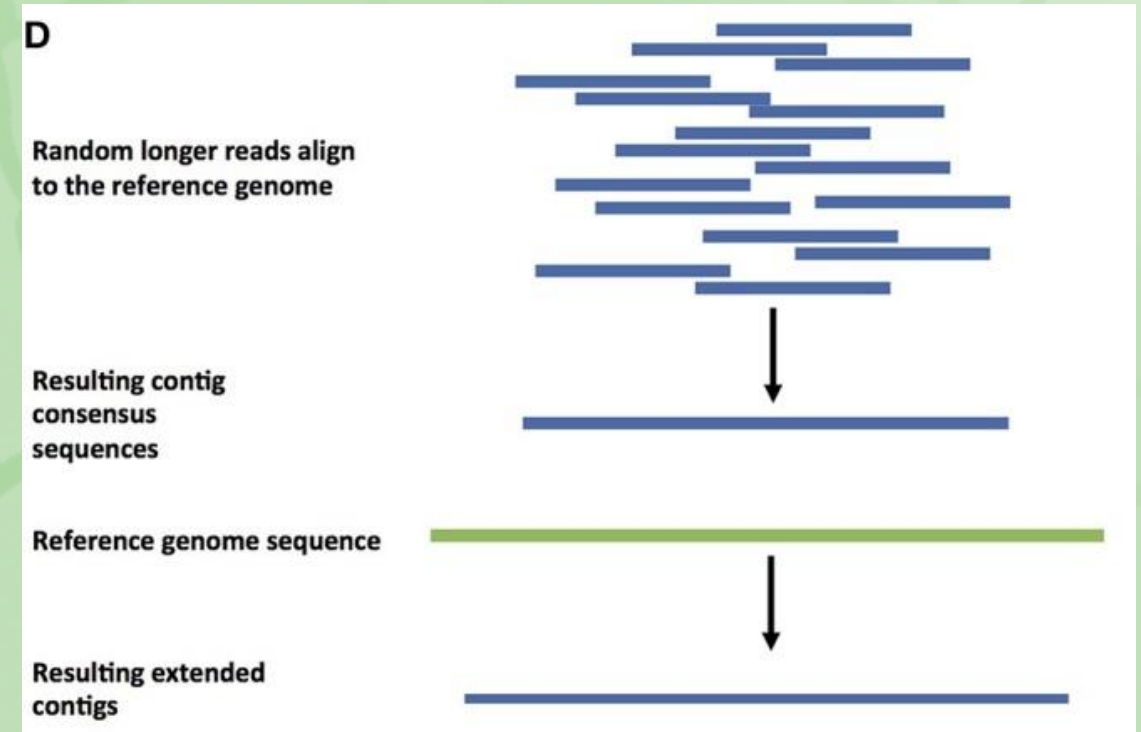
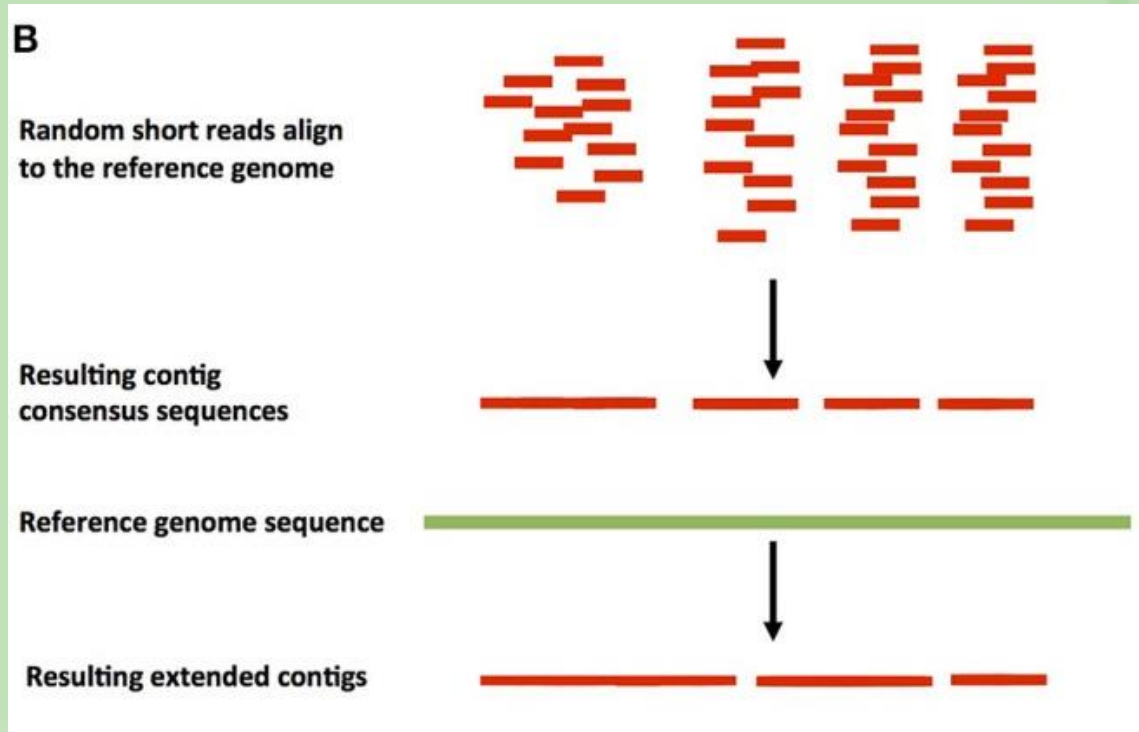
# Montagem guiada por genoma de referência



Adaptado de:  
KYRIAKIDOU et al. 2018. *Frontiers in Plant Science*. DOI: [10.3389/fpls.2018.01660](https://doi.org/10.3389/fpls.2018.01660)

- Alinhamento dos reads à um genoma de referência já montado, e partir dos alinhamentos construir os contigs
- Detecção de variações pontuais, como substituições ou rearranjos mais simples

