Next Generation Sequencing

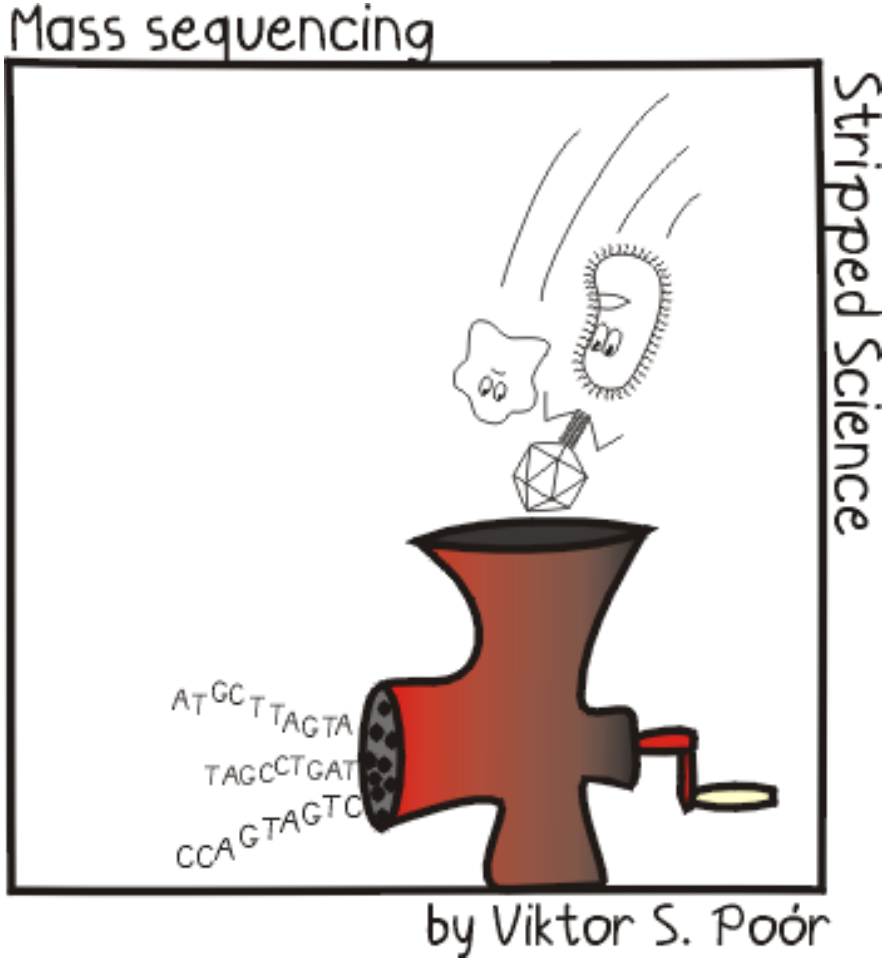
Detecção de variantes – conceitos chave

Diego R. Mazzotti, Ph.D.
Pesquisador
Laboratório de Biologia Molecular do Sono
Departamento de Psicobiologia
Universidade Federal de São Paulo

Supervisor
Molecular Core
Associação Fundo de Incentivo à Pesquisa (AFIP)

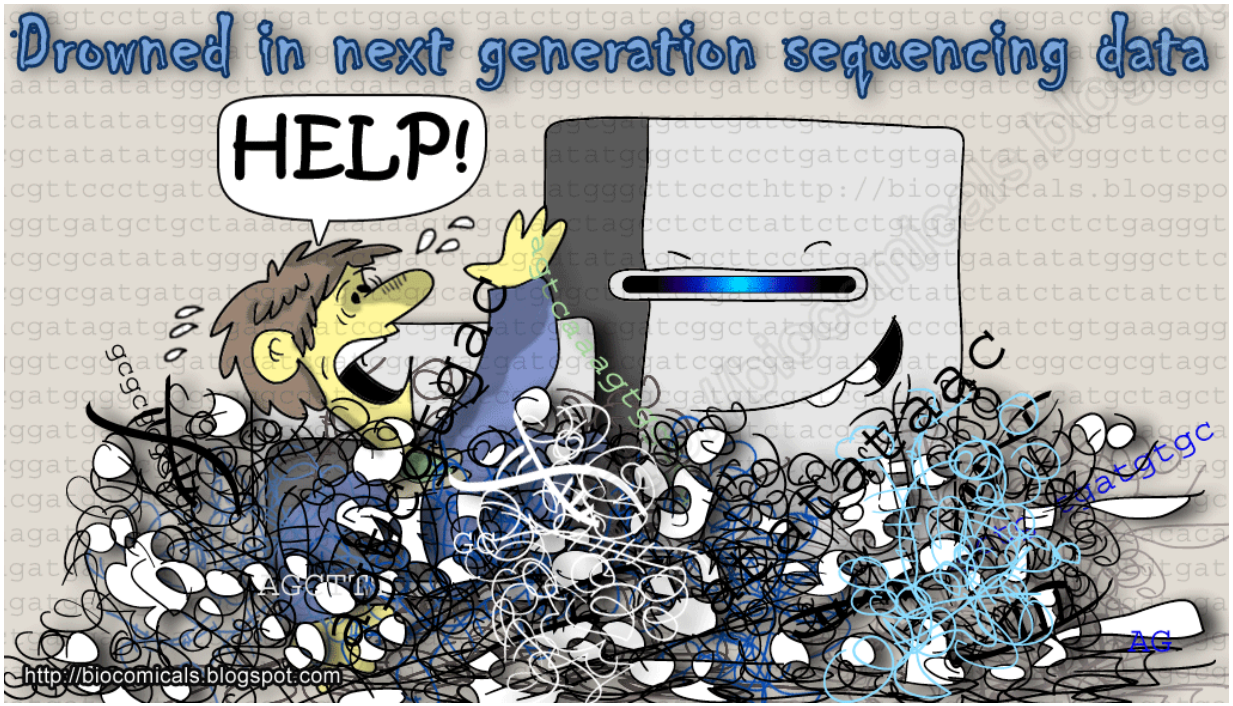
E-mail: mazzottidr@gmail.com

O sequenciamento de nova geração



Análise dos dados de sequenciamento de nova geração

- Independente da metodologia e aplicação:
  - Geração de milhões a bilhões de sequências curtas (50 – 250pb)



