

**Biologia Molecular Computacional**  
**IBI5035/QBQ2507 - 2023**

# **Regulação da expressão gênica na era ômica**

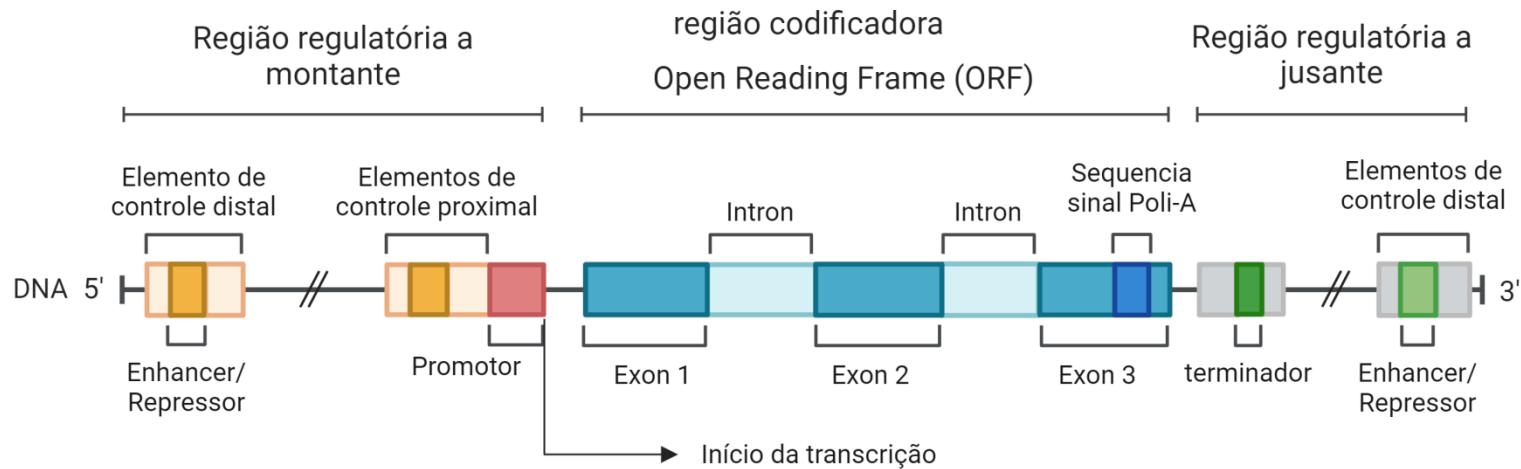
Eduardo Moraes Rego Reis  
Instituto de Química - USP

<b>26 out</b> - Regulação da expressão gênica na era ômica	Eduardo Reis
<b>2 nov</b> - feriado; não haverá aula	
<b>9 nov</b> - análise de transcritomas- RNAseq (tutorial)	Eduardo Reis
<b>16 nov</b> - análise de células únicas (tutorial)	Eduardo Reis
<b>23 nov</b> - bancos de dados genômicos (tutorial)	Eduardo Reis
<b>30 nov</b> - análise de enriquecimento de categorias gênicas (tutorial)	Eduardo Reis
<b>7 dez</b> – estrutura de RNAs (tutorial)	Eduardo Reis
<b>14 dez</b> - microRNAs e redes regulatórias da expressão gênica (tutorial)	Eduardo Reis
<b>21 dez</b> - análise global de elementos regulatórios da expressão gênica (tutorial)	Eduardo Reis
<b>22 dez</b> – prazo de entrega dos exercícios do prof. Eduardo – 23h, hora de SP	

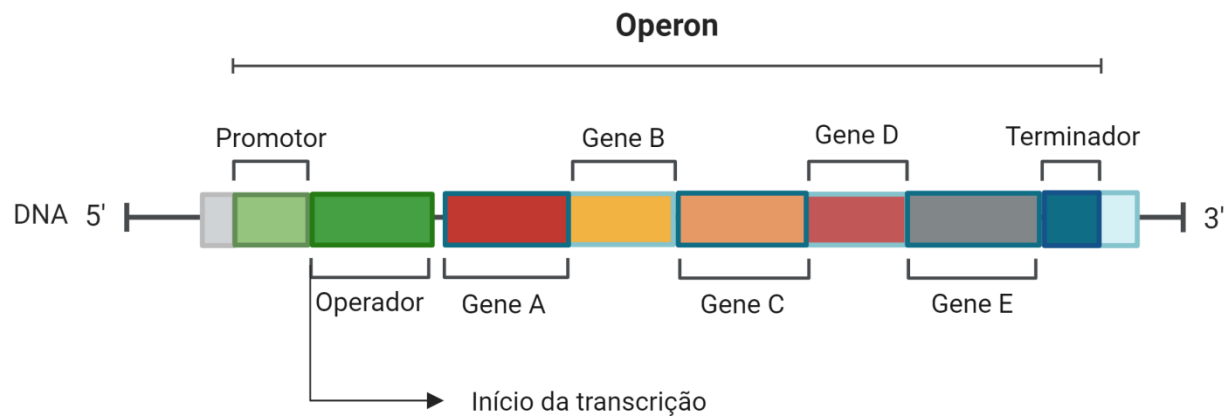
- Relembrando conceitos básicos sobre expressão gênica
- Tecnologias NGS e aplicações no estudo global da regulação da expressão gênica em células e tecidos.
- Análise de transcritomas, Geração, análise de dados e aplicações
- Tutorial de análise de dados de RNAseq utilizando a ferramenta Galaxy (próxima aula)

# Qual é a arquitetura e os componentes de um gene?

## Estrutura de gene eucariótico

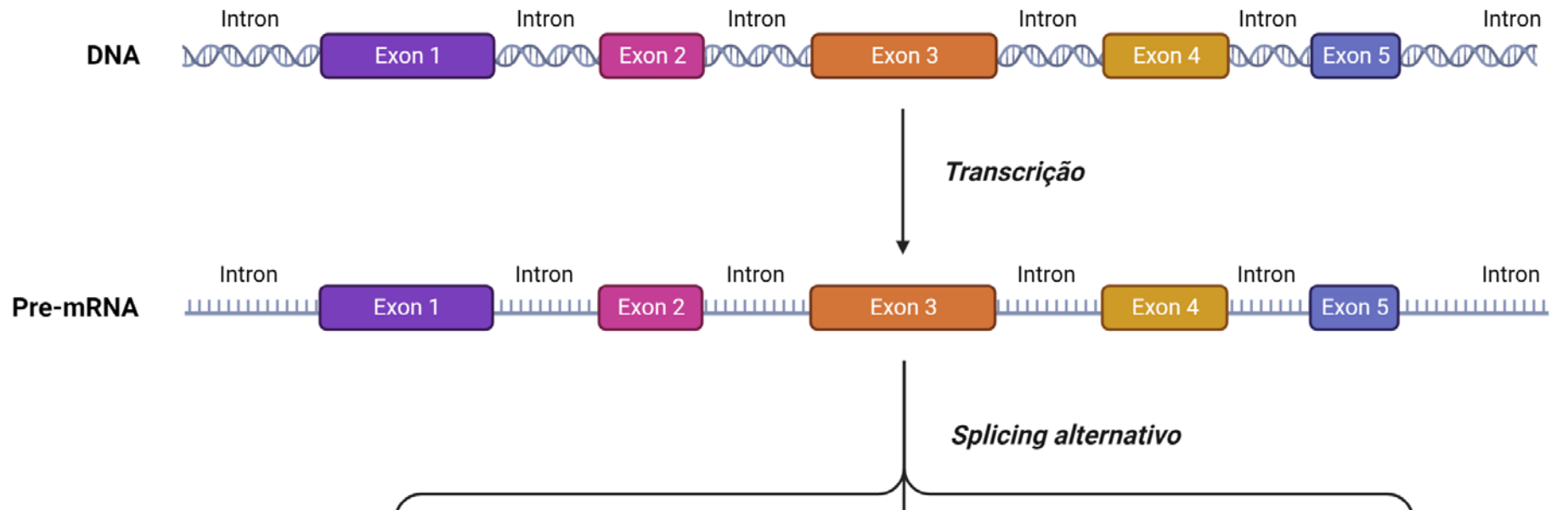


## Estrutura de gene procariótico

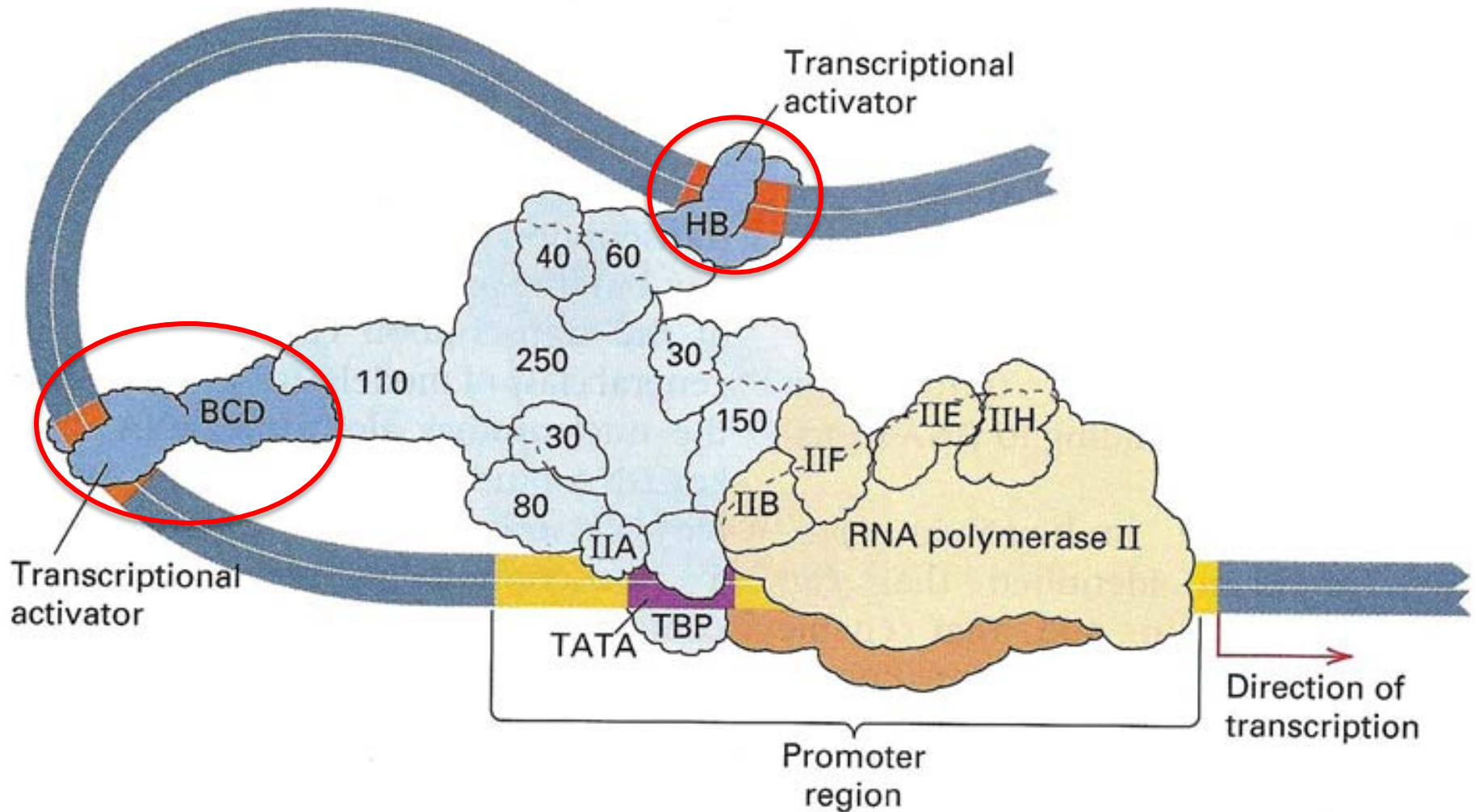


# O que se entende por genes eucarióticos serem interrompidos?

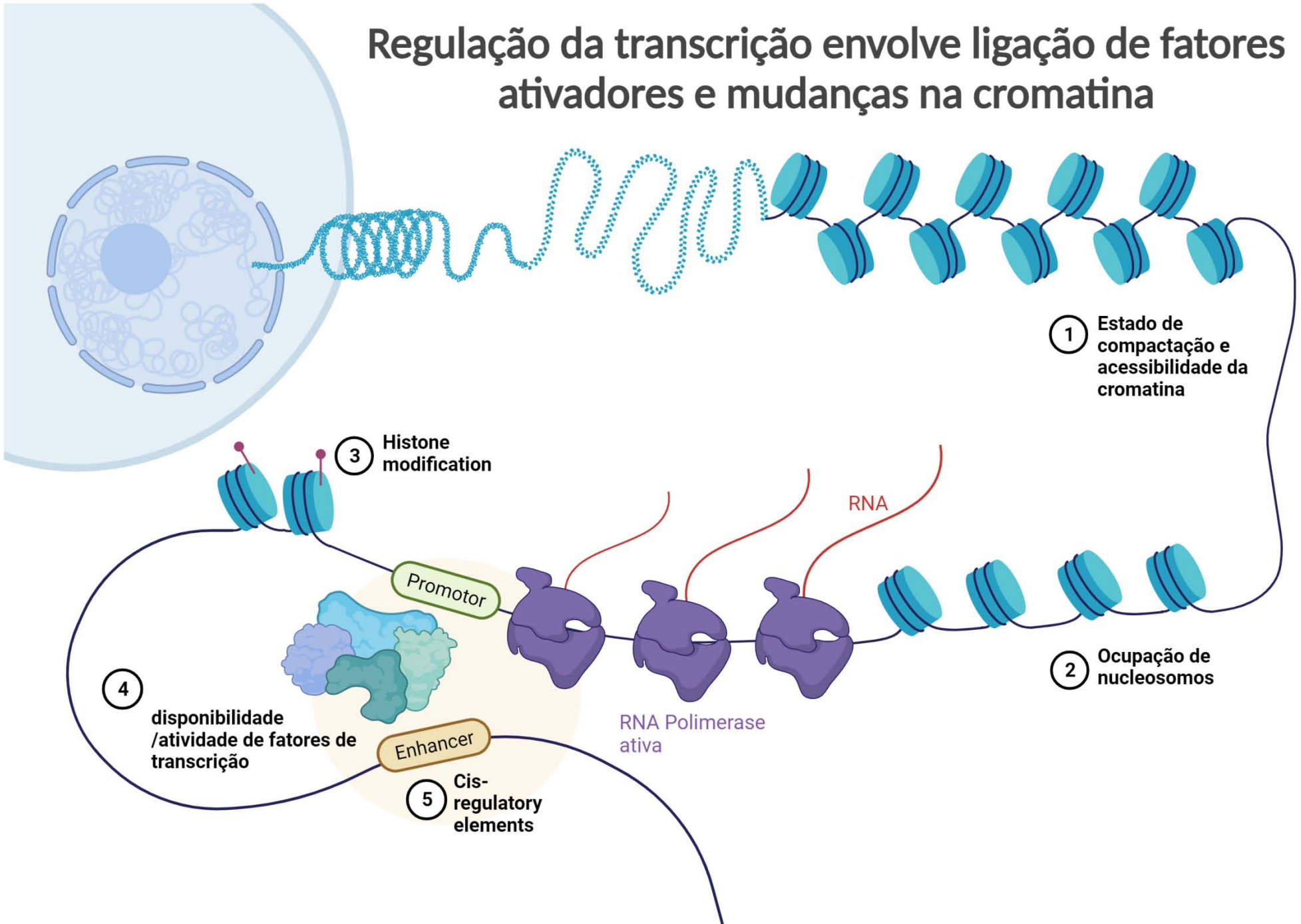
- RNAs precursores (pre-mRNAs) contêm trechos codificadores e UTRs (“exons”) intercalados por trechos não-codificadores (“introns”).
- Introns são removidos após a transcrição (“splicing”) durante o processamento do RNA
- Implica em que o alinhamento de sequências de RNA e cDNA no genoma deve considerar interrupções (“gaps”) devido a ausência de introns nessas moléculas.
- Deve também ser considerado a existência de variantes de splicing alternativo

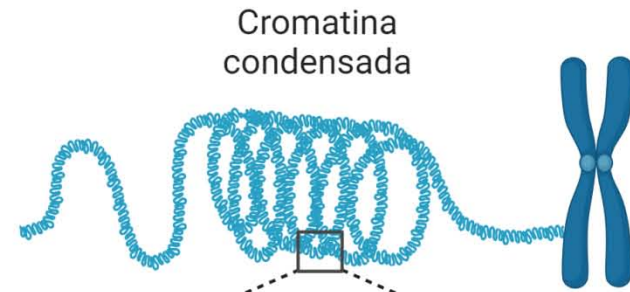
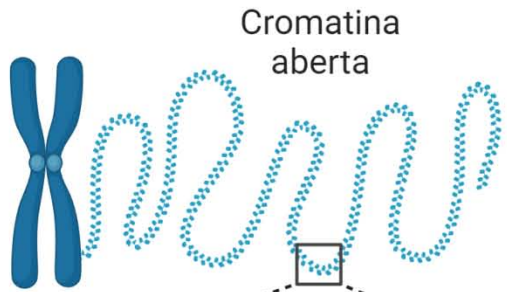


# Ligação de **fatores de transcrição (FT)** a elementos regulatórios no DNA recrutam a RNA Polimerase e ativam a transcrição

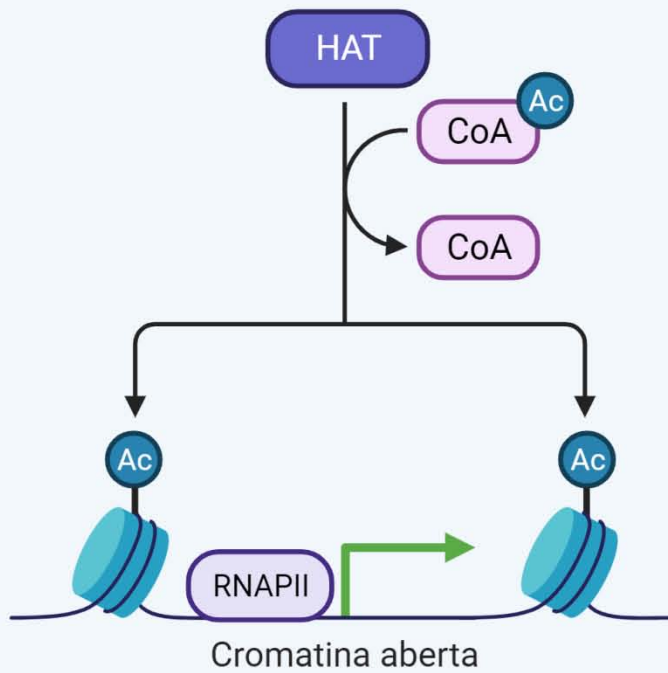


# Regulação da transcrição envolve ligação de fatores ativadores e mudanças na cromatina



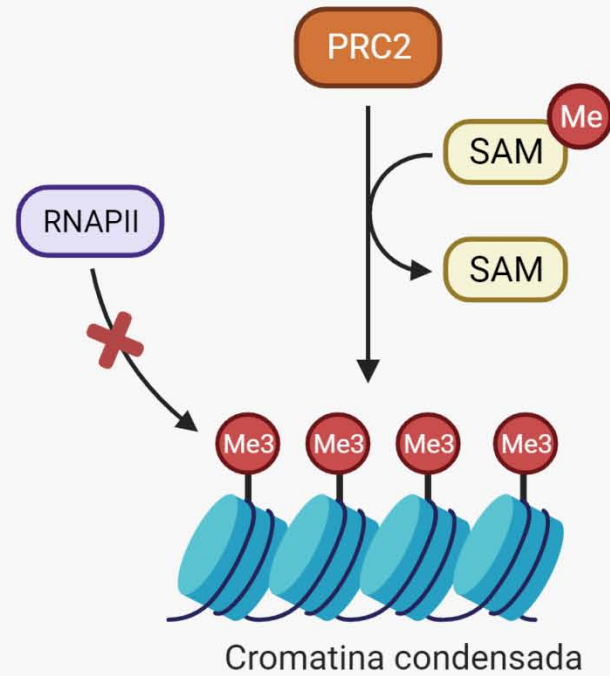


### Acetilação de histonas



**Transcrição ON**

### Metilação de histonas

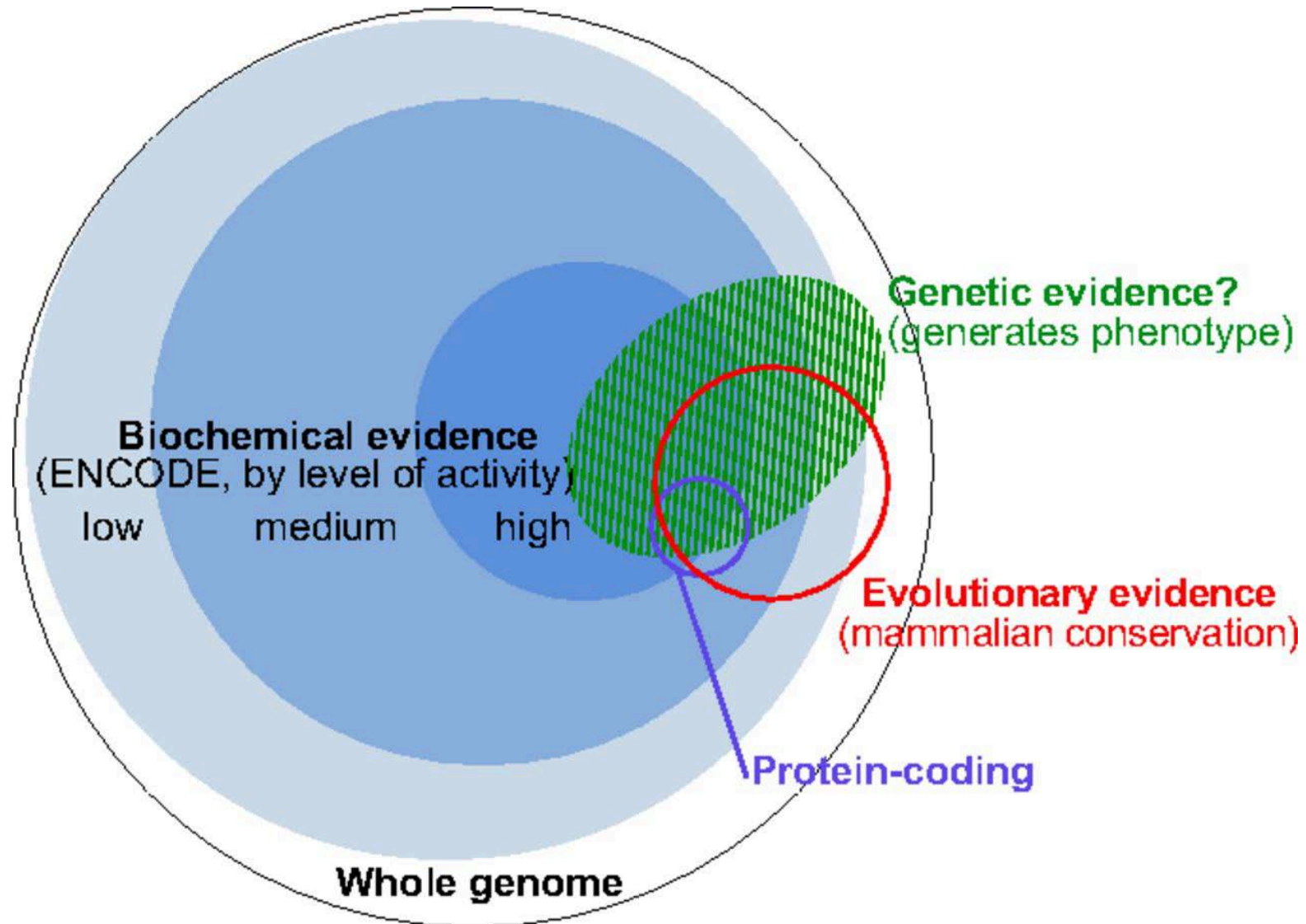


**Transcrição OFF**



# A maior parte do genomas eucariotos é transcrita em RNAs que não codificam proteínas

Projeto ENCODE (Encyclopedia of DNA Elements, 2003 - 2012)

