

CURSO PRÁTICO DE BIOINFORMÁTICA:

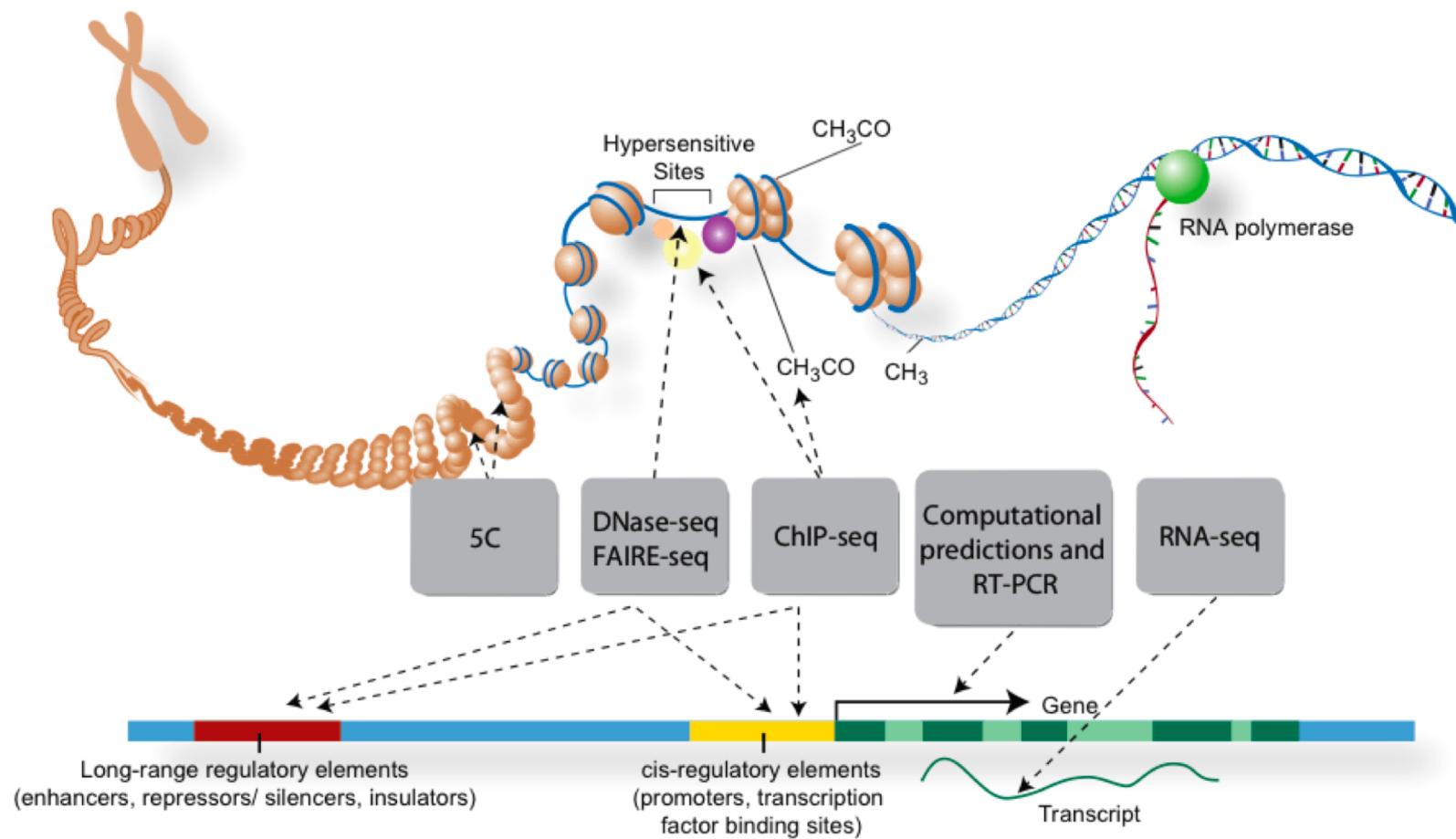
Aula 6 – Outras Aplicações NGS

<https://github.com/rpmartin85/BIOINFO>

Renan Paulo Martin

22/07/19

Outras Aplicações do NGS

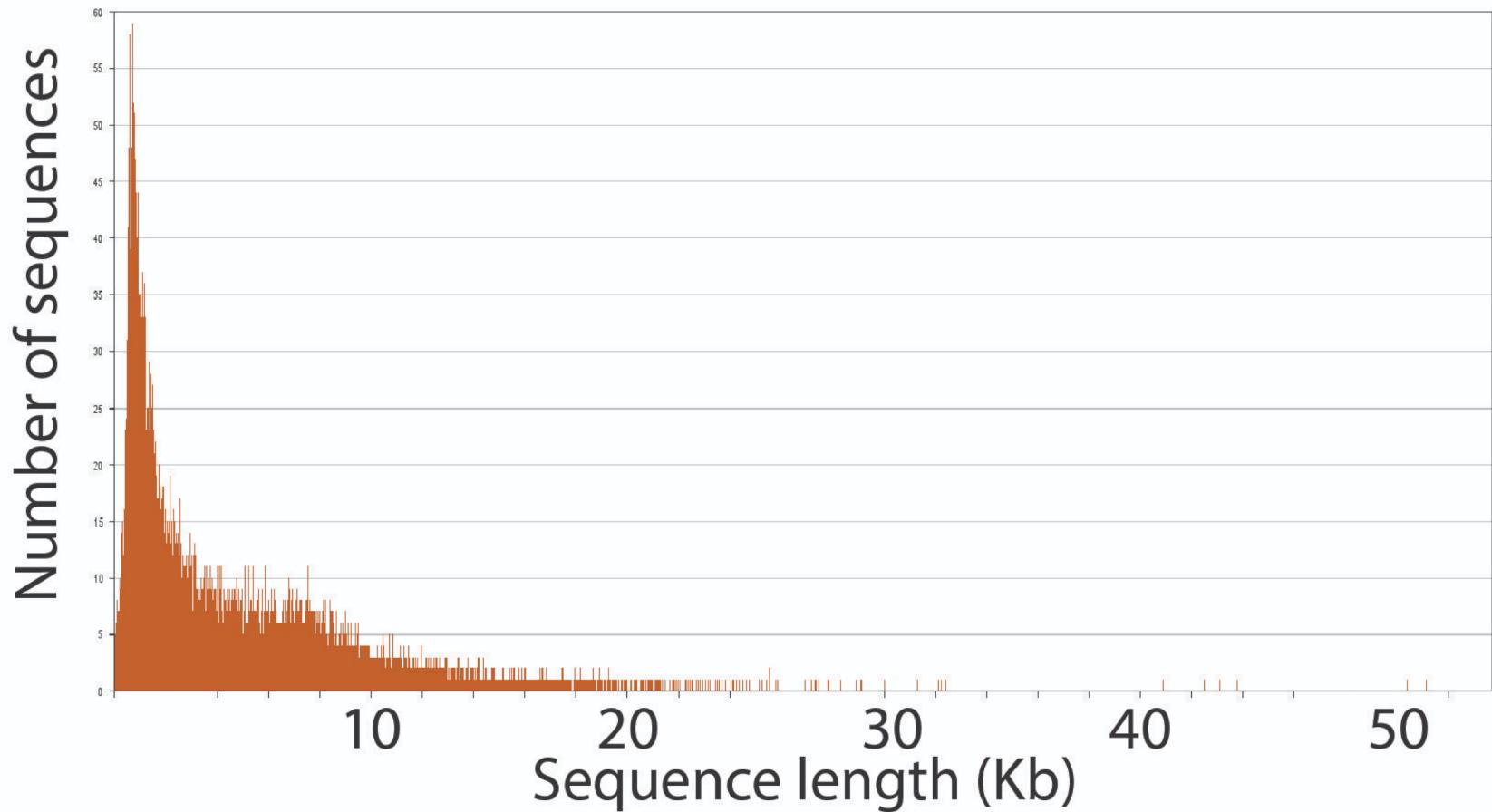


ENCODE Project Consortium. PLoS Biol. 2011 Apr;9(4):e1001046. doi: 10.1371/journal.pbio.1001046.

LRWGS

- Short read
 - 100 bases – 500 bases
 - SNV
 - Pequenos Indels
 - CNV
- Long Read
 - 1 Kb – 50⁺ kb
 - SV
 - Translocações
 - CNVs
 - Grandes Indels

