# Montagem de genomas e avaliação de qualidade de montagens

Drª Desirrê Petters-Vandresen

# Por que 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

# Montagem de genomas

**Cópias de DNA do genoma**

**Reads de sequenciamento**

**Montagem do genoma**

The Call–Chronicle–Examiner

EARTHQUAKE AND FIRE: SAN FRANCISCO IN RUINS

NO HOPE LEFT FOR SAFETY OF ANY BUILDINGS

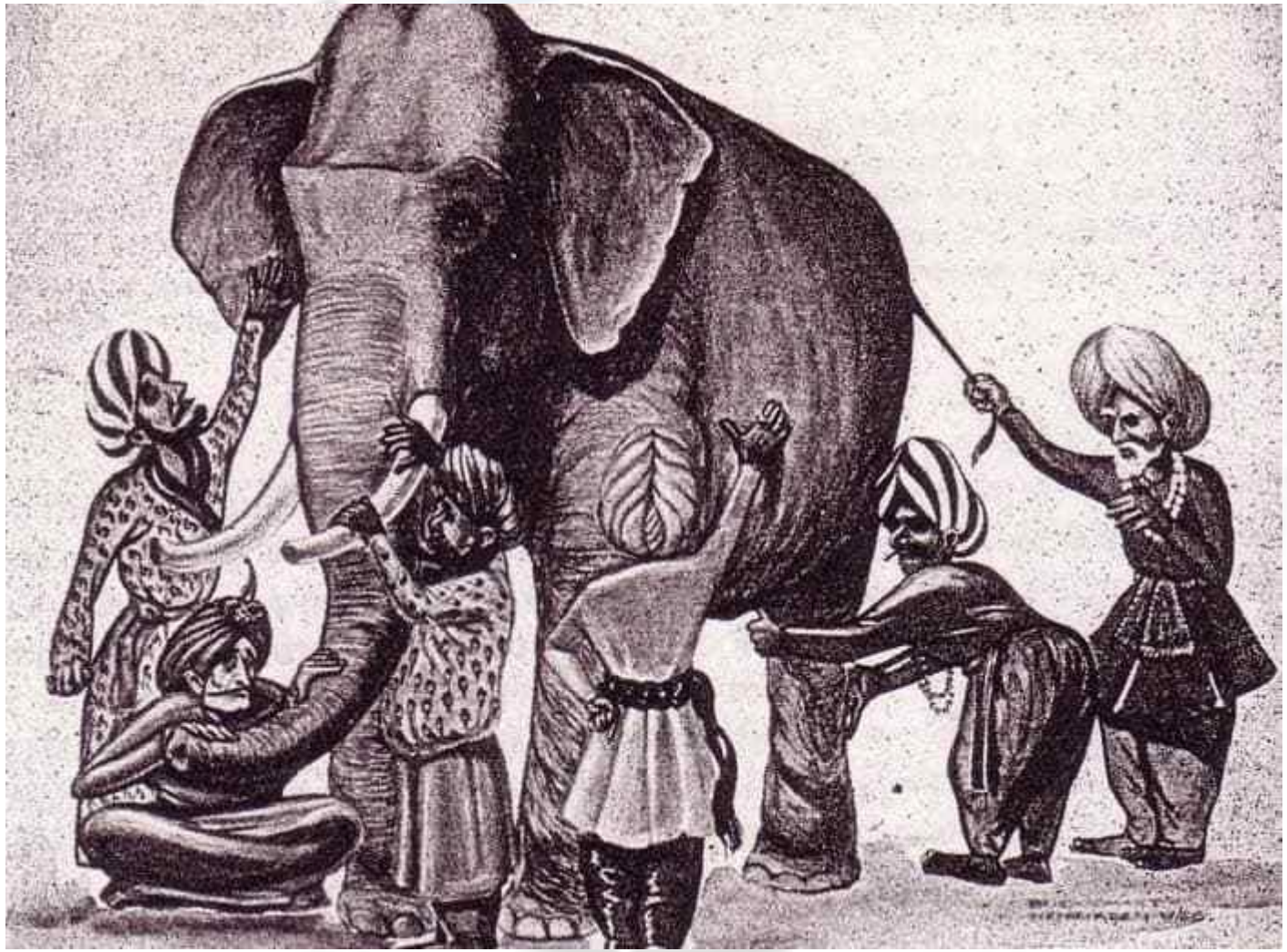
BLOW BUILDINGS UP TO CHECK FLAMES

WHOLE CITY IS ABLAZE

CHURCH OF SAINT IGNATIUS IS DESTROYED

MAYOR CONFERS WITH MILITARY AND CITIZENS

- **Como organizar todos os fragmentos obtidos no sequenciamento na ordem biológica correta e formando uma sequência única e coesa?**

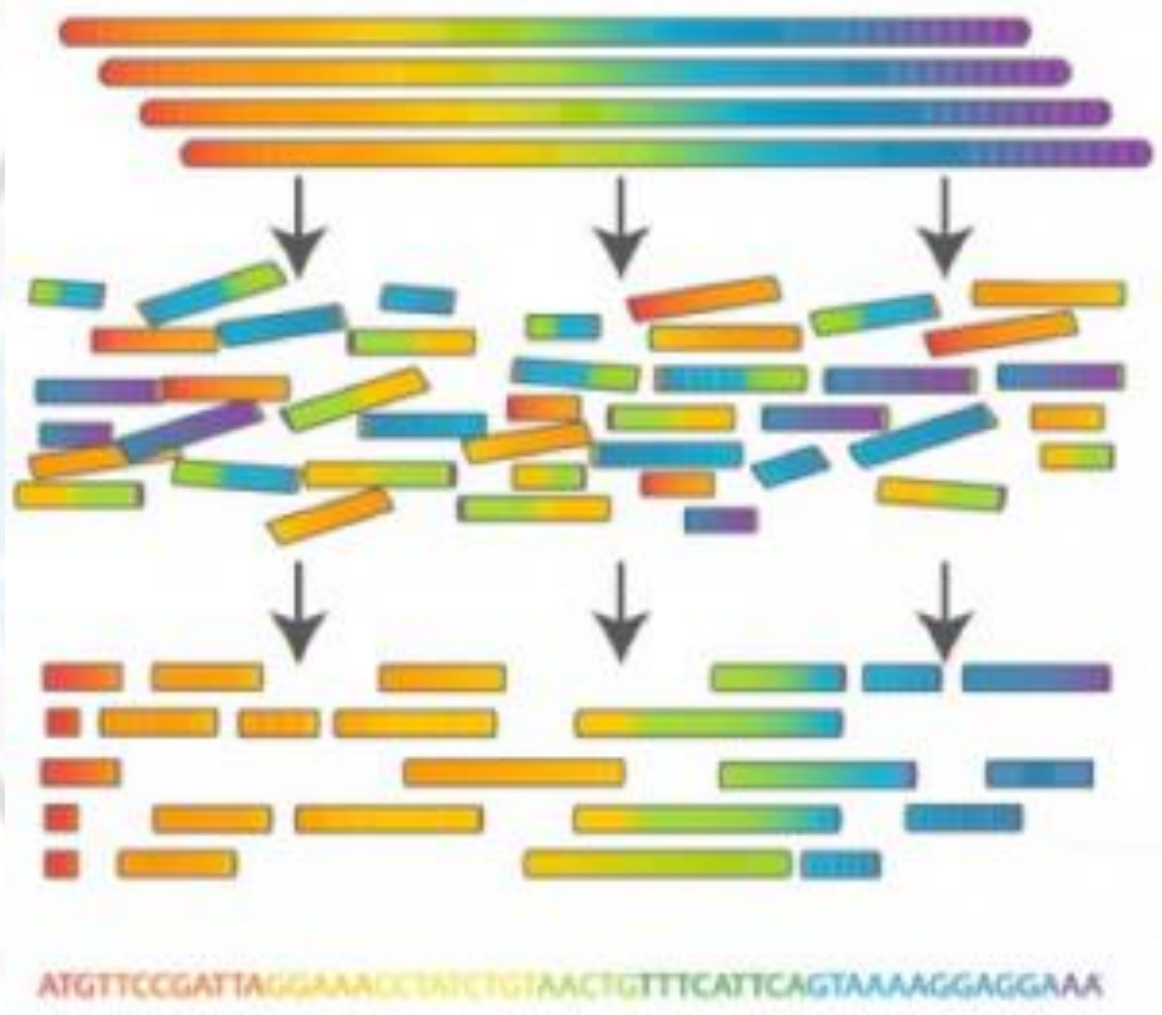


Referência da Imagem

# Montagem de genomas

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

- Diferentes algoritmos e estratégias possíveis

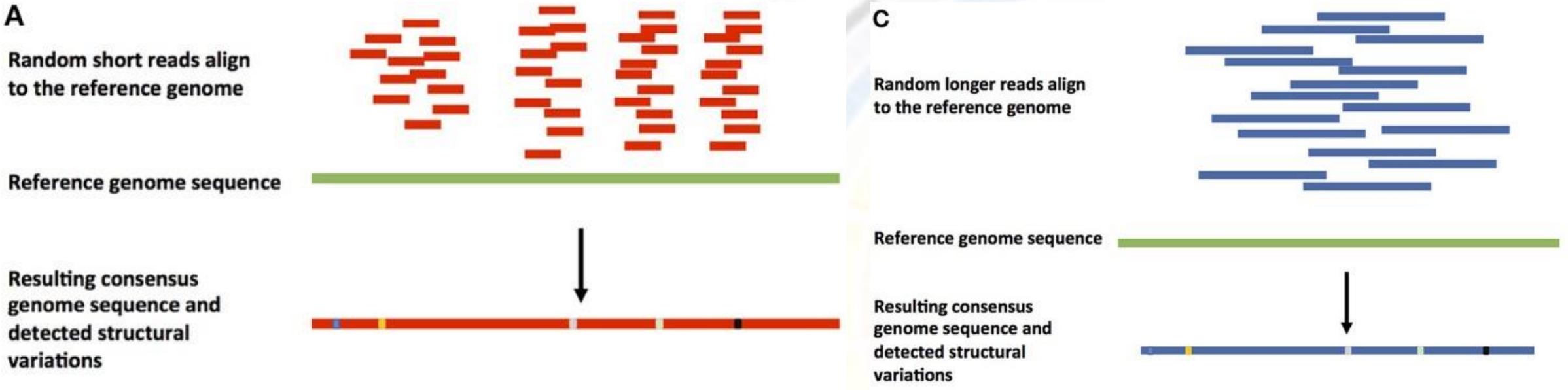



Adaptado de:
COMMINS et al. 2009. **Biological Procedures Online**. DOI: 10.1007/s12575-009-9004-1

# Baseado em genoma de 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

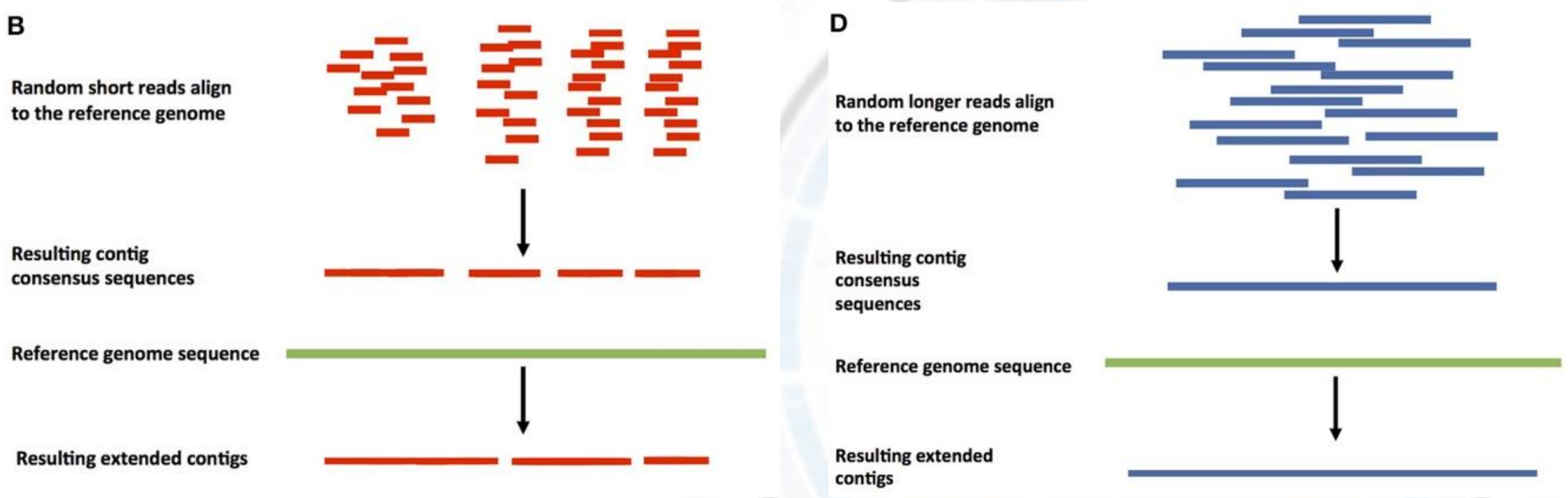
# Montagem guiada por genoma de referência