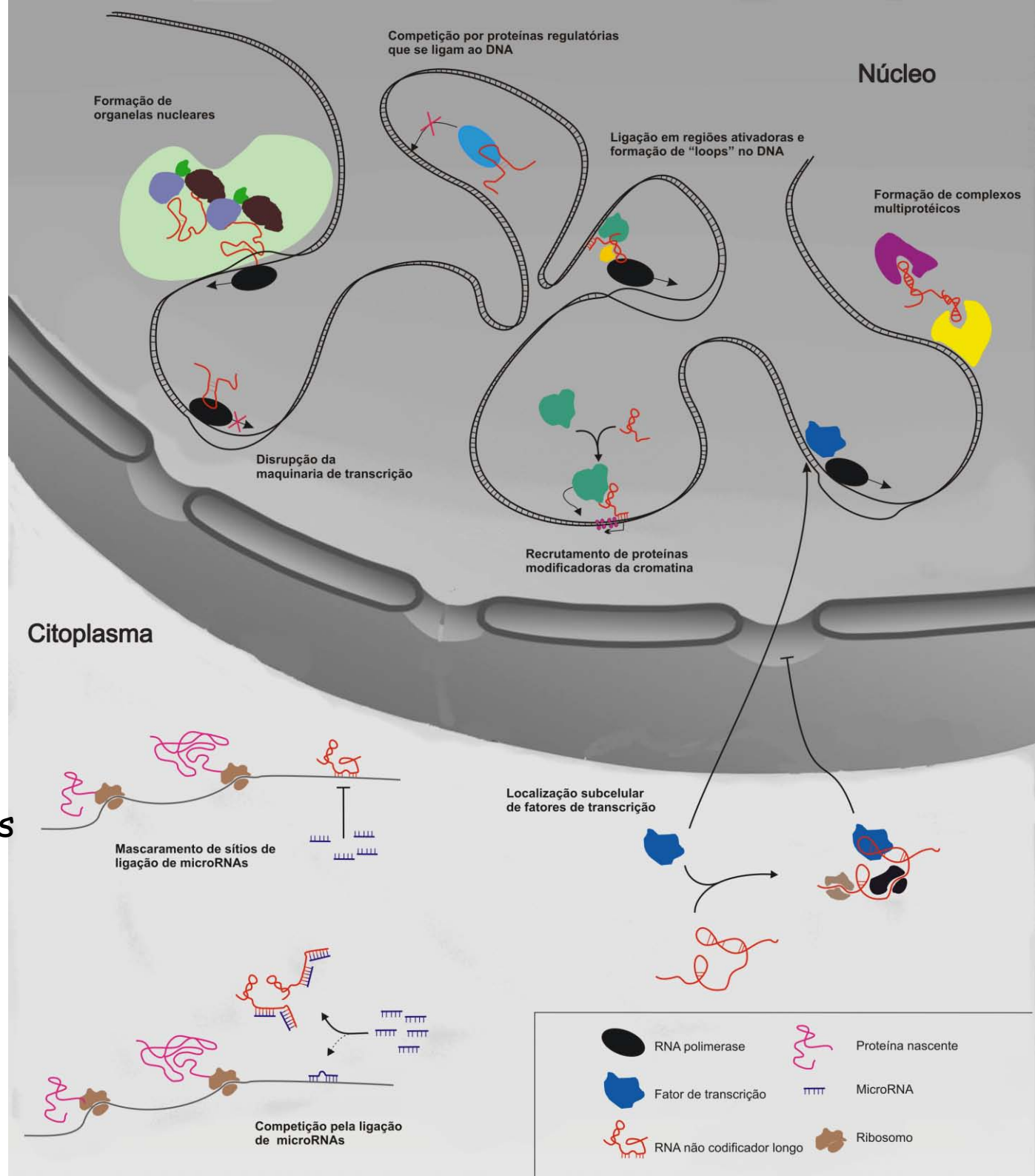


# RNAs não codificadores desempenham papéis centrais na regulação pós-transcricional da expressão gênica

**ncRNAs curtos**  
 < 50-200 nt.  
 ex. microRNAs, piRNAs

**ncRNAs longos (lncRNAs)**  
 > 200 nt, até milhares de bases  
 ex. lincRNAs, antisense RNAs, lncRNAs intrônicos

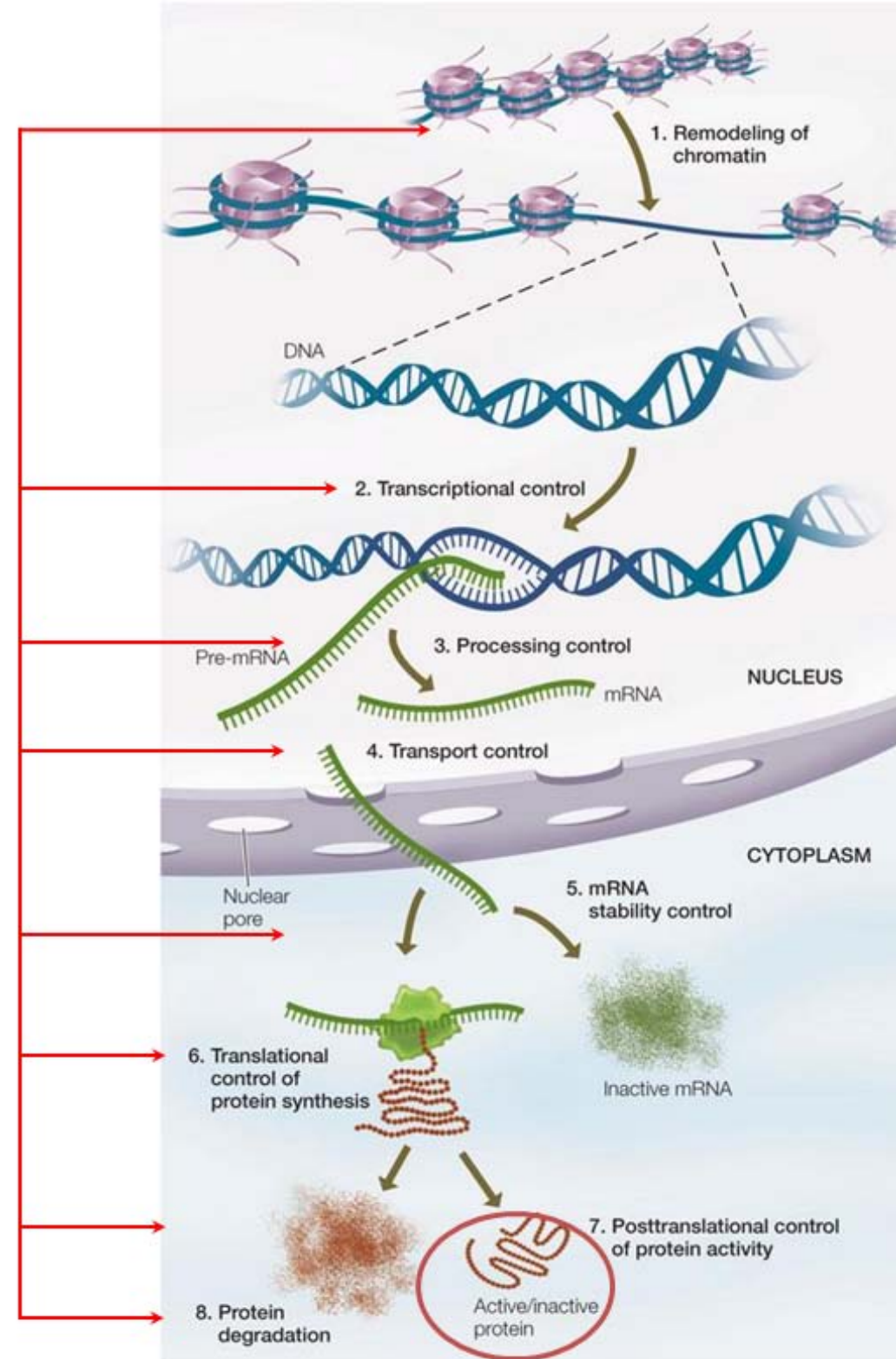
Ayupé e Reis, 2015. Biotecnologia Aplicada a Saúde,  
 Cap. 5 “RNAs não codificadores longos: Genômica, Biogênese, Mecanismos e Função”



# Regulação da expressão gênica é dinâmica e possui múltiplos níveis de controle

- Epigenética
- Transcricional
- Pós-transcricional
- Traducional
- Pós-traducional

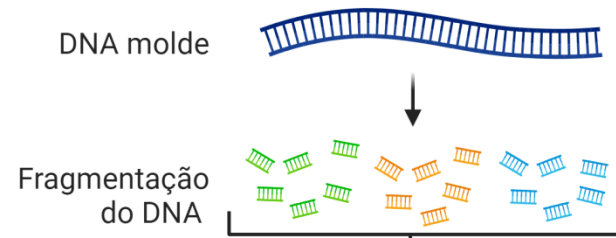
regulação



Como estudar alterações na expressão gênica em escala global?

O surgimento de tecnologias de sequenciamento de DNA de alta-capacidade foi um desenvolvimento essencial

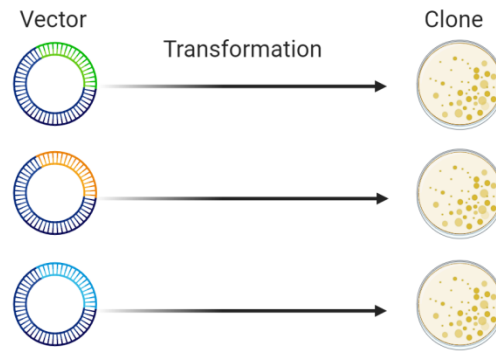
# Sequenciamento de Sanger (dideoxi) vs NGS



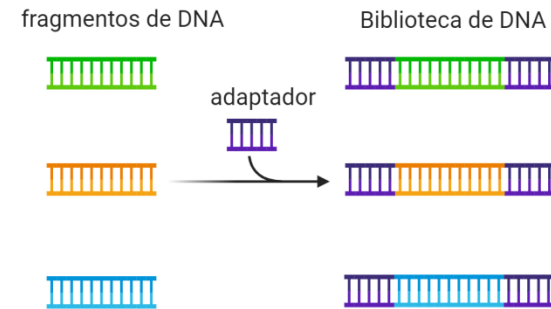
Método Sanger

Métodos NGS

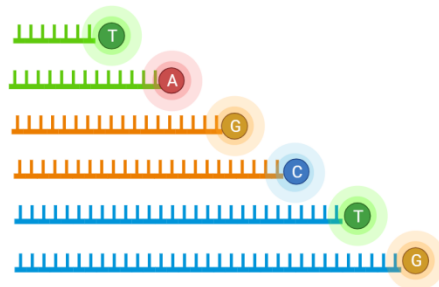
① Clonagem de fragmentos e transformação de bactérias



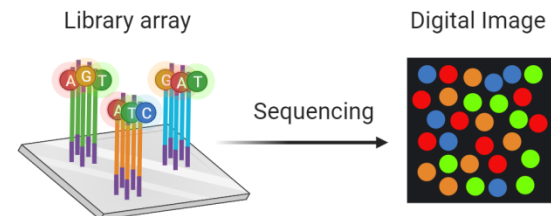
① Ligação de adaptadores nos fragmentos



② sequenciamento Sanger automatizado



② Next Generation Sequencing (NGS)

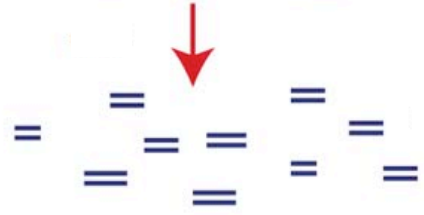


# Illumina - sequenciamento por polinização

DNA genômico ou cDNA

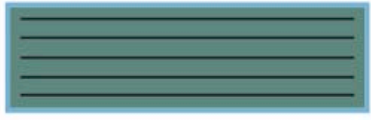
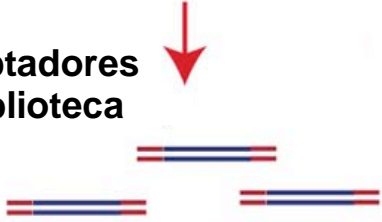


Fragmentação

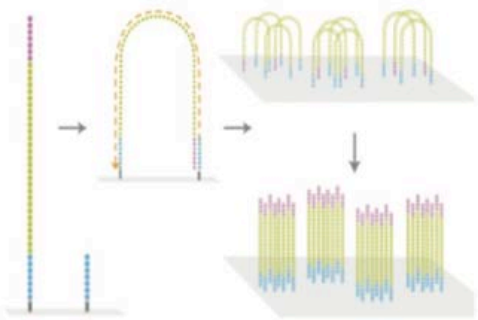


Seleção de fragmentos com 200 a 300 bases

Ligação de adaptadores e geração de biblioteca



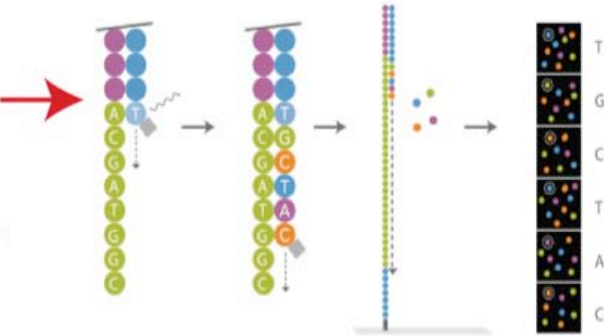
Geração de "clusters" por PCR em fase sólida



Amplificação Clonal ("bridge amplification")

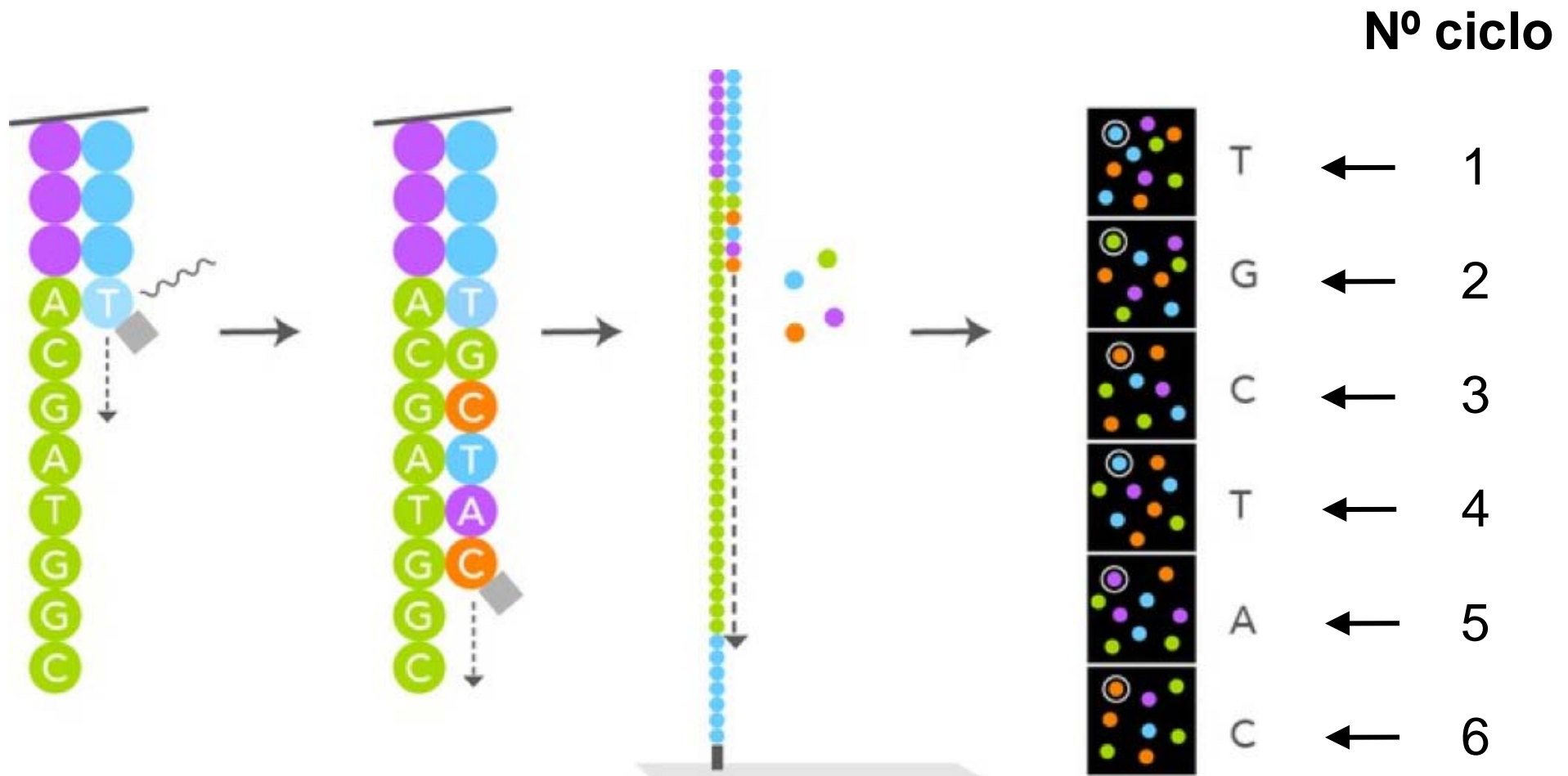


sequencing by synthesis with reversible terminators



Sequenciamento por síntese

# Análise estatística das fluorescências emitidas em cada ciclo permite determinar a sequência do DNA presente no "cluster"



Bilhões de "clusters" em cada corrida

# Estado da arte na acurácia e capacidade de sequenciamento NGS

Alta-capacidade e acurácia, custo decrescente

illumina®



1,2 Gb	7,5 Gb	15 Gb	120 Gb	1.5 Tb	1.8 Tb	<b>6 Tb</b>
4 Mi	25 Mi	25 Mi	400Mi	5 Bi	20 Bi	<b>20 Bi</b>
-	-		~ 20 X	~250 X		~1500 X