Distribuição Típica de Fragmentos



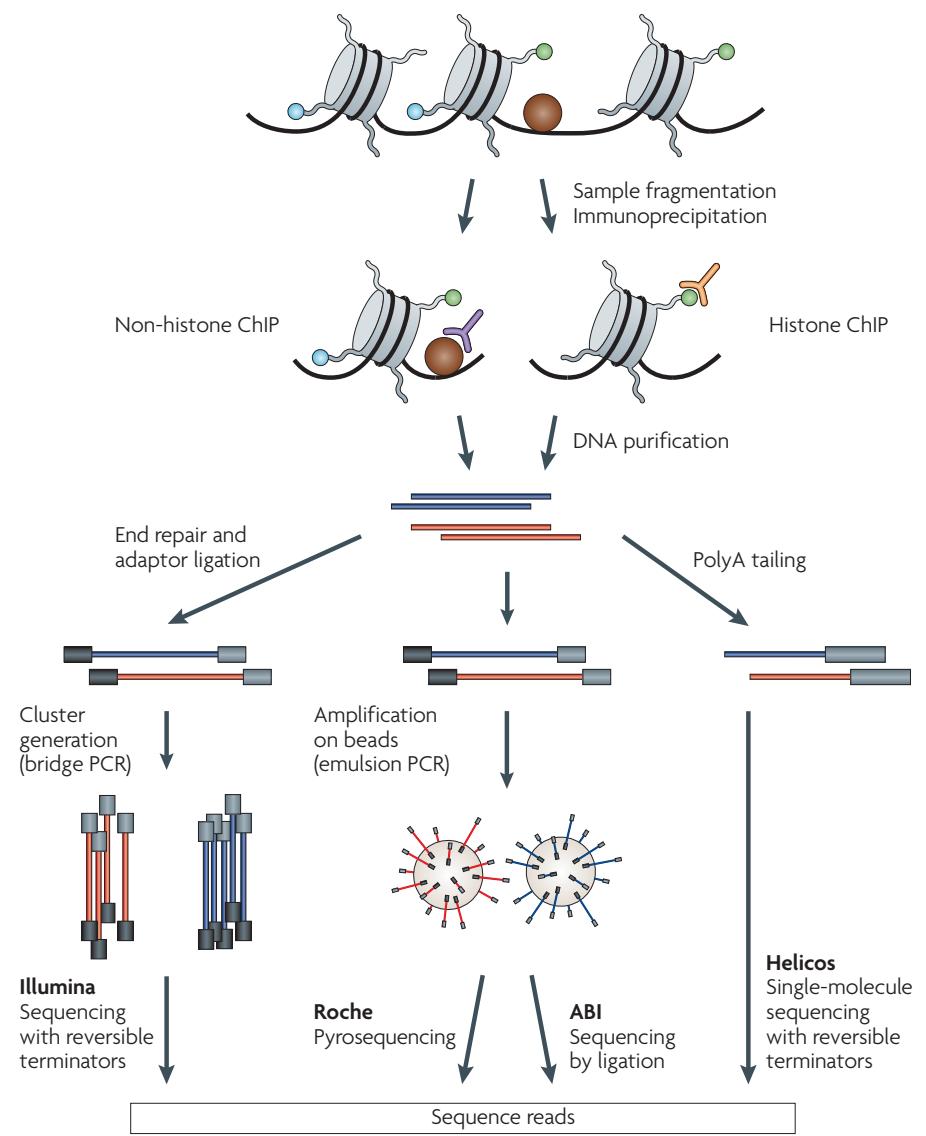
Equipamentos Específicos

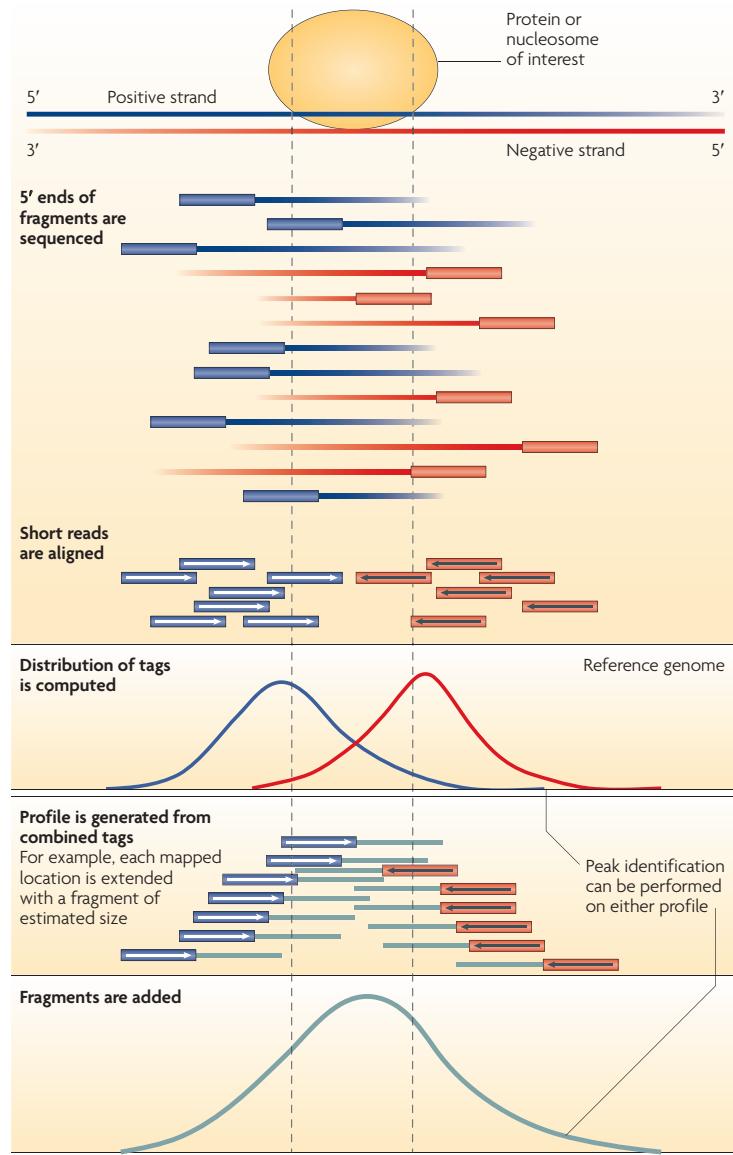
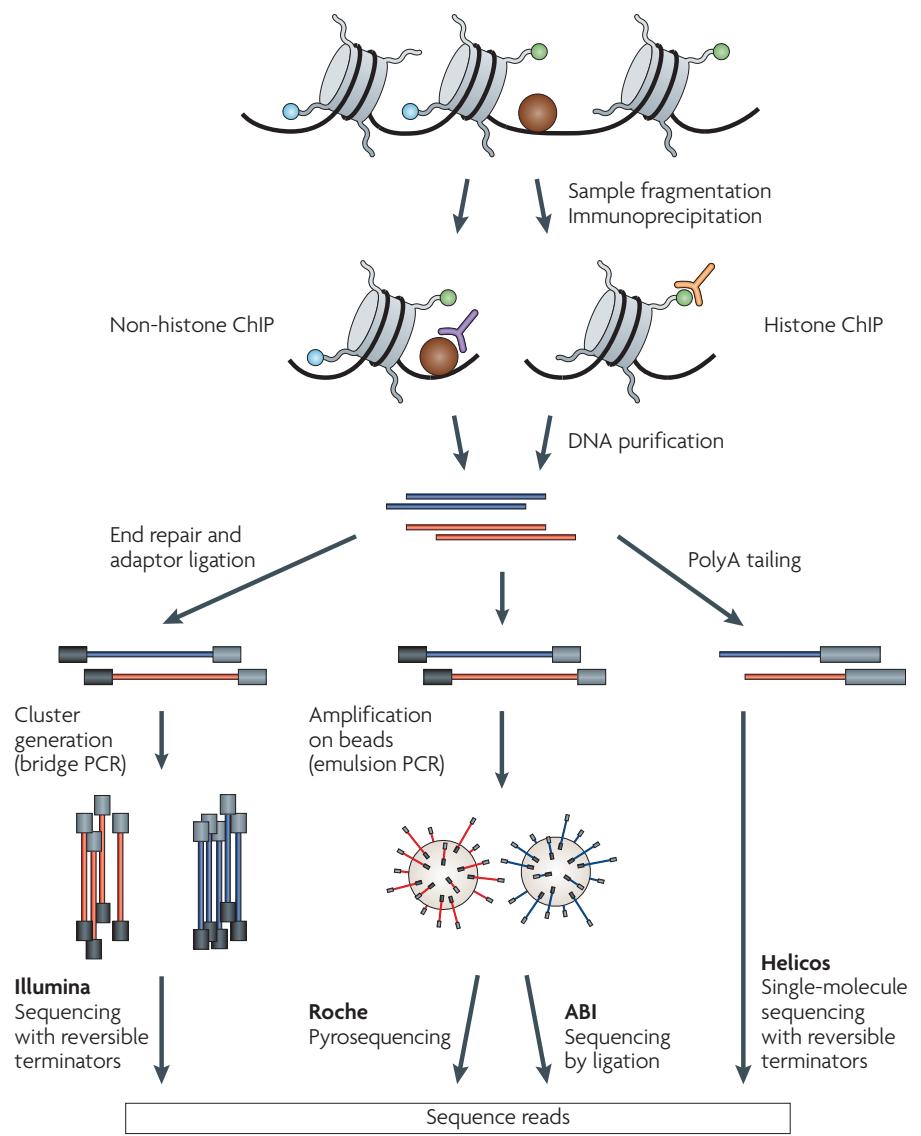


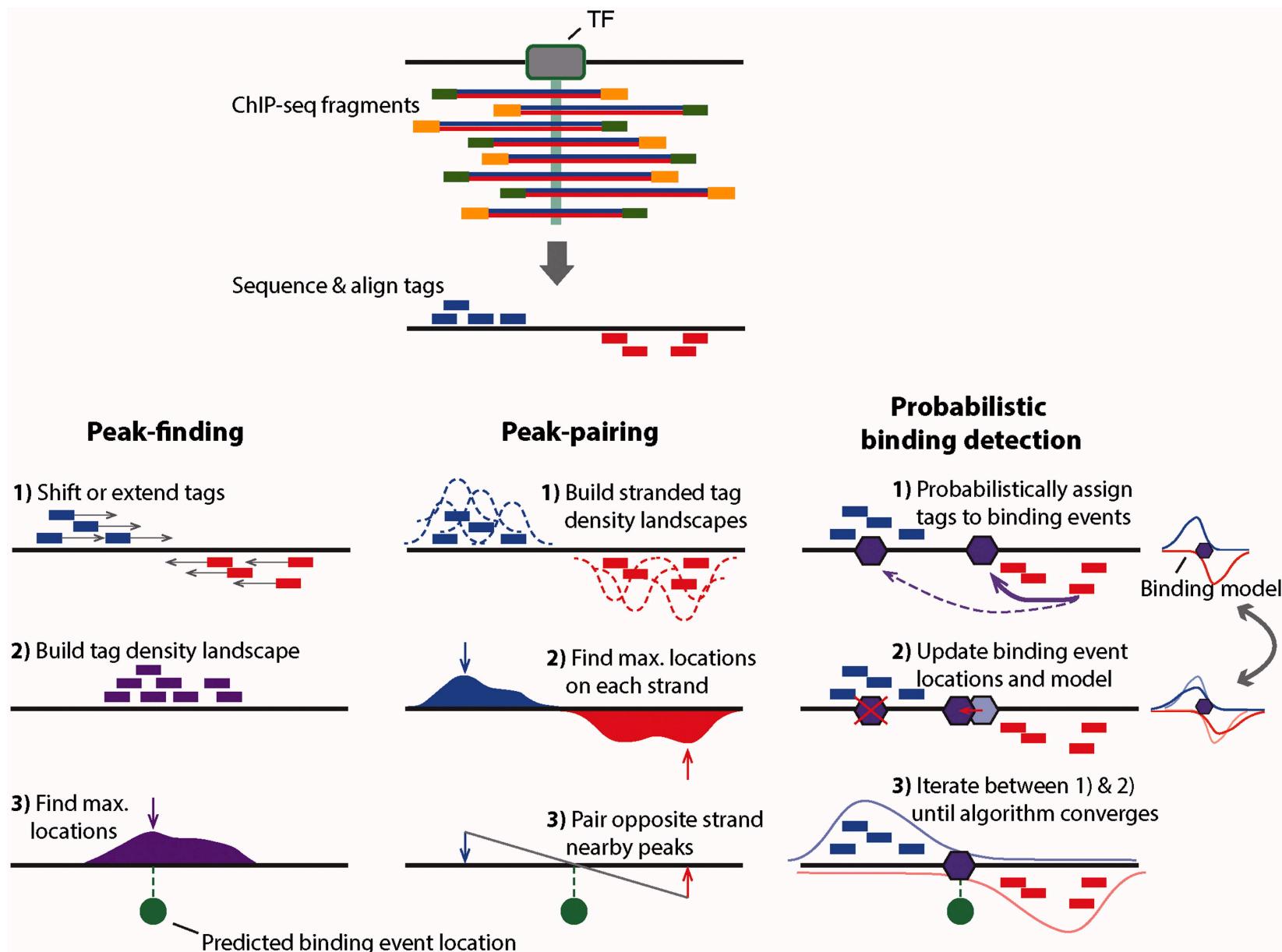
ChIPseq

- Chromatin Imunnoprecipitation
- Enriquecimento do DNA para localização de regiões de ligação em uma preteína de interesse

<https://galaxyproject.org/tutorials/chip/>

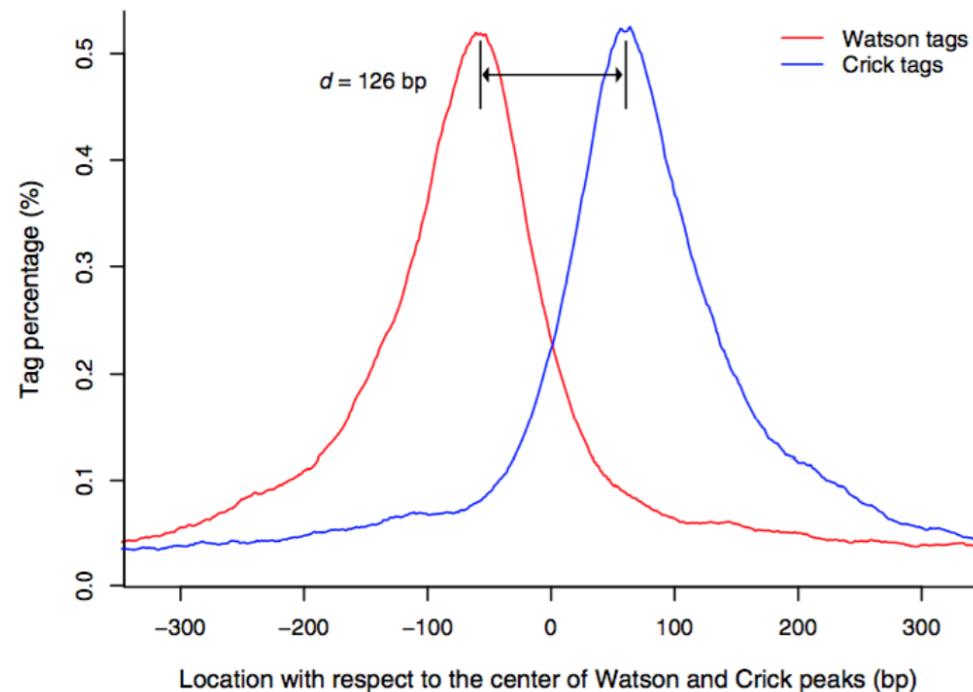






1. enriched with respect to total input (calculated in windows)
2. biggest densities of + and - strand alignments are offset (shift size)
3. enriched with respect to background of ChIP reads

