



# Introdução à Bioinformática

Clovis F. Reis - PhD



# Objetivos

- ▶ Definir e apresentar a Bioinformática;
- ▶ Conhecer os processos de obtenção dos dados
- ▶ Obter noções das ferramentas e conhecimentos necessários

# Sumário

- ▶ O que é Bioinformática
  - Visão geral
  - A magnitude dos dados
  - Hardware necessário
- ▶ Obtendo dados
  - Sequenciamento
  - Repositórios públicos
- ▶ Áreas abrangidas
  - Genômica
  - Transcriptômica
  - Metagenômica
  - Proteômica ...

# O que é Bioinformática

- ▶ Ciência multidisciplinar
- ▶ Biologia e Informática
- ▶ Coletar, armazenar e analisar dados biológicos

# O que é Bioinformática

- Ciência multidisciplinar
- Biologia e Informática
- Coletar, armazenar e analisar dados biológicos

## Sequenciamento de DNA e RNA

### Whole Genome Sequencing of the Pirarucu (*Arapaima gigas*) Supports Independent Emergence of Major Teleost Clades

Ricardo Assunção Vialle<sup>1,†</sup>, Jorge Estefano Santana de Souza<sup>2,†</sup>, Katia de Paiva Lopes<sup>1</sup>, Diego Gomes Teixeira<sup>2</sup>, Pitágoras de Azevedo Alves Sobrinho<sup>2</sup>, André M. Ribeiro-dos-Santos<sup>1,3</sup>, Carolina Furtado<sup>4</sup>, Tetsu Sakamoto<sup>5</sup>, Fábio Augusto Oliveira Silva<sup>6</sup>, Edivaldo Herculano Corrêa de Oliveira<sup>6</sup>, Igor Guerreiro Hamoy<sup>7</sup>, Paulo Pimentel Assumpção<sup>8</sup>, Ândrea Ribeiro-dos-Santos<sup>1,8</sup>, João Paulo Matos Santos Lima<sup>2,9</sup>, Héctor N. Seuánez<sup>4,10</sup>, Sandro José de Souza<sup>2,11</sup>, and Sidney Santos<sup>1,8,\*</sup>

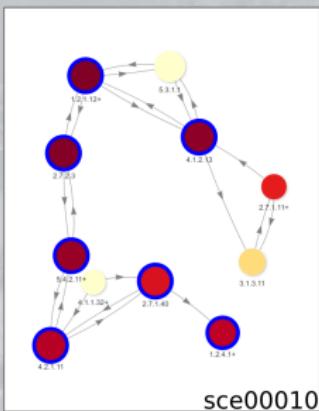
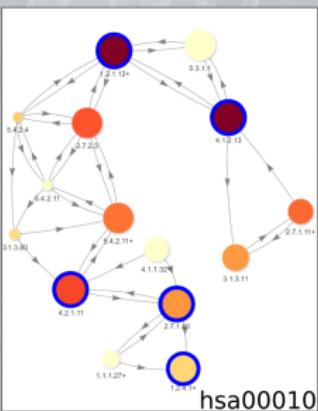
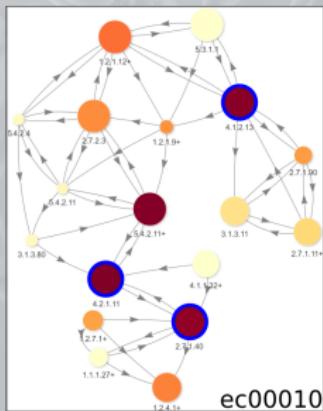
---

2366 *Genome Biol. Evol.* 10(9):2366–2379. doi:10.1093/gbe/evy130 Advance Access publication July 5, 2018

## O que é Bioinformática

- Ciéncia multidisciplinar
  - Biologia e Informática
  - Coletar, armazenar e analisar dados biológicos

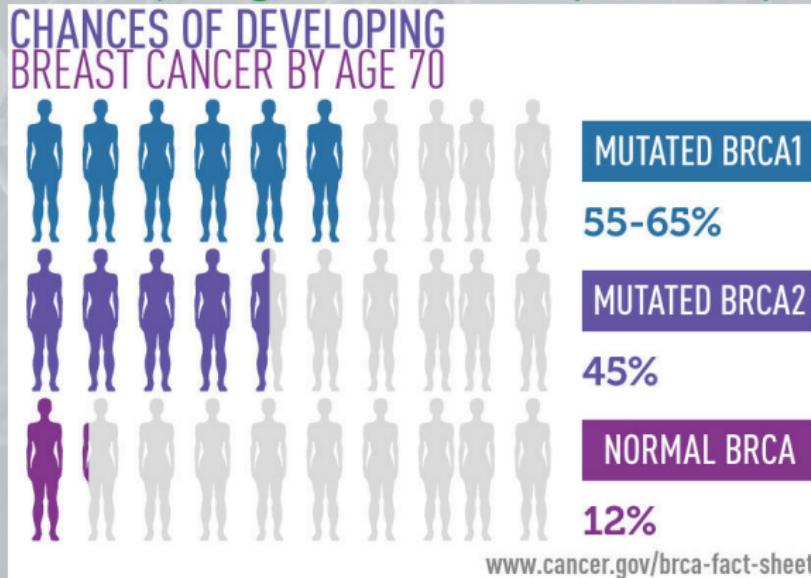
## Compreender funções biológicas complexas



# O que é Bioinformática

- Ciência multidisciplinar
- Biologia e Informática
- Coletar, armazenar e analisar dados biológicos

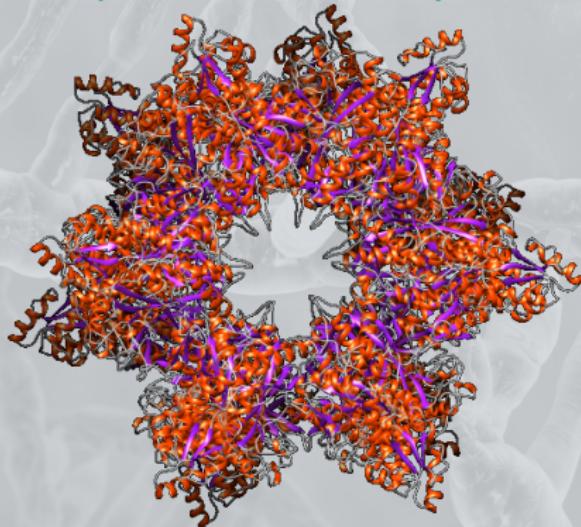
Associar alterações genéticas e de expressão a patologias



# O que é Bioinformática

- ▶ Ciência multidisciplinar
- ▶ Biologia e Informática
- ▶ Coletar, armazenar e analisar dados biológicos

## Predição de estruturas proteicas

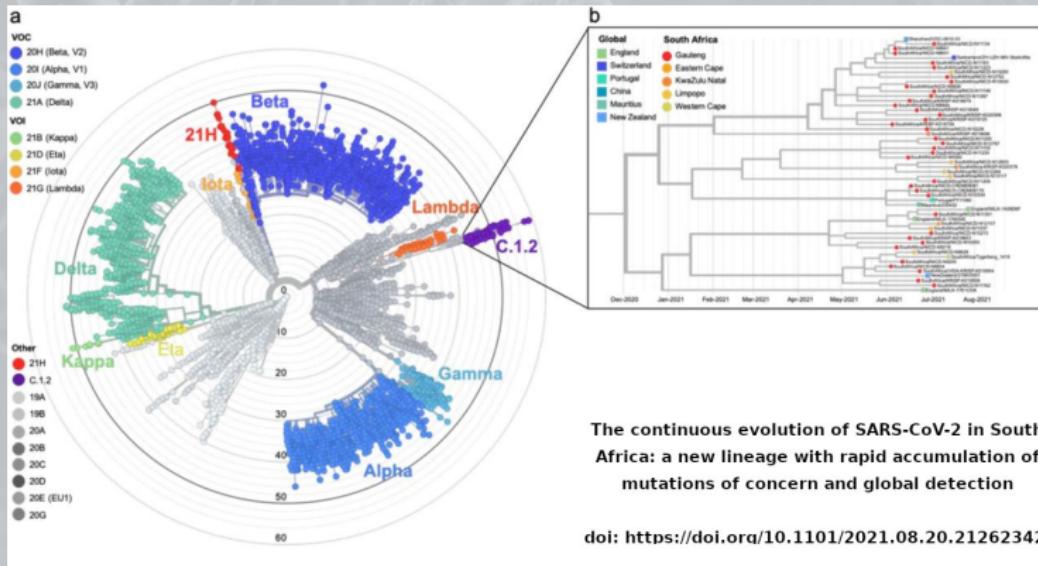


**Glutamine synthetase**

# O que é Bioinformática

- ▶ Ciência multidisciplinar
- ▶ Biologia e Informática
- ▶ Coletar, armazenar e analisar dados biológicos

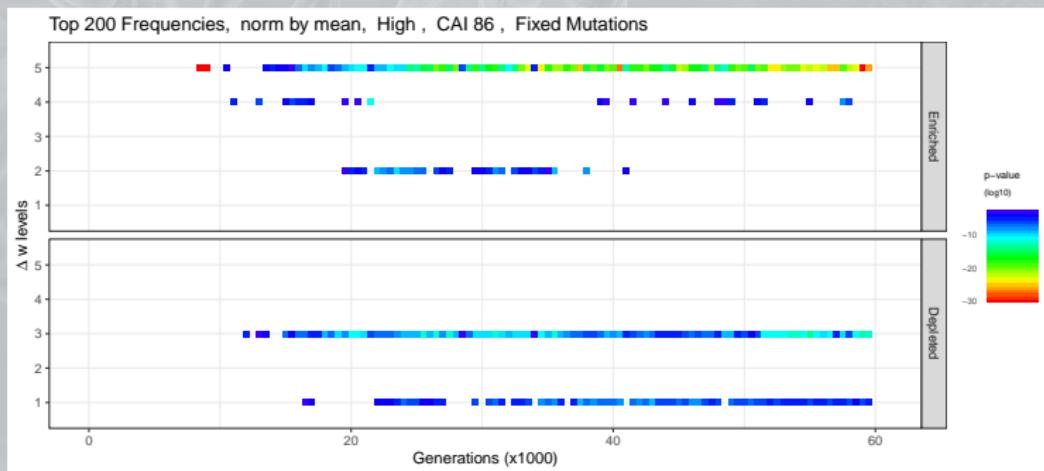
## Construir árvores filogenéticas



# O que é Bioinformática

- ▶ Ciência multidisciplinar
- ▶ Biologia e Informática
- ▶ Coletar, armazenar e analisar dados biológicos

**Estudar a evolução de organismos no tempo**



# O que é Bioinformática

- ▶ Ciência multidisciplinar
- ▶ Biologia e Informática
- ▶ Coletar, armazenar e analisar dados biológicos
  - Sequenciamento de DNA e RNA
  - Compreender funções biológicas complexas
  - Associar alterações genéticas e de expressão a patologias
  - Predição de estruturas proteicas
  - Construir árvores filogenéticas
  - Estudar a evolução de organismos no tempo...
- ▶ Proficiência em:
  - Programação - R e Python
  - Banco de dados
  - Inteligência artificial
  - Estatística
  - Matemática ...

# Aplicabilidade

- ▶ Medicina
  - Prevenção, participação, PREDIÇÃO e PERSONALIZAÇÃO.
  - Dados genéticos ↪ risco de doenças
  - Biomarcadores (BRCA1 e BRCA2 – genes supressores tumorais)
- ▶ Nutrição
- ▶ Desenvolvimento de fármacos
- ▶ Desenvolvimento de vacinas
- ▶ Agricultura
- ▶ Agropecuária
- ▶ Ecologia ...