# Montagem *de novo* guiada por genoma de referência

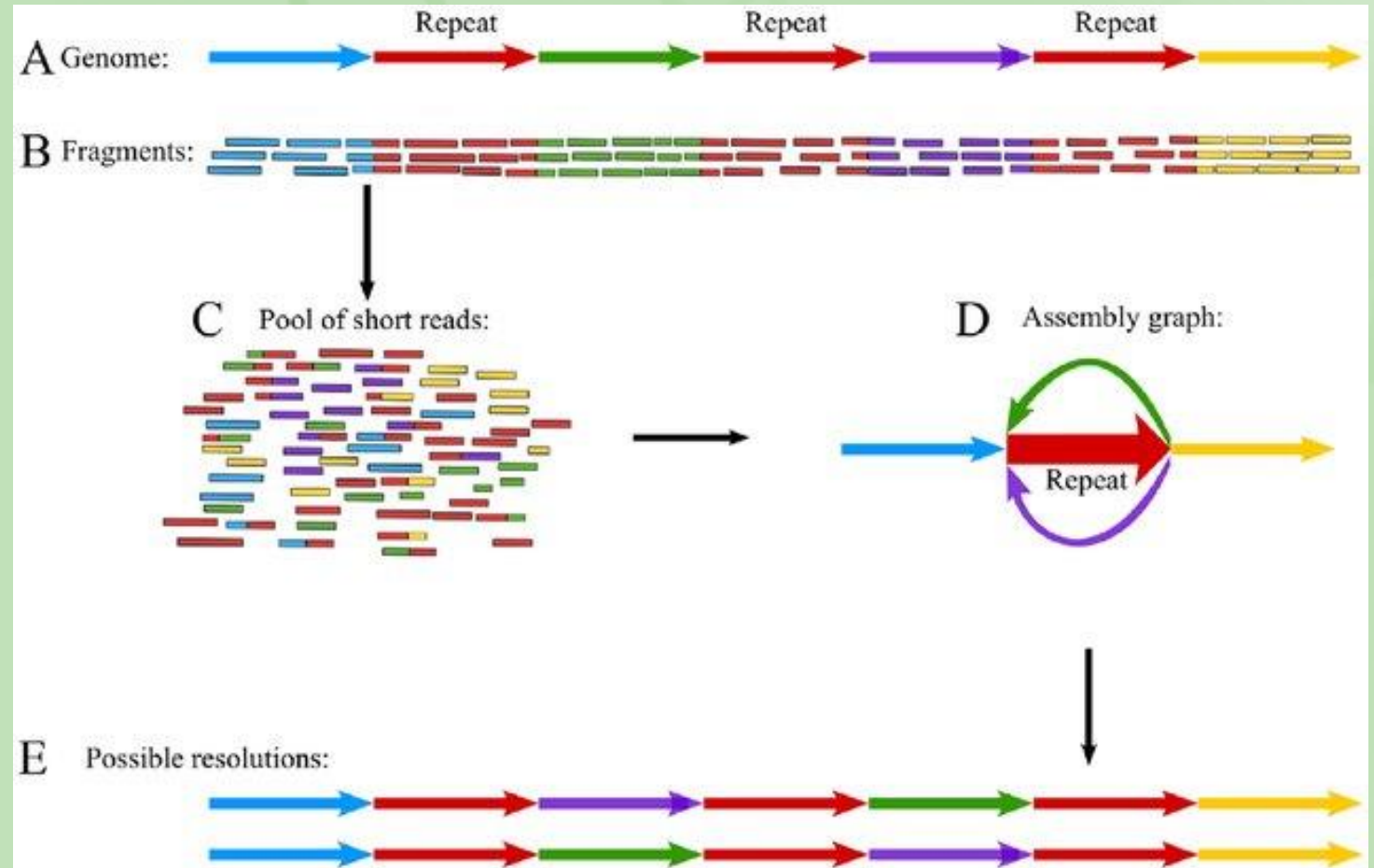


- Montagem inicial dos reads gerando contigs iniciais
- Alinhamento dos contigs à um genoma de referência já montado, e partir dos alinhamentos extender os contigs iniciais em contigs maiores
- Detecção de variações pontuais, como substituições ou rearranjos mais simples

Adaptado de:  
KYRIAKIDOU et al. 2018. *Frontiers in Plant Science*. DOI: [10.3389/fpls.2018.01660](https://doi.org/10.3389/fpls.2018.01660)

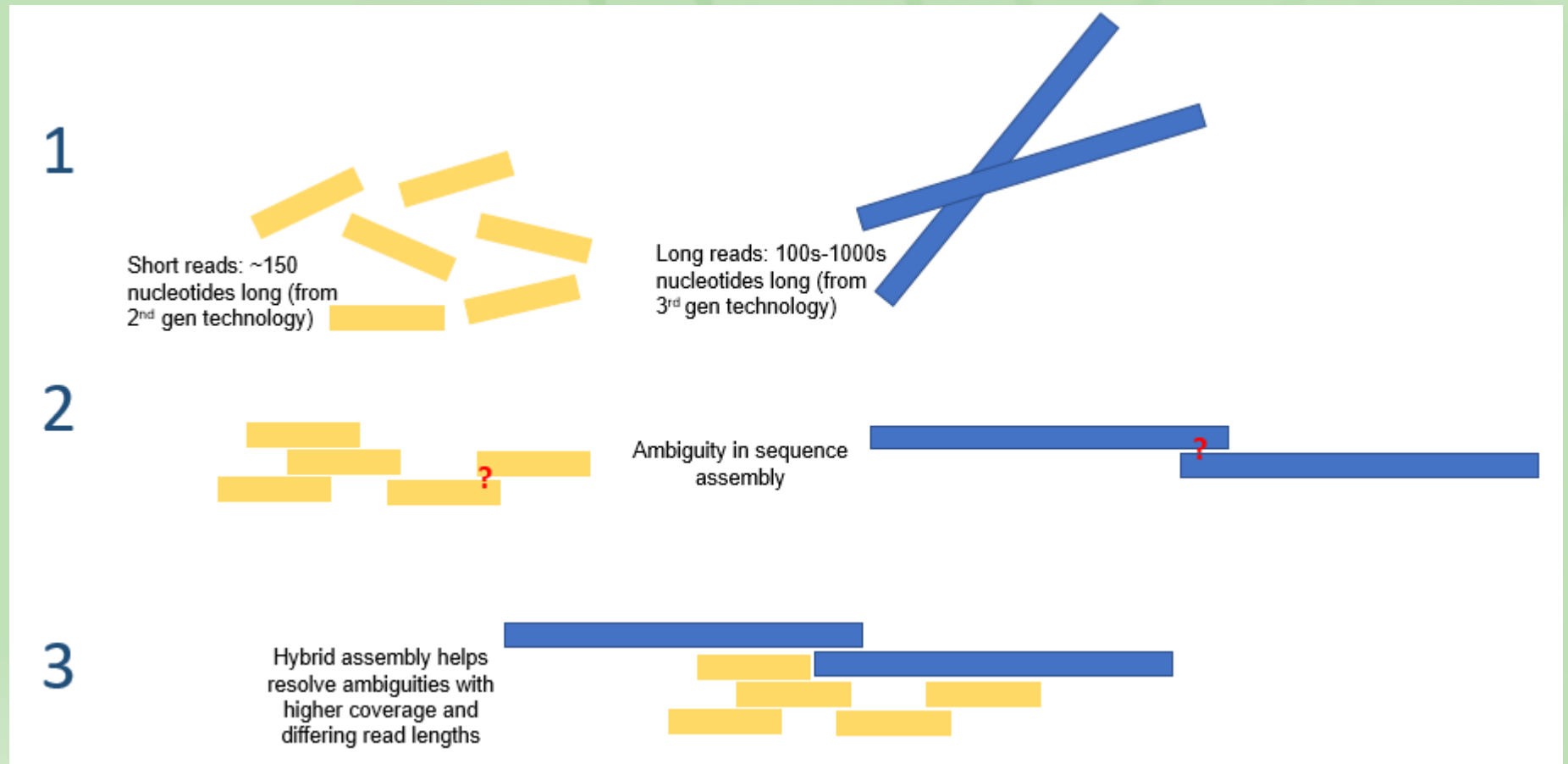
# Problemas de montagem de regiões repetitivas

- Regiões repetitivas mais longas que o tamanho dos reads: ausência de informação sobre as regiões adjacentes para posicionamento correto durante a montagem



# Montagem híbrida (reads longos e reads curtos)

- Reads longos para organizar o genoma em maior escala
- Reads curtos para corrigir erros pontuais e aumentar a confiabilidade de cada base





# Genomas (Formato FASTA)

- **Linha 1:** identificador da sequência após o sinal de maior (>)
- **Linha 2:** sequência

Em geral são arquivos longos e pesados, exigindo o uso de softwares para processar o arquivo completo e obter a informação de interesse

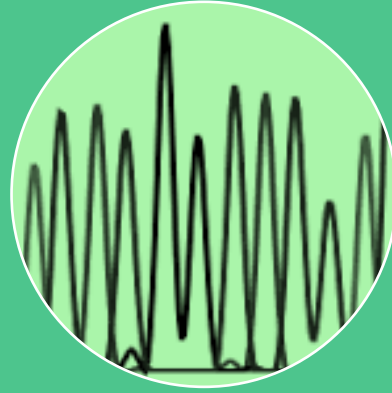
```
1 >scaffold_1
2 CCATGGCTGTCTTGCGATTGTCCAGGGCAGTCTTGACAGCAGGGGCAAGTTGCGCCGCCGCCCGTCCCTT
3 CTCAGTGTCTTCGAAGTTGAGGGAGACGATGACCCTGGTGTTGATGGGACTGTTGGTGTTCGCCGTGGAA
4 GCTTCGTCCTTCTTGCGCTTGGAGCCGGCCGACGCCGCCCTTCTTGCGCTTGGCAATCTCCTCGGGATGCG
5 TGAGAATCTCTTCAATTTTTGCCATGAAGGCGTTCTCTTTCTCGATTTACGAGCGAACTTCCTCGTAGAA
6 TGCCGTGGCTACGCTTGACTCGGTCTTGAAGATGTTGTGGAGAGCCTTGCGGGTGGTCGTTACGACGAA
7 GATGCAGAGAGGGCGCGGTCTGTGGTTCTGCGCGATGGCGTTGCGGGTGTCTTCGTCTTGTTGAACCAGT
8 TGCCGAACCTGAATTACGTCGTCTTTCTTTGCCATCTTTTCCTCGGAGCTCATCGCTTCGATGGTGGCGGC
9 GTCATCCTTGCGCTTTTCGGCGGTCTCGTTTTTCTTCGTCTGGGTGACTTGCAGAAGTGCCTTTGCCCTC
10 AAAGCACTCATTCGGCGACTCTCCTGTTCCGCATCGACGACCTGGCGCCATTCCTGTGACGCATCGCTCA
```