Análise dos dados de sequenciamento de nova geração

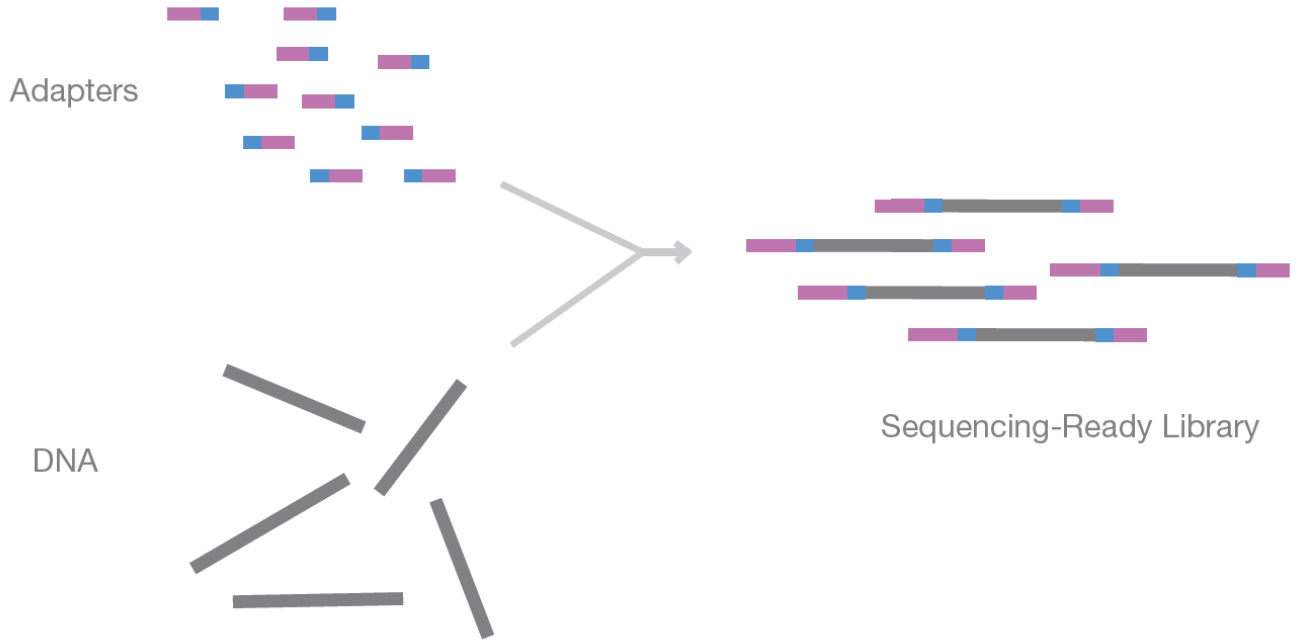
- Milhões de sequências curtas
  - Objetivos:
    - "Juntar" sequências para montar o genoma
    - Mapear as sequências em regiões conhecidas do genoma
    - Identificar o que é diferente em relação a um genoma referência (*Genome/ Exome/ Targeted* Sequencing)
    - Caracterizar níveis de expressão e identificar variantes (*Transcriptome* / RNA-sequencing)

Análise dos dados de sequenciamento de nova geração

- Alguns conceitos chave:
  - *Read*
  - *Single-end versus Paired-End reads*
  - Alinhamento
  - Cobertura (*Coverage / Depth of Coverage*)
  - *Variant Calling*
  - Anotação

Análise dos dados de sequenciamento de nova geração

- *Read*
  - Unidade resultante do sequenciamento de um fragmento de DNA



Análise dos dados de sequenciamento de nova geração

- *Read*
  - O NGS é paralelo, portanto milhões de *reads* são gerados por corrida
  - *Reads* são organizados em um arquivo chamado **.fastq**
  - *Reads* nesse arquivo estão dispostos **aleatoriamente**

Análise dos dados de sequenciamento de nova geração

- Formato **.fastq**

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*(((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

4 linhas para cada:

- 1 – Identificação do *read*
- 2 – Sequência de nucleotídeos
- 3 – Outra identificação do *read* (pouco utilizada)
- 4 – Escores de qualidade por símbolos

Análise dos dados de sequenciamento de nova geração

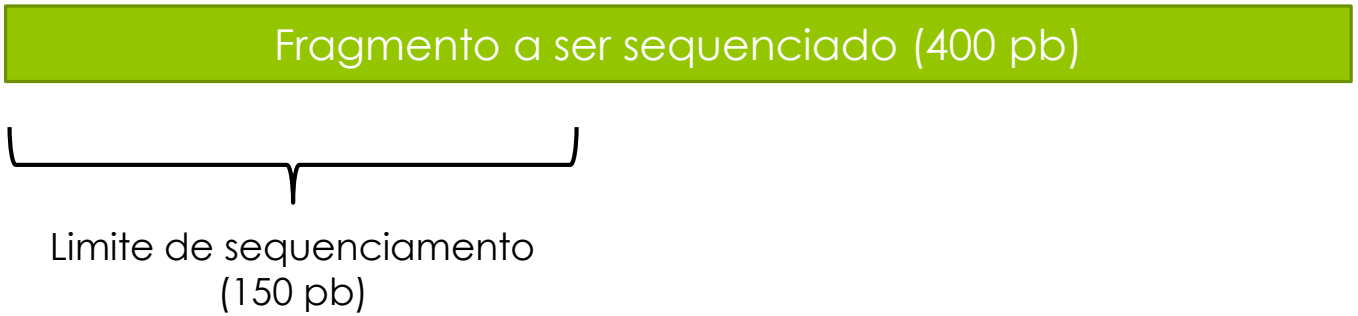
- Escores de qualidade por símbolos

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
 ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
 ..............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
 ..LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                             104                   126
 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).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

