

Montagem de genomas e avaliação de qualidade da montagem

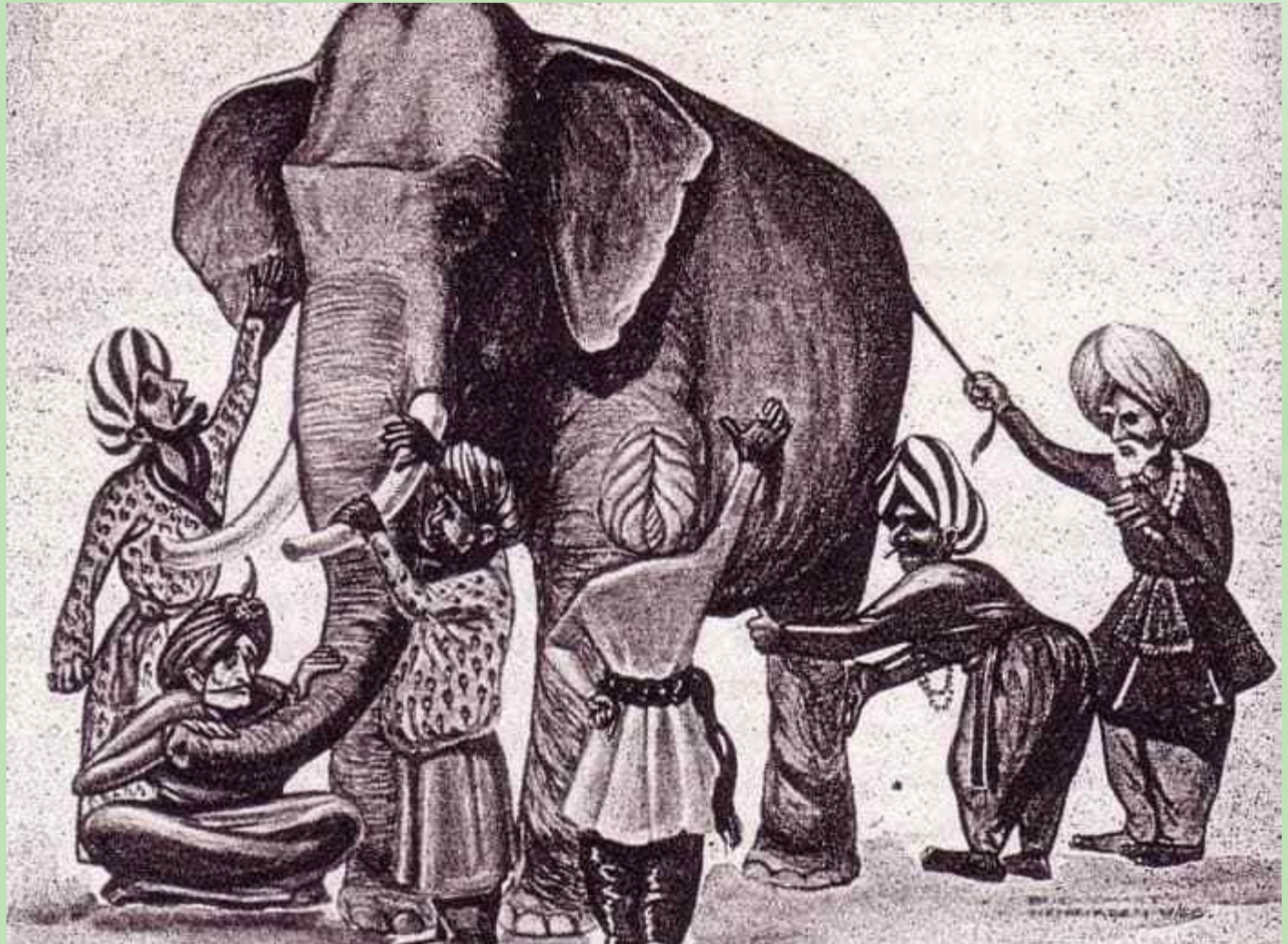
Desirrê Petters-Vandresen

Módulo I – Genômica no Estudo de Microrganismos

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

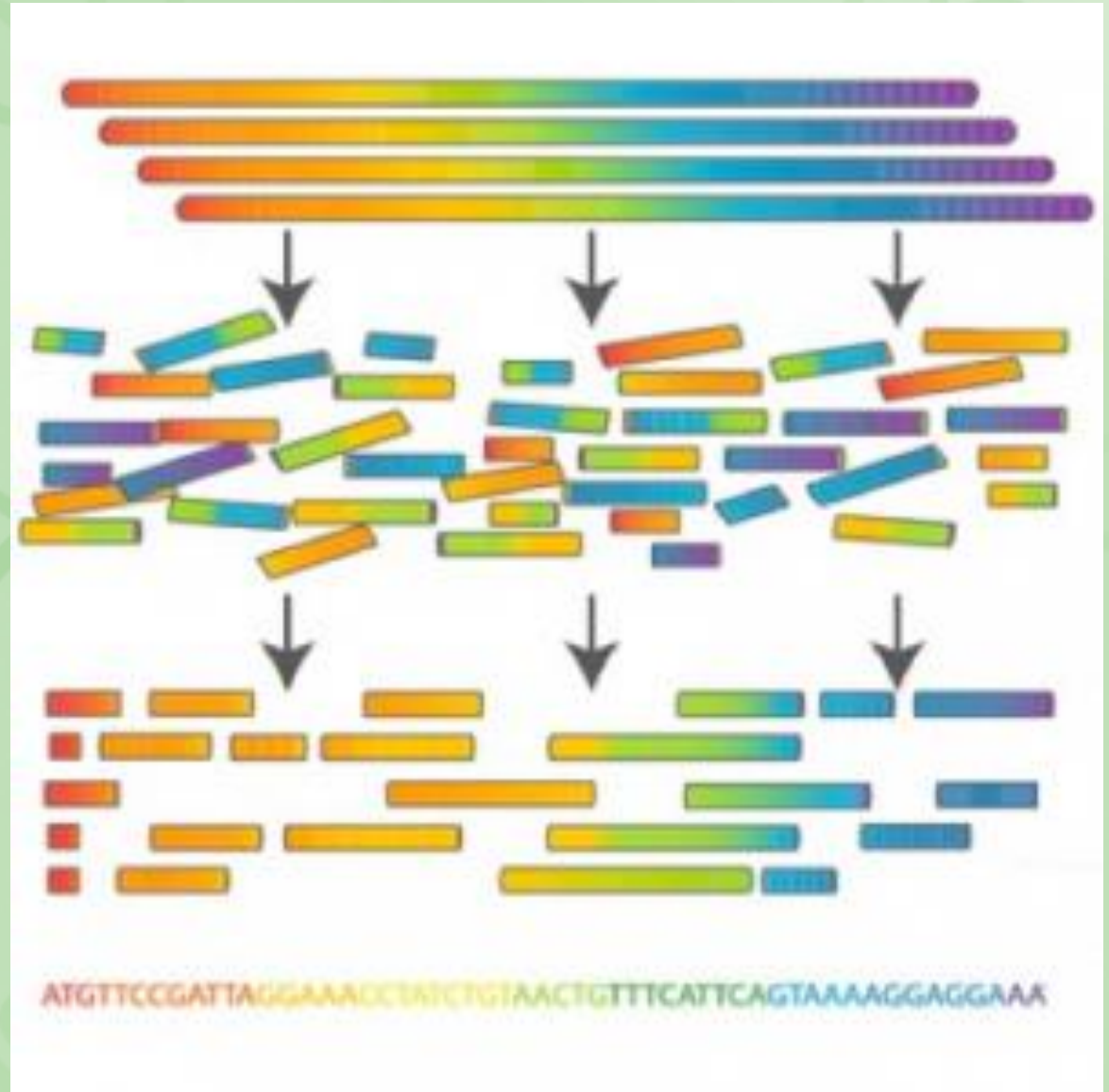
- **Como organizar todos os fragmentos obtidos no sequenciamento na ordem biológica correta e formando uma sequência única e coesa?**



[Referência da Imagem](#)

