Ciências Biomédicas

FMRP/USP

Disciplina
RCB0300- Biotecnologia III

Aula

Princípios de sequenciamento em escala genômica

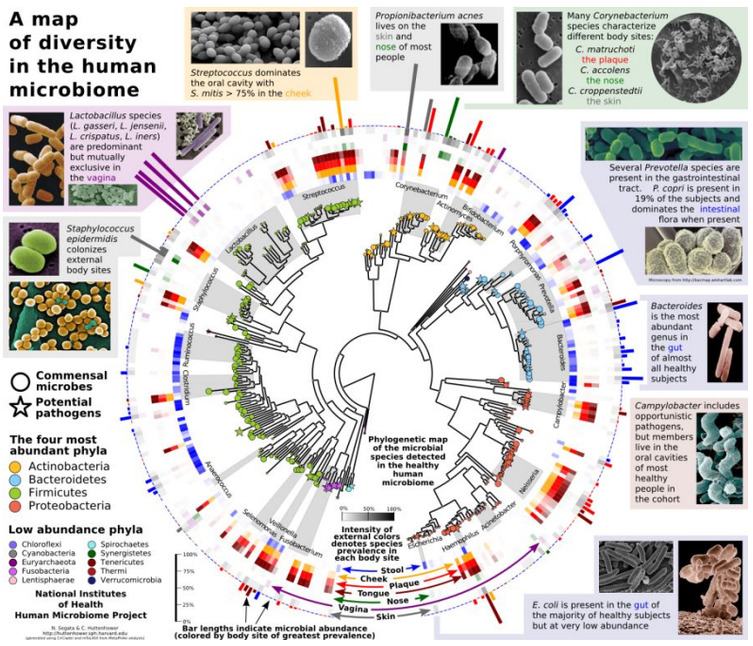
Wilson Araújo da Silva Junior
Departamento de Genética da FMRP/USP
htpp://rge.fmrp.usp.br
wilsonjr@usp.br

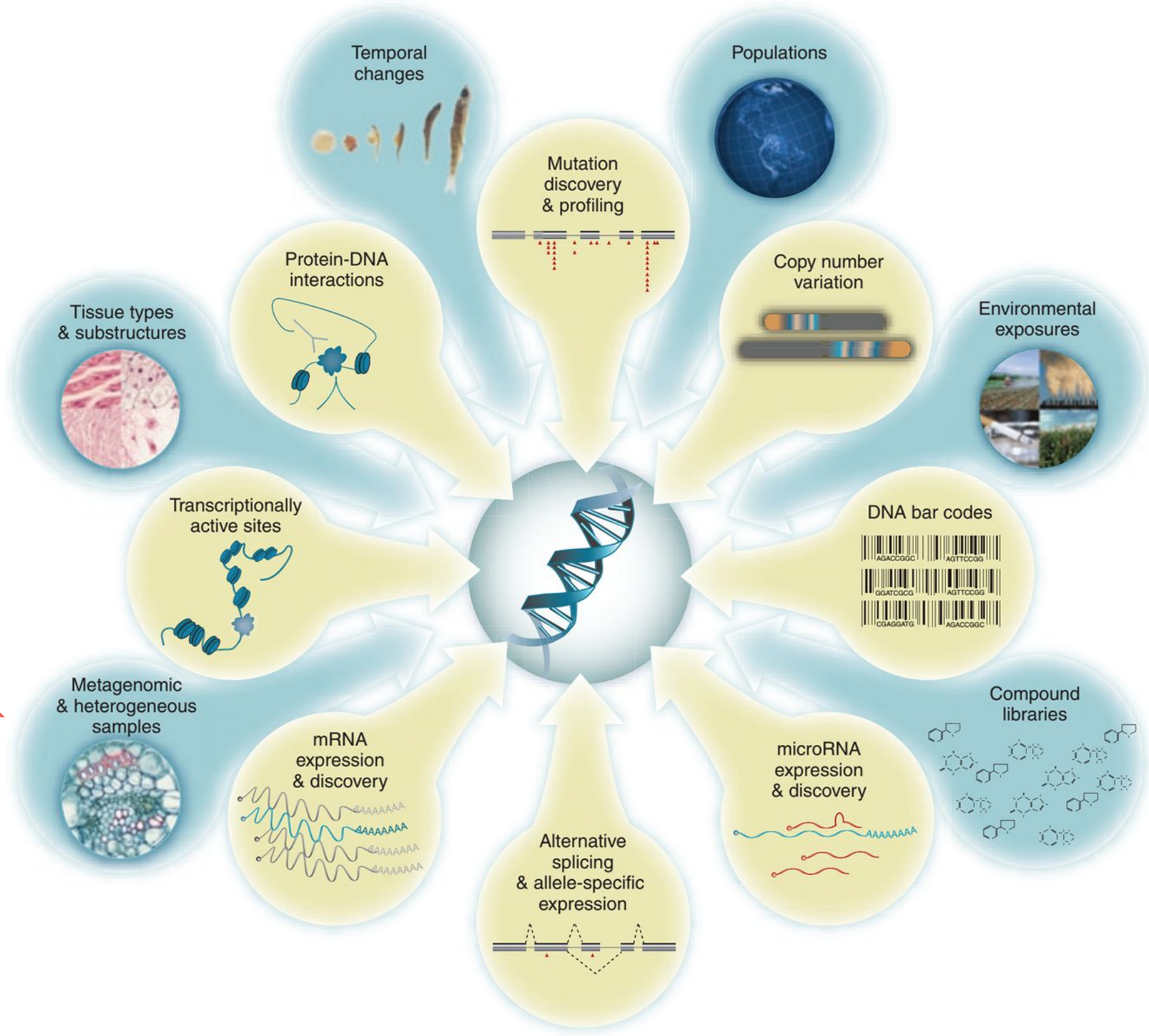
NGS (Next Generation Sequencing)




	iSeq 100 System	MiniSeq System	MiSeq Series ⊕	NextSeq Series ⊕
Run Time	9–17.5 hours	4–24 hours	4–55 hours	12–30 hours
Maximum Output	1.2 Gb	7.5 Gb	15 Gb	120 Gb
Maximum Reads Per Run	4 million	25 million	25 million†	400 million
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 × 300 bp	2 × 150 bp

What would you do if you could sequence everything?





A map of diversity in the human microbiome

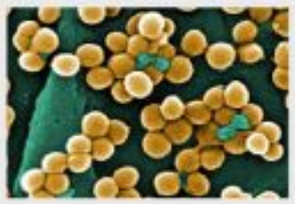
Streptococcus dominates the oral cavity with *S. mitis* > 75% in the cheek

Propionibacterium acnes lives on the skin and nose of most people

Many *Corynebacterium* species characterize different body sites:
C. matruchoti the plaque
C. accolens the nose
C. croppenstedtii the skin

Lactobacillus species (*L. gasseri*, *L. jensenii*, *L. crispatus*, *L. iners*) are predominant but mutually exclusive in the vagina

Staphylococcus epidermidis colonizes external body sites



Several *Prevotella* species are present in the gastrointestinal tract. *P. copri* is present in 19% of the subjects and dominates the intestinal flora when present

Microscopy from http://bacmap.wishartlab.com

Bacteroides is the most abundant genus in the gut of almost all healthy subjects

Phylogenetic map of the microbial species detected in the healthy human microbiome

0% 50% 100%

Intensity of external colors denotes species prevalence in each body site

Stool
Cheek
Plaque
Tongue
Nose
Vagina
Skin

100% 75% 50% 25% 0%

Bar lengths indicate microbial abundance (colored by body site of greatest prevalence)

Campylobacter includes opportunistic pathogens, but members live in the oral cavities of most healthy people in the cohort