# Como avaliar uma montagem?



## Contiguidade

- N50
- L50
- Quantidade de contigs/scaffolds
- Tamanho do maior contig/scaffold



## Análise de bases

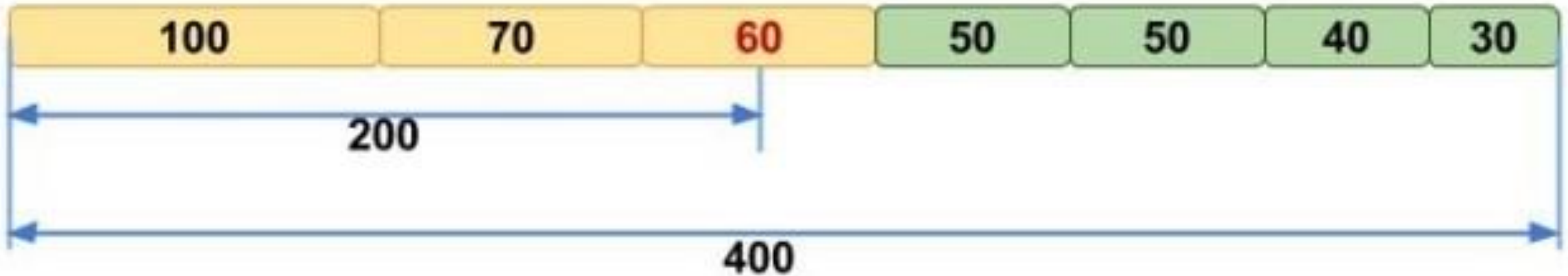
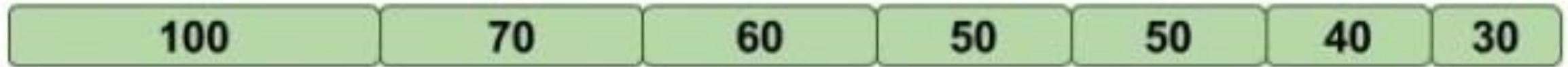
- Cobertura
- Conteúdo GC



## Análise de conteúdo

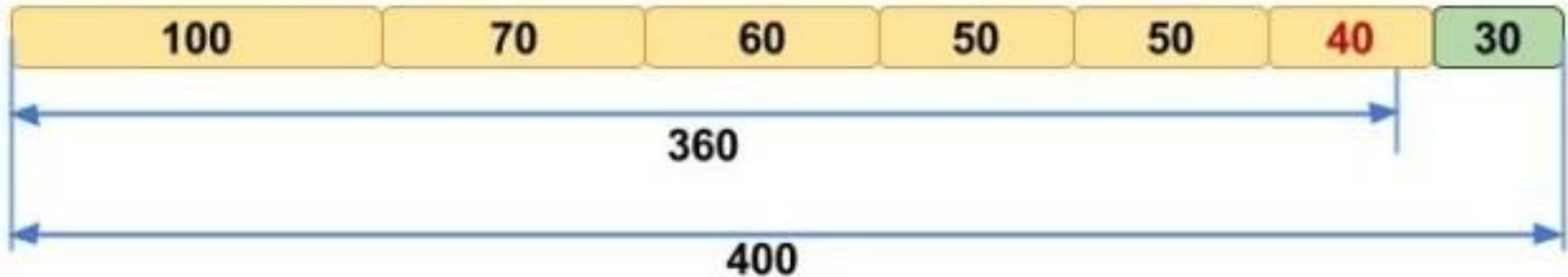
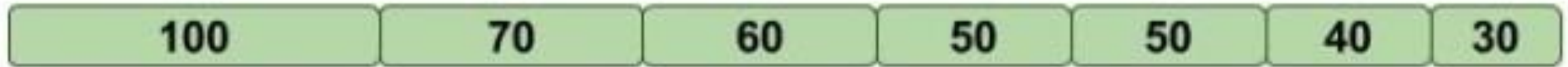
- Presença de telômeros
- Presença de genes conservados
- Comparação com genoma de referência
- Detecção de contaminantes pela distribuição do conteúdo GC
- Detecção de contaminantes por similaridade de sequência

# Contiguidade – N50



- N50: metade da montagem (50%) é representada por contigs/scaffolds com um comprimento igual ou maior que 60Kb

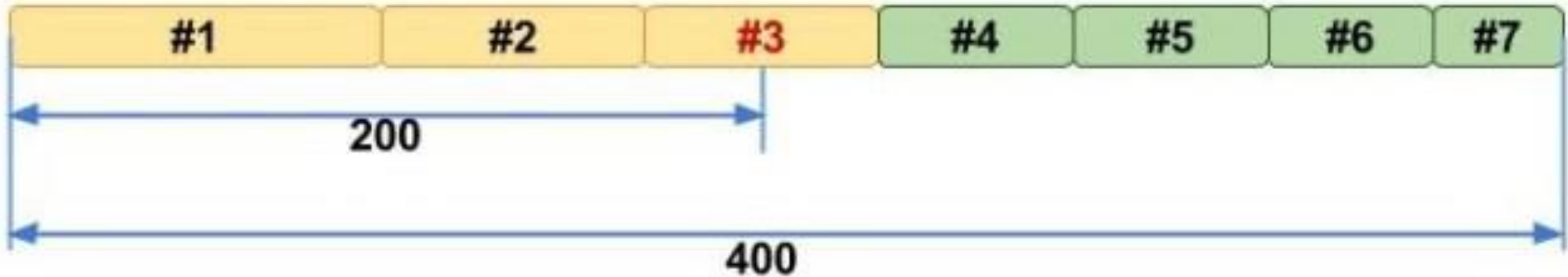
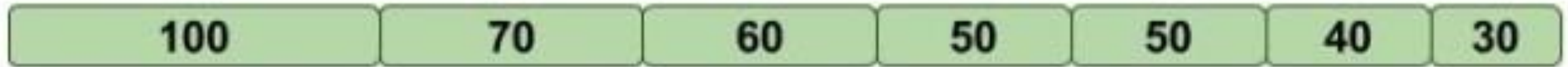
# Contiguidade – N90



- N90: 90% da montagem é representada por contigs/scaffolds com um comprimento igual ou maior que 40Kb



# Contiguidade – L50



- L50: metade da montagem está presente em 3 contigs/scaffolds

# Análise de bases - cobertura

- A cobertura se refere à quantidade de vezes que o genoma foi sequenciado
- Alta cobertura: maior precisão e redução de erros nas montagens
- $Cobertura = \frac{Tamanho\ dos\ reads \times quantidade\ de\ reads}{Tamanho\ total\ do\ genoma}$

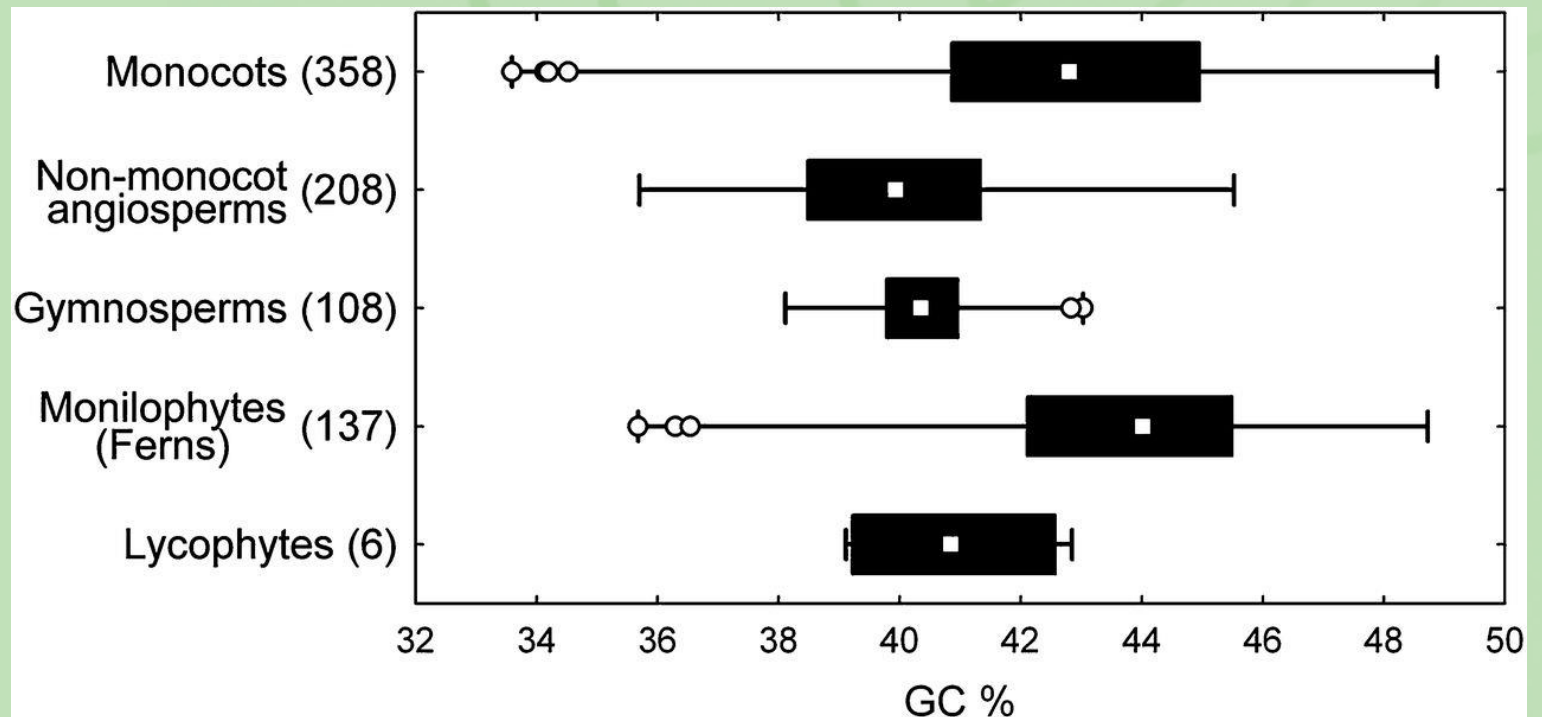
# Análise de bases - cobertura

- Também é possível calcular a cobertura de alinhamento, re-alinhando os reads originais à montagem
- Há muitos reads que não foram alinhados?
- Há regiões da montagem com poucos reads alinhados em relação às outras?



