Bioinformática e o sequenciamento de genomas

Waldeyr Mendes Cordeiro da Silva

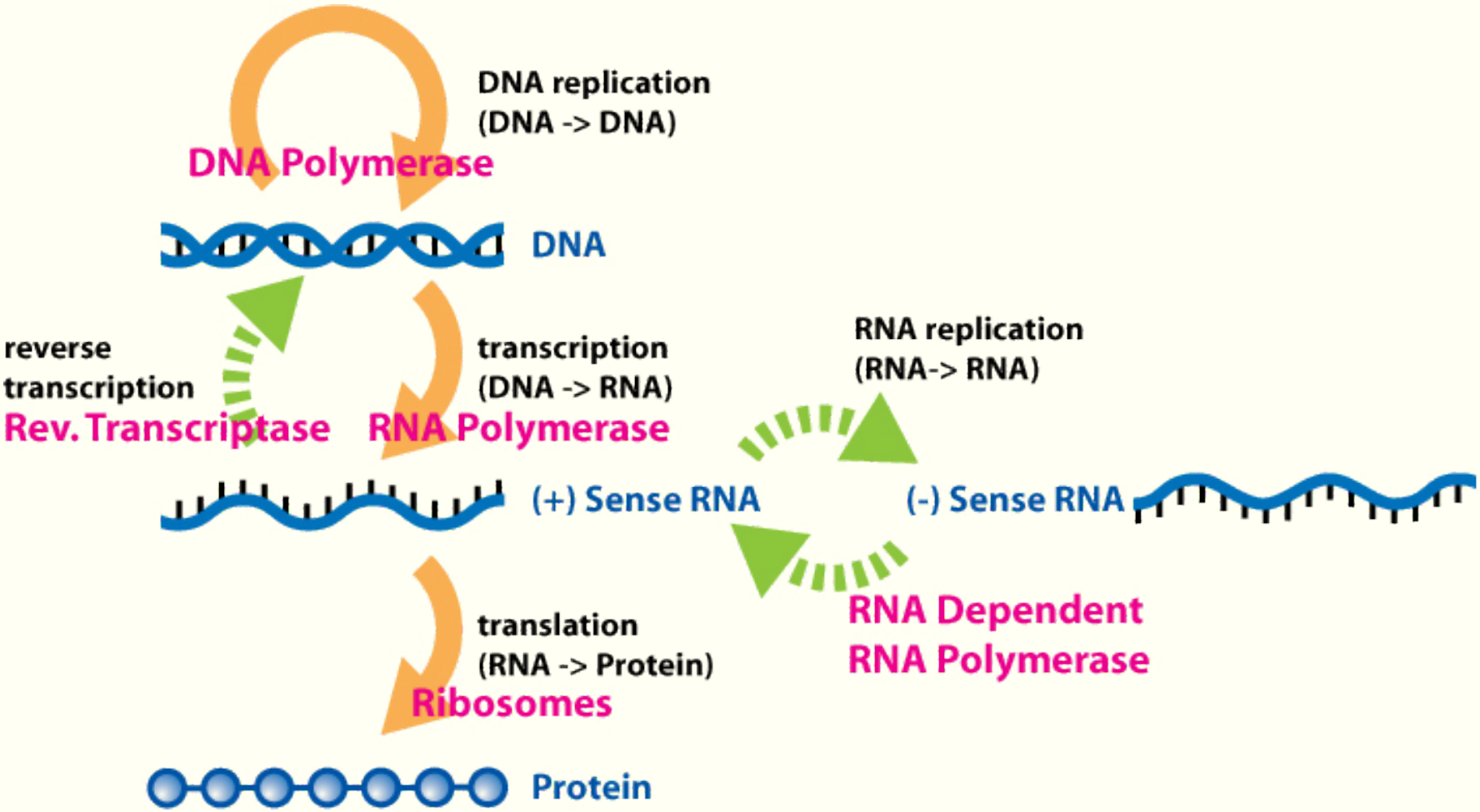
Uma visão geral

Agenda

1. Conceitos Básicos de Biologia Molecular
   - Aspectos biológicos do curso
2. Alinhamentos de Sequências
   - Tipos de alinhamentos
3. Dados de sequenciamento de alto desempenho
   - Formatos de arquivos
   - Filtragem e montagem de fragmentos
     - Controle de qualidade
     - Montagem *de novo*
   - Anotação
     - Significado biológico das sequências montadas
4. Prática (Genoma Sars-Cov-2)