E. coli is present in the gut of the majority of healthy subjects but at very low abundance

Commensal microbes

Potential pathogens

The four most abundant phyla

- Actinobacteria
- Bacteroidetes
- Firmicutes
- Proteobacteria

Low abundance phyla

- Chloroflexi
- Cyanobacteria
- Euryarchaeota
- Fusobacteria
- Lentisphaerae
- Spirochaetes
- Synergistetes
- Tenericutes
- Thermi
- Verrucomicrobia

National Institutes of Health Human Microbiome Project

N. Segata & C. Huttenhower
http://huttenhower.sph.harvard.edu

(generated using GraPhlAn and mTeM400 from MetaPhlAn analysis)

The 16S-based approach

Extract DNA

Microbial community sample

Extract DNA

The shotgun metagenomic approach

Amplify and sequence 16S rRNA

OTU1 GATACAGAGATGCAT GATACAGTGATGCAT GATACAGTAGATGCAT GATACAGTAGATGCAT

OTU2 TACCAGATTTACATAC TACCAGATTACATAC TACCAGATTACATAC

OTU3 CAGGTAGGATAATA CAGGTAGGATAATA CAGGTAGGATAATTA CAGGTAGGATAAATA CAGGTAGGAAATA

OTU4 GATTACAGATTACA GATTACAGATTTACA GATTACAGAATTACA

Group similar sequences into OTUs

NCBI GreenGenes myRDP Silva

Use database to identify OTUs

Community composition: Which organisms are present?

Abundance

OTU

Relative abundance of OTUs in community

OTU phylogeny

Variant sequences and SNPs

GATTACA
GATTACA
GATTTCA
GATTTCA
GATTTCA

Sequence community DNA

NCBI KEGG SEED BLAST

Use database to identify sequences

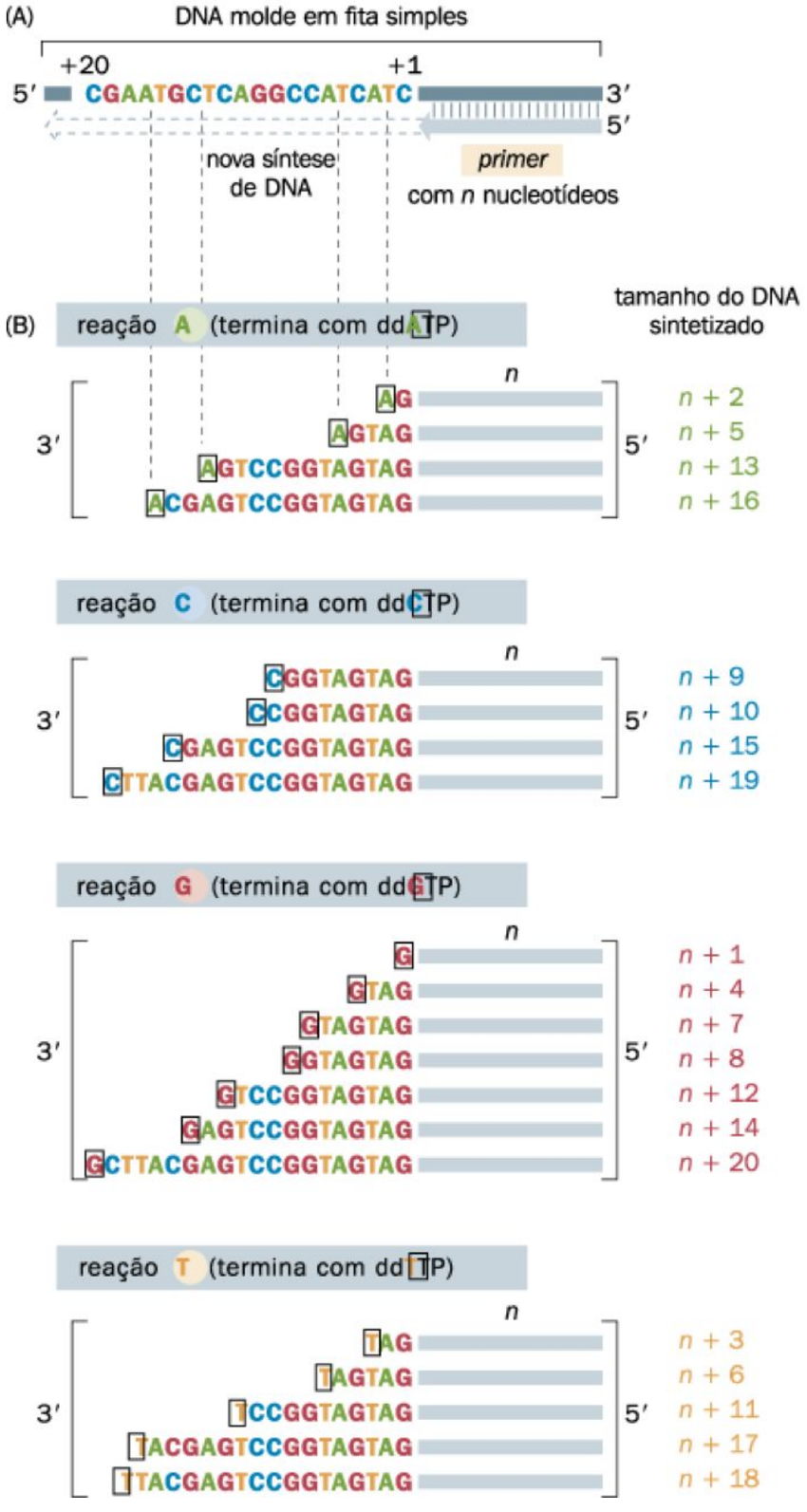
Compare sequences to reference genomes

Community function: What can the community do?