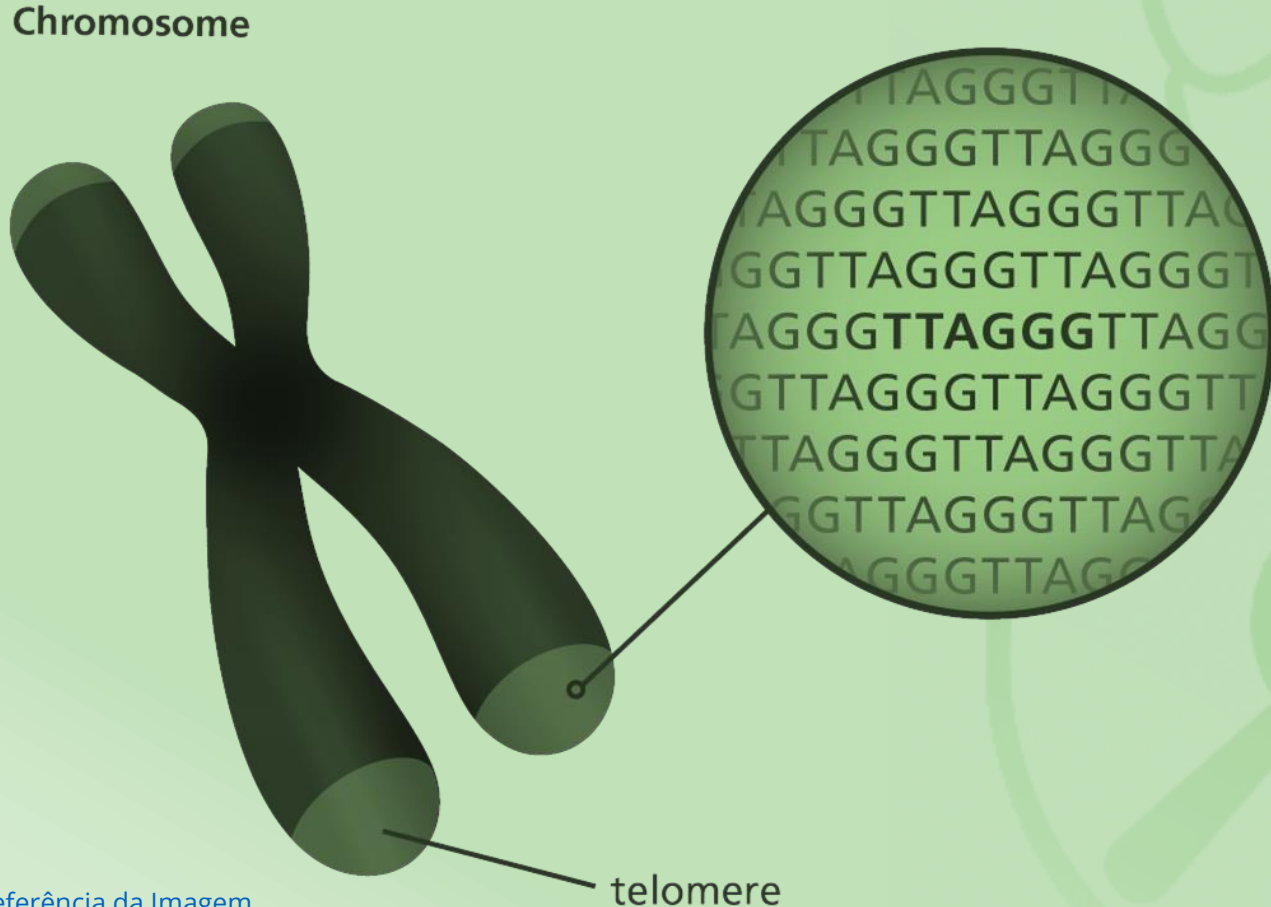
# Análise de bases – Conteúdo GC

- O conteúdo GC da montagem é similar ao conteúdo GC observado para outras linhagens da mesma espécie ou espécies próximas?





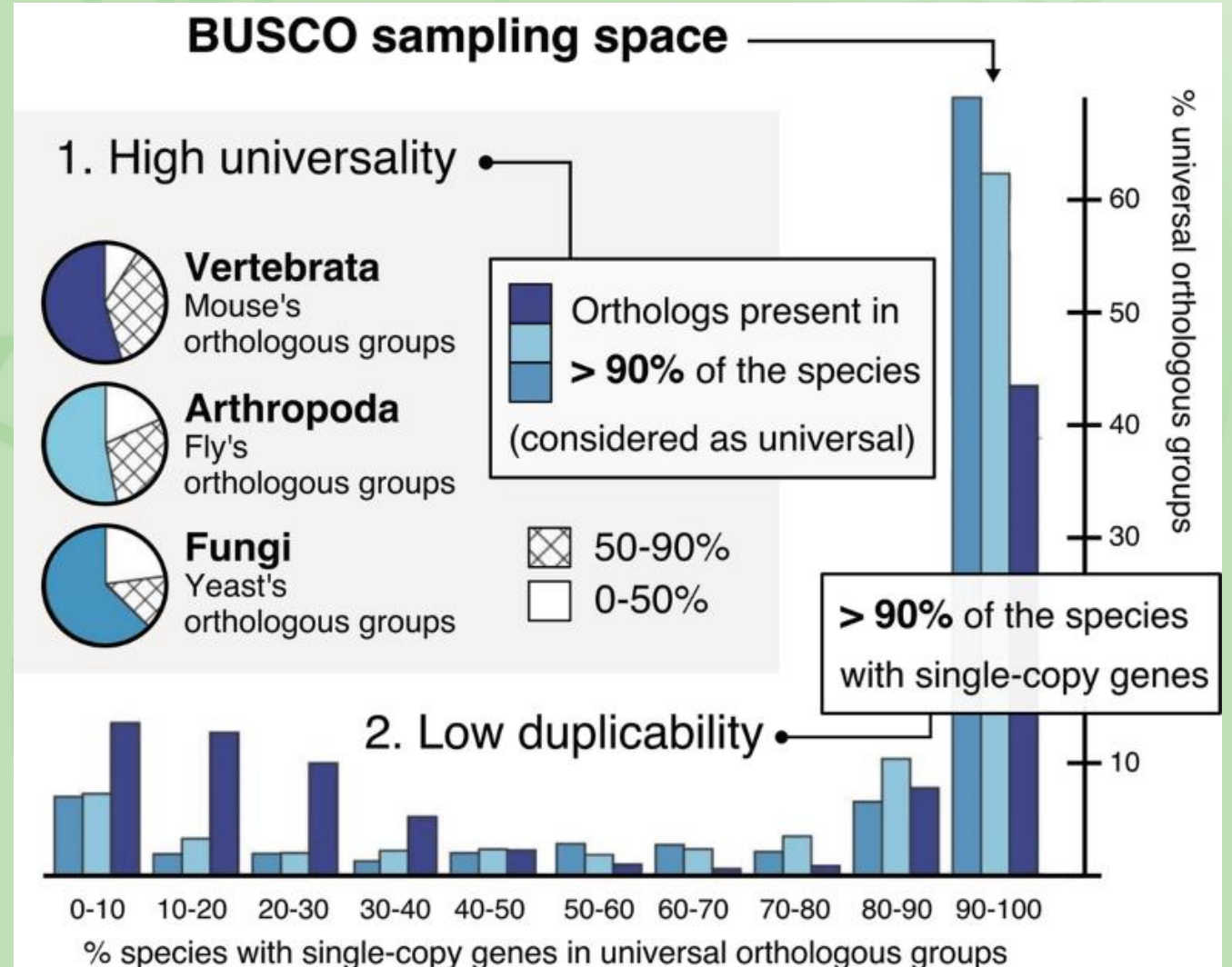
# Análise de conteúdo - Telômeros



- Sequências repetitivas encontradas nas pontas dos cromossomos
- Função protetiva:
  - Impedem que os cromossomos se fusionem nas extremidades
  - Evitam que as sequências de DNA dos cromossomos sejam perdidas (os cromossomos perdem cerca de 25-200 bases por replicação)
- Presença de telômeros no início e fim de um contig/scaffold sugere que se trata de um cromossomo completo