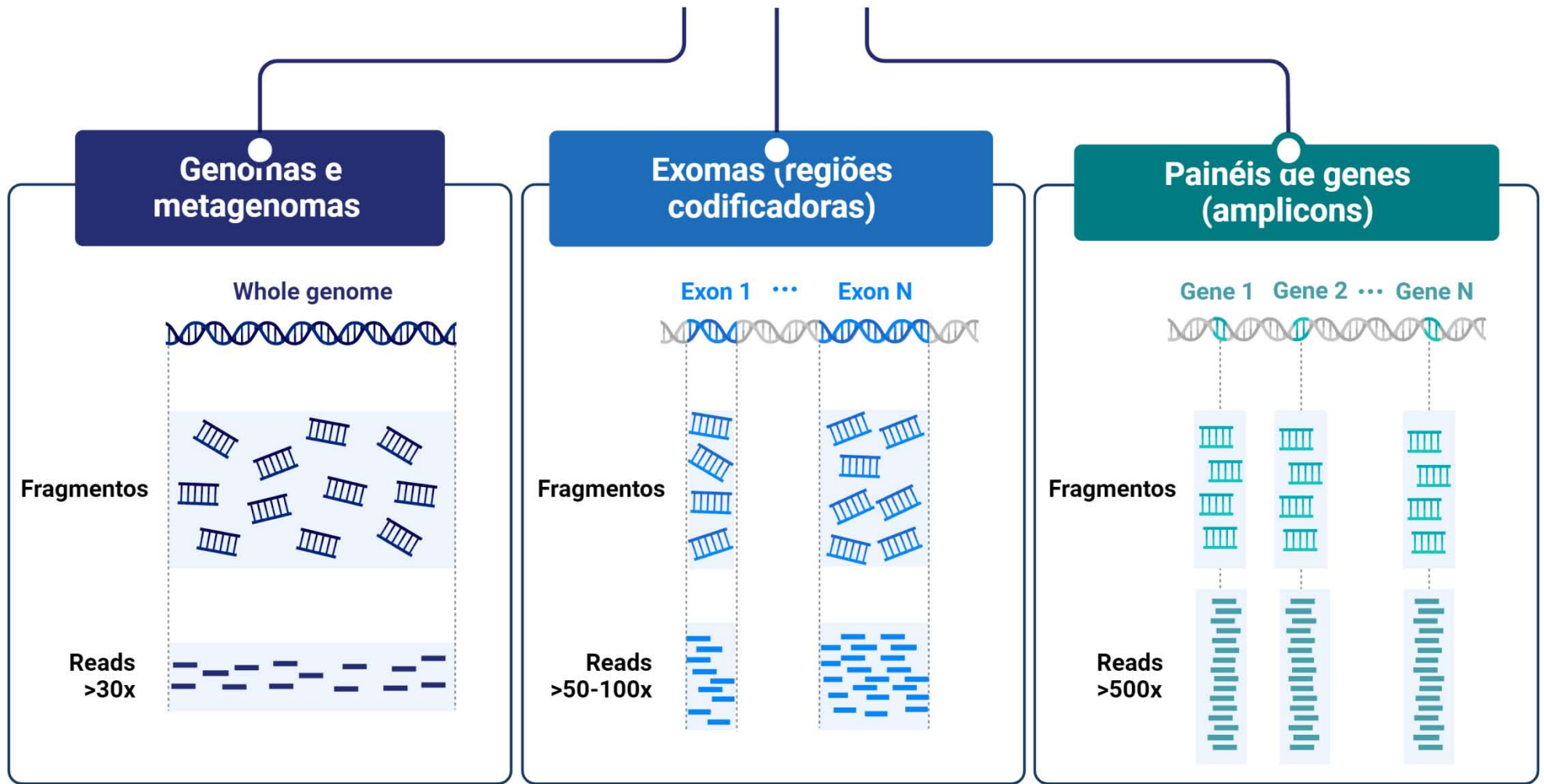
Número de bases de DNA sequenciadas em uma corrida:

Número de reads geradas em uma corrida:

Cobertura do genoma humano (diploide, 6 Gb)



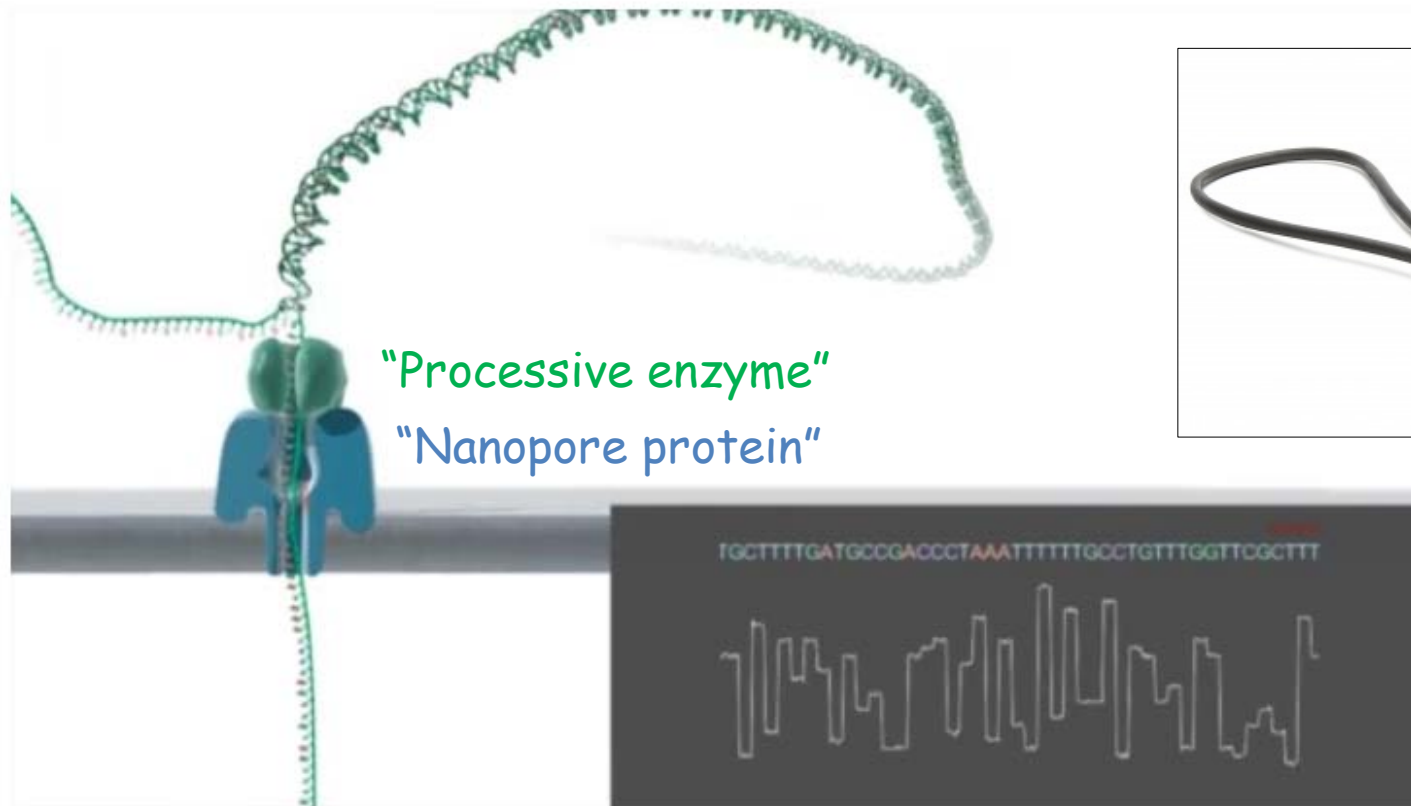
# Next Generation Sequencing (NGS) amplamente aplicado para análise de regiões gênicas e genomas completos



# Limitações do NGS baseado em bibliotecas de fragmentos de DNA/cDNA

- Tecnologias NGS geram **sequencias curtas** (max 300 nt)
- **Ambiguidades** no mapeamento/reconstrução devido a sequencias de baixa complexidade e sequencias repetidas no genoma
- **Informação descontinua** (fragmentação do RNA) dificulta identificar variantes de genes expressos
- Não detectam alterações na sequencia de bases do DNA/RNA (marcas epigenéticas)
- Necessidade de tecnologias que permitam a leitura de cromossomos, genes e RNAs completos

# Nanopore sequencing™ – Oxford Nanopore



Nucleotídeos da cadeia de DNA/RNA são detectados a medida que a fita simples passa pelo nanoporo devido a mudança na condutância dentro do canal

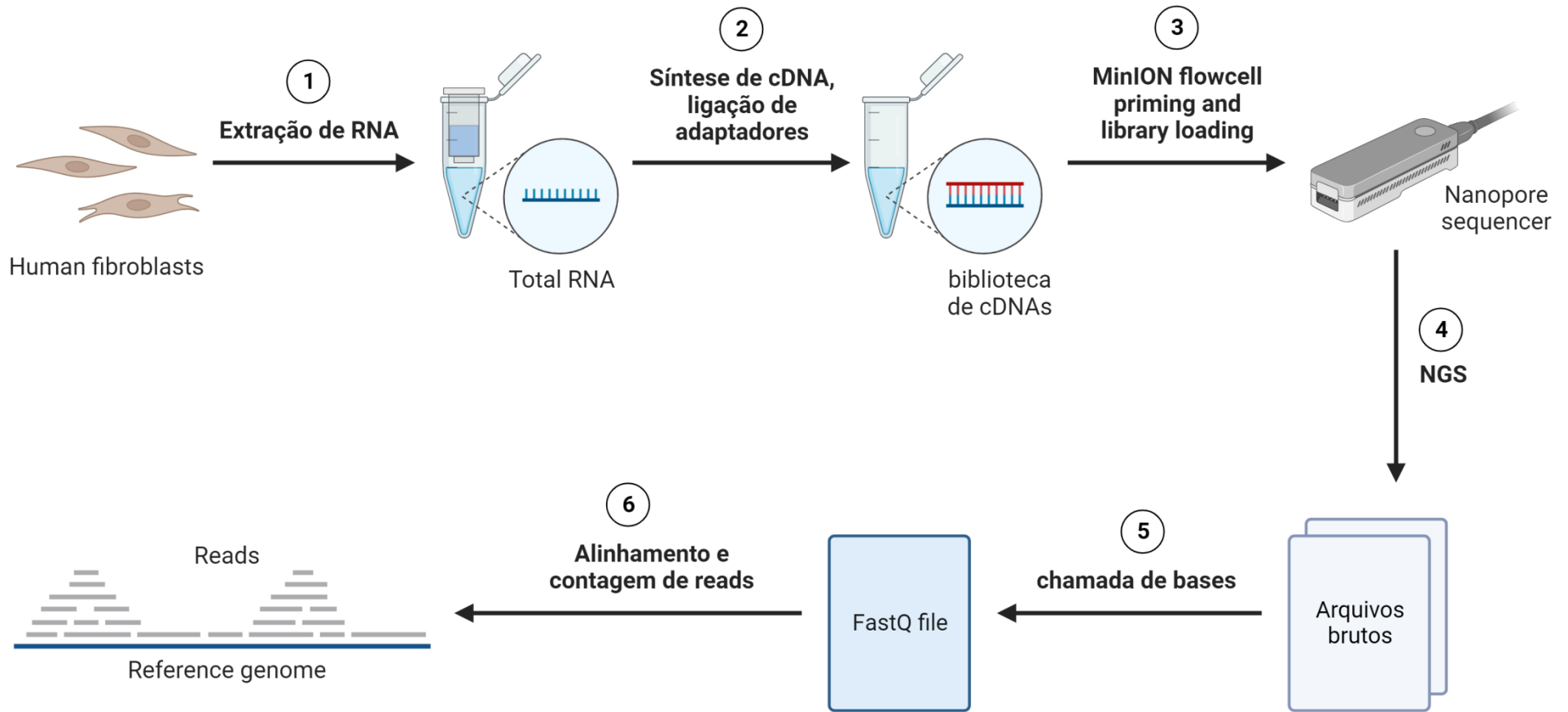
## Vantagens:

- Não requer um equipamento de alto custo
- Sequenciamento de moléculas únicas com até 300 Kb
- Permite detectar modificações químicas nas bases: metilação de DNA, edição de RNAs

Desvantagem: menor capacidade

# NGS na plataforma Oxford Nanopore

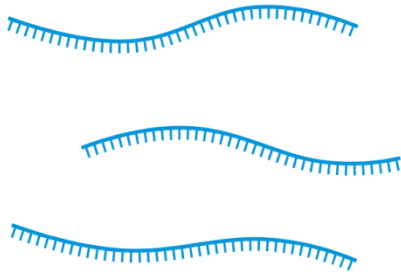
Não requer equipamento caro.



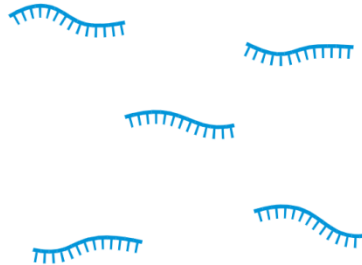
# Dados ômicos gerados a partir de NGS

## RNA-seq: NGS do conjunto de moléculas de RNA de uma amostra

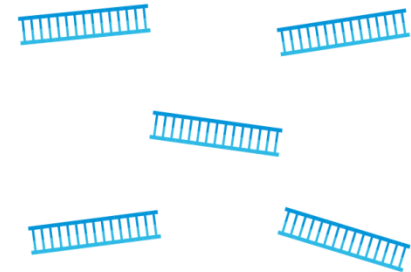
1 Isolar RNA das amostras



2 Fragmentar RNA em pequenos trechos (até 500-1000 nt)



3 Converter fragmentos de RNA em cDNA (transcriptase reversa)



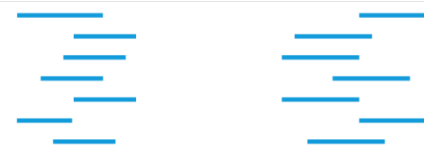
4 Ligação de adaptadores e amplificação



5 sequenciamento NGS

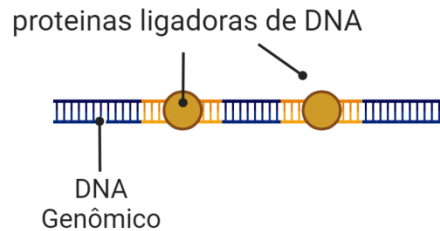


6 Mapear reads no transcriptoma/genoma

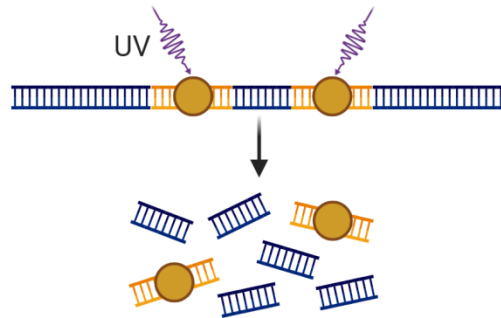


# ChIP-seq: NGS de DNA associado a sítios de ligação de fatores de transcrição e proteínas regulatórias da cromatina

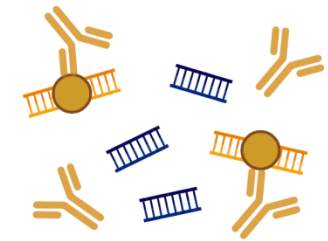
① Proteínas ligadas ao DNA genômico (ex. fatores de transcrição, Histonas)



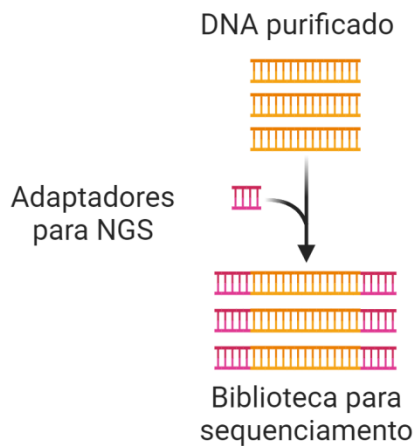
② Crosslinking e fragmentação do DNA



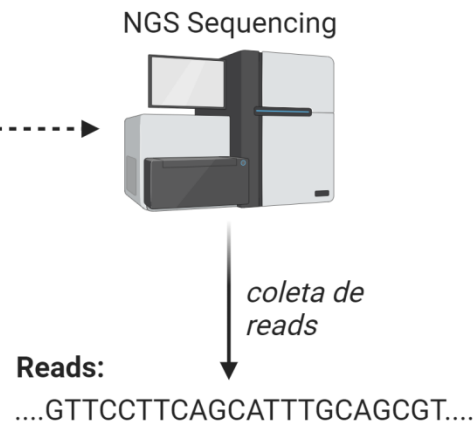
③ Imunoprecipitação do DNA ligado a proteínas com anticorpos específicos



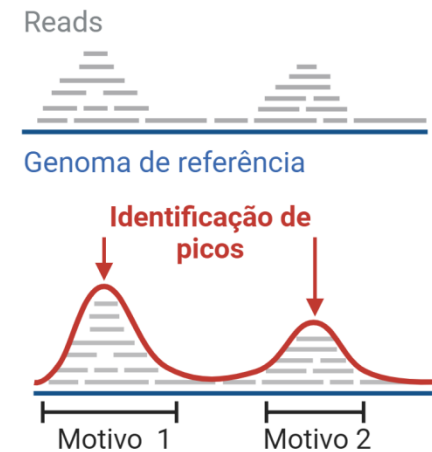
④ Purificação do DNA e ligação de adaptadores



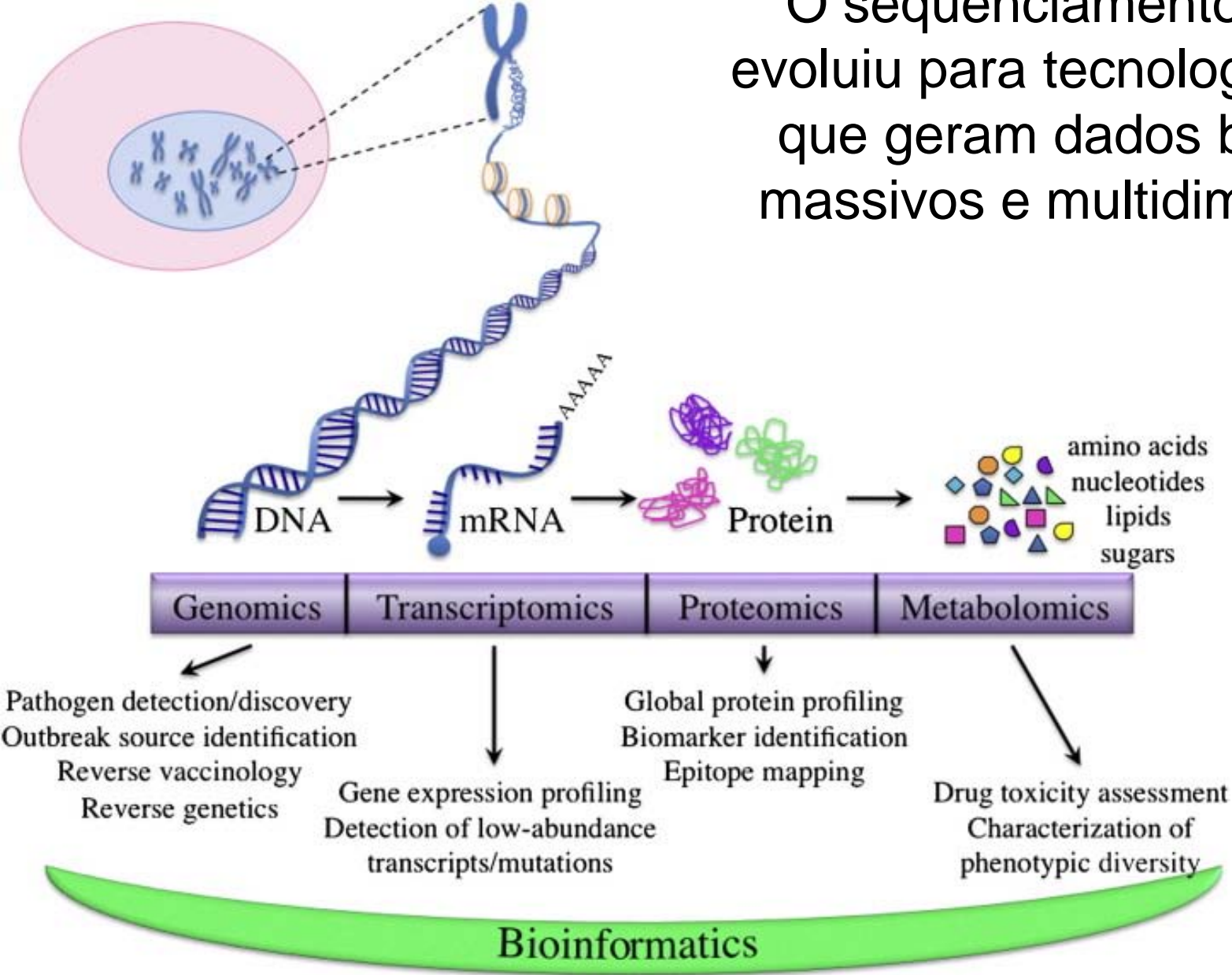
⑤ Sequenciamento NGS



⑥ Análise de sequências e identificação de sítios de ligação no DNA

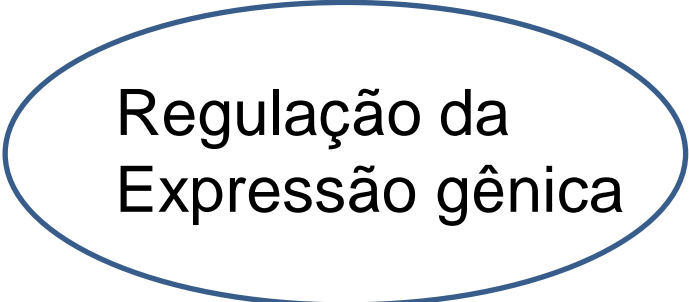


O sequenciamento do DNA evoluiu para tecnologias ômicas que geram dados biológicos massivos e multidimensionais