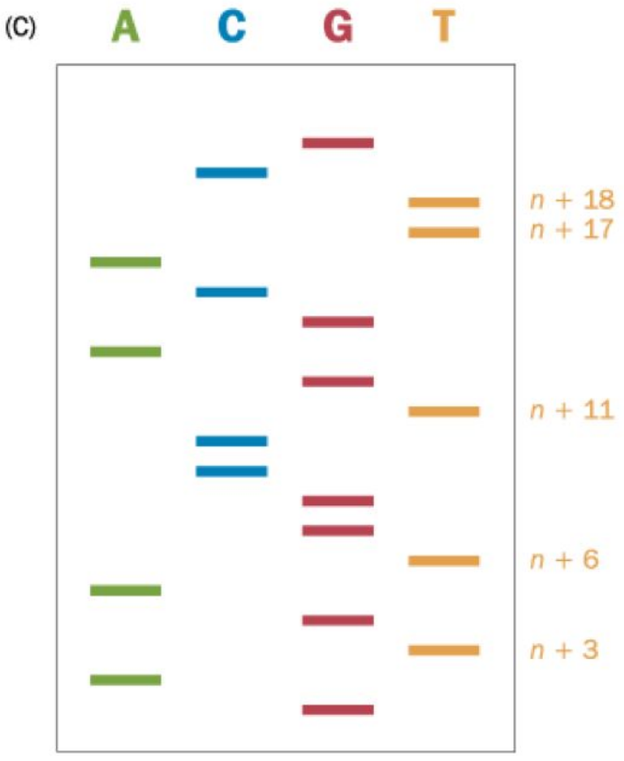
Abundance

Functions

Relative abundance of gene pathways in community



Frederick Sanger
1977



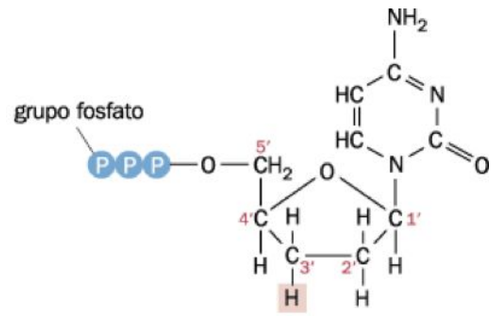
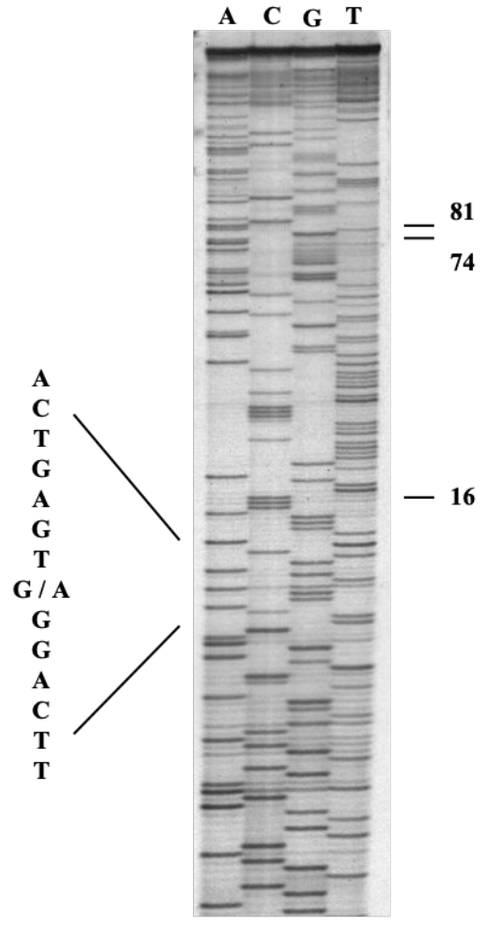
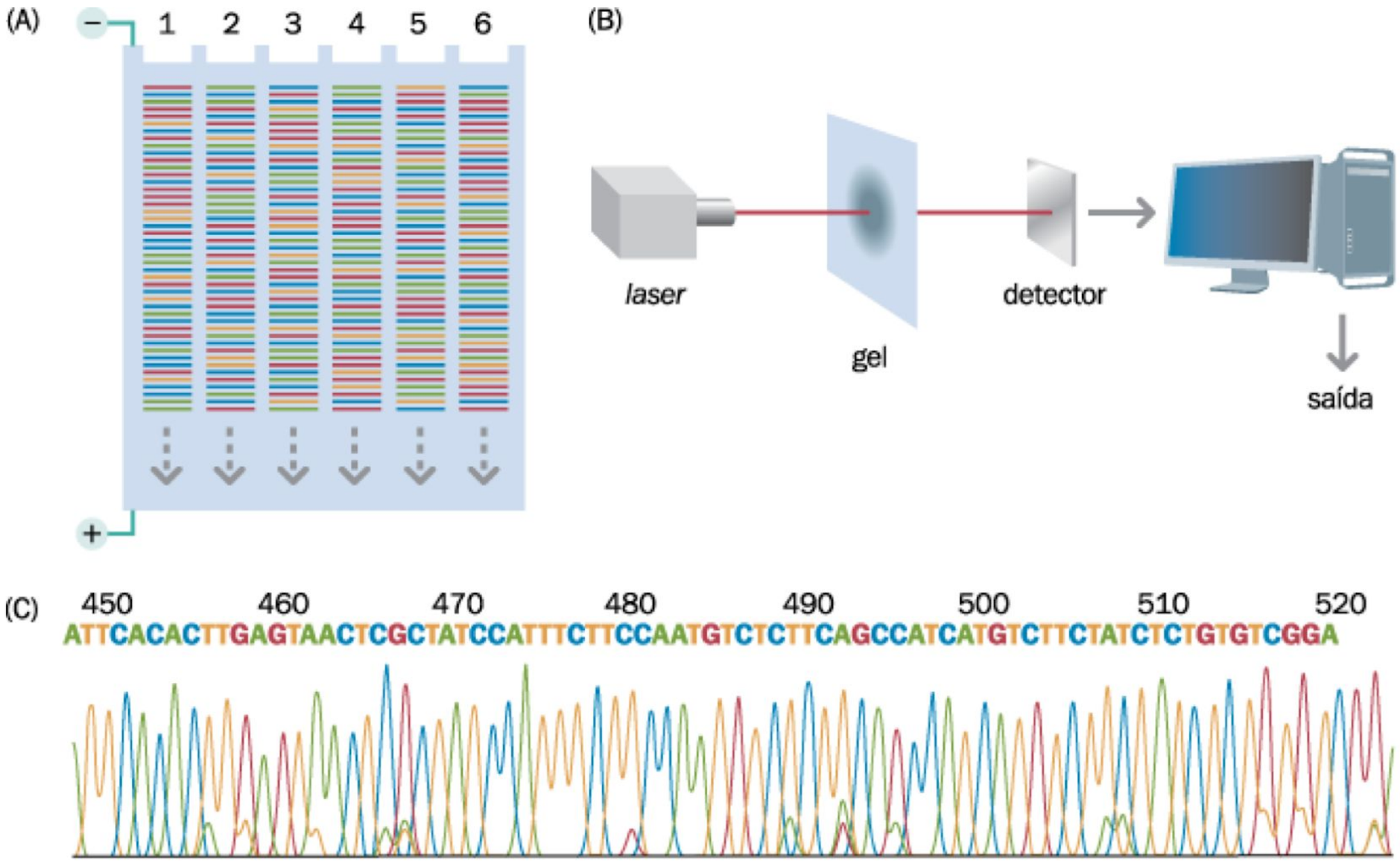


Figura 8.5 Estrutura de um dideoxinucleotídeo: 2′, 3′-dideoxi CTP (ddCTP). O grupo hidroxila ancorado ao carbono 3′ em nucleotídeos normais é trocado por um átomo de hidrogênio (sombreado).

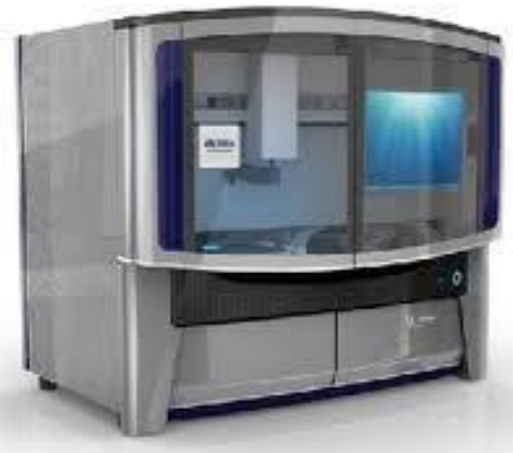


Hb A/ β°-Talassemia (IVS-2 splice junction G-A)

Video: https://www.youtube.com/watch?v=qr6A4eXZOWs


(A)
–
1
2
3
4
5
6
+
(B)
laser
gel
detector
saída
(C)
450
460
470
480
490
500
510
520
ATTCACACTTGAGTAACTCGCTATCCATTTCTTCCAATGTCTCTTCAGCCATCATGTCTTCTATCTCTGTGTCGGA

Next Generation Sequencing



2001 → 2011

Biologia Computacional

Biologia de Sistemas Moleculares

Protocolos:

DNA → Structural and genomic alterations/mutation

- Whole Genome-Seq
- Exome-Seq
- Metagenomics
- 16S Ribossomal diversity

Epigenomics → Regulation of Gene Expression

- Chip-Seq
- Methyl-Seq

RNA → Expression of Genes (canonical and non-canonical)