Montagem de genomas

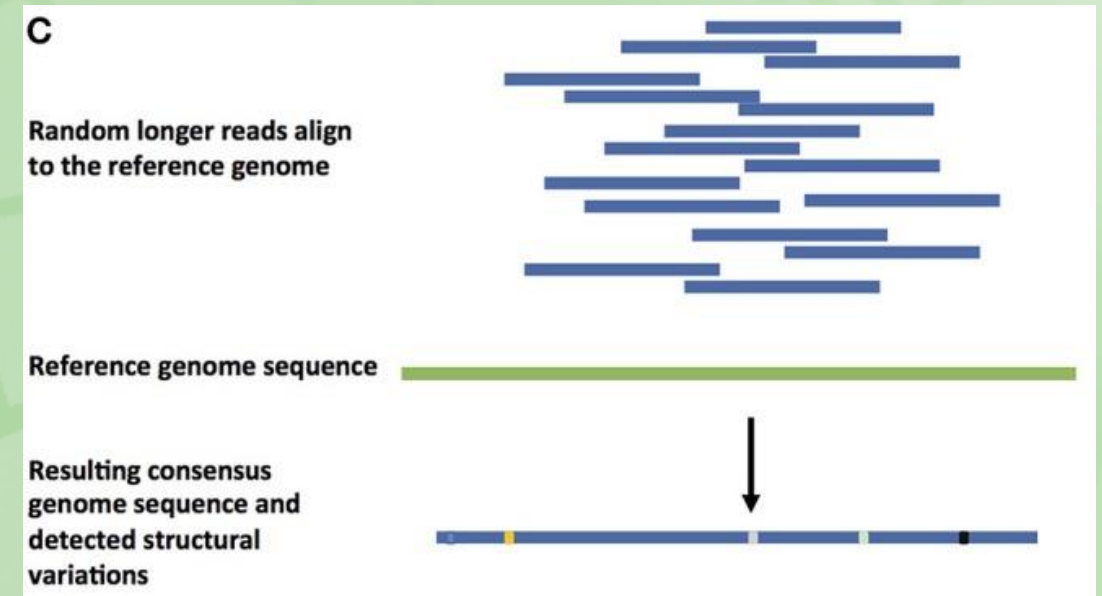
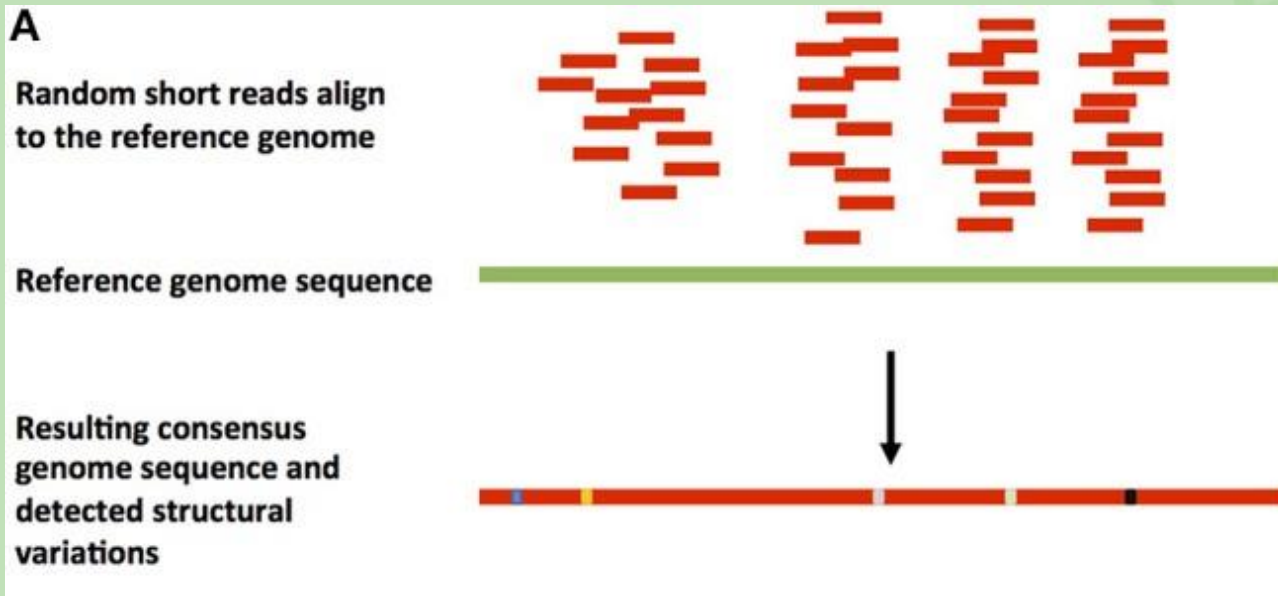
- Utilização dos reads (fragmentos) e informações sobre regiões de sobreposição para produzir sequências únicas e contínuas (contigs)
- Diferentes algoritmos e estratégias possíveis



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

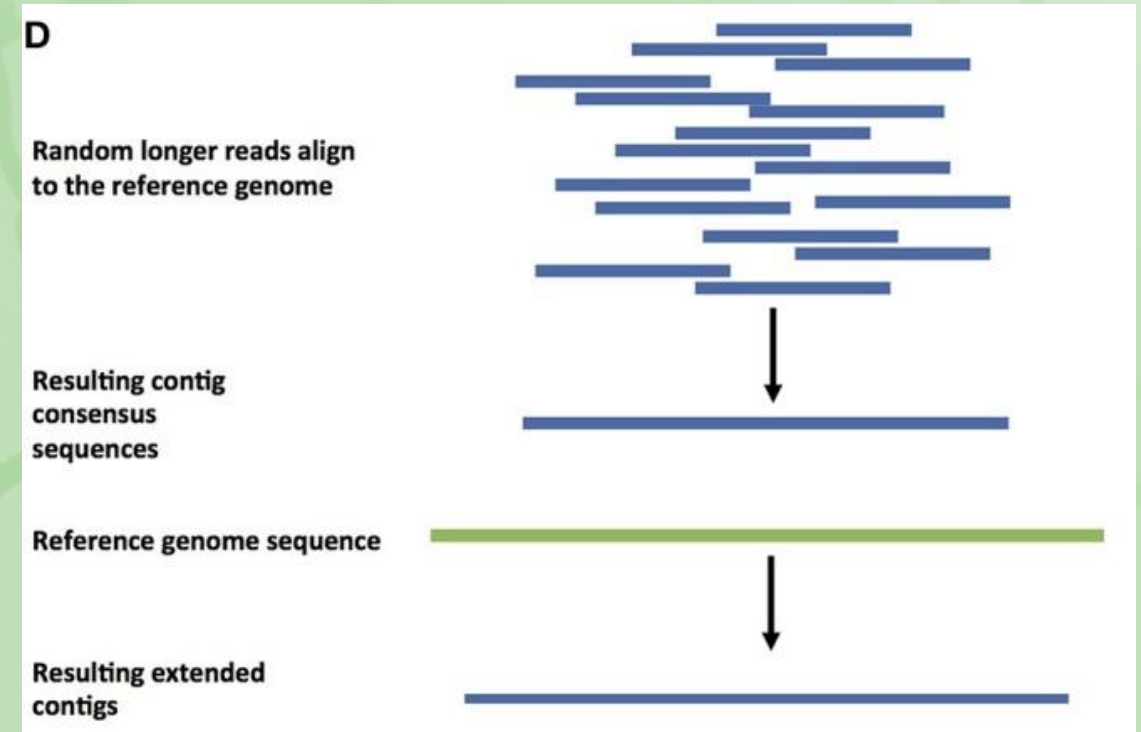
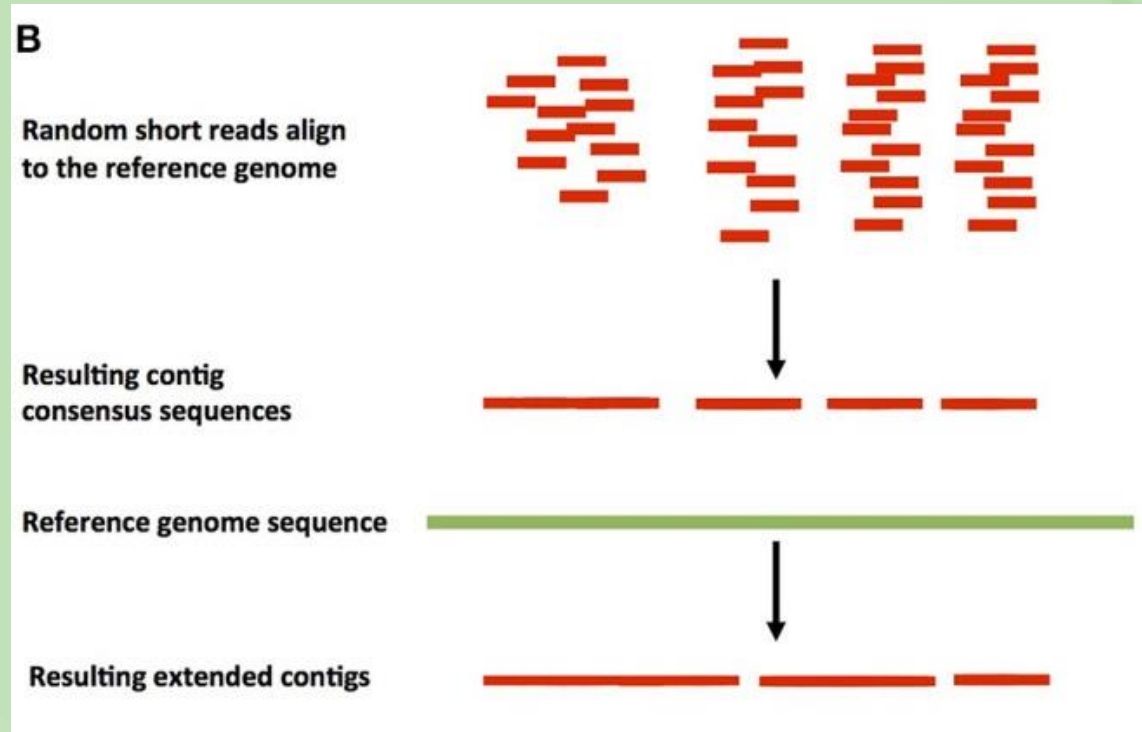
Montagem guiada por genoma de referência