# A magnitude dos dados

- Um genoma humano de referência ↪ 3.2 GB
- Projeto *Arapaima gigas* ↪ 103 GB
- gnomAD v3 ↪ variantes<sup>†</sup> de 76.156 pessoas

Chr	Tamanho	Chr	Tamanho	Chr	Tamanho
1	260,95 GiB	9	148,36 GiB	17	100,7 GiB
2	274,93 GiB	10	160,43 GiB	18	86,69 GiB
3	224,56 GiB	11	154,64 GiB	19	82,66 GiB
4	222,16 GiB	12	152,53 GiB	20	74,09 GiB
5	202,47 GiB	13	110,43 GiB	21	51,26 GiB
6	193,78 GiB	14	106,29 GiB	22	55,38 GiB
7	189,71 GiB	15	98,83 GiB	X	136,42 GiB
8	175,21 GiB	16	110,94 GiB	Y	8,42 GiB
				Total	3.38 TB

<sup>†</sup>Pequenas diferenças normalmente encontrada nos genomas

# Hardware necessário

- ▶ Grande poder de processamento
- ▶ Grande quantidade de memória  
(mínimo 64GB)
- ▶ Grande disponibilidade de  
armazenamento
- ▶ Sistema operacional robusto ↳ Linux

# Hardware necessário

- ▶ Grande poder de processamento
- ▶ Grande quantidade de memória (mínimo 64GB)
- ▶ Grande disponibilidade de armazenamento
- ▶ Sistema operacional robusto ↳ Linux
- ▶ Uso de máquinas do BioME

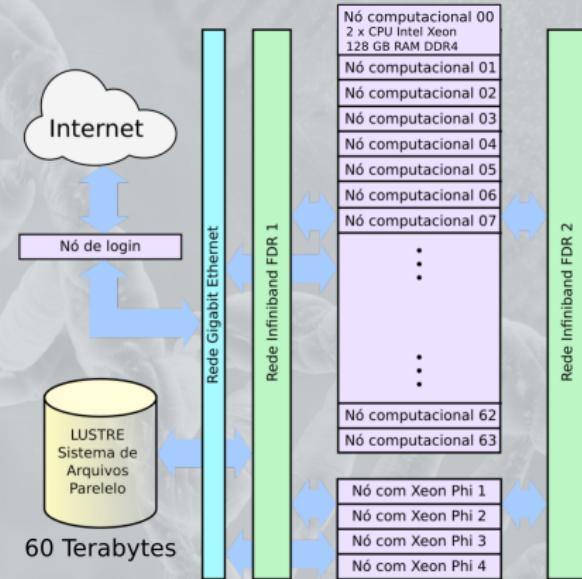
## PROCESSAMENTO

CPU	CORES	MEM (GB)	DISCO
Intel Xeon	48	256	600 GB
Intel Xeon	48	382	600 GB
Intel Xeon	40	240	1.2 TB
Intel Xeon	40	240	1.2 TB
Intel Xeon	24	125	250 GB
AMD Opteron	16	40	292 GB
Intel Xeon	12	40	600 GB
2-Xeon/12	12	64	600 GB
AMD Opteron	12	62	292 GB
Virtual	16	32	20 GB
Virtual	16	32	20 GB
Virtual	16	32	20 GB
Virtual	16	32	20 GB
Total	332	1.6 TB	5.7 TB

**STORAGES:** 130 TB

# Hardware necessário

- Grande poder de processamento
- Grande quantidade de memória (mínimo 64GB)
- Grande disponibilidade de armazenamento
- Sistema operacional robusto ↳ Linux
- Uso de máquinas do BioME
- Uso de núcleos do NPAD



# E por que o Linux?

A maioria dos softwares de Bioinformática rodam Linux e Windows!

Sistema operacional muito robusto

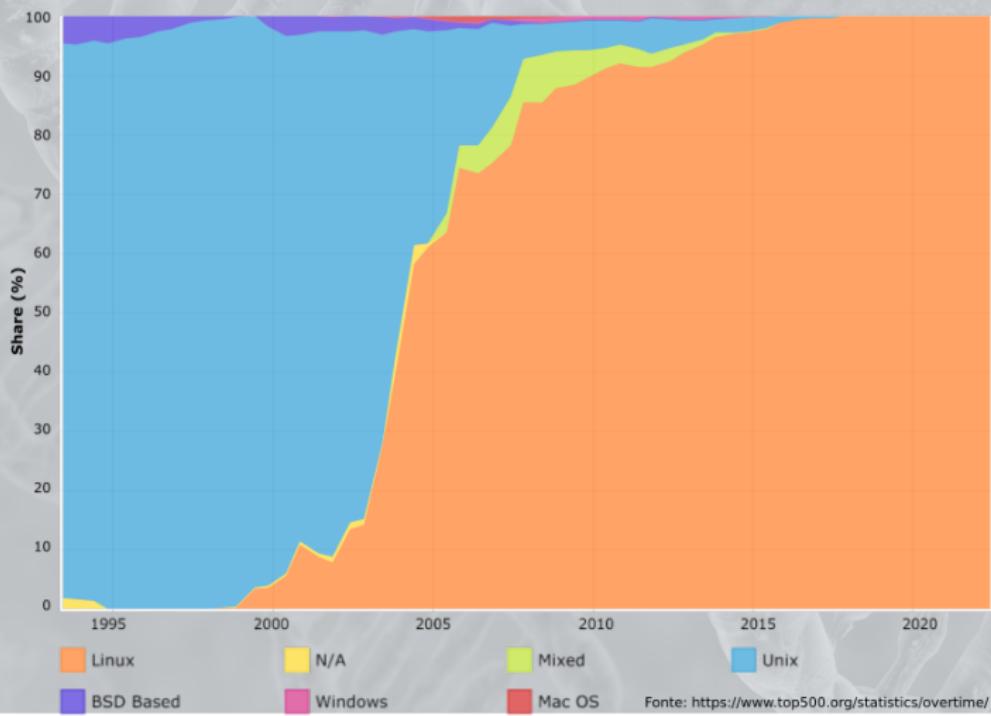
- Baixa vulnerabilidade

- ★ Praticamente imune a vírus
- ★ Menos falhas de segurança
- ★ Correção de falhas mais rápida e eficiente
- ★ Falhas de segurança não comprometem toda a máquina

# E por que o Linux?

► Confiabilidade

Uso em Super computadores



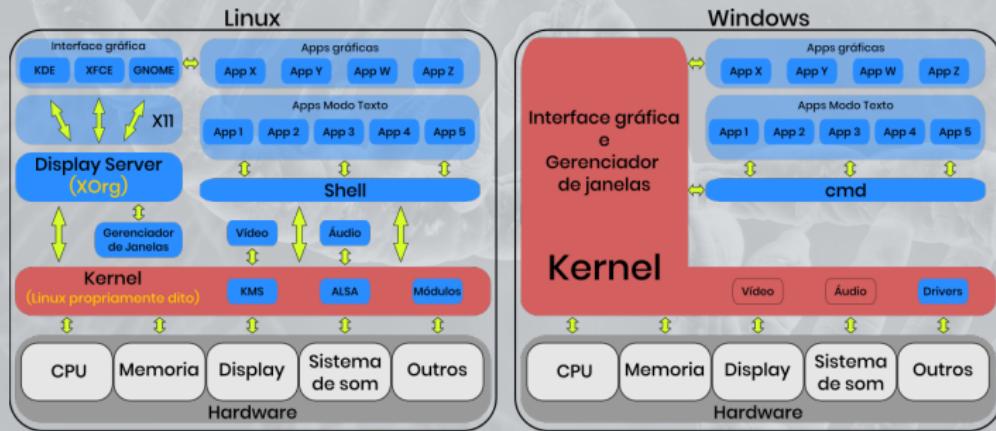
# E por que o Linux?

## ► Estabilidade e disponibilidade

- ★ Funciona sem intervenção por muito de tempo
- ★ Raros problemas de travamento
- ★ Raramente é necessária um reinstalação

# E por que o Linux?

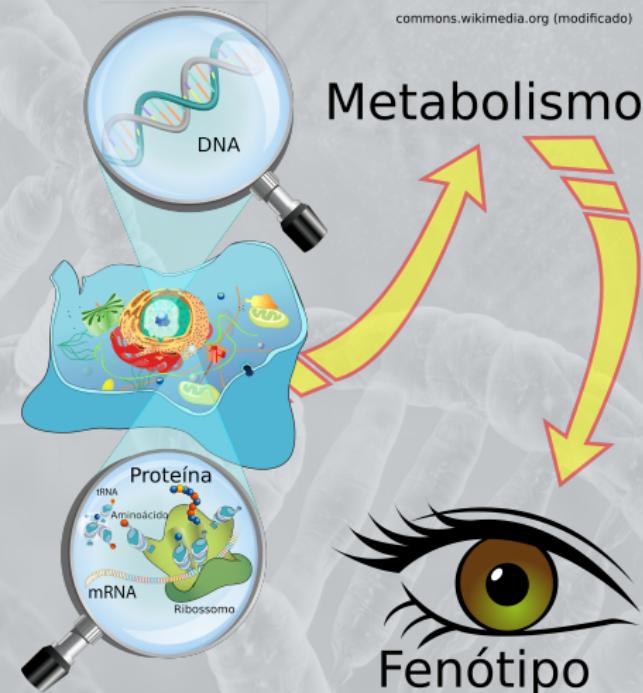
- Modularidade
  - ★ Aproveita melhor o hardware



# Algumas sub-áreas da Bioinformática

commons.wikimedia.org (modificado)

- ▶ Genômica – DNA
- ▶ Transcriptômica – mRNA
- ▶ Proteômica – Proteínas
- ▶ Biologia de Sistemas – Vias Metabólicas
- ▶ Metagenômica – Múltiplos organismos



# Obtenção de dados

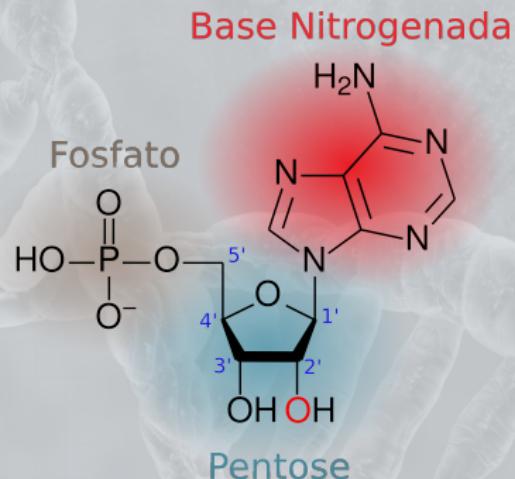


# Fundamentos em Genética

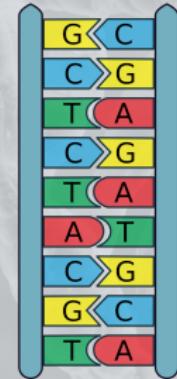
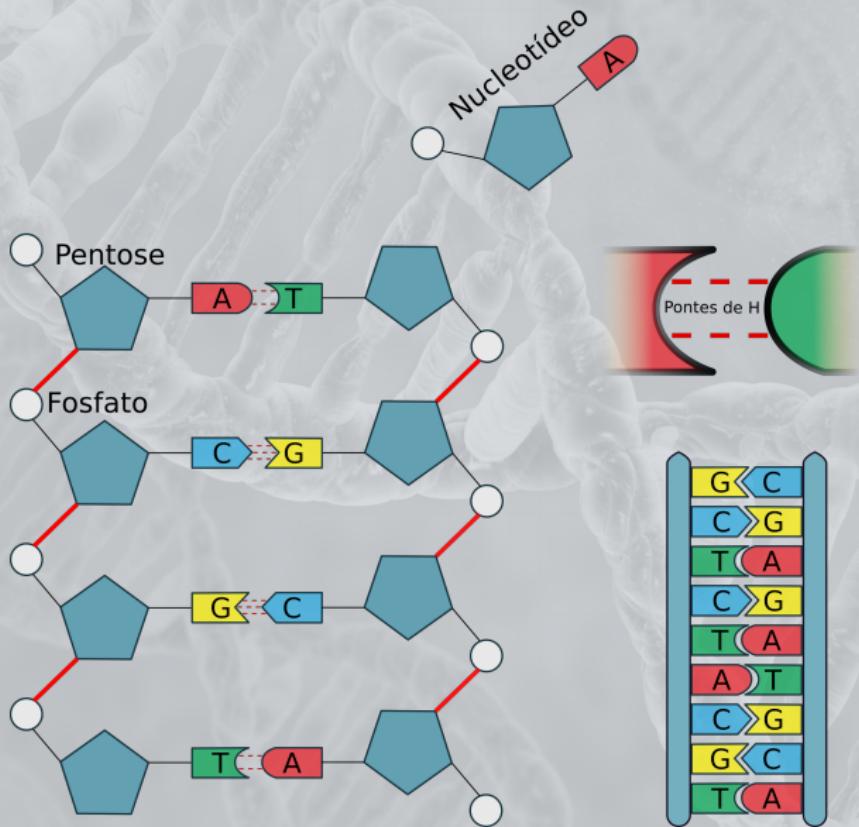
# Fundamentos em Genética – Nucleotídeos

Formado por três moléculas:

- ▶ Base nitrogenada
  - ★ Purinas  
Adenina (A) e  
Guanina (G)
  - ★ Pirimidinas  
Citosina (C),  
Timina (T) e  
**Uracila (U)**
- ▶ Grupo fosfato – não varia
- ▶ Pentose – açúcar com 5 carbonos (ribose)
  - ★ DNA – desoxirribose
  - ★ RNA – ribose

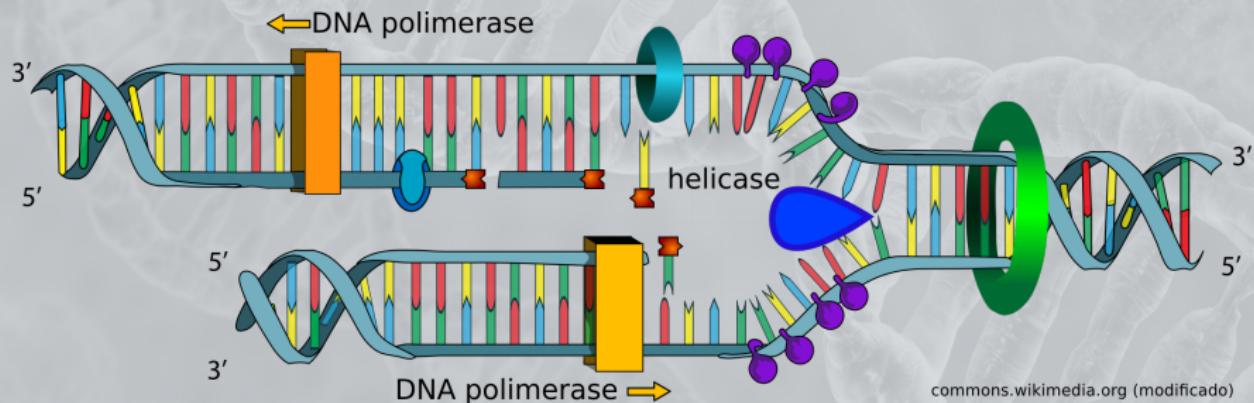


# Fundamentos em Genética – DNA



# Replicação do DNA

- A fita dupla é aberta (desnaturada);
- Nucleotídeos ligam-se aos seus pares;
- A nova fita é polimerizada e fecha-se;
- **Duas fitas duplas idênticas.**