- mRNA-Seq
- miRNA-Seq

Análise de dados em escala genômica

Next Generation Sequencing

1. Definição do problema biológico;

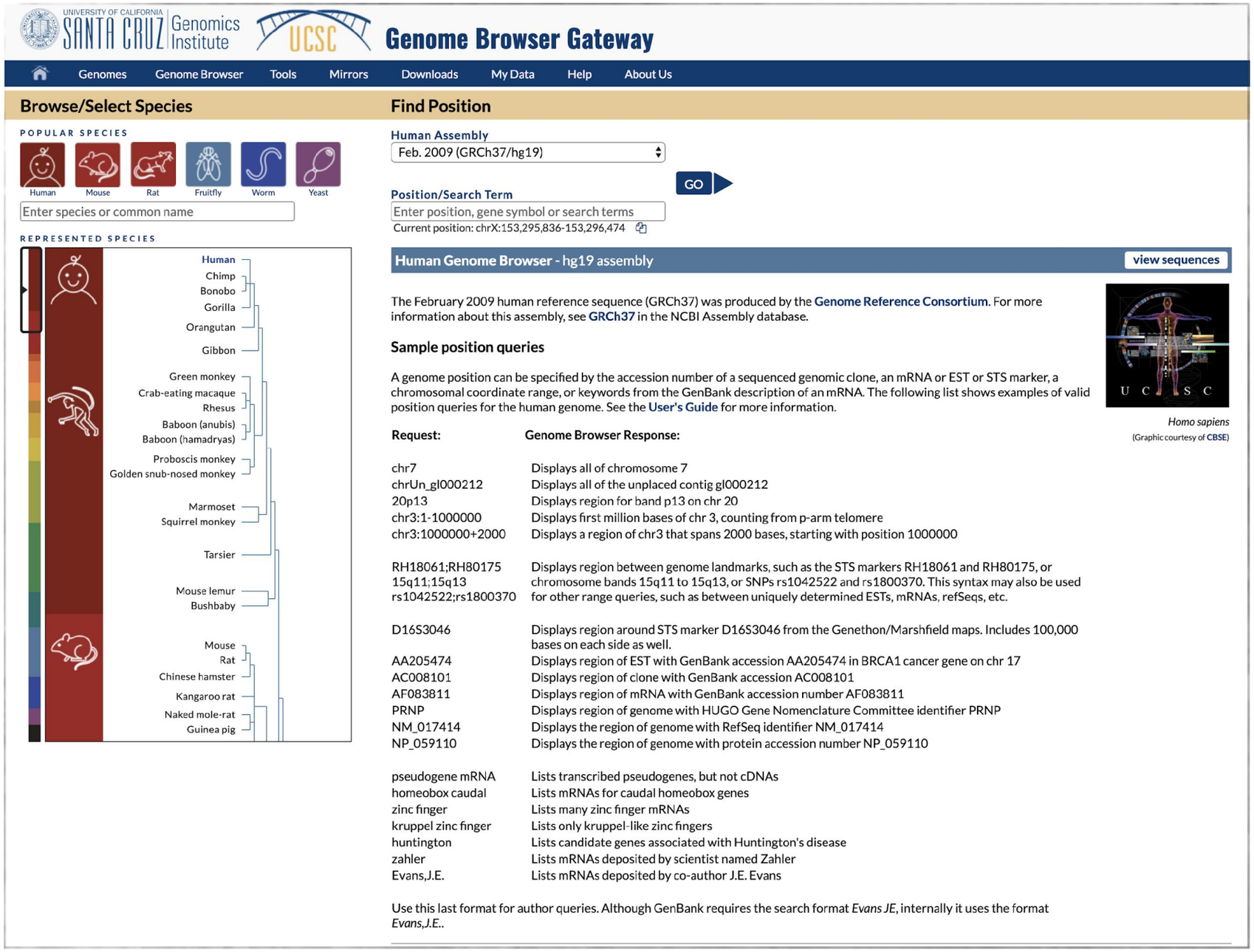
2. Discussão do desenho experimental;
   - Tipo de amostra, número amostral, abordagem de NGS, etc.

4. Discussão dos resultados;

5. Conclusão do estudo.

Análise de dados em escala genômica

Next Generation Sequencing



NGS (Next Generation Sequencing)

Etapas:

1. Extração do DNA ou RNA

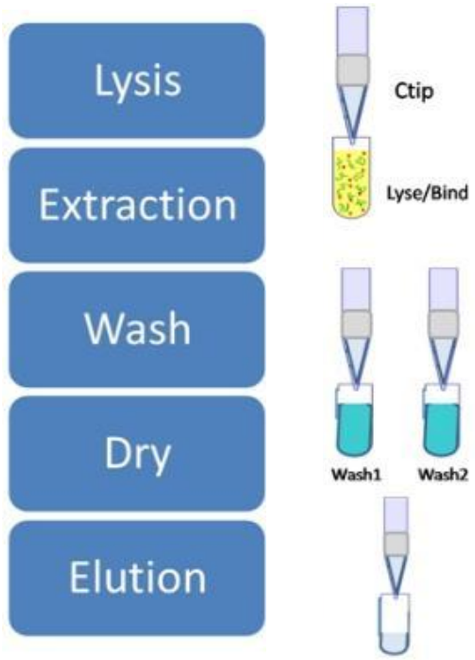
2. Preparo da biblioteca (fragmentação randômica)

3. Sequenciamento da biblioteca

4. Análise dos dados (Bioinformática)

Extração do DNA ou RNA

1. Extração do DNA ou RNA



- Manual ou automatizada

- Qualidade (pureza e integridade)

Pureza: <u>Contaminação com RNA/proteínas</u>

OD 260nm /280nm Razão = 1,8 – 2,0

Amostras com razão acima de 2,0 tem muito RNA

Amostras com razão abaixo de 1,8 tem contaminação com proteína

<u>Contaminação com reagentes</u>

OD 260nm /230nm Razão = 2,0 – 2,2

Amostras com razão abaixo do esperado tem contaminação reagentes

Integridade: <u>Fragmentação do DNA ou RNA</u>

Quantificação do DNA ou RNA

Fator de conversão: Uma absorção de **1,0 a 260** nm corresponde a uma concentração de 50 ng/µL (50 µg/mL) para dsDNA puro.

Obs: Os valores de absorção devem estar dentro da faixa linear do espectrofotômetro.

Método:

1. Diluir uma fração de 10 µL de DNA purificado tratado com RNase em 90 µL de TE (diluição 1/10) (Fator de Diluição).

2. Medir a absorção em 320 nm, 280 nm e 260 nm.

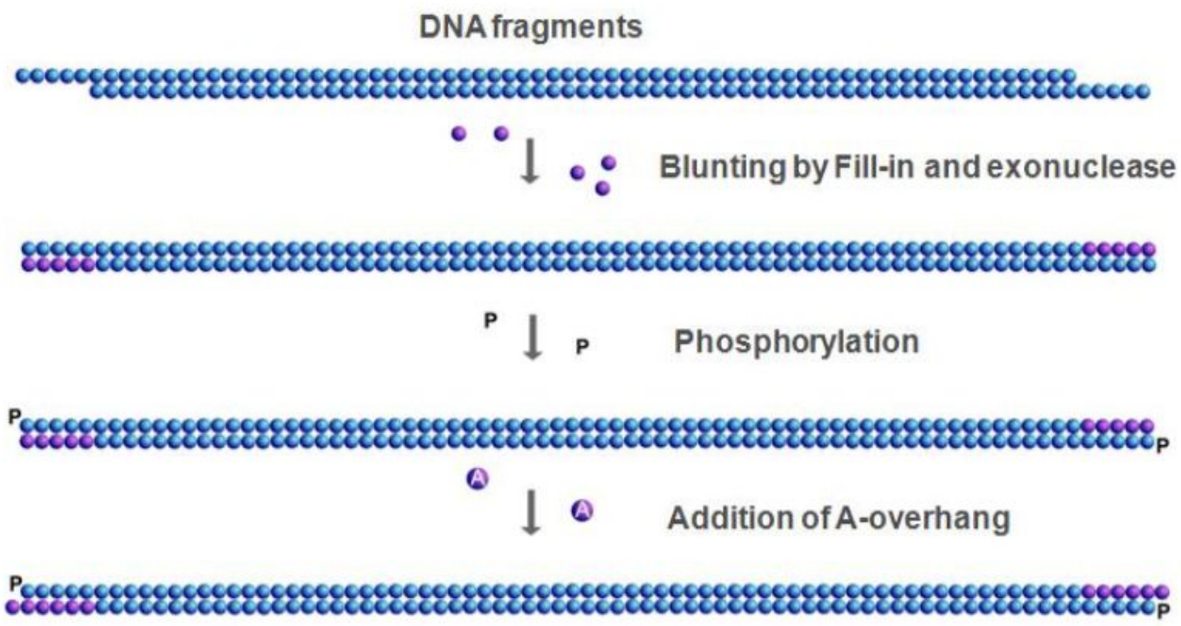
3. Calcular os valores corrigidos de A280 e A260 subtraindo a absorção em 320 nm (A320) dos valores de A280 e A260.