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

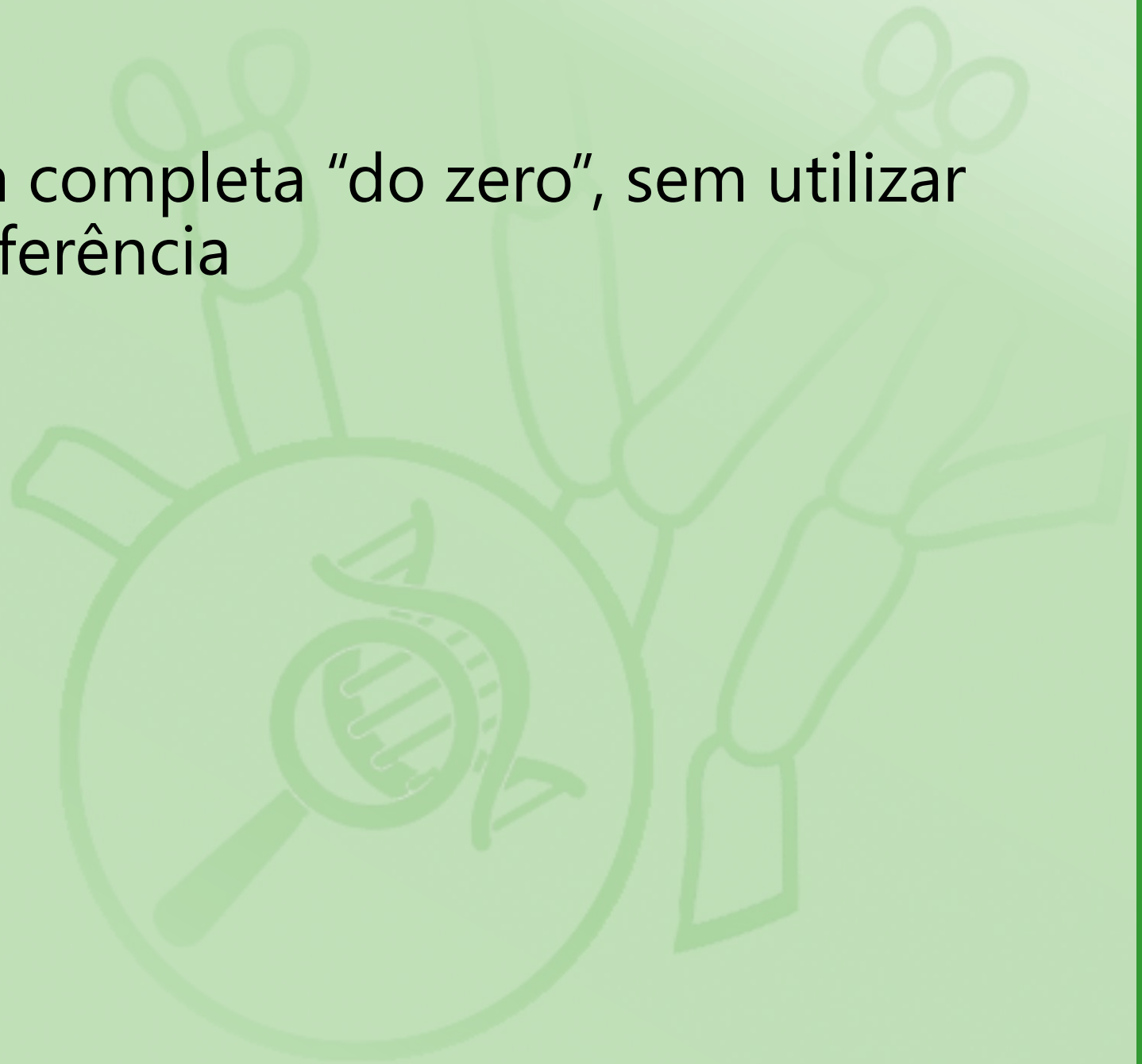


- Montagem inicial dos reads gerando contigs iniciais
- Alinhamento dos contigs à um genoma de referência já montado, e partir dos alinhamentos estender os contigs iniciais em contigs maiores
- Detecção de variações pontuais, como substituições ou rearranjos mais simples

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

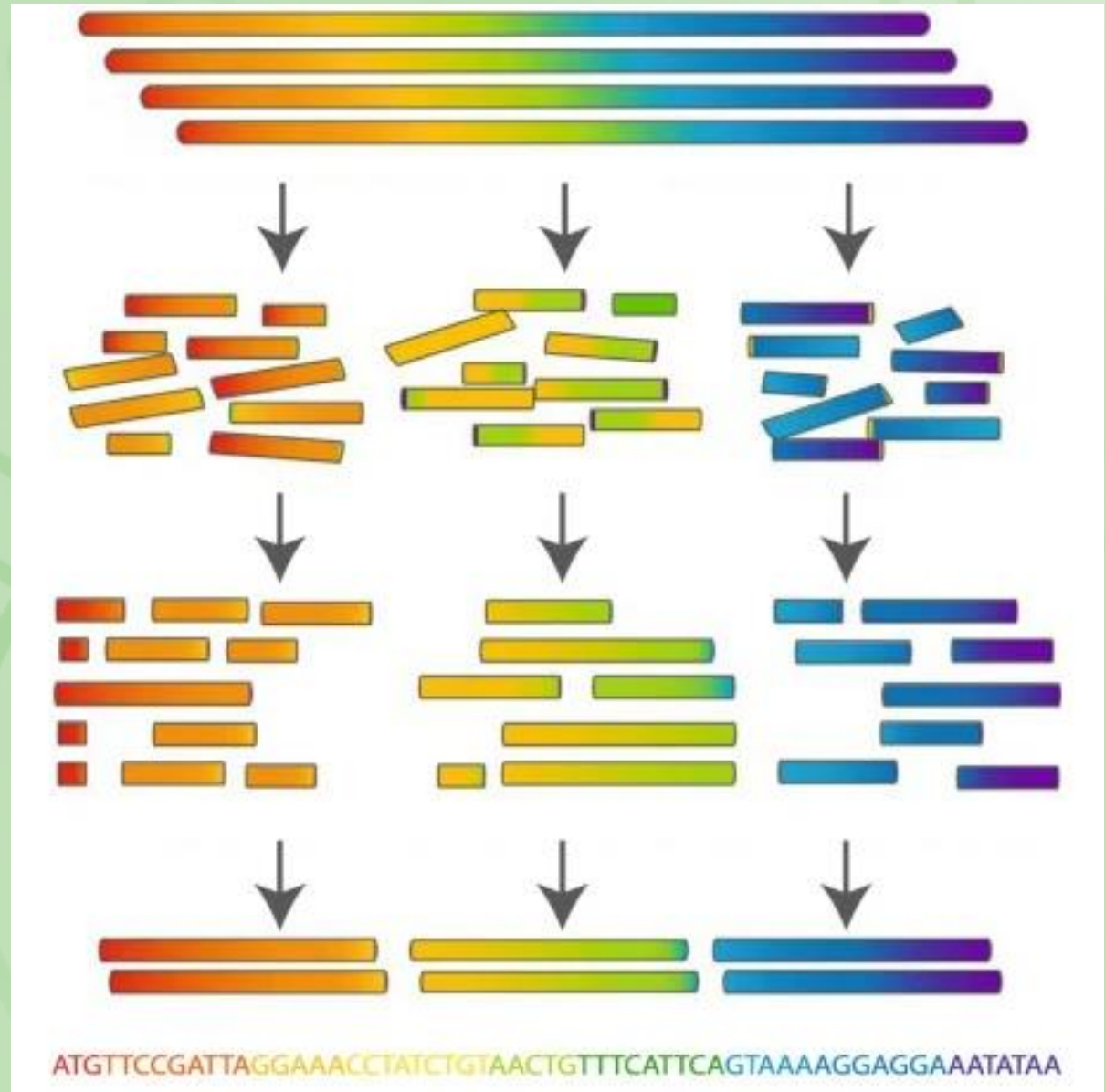
De novo

- Reconstruir a sequência completa “do zero”, sem utilizar outro genoma como referência
- Algoritmos
 - *Greedy*
 - Baseados em grafos