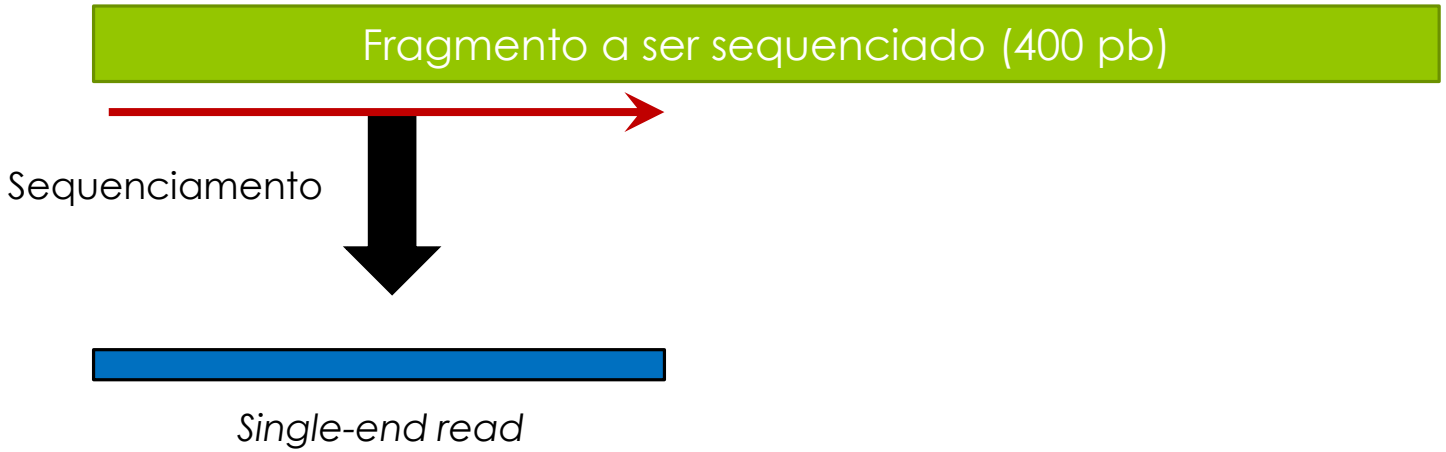
Análise dos dados de sequenciamento de nova geração

- *Single-end versus Paired-end reads*
  - Maneira como cada fragmento de DNA é sequenciado



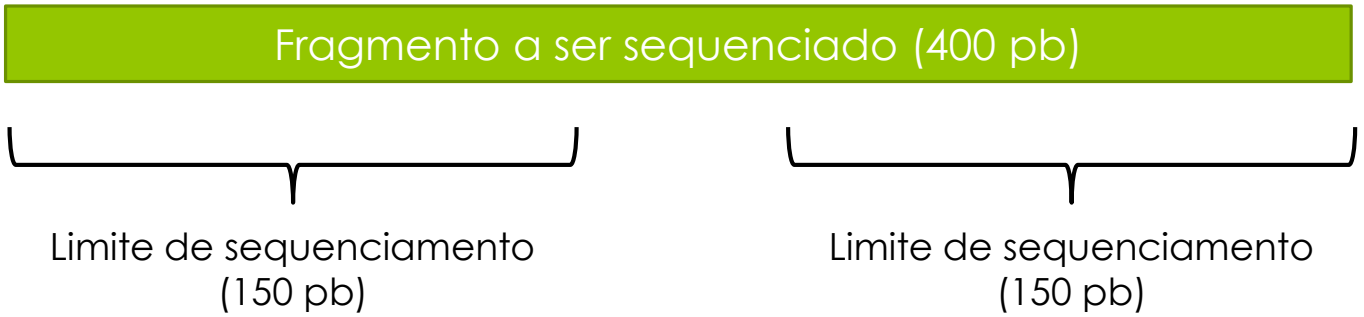
Análise dos dados de sequenciamento de nova geração

- *Single-end versus Paired-end reads*
  - Maneira como cada fragmento de DNA é sequenciado



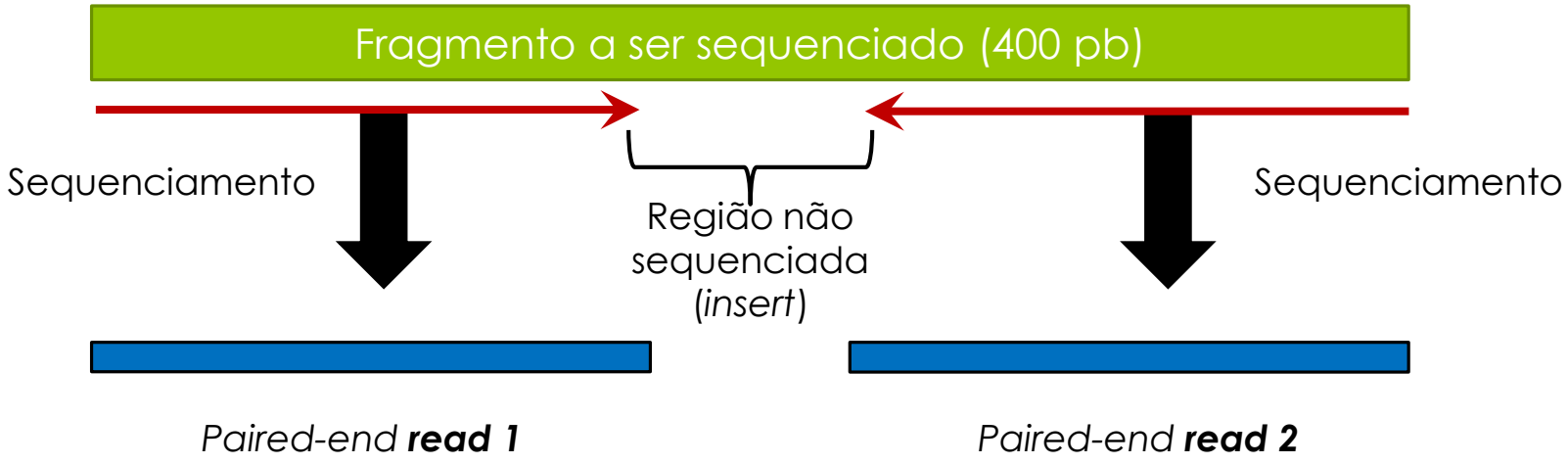
Análise dos dados de sequenciamento de nova geração

- *Single-end versus Paired-end reads*
  - Maneira como cada fragmento de DNA é sequenciado