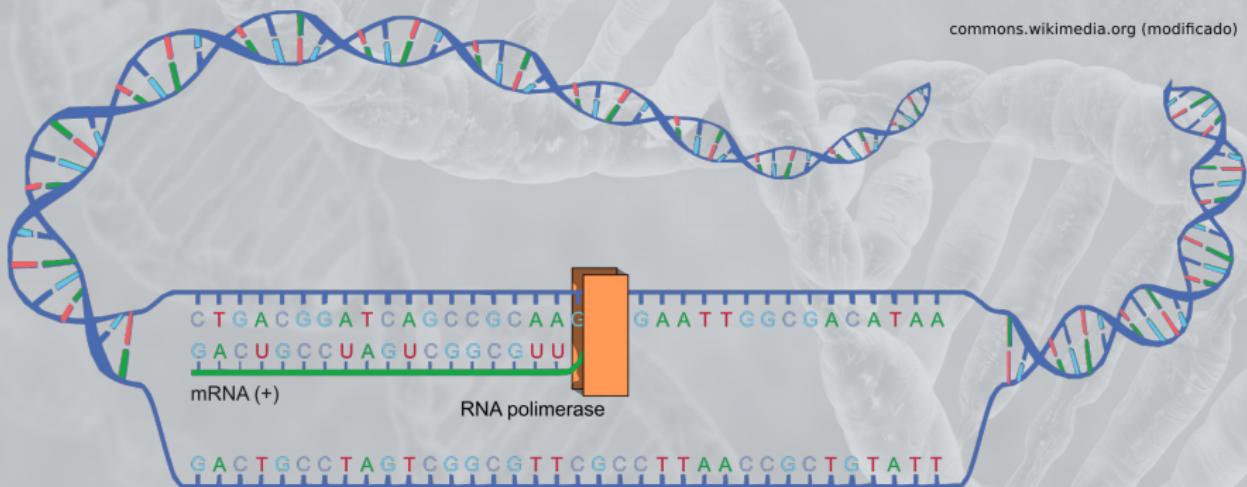


commons.wikimedia.org (modificado)

# Transcrição de RNA

- ▶ A fita dupla é aberta;
- ▶ Nucleotídeos ligam-se aos seus pares:  
Adenina (A) liga-se a uma Uracila (U)
- ▶ O RNA é polimerizado;
- ▶ Cópia invertida de um dos lados da fita.

commons.wikimedia.org (modificado)



# Tradução de proteínas

- Após o *splicing* o mRNA sai do núcleo e se liga a um ribossomo;

- Gene

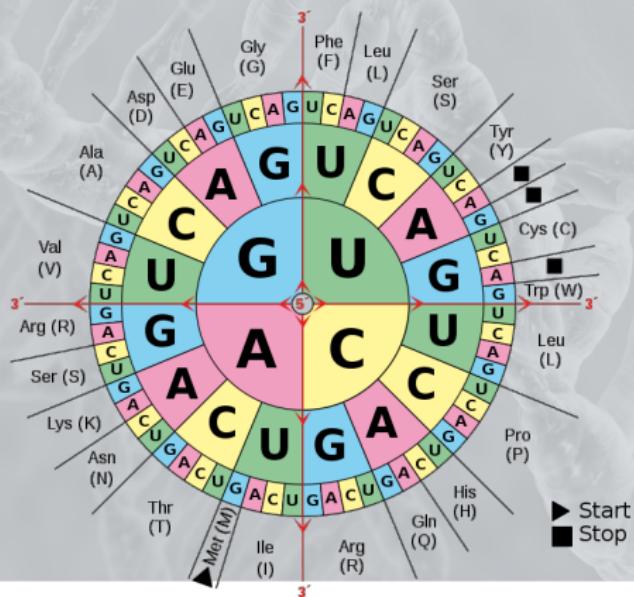
- ★ Grupo de nucleotídeos que codificam uma proteína.

- ★ Grupados de 3 em 3

- ★ Inicia-se por um *start* códon e finaliza em um *stop* códon

- Códon → Grupo de três nucleotídeos que codificam um dos 20 amino-ácidos;

- Proteína → Conjunto de N amino-ácidos.



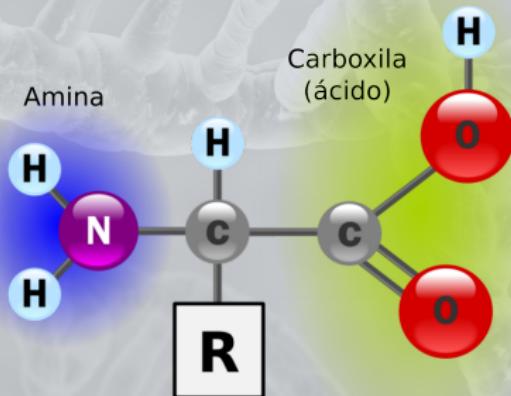
# Tradução de proteínas

- Após o *splicing* o mRNA sai do núcleo e se liga a um ribossomo;

- Gene

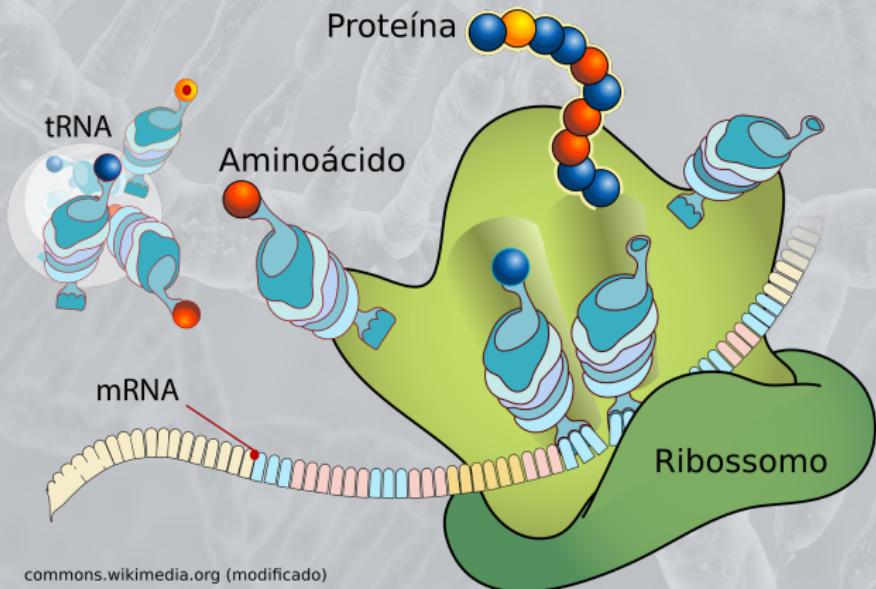
- ★ Grupo de nucleotídeos que codificam uma proteína.
    - ★ Grupados de 3 em 3
    - ★ Inicia-se por um *start* códon e finaliza em um *stop* códon

- Códon  $\mapsto$  Grupo de três nucleotídeos que codificam um dos 20 amino-ácidos;
  - Proteína  $\mapsto$  Conjunto de N amino-ácidos.



# Tradução de proteínas

- Após o *splicing* o mRNA sai do núcleo e se liga a um ribossomo;
- O ribossomo une os amino-ácidos codificados no mRNA e cria uma proteína;
- O mRNA é reciclado.



commons.wikimedia.org (modificado)



# Obtendo os dados

## Sequenciamento Genético

# Operações – Desnaturar o DNA



# Operações – Amplificar o DNA

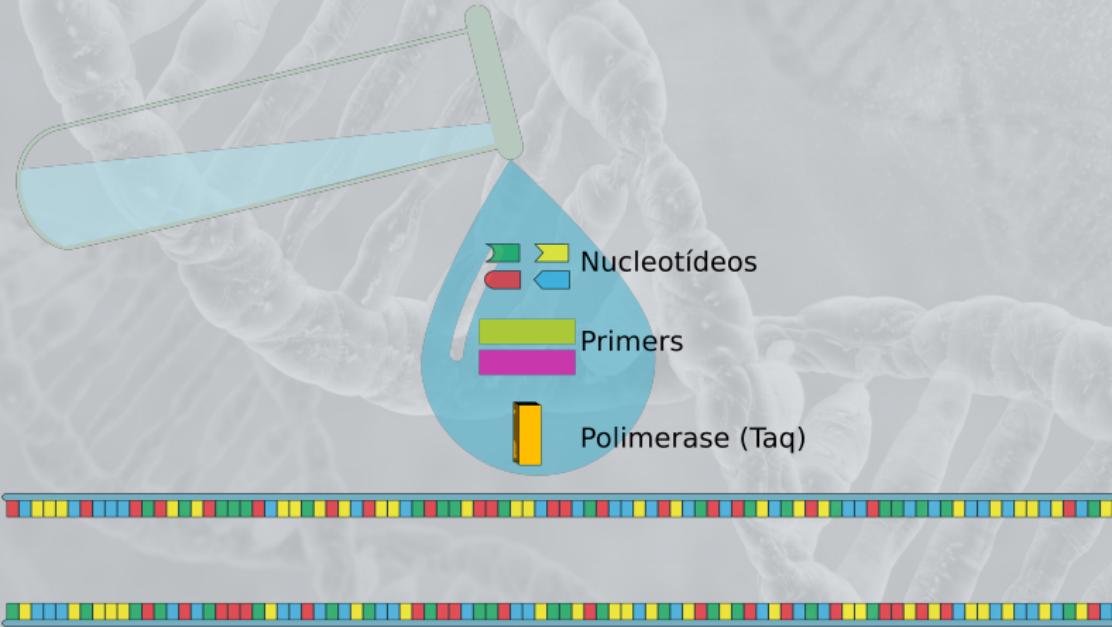
- ▶ Multiplicar o DNA extraído;
- ▶ Técnica da Polymerase Chain Reaction (PCR);
- ▶ Obtenção de N cópias das fitas originais.

# Operações – Amplificar o DNA

Desnaturação (95° C)



# Operações – Amplificar o DNA



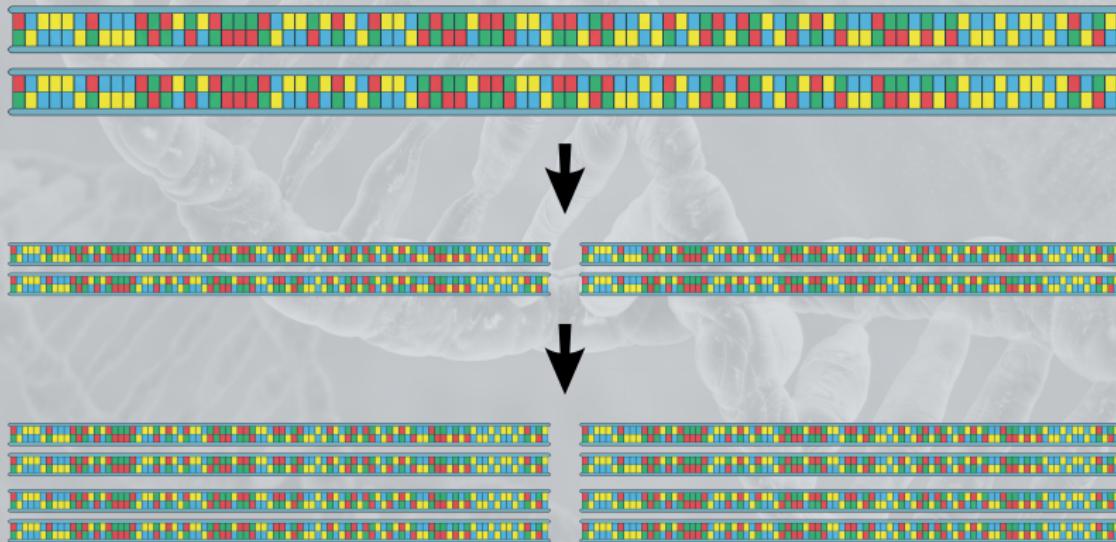
# Operações – Amplificar o DNA

Annealing ( $55^{\circ} \text{ C}$ ) e Síntese ( $72^{\circ} \text{ C}$ )



# Operações – Amplificar o DNA

Repeite-se o ciclo (20 a 40x)



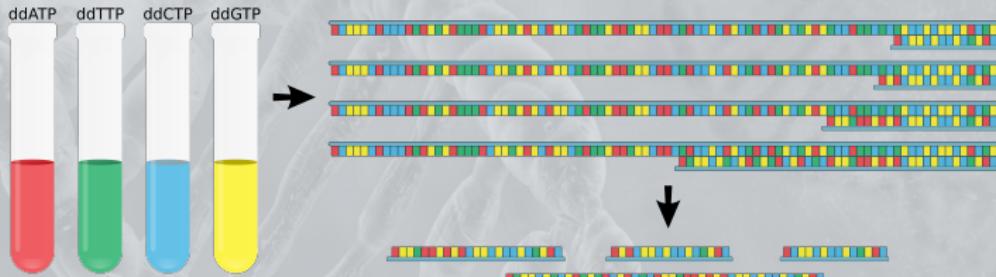
# Um breve histórico (Chain-Terminating PCR)

## ► Frederick Sanger

- 1955 → sequência de aminoácidos de insulina;
- 1975 → método de sequenciamento do DNA (método de Sanger);
- 1977 → sequenciamento completo de um vírus.