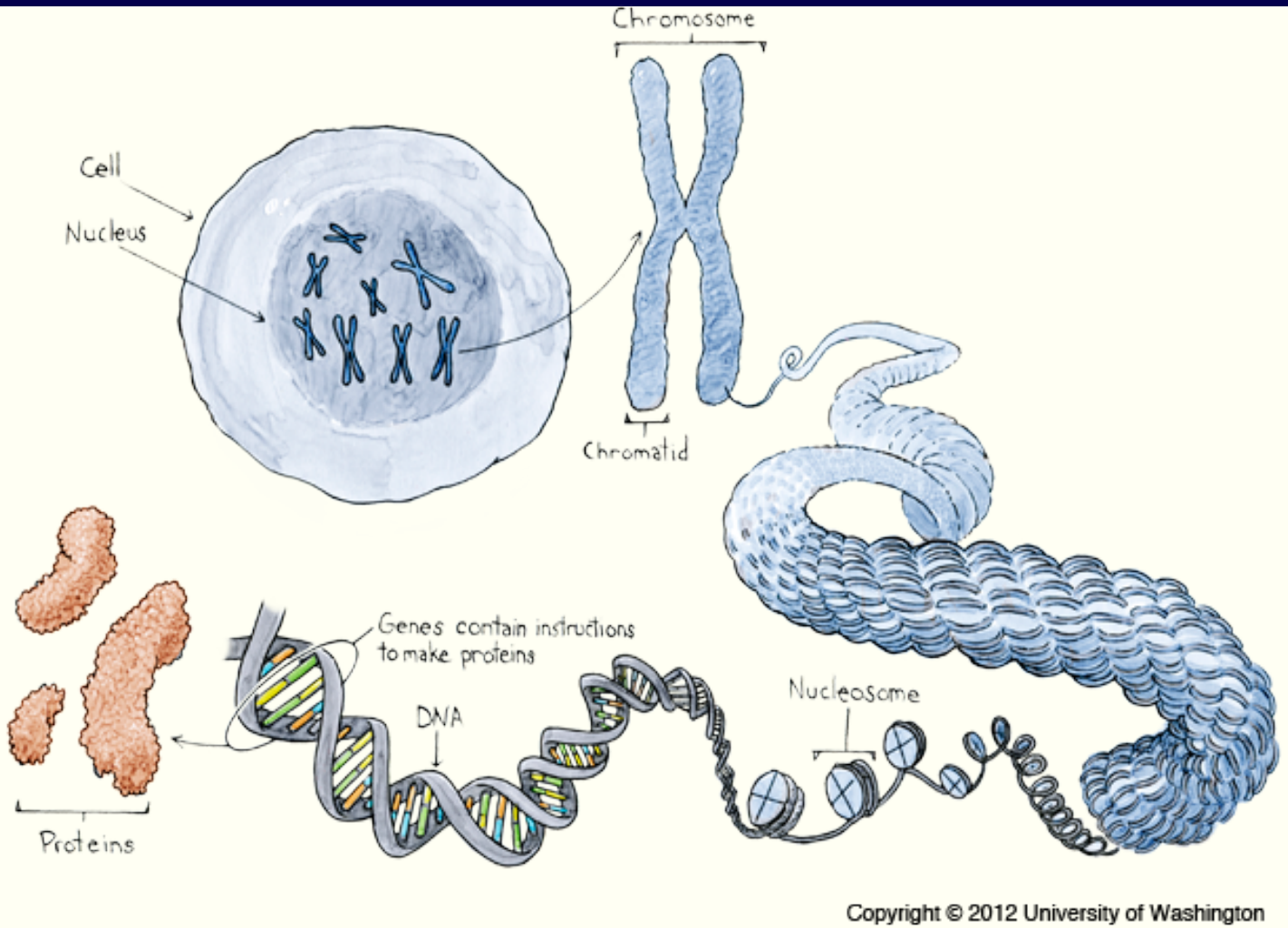
Conceitos Básicos de Biologia Molecular

Dogma



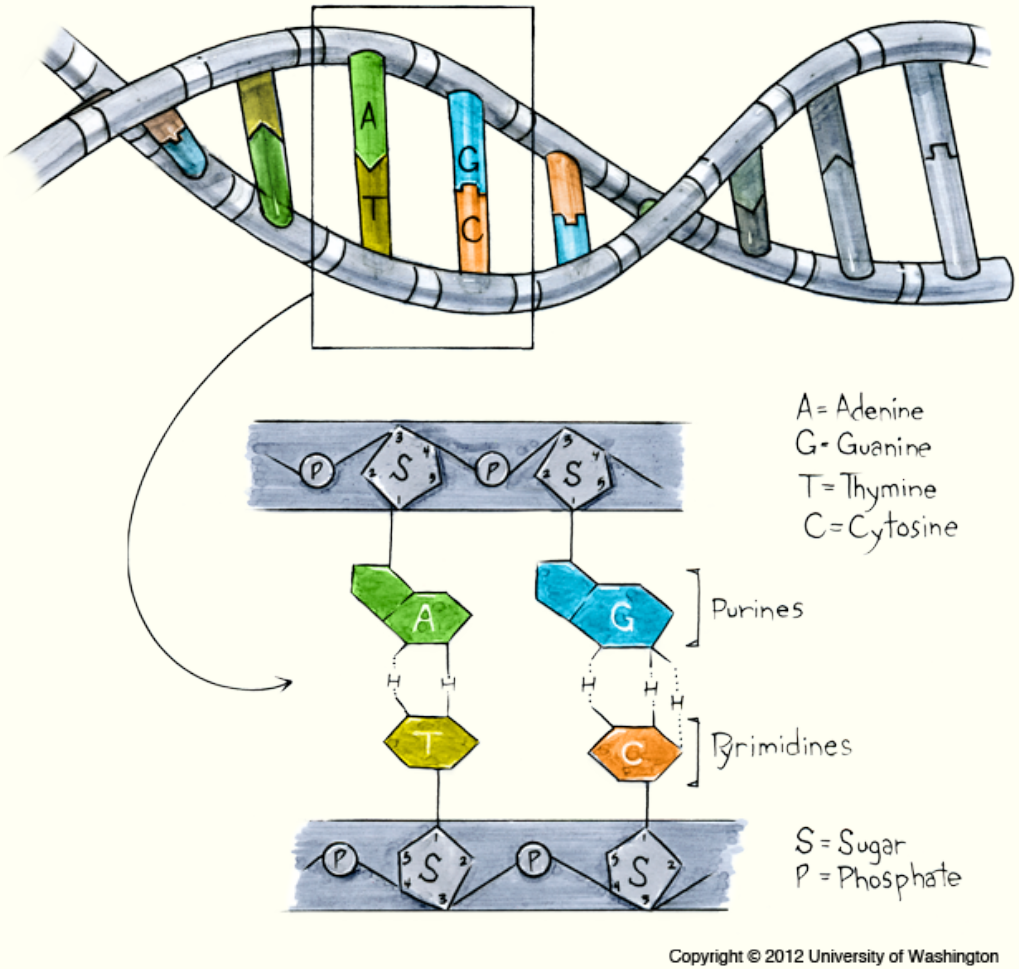
Fonte: https://en.wikipedia.org/wiki/Central_dogma_of_molecular_biology