Adaptado de:
KYRIAKIDOU et al. 2018. **Frontiers in Plant Science**. DOI: 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




Adaptado de:
KYRIAKIDOU et al. 2018. **Frontiers in Plant Science**. DOI: 10.3389/fpls.2018.01660

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

# *De novo*

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

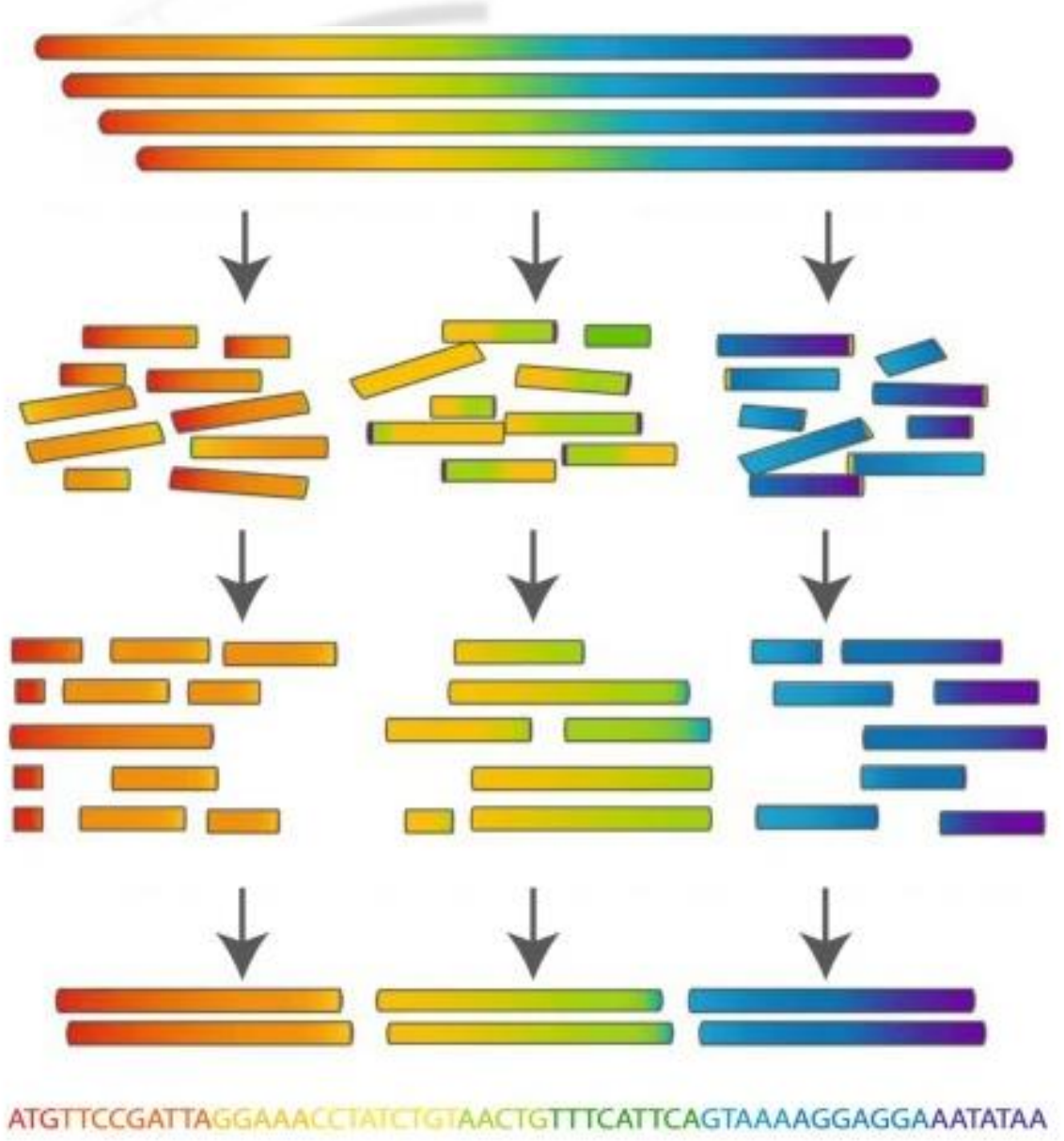
- Algoritmos
  - *Greedy*
  - Baseados em grafos

# Algoritmo Greedy

- Busca de ótimo local (em detrimento de ótimo global)

- **Passos gerais**
  - Cálculo da distância entre reads
  - Clusterização dos reads com maior sobreposição
  - Montagem de reads com sobreposição em contigs
  - Repetição dos passos anteriores até que contigs maiores não possam ser montados

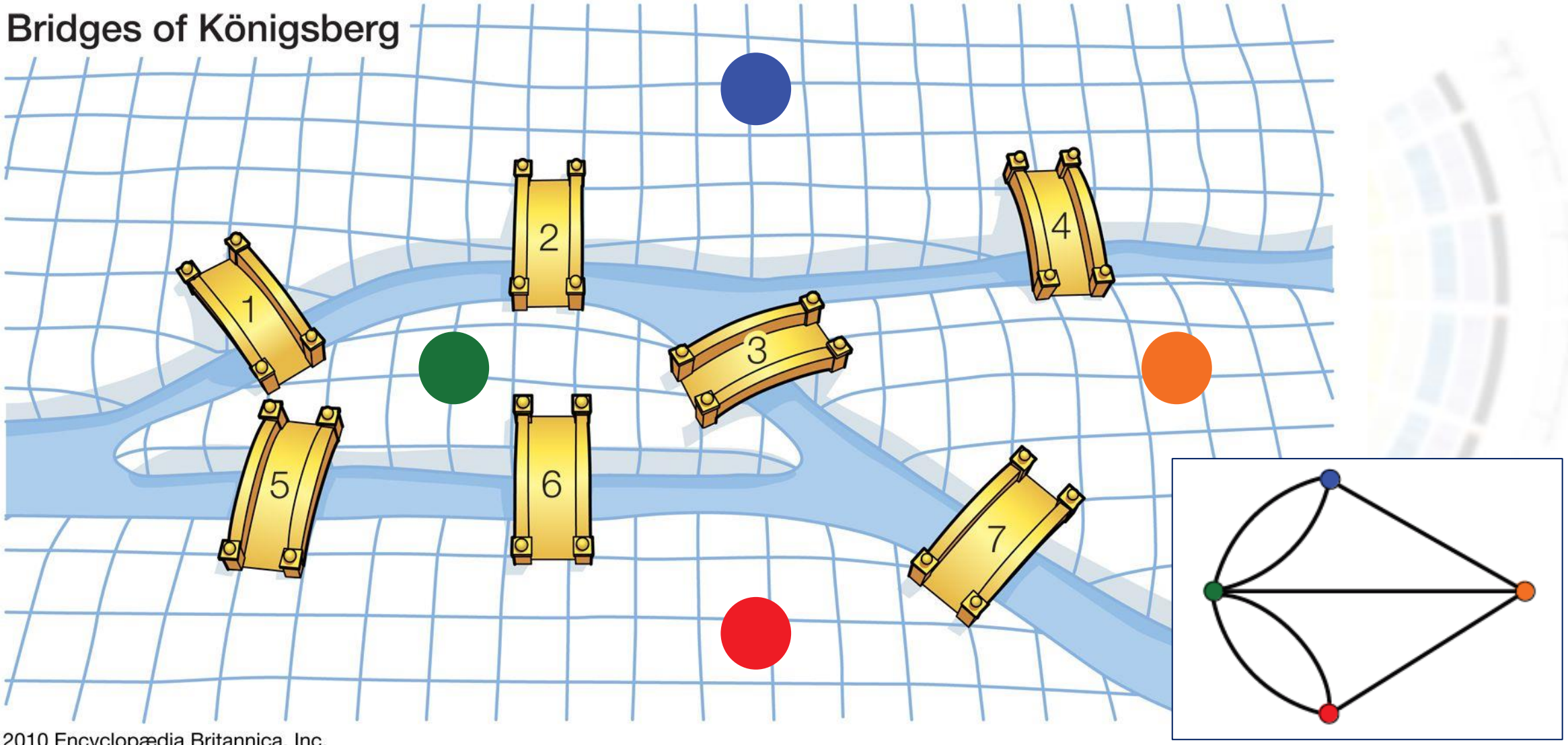
- **Problemas**
  - Não indicados para grandes conjuntos de dados (dificuldade de encontrar o ótimo global)
  - Dificuldade de montagem de regiões repetitivas




Adaptado de:
COMMINS et al. 2009. **Biological Procedures Online**. DOI: 10.1007/s12575-009-9004-1

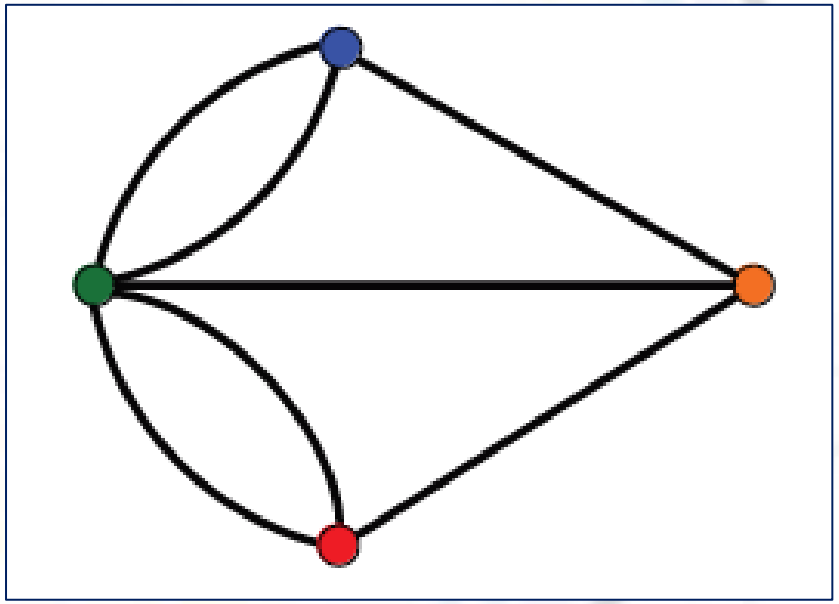
# As sete pontes de Königsberg (Kaliningrad, Rússia)

# As sete pontes de Königsberg (Kaliningrad, Rússia)



# Grafos

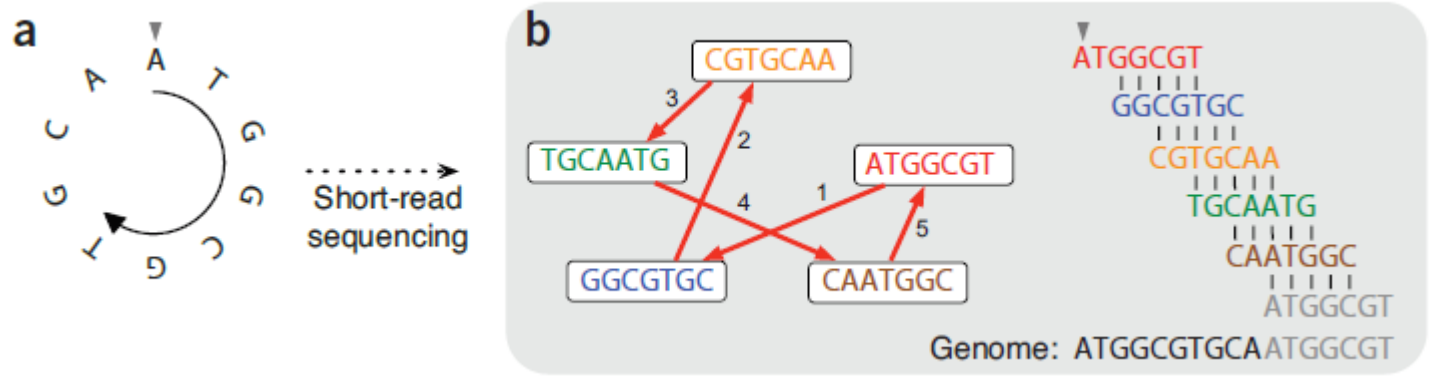
- Um grafo G é definido como um conjunto de vértices (V) e um conjunto de arestas (A) que representam pares únicos dos elementos de V