Nível de expressão  
(RNAseq, scRNAseq)

Estado de ativação da cromatina  
(ChIP-seq)



Regulação da  
Expressão gênica

Enriquecimento  
de categorias gênicas  
(inferências biológicas)

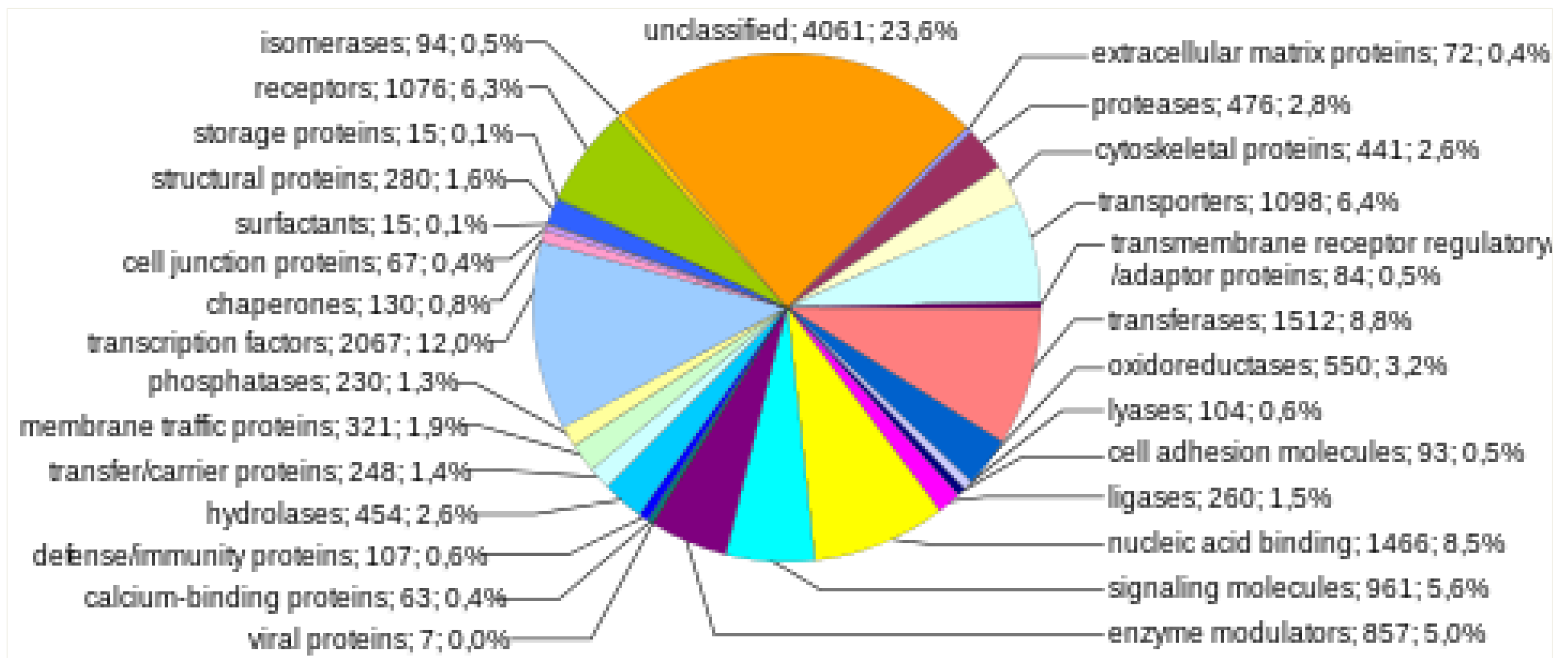
Estrutura secundária  
de RNAs  
(relação com função)

microRNAs  
(redes regulatórias)



# Sequenciamento completo do genoma humano revelou a presença de cerca de 30 mil genes codificadores de proteínas e pelo menos 60 mil RNAs não codificadores

Categorias funcionais de proteínas codificadas no genoma  
(número genes e % do total de genes)



Tão importante quanto o conteúdo gênico é **como, onde e quando cada um dos genes é expresso** (= transcrito pela RNA polimerase)!

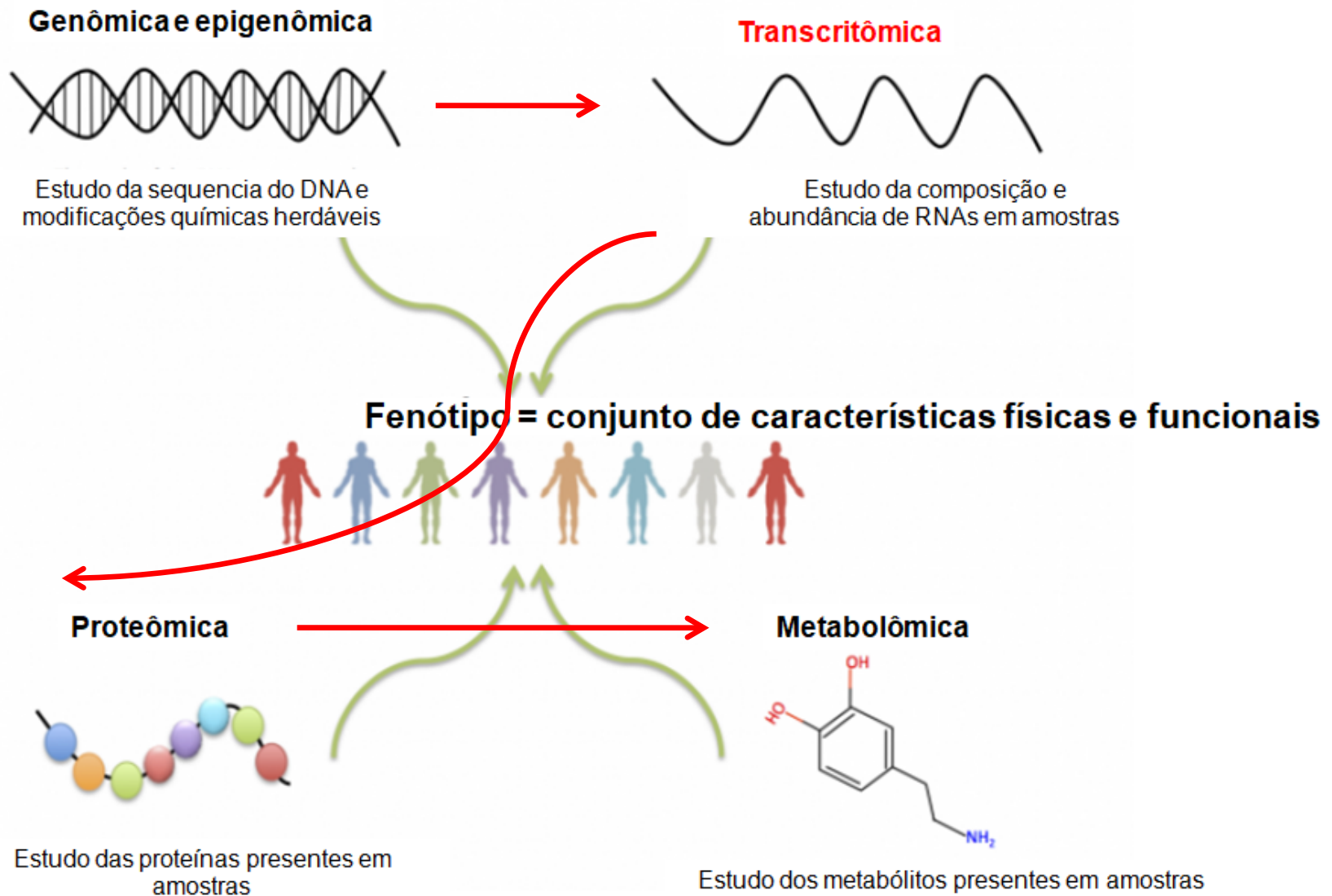


## Transcritoma

Conjunto de **todas** moléculas de RNA (transcritos) existentes nas células em um dado momento/condição

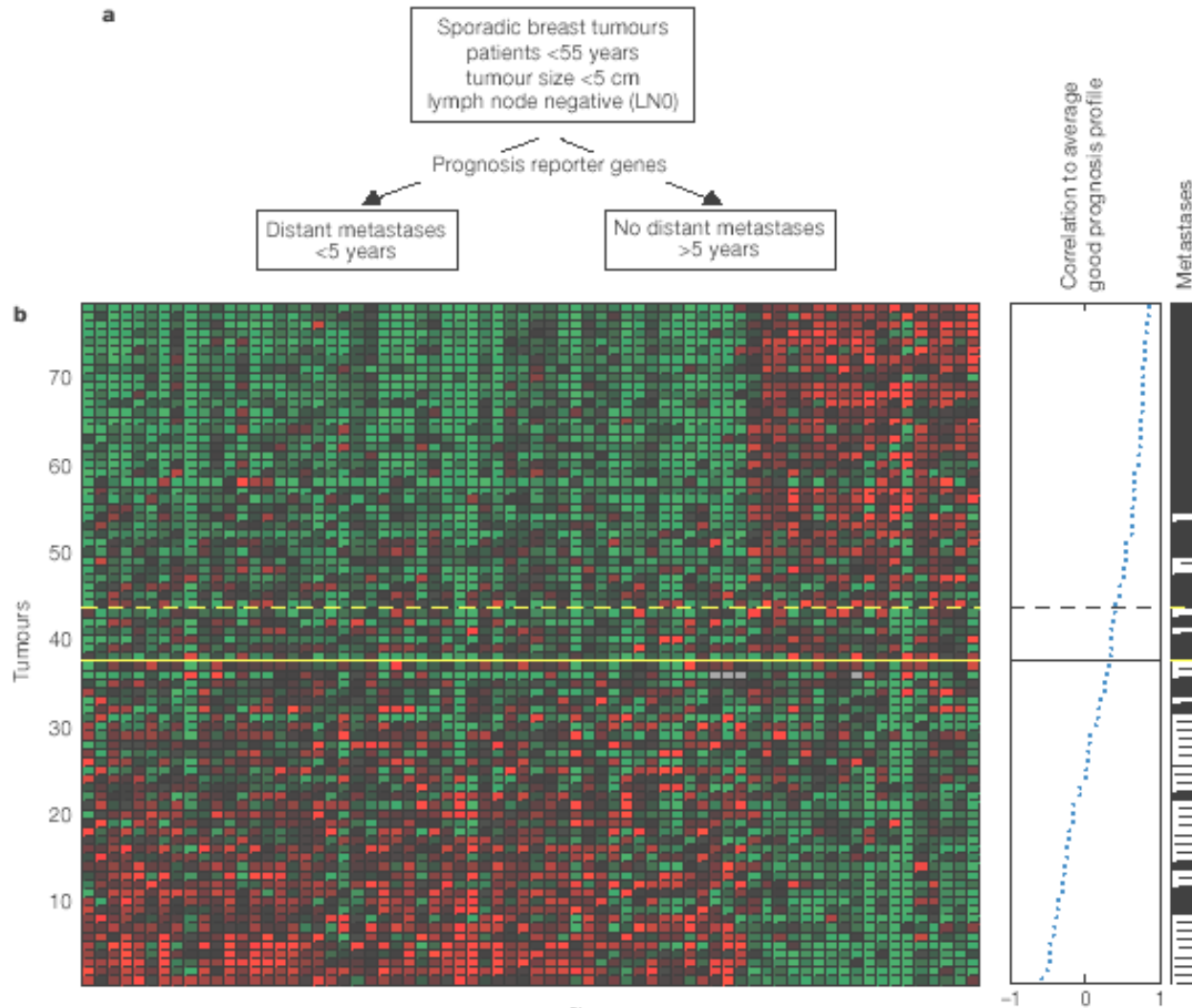
- Dinâmico
- Variável

# A análise do transcriptoma permite **definir as regiões ativas do genoma** que contribuem para a estrutura e função da célula/organismo



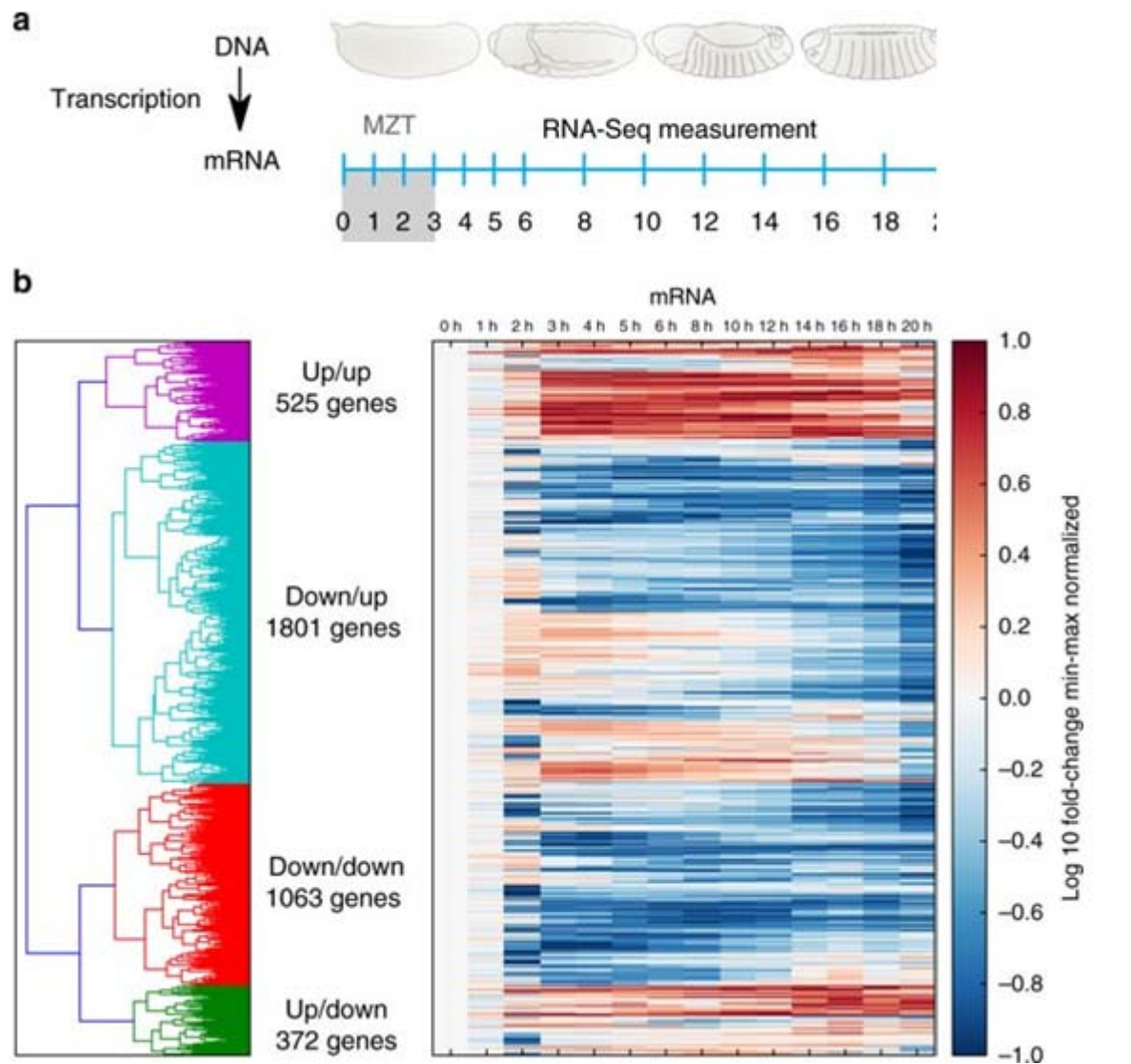
# Composição dinâmica do transcriptoma: varia em estados patológicos

Assinatura de expressão gênica associada ao aparecimento de metástase em pacientes com câncer de mama



# Composição dinâmica do transcriptoma: Varia ao longo do desenvolvimento

Alterações na expressão gênica ao longo do desenvolvimento embrionário da mosca *Drosophila melanogaster*



Becker, et al. Quantifying post-transcriptional regulation in the development of *Drosophila melanogaster*. Nat Commun 9, 4970 (2018).

# Composição dinâmica do transcriptoma: varia com o tipo de tecido

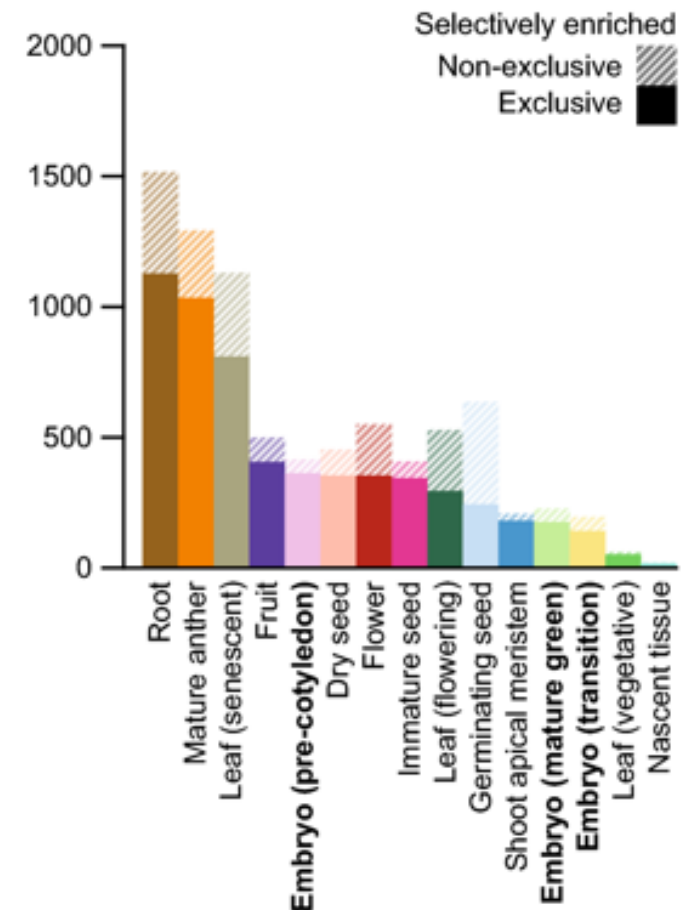
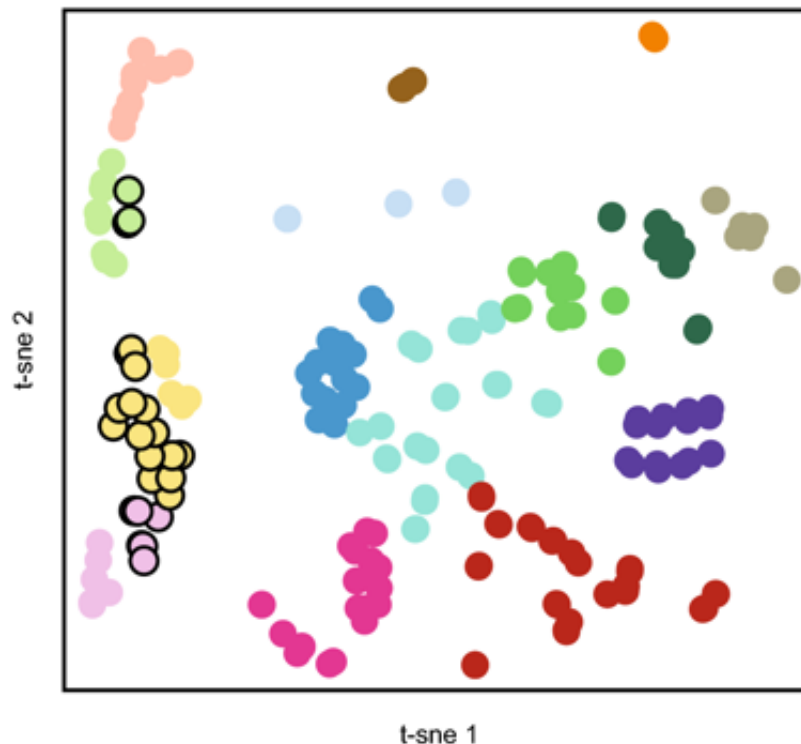
Comparação do transcriptoma de diferentes tecidos da planta *Arabidopsis thaliana*



Tissue cluster

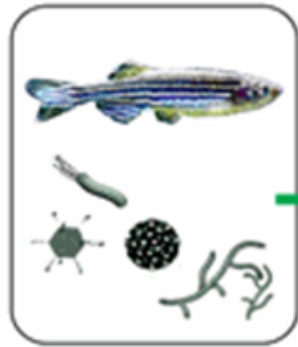
- Embryo (pre-cotyledon)
- Embryo (transition)
- Embryo (mature green)
- Dry seed
- Germinating seed
- Root
- Shoot apical meristem
- Nascent tissue
- Leaf (vegetative)
- Leaf (flowering)
- Leaf (senescent)
- Flower
- Fruit
- Immature seed
- Mature anther