4. Concentração do DNA em ng/µL = A260 corrigido × 10 (**fator de diluição**) × 50 (**fator de conversão**).

5. Razão A260/A280: dividir A260 corrigido por A280 corrigido.

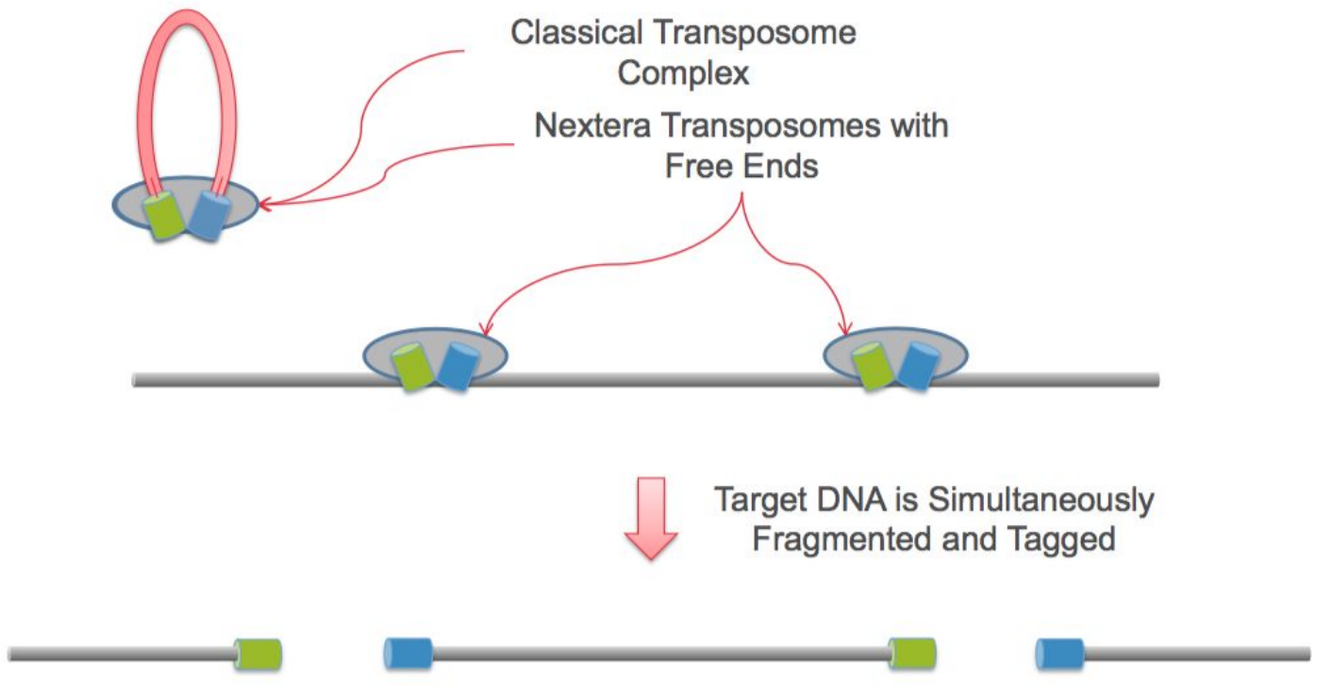
NGS (Next Generation Sequencing)

2. Preparo da biblioteca (fragmentação randômica)



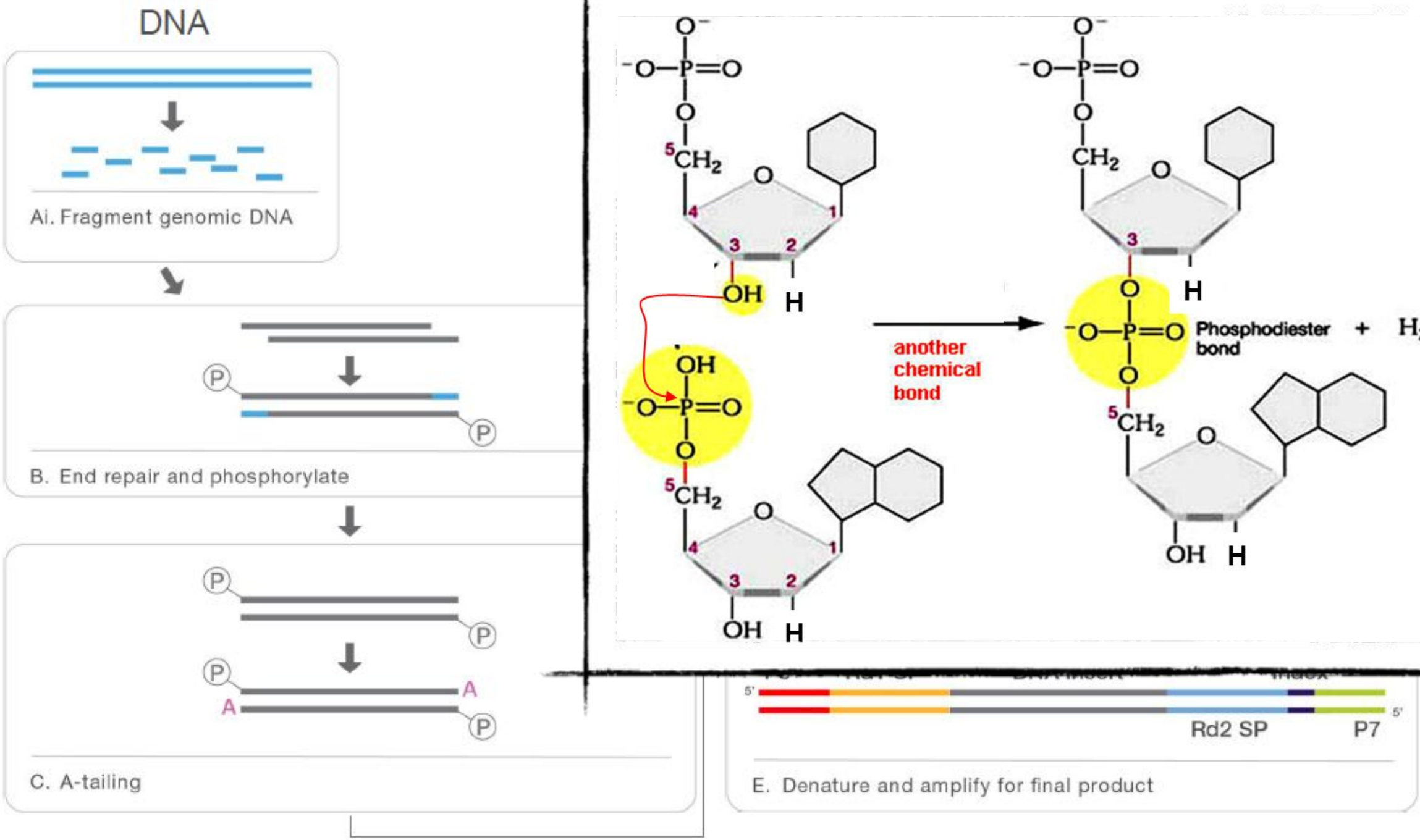
Mecânica ou enzimática

Tagmentação
(Transposase)