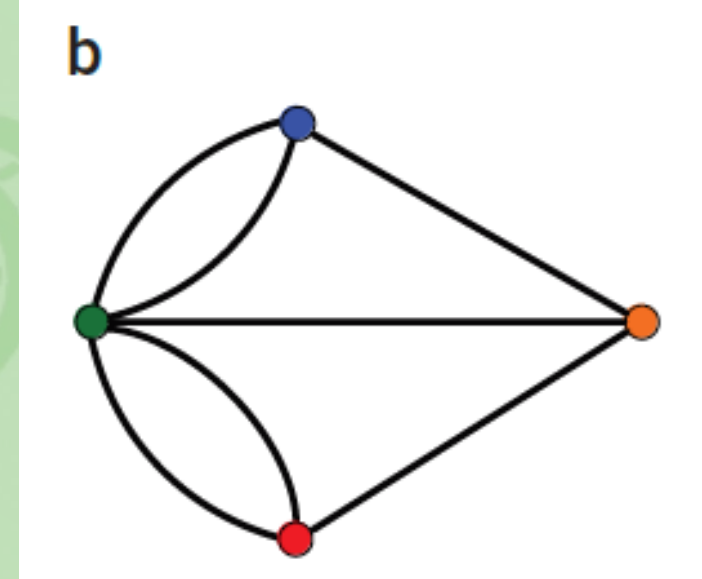
Algoritmo Greedy

- 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



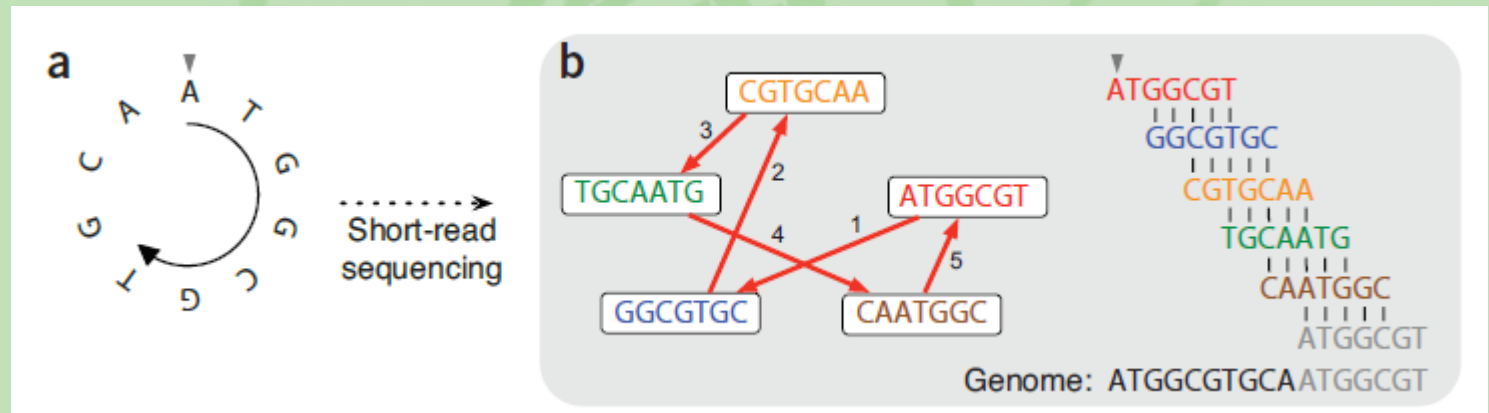
Baseado em grafos

- Problema das pontes de Königsberg (Kaliningrad, Rússia)
- Sete pontes sobre o rio Pregel unindo quatro partes da cidade
- É possível visitar todas as partes da cidade atravessando todas as sete pontes apenas uma vez e retornar ao local de partida?
- Aplicação de diferentes grafos em montagens de genomas



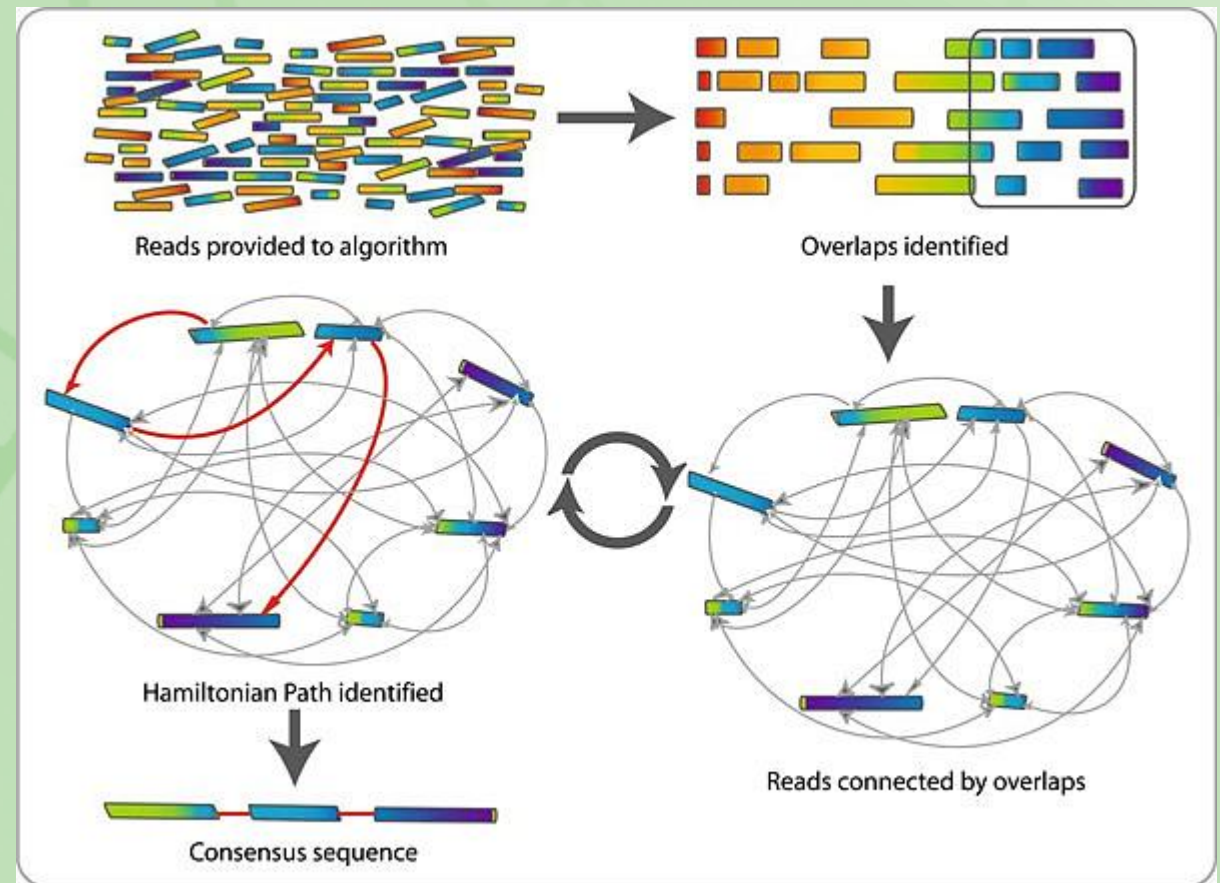
Pensando a montagem de um genoma por meio de grafos

- Cada read é um nó e cada sobreposição entre reads é representada pela seta vermelha, juntando dois nós
- Seguindo pela união entre os nós representada na figura, temos um caminho Hamiltoniano, passando por todos os nós apenas uma vez e terminando no nó inicial, e incluindo todos os reads



Overlap-layout-consensus (OLC)

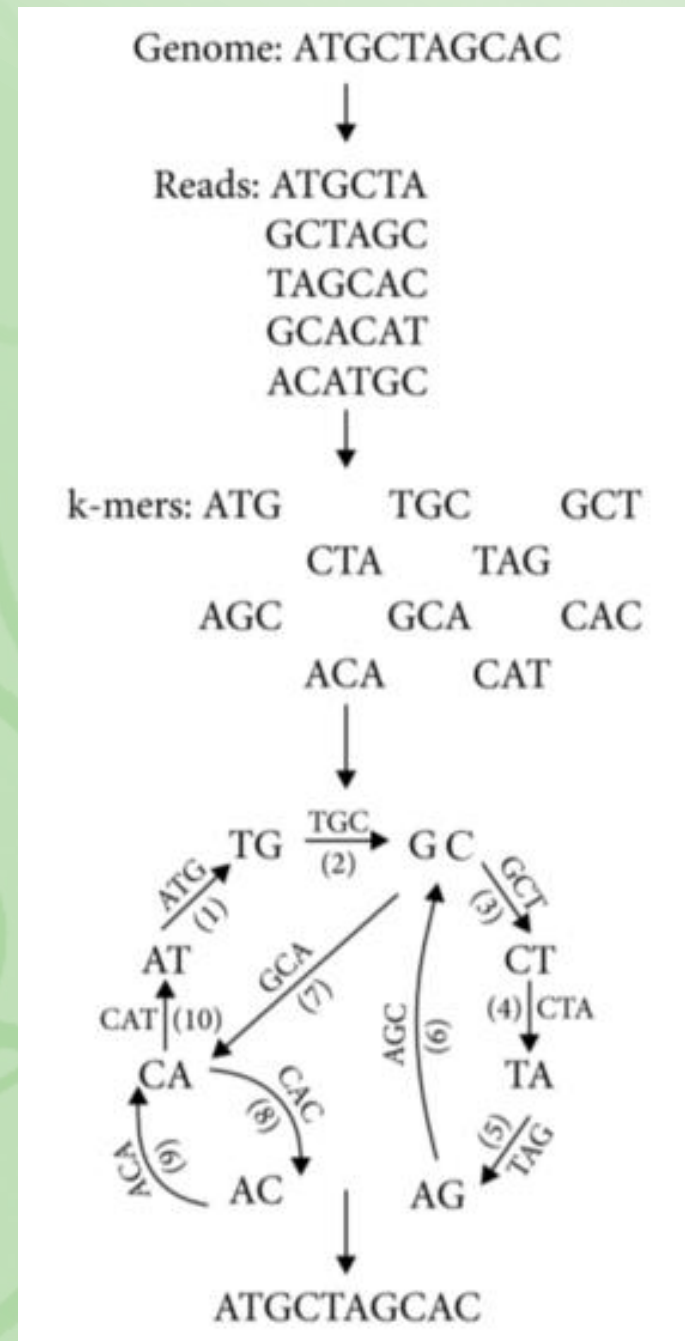
- Sobreposição entre os reads (similar ao algoritmo greedy)
- Grafo de sobreposições, em que cada read é um nó, conectados pelas sobreposições
- Encontrar o caminho passando **por todos os nós** para gerar contigs
- O caminho ideal seria um caminho Hamiltoniano: cada nó seria visitado apenas uma vez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