# Um breve histórico (Chain-Terminating PCR)



Sequência  
A G C T G C T A T T A C C G T

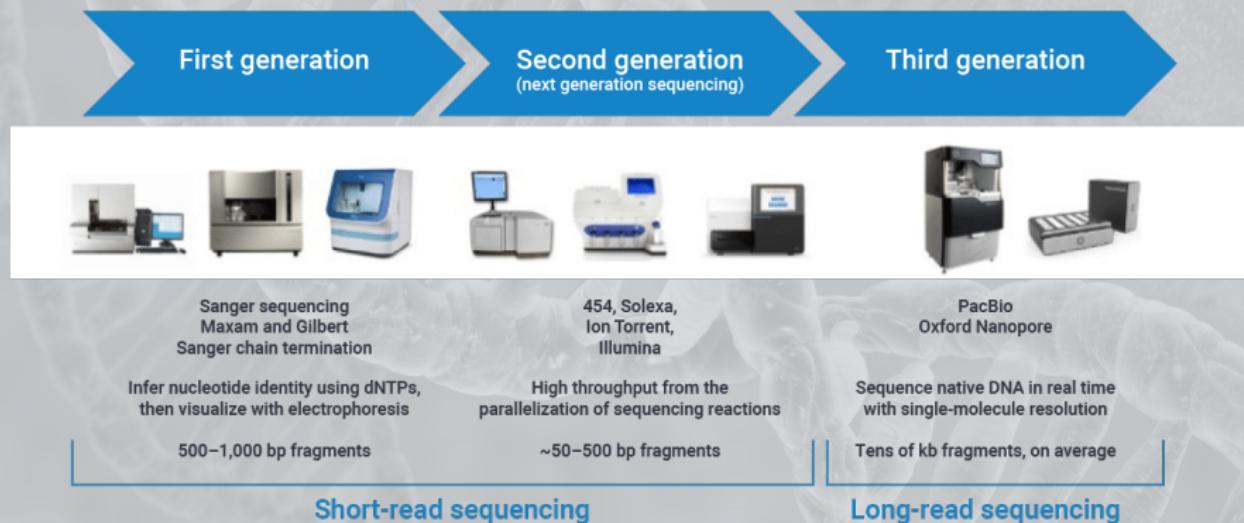
Chain-Terminating PCR



Eletroforese em Gel

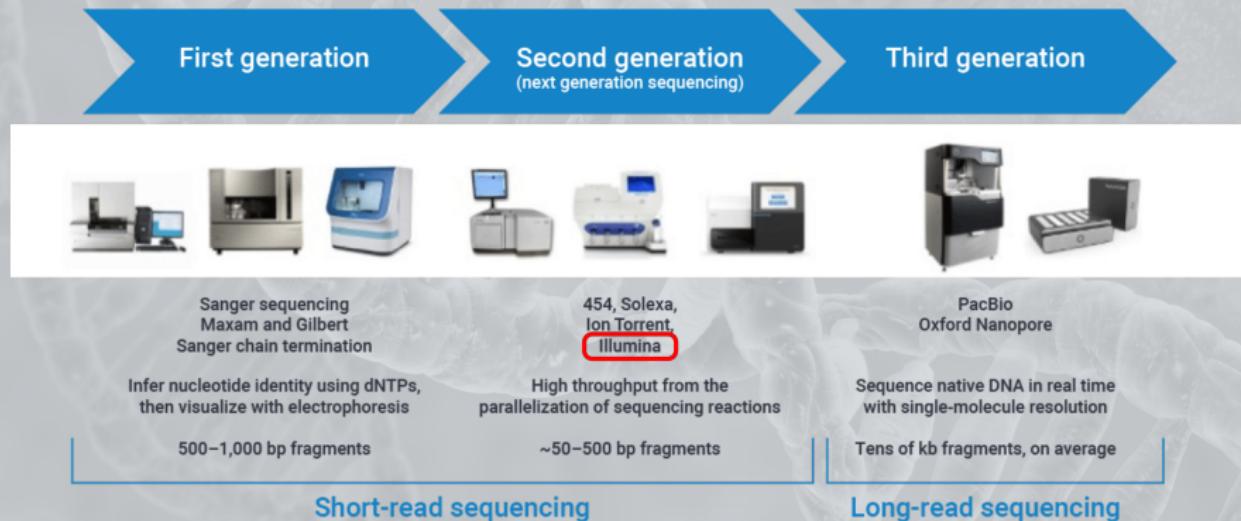


# A evolução dos sequenciadores



<https://www.pacb.com/blog/the-evolution-of-dna-sequencing-tools/>

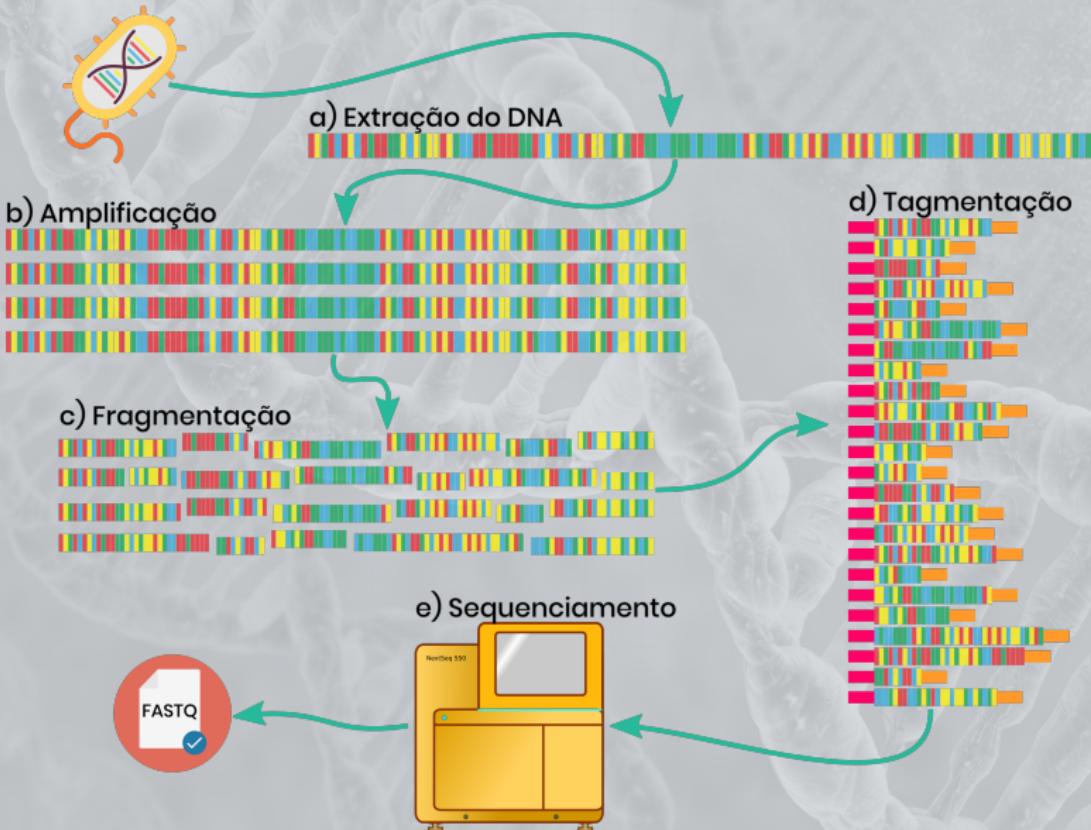
# A evolução dos sequenciadores



<https://www.pacb.com/blog/the-evolution-of-dna-sequencing-tools/>

# Sequenciamento Illumina

# Visão geral



# Coleta de material que contenha DNA

## ► Genoma:

- Material que contenha DNA (sangue, saliva...);
- Único organismo.

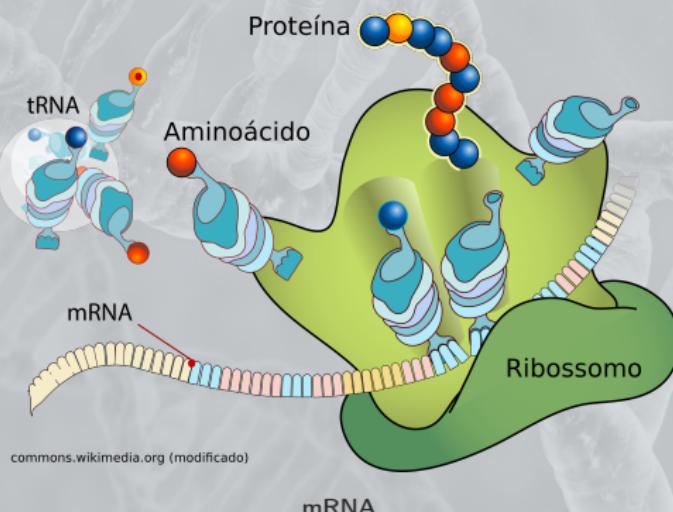


Coleta para genoma

# Coleta de material que contém DNA

## ► Transcriptoma:

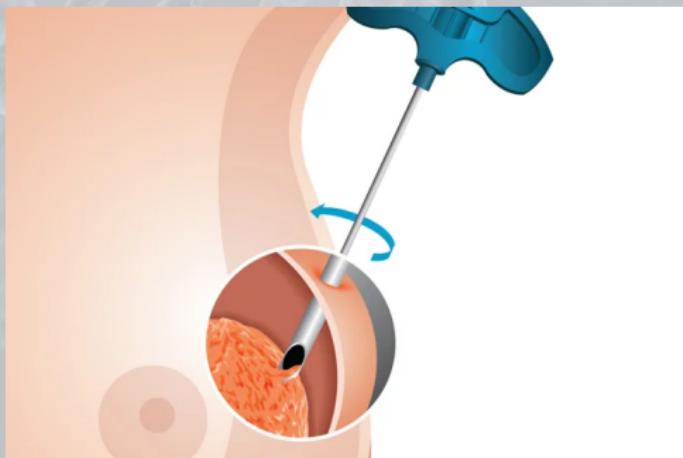
- Observar o que as células estão produzindo no momento, coletando o mRNA;



# Coleta de material que contenha DNA

## ► Transcriptoma:

- Observar o que as células estão produzindo no momento, coletando o mRNA;
- Tecido vivo;
- Coleta-se o tecido saudável e o doente.

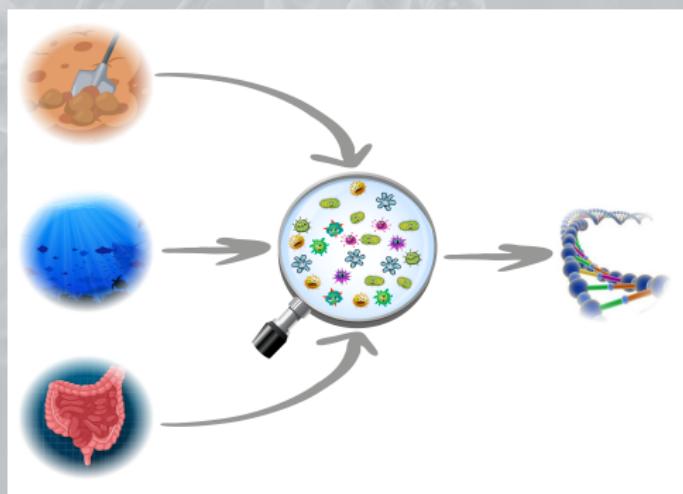


Biópsia

# Coleta de material que contenha DNA

## ► Metagenoma:

- Obtenção de DNA de vários organismos;
- Identificar a microbiota do local;
- Coleta de material que contenha múltiplos tipos de DNA (solo, água, fezes...);



Coleta de múltiplos organismos

# Extração do DNA

(comum a todas áreas)

- ▶ Extração do DNA;



Extração

# Extração do DNA (comum a todas áreas)

- ▶ Extração do DNA;
- ▶ Obtenção de fitas de DNA de várias células/organismos.