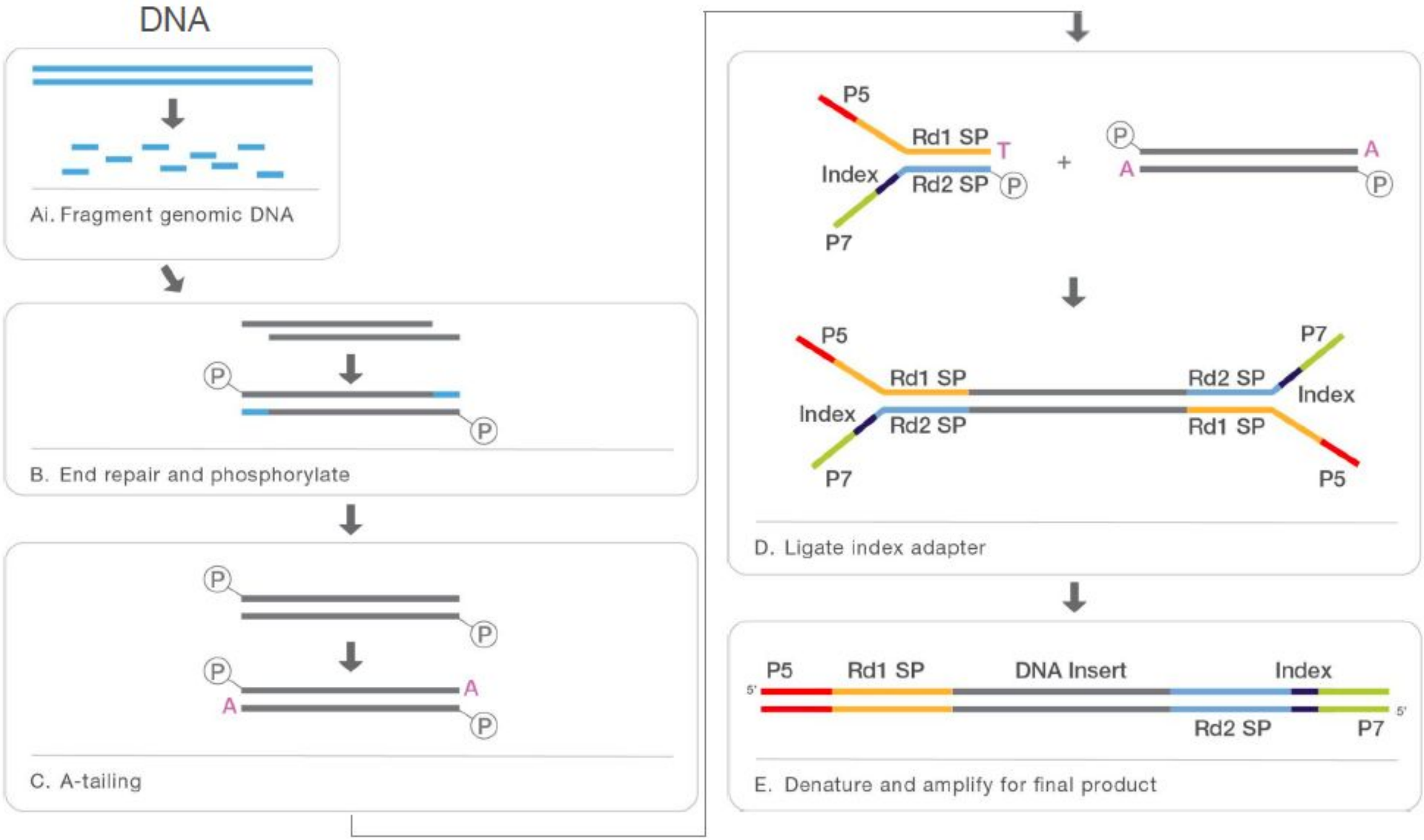
NGS (Next Generation Sequencing)

2. Preparo da biblioteca (fragmentação randômica)



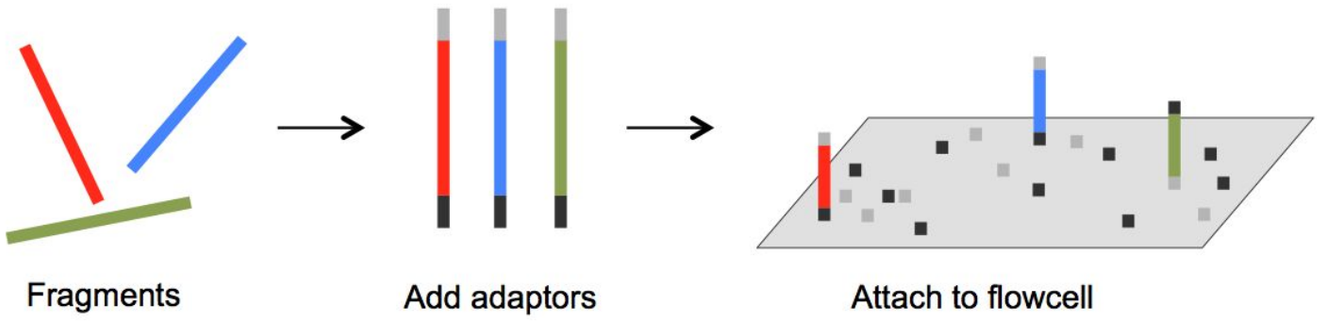
NGS (Next Generation Sequencing)