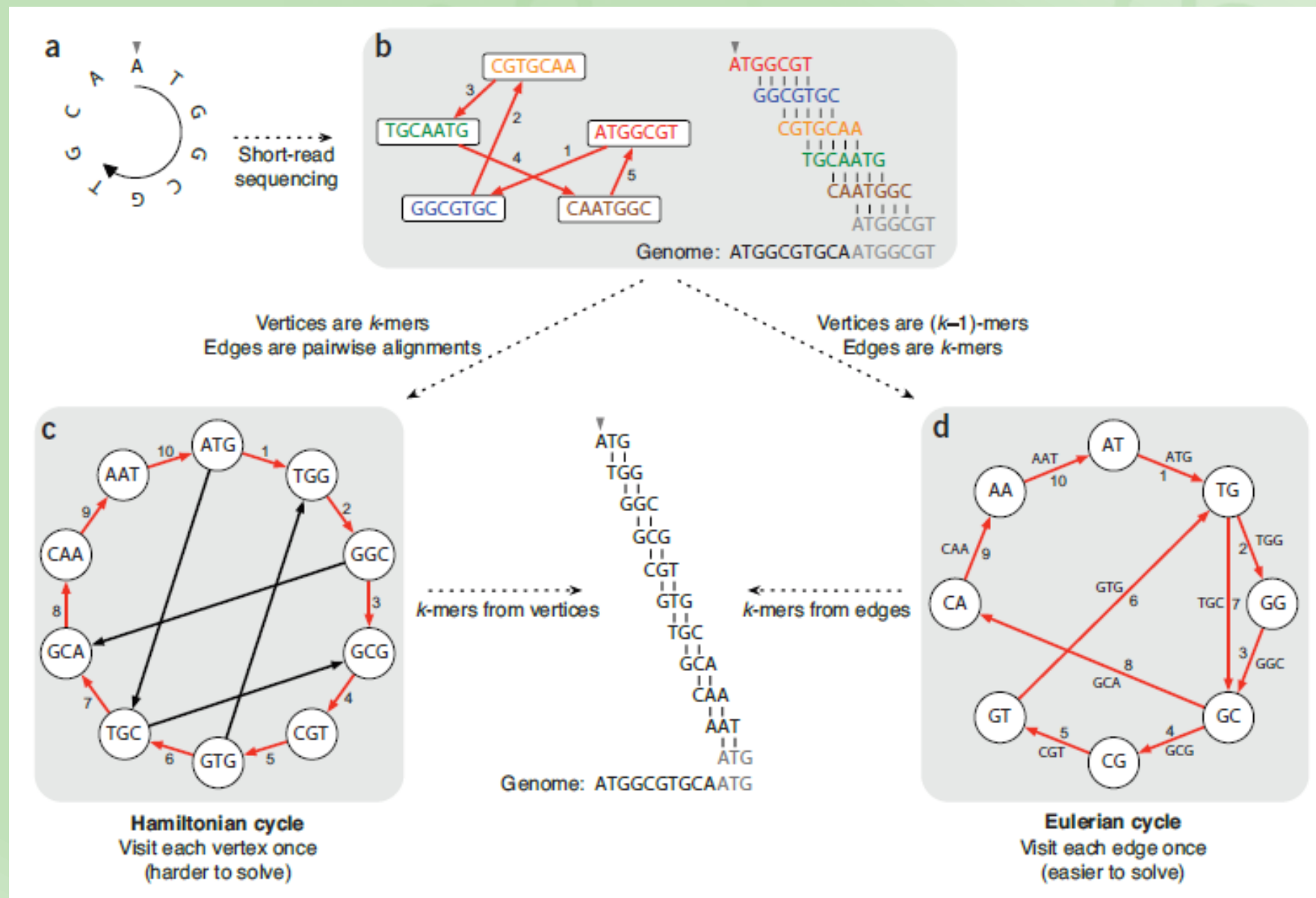
Adaptado de:
COMMINS et al. 2009. **Biological Procedures Online**. DOI: [10.1007/s12575-009-9004-1](https://doi.org/10.1007/s12575-009-9004-1)

Grafos De Brujin

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



Comparativo

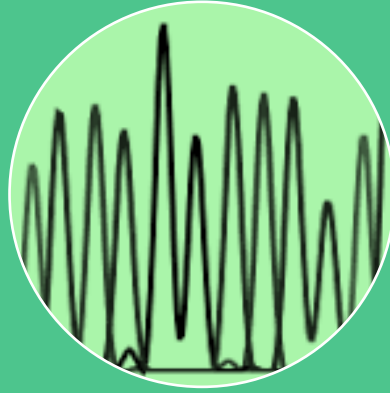


Como avaliar uma montagem?



Contiguidade

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



Análise de bases

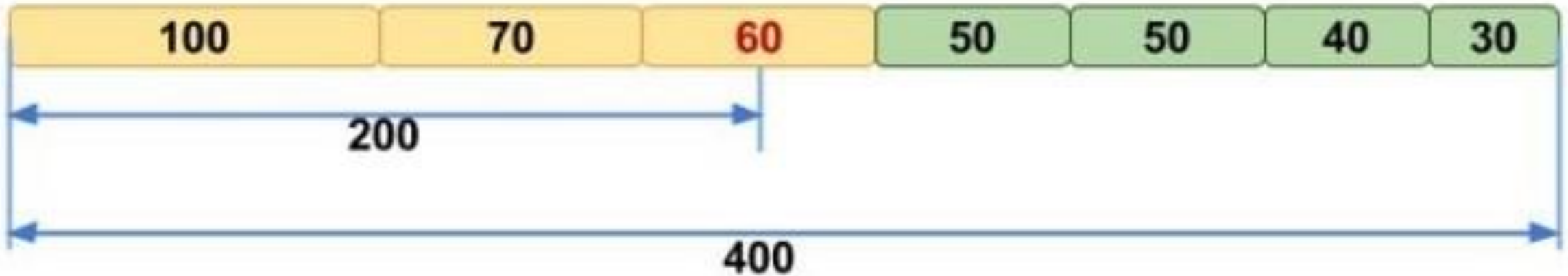
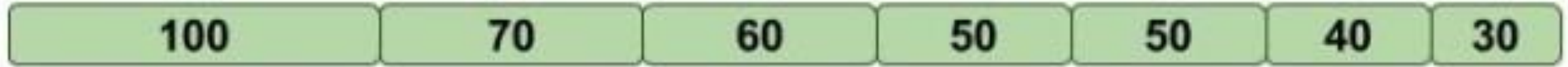
- Cobertura
- Conteúdo GC



Análise de conteúdo

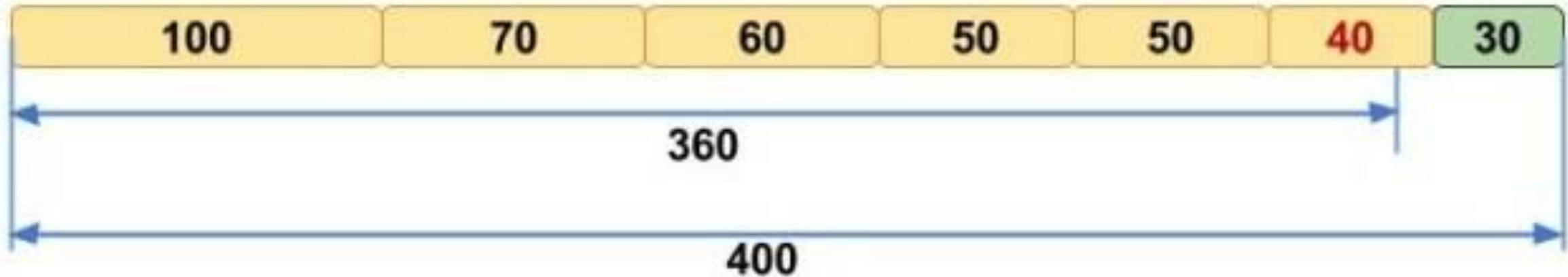
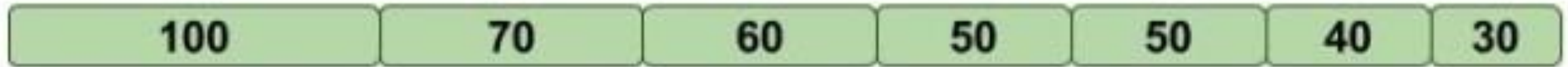
- Presença de telômeros
- Presença de genes conservados
- Comparação com genoma de referência
- Detecção de contaminantes pela distribuição do conteúdo GC
- Detecção de contaminantes por similaridade de sequência