- This work
- Other studies

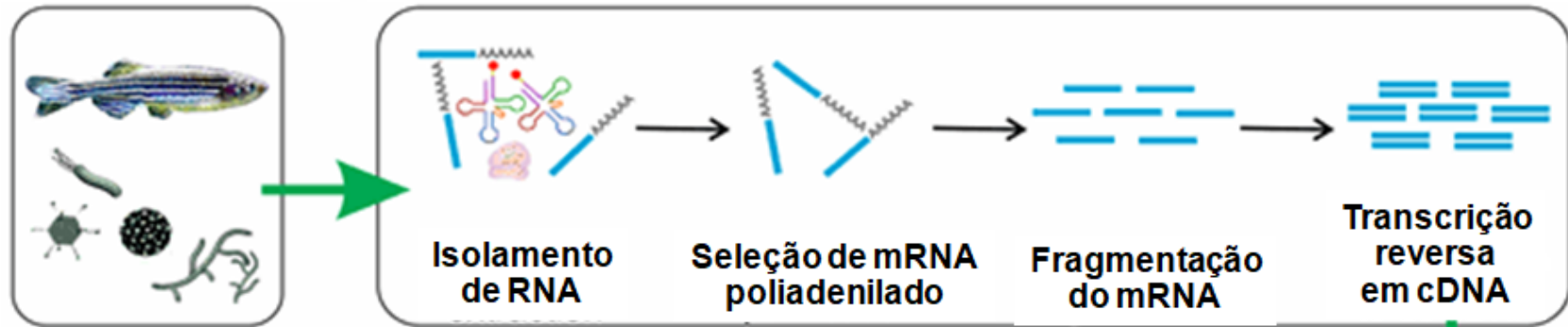


# Etapas na análise do transcriptoma através de RNAseq

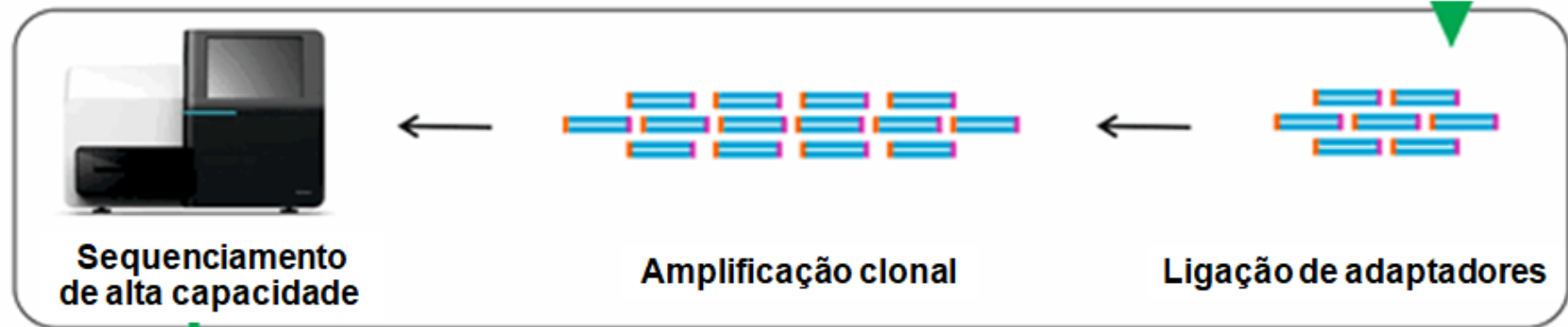
## Amostra de interesse



## Processamento & geração de bibliotecas para sequenciamento



## Sequenciamento de “nova geração” (Next Generation Sequencing)

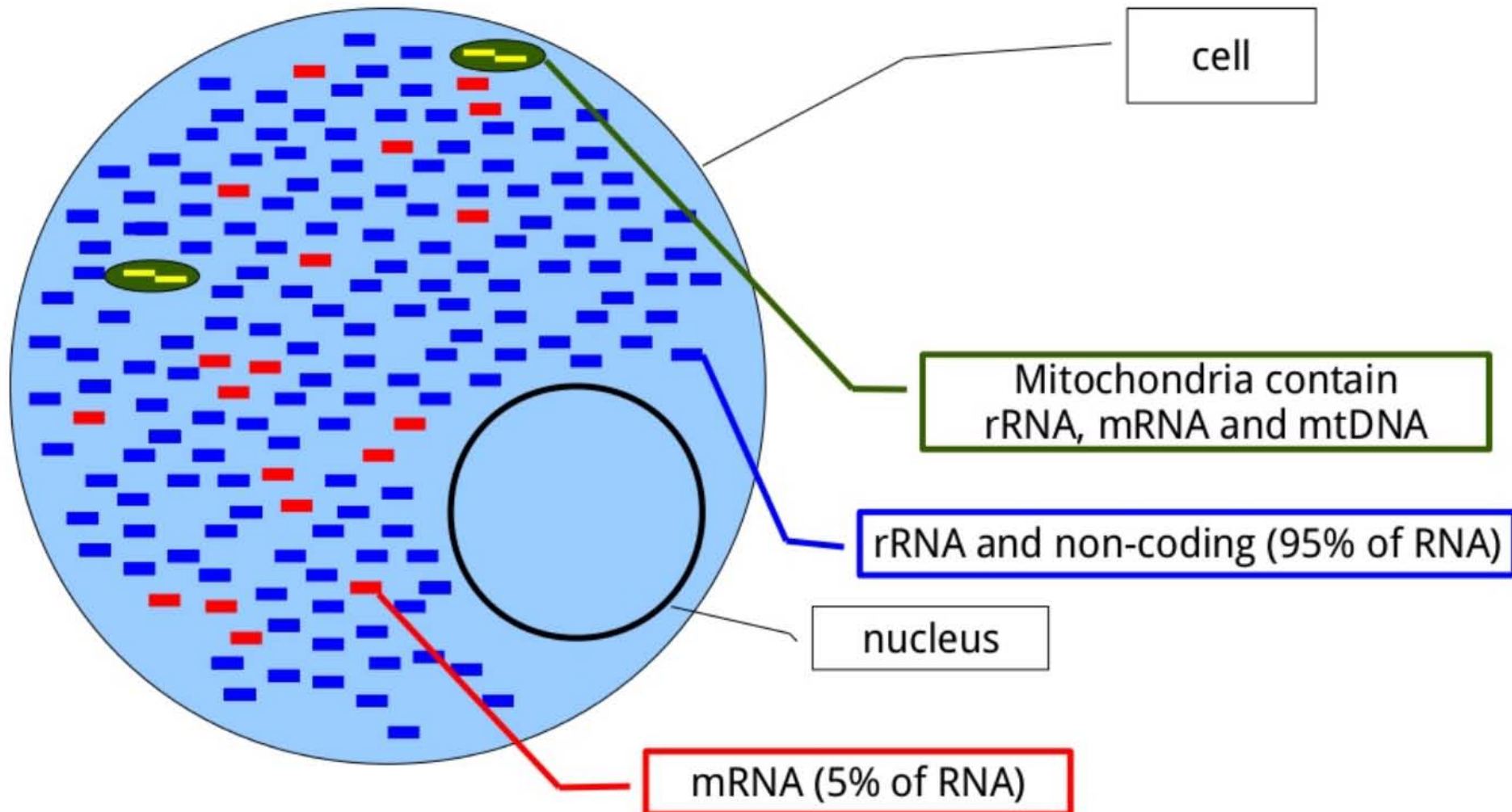


## Análise bioinformática



# Diferentes estratégias para construção de bibliotecas em função do tipo de RNA de interesse

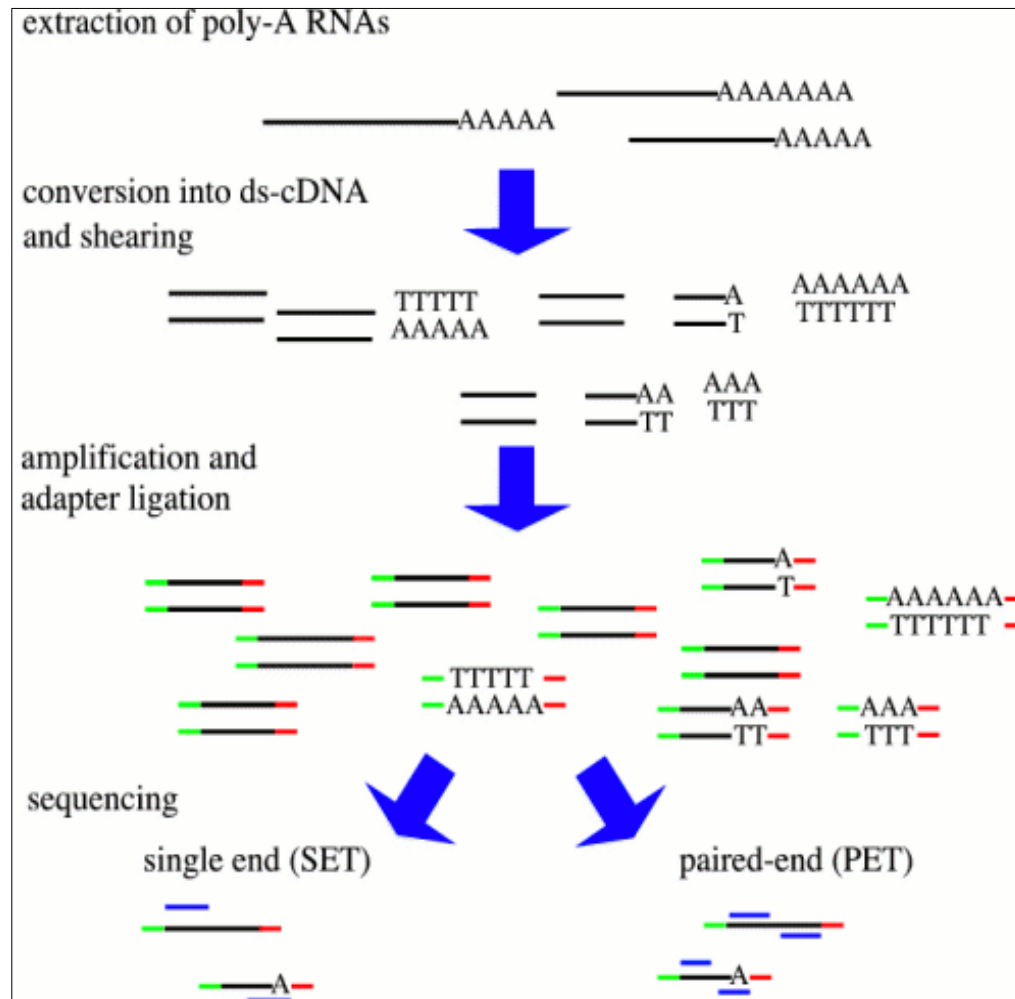
- **Análise global do transcrito**: depleção de RNA ribossomal
- **Análise de RNAs mensageiros**: enriquecimento de RNAs poliadenilados
- **Análise de pequenos RNAs** (ex. microRNAs): seleção por tamanho (<50 nt)





# Bibliotecas para sequenciamento de massivo de RNAs (RNAseq)

- Fragmentação mecânica ou química do RNA
- Ligação de adaptadores (bibliotecas direcionadas ou não-direcionadas)
- Conversão do RNA em DNA por transcrição reversa (cDNA).
- Sequenciamento de uma ou ambas as fitas (“paired-end”)



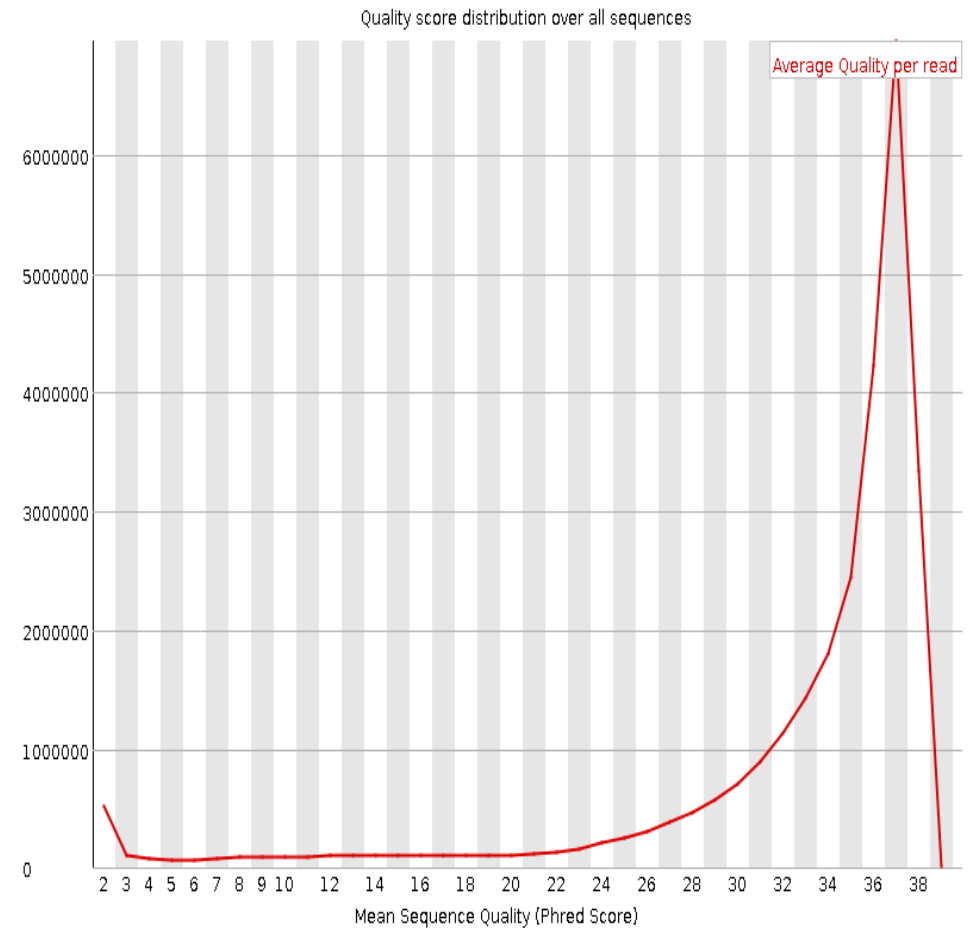
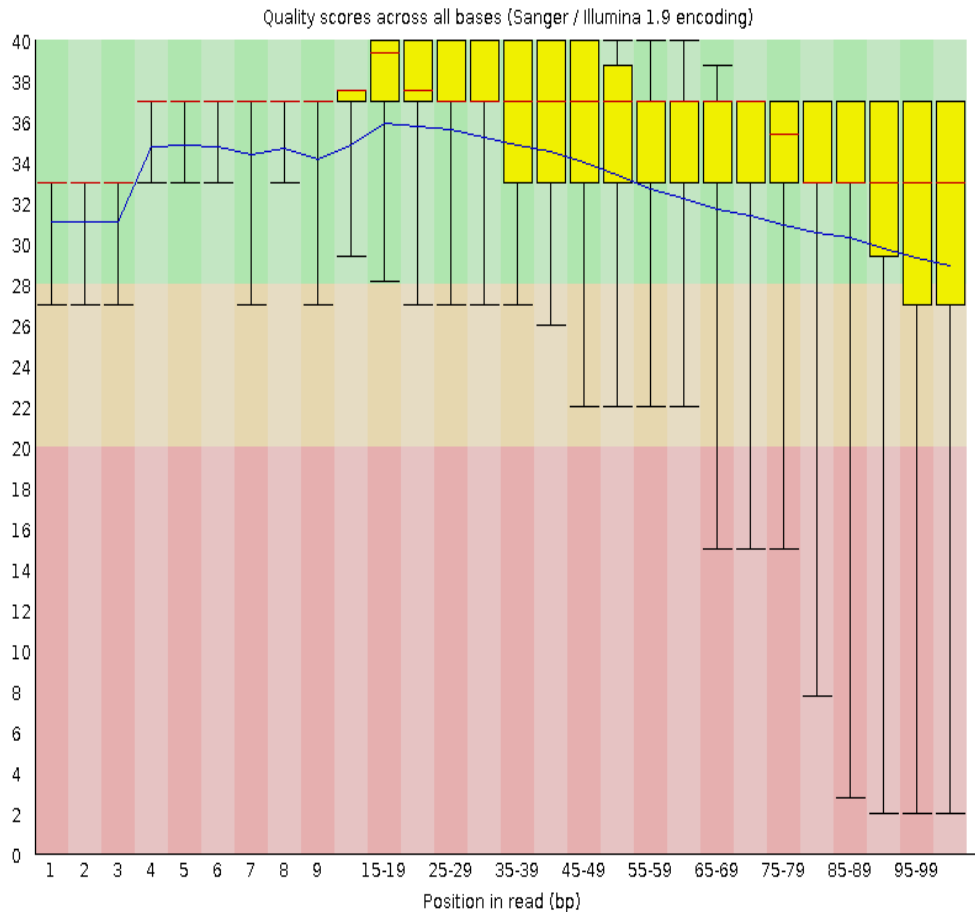
# Etapas na análise de dados de RNA-seq

- Controle de qualidade, pre-processamento, filtragem
- Alinhamento (genoma/transcritoma de referência)
- Reconstrução (=montagem) de transcritomas com ou sem genoma/transcritoma de referência
- Normalização, quantificação, expressão diferencial
- Detecção de *splicing* alternativo, fusões/quimeras
- Análises de enriquecimento de categorias gênicas, redes de co-expressão.

Listagem atualizada de ferramentas para a análise de dados:  
[https://en.wikipedia.org/wiki/List\\_of\\_RNA-Seq\\_bioinformatics\\_tools](https://en.wikipedia.org/wiki/List_of_RNA-Seq_bioinformatics_tools)