Model-based Analysis of ChIP-Seq (MACS) Zhang et al (Liu lab)



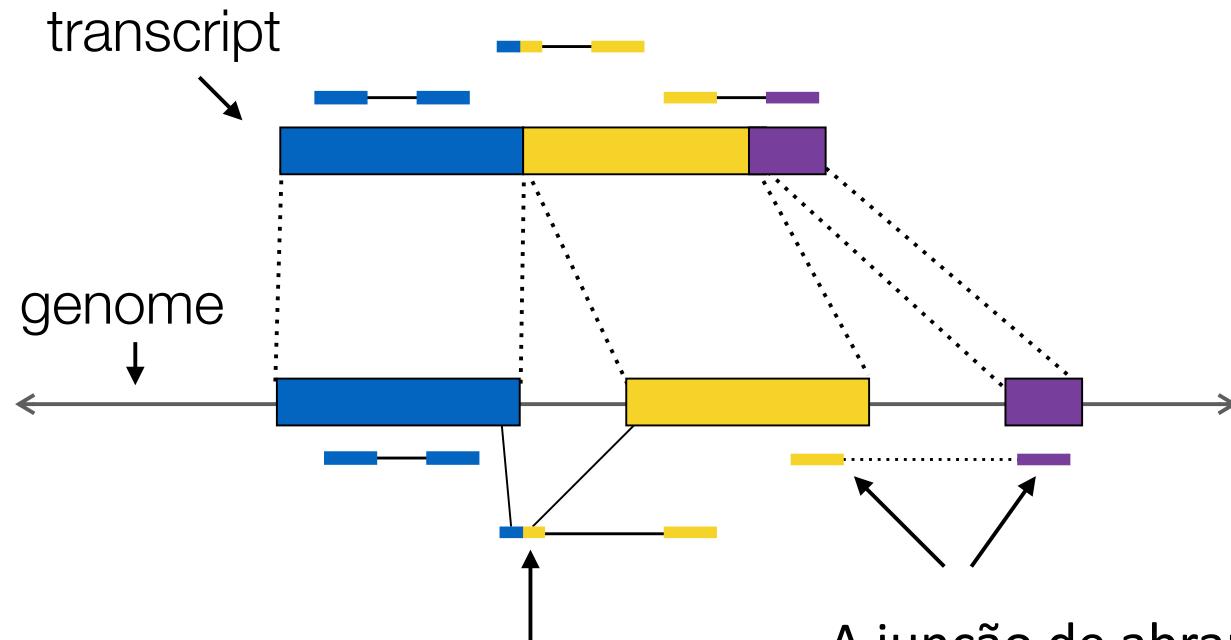
Walkthrough

- Mapeamento dos reads com Bowtie
 - FASTQ -> SAM (BAM ou CRAM)
- Ordenamento do BAM com samtools
 - BAM -> SortedBAM
- deepTools para verificação da cobertura
 - SortedBAM -> bigWig
- Chamada dos picos (Peak calling) usando o MARCS2
 - SortedBAM -> narrowPeak
- Visualização dos Picos no IGV
- Anotação de bases de dados para interpretação
 - Fenótipos associados
 - Vias de sinalização impactadas

RNAseq

- Sequenciamento de transcritos expressos
- Normalmente é usado poly(A+) ou protocolo RiboMinus para depleção de rRNA
- Detecção de expressão gênica
- Detecção de novos transcritos

Problemática do Mapeamento



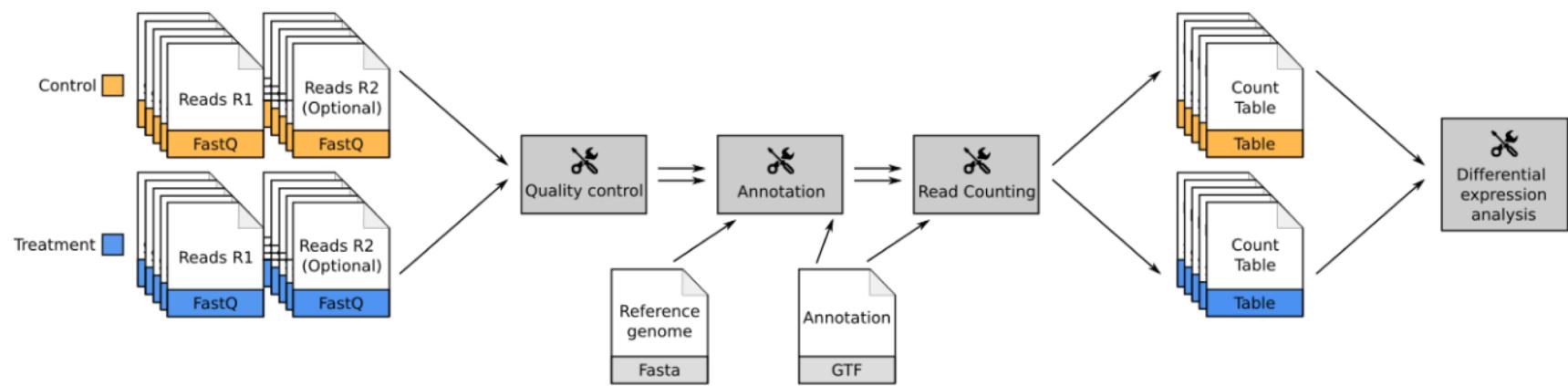
Esse read não está mapeado de forma contínua no genoma de referência

A junção de abrangência dos reads paired ends mapeiam muito longe no genoma de referência

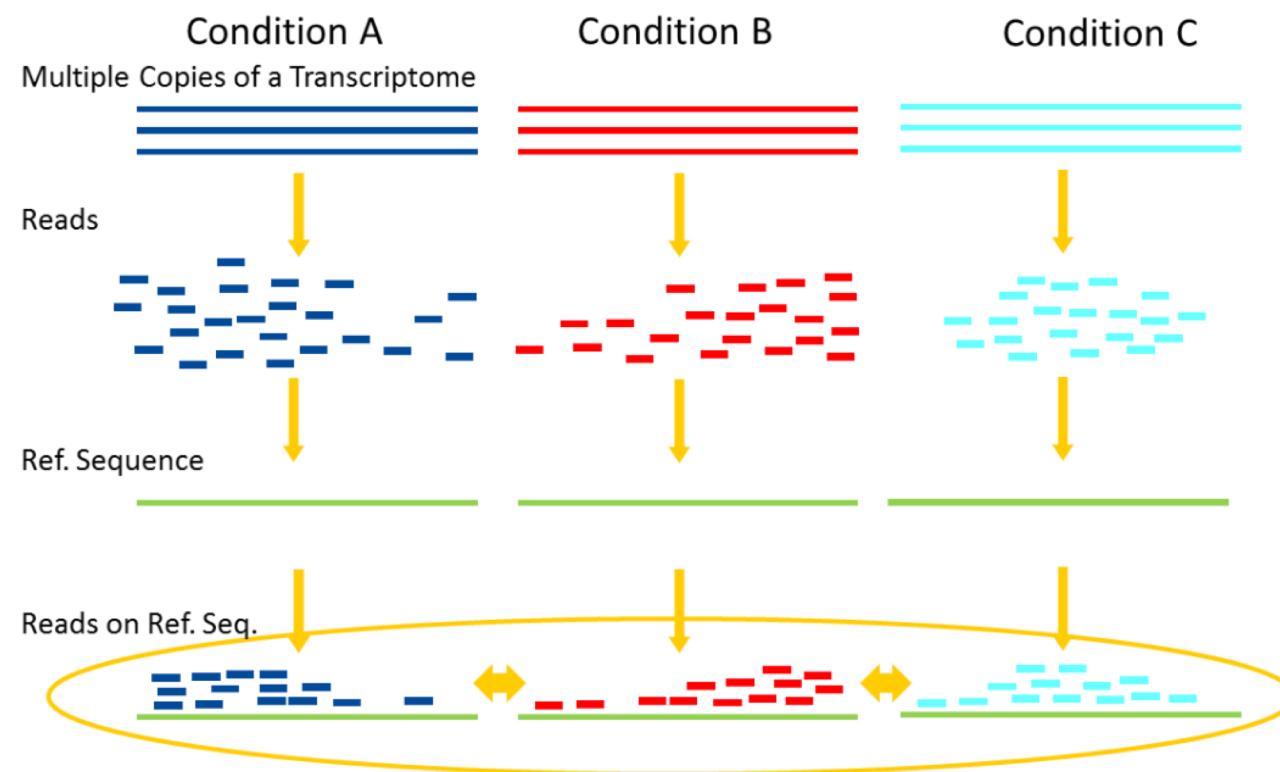
Mapeamento

- Precisa ser feito por um algoritmo de mapeamento diferente do utilizado no DNA-seq
- É necessário usar como referência transcritos conhecidos para poder melhorar a qualidade do alinhamento
 - RefSeq
 - Ensembl
 - KnownGene