Fita de DNA

# Amplificação do DNA



# Fragmentação do DNA

- O equipamento tem restrição quanto ao comprimento do DNA;

# Fragmentação do DNA

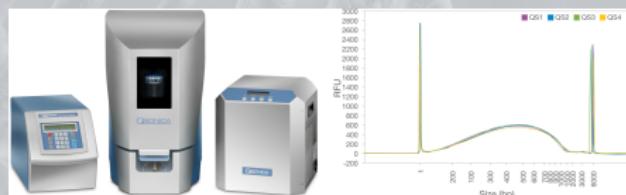
- O equipamento tem restrição quanto ao comprimento do DNA;
- Fragmentação por processos químicos ou mecânicos (Sonicator ↔ ultrassom);



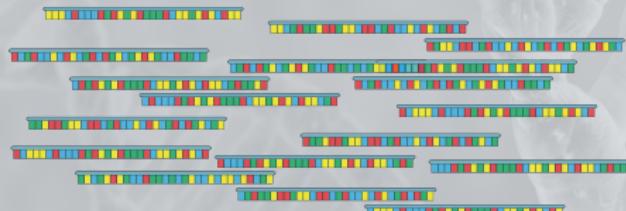
Sonicator

# Fragmentação do DNA

- O equipamento tem restrição quanto ao comprimento do DNA;
- Fragmentação por processos químicos ou mecânicos (Sonicator ↔ ultrassom);
- Obtém-se fitas fragmentadas no tamanho desejado.



Sonicator



Fragmentos de DNA

# Inserção de adaptadores

- Os equipamentos da Illumina sequenciam de forma automática;



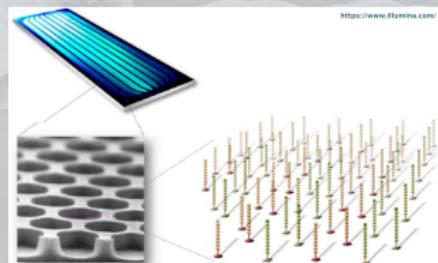
Sequenciador

# Inserção de adaptadores

- ▶ Os equipamentos da Illumina sequenciam de forma automática;
- ▶ O DNA adere a uma lâmina especial (Flow Cell);



Sequenciador



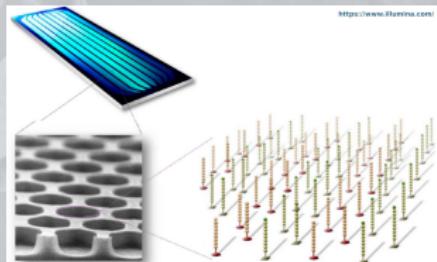
Flow Cell

# Inserção de adaptadores

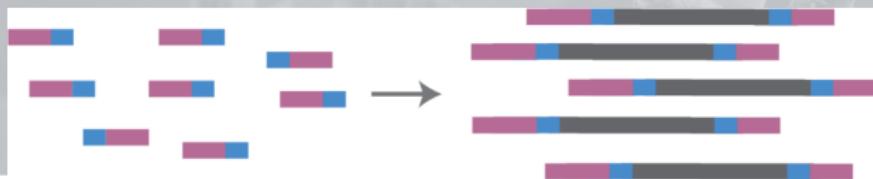
- ▶ Os equipamentos da Illumina sequenciam de forma automática;
- ▶ O DNA adere a uma lâmina especial (Flow Cell);
- ▶ Adaptadores ↳ fragmentos de DNA com sequência específica. Ligam-se ao DNA da Flow Cell.



Sequenciador



Flow Cell

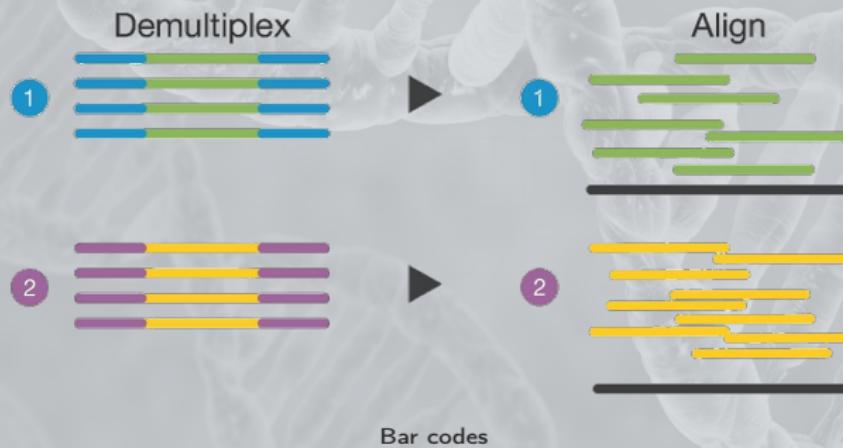


[www.illumina.com/technology/next-generation-sequencing.html](http://www.illumina.com/technology/next-generation-sequencing.html)



# Inserção de adaptadores

- ▶ Os equipamentos da Illumina sequenciam de forma automática;
- ▶ O DNA adere a uma lâmina especial (Flow Cell);
- ▶ Adaptadores ↳ fragmentos de DNA com sequência específica. Ligam-se ao DNA da Flow Cell.
- ▶ Bar codes ↳ permite o sequenciamento de mais de uma amostra por corrida (multiplexação).



# Preparação da *Flow Cel* e amplificação dos clusteres

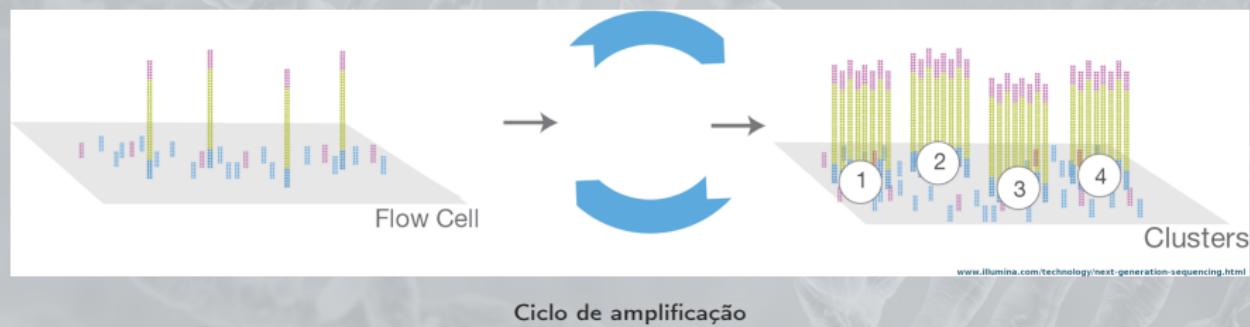
- Realizado automaticamente pelo sequenciador;
- Kit sob a forma de cartucho.



Kit de reagentes

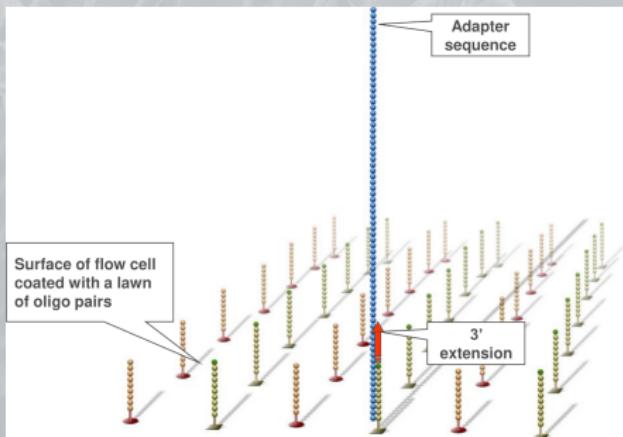
# Preparação da *Flow Cel* e amplificação dos clusteres

- O DNA é amplificado dentro dos clusteres ↳ utiliza emissão de luz;



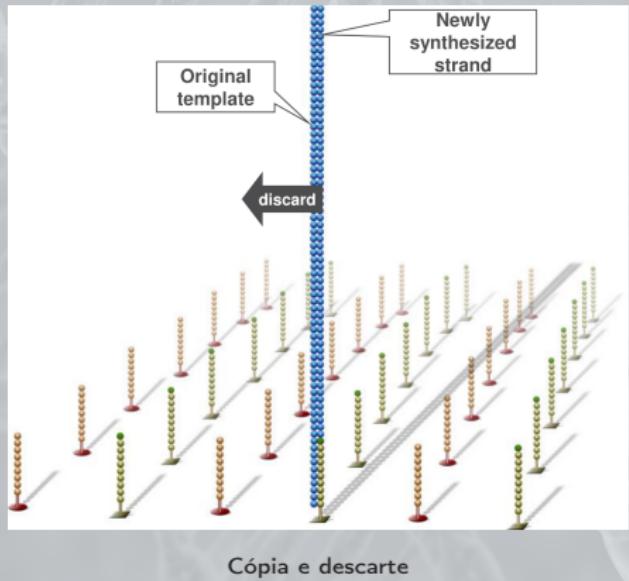
# Preparação da *Flow Cel* e amplificação dos clusteres

- Cada cluster conterá um fragmento de DNA ligado ao adaptador;



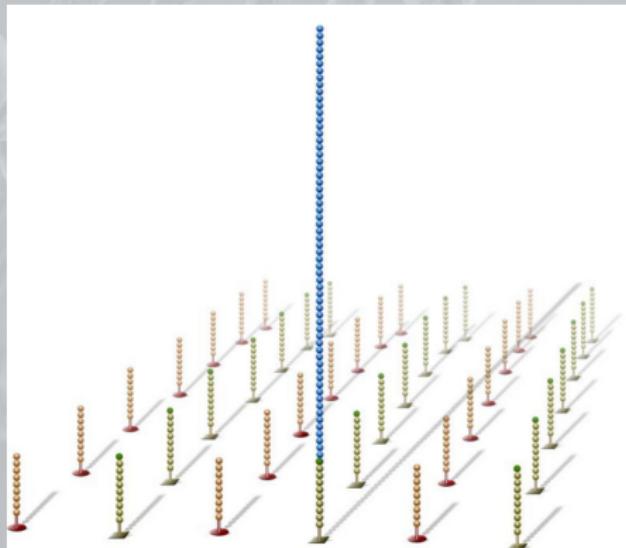
# Preparação da *Flow Cel* e amplificação dos clusteres

- Uma fita complementar é criada e a original é descartada;



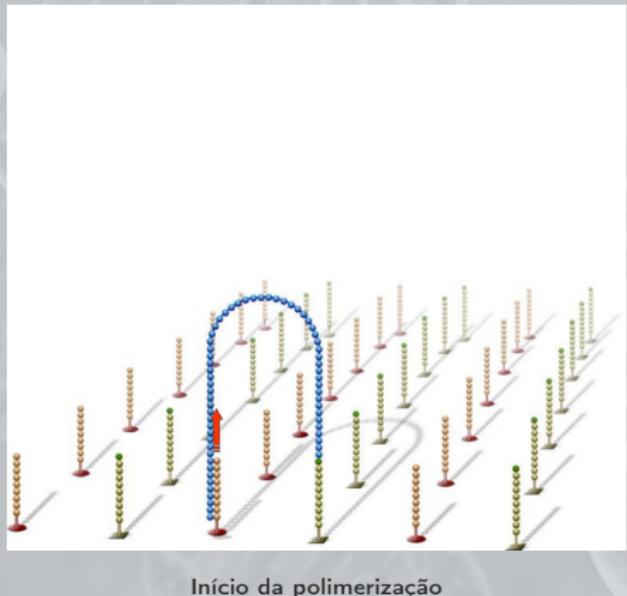
# Preparação da *Flow Cel* e amplificação dos clusteres

- Inicia-se a amplificação;



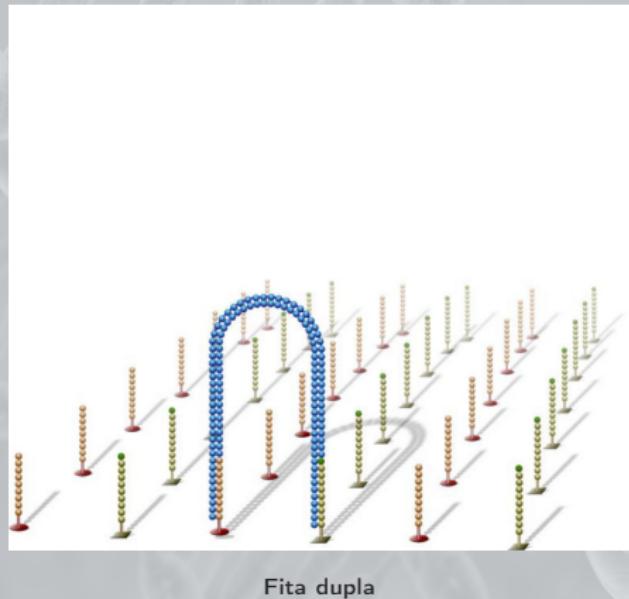
# Preparação da *Flow Cel* e amplificação dos clusteres

- A fita é curvada e a polimerização é realizada;