DNA



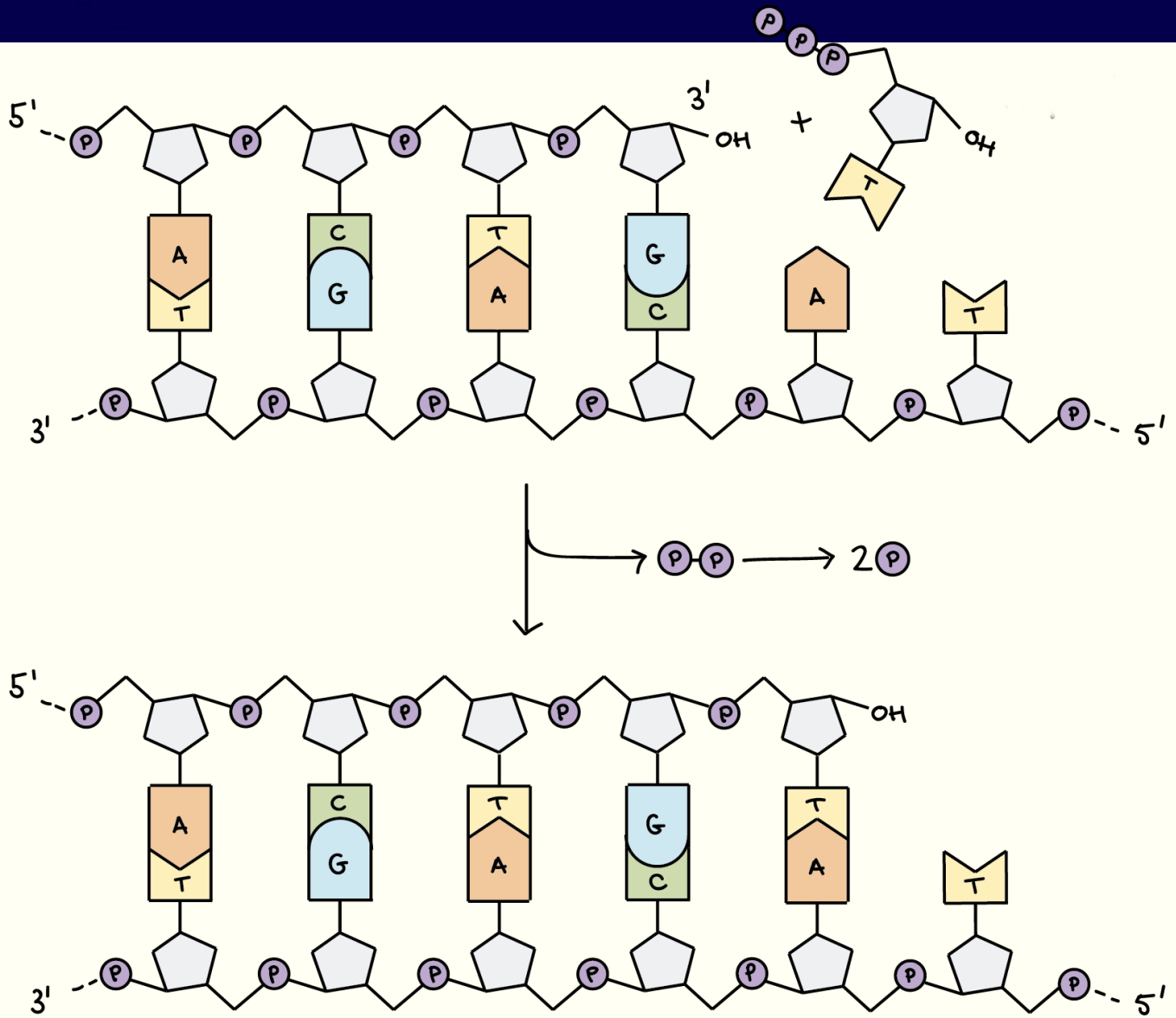
Fonte: https://www.my46.org/intro/what-is-dna

DNA



Fonte: https://www.my46.org/intro/what-is-dna

DNA



Fonte: https://www.khanacademy.org

Sequenciamento de DNA

- Obter string(s) representando as moléculas que compõem o DNA
- Ainda não é possível sequenciar toda a molécula diretamente
- Sequenciar pedaços da molécula, começando em alguma posição na direção 5' $\rightarrow$ 3'
- Fragmento (*read*): substring de uma das fitas da molécula alvo de DNA
- Não sabemos:
  - A que fita pertence
  - A posição relativa ao início da fita