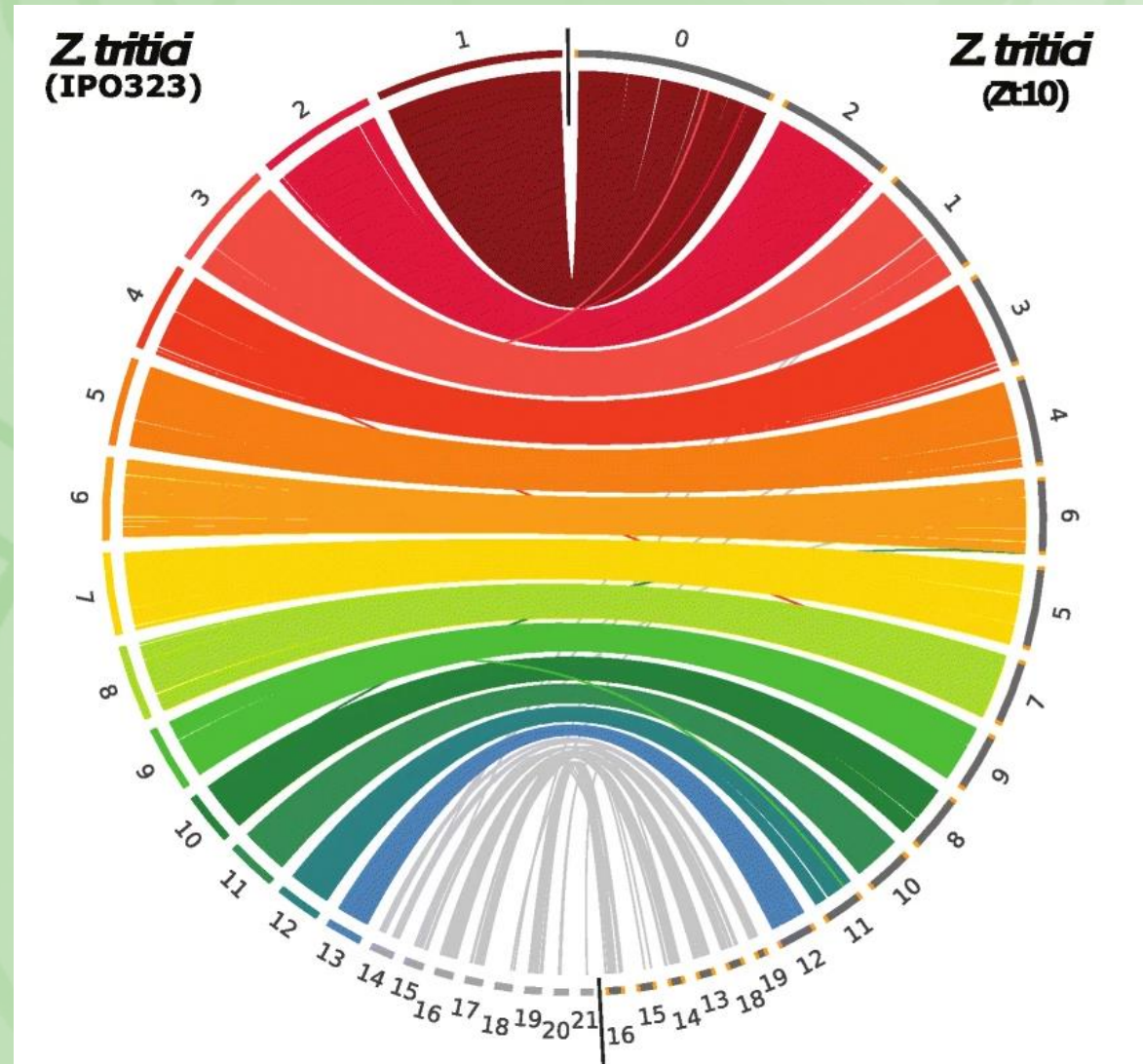
# Análise de conteúdo – Genes conservados

- Avaliação do conteúdo gênico que seria o mínimo esperado em uma montagem ao considerar as relações evolutivas entre os organismos
- BUSCO (Benchmarking Universal Single-Copy Orthologs), <http://busco.ezlab.org/>
  - Alta universalidade: Presente em 90% das espécies do grupo analisado
  - Baixa duplicabilidade: presente em cópia única em 90% das espécies do grupo analisado



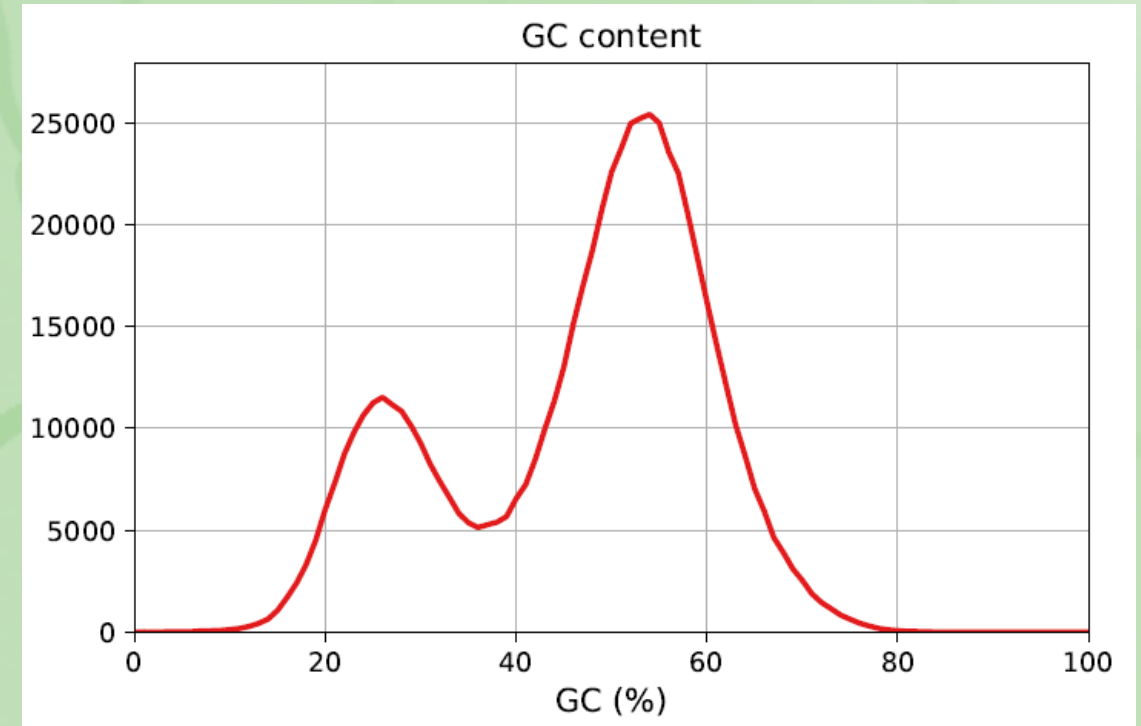
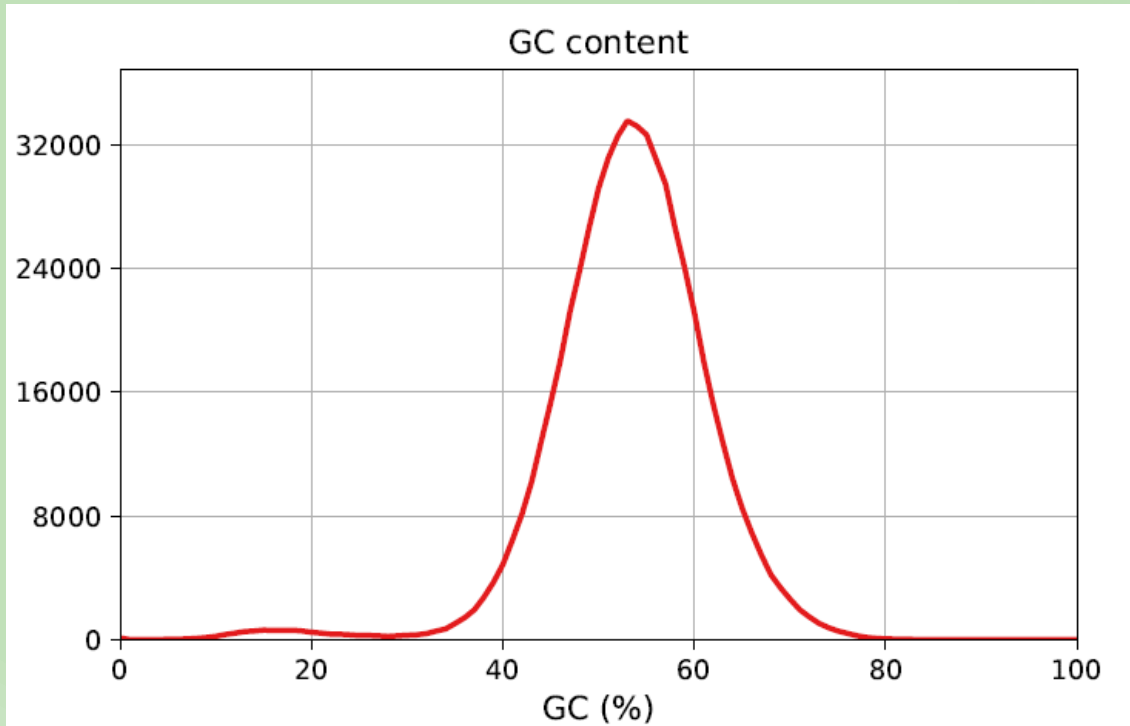
# Análise de conteúdo – Comparação com genoma de referência