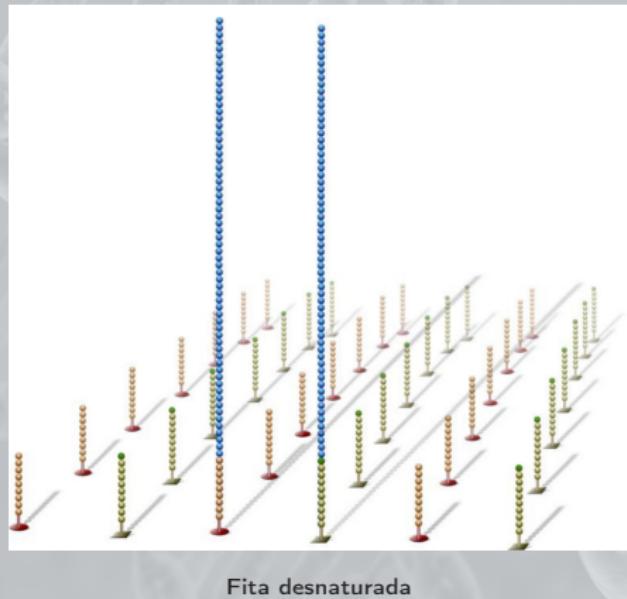
# Preparação da *Flow Cel* e amplificação dos clusteres

- Ao final do processo obtém-se uma fita dupla;



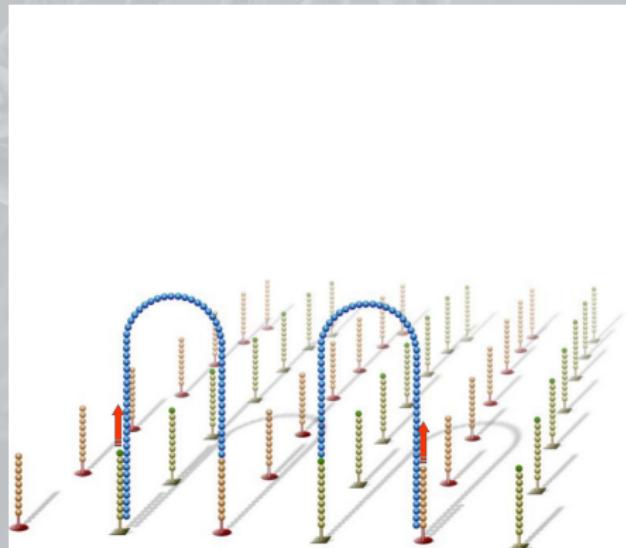
# Preparação da *Flow Cel* e amplificação dos clusteres

- A fita dupla é desnaturada (separada) e obtém-se duas fitas simples;



# Preparação da *Flow Cel* e amplificação dos clusteres

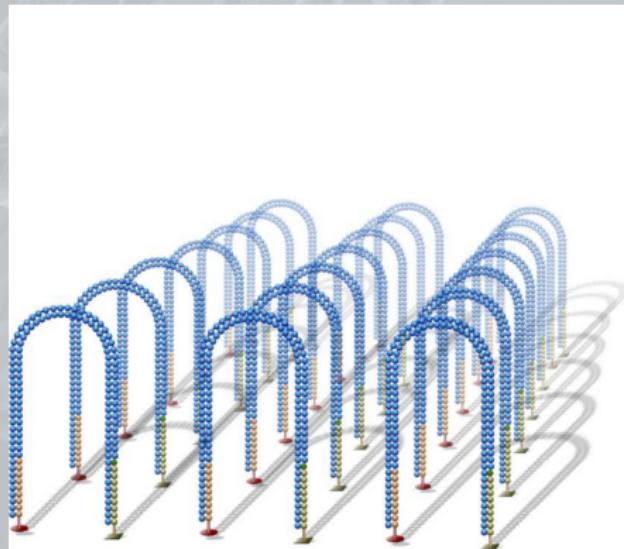
- Nova curvatura e polimerização é realizada;



Curvatura e polimerização

# Preparação da *Flow Cel* e amplificação dos clusteres

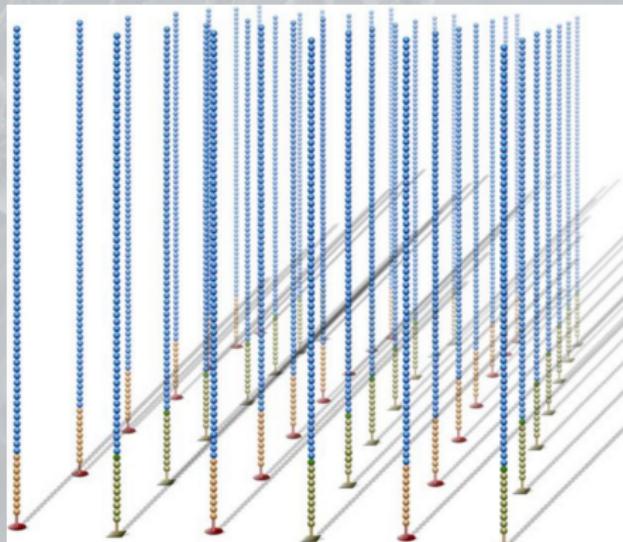
- O processo se repete;



Vários ciclo de amplificação

# Preparação da *Flow Cel* e amplificação dos clusteres

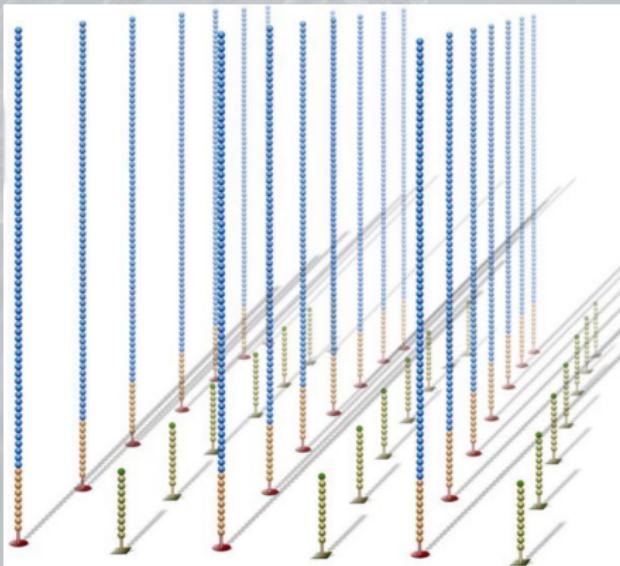
- O cluster é amplificado com *strands forward* e reversas;



Cluster amplificado

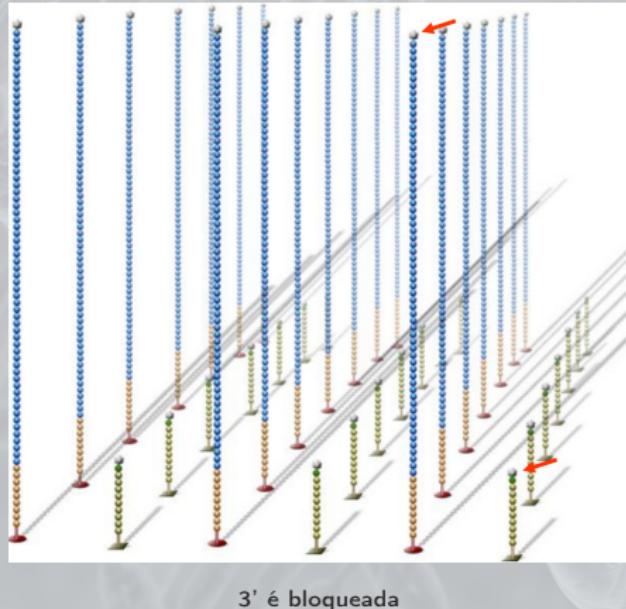
# Preparação da *Flow Cel* e amplificação dos clusteres

- As *strands* reversas são eliminadas;



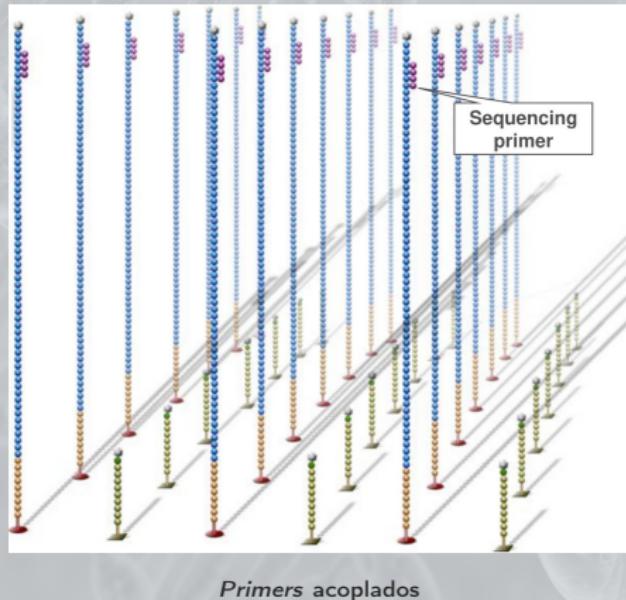
# Preparação da *Flow Cel* e amplificação dos clusteres

- A extremidade 3' é bloqueada para evitar primerização indesejada;



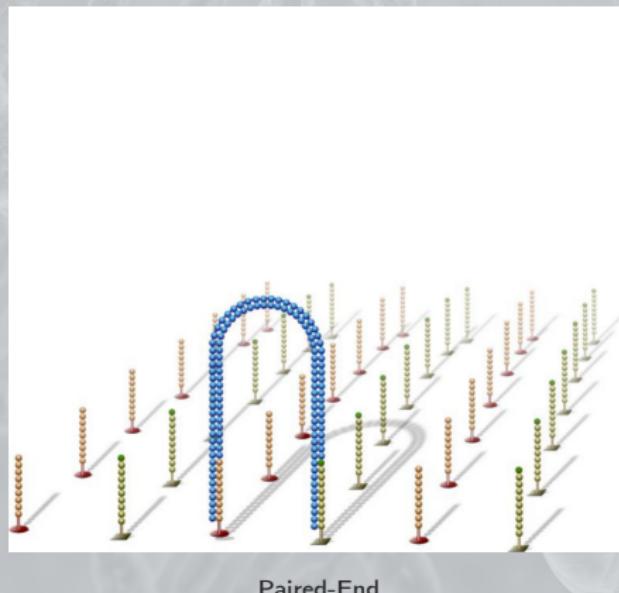
# Preparação da *Flow Cel* e amplificação dos clusteres

- Os primers são adicionados e o sequenciamento por síntese será realizado.



# Preparação da *Flow Cel* e amplificação dos clusteres

- Caso o sequenciamento seja *Paired-End*, após o sequenciamento da *strand forward* o processo é reiniciado, sendo que a *strands forward* será descartada, sequenciando-se a *strand reversa*.



# Sequenciamento por Síntese (Sequencing By Synthesis –SBS)

- Também é realizado automaticamente pelo sequenciador;

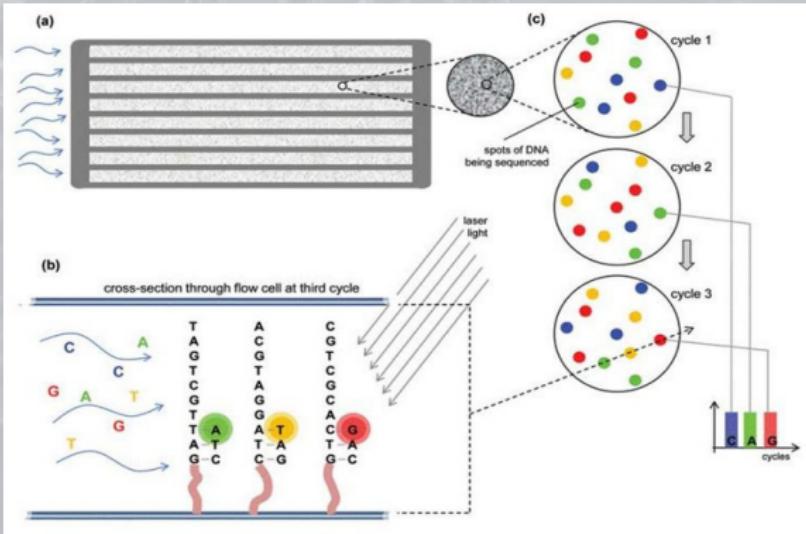
# Sequenciamento por Síntese (Sequencing By Synthesis –SBS)

- Cada nucleotídeo está associado a um fluoróforo (emissor de luz) de cor específica;



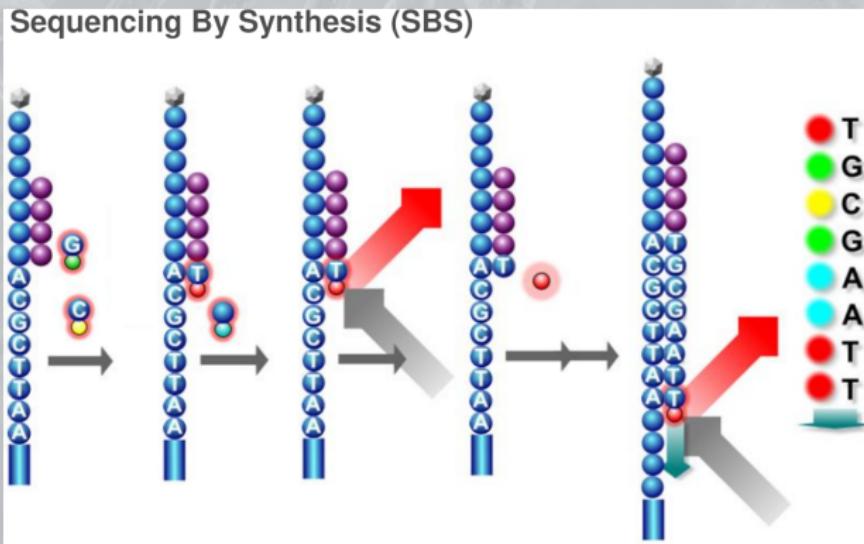
# Sequenciamento por Síntese (Sequencing By Synthesis –SBS)