Alinhamentos de Sequências

Alinhamentos de Sequências

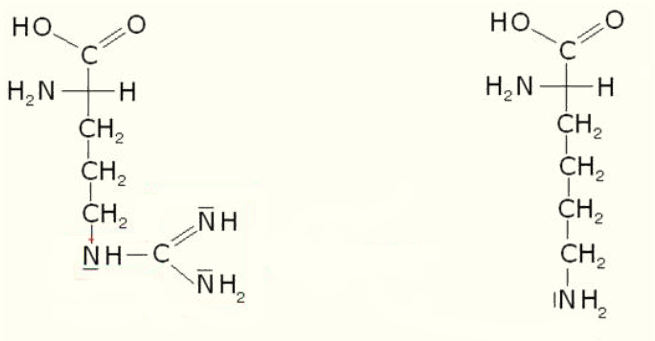
Posicionamento das sequências, preservando a ordem dos nucleotídeos ou aminoácidos e indicando as posições em que as sequências são iguais ou diferentes

- Ferramenta básica da Bioinformática
- Alfabeto
  - DNA/RNA - 4 nucleotídeos (ACGT/ACGU)
  - Proteínas - 20 aminoácidos (A, R, N, D, E, C, G, Q, H, I, L, K, M, F, P, S, Y, T, W, V)
- Interesse no alinhamento ótimo: o máximo de similaridade e o mínimo de diferenças

Alinhamentos de Sequências

- Identidade $\rightarrow$ Porcentagem de aminoácidos (ou nucleotídeos) com um *match* direto no alinhamento
- Similaridade $\rightarrow$ Porcentagem de *matches* idênticos e similares (substituição conservativa)

  Exemplo: arginina $\leftrightarrow$ lisina



- Homologia $\rightarrow$ Similaridade entre sequências que dividem ancestral comum

Alinhamento

Tipos de alinhamento

Quanto à quantidade de entradas

a) Pairwise - pareamento de 2 sequências
b) Alinhamento múltiplo - múltiplas sequências

Quanto à estratégia de alinhamento

a) Global
b) Local

Quanto ao tipo de entrada

DNA x RNA x Proteína

Alinhamento

Exemplo

R O S A V E R M E L H A
A M O R O S O V E R M E

8% de identidade (1 em 12).

Alinhamento

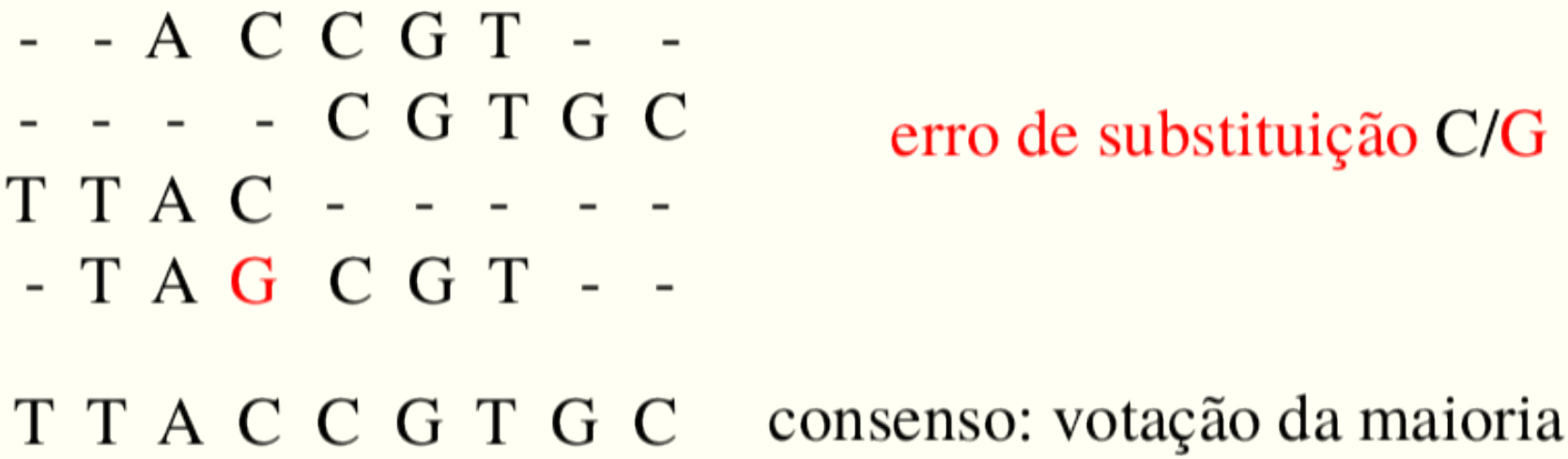
Exemplo

⊖ ⊖ ⊖ R O S A V E R M E L H A
A M O R O S O V E R M E ⊖ ⊖ ⊖

53% de identidade (8 em 15).