- Genes essenciais e conservados presentes na linhagem de referência estão presentes na nova montagem?
- A organização da nova montagem é similar à montagem de referência?





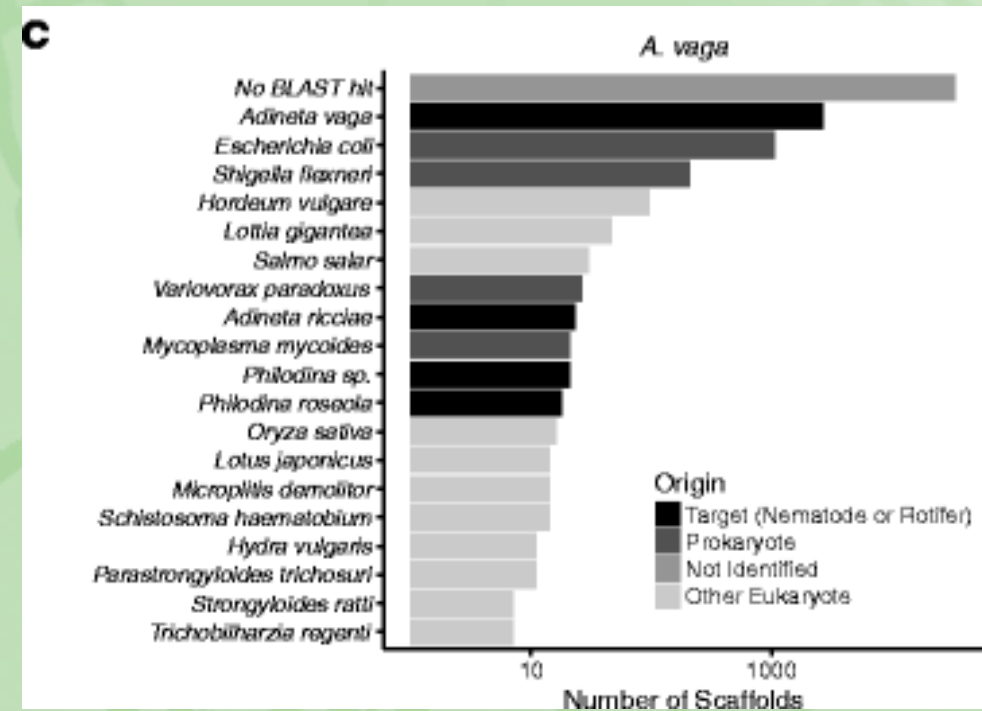
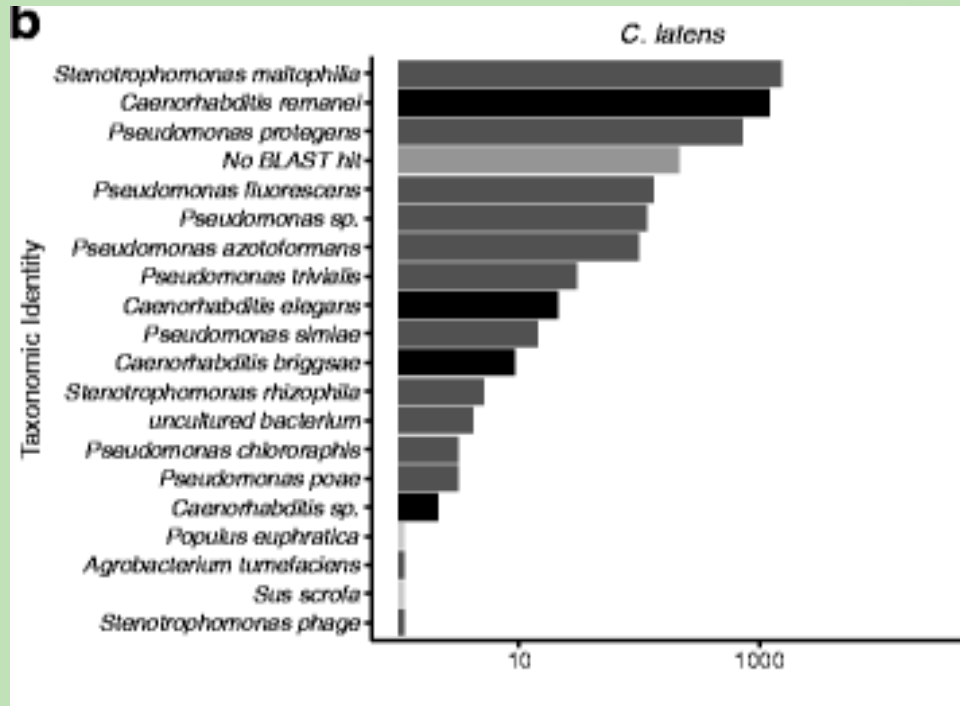
# Análise de conteúdo – Contaminantes (Distribuição do conteúdo GC)



- Quantos picos são observados na distribuição de conteúdo GC?
- Mitocôndria, sequências repetitivas ou contaminação?

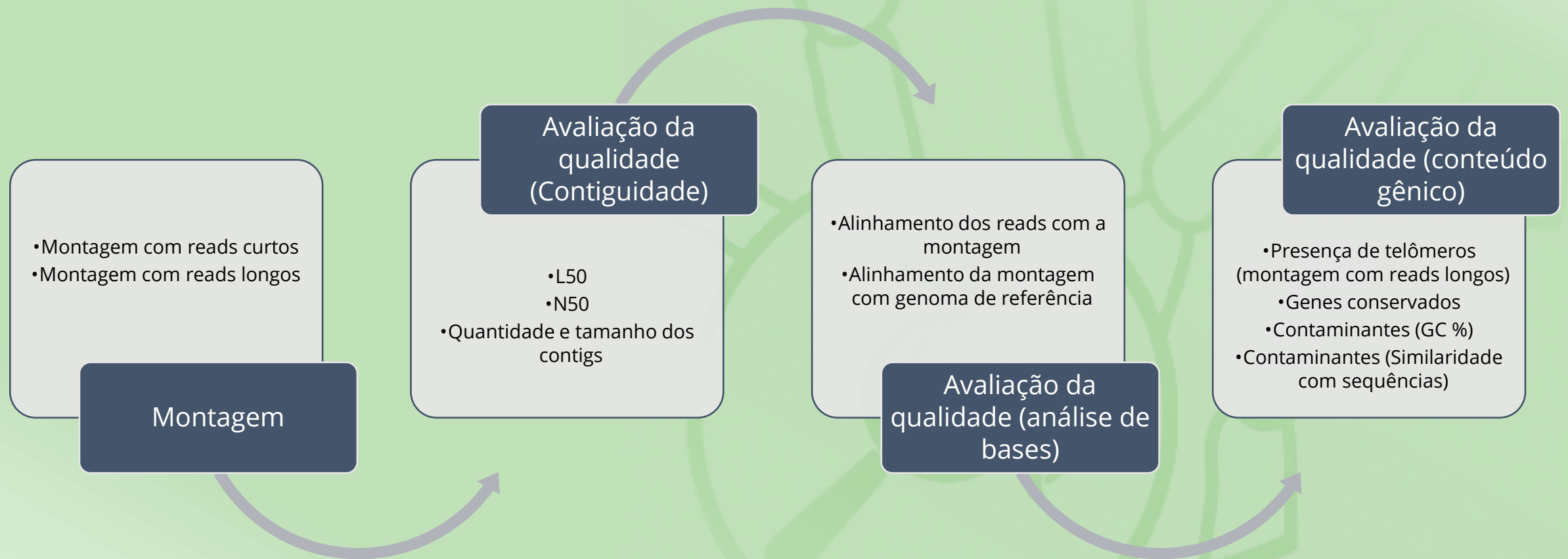


# Análise de conteúdo - Contaminantes



Adaptado de:  
FIERST et al. 2017. *BMC Bioinformatics*. DOI: [10.1186/s12859-017-1941-0](https://doi.org/10.1186/s12859-017-1941-0)