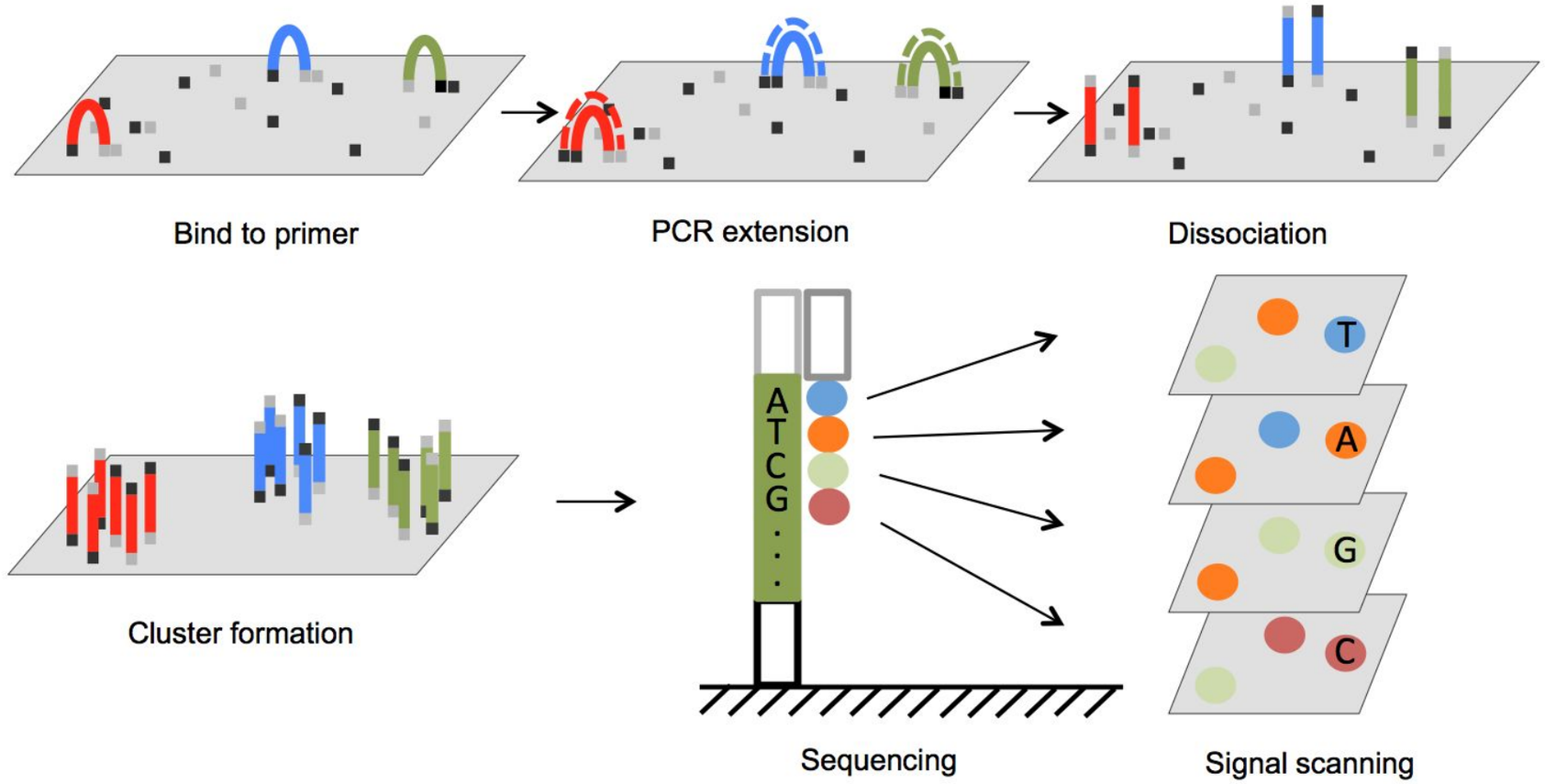
2. Preparo da biblioteca (fragmentação randômica)



NGS (Next Generation Sequencing)