- Cada cluster conterá um fragmento de DNA;



# Sequenciamento por Síntese (Sequencing By Synthesis –SBS)

- A fita é curvada e a polimerização iniciada;



# Sequenciamento por Síntese (Sequencing By Synthesis –SBS)

- Ao final do processo obtém-se um arquivo texto (FASTQ) contendo:
  - A sequência lida em cada cluster;
  - Informações sobre a qualidade da leitura (certeza da informação).

```
@ML-P2-14:9:000H003HG:1:11102:17290:1073 1:N:0:TCCTGAGC+GCGATCTA
TTGGTAACAGCATGAATTATTCTAGCCACTAAACTCTATGAACATCTTGAGGTTTAGATAGAGCCTGAAGTACACAGAGAACATTCTAAAAAA
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE<AAAAA
```

# Repositórios públicos

- ▶ A maioria das revistas exigem a disponibilização dos *datasets* de sequenciamento;
- ▶ Repositórios públicos ↳ NCBI;
- ▶ Alternativa para pesquisas com baixo orçamento:
  - Uma infinidade de dados disponíveis;
  - O experimento nem sempre é o ideal para nossa pesquisa

# NCBI (*National Center for Biotechnology Information*)

**NIH** National Library of Medicine  
National Center for Biotechnology Information

All Databases  Search

## Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News & Blog](#)

<b>Submit</b> Deposit data or manuscripts into NCBI databases	<b>Download</b> Transfer NCBI data to your computer	<b>Learn</b> Find help documents, attend a class or watch a tutorial
		
<b>Develop</b> Use NCBI APIs and code libraries to build applications	<b>Analyze</b> Identify an NCBI tool for your data analysis task	<b>Research</b> Explore NCBI research and collaborative projects
		

### COVID-19 Information

[Public health information \(CDC\)](#) | [Research information \(NIH\)](#) | [SARS-CoV-2 data \(NCBI\)](#) | [Prevention and treatment Information \(HHS\)](#) | [Español](#)

#### Popular Resources

PubMed  
Booksshelf  
PubMed Central  
BLAST  
Nucleotide  
Genome  
SNP  
Gene  
Protein  
PubChem

#### NCBI News & Blog

Introducing NLM's new NCBI Datasets genome page! 29 Jun 2022

As part of the NIH Comparative

Join Us at the ISMB Codeathon- Tools for Sharable Protein Analysis 28 Jun 2022

NLM's NCBI is gearing up for the Tools

New RefSeq annotations are available! 27 Jun 2022

In April and May, the NCBI Eukaryotic Genome Annotation Pipeline released

[More...](#)

# NCBI (National Center for Biotechnology Information)

Display Settings: ▾

Send to: ▾

**Escherichia coli REL606**

Accession: PRJNA380528 ID: 380528

LTEE metagenomic sequencing over 60,000 generations

Time-resolved metagenomic sequencing of Richard Lenski's long-term evolution experiment with Escherichia coli over 60,000 generations.

Accession	PRJNA380528
Data Type	Raw sequence reads
Scope	Multisolate
Organism	Escherichia coli REL606 [Taxonomy ID: 1284382] Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Escherichia; Escherichia coli; Escherichia coli REL606
Publications	Good BH et al., "The dynamics of molecular evolution over 60,000 generations.", Nature, 2017 Nov 2;551(7678):45-50
Submission	Registration date: 26-Mar-2017 Harvard University
Relevance	Evolution

**Project Data:**

Resource Name	Number of Links
SEQUENCE DATA	
SRA Experiments	2443
PUBLICATIONS	
PubMed	1
PMC	1
OTHER DATASETS	
BioSample	1512

See Genome Information for Escherichia coli

NAVIGATE UP

This project is a component of the Long-Term Evolution Experiment with E. coli

NAVIGATE ACROSS

13 additional projects are components of the Long-Term Evolution Experiment with E. coli.

7759 additional projects are related by organism.

Related information

BioProject

BioSample

Full text in PMC

Genome

PubMed

SRA

Taxonomy

Umbrella projects

Recent activity

Turn Off Clear

Escherichia coli REL606 BioProject

PRJNA380528 (1) BioProject

Arapaima gigas (54) SRA

SRX1970305 (1) SRA

Homo sapiens Genome

See more...

# NCBI (*National Center for Biotechnology Information*)

Full ▾

## **SRX3295290: REL11294**

1 ILLUMINA (Illumina HiSeq 2500) run: 975,777 spots, 292.7M bases, 151.8Mb downloads

**Design:** Ara-1 generation 49000 mixed population sample

**Submitted by:** Harvard University

**Study:** LTEE metagenomic sequencing over 60,000 generations

[PRJNA380528](#) • [SRP119922](#) • [All experiments](#) • [All runs](#)

[show Abstract](#)

**Sample:** Ara-1 generation 49000 mixed population sample

[SAMN07763244](#) • [SRS2604012](#) • [All experiments](#) • [All runs](#)

**Organism:** *Escherichia coli* B str. REL606

**Library:**

**Name:** CL3\_1\_TACCTGAC-CGTACCGG.L2

**Instrument:** Illumina HiSeq 2500

**Strategy:** WGS

**Source:** GENOMIC

**Selection:** RANDOM

**Layout:** PAIRED

**Runs:** 1 run, 975,777 spots, 292.7M bases, [151.8Mb](#)

Run	# of Spots	# of Bases	Size	Published
<a href="#">SRR6184990</a> 	975,777	292.7M	151.8Mb	2017-10-18

ID: 4619193

# NCBI (*National Center for Biotechnology Information*)

**National Library of Medicine**  
National Center for Biotechnology Information

**Sequence Read Archive**

Run Browser > SRR6184990

## REL11294 (SRR6184990)

**Metadata** Analysis Reads Data access FASTA/FASTQ download

### Run

Run	Spots	Bases	Size	GC Content	Published	Access Type
SRR6184990	975.8k	292.7M	151.8M	50%	2017-10-18	public

Quality graph (bigger)

This run has 2 reads per spot:

L=150, 100% L=150, 100%

Legend

### Experiment

Experiment	Library Name	Platform	Strategy	Source	Selection	Layout	Action
SRX3295290	CL3_1_TACCTGAC-CGTACCGG.L2	Illumina	WGS	GENOMIC	RANDOM	PAIRED	BLAST

Design:  
Ara-1 generation 49000 mixed population sample

**National Library of Medicine**  
National Center for Biotechnology Information

Sequence Read Archive    Search    Run Browser    Analyses    Study    Provisional SRA    Documentation    Mirroring

Run Browser > SRR6184990

## REL11294 (SRR6184990)

Metadata    Analysis    Reads    Data access    **FASTA/FASTQ download**

### Download for Experiment SRX3295290

Accession	Total Bases	Spots	
		Total	Filtered
SRR6184990	292.7M	975.8k	

**Filter Runs**

Search by sub-sequence, spo   

What can the filter be applied to?

**Download**

Filtered     Clipped    FASTA    OR    FASTQ

Existem ferramentas para download em lote

# NCBI (National Center for Biotechnology Information)



National Library of Medicine  
National Center for Biotechnology Information

Log in

SRA

SRA

SRX1970305

Create alert Advanced

Search

Help

Full ▾

Send to: ▾

**SRX1970305: GSM2248122: Control\_D0; Homo sapiens; RNA-Seq**

1 ILLUMINA (Illumina HiSeq 2500) run: 12.4M spots, 631.8M bases, 370.2Mb downloads

**Submitted by:** NCBI (GEO)

**Study:** RNA-seq of Human neural progenitor cells exposed to lead (Pb) reveals transcriptome dynamics, splicing alterations and Pb disease risk associations

[PRJNA330909](#) • [SRP079342](#) • All experiments • All runs

show Abstract

[Detailed description](#)

**Sample:** Control\_D0

[SAMN05430048](#) • [SRS1579362](#) • All experiments • All runs

**Organism:** *Homo sapiens*

**Library:**

**Instrument:** Illumina HiSeq 2500

**Strategy:** RNA-Seq

**Source:** TRANSCRIPTOMIC

**Selection:** cDNA

**Layout:** SINGLE

**Construction protocol:** Samples were collected at indicated time points by lysing cells directly on the plate with 350µl RLT buffer. Cells were lysed in 350µl of the QIAGEN Buffer RLT (79216). Total RNA was isolated using the QIAGEN RNeasy 96 Kit (74181) and quantified with the ThermoFisher Scientific Quant-IT RNA Assay Kit (Q33140) and the Agilent Bioanalyzer RNA 6000 Pico Kit (5067-1513). cDNA libraries were prepared with 100 ng of input total RNA and indexed with the Illumina TruSeq RNA Prep Kit v2 (RS-122-2001 and RS-122-2002). Indexed cDNA libraries were pooled and sequenced on an Illumina HiSeq2500 with single end 51 bp reads.

**Experiment attributes:**

**GEO Accession:** GSM2248122

**Links:**

**Runs:** 1 run, 12.4M spots, 631.8M bases, [370.2Mb](#)

Run	# of Spots	# of Bases	Size	Published
<a href="#">SRR3944315</a>	12,387,706	631.8M	370.2Mb	2018-01-02

ID: 2834712

## Related information

BioProject

BioSample

GEO DataSets

PMC

PubMed

Taxonomy

## Search details

SRX1970305 [All Fields]

Search

See more...

## Recent activity

[Turn Off](#) [Clear](#)

[SRX1970305 \(1\)](#)

SRA

[Escherichia coli REL606](#)

BioProject

[PRJNA380528 \(1\)](#)

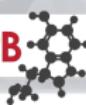
BioProject

[Arapaima gigas \(54\)](#)

SRA

[Homo sapiens](#)

Genome

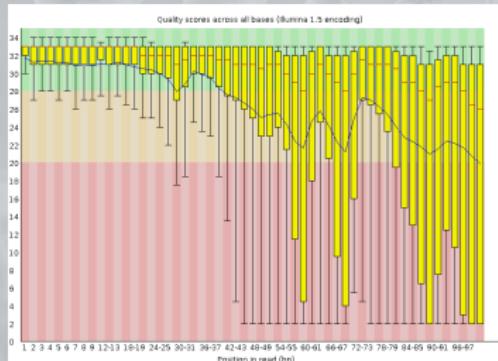


# Início dos Pipelines de Bioinformática

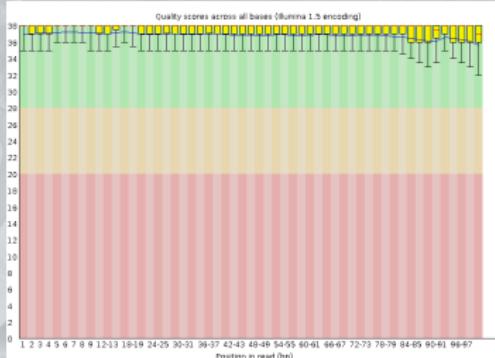
- ▶ Remoção de adaptadores (Trimmomatic, AdapterRemoval, Cutadapt...)
- ▶ Demultiplexação, se for o caso
- ▶ Remoção de contaminantes (Genomas indesejados, Erros de protocolo...)
- ▶ Testes de qualidade
  - Verificar a qualidade do sequenciamento
  - FastQC e MultiQC

# Testes de qualidade

- ▶ Verificar a qualidade do sequenciamento
- ▶ FastQC e MultiQC



FastQC Ruim



FastQC Bom

FastQC: Total Sequences (1000000)

Sample Name	% Dupes	% QC	Length	% Failed	M Sequences
RNA-Seq	4.9%	47%	100 bp	18%	0.1
RRBS	15.1%	39%	100 bp	18%	0.1
bad	28.4%	47%	100 bp	18%	0.5
good	0.0%	45%	100 bp	0%	0.2

Sumário do MultiQC



Qualidade MultiQC

# Montagem ou Alinhamento dos Dados

- ▶ Feito com softwares próprios
  - Blast e seus derivados, Hisat2, STAR...
  - **Medusa**, Megan6, GO FEAT, eggNOG-mapper... (Metagenomas)
- ▶ Criam índices de referência
- ▶ Consomem muito tempo e poder de processamento
- ▶ Múltiplos genomas no caso dos **metagenomas**