# Pensando a montagem de um genoma por meio de grafos

- Cada read é um vértice e cada sobreposição entre reads (aresta) é representada pela seta vermelha

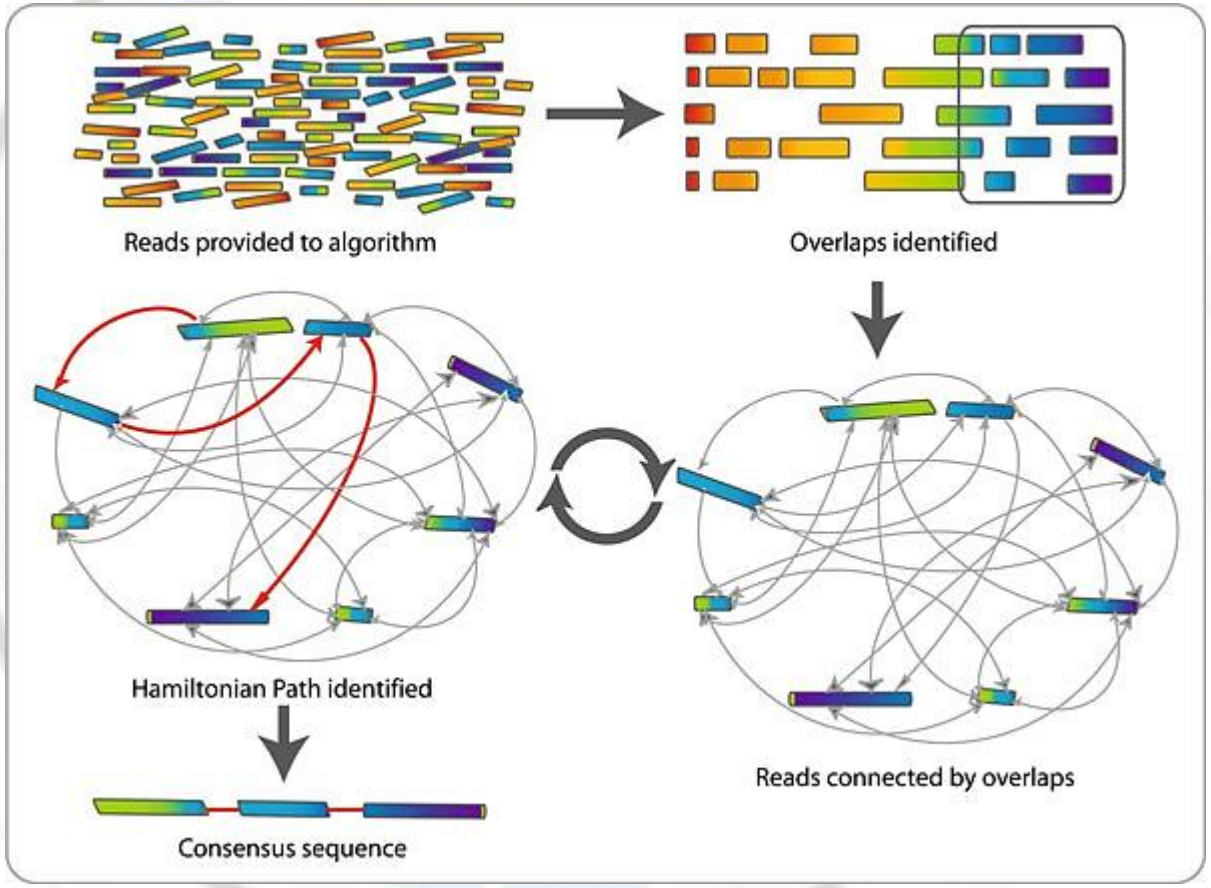
- **<u>Caminho Hamiltoniano:</u>** passa por cada um dos reads apenas uma vez e termina no read inicial, e inclui todos os reads




Adaptado de:
COMPEAU et al. 2011. **Nature Biotechnology**. DOI: 10.1038/nbt.2023

# Overlap-layout-consensus (OLC)

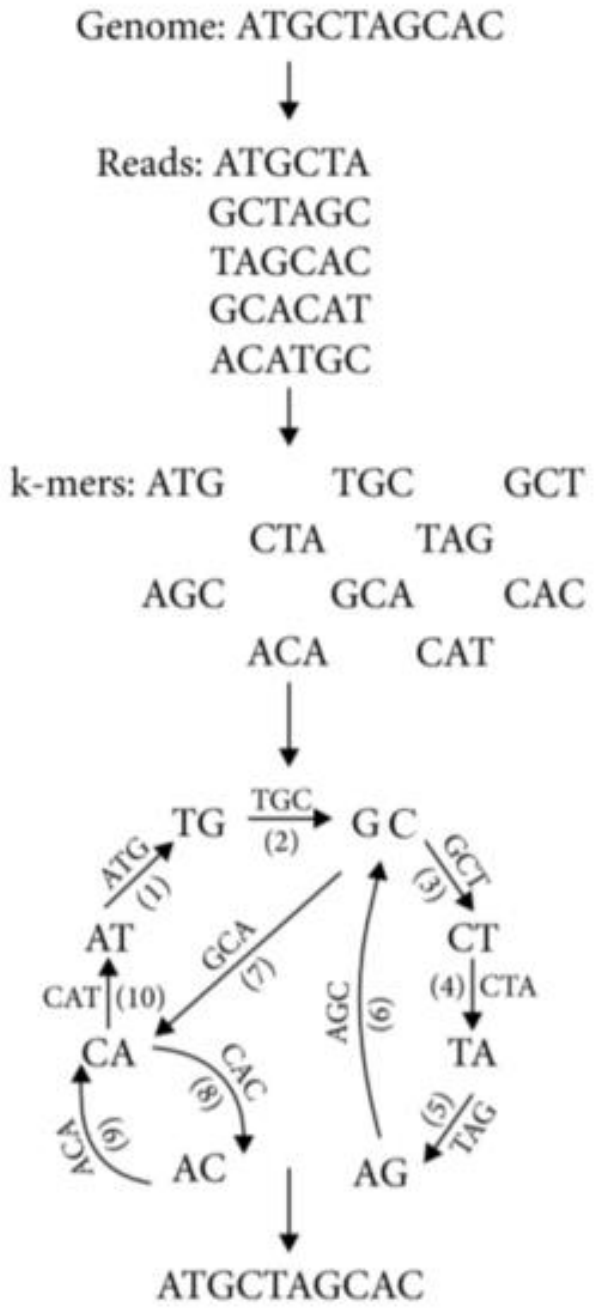
- Sobreposição entre os reads (similar ao algoritmo greedy)
- Grafo de sobreposições, em que cada read é um vértice conectados pelas sobreposições (arestas)
- Encontrar o caminho passando **por todos os vértices** para gerar contigs
- O caminho ideal seria um caminho Hamiltoniano: cada vértice seria visitado apenas uma vez
- Computacionalmente difícil




Adaptado de:
COMMINS et al. 2009. **Biological Procedures Online**. DOI: 10.1007/s12575-009-9004-1

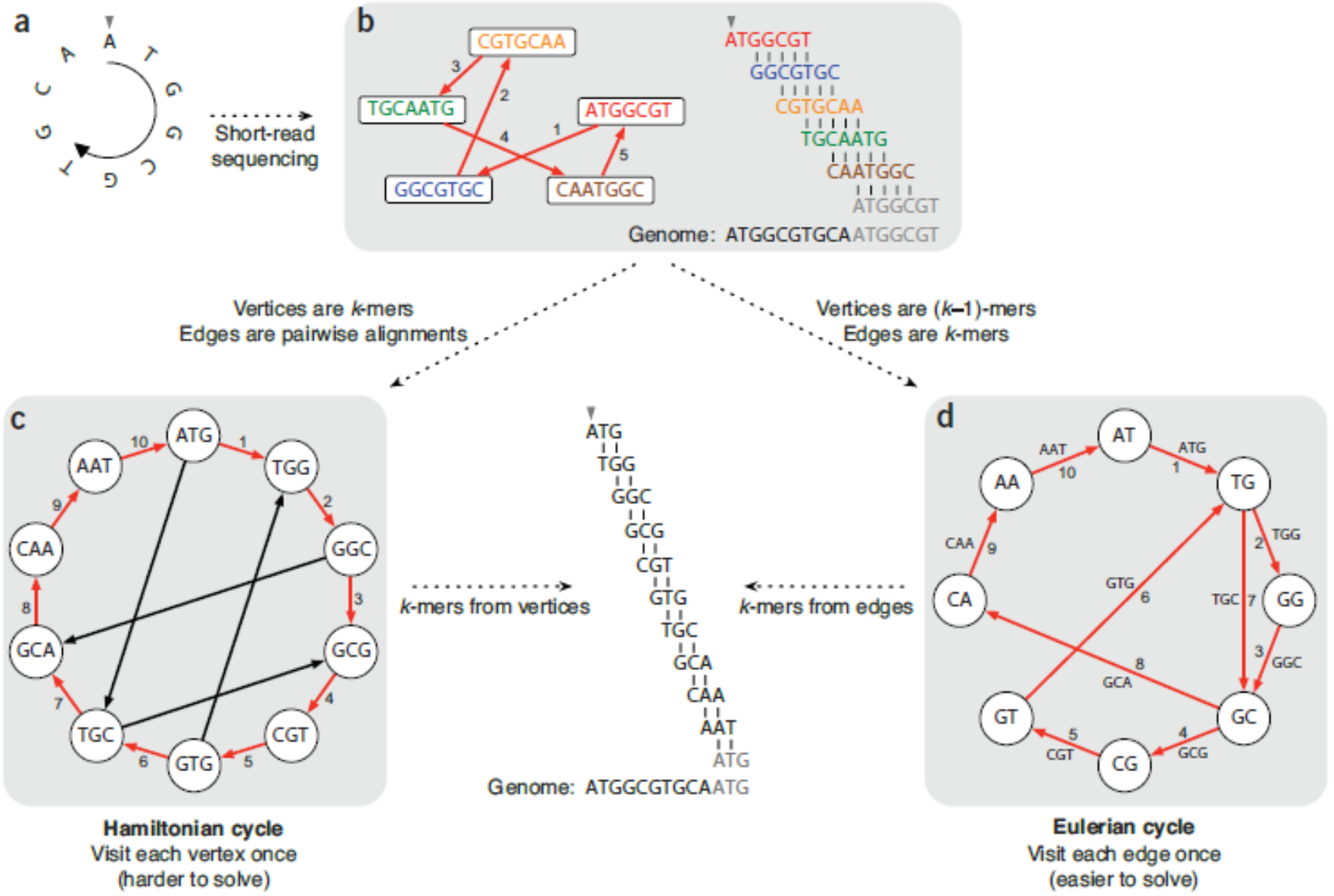
# Grafos de '*De Bruijn*'

- Reads são quebrados em fragmentos de tamanhos específicos (k-mers)
- K-mers - 1 utilizados como vértices na montagem do grafo
- Arestas entre os vértices representadas pelos k-mers
- Encontrar o caminho passando **por todas as arestas** para gerar os contigs
- O caminho ideal seria um caminho Euleriano: cada aresta seria visitada apenas uma vez
- Computacionalmente mais fácil, há vários algoritmos eficientes para encontrar caminhos Eulerianos




Adaptado de:
MADURANGA et al. 2020. **Towards Data Science (Blog)**. Link do Post:

# Comparativo




Adaptado de:
COMPEAU et al. 2011. **Nature Biotechnology**. DOI: 10.1038/nbt.2023