Alinhamento

Erros



Alinhamento de sequências

Erros

erro de inserção de A

consenso: votação da maioria

impressão: - não aparece

Alinhamento de sequências

Erros

```
- - A C C G T - -
- - - - C G T G C
T T A C - - - - -
- T A C - G T - -

T T A C C G T G C
```

erro de remoção de C
no último fragmento

consenso: votação da maioria

Alinhamento de sequências

Modelos de pontuação

- Substituições
- Gaps (inserções/deleções)
- Matriz de substituição

Alinhamento

Modelos de pontuação

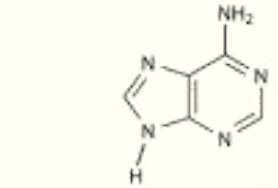
- Tomando as sequências: GACGGATTAG e GATCGGAATAG
- *Match* = +1
- *Mismatch* = -1
- *Gap* = -2

G	A	-	C	G	G	A	T	T	A	G
G	A	T	C	G	G	A	A	T	A	G
+1	+1	-2	+1	+1	+1	+1	-1	+1	+1	+1 = 6

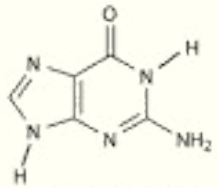
- Obs: Valores das penalidades podem ser escolhidos

Alinhamento

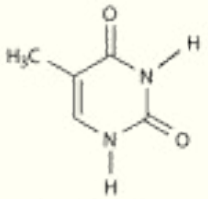
Matrizes de susbstituição



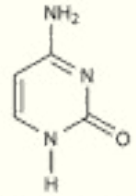
Adenina (A)



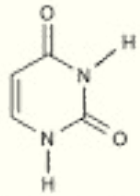
Guanina (G)



Timina (T)



Citosina (C)



Uracila (U)

	A	C	G	T
A	+20	+5	+10	+5
C	+5	+20	+5	+10
G	+10	+5	+20	+5
T	+5	+10	+5	+20

Dados de sequenciamento de alto desempenho

Formatos (FASTQ)

```
@SEQ_ID
TTCAACTCGTTAGTAAATATCAAACGATCAGTACCATTTTGGGGTTCAAAGTGACAGTTT
+
!'>>>>CCC'*((((***(***-+*'')+))%%%++))**55CCF>>%%%%).1CCCC65
```

Exemplo Illumina: **@HWUSI-EAS100R:6:73:941:1973#0/1**

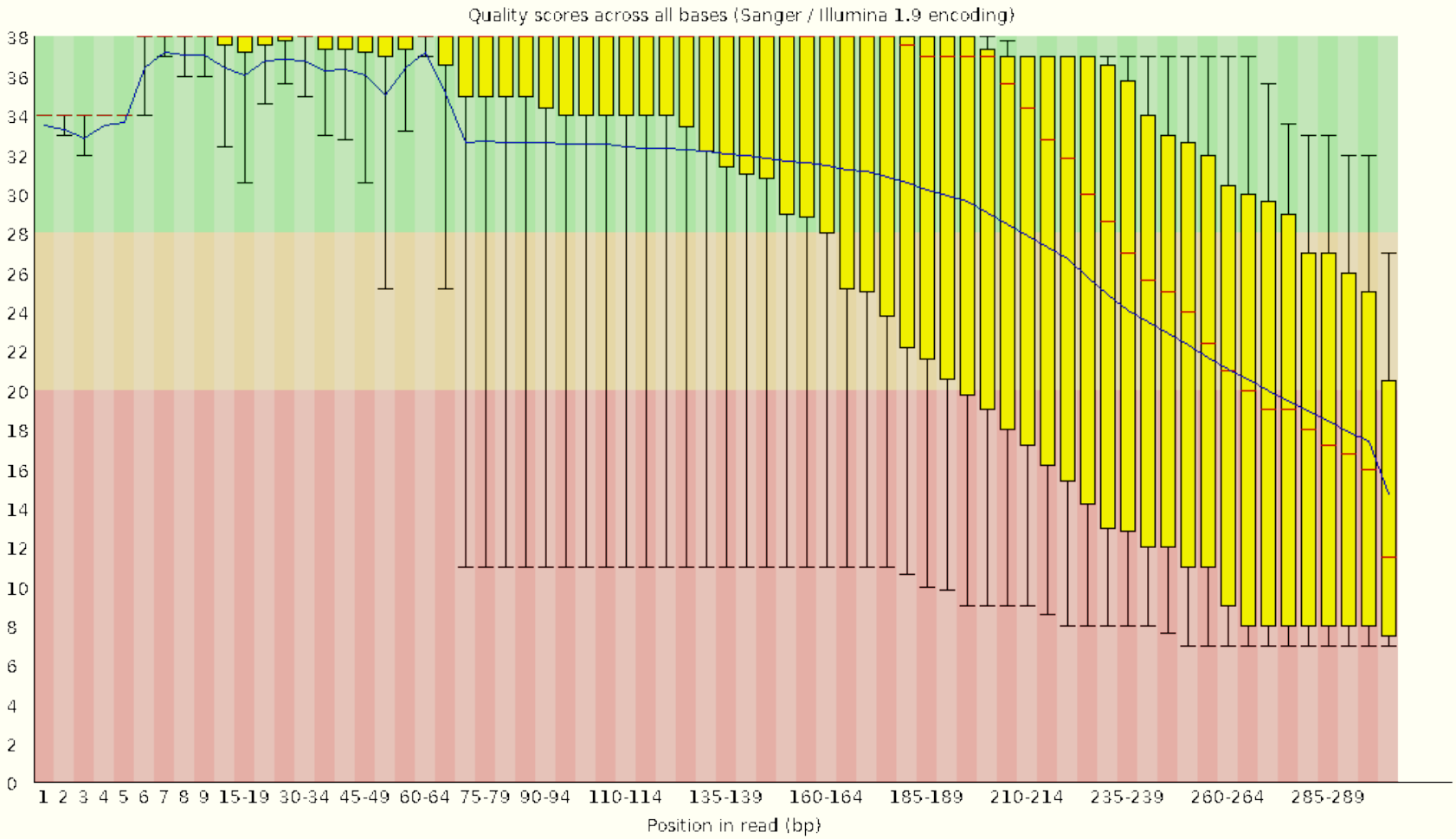
- HSWUSI-EAS100R $\rightarrow$ Unique instrument name
- 6 $\rightarrow$ Flowcell lane
- 73 $\rightarrow$ Tile number within the flow cell lane
- 941 $\rightarrow$ x-coordinate of the cluster within the tile
- 1973 $\rightarrow$ y-coordinate of cluster within the tile
- #0 $\rightarrow$ Index number for multiplexed sample
- /1 $\rightarrow$ Member of a pair