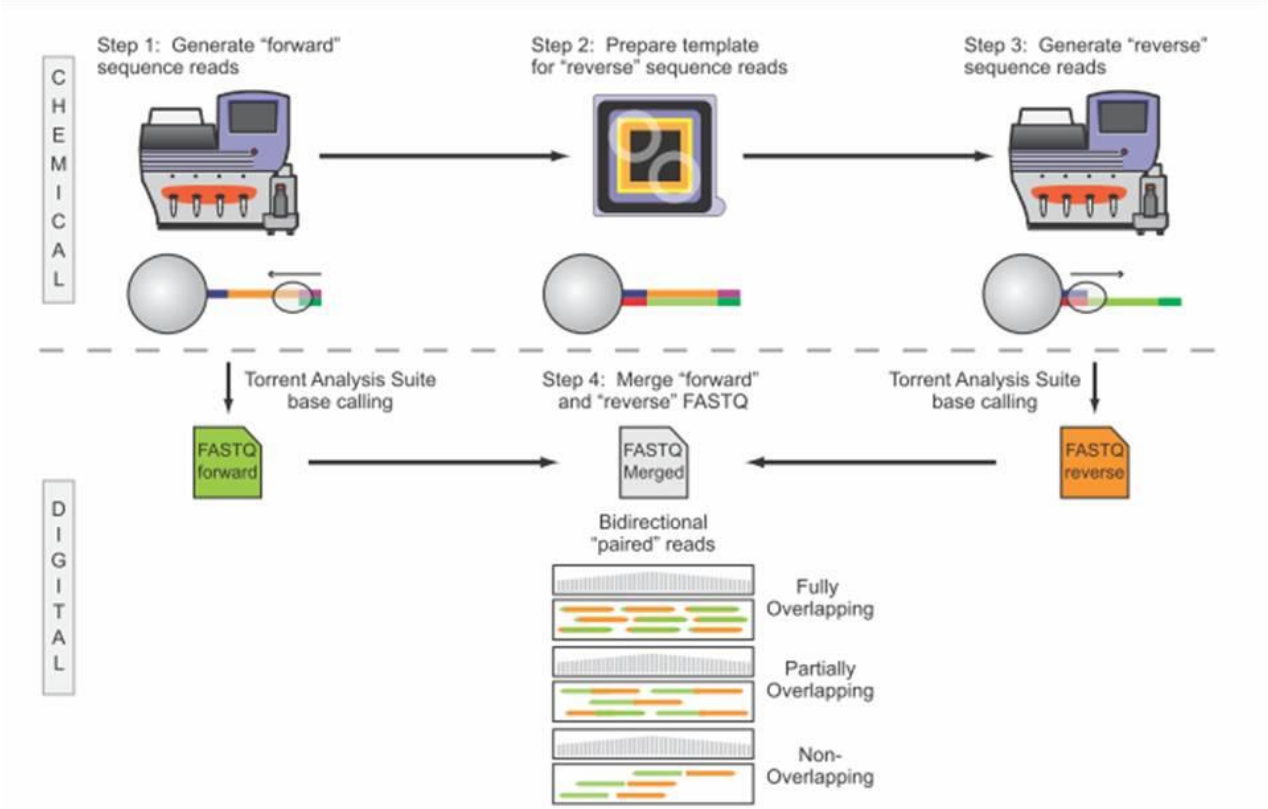
Análise dos dados de sequenciamento de nova geração

- *Single-end versus Paired-end reads*
  - Maneira como cada fragmento de DNA é sequenciado



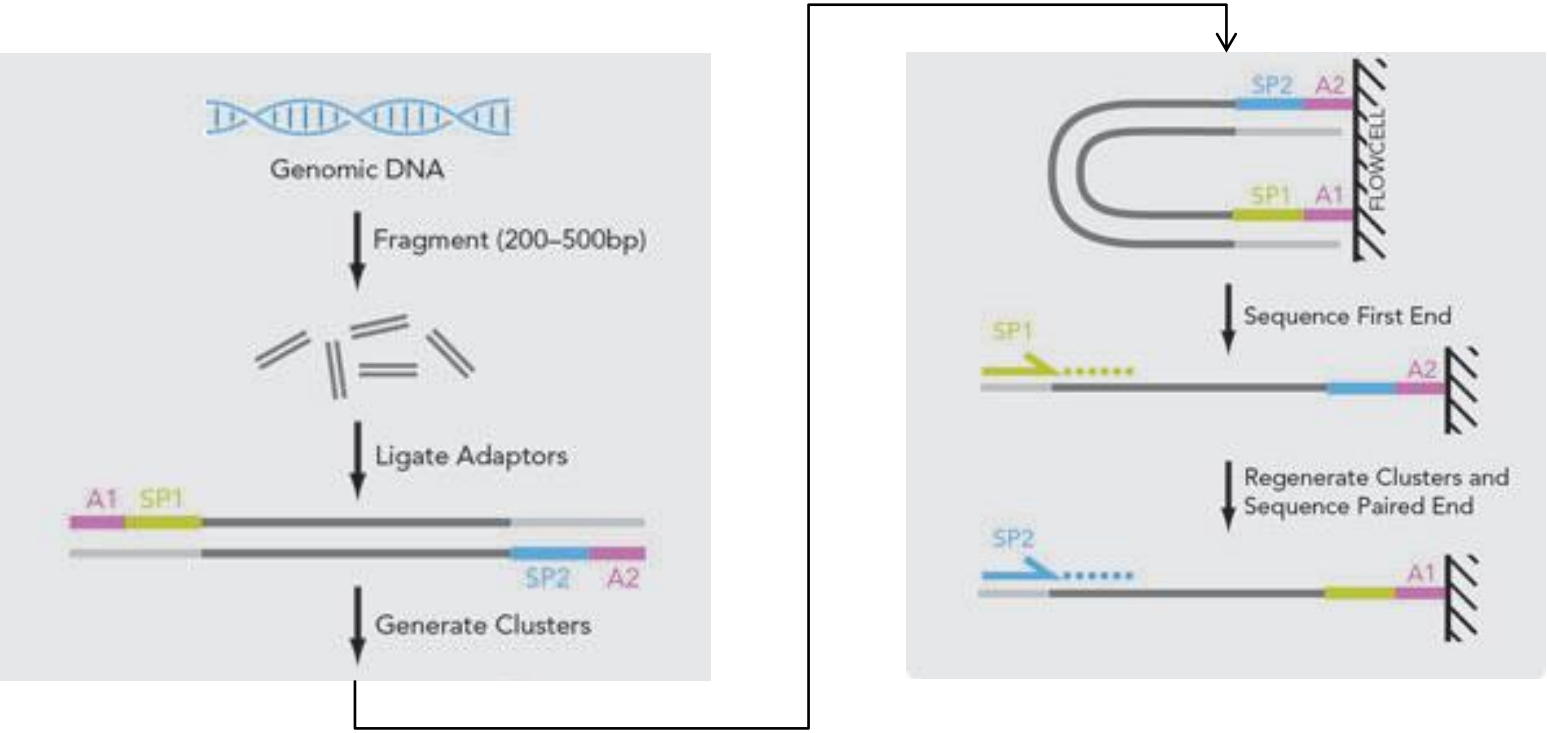
Análise dos dados de sequenciamento de nova geração