# Etapas da montagem *de novo*

Reads

↓

*Contigs*

↓

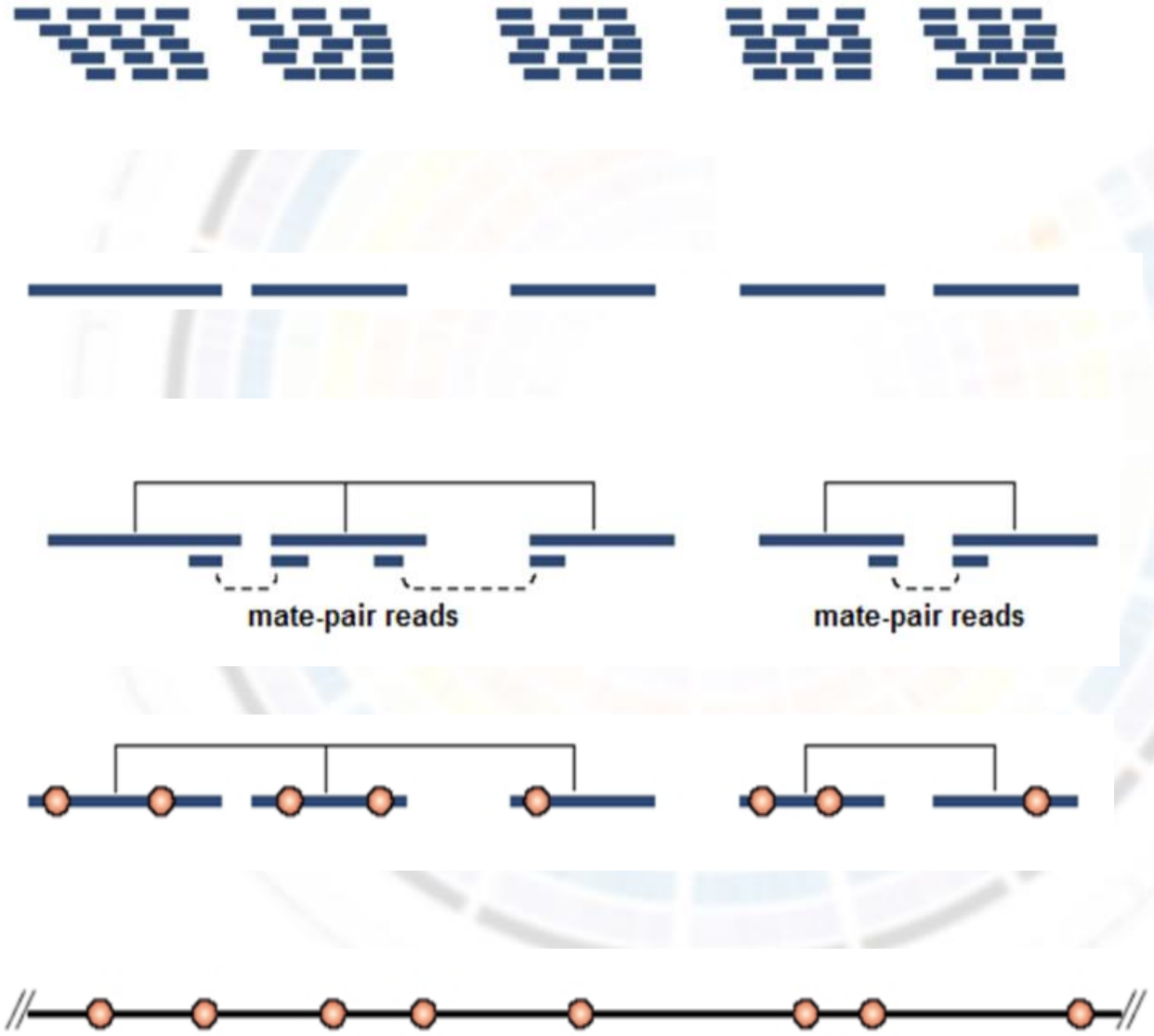
Ordenação dos *contigs*

↓

*Scaffolds*

↓

Sequência do genoma



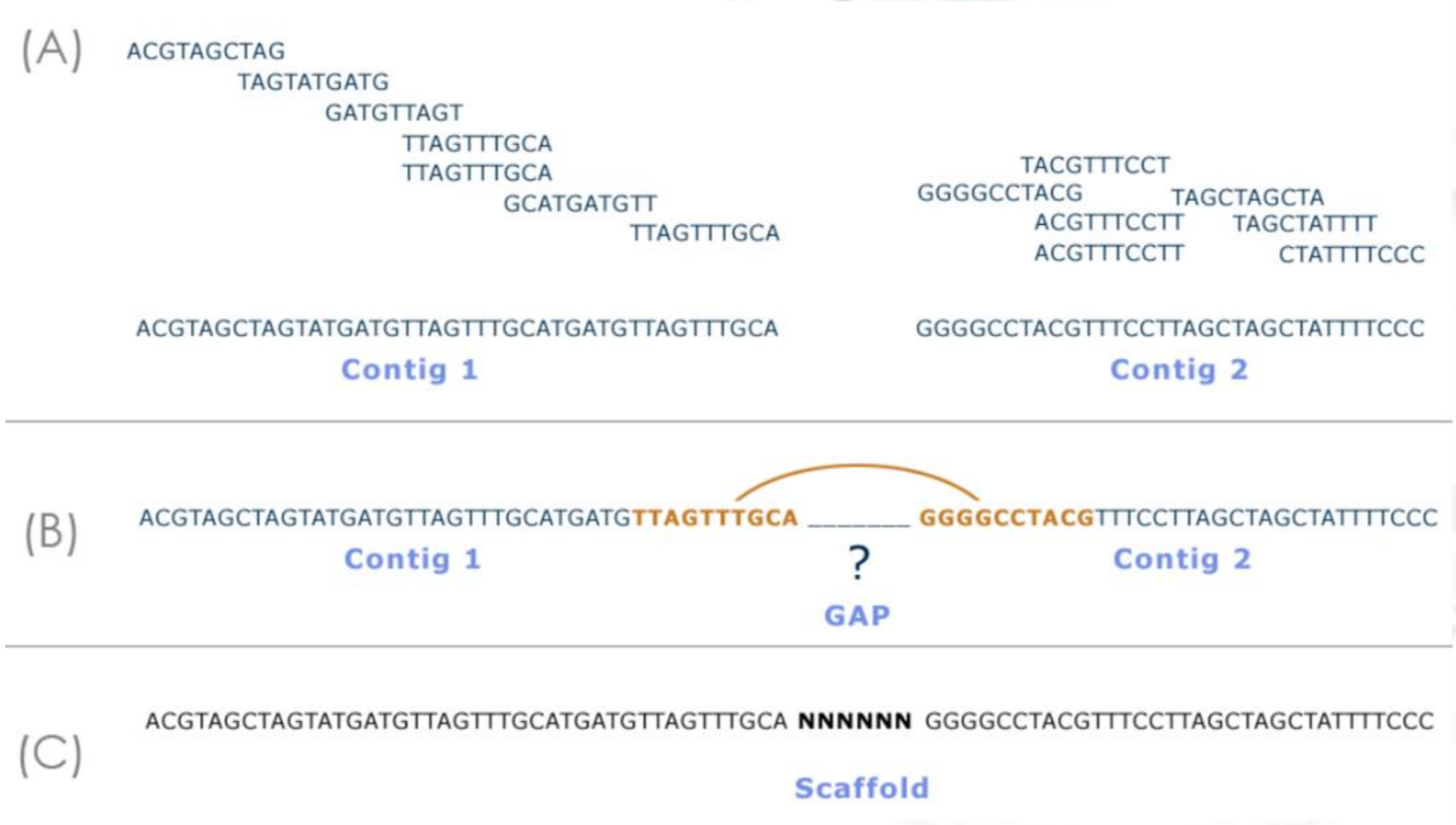
# Obtenção dos contigs

- *Contigs* (sequências contíguas) são geradas a partir da sobreposição de um conjunto de reads


- A sequência do *contig* é a sequência consenso dos reads sobrepostos

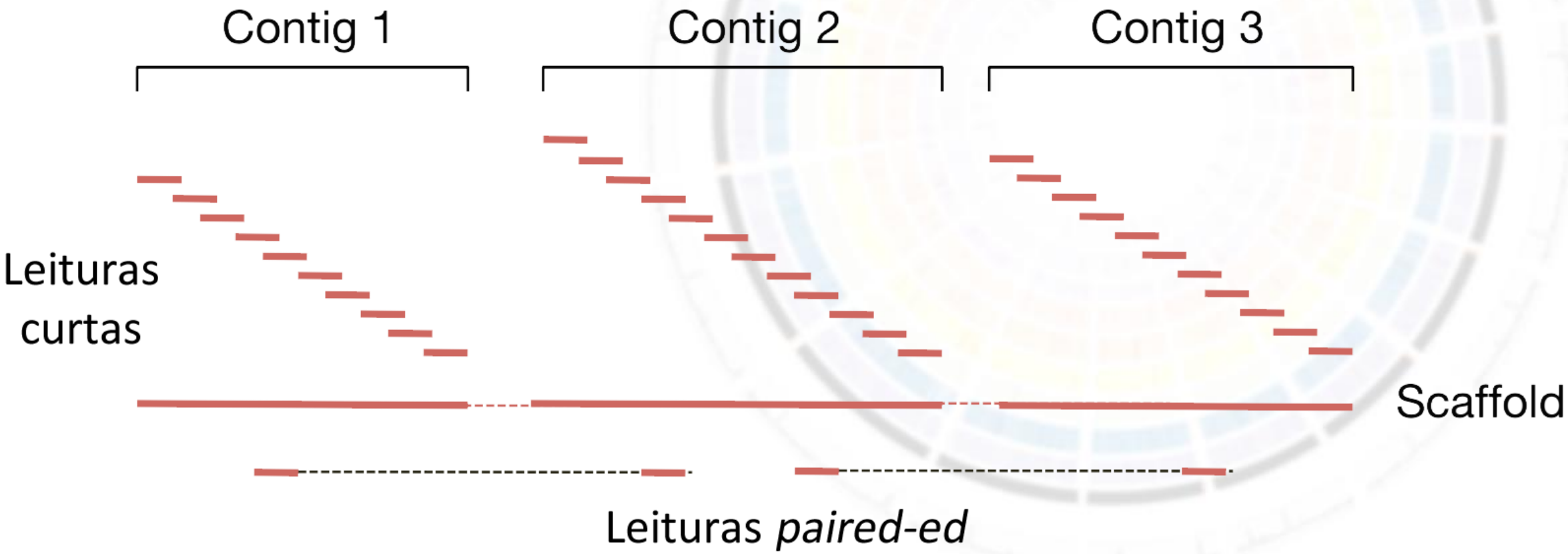
```
Seq4    TTCACACACCCTATACCAATAGTTTTCTGGCTCCTGACCATCAAACTG
Seq5                          TTTTCTGGCTCCTGACCTTCAAACTGCCTCCATATGACTGTGCTCT
Seq6                  TACCAATAGTTTACTGGCTCCTGACCCTCAAACTGCCTCC
Seq7                       ATAGTTTTCTGGCTCCTGACCGTCAAACTGCCTCCATATGA

Cons    TTCACACACCCTATACCAATAGTTTTCTGGCTCCTGACCNTCAAACTGCCTCCATATGACTGTGCTCT
```