Pipeline de Análise



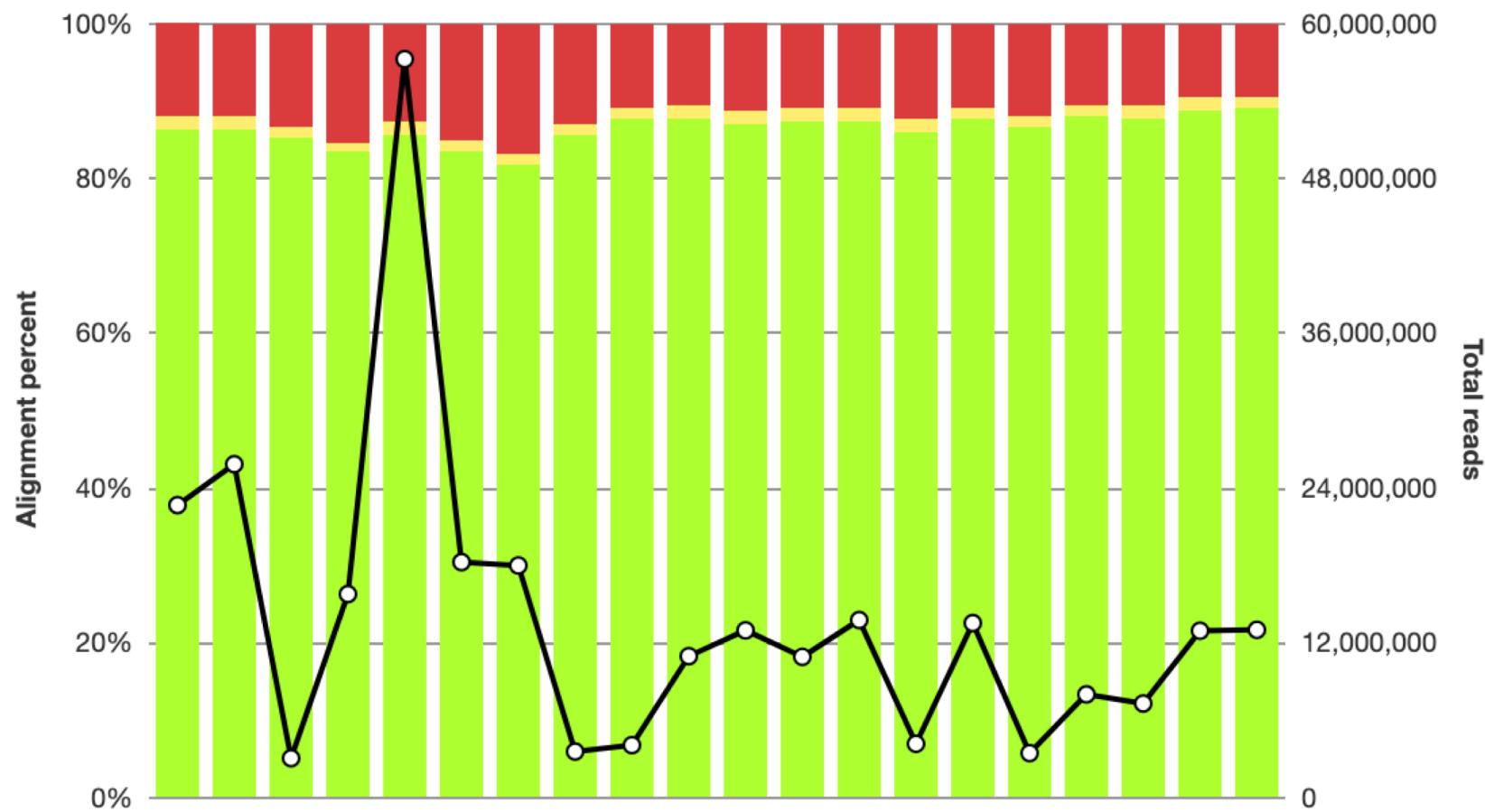
Overview



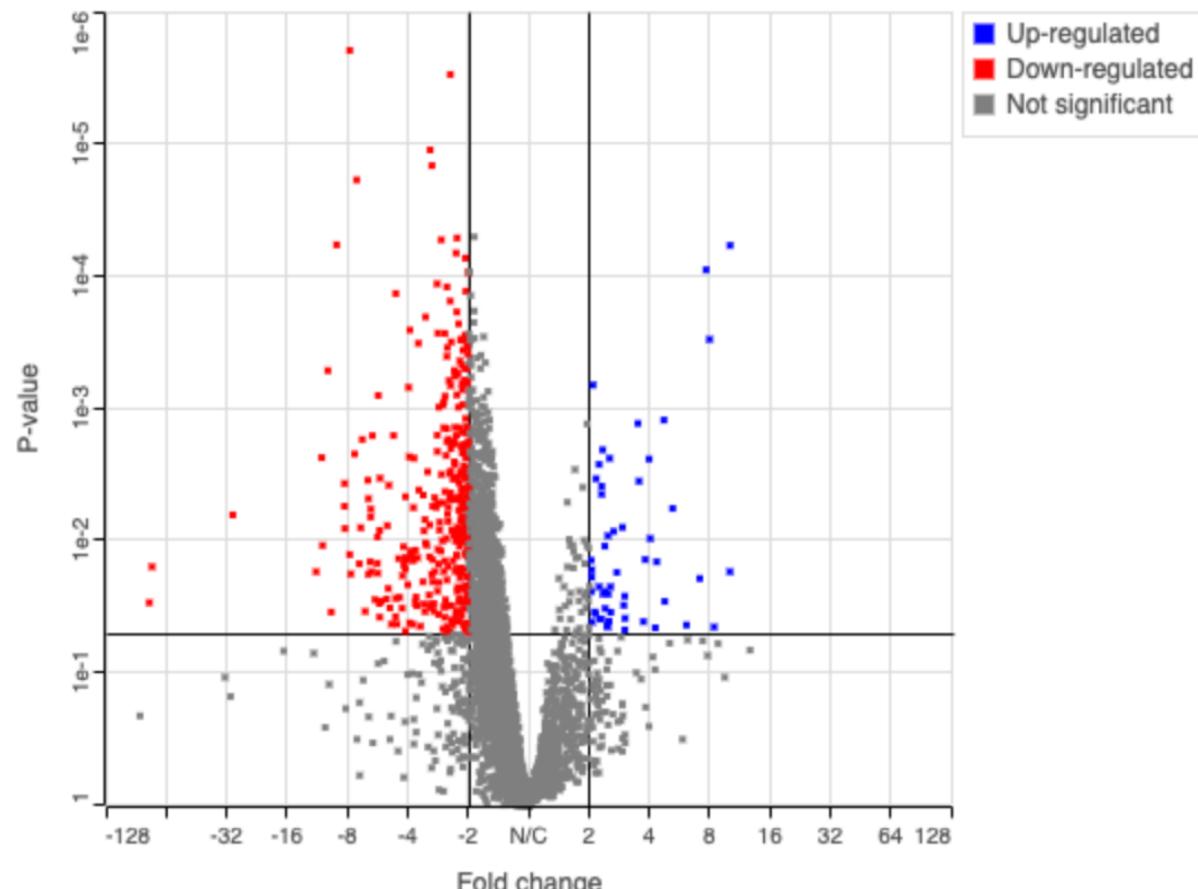
Walkthrough

- Mapeamento dos reads
 - Bowtie
 - STAR
 - FASTQ -> BAM
- Quantificação dos reads
 - Cufflinks
 - Partek
 - BAM -> Gene counts || Transcript counts (inclusive novos transcritos)
- Normalização
- Análise estatística
 - Análise de expressão diferencial
 - GSA (partek)
 - ANOVA
- PCA