- *Single-end versus Paired-end (Ion Torrent)*



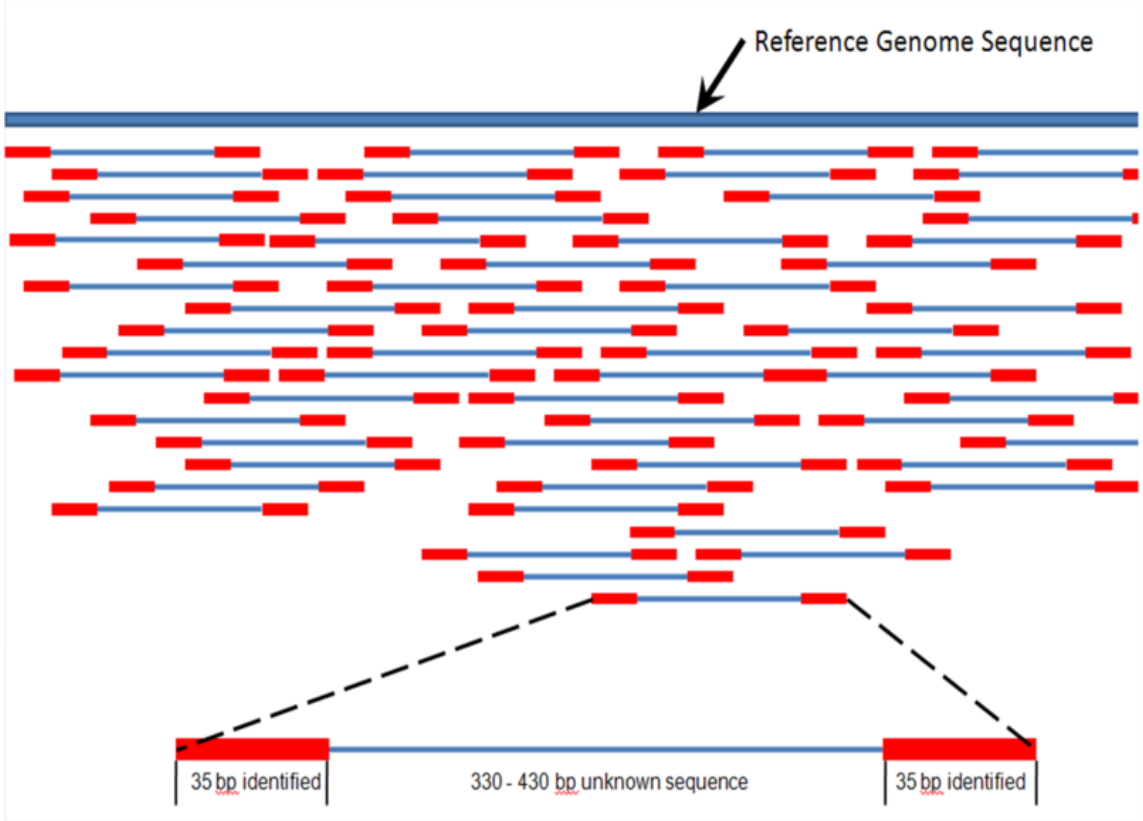
Análise dos dados de sequenciamento de nova geração

- *Illumina paired-end*



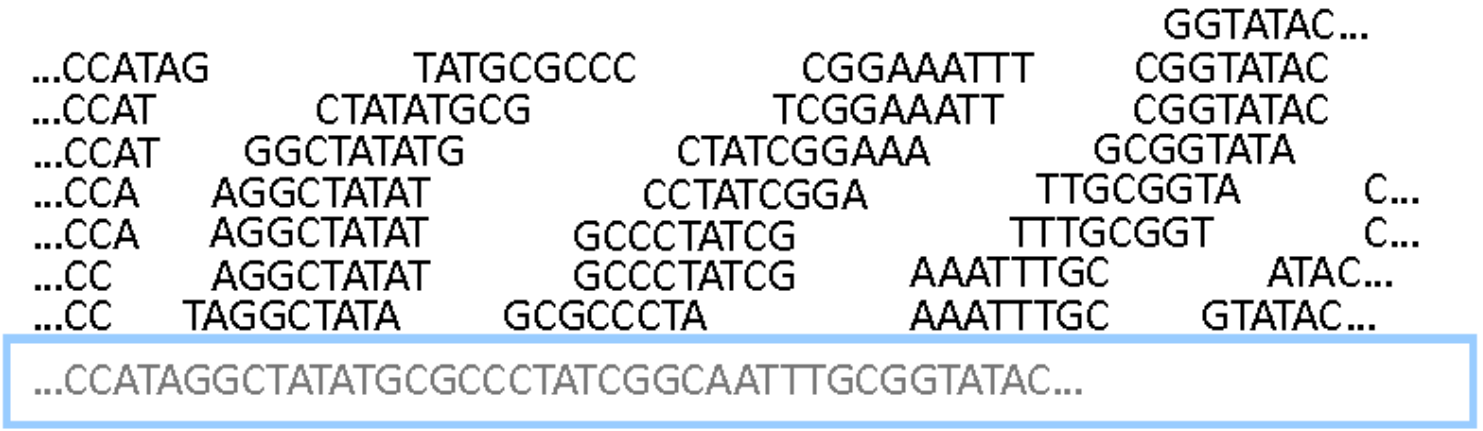
Análise dos dados de sequenciamento de nova geração