# Ordenação dos contigs

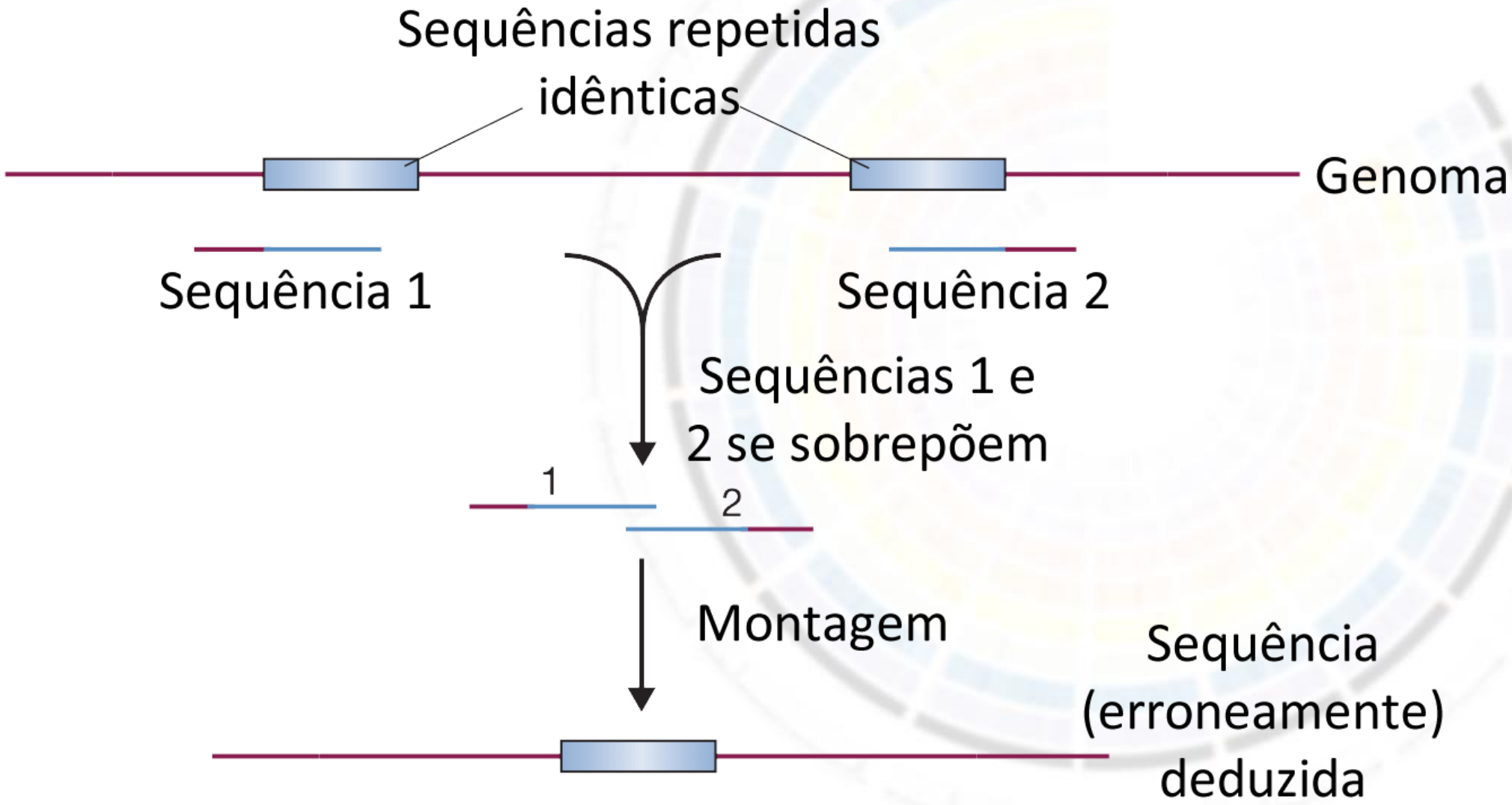


# Ordenação dos contigs

- O uso de sequências em pares (pair-end ou mate-pair) permite a ordenação dos contigs em scaffolds

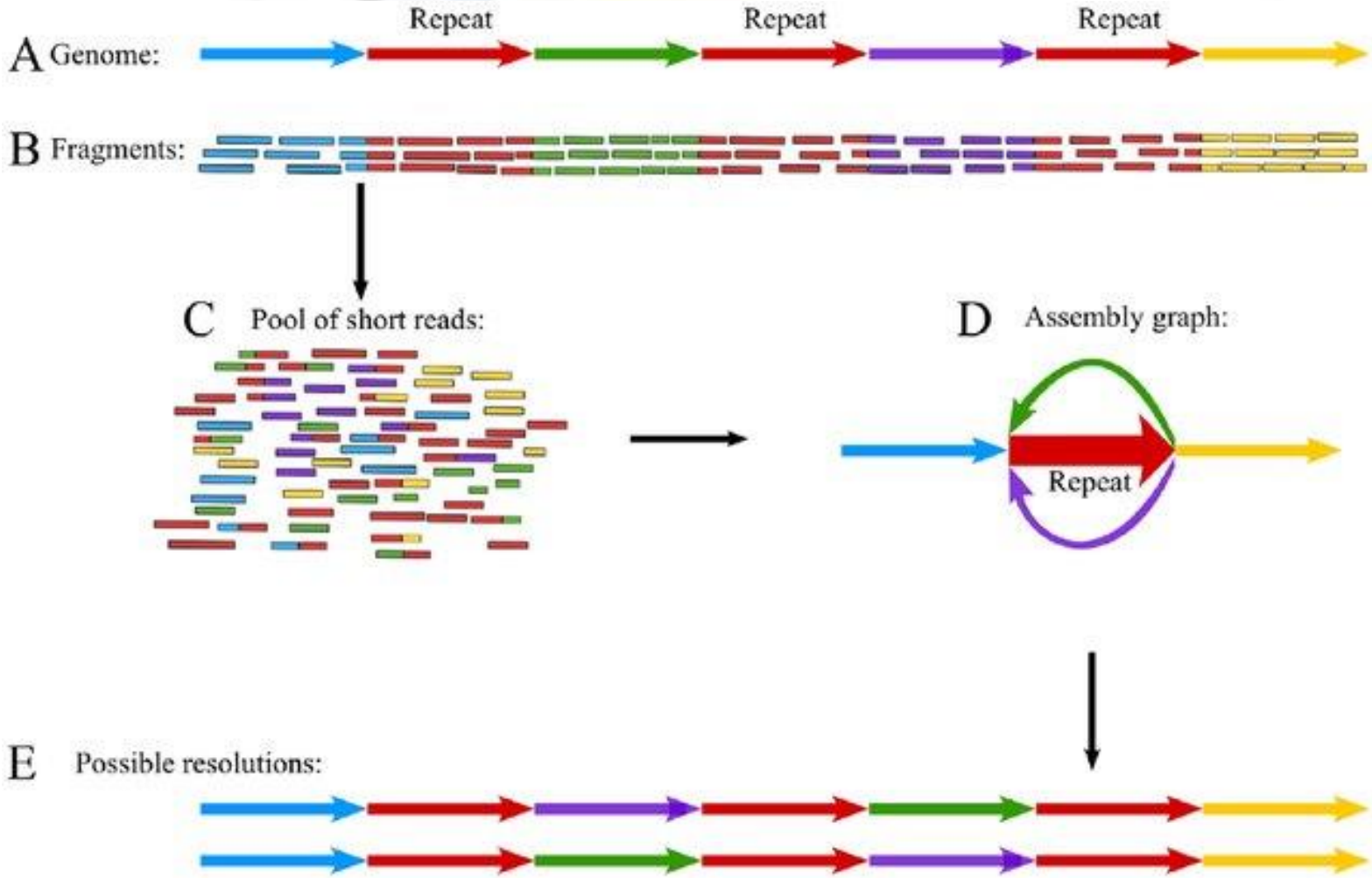


# Dificuldades: montagem de regiões repetitivas



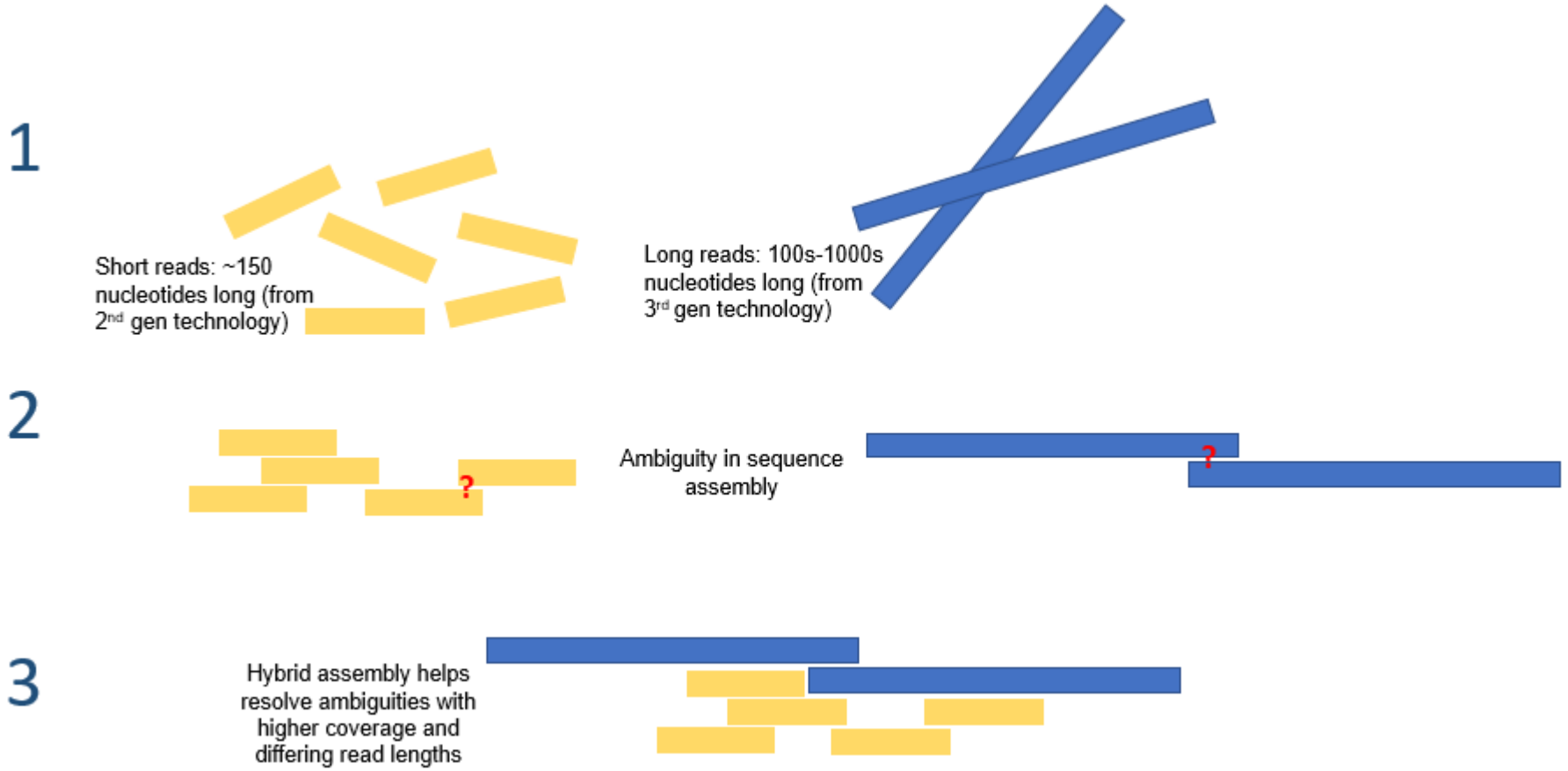
# Dificuldades: montagem de regiões repetitivas

- **Regiões repetitivas mais longas que os 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

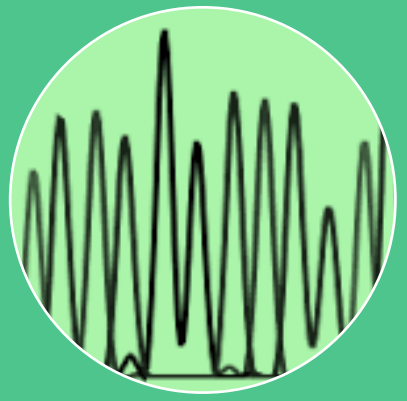
Em geral são arquivos longos e pesados, exigindo o uso de softwares para processar o arquivo completo e obter a informação de interesse

```
1  >scaffold_1
2  CCATGGCTGTCTTGCGATTGTCCAGGGCAGTCTTGACAGCAGGGGCAAGTTGCGCCGCCGCCCGTCCCTT
3  CTCAGTGTCTTCGAAGTTGAGGGAGACGATGACCCTGGTGTTGATGGGACTGTTGGTGTTCGCCGTGGAA
4  GCTTCGTCCTTCTTGCGCTTGGAGCCGGCCGACGCCGCCTTCTTGCGCTTGGCAATCTCCTCGGGATGCG
5  TGAGAATCTCTTCAATTTTTGCCATGAAGGCGTTCTCTTTCTCGATTTACGAGCGAACTTCCTCGTAGAA
6  TGCCGTGGCTACGCTTGACTCGGTCTTGAAGATGTTGTGGAGAGCCTTGCGGGTGGTCGTTCACGACGAA
7  GATGCAGAGAGGGCGCGGTCGTGGTTCTGCGCGATGGCGTTGCGGGTTGTCTTCGTCTTGTTGAACCAGT
8  TGCCGAACTGAATTACGTCGTCTTTCTTTGCCATCTTTTCCTCGGAGCTCATCGCTTCGATGGTGGCGGC
9  GTCATCCTTGCGCTTTTCGGCGGTCTCGTTTTTCTTCGTCTGGGTGACTTGCAGAAGTGCCTTTGCCCTC
10 AAAGCACTCATTCGGCGACTCTCCTGTTCCGCATCGACGACCTGGCGCCATTCCTGTGACGCATCGCTCA
```