Formatos (FASTQ)

A qualidade (varia de 33 a 126) de cada nucleotídeo sequenciado é representado pelo caractere correspondente da tabela ASCII.
Os valores *shifted down* para 0 a 93 por compatibilidade com a escala PHRED de qualidade (0 a 60)

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Formatos (FASTQ)



Formatos (FASTA)

```
>gi|13959657|sp|Q9PTU8|VSP3_BOTJA Venom serine proteinase A precursor
MVLIRVIANLLILQLSNAQKSSELVIGGDECNITEHRFLVEIFNSSGLFCGGTLIDQEWVLSAAHCDMRN
MRIYLGVHNEGVQHADQQRRFAREKFFCLSSRNYTKWDKDIMLIRLNRPVNNSEHIAPLSLPSNPPSVGS
VCRIMGWGTITSPNATFPDVPHCANINLFNYTVCRGAHAGLPATSRTLCAGVLQGGIDTCGGDSGGPLIC
NGTFQGIVSWGGHPCAQPGEPALYTKVFDYLPWIQSIIAGNTTATCPP
```

1. Cabeçalho
   - GenBank/EMBL $\rightarrow$ gi|gi_number|*|accession.version|locus
   - NCBI refseq $\rightarrow$ ref|accession|locus
   - PRF Protein Research Foundation $\rightarrow$ pir|entry
   - SWISS-PROT $\rightarrow$ sp|accesion|locus
   - PDB Protein Data Bank $\rightarrow$ pdb|entry|chain
2. Sequência
   - nucleotídeos ou aminoácidos

Formatos (FASTA)

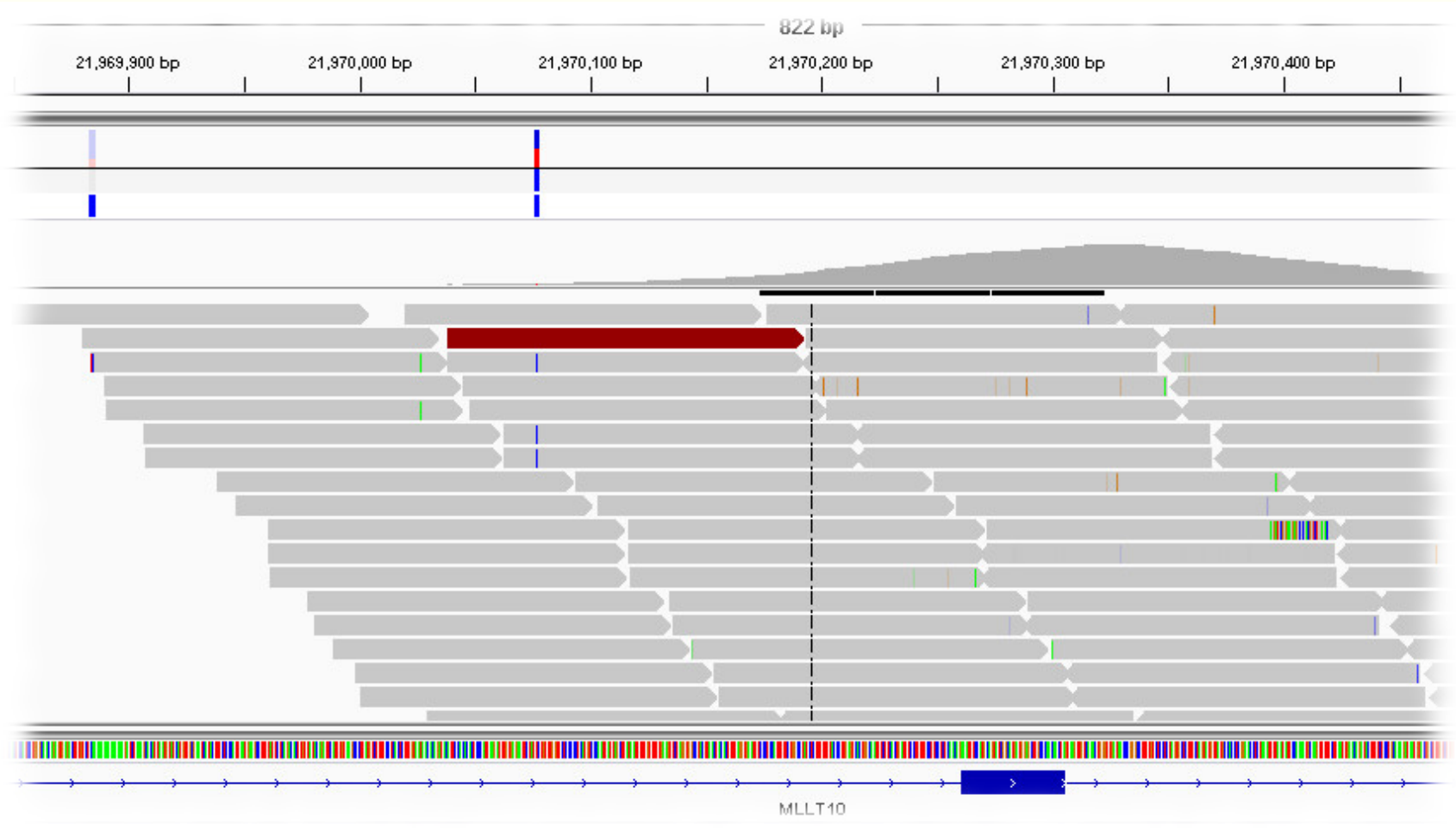
- .fasta, .fa $\rightarrow$ arquivo fasta genérico
- .fna $\rightarrow$ FASTA nucleotídeos
- .ffn $\rightarrow$ FASTA regiões codificadoras (nucleotídeos)
- .faa $\rightarrow$ FASTA aminoácidos
- .frn $\rightarrow$ FASTA RNA não codificador
- Multi-fasta $\rightarrow$ múltiplas sequências em um único arquivo