- *Paired-end reads*



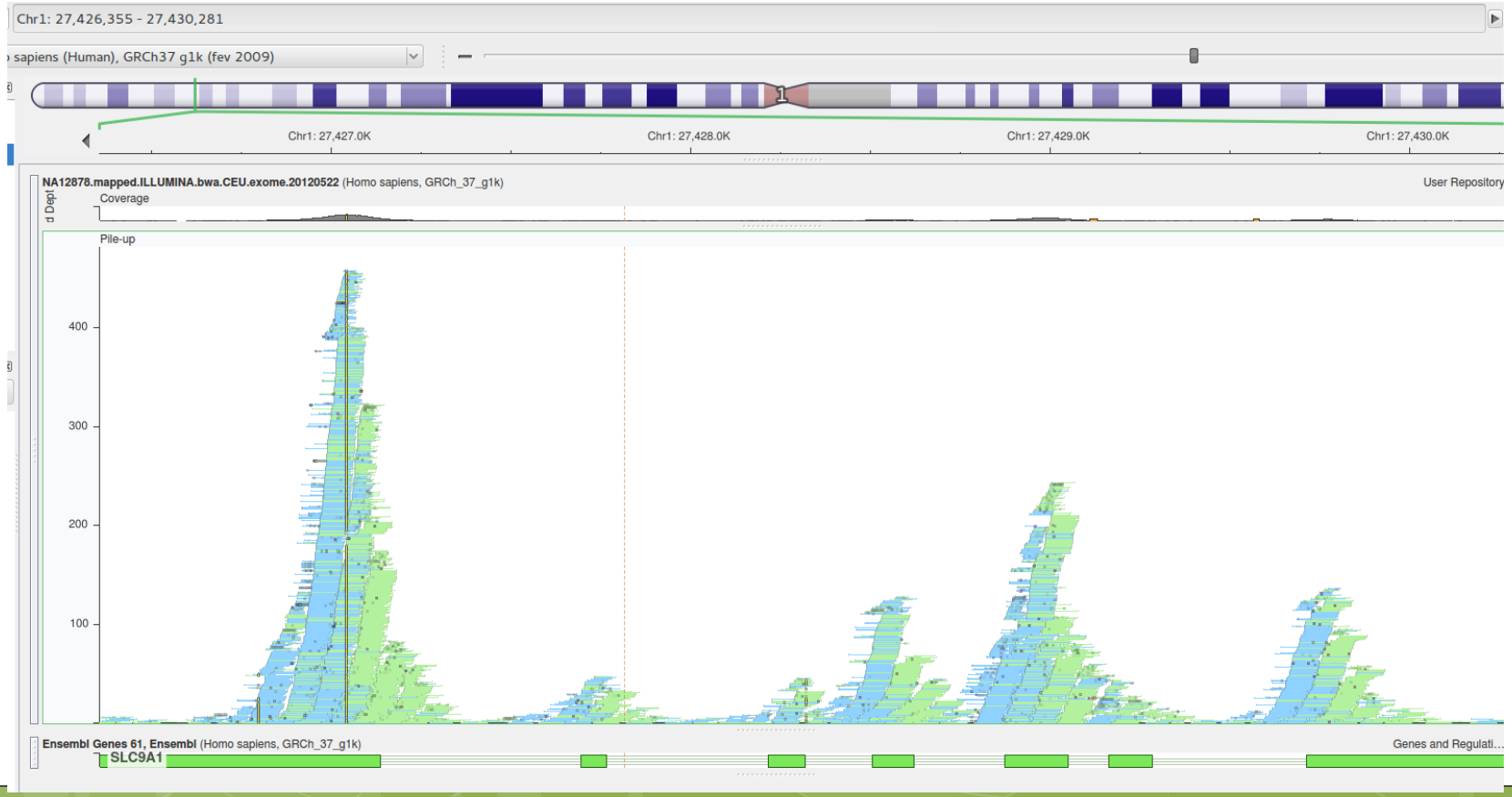
Análise dos dados de sequenciamento de nova geração

- Alinhamento (mapeamento)
  - Identificar a coordenada genômica de cada *read*
  - Deve ser computacionalmente rápida e permitir *mismatches*



Análise dos dados de sequenciamento de nova geração

- Exemplo de um resultado de alinhamento



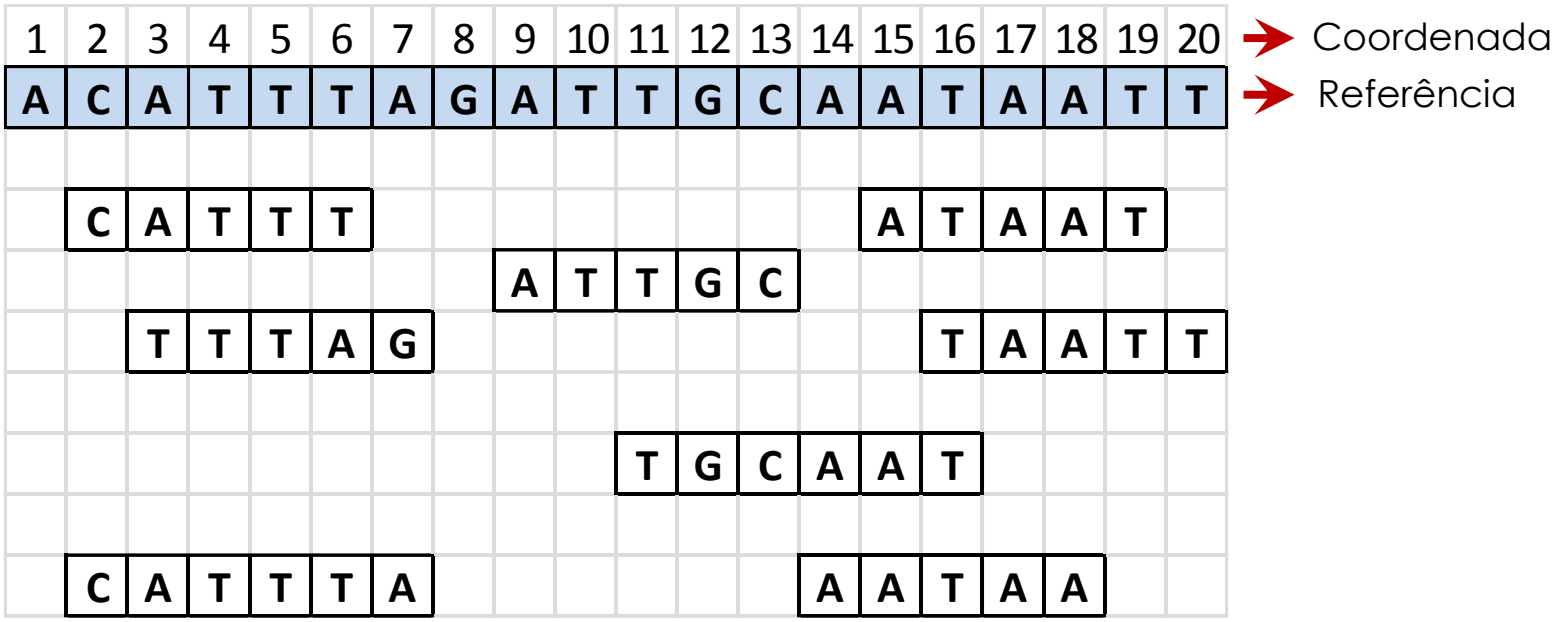
Análise dos dados de sequenciamento de nova geração