Contiguidade – N50



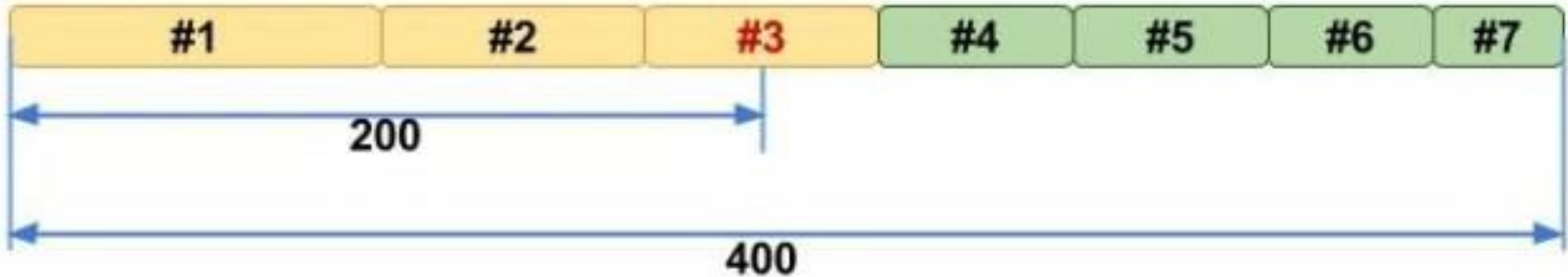
- N50: metade da montagem (50%) é representada por contigs/scaffolds com um comprimento igual ou maior que 60Kb

Contiguidade – N90



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

Contiguidade – L50



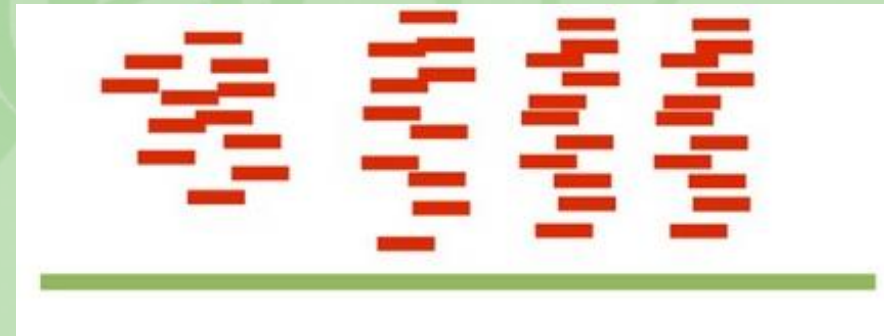
- L50: metade da montagem está presente em 3 contigs/scaffolds