Formatos (SAM, BAM)

SAMTools fazem pós-processamento de alinhamentos de *reads*, as quais são sequências de DNA em formato FASTQ.

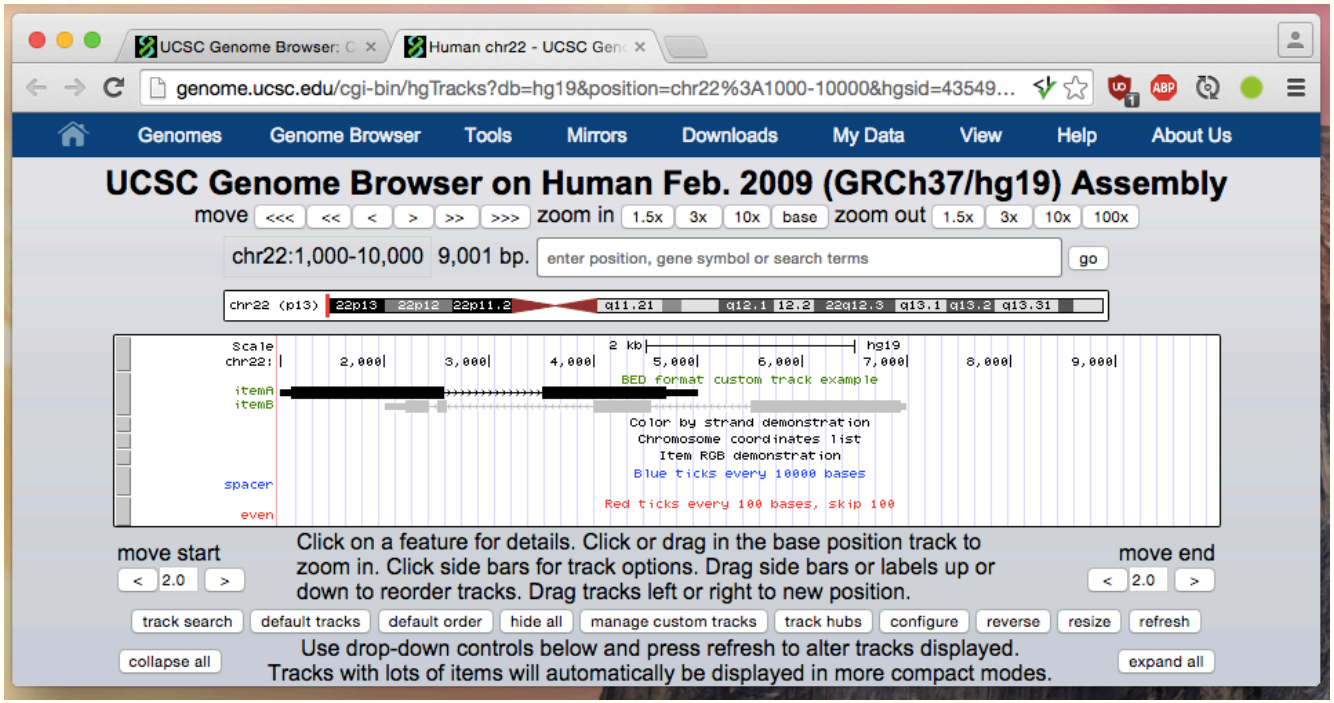
- SAM (Sequence Alignment/MAP) guarda o alinhamento das *reads* e pode ser lido por diversos softwares como o IGV (Integrated Genome Viewer).
- BAM (Binary Alignment/MAP) é uma versão comprimida de um alinhamento das *reads*. Pode ser obtido diretamente do alinhamento ou convertido a partir de um arquivo SAM.

Formatos (SAM, BAM)



Formatos (BED)

- BED é um arquivo organizado em colunas separadas por tabulação (tab) com anotações da sequência
- Pode ser aberto em um genome browser



Formatos (BED)

- Arquivos BED têm 12 colunas, 1-3 obrigatórias, 4-12 opcionais
  1. **chrom** $\rightarrow$ nome do cromossomo no qual a *feature* existe
  2. **start** $\rightarrow$ posição inicial na sequência
  3. **end** $\rightarrow$ posição final na sequência
  4. name $\rightarrow$ nome da *feature*
  5. score $\rightarrow$ 0 and 1000 (nível de cinza[1])
  6. strand $\rightarrow$ direção da fita "+" ou "-"
  7. thickStart $\rightarrow$ posição inicial onde a *feature* é desenhada
  8. thickEnd $\rightarrow$ posição final onde a *feature* é desenhada
  9. itemRgb $\rightarrow$ determina a cor dos dados
  10. blockCount $\rightarrow$ número de bloco (exons)
  11. blockSizes $\rightarrow$ lista de blocos separados por vírgula
  12. blockStarts $\rightarrow$ lista de posições iniciais dos blocos

1) Pode ser usado para outras medidas como p-value, up/down, ...