- Exemplo de um resultado de alinhamento

Análise dos dados de sequenciamento de nova geração

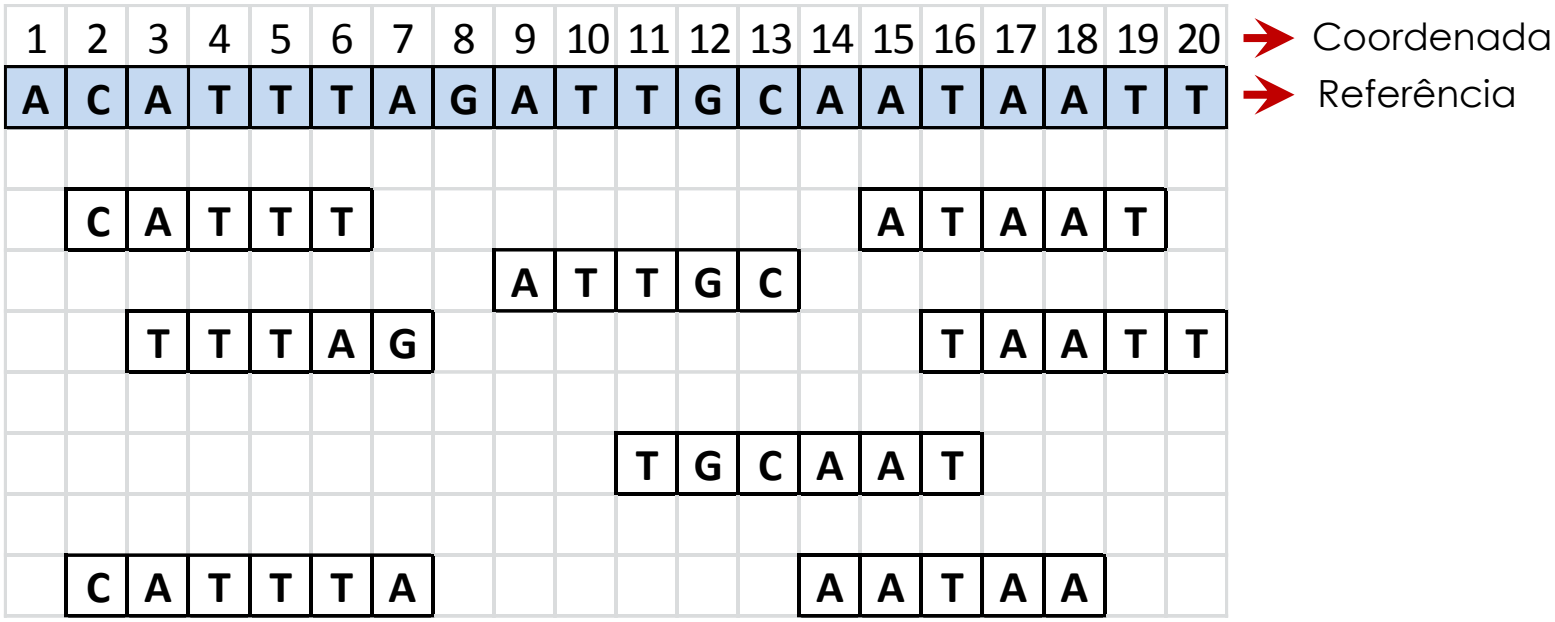
- Cobertura / *Coverage / Depth of Coverage*
  - Três definições:
    - Quantas vezes em média cada base foi "lida" (sequenciada). Ex: a cobertura foi 30 X
    - Quantos % do genoma foi coberto pelo meu sequenciamento. Ex: 95% de cobertura
    - Quantos % do genoma foi lido pelo menos X vezes. Ex: 80% do genoma foi lido pelo menos 30 X

Análise dos dados de sequenciamento de nova geração



Quantos reads?
Cobertura média?
% do genoma coberto?
Coordenada e cobertura da base com maior cobertura?

Análise dos dados de sequenciamento de nova geração



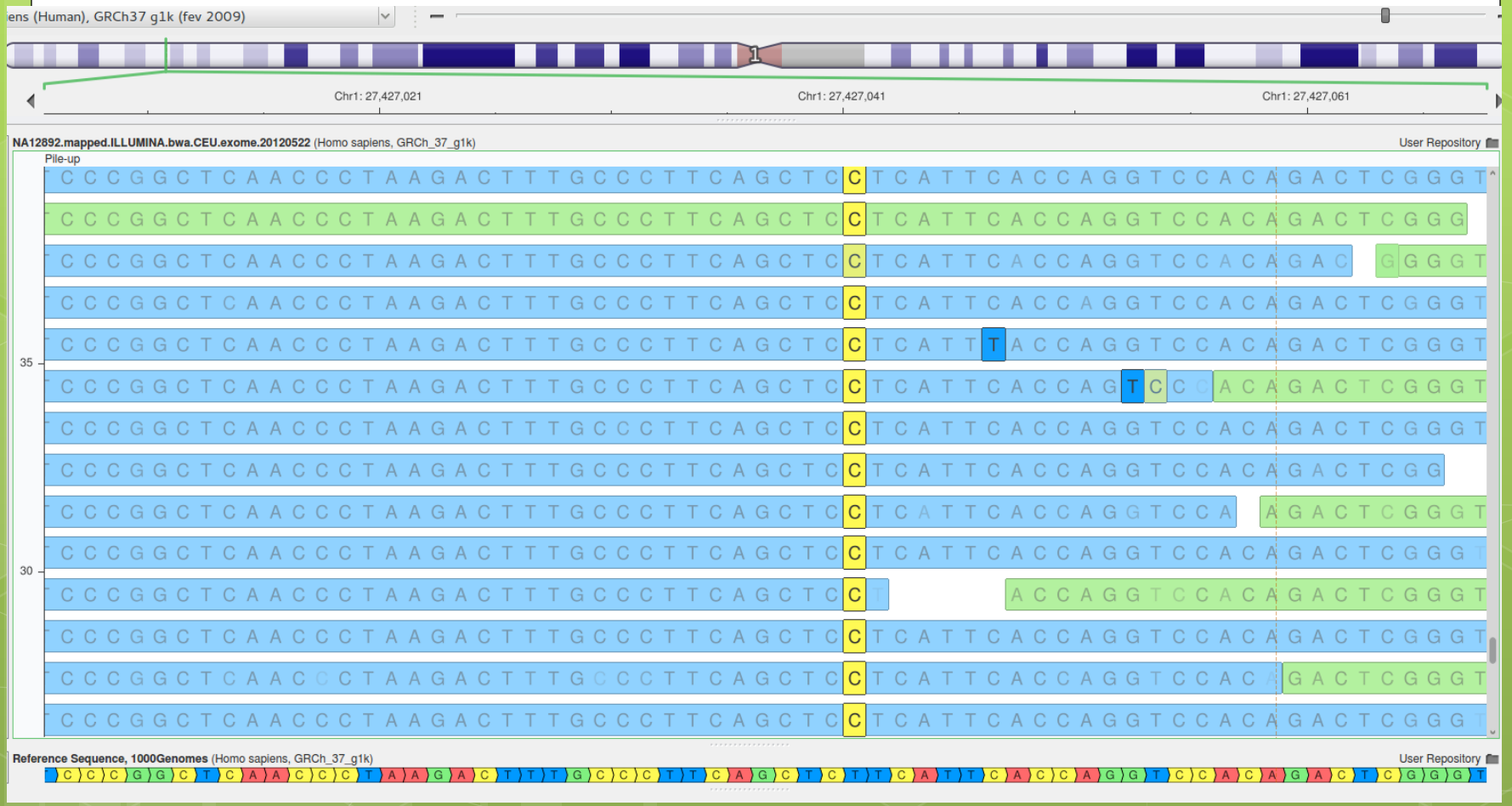
Quantos reads? → **8**
Cobertura média? → (0+2+3+3+3+...+3+2+1) / 20 = **2,1 X**
% do genoma coberto? → 18 / 20 = **90%**
Coordenada e cobertura da base com maior cobertura? → **16 (4 X)**