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

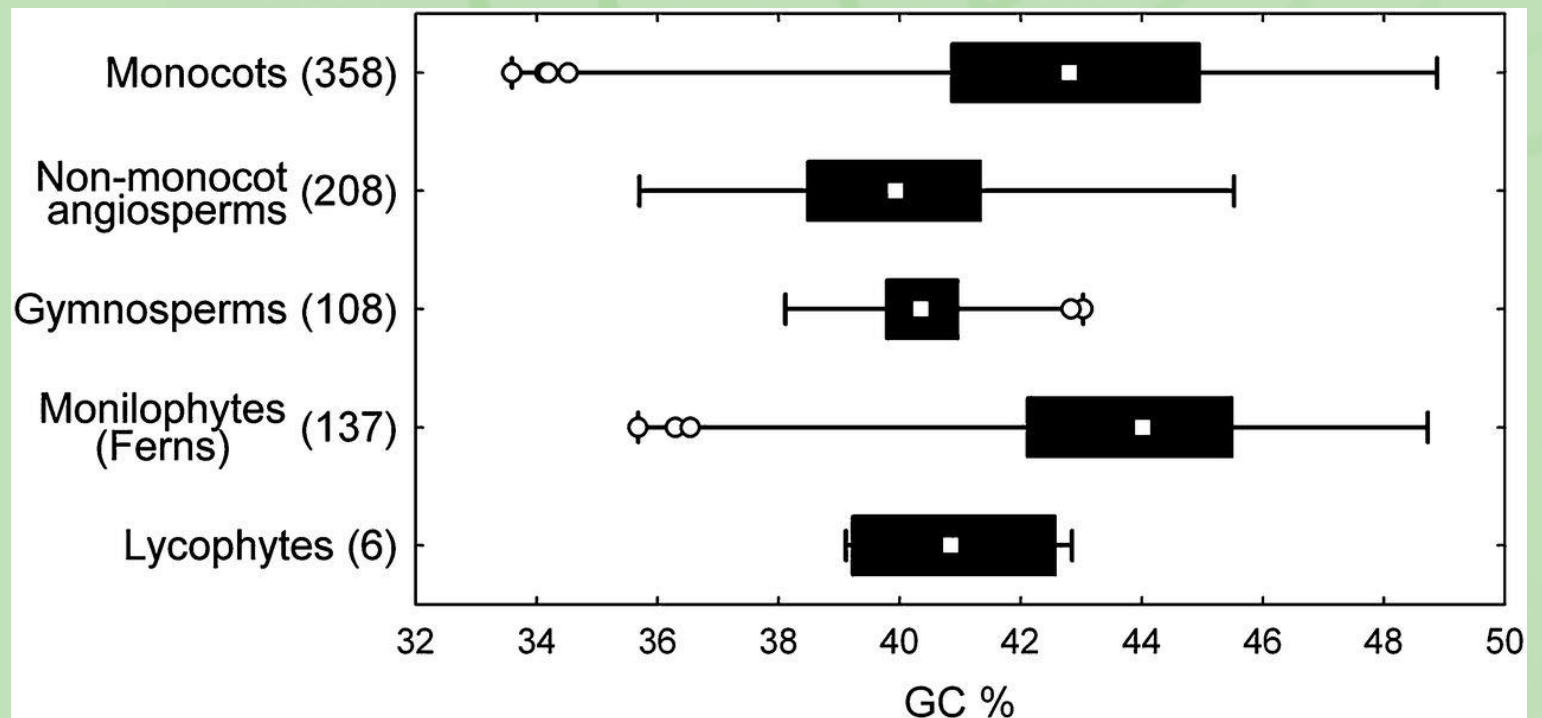
Análise de bases - cobertura

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



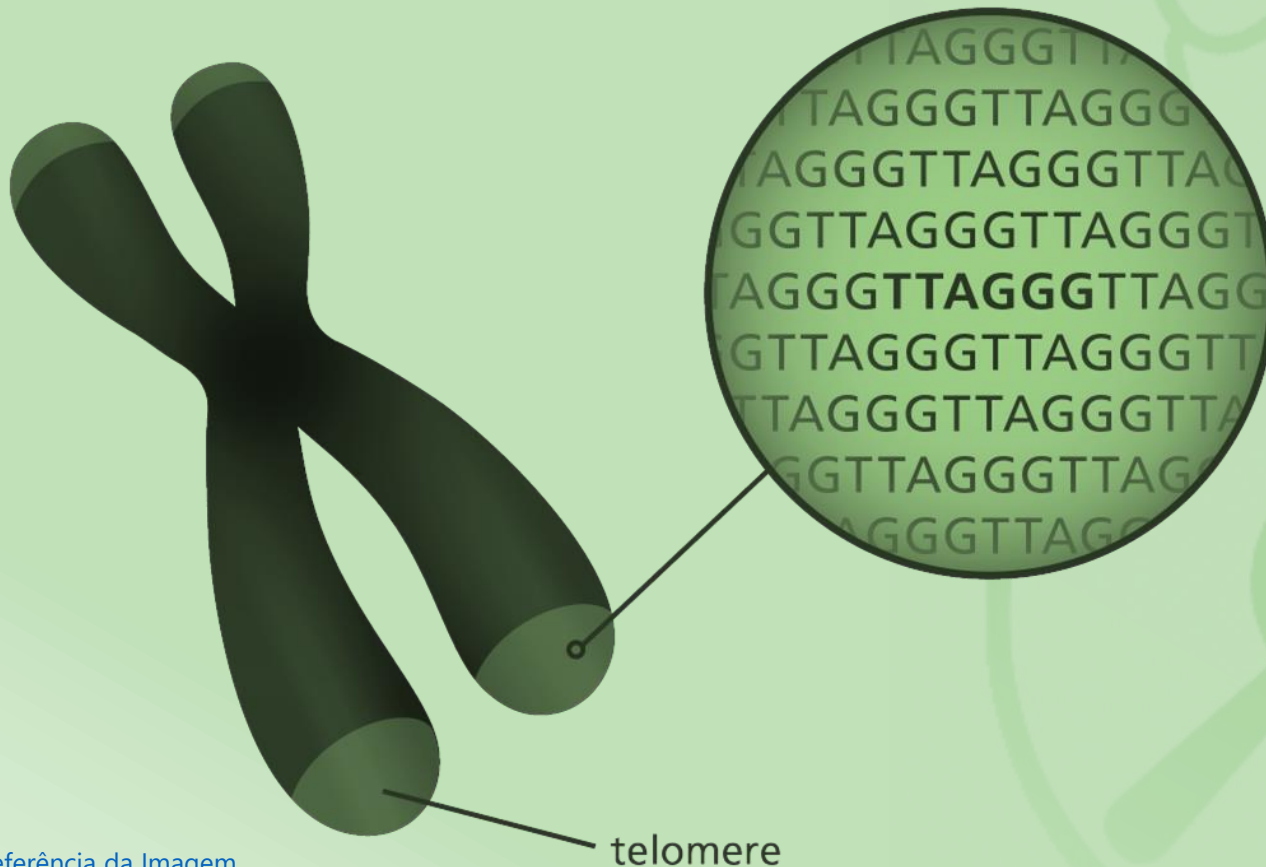
Análise de bases – Conteúdo GC

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



Análise de conteúdo - Telômeros

Chromosome

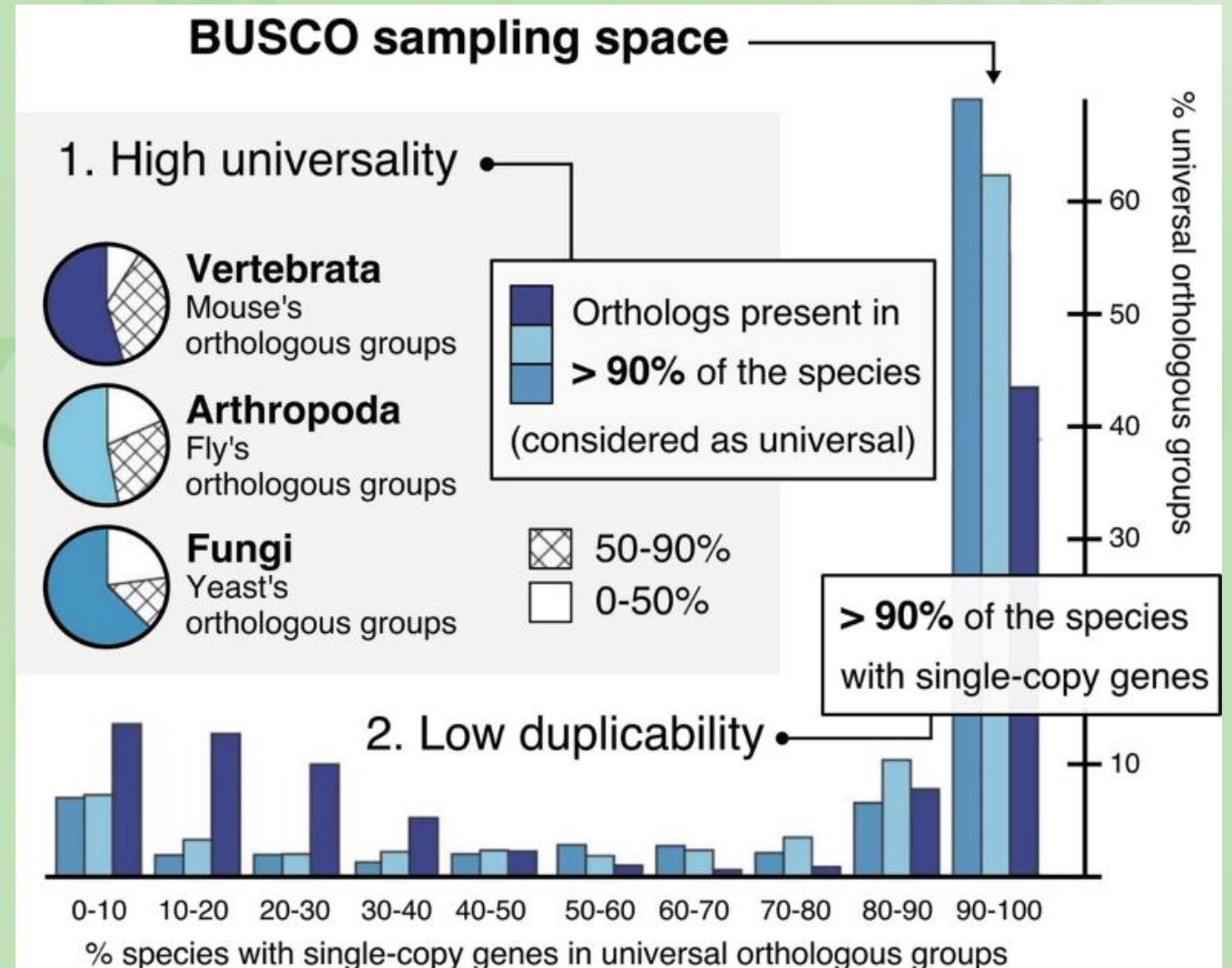


Referência da Imagem

- 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

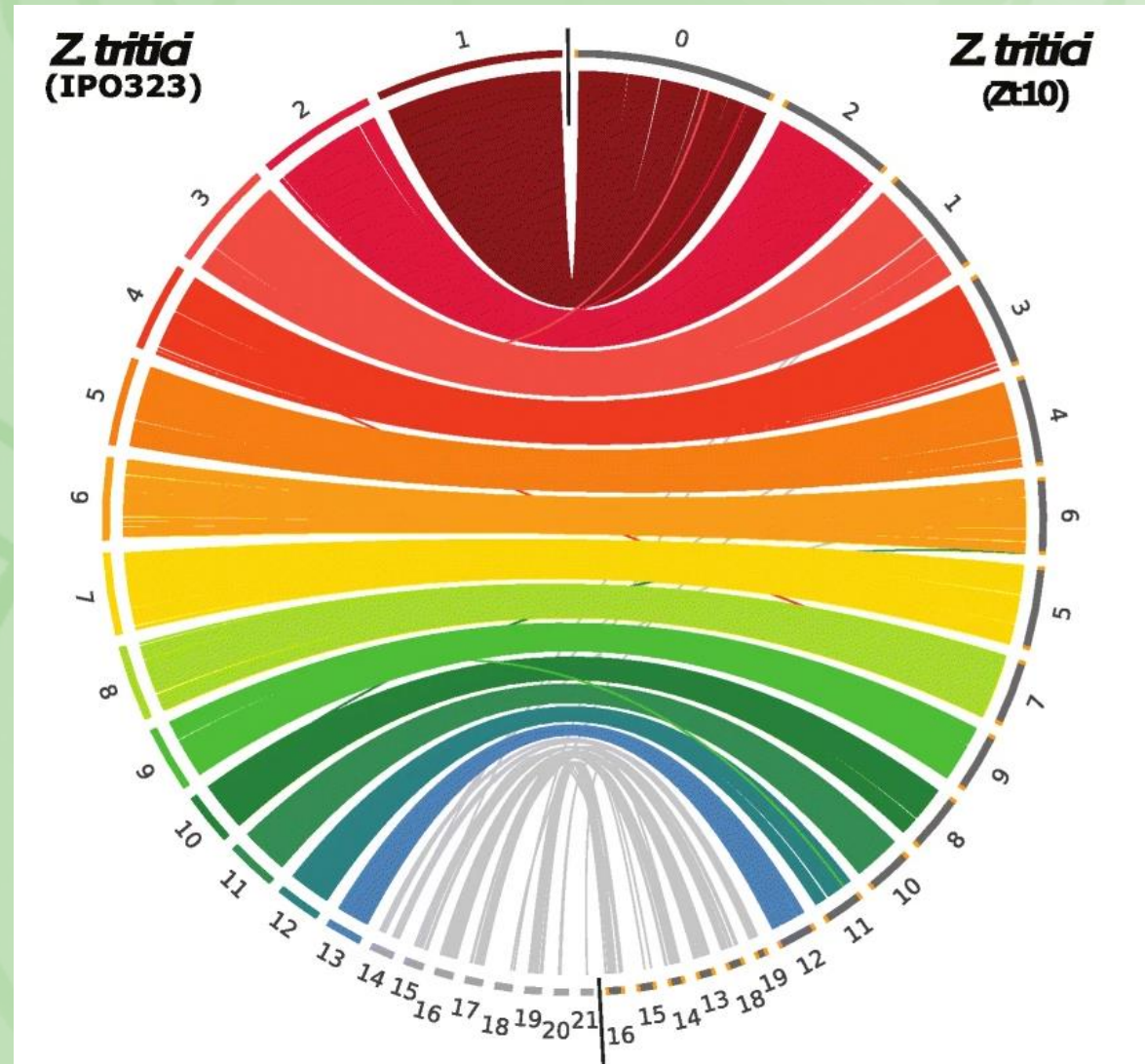
Análise de conteúdo – Genes conservados