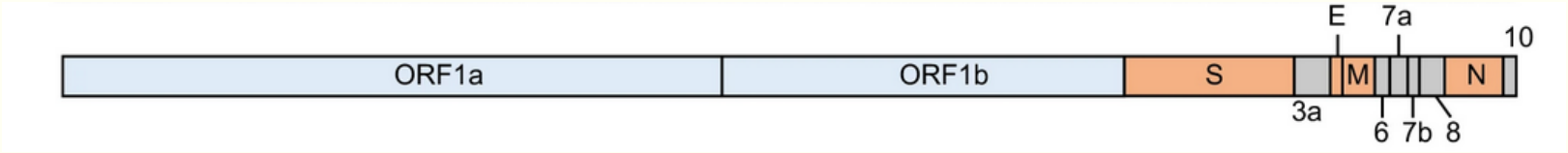
Formatos (GFF)

- GFF são similares aos BED e têm 9 colunas obrigatórias
  1. seqname $\rightarrow$ nome da sequência
  2. source $\rightarrow$ origem da *feature*
  3. feature $\rightarrow$ tipo de *feature*, equivalente ao campo *name* do BED
  4. start $\rightarrow$ posição inicial
  5. end $\rightarrow$ posição final
  6. score $\rightarrow$ assim como o arquivo BED permite níveis de valores representando a expressividade da anotação
  7. strand $\rightarrow$ direção da fita “+” ou “-”
  8. frame $\rightarrow$ frame da sequência codificadora: “0”,“1”,“2” ou “.”;
  9. attribute $\rightarrow$ muda conforme a versão do GFF (GFF1, GFF2, GFF3) e denota texto livre com algum significado biológico

Prática

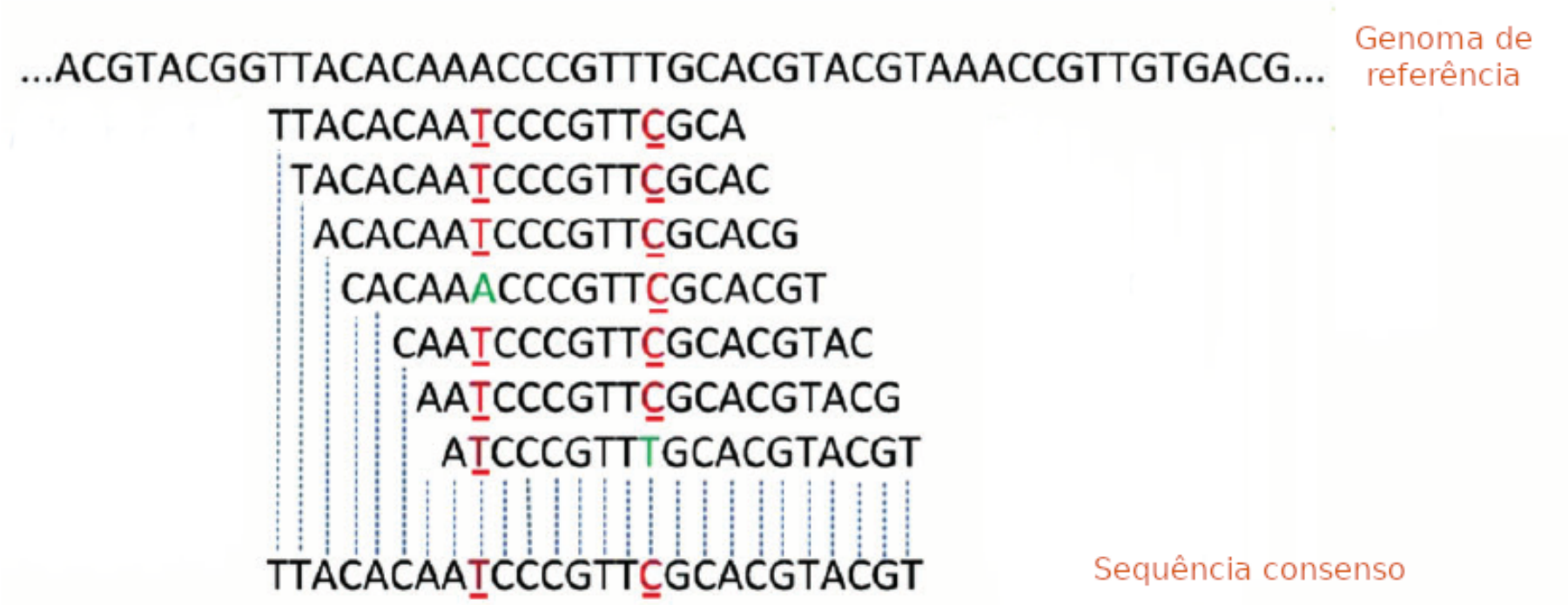
Prática...

Genoma do vírus Sars-Cov-2



Prática...

Montagem com genoma de referência



Prática...

Montagem com genoma do Sars-Cov-2 (Maranhão)