- Há sequências de outros organismos na montagem?
- Uso do BLAST (sequência completa) ou Kraken (k-mers)



# Anotação e detecção de sequências de interesse

- Sequência em si sem anotações e informações associadas: baixa aplicabilidade em abordagens práticas e funcionais
- Busca por padrões e características que identifiquem genes e regiões de interesse em uma sequência



# Como avaliar a qualidade de uma anotação?

- Avaliação de conteúdo (BUSCO)
- Comparação com anotações prévias confiáveis
  - A quantidade de genes é similar?
  - Anotações funcionais resultam em informações similares?
- Avaliação manual



# Formato GFF3 (General Feature Format)

- Uma feature por linha, 9 colunas delimitadas por tabulações
- **1 (seqid):** nome do cromossomo, contig ou scaffold em que a feature está localizada
- **2 (source):** nome do programa que gerou a anotação, ou da base de dados em que a anotação foi obtida
- **3 (type):** categoria da feature (ex: gene, CDS, mRNA, exon)
- **4 (start):** posição do início da feature no cromossomo, contig ou scaffold
- **5 (end):** posição do final da feature no cromossomo, contig ou scaffold
- **6 (score):** score de confiabilidade, mas muitos softwares não atribuem nenhum valor (.)
- **7 (strand):** indica se a feature está na fita direta/forward (+) ou reversa/reverse (-)
- **8 (phase):** fase de leitura
- **9 (attributes):** informações adicionais sobre a feature, como nome, ou vínculo com outra feature anterior

```
scaffold_10 EVM gene      2697      4438      .      +      .      ID=scaffold_10.1;Name=scaffold_10.1
scaffold_10 EVM mRNA      2697      4438      .      +      .      ID=scaffold_10.1;Parent=scaffold_10.1;Name=scaffold_10.1
scaffold_10 EVM exon      2697      3078      .      +      .      ID=scaffold_10.1.exon1;Parent=scaffold_10.1
scaffold_10 EVM CDS 2697      3078      .      +      0      ID=cds.scaffold_10.1;Parent=scaffold_10.1
scaffold_10 EVM exon      3145      4023      .      +      .      ID=scaffold_10.1.exon2;Parent=scaffold_10.1
scaffold_10 EVM CDS 3145      4023      .      +      2      ID=cds.scaffold_10.1;Parent=scaffold_10.1
scaffold_10 EVM exon      4068      4438      .      +      .      ID=scaffold_10.1.exon3;Parent=scaffold_10.1
scaffold_10 EVM CDS 4068      4438      .      +      2      ID=cds.scaffold_10.1;Parent=scaffold_10.1
```

# Regiões codificantes (CDS) em formato FASTA

- **Linha 1:** identificado r da sequência após o sinal de maior (>)
- **Linha 2:** sequência
- Cada sequência corresponde a um gene

```
1 >scaffold_10.1
2 ATGACGGACCGCTCTCGACGCTGCTCTTACTCCAACGCCATTTTGCTGGTGCTGGGCCTCGCCGGGC
3 TGGCATACATCAGCTTCCGCGCCGCGTACGGCACCGACGTGGGGCGCATCACGGGCATTCTTGAGCCGGG
4 GCACGCGGTGGCGTTCTACGGACACCTCAACTCCAAGGCGCTCGGCAGCGACCAACCCCACTGCGCTGCAG
5 GAGTATTCGGTGAAGAATGGGTGGCCGTTGGTGCAGGTGCGGTTTGGGCAGCGGCGGGTCGTGGTGCTGA
6 ATACGTTTGCGGCGGCGCAGCATTTTCATCATTCGCAATGGGGGGGCGACAATTGACCGCCCGCTGTTTTG
7 GACATTTTCATAAGTTTGTGAGCAATACGCAGGGCGCAACCATTGGCACGTGCGCGTGGGACGCATCGTGC
8 AAGCGCAAGCGCACGGCAATCGGGGCGTACATGACGCGGCCGGCCATCCAGCGCAATGCGCCACTCATCG
9 ACATCGAGGCGCTGGGGCTGGTCGAAGGCATCTTTAACGCCTCGCTGGACGACAACAACAACCCCAAGTGT
10 CGAAGTGGACCCCCGCTCTTTTTCCAGCGGGCTTCGCTCAACTTTGTGCTCATGCTCTGCTACGCGTCG
11 CGGTTCCCGGACATTGACGACCCGCTGCTGCACGAGATTCTGGCCACGGGCAAGACGGTCAGCACGTTTC
12 GCAGCACCAACAACAACATGGCCGACTACGTGCCGCTGCTGCGGTACCTGCCCAACGCGCGGACGGCGAT
13 GGCCAAGCAGGTGACCAAGAAGCGCGACGTGTGGCTCGAGGCGCTGCTGGAGCGCGTGCGCAAAGCCGTG
14 GCGGCCGGCAAGCCCGTGTCTGTCATTGCATCGTCGCTGCTCAAGGAAAAGGGGTCCGAGAAGCTGACAG
15 AGGCCGAGATTCGCTCCATCAACGTCGGGCTCGTCTCGGGCGGCAGCGACACGATTGCGACGACGGGGCT
16 CGGCGGGCTTGGGTTCTCGCGTCCAAGGAGGGCCAGGCGATTGAGCAAAAGGCGTACGACGAGATTATG
17 AAGGTCTACGCGACGGCCGAGGAGGCGTGGGAGAATTGCGTGCTCGAGGAGAATGTCGAGTACGTGTCG
18 CGCTCGTGCGCGAGATGCTGCGGTACTACTGCGCGATACAGCTGCTGCCACCGCGCAAGACGTGCAAGCC
19 GTTTGAGTGGCATGGCGCACAAATCCCTGCTGGTGTACGGTATACATGAACGCGCAGGCTATCAATCAC
20 GACAAAACCGCATACGGACCAGACGCGCACATTTTCCGACCAGAGCGTTGGCTCGATCCCAGCAGTCCGT
21 ACCAGGTGCGGGCTTCCCTACCACTACTCGTATGGCGCGGGCTCGCGAGCATGCACGGCCGTGGCGCTGTC
22 GAACCGGATTCTCTACTGCTACTTTGTGAGGCTGATTGTTTTCGTTCCGCTTCACGGCCAGCGCAGACGCG
23 CCGCCGACGCTGGATTACATTGGATTCAACGAGAACCCGCAGGCGGCGACGGTCGTCCCAAAGACGTTTC
24 GGGTTAACATTGAGGAGAGGCGGCCGAGGGAGGAGCTGGCCAAGAATTTGAGGCGAGTCGAAAGGCCAC
25 TTCTCACCTCGTCTTTACTTAG
```

