Conceitos fundamentais sobre a metodologia para análise de RNA-Seq

Pablo Rodrigo Sanches

Departamento de Genética – FMRP/USP

psanches@usp.br

Pablo Rodrigo Sanches

- É **Analista de Sistemas** e **Bioinformata** desde 2003 na Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo (FMRP-USP), onde desenvolve soluções em tecnologia para projetos de pesquisa, entre outros.
- É **Professor** desde 2011 em cursos de graduação na Universidade de Ribeirão Preto (UNAERP), em disciplinas correlatas a tecnologia de informação e engenharia de software.

- TITULAÇÃO: Doutor em Genética pela FMRP-USP, Mestre em Física Computacional pelo Instituto de Física de São Carlos da USP (IFSC-USP), MBA em Tecnologia da Informação e Gestão de Negócios pela Fundação Getulio Vargas (FGV) e Bacharel em Análise de Sistemas pela UNAERP.

O que é isso para você? 🤔

```
>Sequence_1 assembly1
CCCTAAACCCTAAACCCTAAACCCTAAACCTCTGAATCCTTAATCCCTAAATCCCTAAAT
CTTTAAATCCTACATCCATGAATCCCTAAATACCTAATTCCCTAAACCCGAAACCGGTTT
CTCTGGTTGAAAATCATTGTGTATATAATGATAATTTTATCGTTTTTATGTAATTGCTTA
TTGTTGTGTGTAGATTTTTTAAAAATATCATTTGAGGTCAATACAAATCCTATTTCTTGT
GGTTTTCTTTCCTTCACTTAGCTATGGATGGTTTATCTTCATTTGTTATATTGGATACAA
GCTTTGCTACGATCTACATTTGGGAATGTGAGTCTCTTATTGTAACCTTAGGGTTGGTTT
ATCTCAAGAATCTTATTAATTGTTTGGACTGTTTATGTTTGGACATTTATTGTCATTCTT
```

Uma sequência de caracteres?
Um arquivo texto?

vs.

Um Gene?
A parte de um genoma?

E isso?

```
while(my $seq = $seqio->next_seq) {
  @idseqs =  `cat $ARGV[0]`;
  foreach $idseq (@idseqs)
  {
    chomp $idseq;
    if($seq->desc =~ /$idseq/)
    {
      if($seq->desc =~ m/$ARGV[3]/) {
        $pos = $-[0];
      }
      $desc = substr($seq->desc,$pos);
      print INFO ">" . $idseq . " " . $desc . "\n";
      print INFO $seq->seq . "\n";
    }
  }
}
```

Um texto em uma língua desconhecida?
Palavras sem sentido organizadas?

vs.

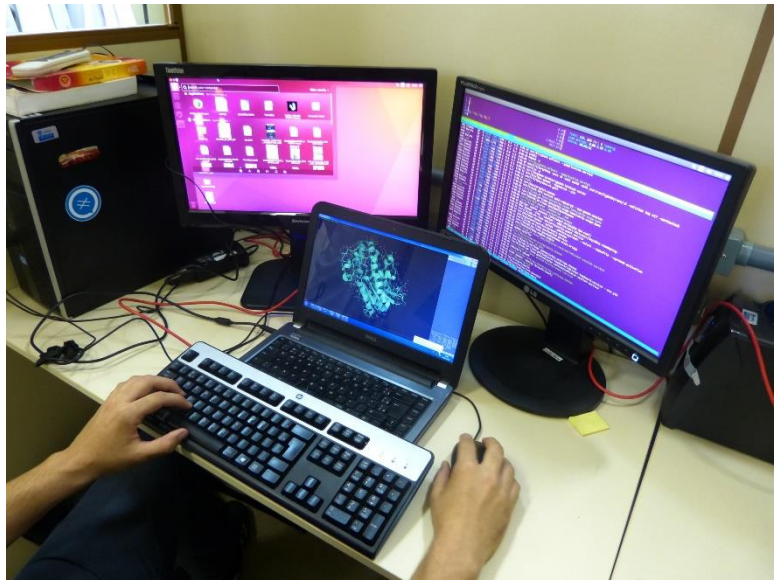
Um algoritmo?
Um código-fonte escrito em linguagem Perl?

Mercado de trabalho

Piloto de drones, Engenheiro de robôs, Youtuber, Streamer gamer, Cyber atleta, Influenciador digital, Cientista de dados, Técnico de saúde assistida por Inteligência Artificial, Walker/Talker, ...

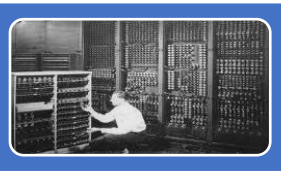
E ainda...

- Bioinformata → Bioinformática
  - Etimologia
    - Bio = "*bios*" (vida) + Informática = "*informatik*" (informação automática)
  - Não é exatamente nova, porém pouco conhecida... Ainda 😮



Existe uma relação entre Biologia e Informática?