QA/QC



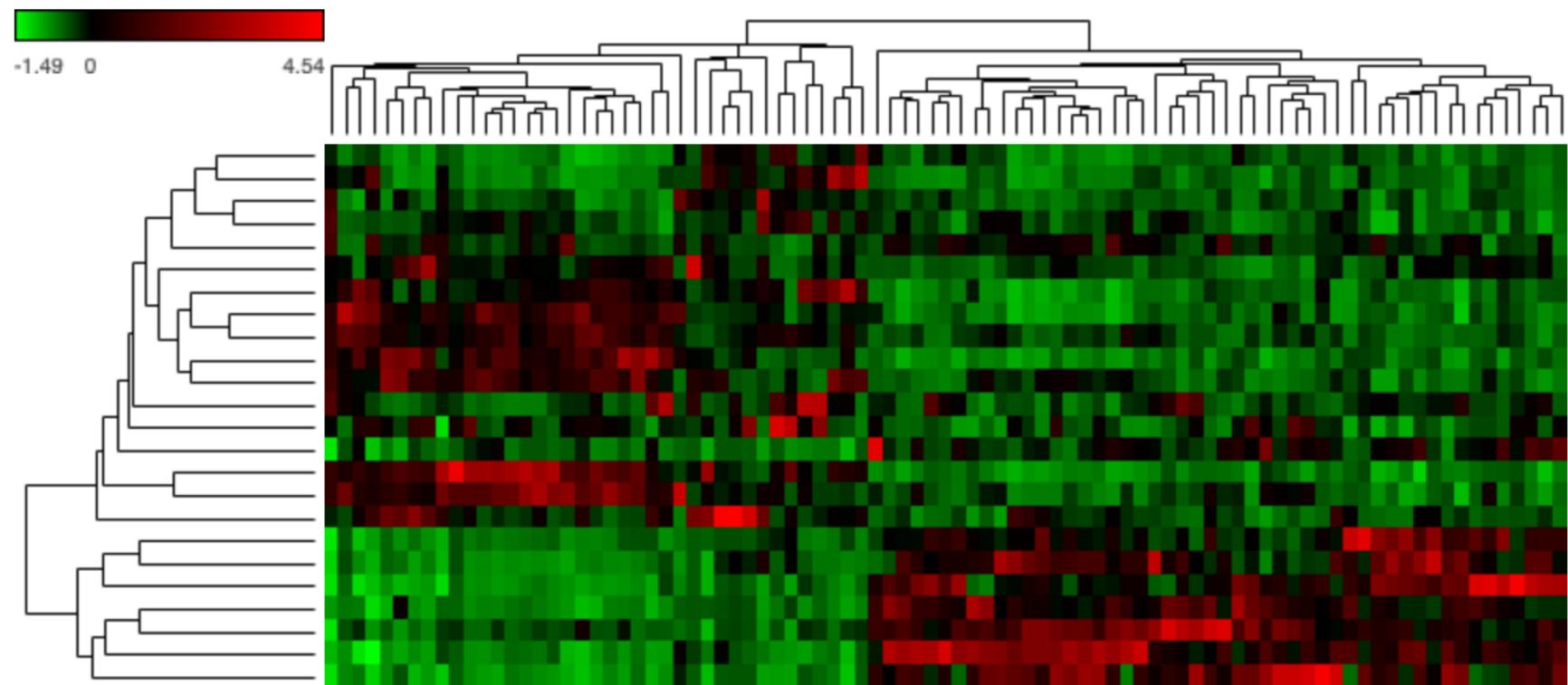
Volcano Plot



Up-regulated: 53

Down-regulated: 341

Clusterização Hierárquica



RNA-Seq + DNA-Seq

SCIENCE TRANSLATIONAL MEDICINE | RESEARCH ARTICLE

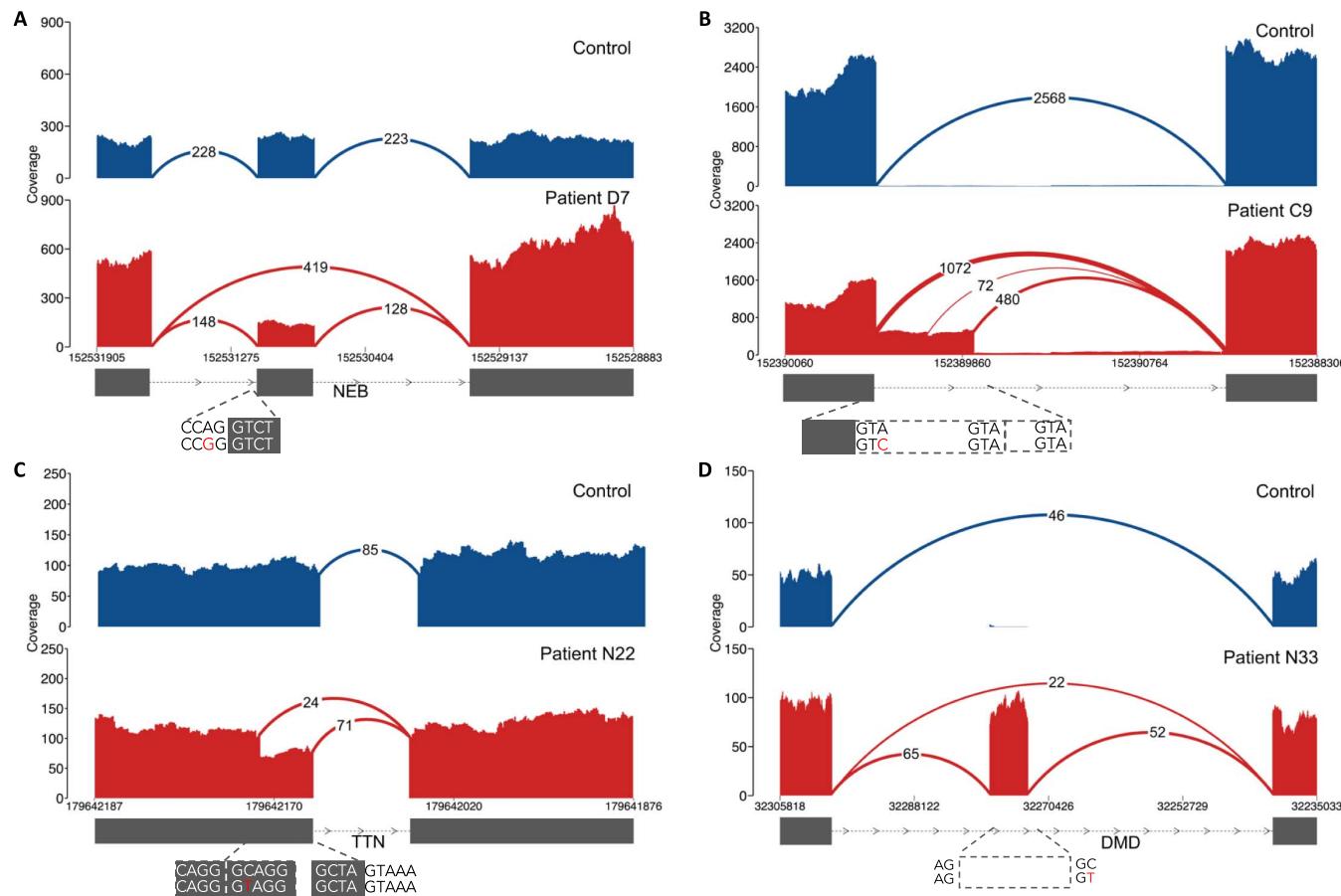
GENETIC DIAGNOSIS

Improving genetic diagnosis in Mendelian disease with transcriptome sequencing

Beryl B. Cummings,^{1,2,3} Jamie L. Marshall,^{1,2} Taru Tukiainen,^{1,2} Monkol Lek,^{1,2,4,5}
Sandra Donkervoort,⁶ A. Reghan Foley,⁶ Veronique Bolduc,⁶ Leigh B. Waddell,^{4,5}
Sarah A. Sandaradura,^{4,5} Gina L. O'Grady,^{4,5} Elicia Estrella,⁷ Hemakumar M. Reddy,⁸ Fengmei Zhao,^{1,2}
Ben Weisburd,^{1,2} Konrad J. Karczewski,^{1,2} Anne H. O'Donnell-Luria,^{1,2} Daniel Birnbaum,^{1,2}
Anna Sarkozy,⁹ Ying Hu,⁶ Hernan Gonorazky,¹⁰ Kristl Claeys,¹¹ Himanshu Joshi,⁵ Adam Bournazos,^{4,5}
Emily C. Oates,^{4,5} Roula Ghaoui,^{4,5} Mark R. Davis,¹² Nigel G. Laing,^{12,13} Ana Topf,¹⁴
Genotype-Tissue Expression Consortium, Peter B. Kang,^{7,8} Alan H. Beggs,⁷ Kathryn N. North,¹⁵
Volker Straub,¹⁴ James J. Dowling,¹⁰ Francesco Muntoni,⁹ Nigel F. Clarke,^{4,5*} Sandra T. Cooper,^{4,5}
Carsten G. Bönnemann,⁶ Daniel G. MacArthur^{1,2†}

2017 © The Authors,
some rights reserved;
exclusive licensee
American Association
for the Advancement
of Science.

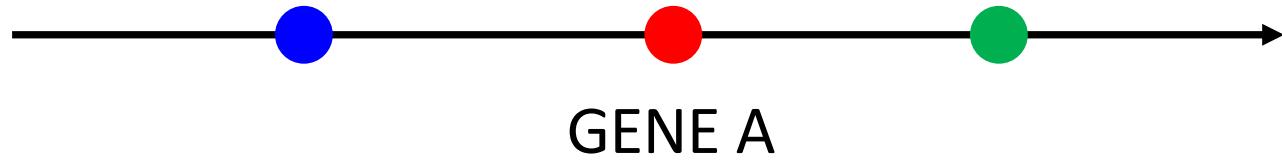
Aberrações de Splicing



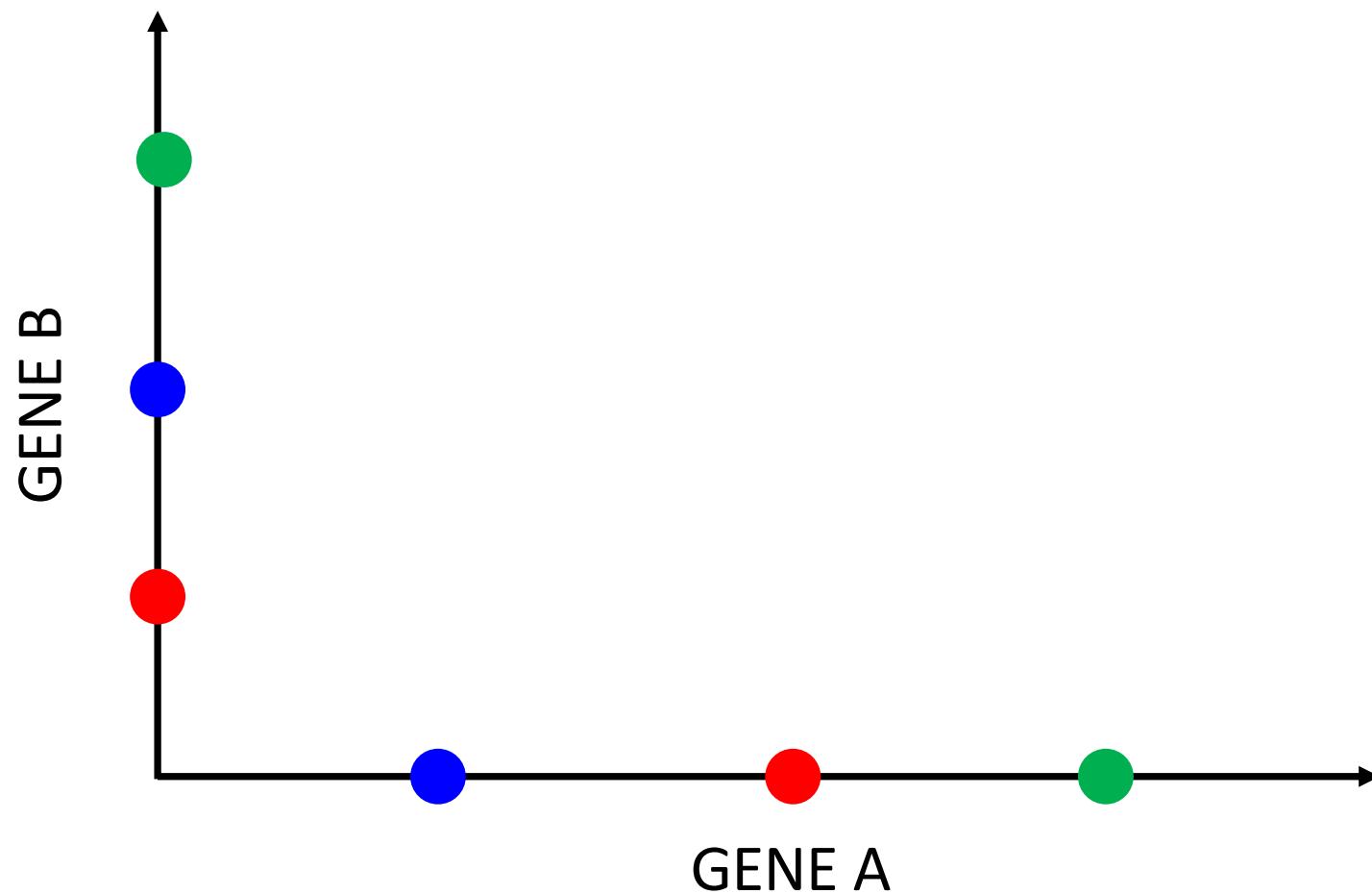
PCA

- Dados de genômica são muito grandes
- Normalmente dezenas de colunas por milhares ou até milhões de linhas
- O que é conhecido como a maldição da dimensionalidade
- Análise estatística mais difícil
- Visualização impossível