# Passo 1: Avaliação da qualidade das sequencias

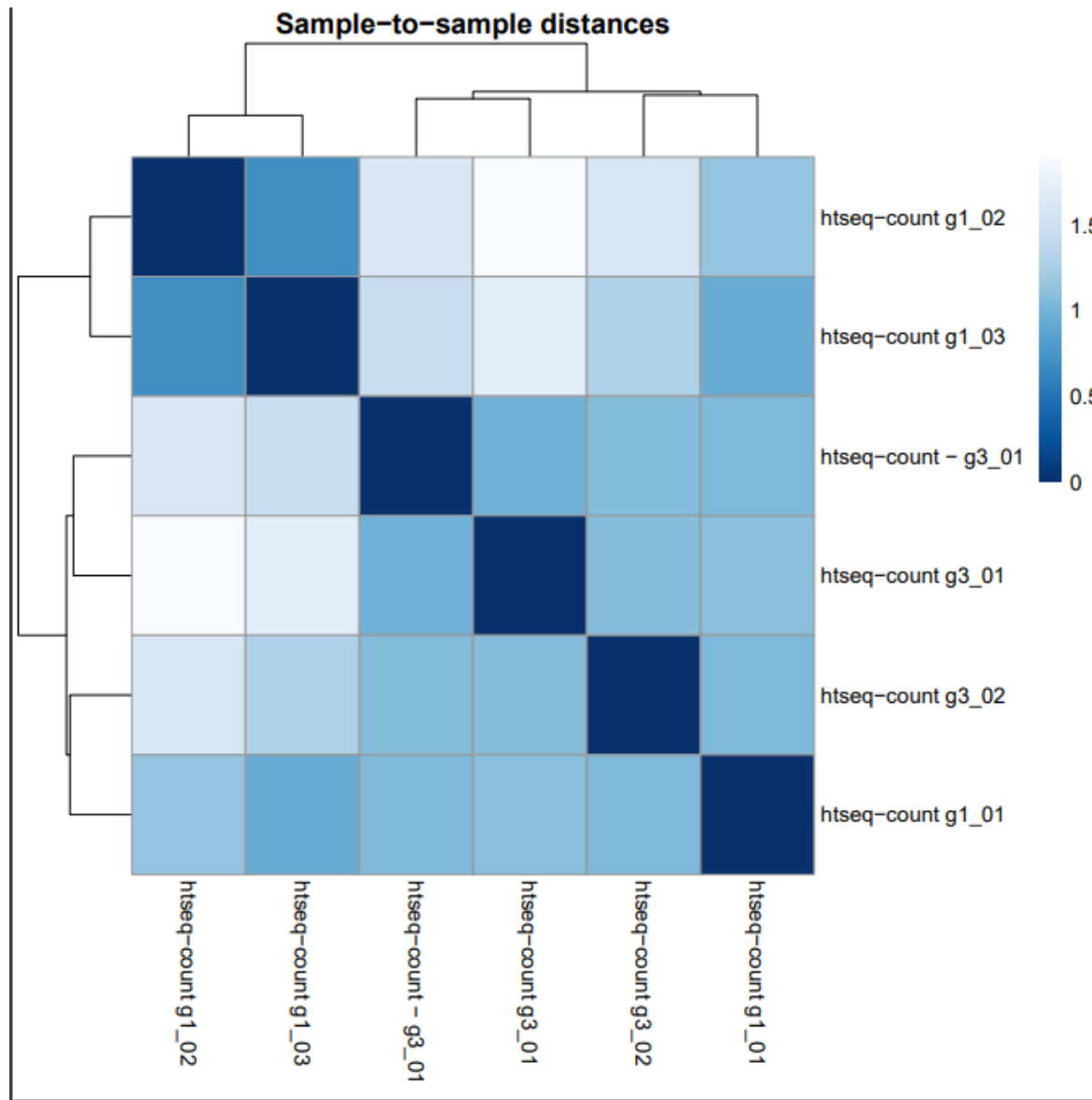
## Programa FastQC



Phred score =  $-10 * \log_{10} (\text{prob. de erro})$

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%

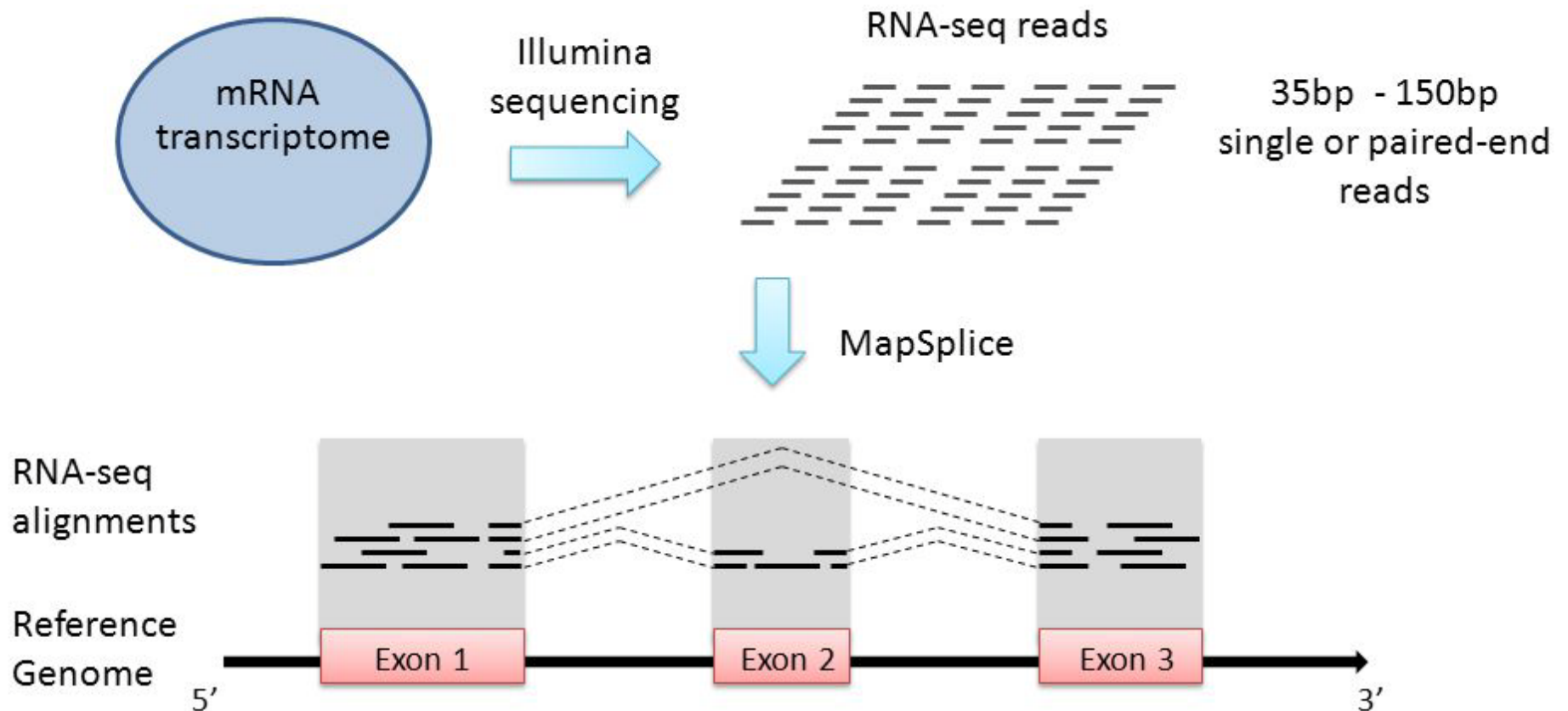
# DESeq2: Correlação entre expressão gênica nas amostras



# Passo 2: Alinhamento de reads no genoma

("gapped alignment") considera a presença de bordas exon/intron

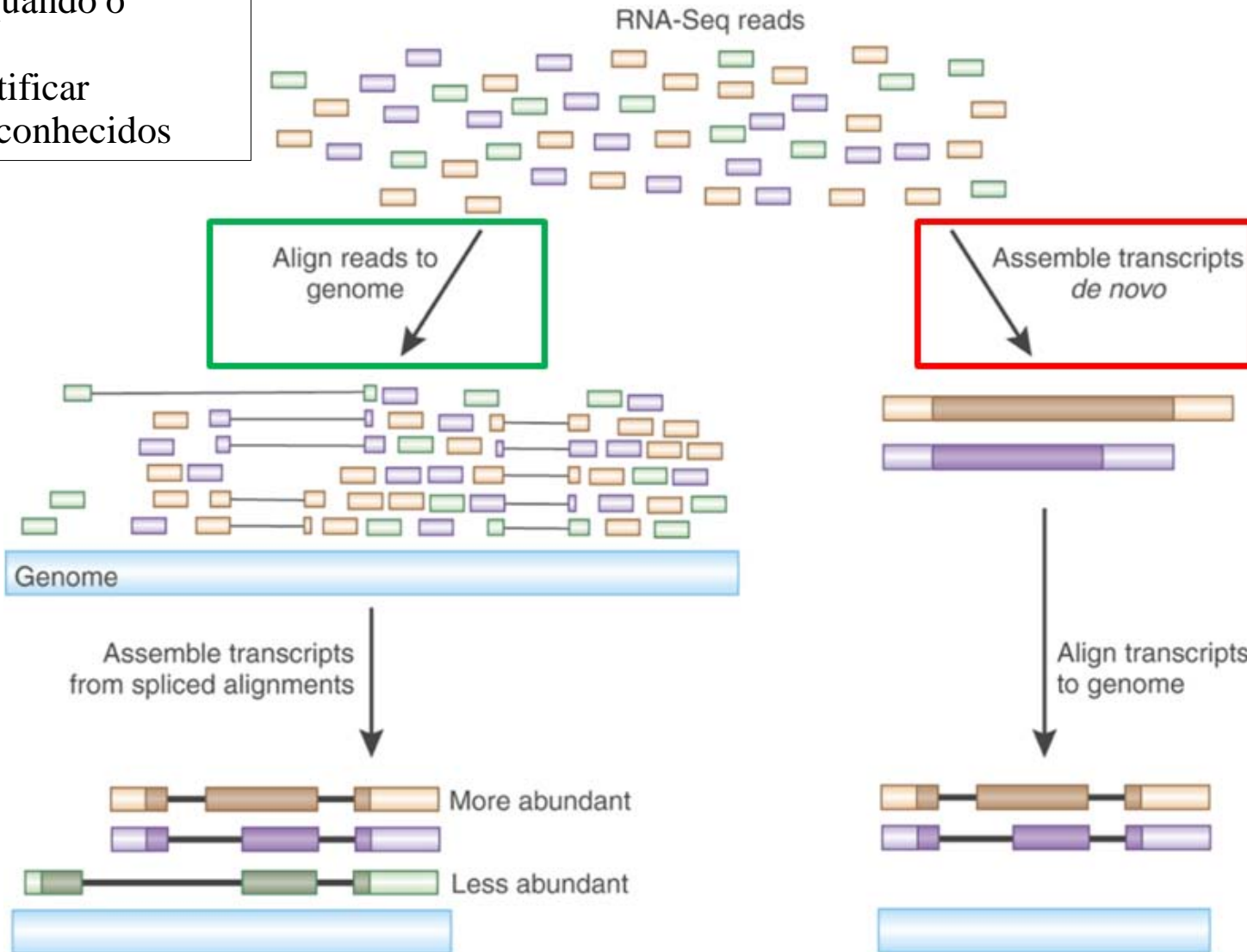
Exemplos de programas: STAR, TopHat, HISAT, RSEM



1. Wang K., et al., *MapSplice: Accurate Mapping of RNA-seq Reads for Splice Junction Discovery*, Nucleic Acids Research, 2010.
2. Hu Y., et al., *A Probabilistic Framework for Aligning Paired-end RNA-seq Data*, Bioinformatics, 2010

# Passo 3: Reconstrução e quantificação de transcriptomas a partir de dados de RNAseq

A etapa de reconstrução não é necessária quando o objetivo é detectar/quantificar transcritos já conhecidos

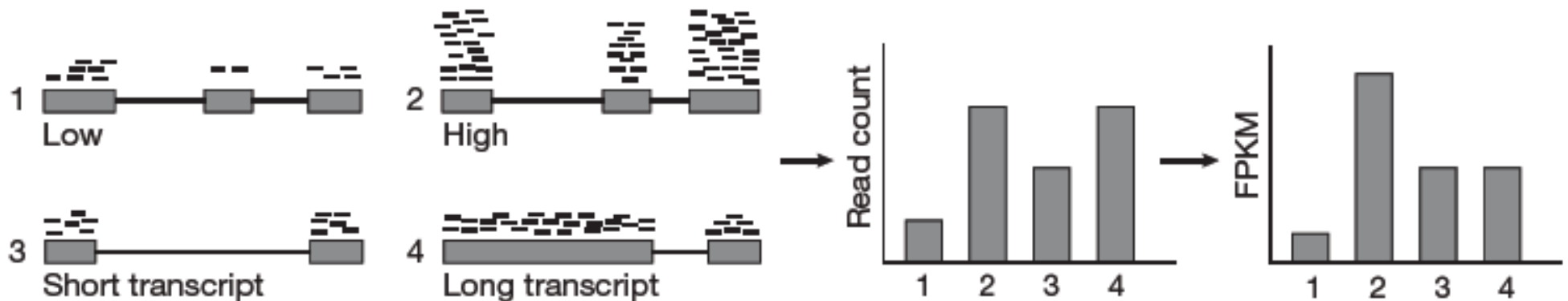


- **organismos sem genoma conhecido**

- **meta-transcritomas**

# Quantificação da expressão gênica através de RNA seq

Normalização pelo tamanho do transcrito (nº de pares de base) e tamanho da biblioteca (nº de reads) permite comparar níveis de expressão entre genes com tamanhos diferentes



*Fragments per kilobase of exon model per million mapped reads*

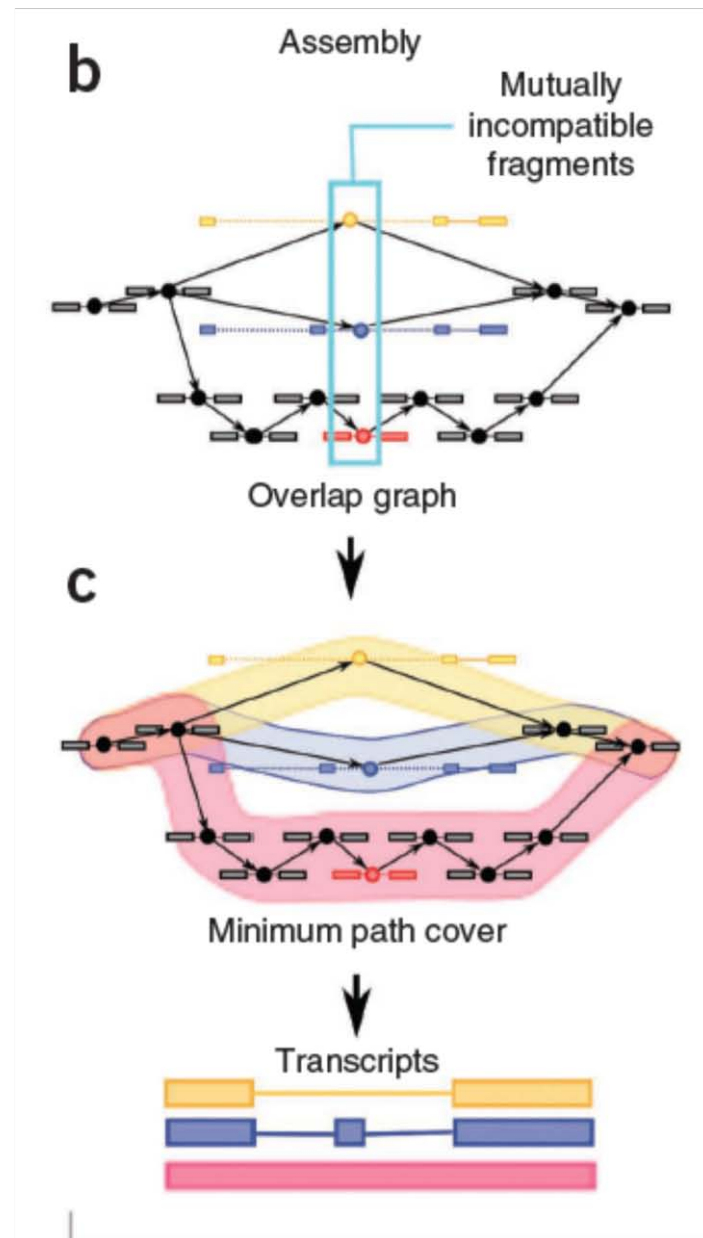
$$\text{FPKM} = \frac{\text{n}^\circ \text{ de fragmentos (reads) que mapeiam em exons do transcrito}}{\text{n}^\circ \text{ total de sequencias (milhões)} \times \text{tamanho dos exons (KB)}}$$

Outras formas de normalização são recomendadas para comparação entre amostras: Transcripts Per Million (TPM), Trimmed Mean of M-values (TMM)

# Como lidar com a expressão de RNA com variantes de splicing?

## Cufflinks

- Programa de montagem de sequências de NGS
- Utiliza as posições das coordenadas genômicas de reads alinhados
- Faz a reconstrução de transcritos completos a partir dos fragmentos sequenciados
- Utiliza uma abordagem de parsimônia para distribuir os reads mapeados no genoma entre as diferentes isoformas
- Permite identificar e estimar a abundância relativa de genes e formas de *splicing*

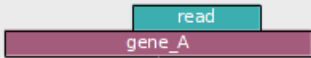
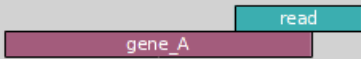



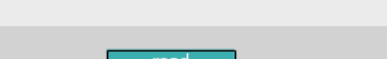






# HTseq

[https://htseq.readthedocs.io/en/release\\_0.9.1/](https://htseq.readthedocs.io/en/release_0.9.1/)

- Para cada gene, conta o número de reads alinhados no genoma que se sobrepoe com seus exons.
- Contagens não normalizadas são mais apropriadas para a análise de genes diferencialmente expressos usando programas que empregam modelos de inferência estatística (ex. DESeq2)
- Não são adequados para comparar com precisão a abundância de genes com tamanhos diferentes

	<b>union</b>	<b>intersection_strict</b>	<b>intersection_nonempty</b>
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous (both genes with --nonunique all)	gene_A	gene_A
	ambiguous (both genes with --nonunique all)		
	alignment_not_unique (both genes with --nonunique all)		

# Quantificação da expressão gênica através de RNA seq

## Nova geração de programas de quantificação


- Pseudoalinhamento de reads com transcrito de referencia
- não necessitam fazer o alinhamento dos reads no genoma
- Ultra-rápidos

## Salmon

Brief Communication | Published: 06 March 2017

### Salmon provides fast and bias-aware quantification of transcript expression

Rob Patro , Geet Duggal, Michael I Love, Rafael A Irizarry & Carl Kingsford 


*Nature Methods* **14**, 417–419 (2017) | [Download Citation](#) 

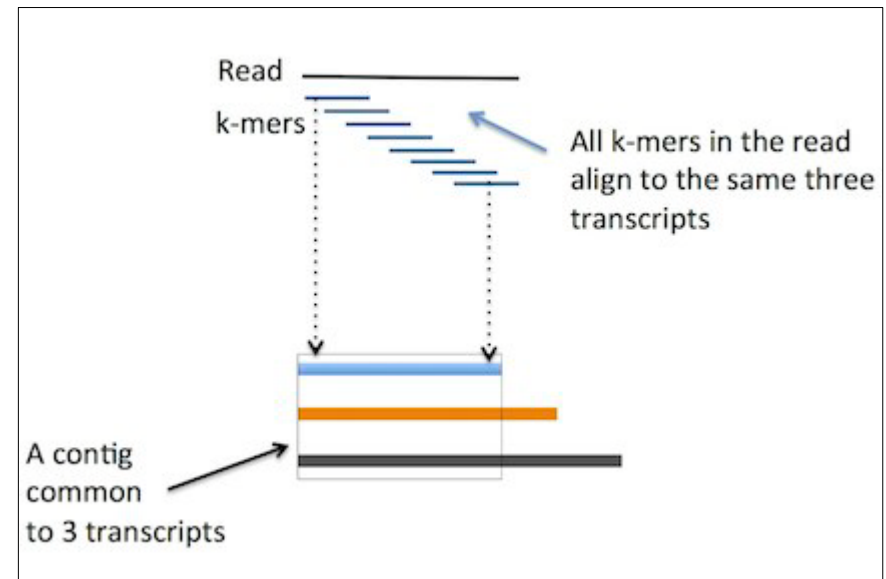
## Kallisto

Brief Communication | Published: 04 April 2016

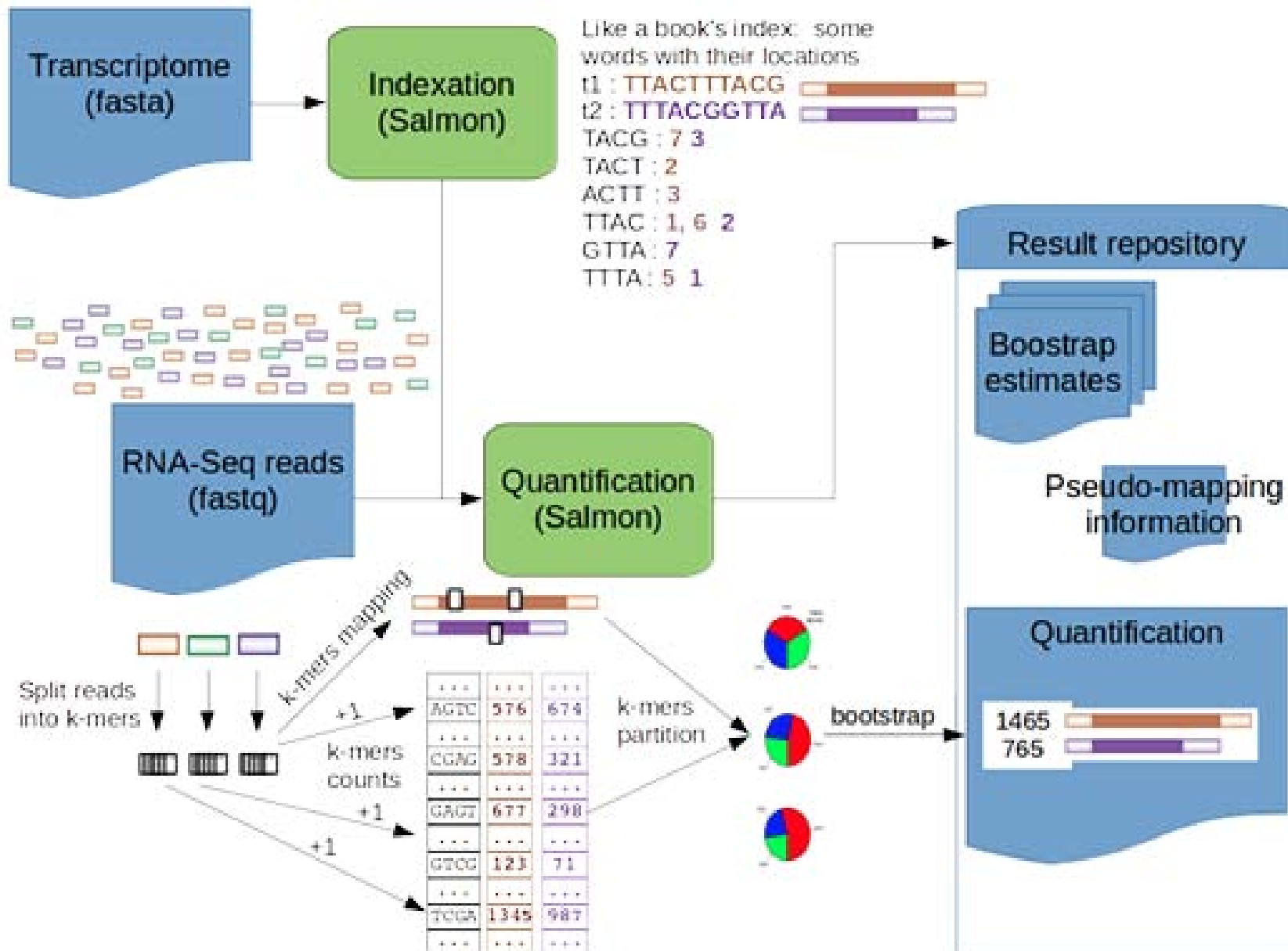
### Near-optimal probabilistic RNA-seq quantification