# Como avaliar uma montagem?

## Contiguidade

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



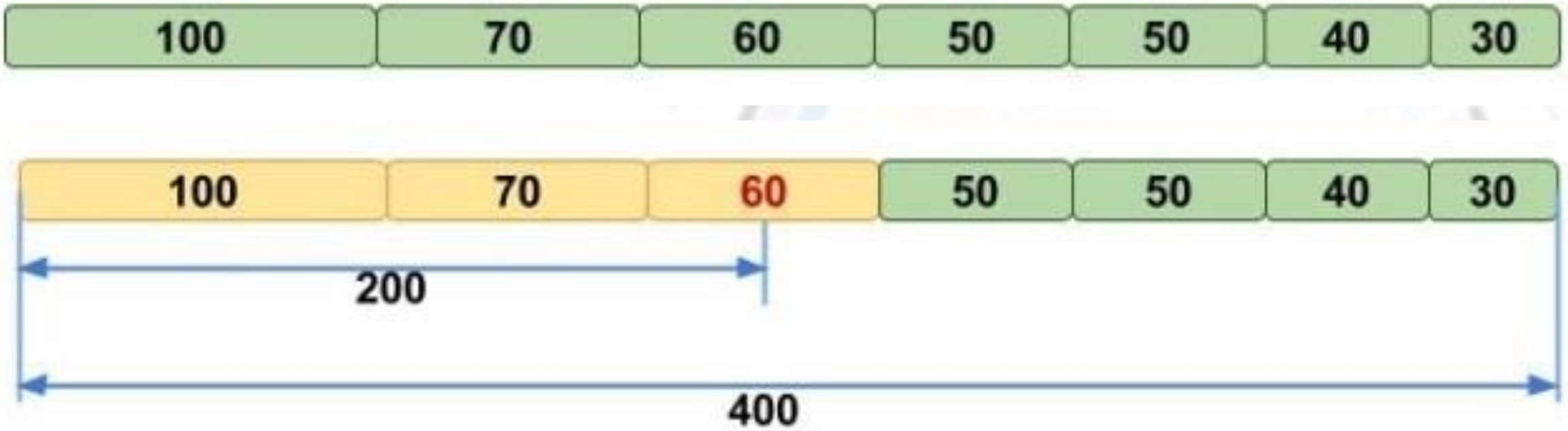
## Análise de bases

- 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

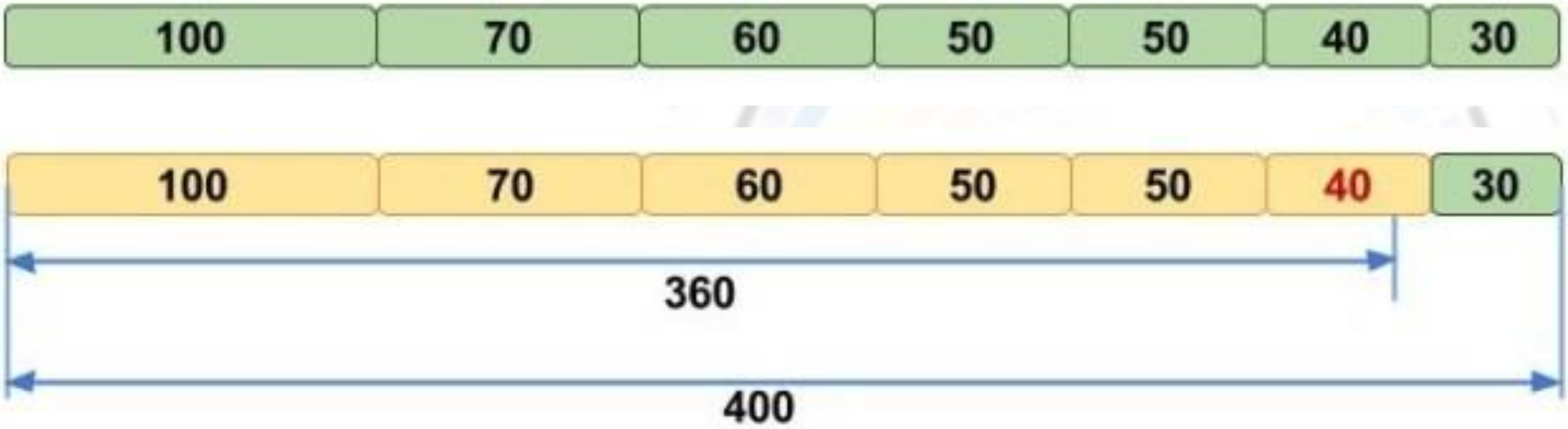
# Contiguidade – N50




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

Adaptado de:
VIDEVALL et al. 2017. **The Molecular Ecologist (Blog)**. Link do Post:

# Contiguidade – N90




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

Adaptado de:
VIDEVALL et al. 2017. **The Molecular Ecologist (Blog)**. Link do Post:

# Contiguidade – L50


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

Adaptado de:
VIDEVALL, 2017. **The Molecular Ecologist (Blog)**. Link do Post:

# Análise de bases - cobertura

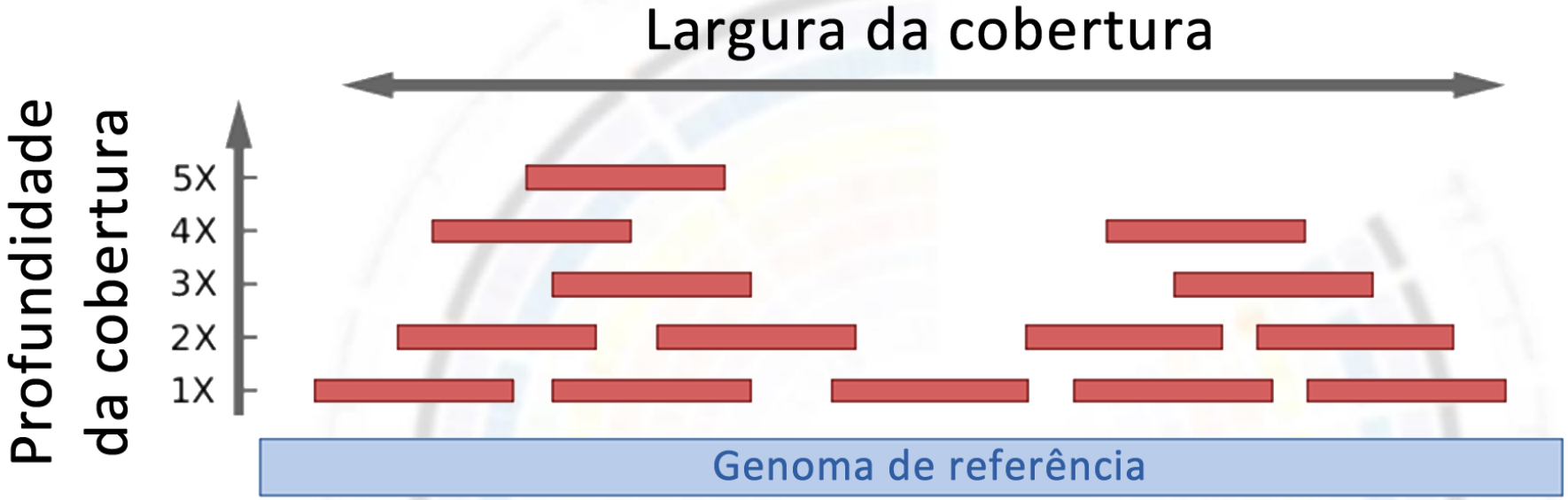
- A cobertura se refere à quantidade de vezes que o genoma foi sequenciado

- Alta cobertura: maior precisão e redução de erros nas montagens

- $$Cobertura = \frac{Tamanho\ dos\ reads \times quantidade\ de\ reads}{Tamanho\ total\ do\ genoma}$$

# Cobertura

- Boa cobertura garante a qualidade da montagem final




```
                                AGCATCGTAGCTTCAGTATGATGATGCTAG   Read 1
ATGATCGTAGCTAGCATCGTAGCTAGC                                      Read 2
   ATCGTAGCTAGCATCGTAGCTAGCATCGTAGCTT                            Read 3
                                TTGTAGCTTCAGTATGATGATGCTAG        Read 4
              GCATCGTAGCTAGCATCGTAGCTTCAGT                       Read 5
ATGATCGTAGCTAGCATCGTA                                            Read 6
ATGATCGTAGCTAGCATCGTAGCTAGCATCGTAGCTTCAGTATGATGATGCTAG           Deduced sequence
```

# Análise de bases - cobertura

- Também é possível calcular a cobertura de alinhamento, re-alinhando os reads originais à montagem

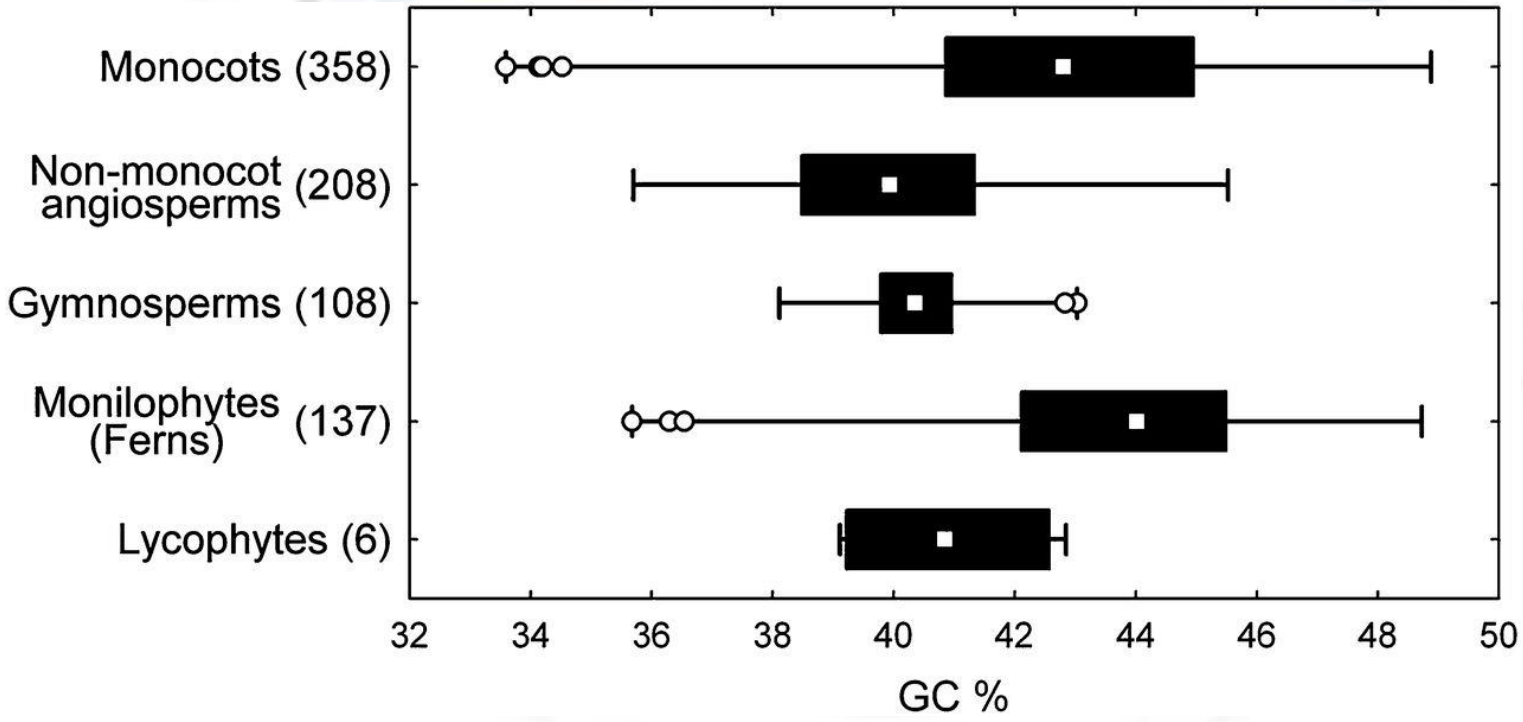
- Há muitos reads que não foram alinhados?