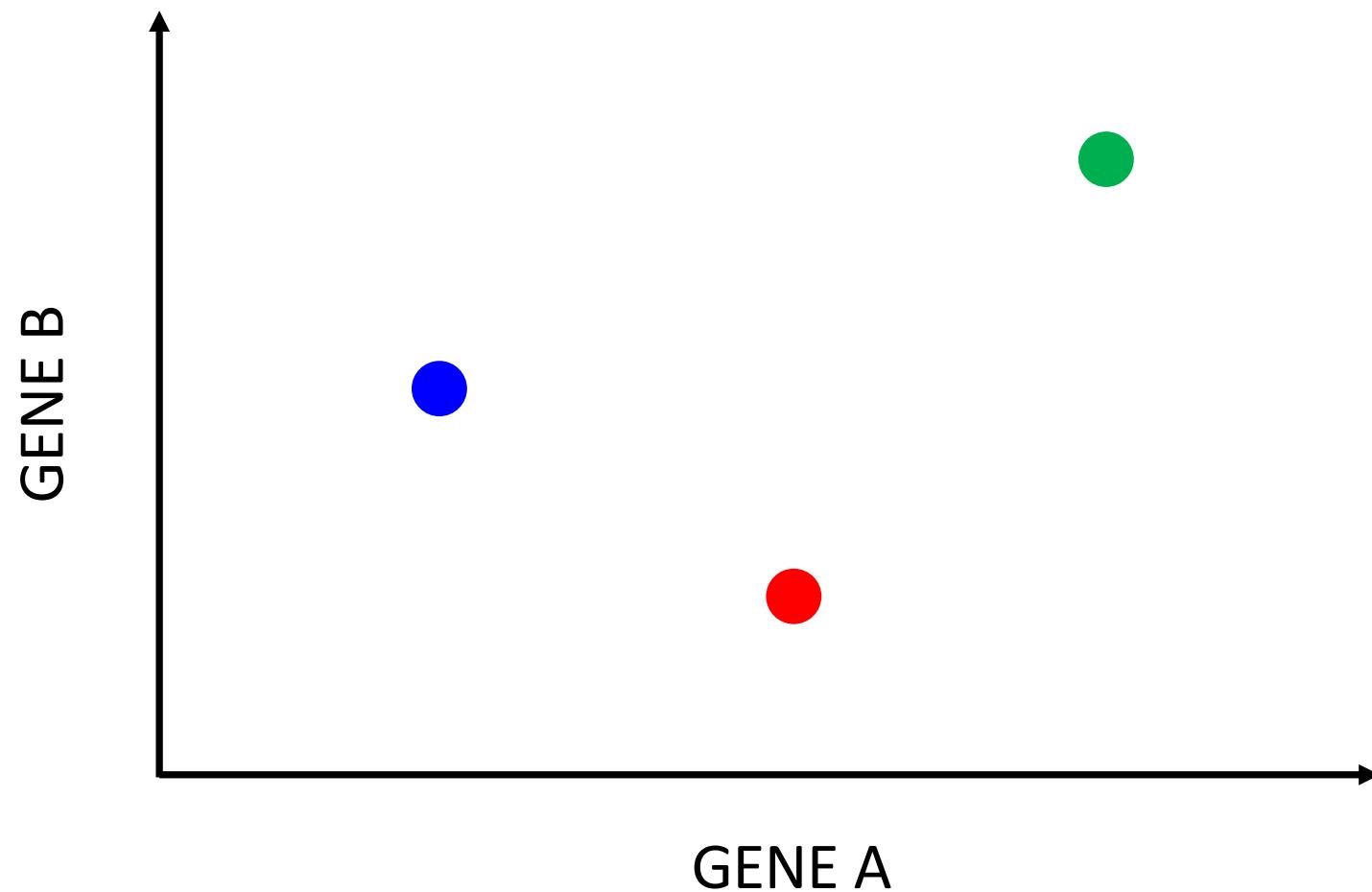
Expressão Gênica



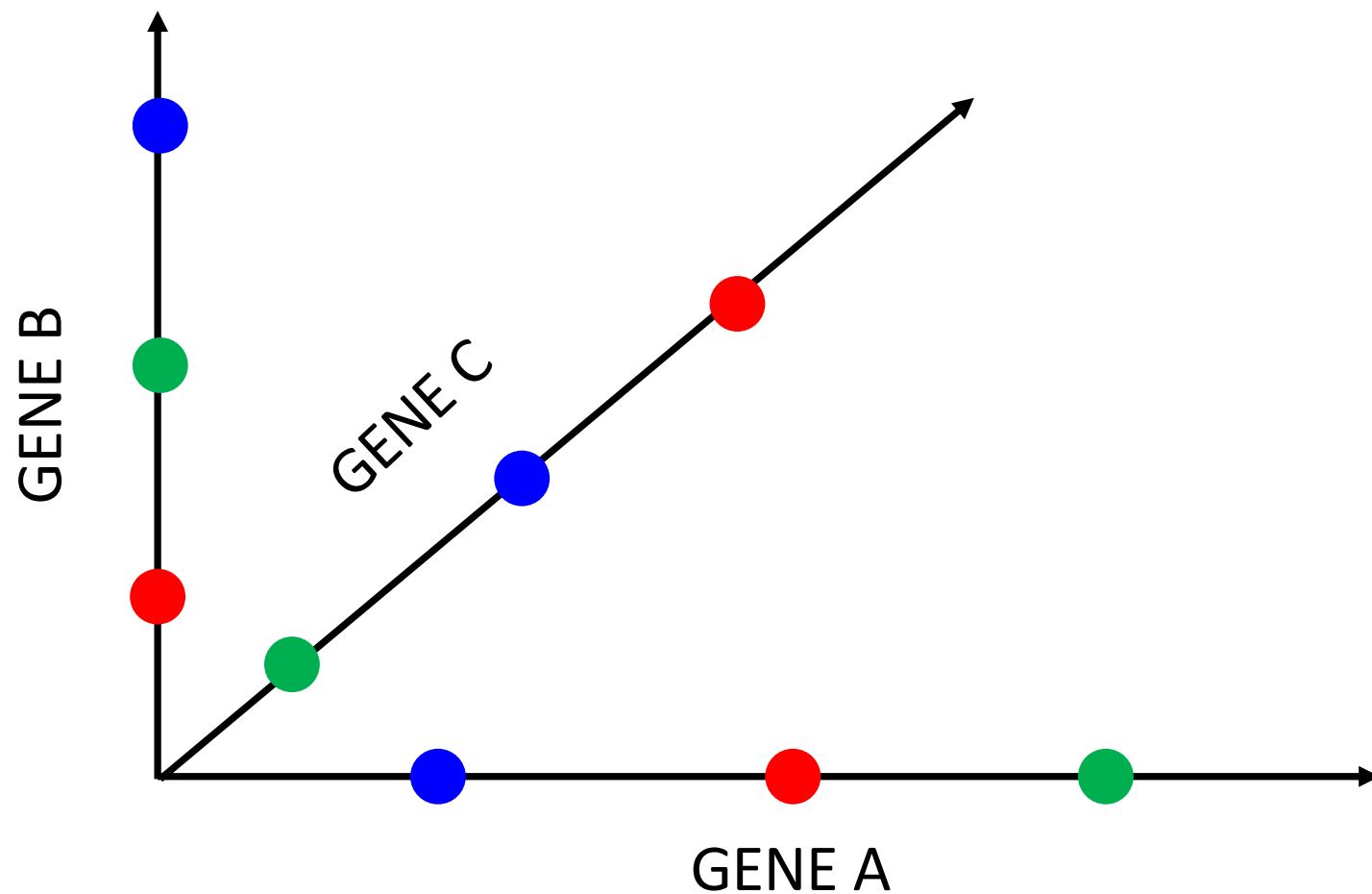
Expressão Gênica



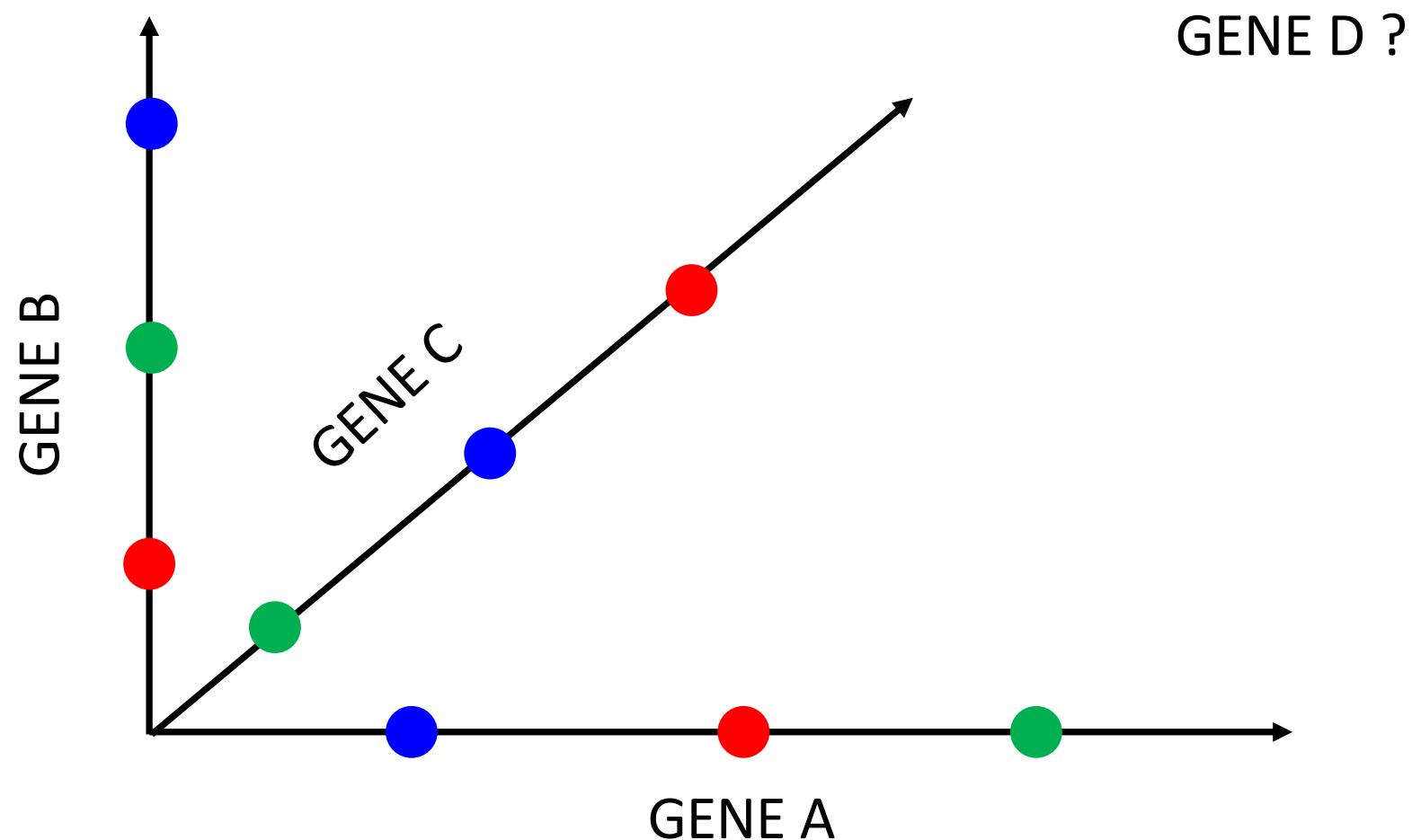
Expressão Gênica



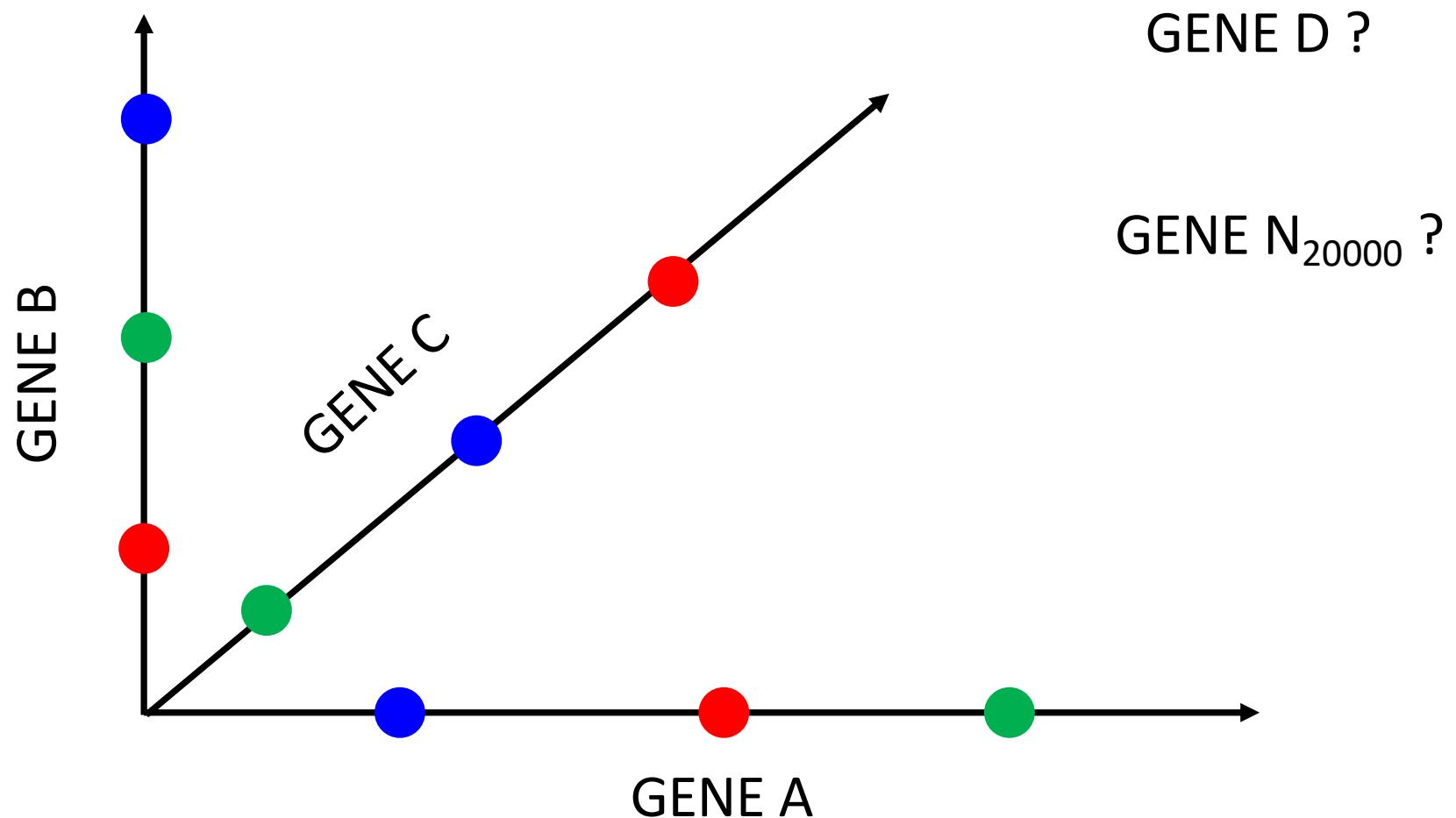
Expressão Gênica



Expressão Gênica



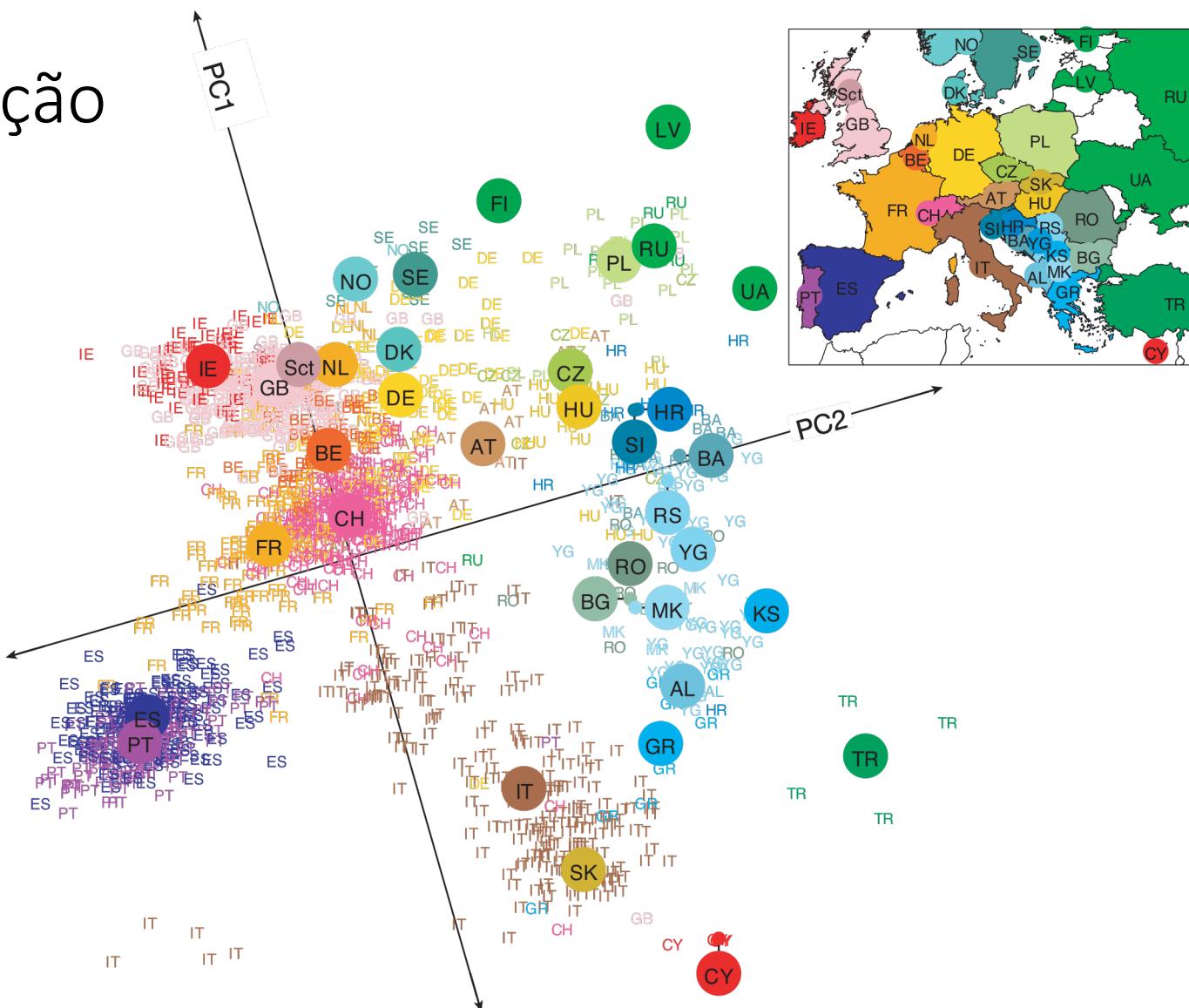
Expressão Gênica



PCA

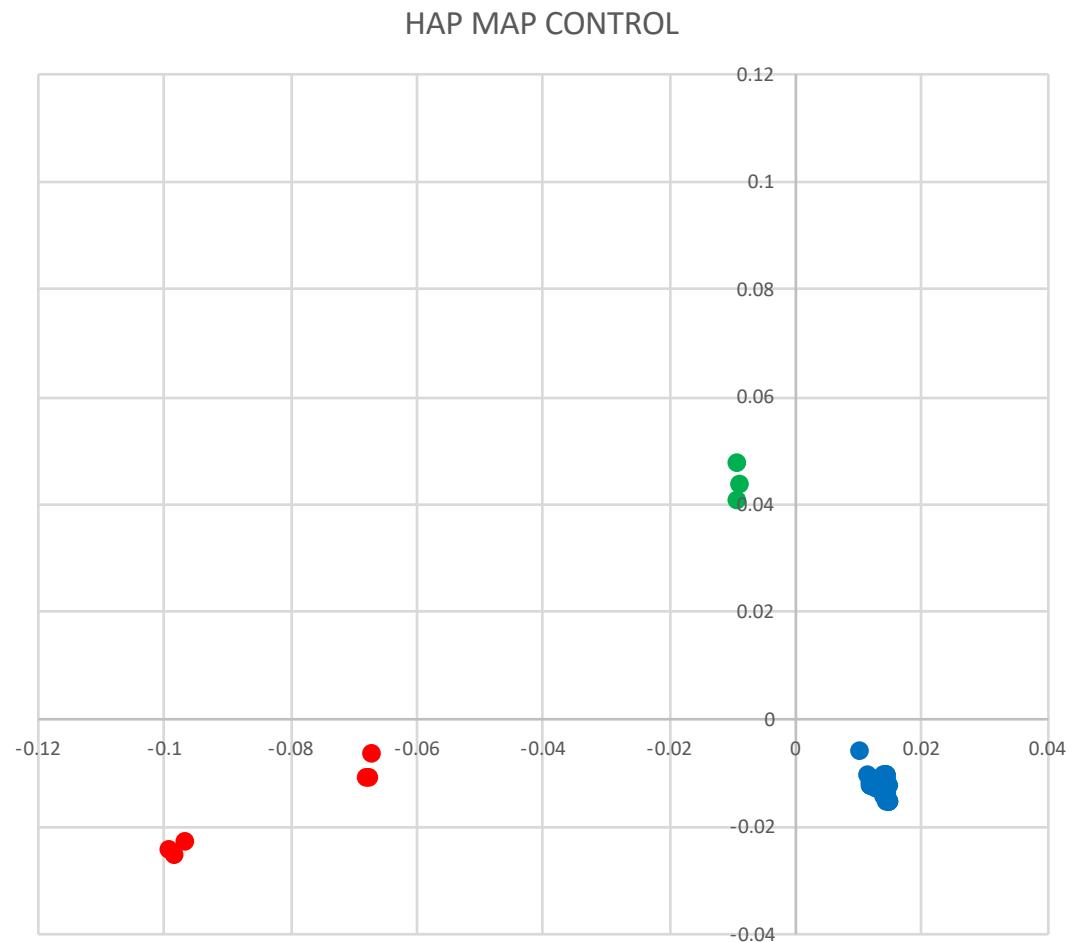
- Redução de dimensões
- Transformar um conjunto de variáveis em um novo conjunto de variáveis não correlacionadas entre si
- Transformação linear orthogonal
- Ordenação dessas variáveis
 - Da variável que explica o maior número de variáveis originais
 - PC1
 - Para variável que explica o menor número de variáveis originais
 - PCN