Nicolas L Bray, Harold Pimentel, Páll Melsted & Lior Pachter 

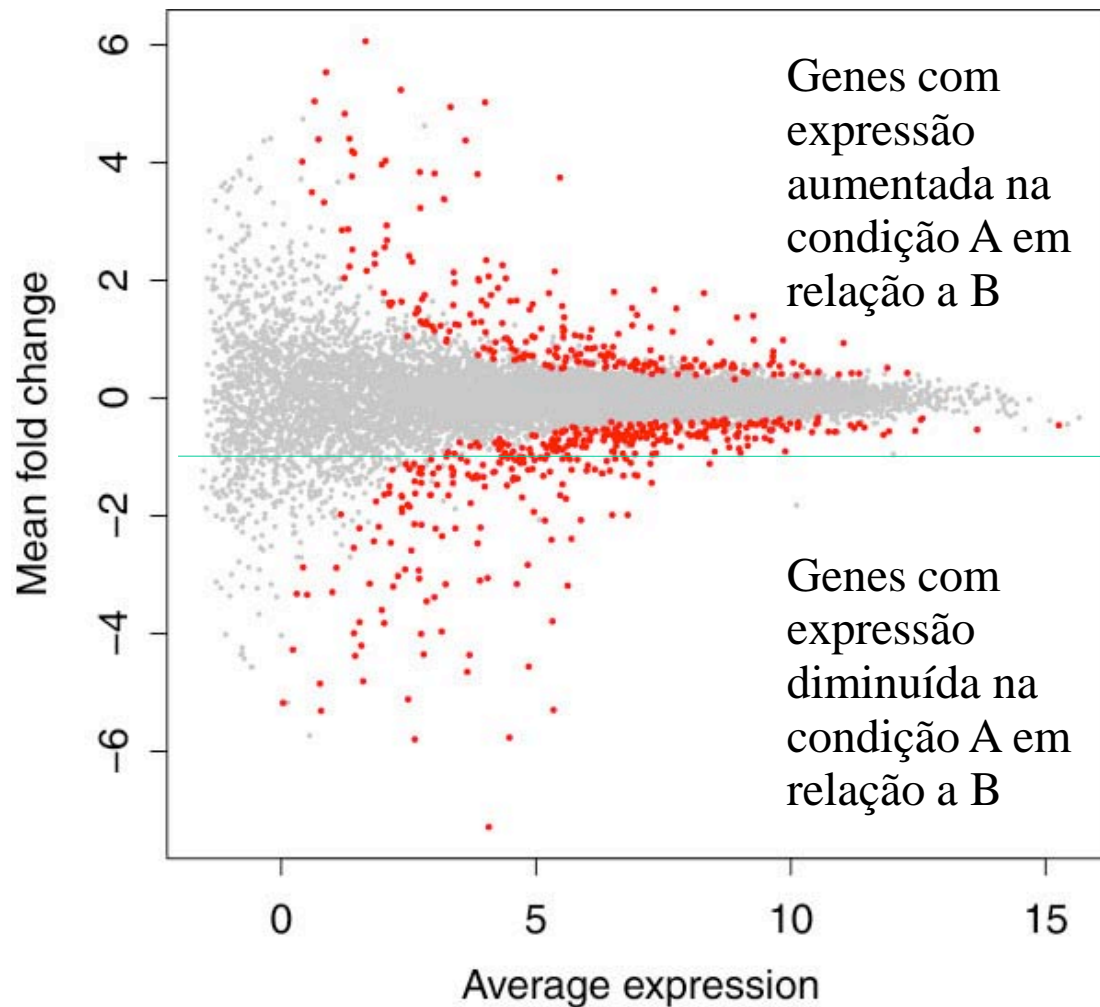
*Nature Biotechnology* **34**, 525–527 (2016) | [Download Citation](#) 



# Esquema geral do pipeline de quantificação do Salmon



# Passo 4: Identificação de genes com expressão alterada entre duas ou mais condições



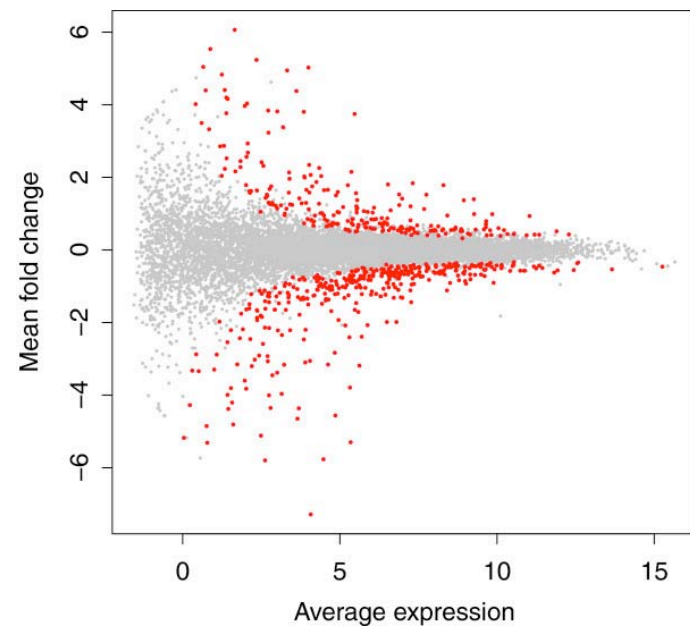
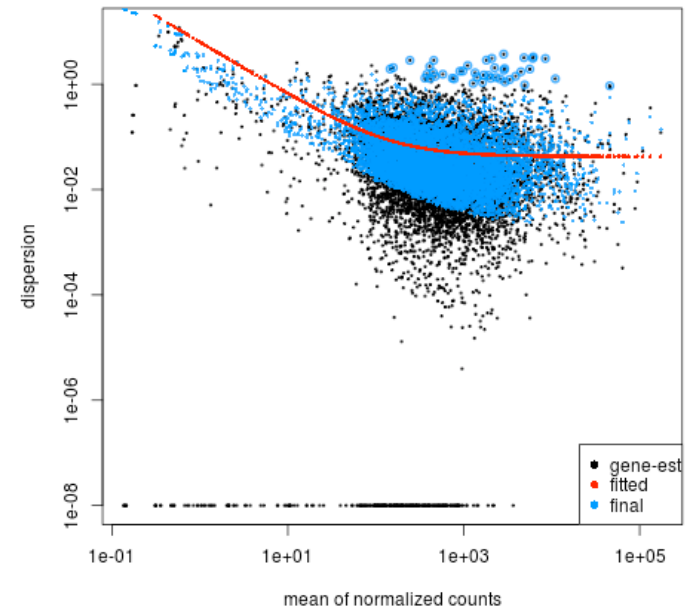
**Exemplos de programas:**  
**DESeq2**  
**CufDiff**  
**EdgeR**

- Utilizam como entrada dados de contagem
- Aplicam correções para o tamanho das bibliotecas
- Utilizam diferentes modelos para estimar e fazer inferência estatística na diferença de expressão entre condições

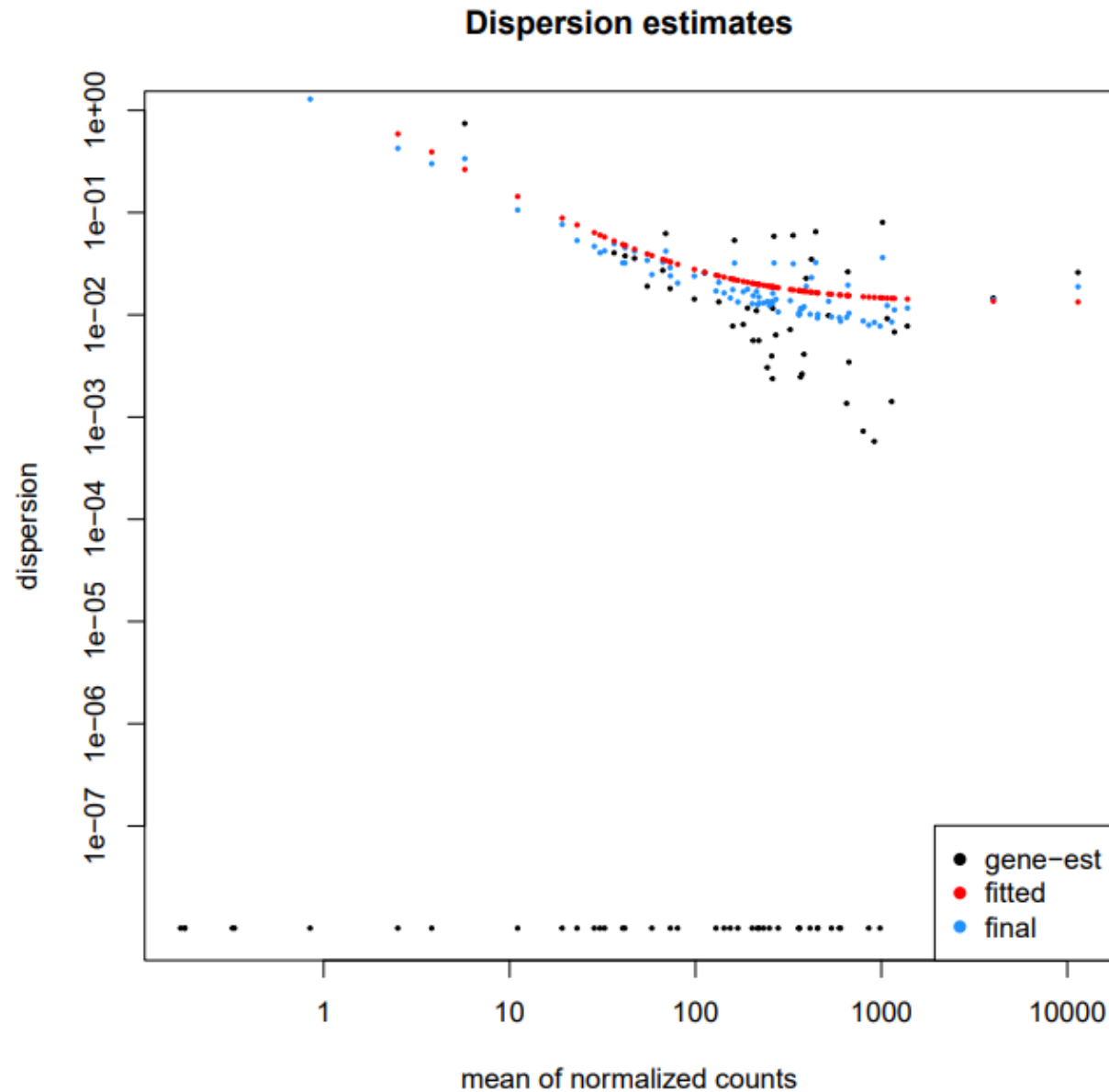
# DESeq2

[https://htseq.readthedocs.io/en/release\\_0.9.1/](https://htseq.readthedocs.io/en/release_0.9.1/)

- Utiliza como entrada dados de contagem não normalizados (ex. HTseq).
- Aplica correções para o tamanho das bibliotecas
- Utiliza um modelo linear generalizado baseado em distribuição binomial negativa para estimar e fazer inferência estatística na diferença de expressão entre condições



# Precisão da medida da expressão gênica proporcional a contagem de reads

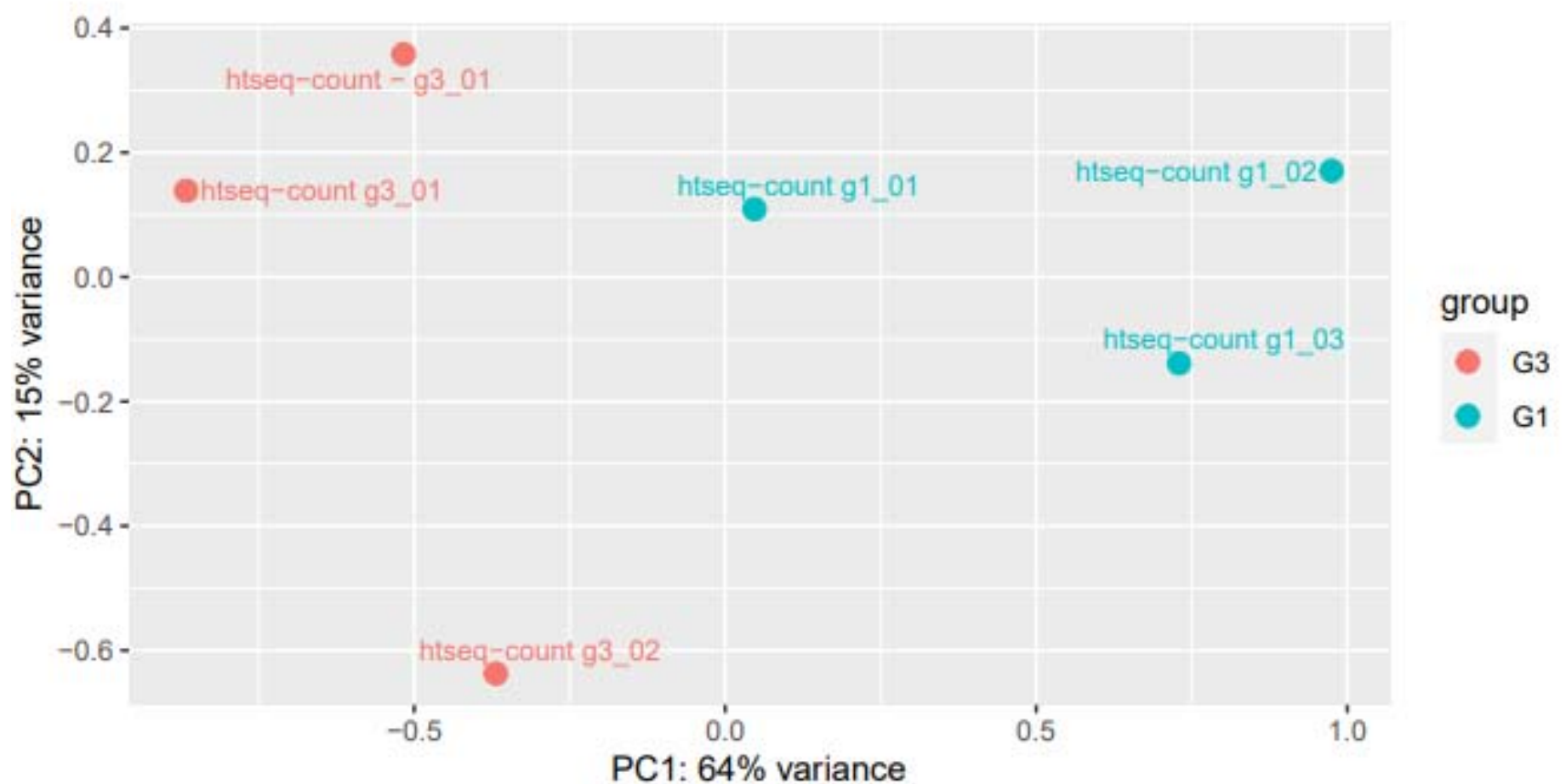


# Saida do DESeq2

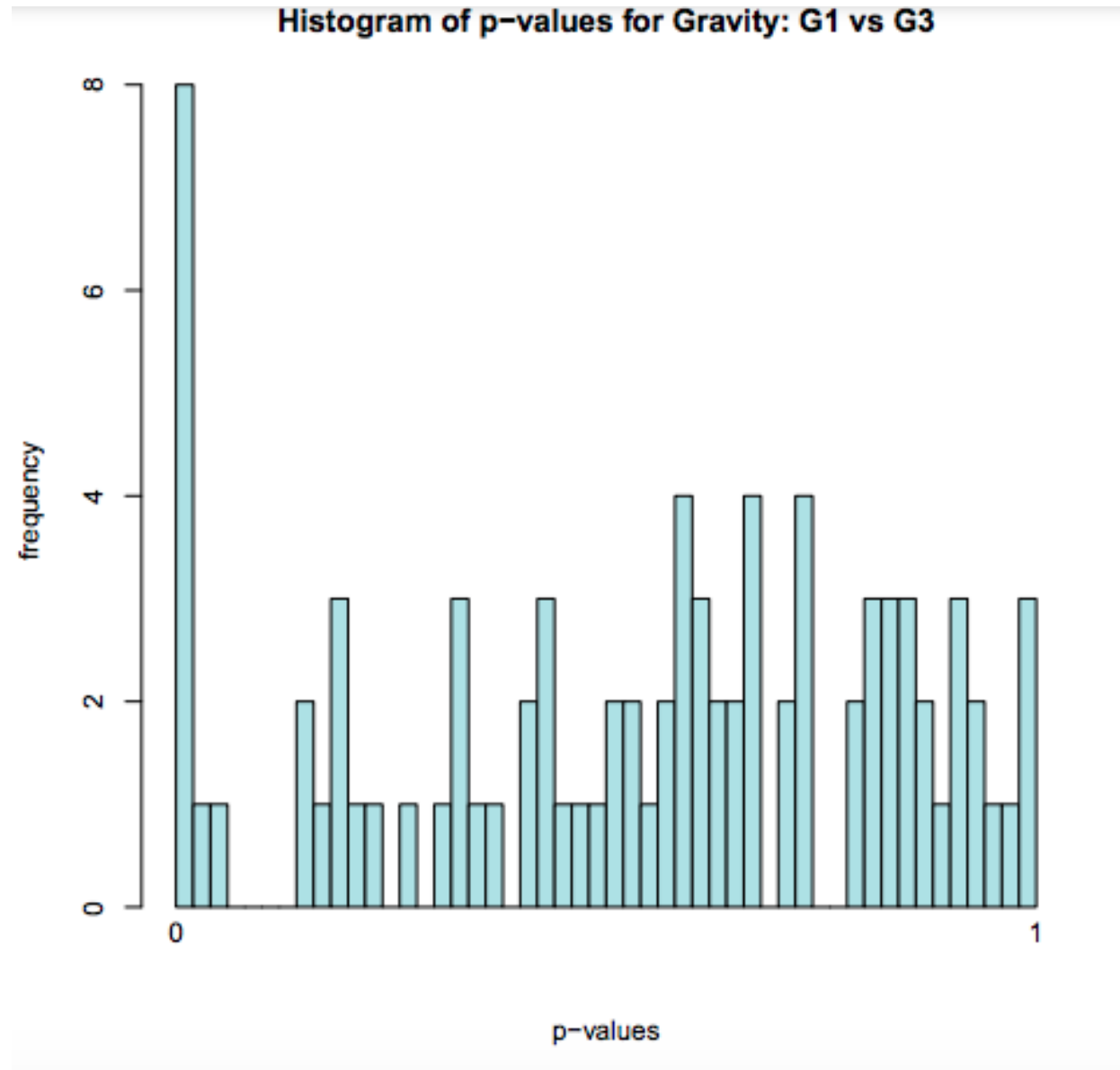
## Análise de Componentes Principais

Técnica de redução de dimensionalidade dos dados de expressão gênica.

A expressão de todos os genes detectados (N vetores) é projetada em nas duas dimensões que melhor representam a variabilidade dos dados (PC1 e PC2, eixos X e y) .



# Distribuição de valores de significância (p-valor) para a diferença de expressão de genes entre grupos teste e controle





---

**PROTOCOL**

# Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks

Cole Trapnell<sup>1,2</sup>, Adam Roberts<sup>3</sup>, Loyal Goff<sup>1,2,4</sup>, Geo Pertea<sup>5,6</sup>, Daehwan Kim<sup>5,7</sup>, David R Kelley<sup>1,2</sup>, Harold Pimentel<sup>3</sup>, Steven L Salzberg<sup>5,6</sup>, John L Rinn<sup>1,2</sup> & Lior Pachter<sup>3,8,9</sup>

---

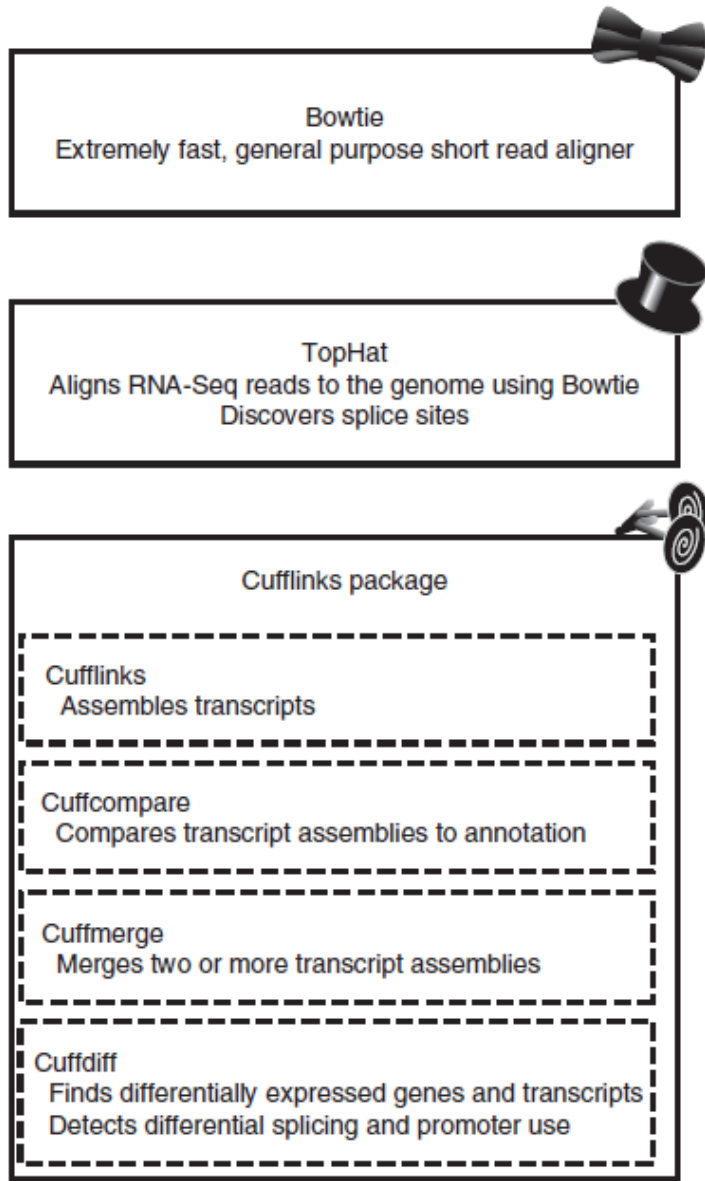
<sup>1</sup>Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. <sup>2</sup>Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts, USA. <sup>3</sup>Department of Computer Science, University of California, Berkeley, California, USA. <sup>4</sup>Computer Science and Artificial Intelligence Lab, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. <sup>5</sup>Department of Medicine, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. <sup>6</sup>Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland, USA. <sup>7</sup>Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland, USA. <sup>8</sup>Department of Mathematics, University of California, Berkeley, California, USA. <sup>9</sup>Department of Molecular and Cell Biology, University of California, Berkeley, California, USA. Correspondence should be addressed to C.T. (cole@broadinstitute.org).