Análise dos dados de sequenciamento de nova geração

- *Variant Calling*
  - Identificar o que realmente é diferente do genoma de referência, ou seja, **identificar variantes**
  - Deve conseguir diferenciar:
    - Erro de sequenciamento versus variante real
    - Alteração em homozigose x heterozigose

Análise dos dados de sequenciamento de nova geração

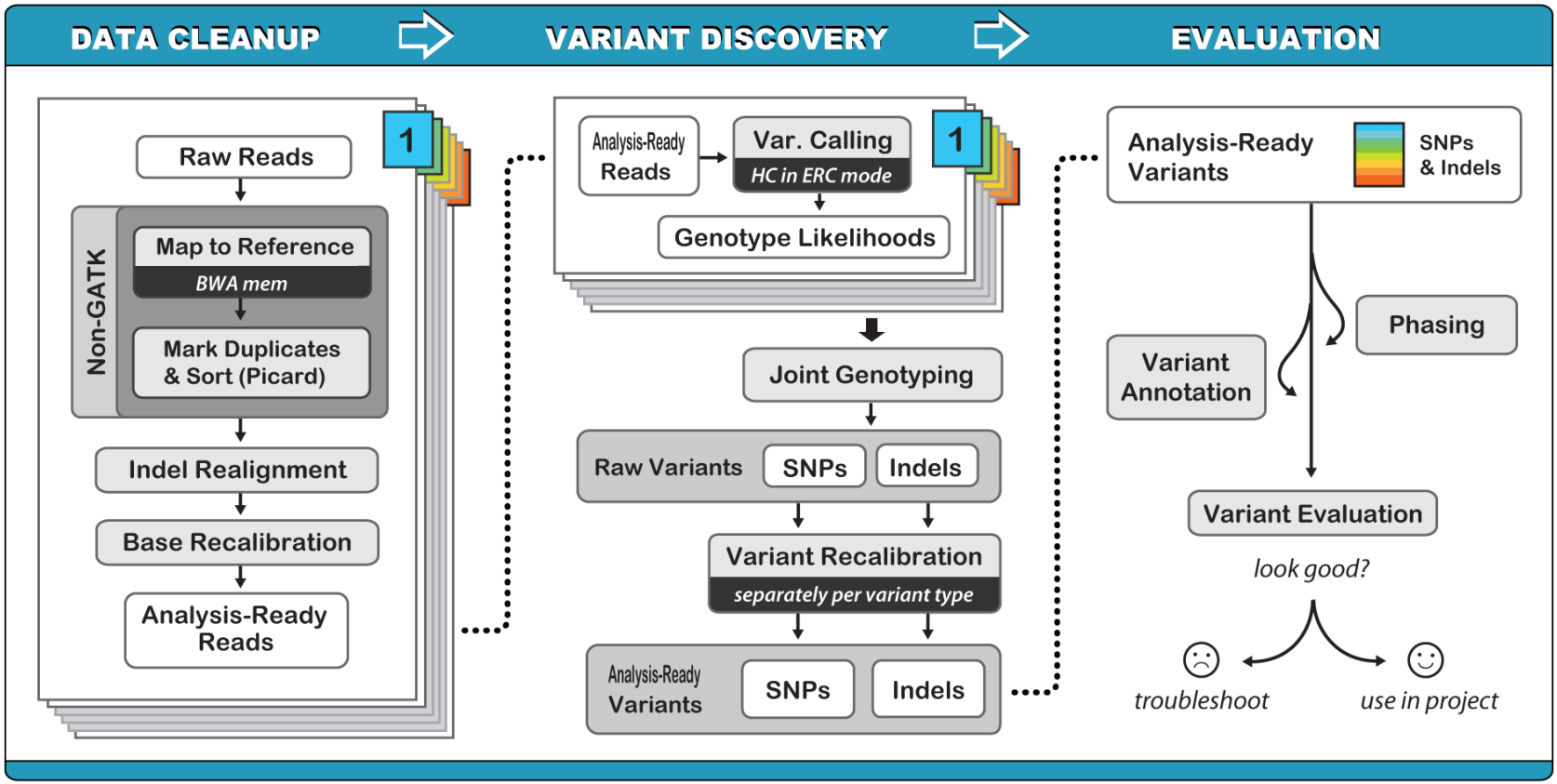
- *Variant Calling*



Análise dos dados de sequenciamento de nova geração

- Anotação
  - Atribuir o máximo possível de informação sobre cada variante identificada
  - Nome do gene, troca de base, troca de aminoácido, classe funcional, escore de predição de patogenicidade, frequência em estudos populacionais, etc.

Análise dos dados de sequenciamento de nova geração



Análise dos dados de sequenciamento de nova geração

- Etapas da análise (tutorial):