# Sequências de aminoácidos em formato FASTA

- **Linha 1:** identificador da sequência após o sinal de maior (>)
- **Linha 2:** sequência
- Cada sequência corresponde a um gene

```
1 >scaffold_10.1
2 MDGPLSTLLSYSNAILLVLGLAGLAYISFRAAYGTDVGRITGIPEPGHAVAFYGHLNSKALGSDHPTALQ
3 EYSVKNGWPLVQVRFGQRRVVVLTFAAAQHFIIRNGGATIDRPLFWTFHKFVSNTQGATIGTSPWDASC
4 KKRRTAIGAYMTRPAIQRNAPLIDIEALGLVEGIFNASLDDNNNPSVEVDPRLFFQRASLNFVLMICYAS
5 RFPDIDDPLLHEILATGKTVSTFRSTNNNMADYVPLLRYLPNARTAMAKQVTKKRDVWLEALLERVRKAV
6 AAGKPVSCIASSLLKEKGSEKLTEAEIRSINVGLVSGGSDTIATTGLGGLGFLASKEGQAIQQKAYDEIM
7 KVYATAEEAWENCVLEENVEYVVALVREMLRYYCAIQLLPPRKTCKPFEWHGAQIPAGVTVYMNAQAINH
8 DKTAYGPDHIFRPERWLDPSQVGLPYHYSYGAGSRCTAVALSNRILYCYFVRLIVSFRFTASADA
9 PPTLDYIGFNENPQAATVVPKTFRVNIEERRPREELAKNFEASRKATSHLVFT
10 >scaffold_10.2
11 MALQTCRRCRKRRKCDLQLPACTSCQLVDLECLYFDDSLGHDVPRSYLHALSKKVENLESTINAIKSPA
12 AAAPSPTPFSQSDCPTPLQASLDPRGSSASSLGLGTSAGLLENLLKTLVQRSSTQDQSALS RFASRTRDV
13 EDDSALAFPPLKVNFSKLDTSLSLQQPHLQRALIEYYAKTVQSSFPLLSKAQIDSLLRYEHPLRQCTAAER
14 LPIYGIFALASNLVSRDLDDKQDQSIASMTWTERFHSYIAGFDSSNAHGAVRMKQNILALCFLALLDLVSPL
15 SPKGGVWEVVGAASRSYVKVLDLSDVSSPEIDDEFERLGHCIYLLESTLSIHFRIPSLYCNSAPTVIPSG
16 LSEPLVYHTLYTLTQLLNFPKDVSVDMESSIPACLRINLESGPSDVSLGQAQVYLTTLHPLFTSPGAGIHC
17 CSPDLLSKIALAAAFITHTHKLNKERRVVSIVVTAENVLQAGAAWAAYLMLHSQRDSPLHDYHVPKPID
18 KLPPMEPIVRCSSLLASFAERWKGGRRFCQAWAEFTELLADDLSKMATAPQA
```



Anotação  
gênica

```
graph TD; A[Anotação gênica] --> B[Anotação de elementos transponíveis]; B --> C["Anotação funcional:  
- Efetores  
- CAZymes  
- Clusters de metabólitos secundários"]; C --> D["Anotação funcional:  
- Função molecular  
- Localização celular  
- Processo biológico"]; D --> E[Análises comparativas];
```

Anotação de  
elementos  
transponíveis

Anotação funcional:  
- Efetores  
- CAZymes  
- Clusters de metabólitos secundários

Anotação funcional:  
- Função molecular  
- Localização celular  
- Processo biológico

Análises  
comparativas

- Possibilidade de anotações adicionais:
  - Responder à pergunta inicial
    - Testar as hipóteses
  - Sugerir novas perspectivas
  - Fornecer bases para estudos futuros



# Que tipo de genoma eu preciso encontrar?

- As perguntas de um projeto podem ser respondidas com diferentes tipos de “montagem”:
  - Genoma completo
  - Genoma rascunho
  - Somente genes
- Montagens mais completas permitem com que mais perguntas sejam respondidas, mas aumentam os custos envolvidos e o tempo de execução do projeto

# Genoma completo

- Um genoma é considerado completo quando a qualidade da sequência atende aos “Bermuda standards”:
  - Taxa de erro de nucleotídeo: 1 a cada 10.000 bases ou menos para a maior parte da sequência (99,99% de acurácia)
  - Ausência de gaps na montagem
- Muitos genomas eucarióticos apresentam regiões complexas (e. g. regiões repetitivas) que são difíceis de sequenciar e montar: dificilmente um genoma completo (de acordo com a definição) é obtido
- Termos como “*working draft*” e “*essencialmente completo*” são usados para descrever genomas de qualidade superior à genomas rascunho, mas que ainda não são genomas completos


# Genoma rascunho

- Não há um padrão absoluto para reconhecer genomas rascunho (como no caso de genoma completo)
- Em geral, um bom genoma rascunho apresenta sequências com boa qualidade, cobertura e completude, e baixo nível de fragmentação
- Não é incomum encontrar genomas rascunho altamente fragmentados, com má qualidade e incompletos

# Somente genes

- Sequenciamento de RNA mensageiro (regiões transcritas) e produção de transcriptoma
- Custo mais baixo se comparado à genoma completo
- Melhor aproveitado se já existe um genoma de referência para a espécie
- Limitações:
  - Somente genes transcritos na condição avaliada serão sequenciados
  - Ausência de informação sobre regiões e sequências repetitivas, arquitetura genômica, organização cromossômica ou coordenadas dos genes

# Quais perguntas serão respondidas com estes dados?



Genoma essencialmente completo	<ul style="list-style-type: none"><li>• Regiões repetitivas</li><li>• Elementos transponíveis</li><li>• Organização cromossômica</li><li>• Arquitetura genômica</li><li>• Sintonia e rearranjos cromossômicos</li></ul>
Genoma rascunho	<ul style="list-style-type: none"><li>• Anotação gênica</li><li>• Anotação funcional</li><li>• Variantes populacionais (em regiões não codificantes)</li></ul>
Somente genes	<ul style="list-style-type: none"><li>• Estudos funcionais de genes específicos</li><li>• Variantes populacionais (em regiões codificantes)</li></ul>



# Qual a cobertura de sequenciamento necessária?

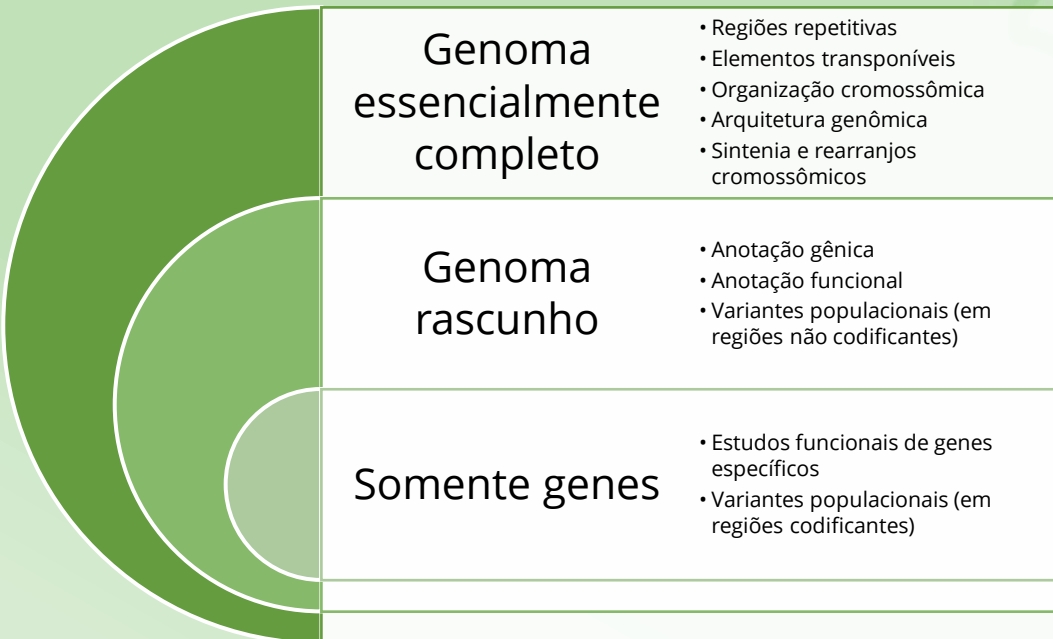
- A cobertura se refere à quantidade de vezes que o genoma foi sequenciado
- Alta cobertura: maior precisão e redução de erros nas montagens
- $Cobertura = \frac{Tamanho\ dos\ reads \times quantidade\ de\ reads}{Tamanho\ total\ do\ genoma}$

# Qual a contiguidade que montagem deve apresentar?

- A contiguidade está relacionada ao grau de fragmentação do genoma. Quanto maior a fragmentação, menor a contiguidade
- Para maior contiguidade, é necessária uma boa cobertura e uma técnica de sequenciamento com reads longos e capaz de resolver regiões repetitivas

# A melhor estratégia de sequenciamento depende do contexto

- **Somente genes e/ou genoma rascunho:** Illumina
- **Genoma essencialmente completo:** PacBio ou Nanopore, ou combinações entre dois ou três métodos



Método	Illumina	PacBio	Nanopore
Comprimento dos reads	100 – 300 pb	10 – 100 kb	Variável (até 1000 kb)
Taxa de erro	0.1%	5 – 15%*	5 – 20%*
Eficiência (bases por corrida)	200 – 600 Gb	10 – 20 Gb	5 – 10 Gb
Prós	Alta confiabilidade e	Reads longos, velocidade e alta eficiência	Reads longos, velocidade e alta eficiência
Contras	Reads curtos, velocidade	Taxa de erro elevada	Taxa de erro elevada

\* Em baixa cobertura

# Onde encontrar reads, montagens e anotações?



RefSeq