Published online 1 March 2012; doi:10.1038/nprot.2012.016

Recent advances in high-throughput cDNA sequencing (RNA-seq) can reveal new genes and splice variants and quantify expression genome-wide in a single assay. The volume and complexity of data from RNA-seq experiments necessitate scalable, fast and mathematically principled analysis software. TopHat and Cufflinks are free, open-source software tools for gene discovery and comprehensive expression analysis of high-throughput mRNA sequencing (RNA-seq) data. Together, they allow biologists to identify new genes and new splice variants of known ones, as well as compare gene and transcript expression under two or more conditions. This protocol describes in detail how to use TopHat and Cufflinks to perform such analyses. It also covers several accessory tools and utilities that aid in managing data, including CummeRbund, a tool for visualizing RNA-seq analysis results. Although the procedure assumes basic informatics skills, these tools assume little to no background with RNA-seq analysis and are meant for novices and experts alike. The protocol begins with raw sequencing reads and produces a transcriptome assembly, lists of differentially expressed and regulated genes and transcripts, and publication-quality visualizations of analysis results. The protocol's execution time depends on the volume of transcriptome sequencing data and available computing resources but takes less than 1 d of computer time for typical experiments and ~1 h of hands-on time.

# Protocolo Tuxedo

## Exemplo de *pipeline* para análise de dados de RNAseq



### Step 1 Alinhamento

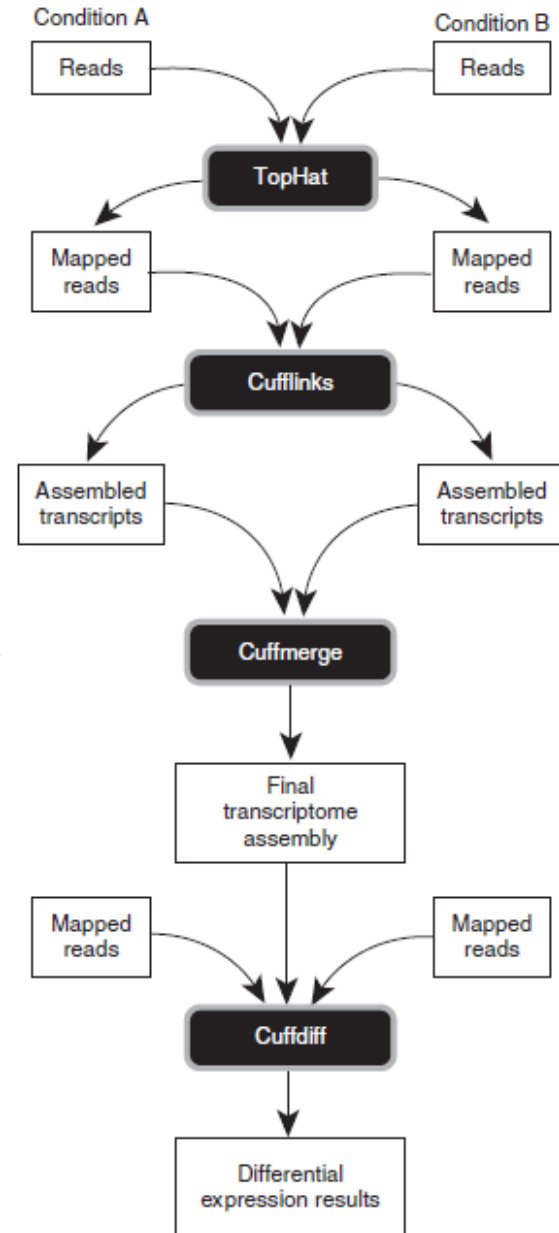
### Step 2

### Steps 3-4 Montagem/ quantificação

### Steps 3-4

### Step 5

### Expressão diferencial



# Galaxy: A Web-Based Genome Analysis Tool for Experimentalists

Daniel Blankenberg,<sup>1,5</sup> Gregory Von Kuster,<sup>1,5</sup> Nathaniel Coraor,<sup>1,5</sup>  
Guruprasad Ananda,<sup>1,5</sup> Ross Lazarus,<sup>2,5</sup> Mary Mangan,<sup>3</sup>  
Anton Nekrutenko,<sup>1,5</sup> and James Taylor<sup>4,5</sup>

<sup>1</sup>The Huck Institutes for the Life Sciences, Pennsylvania State University,  
University Park, Pennsylvania

<sup>2</sup>Channing Laboratory, Harvard Medical School, Boston, Massachusetts

<sup>3</sup>OpenHelix LLC, Bellevue, Washington

<sup>4</sup>Emory University, Atlanta, Georgia

<sup>5</sup>The Galaxy Team, Pennsylvania State University, University Park, Pennsylvania

## ABSTRACT

High-throughput data production has revolutionized molecular biology. However, massive increases in data generation capacity require analysis approaches that are more sophisticated, and often very computationally intensive. Thus, making sense of high-throughput data requires informatics support. Galaxy (<http://galaxyproject.org>) is a software system that provides this support through a framework that gives experimentalists simple interfaces to powerful tools, while automatically managing the computational details. Galaxy is distributed both as a publicly available Web service, which provides tools for the analysis of genomic, comparative genomic, and functional genomic data, or a downloadable package that can be deployed in individual laboratories. Either way, it allows experimentalists without informatics or programming expertise to perform complex large-scale analysis with just a Web browser. *Curr. Protoc. Mol. Biol.* 89:19.10.1-19.10.21. © 2010 by John Wiley & Sons, Inc.

Keywords: Galaxy • analysis • bioinformatics • workflow • algorithm • pipeline • genomics • SNPs

# https://usegalaxy.org/

The screenshot shows the Galaxy web interface. At the top, there's a browser address bar with the URL <https://main.g2.bx.psu.edu>. Below the browser, there's a navigation bar with tabs for 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Cloud', 'Help', and 'User'. The main content area features a central banner titled 'Running Your Own Understanding how Galaxy works' with the subtitle 'An in-depth tutorial'. Below this is a section for 'Live Quickies' with eight cards representing different tasks: 'Mapping against custom genome', 'Illumina mapping: Single Ends', 'Illumina mapping: Paired Ends', 'Basic fastQ manipulation', 'Advanced fastQ manipulation', '454 Mapping: Single End', 'Uploading Data using FTP', and 'Managing account histories'. To the left is a 'Tools' sidebar with a search bar and a list of tool categories. To the right is a 'History' sidebar showing a list of recent jobs with their names and sizes. At the bottom, there's a text box with the Galaxy build ID '\$Rev 7721:da9d740fce31\$' and a Twitter link for 'galaxyproject'. A disclaimer at the very bottom states that the resource is free and public, and that data is not encrypted.

Galaxy

Analyze Data Workflow Shared Data Visualization Cloud Help User Usina 22%

Tools

search tools

Get Data  
Send Data  
ENCODE Tools  
Lift-Over  
Text Manipulation  
Convert Formats  
FASTA manipulation  
Filter and Sort  
Join, Subtract and Group  
Extract Features  
Fetch Sequences  
Fetch Alignments  
Get Genomic Scores  
Operate on Genomic Intervals  
Statistics  
Graph/Display Data  
Regional Variation  
Multiple regression  
Multivariate Analysis  
Evolution  
Motif Tools  
Multiple Alignments  
Meta-genomic analyses  
Phenotype Association  
Genome Diversity  
EMBOSS

NGS TOOLBOX BETA  
NGS: QC and manipulation  
NGS: Mapping  
NGS: SAM Tools  
NGS: GATK Tools (beta)  
NGS: Variant Detection  
NGS: Indel Analysis  
NGS: Peak Calling  
NGS: RNA Analysis

Running Your Own  
Understanding how Galaxy works  
An in-depth tutorial

Live Quickies

Mapping against custom genome  
Illumina mapping: Single Ends  
Illumina mapping: Paired Ends  
Basic fastQ manipulation:  
Advanced fastQ manipulation:  
454 Mapping: Single End  
Uploading Data using FTP  
Managing account histories

Galaxy is an open, web-based platform for data intensive biomedical research. Whether on this free public server or [your own instance](#), you can perform, reproduce, and share complete analyses. The [Galaxy team](#) is a part of [BX at Penn State](#), and the [Biology](#) and [Mathematics and Computer Science](#) departments at [Emory University](#). The [Galaxy Project](#) is supported in part by [NSF](#), [NHGRI](#), [The Huck Institutes of the Life Sciences](#), [The Institute for CyberScience at Penn State](#), and [Emory University](#).

Galaxy build: \$Rev 7721:da9d740fce31\$

galaxyproject

more ...

This is a free, public, internet accessible resource. Data transfer and data storage are not encrypted. If there are restrictions on the way your research data can be stored and used, please consult your local institutional review board or the project PI before uploading it to any public site, including this Galaxy server. If you have protected data, large data storage requirements, or short deadlines you are encouraged to setup your own [local Galaxy instance](#) or run [Galaxy on the cloud](#).

History

Mapping/Annotation RIP 22.4 GB

167: best-hit allreads.txt

162: Sort on data 161

161: Compute on data 160

160: Cluster on data 158

158: Sort on data 157

157: Sort on data 156

156: Compute on data 155

155: Cluster on data 154

154: Subtract on data 152 and data 153

153: Subtract on data 150 and data 97

152: CCSD - 5Kb flanks

150: UCSC Main on Human: ccdsGene (genome)

140: AK023338 edited in ovary tumors IPPsp1


139: AHNAK-3UTR IPPsp1

138: PTP4A1-3UTR IPPsp1

137: BMP2-3UTR IPPsp1

Tools Options

- Get Data
- Send Data
- ENCODE Tools
- Lift-Over
- Text Manipulation
- Filter and Sort
- Join, Subtract and Group
- Convert Formats
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- Statistics
- Graph/Display Data
- Regional Variation
- Multiple regression
- Multivariate Analysis
- Evolution
- Motif Tools
- Multiple Alignments
- Metagenomic analyses
- FASTA manipulation
- NCBI BLAST+
- NGS: QC and manipulation
- NGS: Picard
- NGS: Assembly
- NGS: Mapping
- NGS: Indel Analysis
- NGS: RNA Analysis
- NGS: SAM Tools
- NGS: GATK Tools
- NGS: Peak Calling
- SNP/WGA: Data: Filters

 This dataset is large and only the first megabyte is shown below.  
[Show all](#) | [Save](#)

```
@61CC3AAXX100125:7:118:2538:5577/1
GACACCTTTAATGTCTGAAAAGAGACATTACCACATCTATTCTCTGGAGGGCTACCACCTAAGAGCCTTCA
+
?>CADFEEEBEDIEHHIDGGGEEEEHFFGIGIIFFIIEFHIIHIIFFIIIDEIIIGIIIEHFFPIIEHIF
@61CC3AAXX100125:7:1:17320:13701/1
CTCAGAAGACCCCTGAGAACATGTGCCAAGGTGGTCACAGTGCATCTTAGTTTTGTACATTTTAGGGAGAT
+
?BCAAADBBGGHGIDDDGHFEIIFIIIFGEIIFIIFIGIFEIIFGIIIEHFFHHIHEIFGHHIEFIIIEECE
@61CC3AAXX100125:7:93:5100:14497/1
CTCAACTGGCTGAAAGTATTATCAATAGAAAGGAATGTTCAAGTTCCTCAATTTTAGAGTGCCTGGCCTA
+
?BCACEEGGGFICFFDECEGDEHFGFDEEGGEIEGFIFHIGEIGHIIGHGHHEHFF@GIIIIIIIIHID
@61CC3AAXX100125:6:92:7549:15004/1
CTTTTGCCAGTGACTCATCTGGCAGGTATCTCAAGTCAGCCCTTGCTGGCCTGGCACCTTGCTGTGGTCT
+
?BBBCGFDDCHHHHFEHIIIFIEIDFIIIGGEFIEGIIIIHIIIIIIIIIIIGIHHIIIIIIIGFH
@61CC3AAXX100125:5:7:1488:7780/1
CCTGAGCTGCAGCACAGAGTGGAGGTAGTGGGGAGCTGTCACTGGGTATGCCCCCTTTCCCTGTGCCA
+
9==>.<CDEEB@FCFC@?@G=;AF<9<8@>;4.;G@DAE@9HCIH@<?728$'=B8@:68CB8>>8<8D
@61CC3AAXX100125:7:72:14903:20386/1
TTCCCTCTGAGGGCCCCACCCACTATACATCATCCCTTCATGGTGAGGGAGACTTCAGCCCTCAATGCCACC
+
?ACDDEFFHBCHHHHFHGGCHHDFFIIFIIIIHIGFIIIFIEEIEFEIHHIGFIIIIIGHCIIIFII
@61CC3AAXX100125:7:88:9942:19183/1
CCTGTCTCTGAGCTCCTTCTGGTCTGCTGCAGGGACAGGGCCACAGCCACCTGTCTGGCTTCTCTCT
+
?CCEAEFFHCCHHHGHIIHIIIEHHIIIIIFEIEIEIGEIIIGIGFIIIGIHHIIEIIDIHIGIHH
@61CC3AAXX100125:7:76:1585:2024/1
AAATTACAGTGGCTGGCAGAAGGAGAGAGAGCCAGGACAGGGGGCTGGGGCCTGTCCCCGCAGCAGAG
+
9.:8?;B@?CEEEFEAA;>;3:#####
@61CC3AAXX100125:6:26:17654:5573/1
AGGGCAGGGGGTTTGTCTGGGTTCAAGACCATGGAAGGAAGGGGTAGAGAAGGAGGCCAACAGTGAGG
+
??DEECBGGHDFGIEGHIHIIIFGIGGEEIIGFIIIEHEIEHFIIIFEFFHFIFFIIIGHIGCFGIEEI
@61CC3AAXX100125:7:117:7805:10957/1
AGCTGACCTTTCAAAGTGCATTCATGGGTCCTCCATGGAGAATCCAGTGCCTGAGCCATCACCTACAGCC
+
??DCFAFFPEFEGFEGDEIIFEGHFFIIIFIIIIIGFIIIEEHFIIGFFIIIIIF@IHHIGIHHIIEIGEII
@61CC3AAXX100125:7:36:11248:16392/1
GGTCTCAGCTGGCTGGGCCATTGGGAGTTCTATCTTGGGAAGGATGCTTTCCTAGGCTCTGAGATCG
+
?D@DDEDBGFHHHHIIIIIFPHGIII?EPHIIDPIHHIIIIAEPFI@GIIIIIIHAFIIIIICFBBIID
@61CC3AAXX100125:6:80:10088:8830/1
CAGGCTCAGGAATACACTGGCTCCCTCAAACCTGGGAATGTGCCAACCTGCCCTCCCAGCCTTCCAGCCC
+
?@?EEEFDBHCFDCHPHIHHIIHIIIIIFHDIIIII>HPIGIIIGHIIIIIIIIHIC9IIFC;IIDCIII
@61CC3AAXX100125:6:115:5701:20053/1
CTCTATCCATCTTGGGAGGTTGTATCCCTGAACTTCTAGAGCACATGTCCAGGCCCCCAAGCCTC
+
?BCC@BEFDCGGFIIIDIEGIEHFIIIIHIEGIIHIIIEEEEIGIGPHIIFIIGFIIIIIGCBFIHH
@61CC3AAXX100125:5:20:10205:7274/1
GCAACAGTGCCTGTCACTGCATCTGCTGGCTGCATCAAAATGCATGAGTTGGAGGCTTGGGGGCC
+
9=7;A?>CEE>AA?;?34=>;<99<?==#####
@61CC3AAXX100125:6:22:16350:6073/1
AACTCATGCACTTGTATCAGGCAGCCAGCAGATGCAGGTGACAGGCACTGTTGCCACCAGCCAGGATGGCC
```

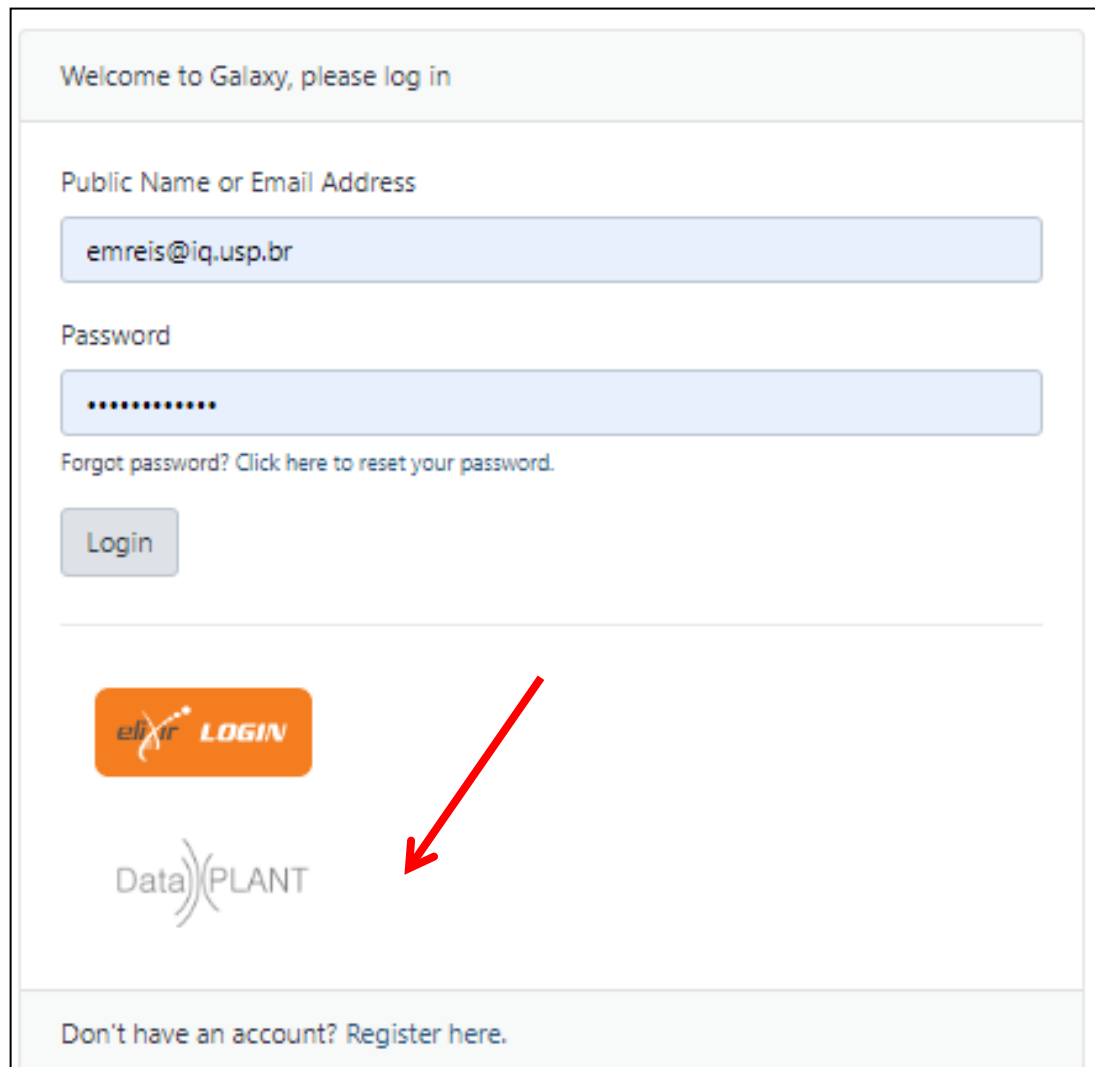
History Options

imported: 1.3 Gb  
 BioInfoSummer\_NGSworkshop

- 16: UCSC Main on Human: knownGene (chr22:1-51304566)
- 15: NA12878 high confidence Indels
- 14: Extract indels on data 4
- 13: NA12878 high confidence SNPs
- 12: Filter pileup on data 9
- 10: Filter pileup on data 9
- 9: Generate pileup on data 5: converted pileup
- 5: NA12878.chr22\_exome.BWA mapped.chr22\_filtered  
 57.6 Mb  
 format: bam, database: hg19  
 Info: Samtools Version: 0.1.12a (r862)SAM file converted to BAM  
 display at UCSC main  
 display at Ensembl Current  
 Binary bam alignments file

# Criar conta no servidor **Galaxy-Europa** para executar o tutorial:

<https://usegalaxy.eu/>




Welcome to Galaxy, please log in


Public Name or Email Address

Password

[Forgot password? Click here to reset your password.](#)

---





[Don't have an account? Register here.](#)

A red arrow points from the top right towards the Data) (PLANT logo.