- 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

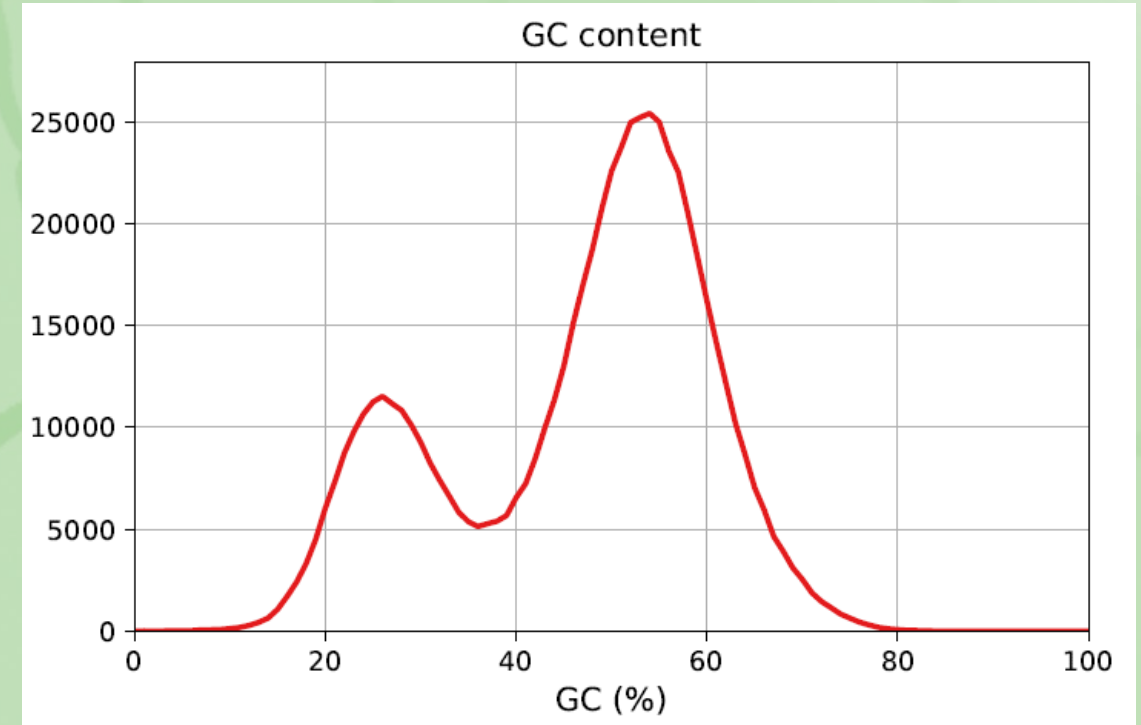
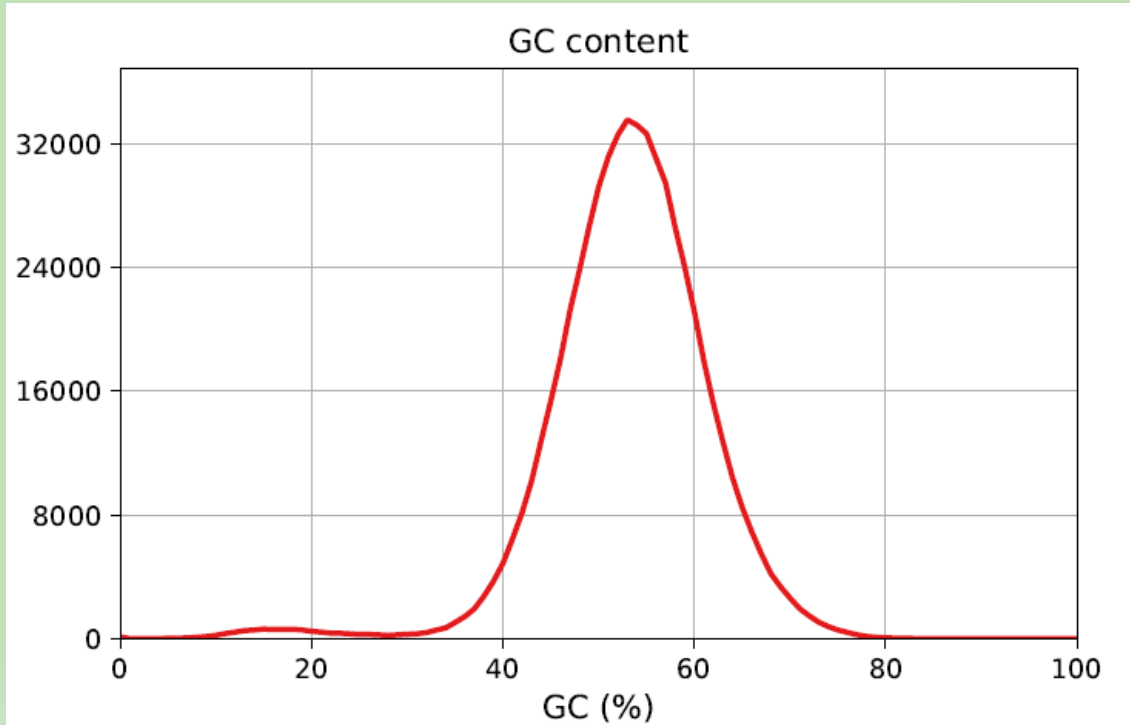


Análise de conteúdo – Comparação com genoma de referência

- 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?

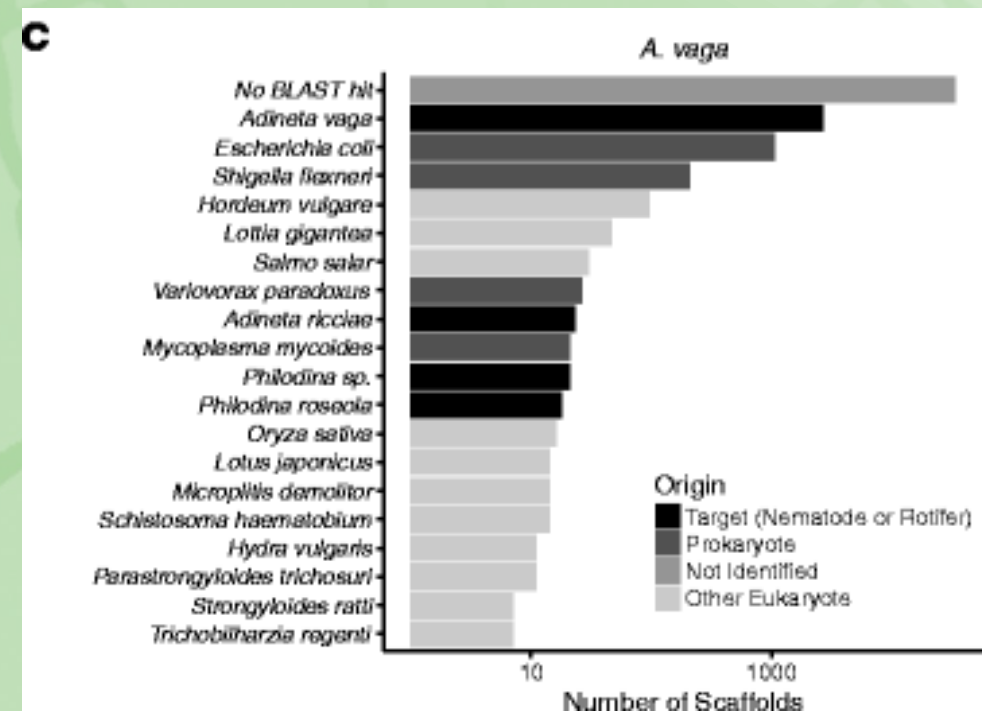
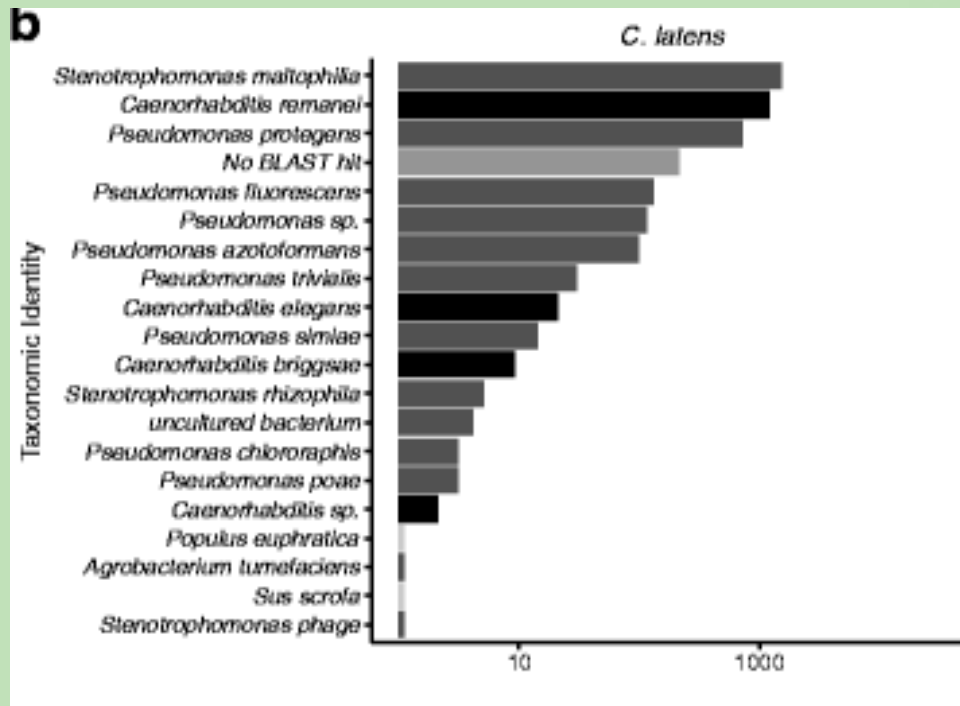


Análise de conteúdo – Contaminantes (Distribuição do conteúdo GC)



- Quantos picos são observados na distribuição de conteúdo GC?
- Mitocôndria 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)