- Há regiões da montagem com poucos reads alinhados em relação às outras?

Adaptado de:
KYRIAKIDOU et al. 2018. **Frontiers in Plant Science**. DOI: 10.3389/fpls.2018.01660

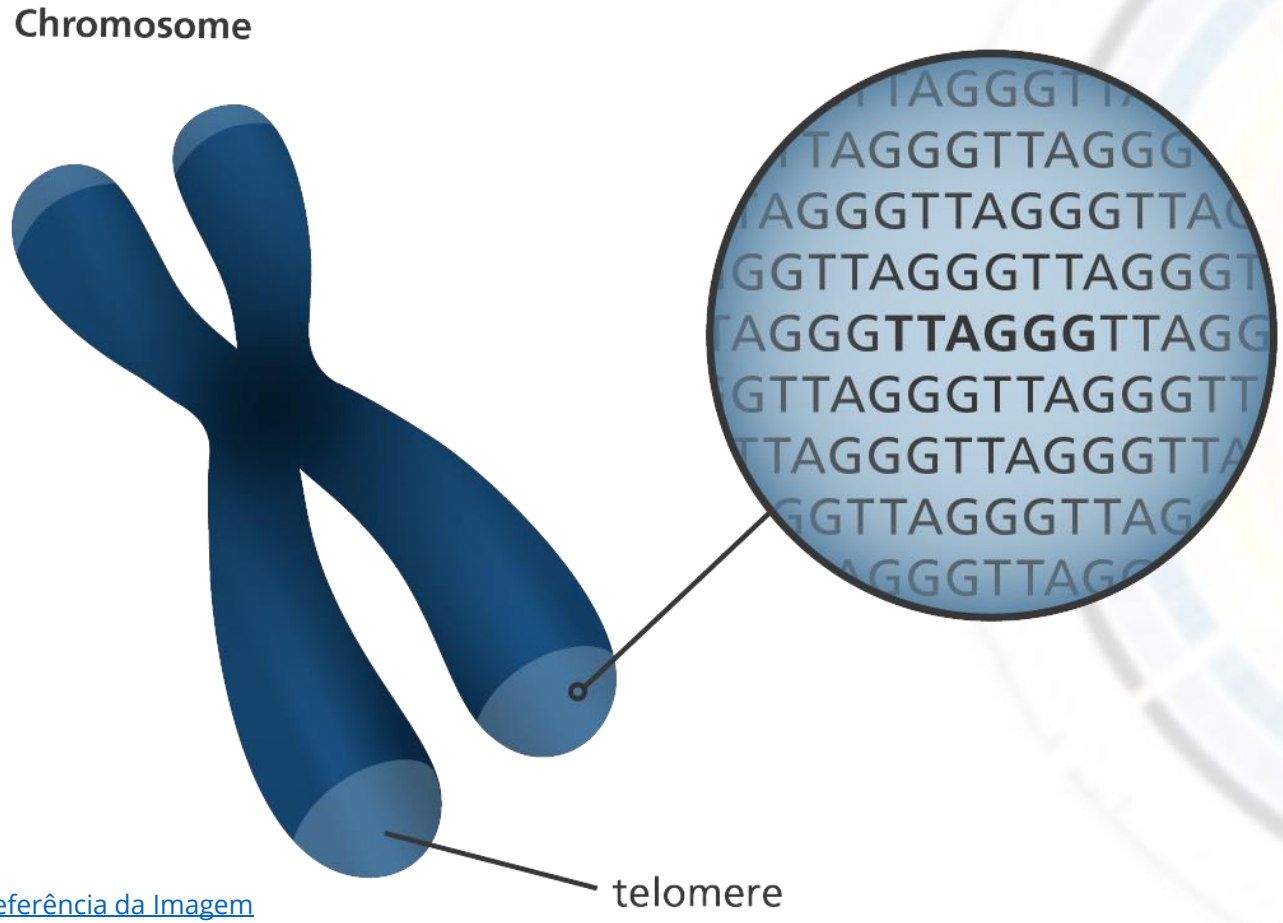
# Análise de bases – Conteúdo GC

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




Adaptado de:
ŠMARDA et al. 2014. **PNAS**. DOI: 10.1073/pnas.1321152111

# Análise de conteúdo - Telômeros




Referência da Imagem

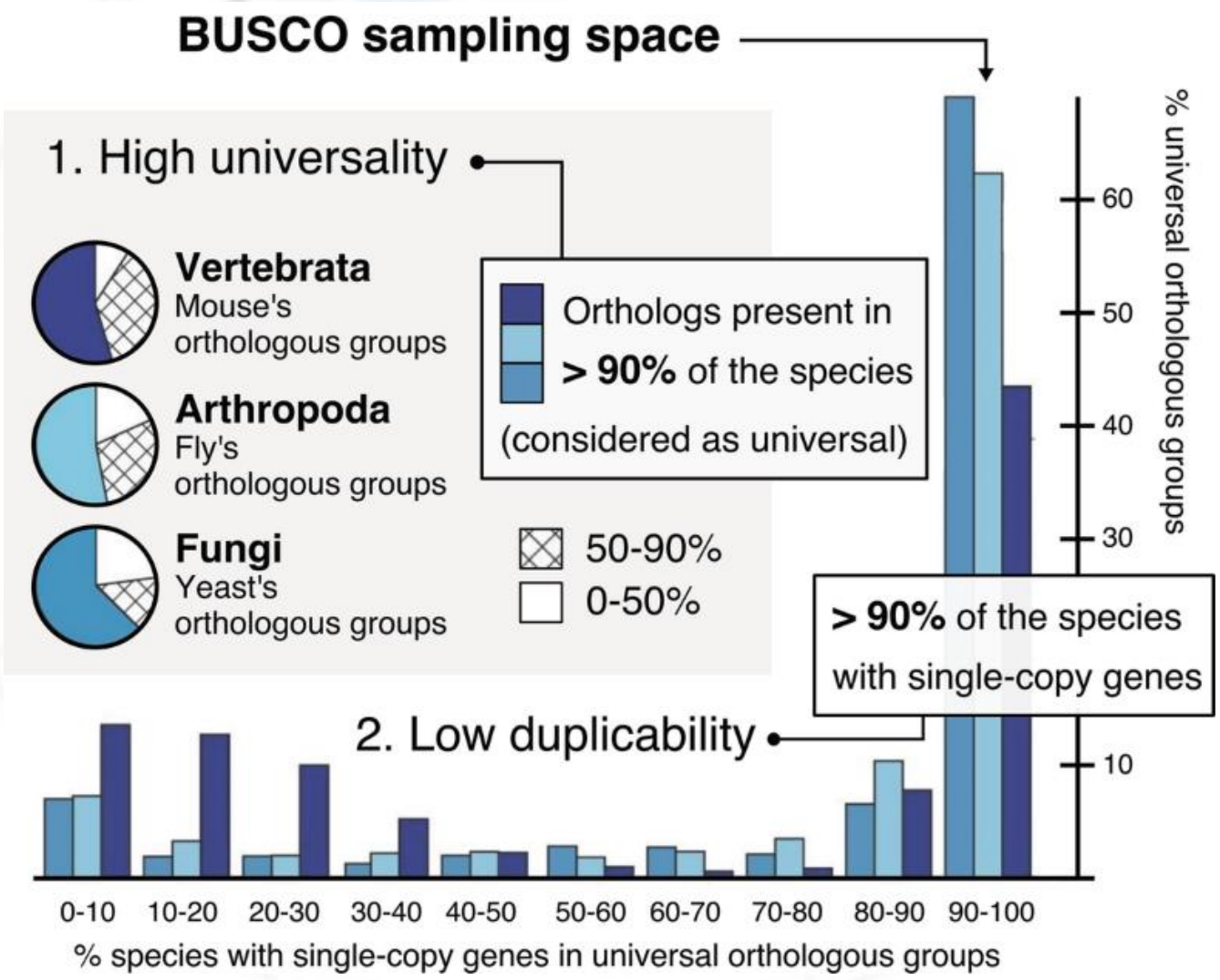
- Sequências repetitivas encontradas nas extremidades dos cromossomos eucarióticos
- Função protetiva:
  - Impedem que os cromossomos se fusionem nas extremidades
  - Evitam que as sequências codificantes sejam perdidas a cada rodada de replicação
- Presença de telômeros no início e fim de um contig/scaffold 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

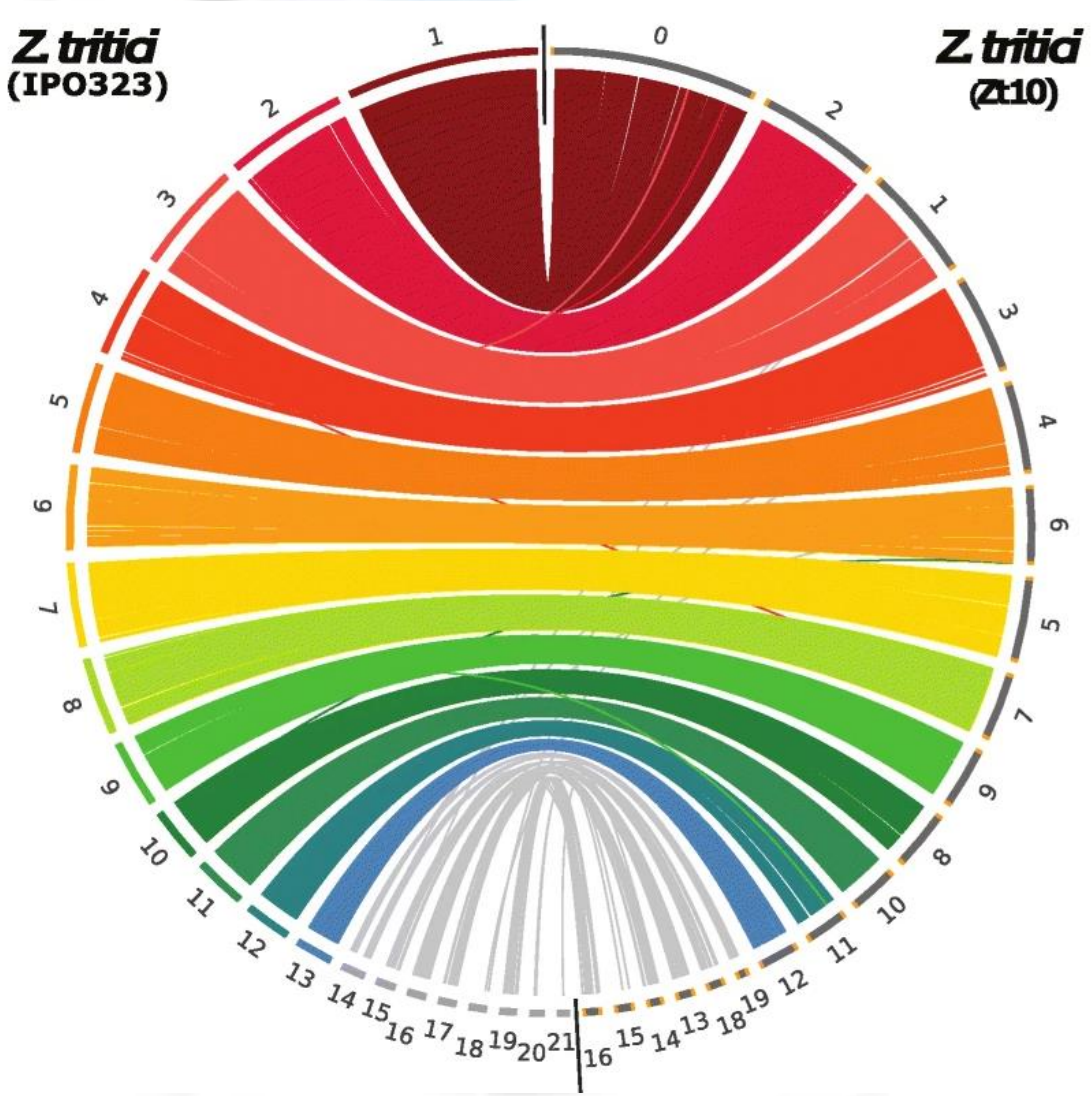



Adaptado de:
SEPPEY et al. 2019. **Methods in Molecular Biology**. DOI: 10.1007/978-1-4939-9173-0_14