# Processamento dos dados alinhados (entre outras possibilidades)

## ► Genoma – chamada de variantes

- Lista as diferenças entre a referência e a amostra
- Mutações prejudiciais

CustomCell Annotation	Consent Info						Strand Bias						impCT					
	Chrom	Position	Reference	Variant	Sample Coverage		Lisher's	Odds Ratio	Gene	Starts	Region	Region Descrip	Locs	Var type	Impact	Alteration	changePer	change/locN
1 ★	chr2	211679193	C	T	0%	0.000	0.000	0.000	ENSG000005235.2	intron	INTRON	12	SNV	DEleIntr		C.1450=GG>A		
1 ★	chr2	224181971	C	T	0%	0.000	0.000	0.000	ENSG000005235.2	intron	INTRON	14	SNV	DEleIntr	DEleIntr	p.His197Asn	p.His197Asn	
1 ★	chr2	224567027	T	G	0%	0.000	0.000	0.000	ENSG00000521703.2	intron	INTRON	1	SNV	DEleIntr		C.041>GGGAA-C		
1 ★	chr9	334567009	T	C	0%	0.000	0.000	0.000	ENSG00000521714.1	intron	INTRON	1	SNV	DEleIntr		C.041>GGGAA-C		
1 ★	chr2	102624285	C	A	0%	0.000	0.000	0.000	ENSG00000520844.4	intron	INTRON	19	SNV	DEleIntr		C.1767>GGGAA		
1 ★	chr2	100621361	A	T	0%	0.000	0.000	0.000	ENSG00000520844.4	intron	INTRON	19	SNV	DEleIntr		C.1767>GGGAA		
1 ★	chr2	102641881	C	A	0%	0.000	0.000	0.000	ENSG00000520844.4	intron	NON_SYNONYM_CDS	27	SNV	DEleIntr	DEleIntr	p.Arg184Ter	p.Arg184Ter	
1 ★	chr3	100804113	C	T	0%	0.000	0.000	0.000	ENSG00000520844.4	intron	INTRON	34	SNV	DEleIntr		C.3457>GGC>G		
1 ★	chr2	100804113	C	T	0%	0.000	0.000	0.000	ENSG00000520825	intron	INTRON	34	SNV	DEleIntr		N.100804113del		
1 ★	chr2	12603544	C	T	0%	0.000	0.000	0.000	ENSG00000520830.2	intron	INTRON	7	SNV	DEleIntr		C.0341>GGGAC-T		
1 ★	chr1	370176313	G	T	0%	0.000	0.000	0.000	ENSG00000519313	intron	INTRON	10	SNV	DEleIntr		C.0864>GGGAC-T		
1 ★	chr2	470624298	C	T	0%	0.000	0.000	0.000	ENSG00000519313	intron	NON_SYNONYM_CDS	14	SNV	HIGH	DEleIntr	p.Arg204Ter	C.0864>GGGAC-T	
1 ★	chr2	4706322	G	T	0%	0.000	0.000	0.000	ENSG00000519313	intron	NON_SYNONYM_CDS	14	SNV	HIGH	DEleIntr	C.0864>GGGAC-T		
1 ★	chr2	4706240	T	G	0%	0.000	0.000	0.000	ENSG00000519313	intron	NON_SYNONYM_CDS	10	SNV	HIGH	DEleIntr	C.0864>GGGAC-T		
1 ★	chr2	4706233	A	T	0%	0.000	0.000	0.000	ENSG00000519313	intron	NON_SYNONYM_CDS	9	SNV	HIGH	DEleIntr	C.0864>GGGAC-T		
1 ★	chr2	37064186	G	T	0%	0.000	0.000	0.000	ENSG00000519313	intron	NON_SYNONYM_CDS	24	SNV	DEleIntr	DEleIntr	p.Trp103Ter	p.Trp103Ter	
1 ★	chr3	52506470	G	T	0%	0.000	0.000	0.000	ENSG00000519313	intron	NON_SYNONYM_CDS	21	SNV	DEleIntr	DEleIntr	p.Arg109Ter	C.52506470A	
1 ★	chr1	53788051	A	T	0%	0.000	0.000	0.000	ENSG00000519313	intron	NON_SYNONYM_CDS	19	SNV	DEleIntr	DEleIntr	p.Thr76Ter	C.53788051A	
1 ★	chr2	52859252	C	T	0%	0.000	0.000	0.000	ENSG00000519313	intron	NON_SYNONYM_CDS	19	SNV	DEleIntr	DEleIntr	p.Arg260Ter	C.52859252G	
1 ★	chr3	52644652	A	T	0%	0.000	0.000	0.000	ENSG00000519313	intron	NON_SYNONYM_CDS	9	SNV	DEleIntr	DEleIntr	C.0999>GGT-C		
1 ★	chr1	57802778	T	A	0%	0.000	0.000	0.000	ENSG00000519313	intron	NON_SYNONYM_CDS	4	SNV	HIGH	DEleIntr	C.0999>GGT-C		
1 ★	chr3	70072700	A	T	0%	0.000	0.000	0.000	ENSG00000519313	intron	NON_SYNONYM_CDS	17	SNV	DEleIntr	DEleIntr	C.1531>GGT-C		
1 ★	chr1	70172700	T	A	0%	0.000	0.000	0.000	ENSG00000519313	intron	NON_SYNONYM_CDS	17	SNV	DEleIntr	DEleIntr	C.1531>GGT-C		
1 ★	chr2	105670323	C	A	0%	0.000	0.000	0.000	ENSG00000519313	intron	NON_SYNONYM_CDS	18	SNV	DEleIntr	DEleIntr	p.Val359Ter	C.105670323C	
1 ★	chr2	142457783	C	T	0%	0.000	0.000	0.000	ENSG00000519313	intron	NON_SYNONYM_CDS	44	SNV	DEleIntr	DEleIntr	C.142457783T		
1 ★	chr2	142448012	A	T	0%	0.000	0.000	0.000	ENSG00000519313	intron	NON_SYNONYM_CDS	29	SNV	DEleIntr	DEleIntr	C.142448012A	C.142448012A	
1 ★	chr3	142196550	G	A	0%	0.000	0.000	0.000	ENSG00000519313	intron	NON_SYNONYM_CDS	33	SNV	DEleIntr	DEleIntr	C.0739>GGT-C		
1 ★	chr2	142491145	C	T	0%	0.000	0.000	0.000	ENSG00000519313	intron	NON_SYNONYM_CDS	51	SNV	DEleIntr	DEleIntr	C.142491145A		
1 ★	chr3	142196116	G	A	0%	0.000	0.000	0.000	ENSG00000519313	intron	NON_SYNONYM_CDS	31	SNV	DEleIntr	DEleIntr	C.0531>GGT-C		
1 ★	chr2	142511906	A	T	0%	0.000	0.000	0.000	ENSG00000519313	intron	NON_SYNONYM_CDS	23	SNV	DEleIntr	DEleIntr	C.142511906A		

# Processamento dos dados alinhados (entre outras possibilidades)

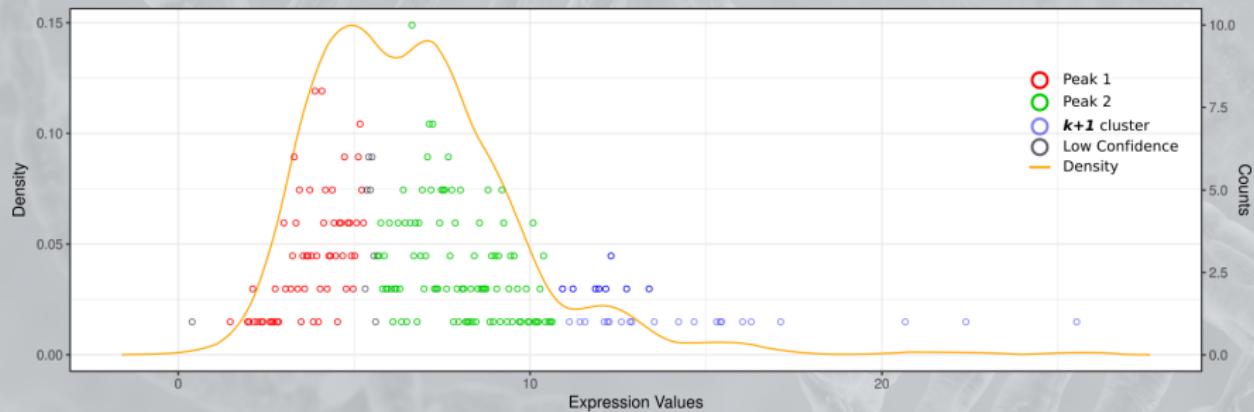
## ► Genoma – chamada de variantes

- Lista as diferenças entre a referência e a amostra
- Mutações prejudiciais

snpEFF						ExAC Freque			
Var Type	Impact	Alteration	changeProt	changecDNA	codon	dbSNP	Total	African	American
SNV	MODERATE	MISSENSE	p.Ser192Tyr	c.575C>A	tCt/tAt	rs1042602	0.2518	0.0620	0.1918
SNV	MODERATE	MISSENSE	p.Leu55Arg	c.164T>G	cTt/cGt	rs17127947	0.1772	0.2530	0.0931
SNV	MODERATE	MISSENSE	p.Tyr131His	c.391T>C	Tac/Cac	rs4057750	0.2324	0.3465	0.1323
SNV	MODERATE	MISSENSE	p.Cys259Arg	c.775T>C	Tgt/Cgt	rs17277228	0.4324	0.5115	0.2691
SNV	MODERATE	MISSENSE	p.Ser139Leu	c.416C>T	tCg/tTg		0.0005	0.0003	0.0009
SNV	MODERATE	MISSENSE	p.Gly16Ser	c.46G>A	Ggc/Agc	rs1218762	0.2627	0.3520	0.2036

# Processamento dos dados alinhados (entre outras possibilidades)

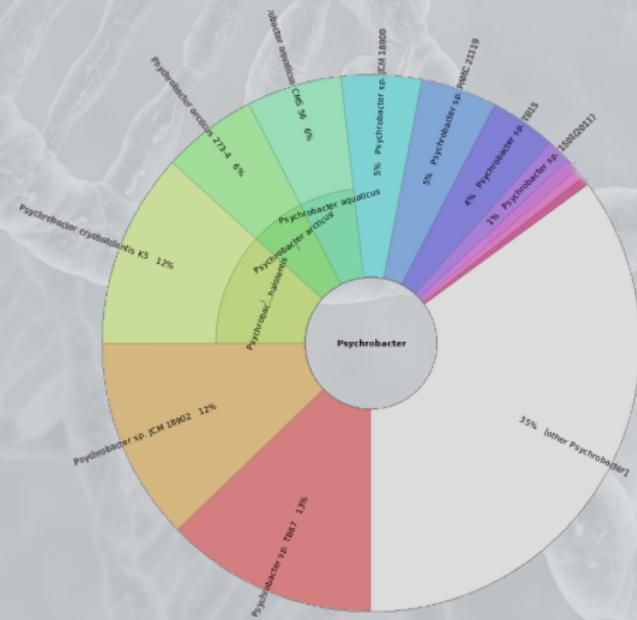
- Transcriptoma – contagem de transcritos
  - Mapa do que está acontecendo dentro das células



# Processamento dos dados alinhados (entre outras possibilidades)

## ► Metagenoma

- Lista das espécies contidas na amostra
- Taxonomia (Kaiju) e funcionalidade



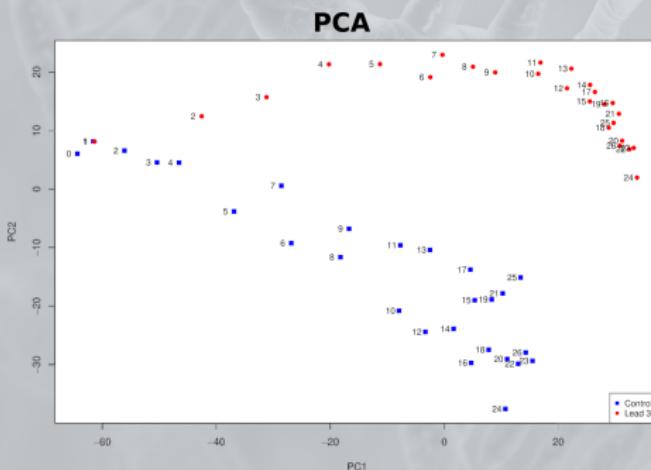
# A imaginação é o limite

- ▶ Elaborar uma hipótese
- ▶ Tratar os dados
- ▶ Testar a hipótese
  - Ferramentas biológicas (BioConductor e R)



# A imaginação é o limite

- Elaborar uma hipótese
- Tratar os dados
- Testar a hipótese
  - Ferramentas biológicas (BioConductor e R)
  - Ferramentas estatísticas



**Teste Hipergeométrico**

Faixa	white	black	drawn	p-val	FDR
1	41	288	761	200	0.008066 0.03226
2	30	196	853		0.080954 0.24286
3	38	335	714		0.000005 <b>0.00002</b>
4	51	150	899		0.999999 1.00000
5	40	80	969		1.000000 1.00000
Total	200	1049			

# A imaginação é o limite

- ▶ Elaborar uma hipótese
- ▶ Tratar os dados
- ▶ Testar a hipótese
  - Ferramentas biológicas (BioConductor e R)
  - Ferramentas estatísticas
- ▶ Fundamentação biológica
- ▶ Conclusões

# Conclusão

- ▶ A Bioinformática é uma ciência nova
- ▶ As possibilidades são ilimitadas
- ▶ Há uma enorme quantidade de dados a serem explorados



Perguntas?