Aplicação

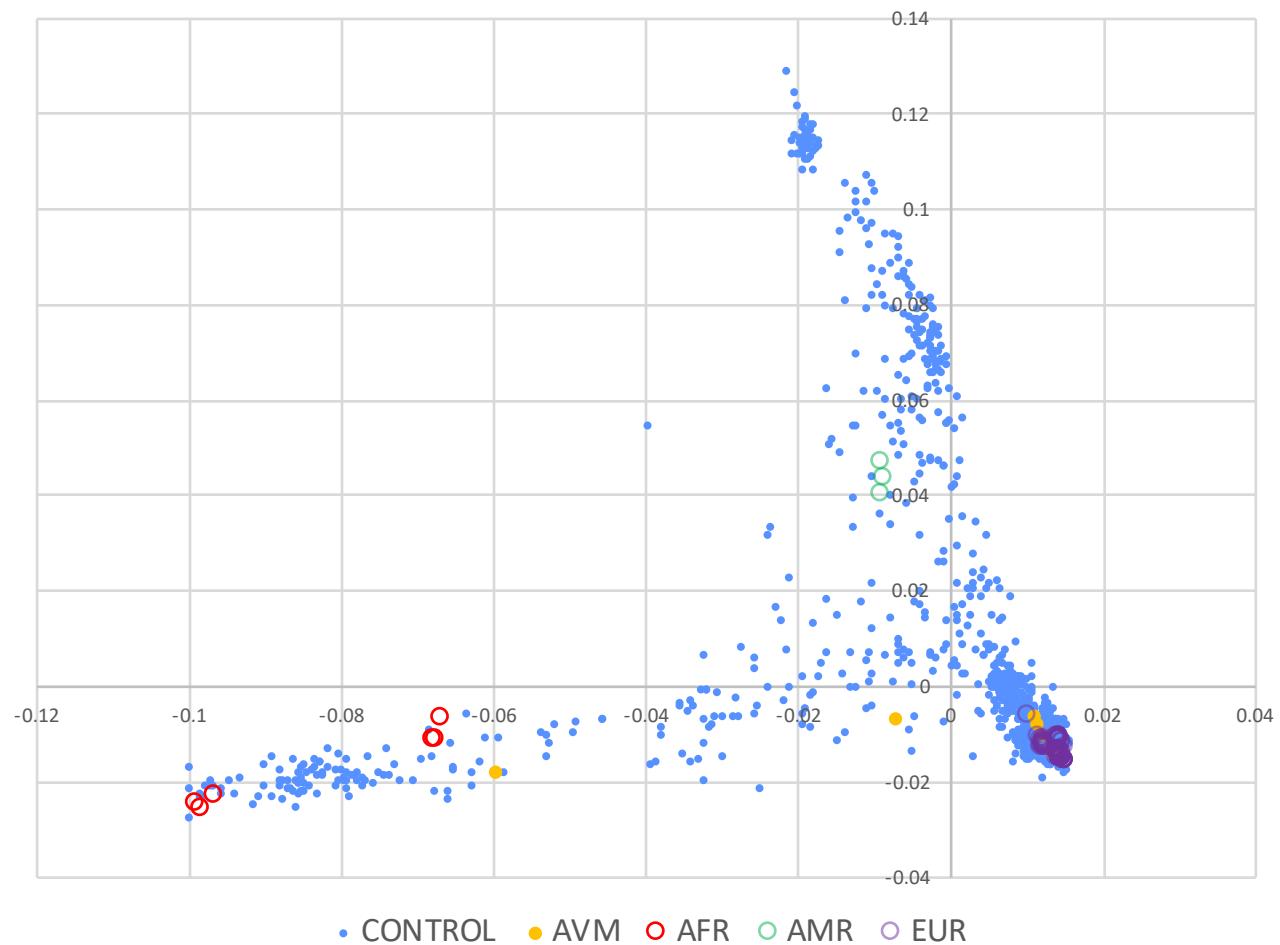


Classificação de População com PCA

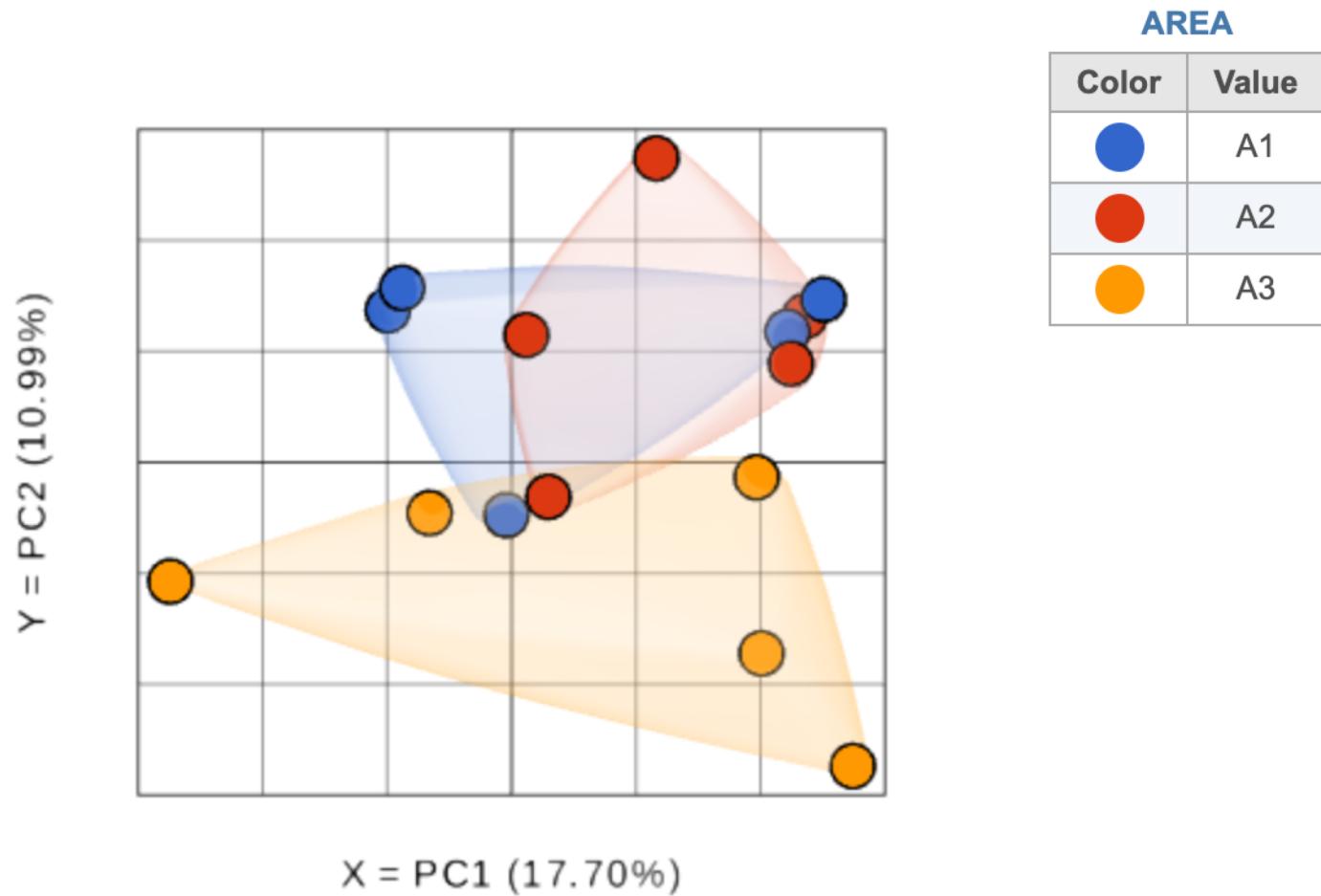
- 94 samples:
 - 5 populations;
 - 3 Super populations:
- **85 - EUR**
 - 83 - CEU
 - From 8 different subjects
 - 2 - TSI
- **3 - AMR**
 - 3 - MXL
- **6 - AFR**
 - 3 - YRI
 - 3 - LWK



1320 Amostras WES



Aplicação PCA RNA-seq



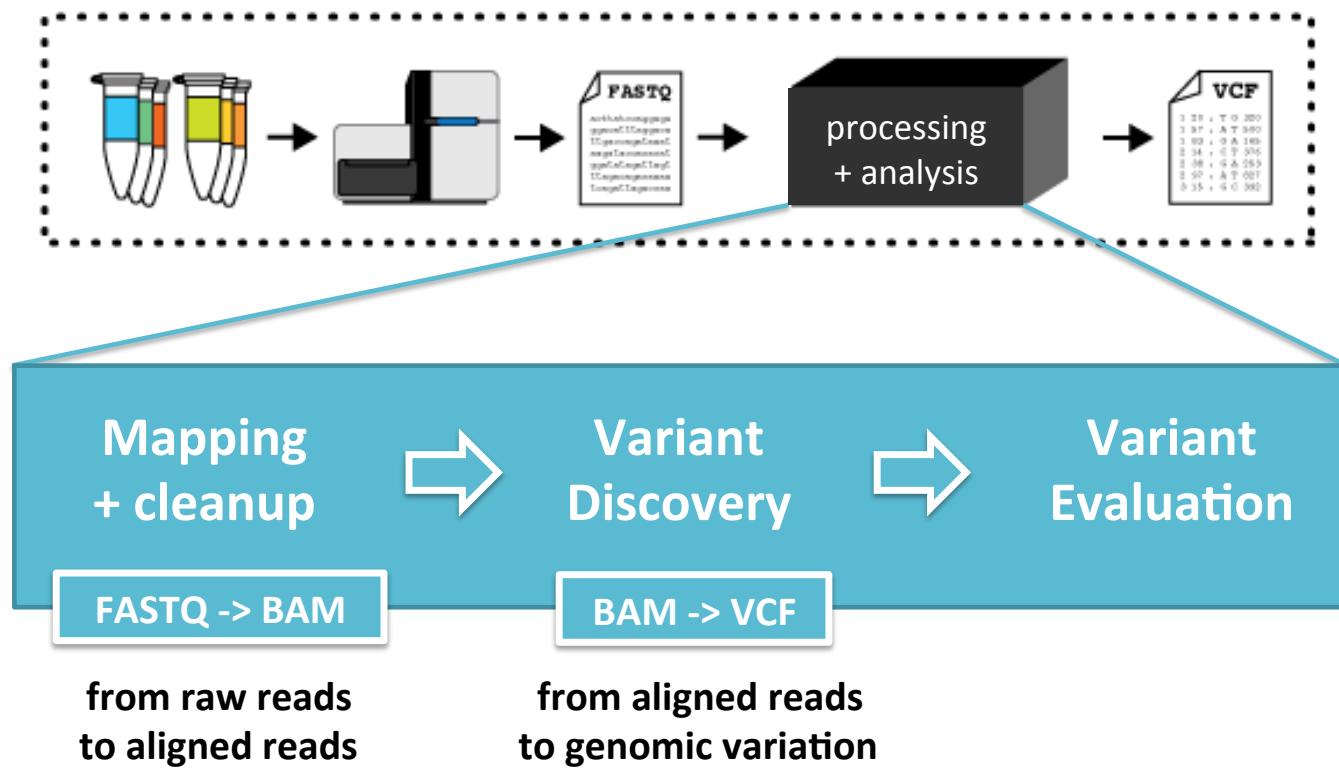
Walkthrough

- Obter um conjunto de SNVs (Painel, WES, WGS, SNP array)
- Obter um conjunto de SNVs de população conhecida
 - Hapmap control
 - 1000g
- Selecionar SNVs com alta frequencia alelica
 - AF > 0,01
- Utilizar o software PLINK para gerar os PCs (PCA)
- Plotar os dados
 - Excel
 - R

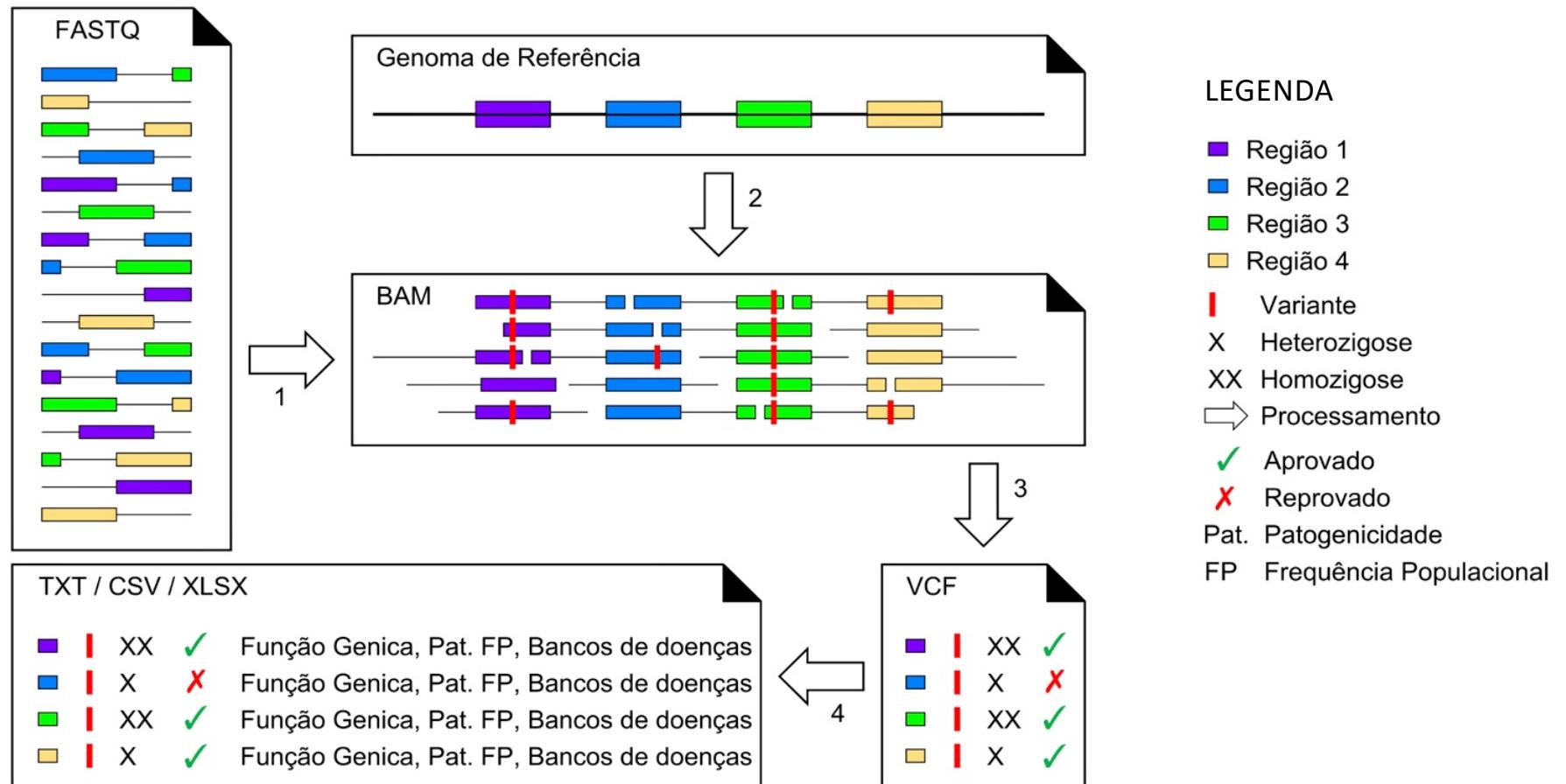
PhenoDB

- Cadastro no site do phenoDB para a próxima aula

<https://phenodb.org/>



Pipeline DNASeq



Pipeline com Excel

- Mapeamento BWA
- samtools sort + index
- Picard mark duplicates + index
- GATK base recalibrator + index
- GATK print reads + index
- GATK Haplotype caller
- Conversão VCF para ANNOVAR
- Anotação do VCF com ANNOVAR