# Análise de conteúdo – Comparação com genoma de referência

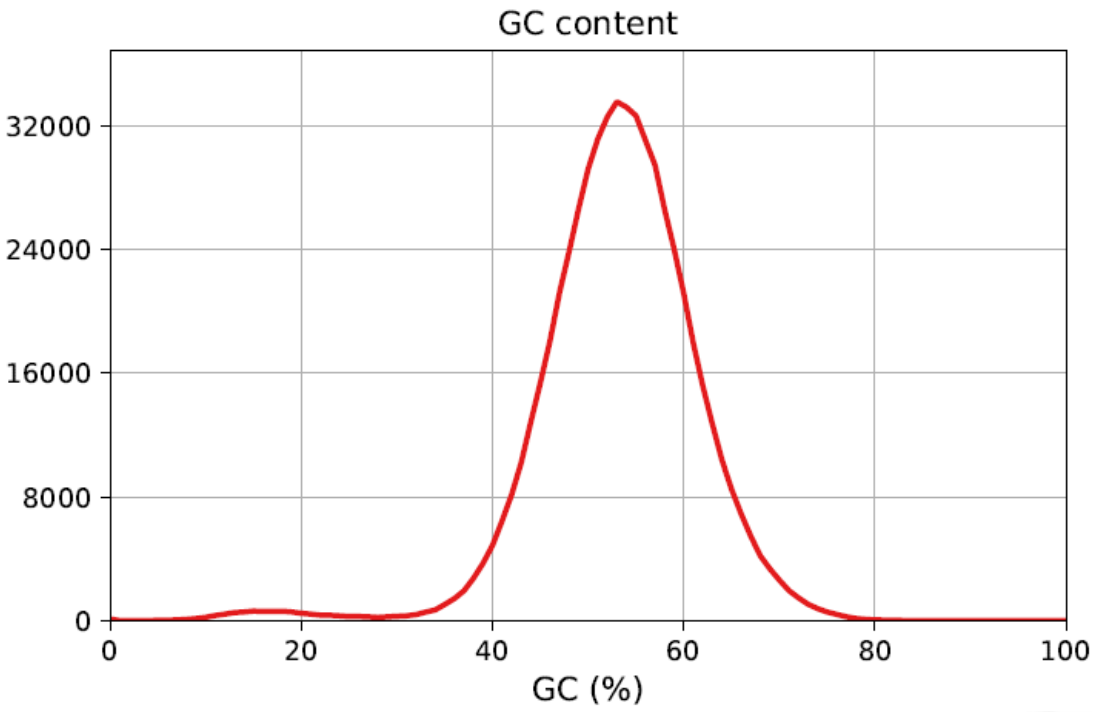
- Genes essenciais e conservados presentes na linhagem de referência estão presentes na nova montagem?

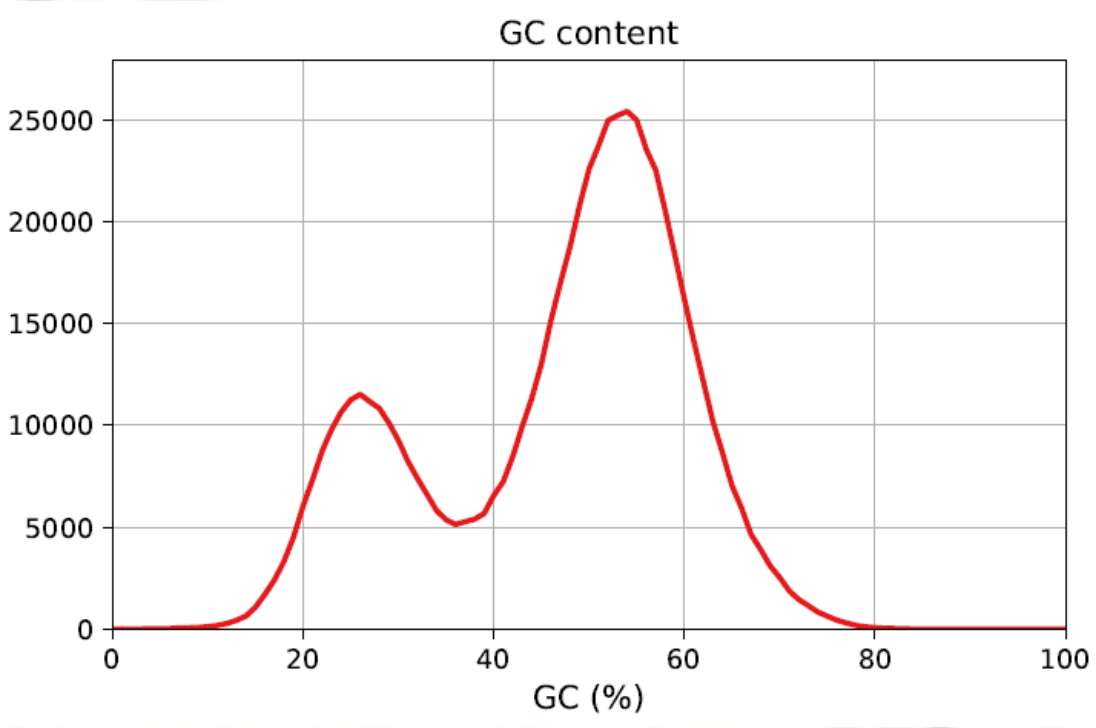
- A organização da nova montagem é similar à montagem de referência?




Adaptado de:
FEURTEY et al. 2020. **BMC Genomics.** DOI: 10.1186/s12864-020-06871-w

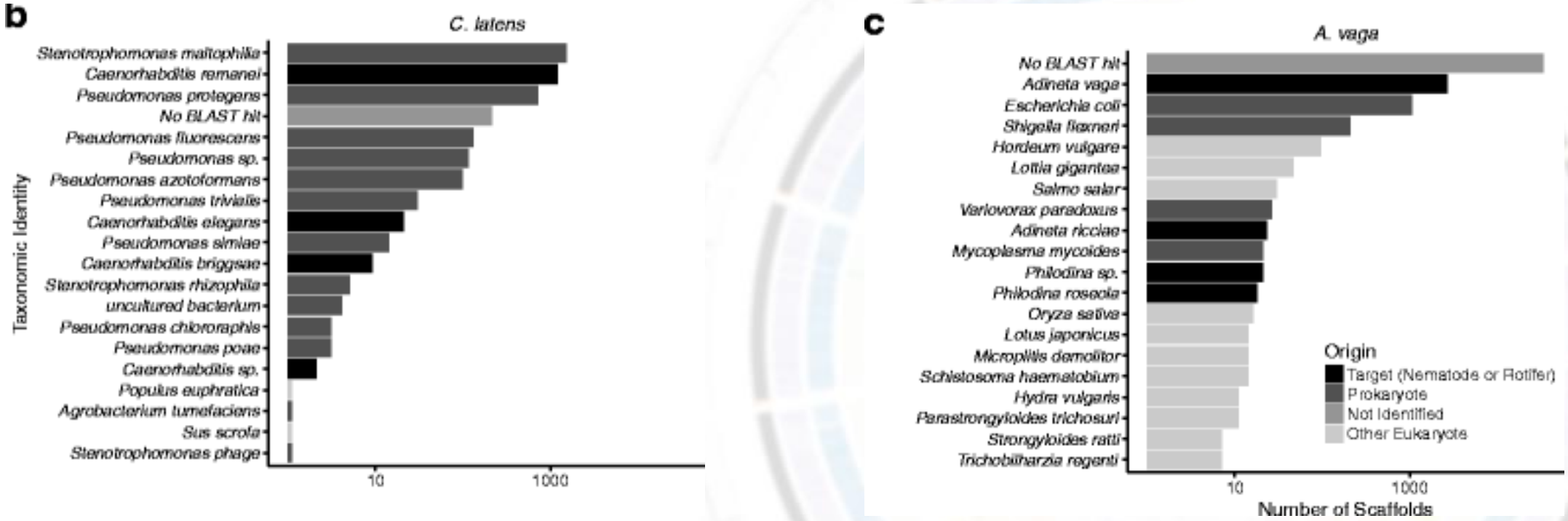
# Análise de conteúdo – Contaminantes (Distribuição do conteúdo GC)



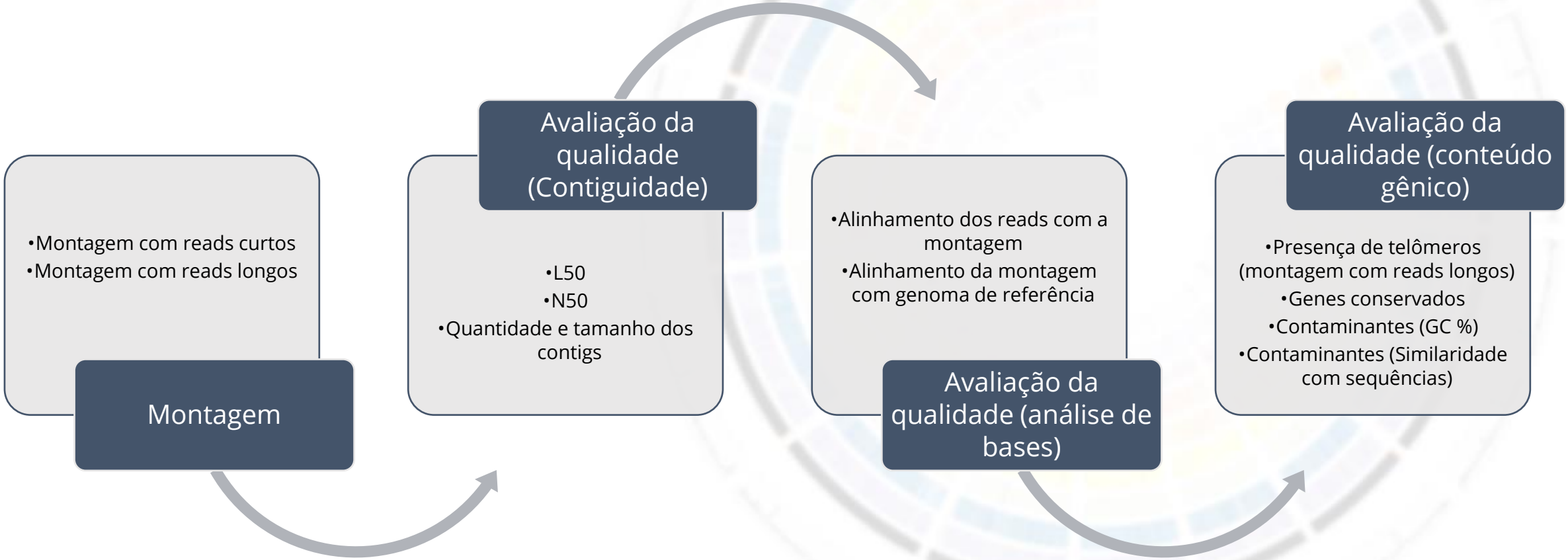

- Quantos picos são observados na distribuição de conteúdo GC?
- Mitocôndria, sequências repetitivas ou contaminação?

# Análise de conteúdo - Contaminantes




Adaptado de:
FIERST et al. 2017. **BMC Bioinformatics**. DOI: 10.1186/s12859-017-1941-0

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


•Montagem com reads curtos
•Montagem com reads longos
Montagem
Avaliação da qualidade (Contiguidade)
•L50
•N50
•Quantidade e tamanho dos contigs
•Alinhamento dos reads com a montagem
•Alinhamento da montagem com genoma de referência
Avaliação da qualidade (análise de bases)
Avaliação da qualidade (conteúdo gênico)
•Presença de telômeros (montagem com reads longos)
•Genes conservados
•Contaminantes (GC %)
•Contaminantes (Similaridade com sequências)