3. Sequenciamento da biblioteca

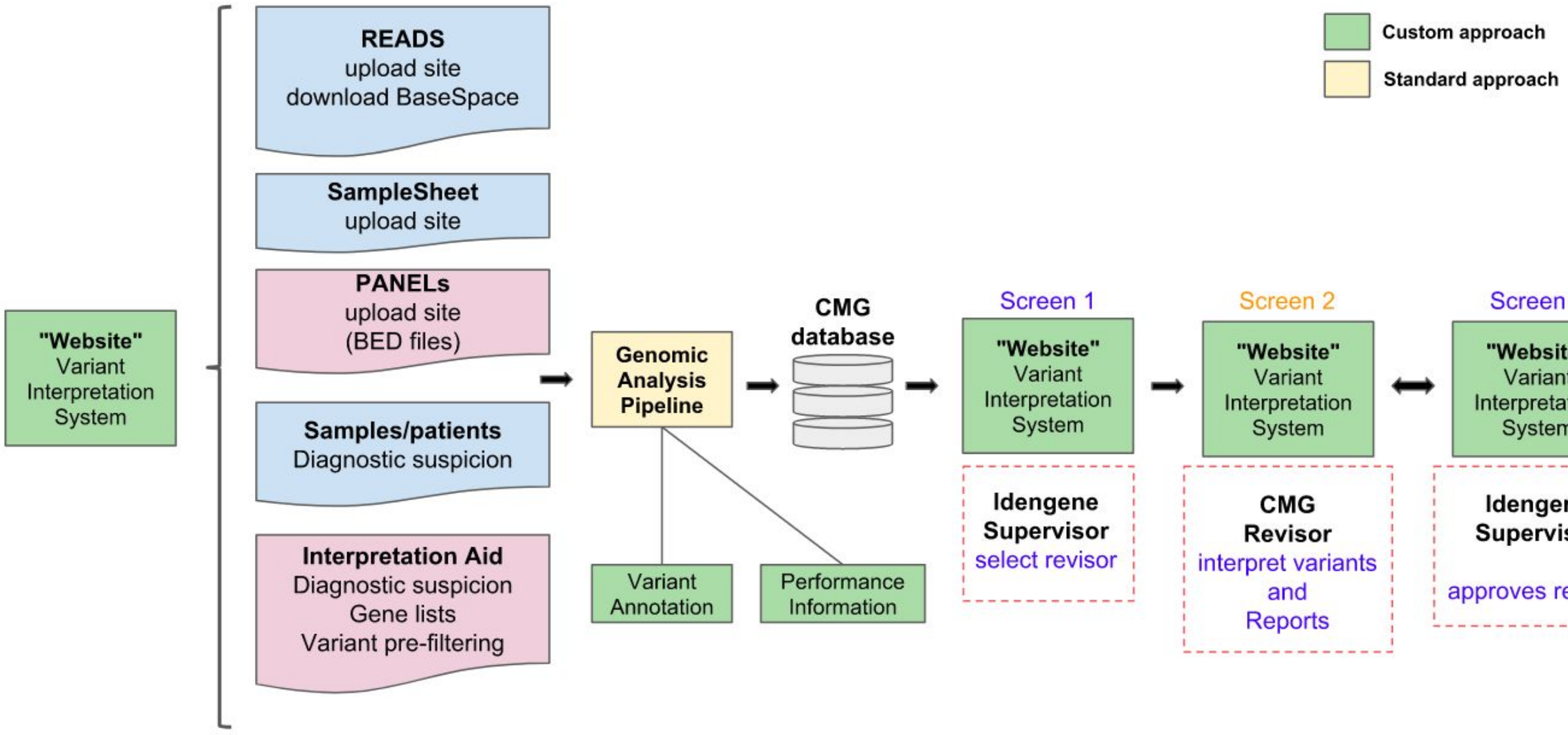




Video: https://www.youtube.com/watch?v=fCd6B5HRaZ8

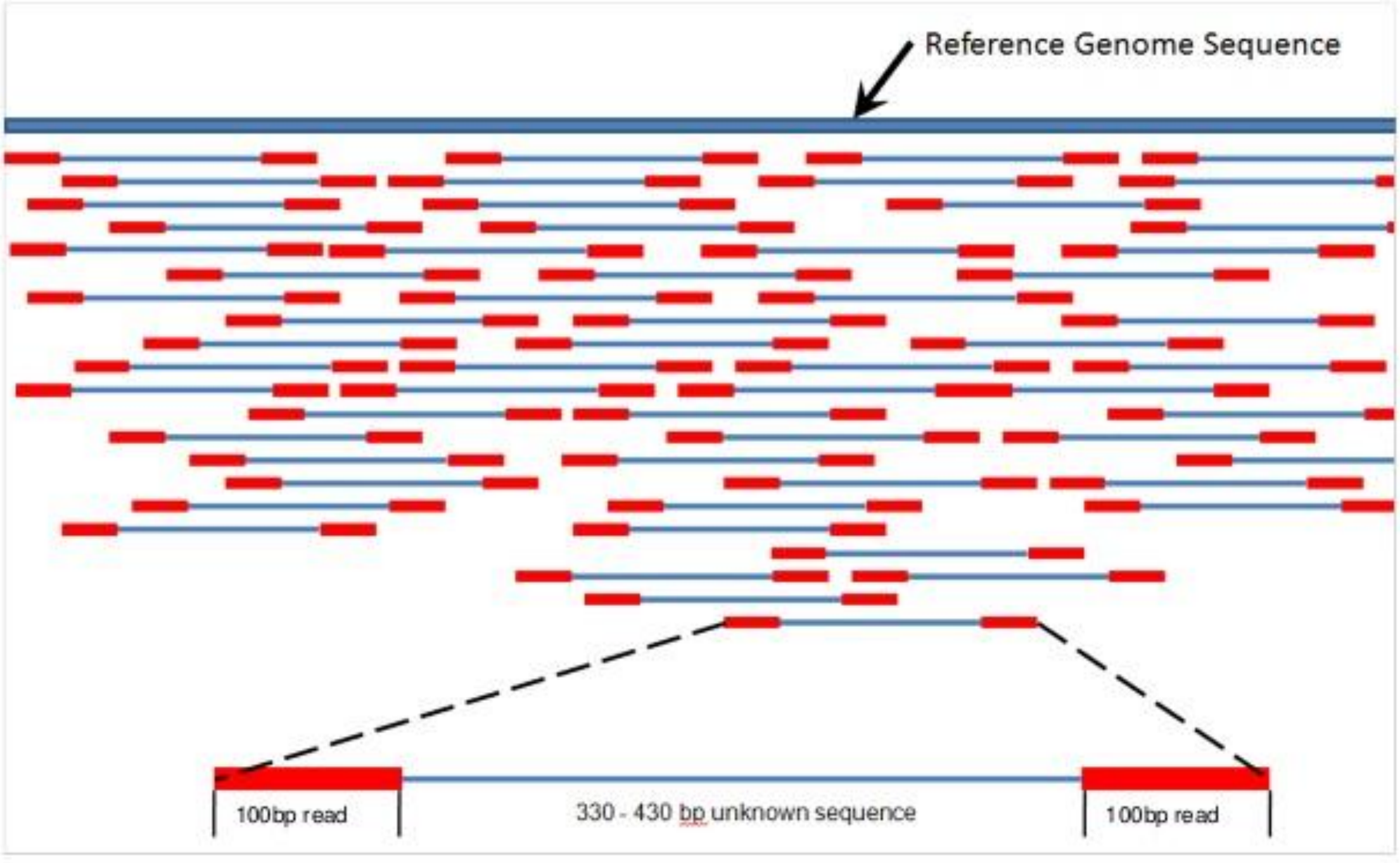
NGS (Next Generation Sequencing)

4. Análise dos dados (Bioinformática)



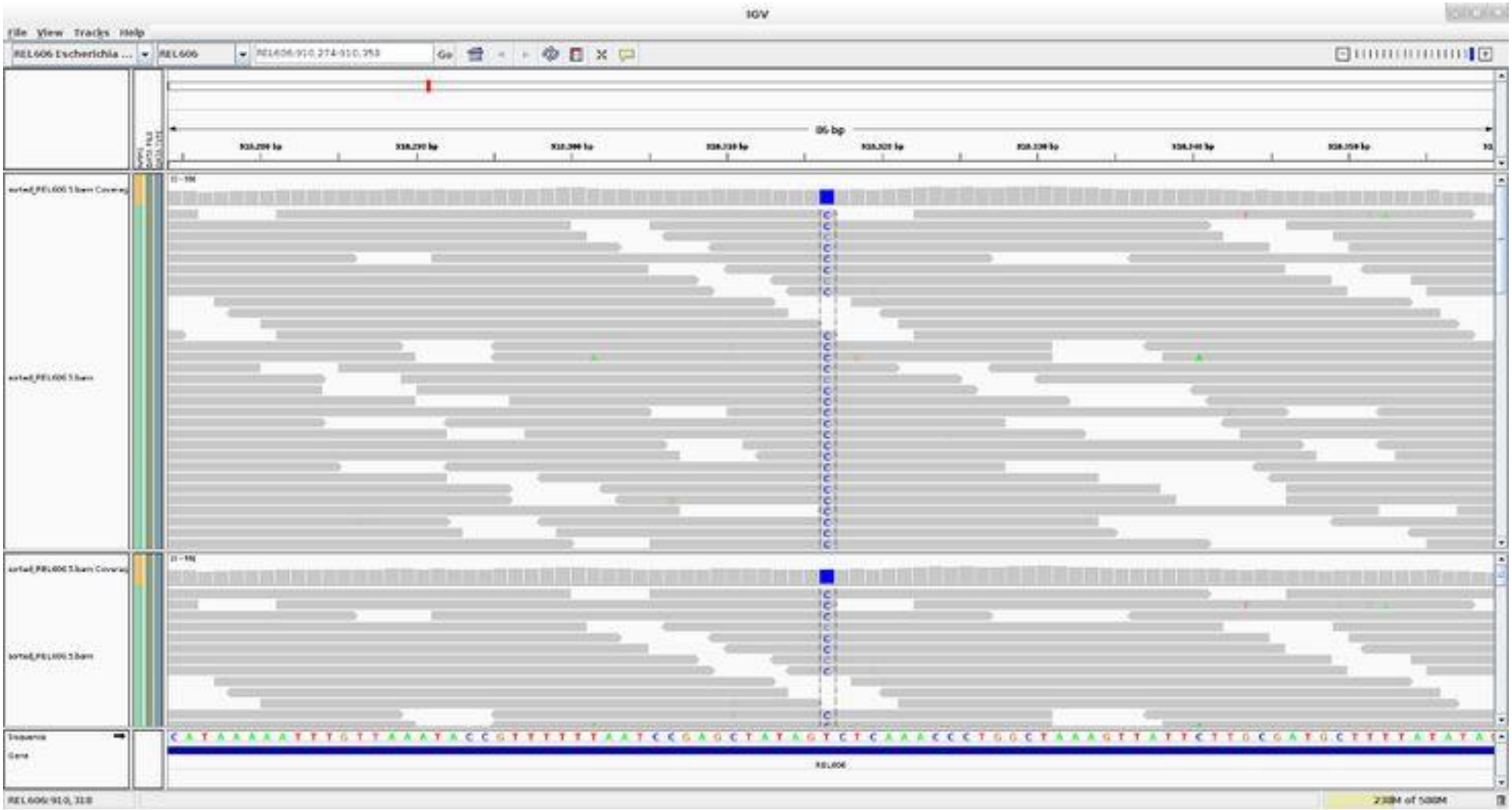
Parâmetros importantes

Cobertura!!!



Parâmetros importantes

Cobertura!!!



Parâmetros importantes

The Lander/Waterman equation

Equação Geral: **C = LN/G**

C: cobertura
L: tamanho da read (sequência)
N: número de reads
G: tamanho do genoma haploide

$C = (100\ bp)*(189\times10^6)/(3\times10^9\ bp) = 6.3$

	 iSeq 100 System	 MiniSeq System	 MiSeq Series	 NextSeq Series
Run Time	9–17.5 hours	4–24 hours	4–55 hours	12–30 hours
Maximum Output	1.2 Gb	7.5 Gb	15 Gb	120 Gb
Maximum Reads Per Run	4 million	25 million	25 million†	400 million
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 × 300 bp	2 × 150 bp

*Com base na equação de Lander e Waterman (C=LN/G), qual seria a cobertura de 10 exomas sequenciados na plataforma NextSeq com tamanho de "reads" (sequencias) de **150** nucleotídeos?*

Informações adicionais:

*1. Os exomas serão sequenciados na plataforma NextSeq, que tem capacidade de gerar **120 Gb** de dados.*

2. Um exoma humano tem o tamanho de 46 Mb.

*Com base na equação de Lander e Waterman (C=LN/G), e considerando uma cobertura do genoma completo de **30x**, quantas amostras podem ser sequenciadas numa corrida com tamanho de "reads" (sequencias) de **300** nucleotídeos?*

Informações adicionais:

*1. Os genomas serão sequenciados na plataforma NextSeq, que tem capacidade de gerar **120 Gb** de dados.*

2. Um genoma humano tem o tamanho de 3 Gb.