Histórico

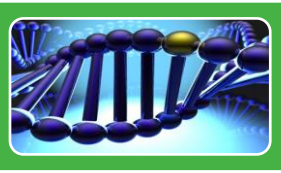


A história começa na década de 1940 com a invenção do moderno computador digital

Ele se chama digital, pois os dados são armazenados com um alfabeto binário (0 e 1)

A descoberta da dupla hélice, em 1953, por Watson e Crick, mostrou que a informação genética também é armazenada de forma digital

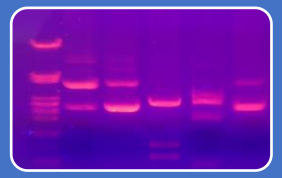
Mas diferente do alfabeto binário dos computadores, os dados genéticos são armazenados com um alfabeto quaternário (A, C, G e T)



Mais tarde se descobriu que a forma dos genes operarem também é digital. Até certo ponto, os genes podem ser “ligados” ou “desligados”

Apenas estas observações já seriam suficientes para prever, na década de 1950, que um dia informática e biologia molecular iriam juntas fazer nascer uma nova área de conhecimento

Histórico



Apesar da estrutura do DNA ter sido desvendada em 1953, a informação nela contida não podia ser “lida”

Foi preciso esperar até fins da década de 1980 para que aparecesse uma “lente de aumento” suficientemente boa que permitisse a leitura do DNA em grande quantidade

Na década de 1970 a unidade básica de armazenamento de informação era o kilobyte - aproximadamente 1000 letras;

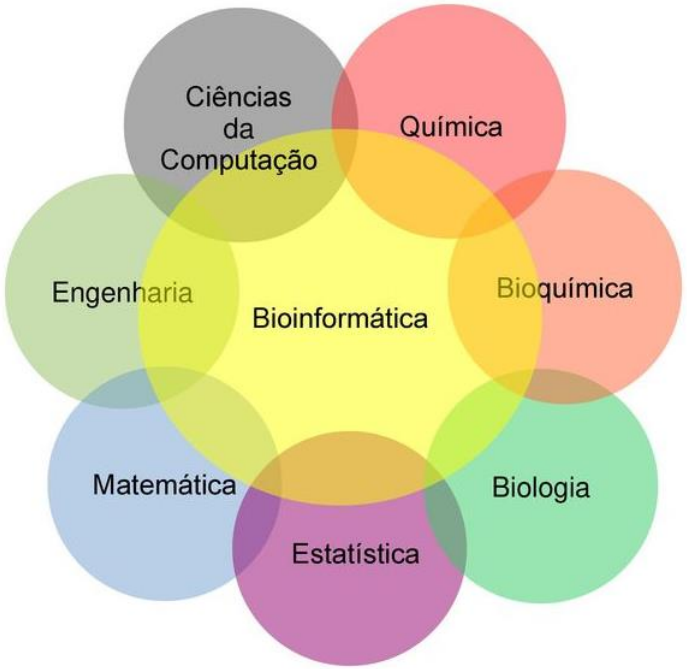
Um computador da época tinha alguns kbytes de memória;

Com tal memória um computador desses não seria capaz de processar nem sequer o genoma de um vírus (20 kb), ou 20 mil letrinhas; que dirá o genoma humano, com seus 3 bilhões de letrinhas.

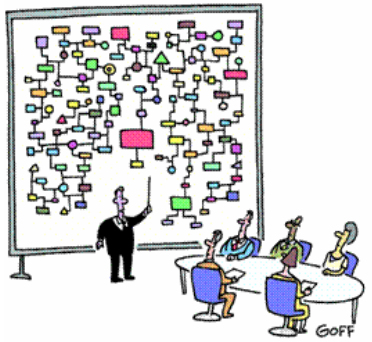
Foi preciso esperar alguns anos para que essas duas áreas alcançassem formas de produzir a biologia em larga-escala. Produção de dados em massa gera demanda para análises computacionais => Bioinformática.

Bioinformática

- "…aplicação das **técnicas da informática**, no sentido de **análise da informação** na área de estudo da **biologia**…";
- "...a utilização de **técnicas computacionais** e **matemáticas** relacionadas ao **conhecimento químico, físico, biológico, biomédico**,...".



Desafios

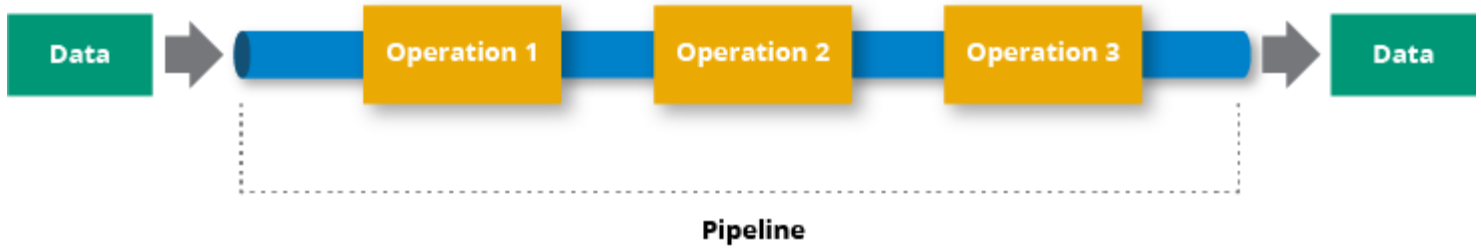


"And that's why we need a computer."

- Sequenciamento
- *Base Calling*
- Qualidade do sequenciamento
- Alinhamento/Montagem
- Predição/Anotação
- Vias metabólicas
- Expressão diferencial
- *Splicing* alternativo
- Identificação de mutações
- Filogenia
- Regiões promotoras

- Regiões não-codificantes
- RNA-Seq, miRNA-Seq, ChiP-Seq
- Domínios de proteínas
- Bioinformática estrutural
- Bancos de dados biológicos
- *Big Data*
- *Machine Learning*
- Biologia sintética
- Medicina personalizada
- ...

Como processar os dados de RNA-Seq?



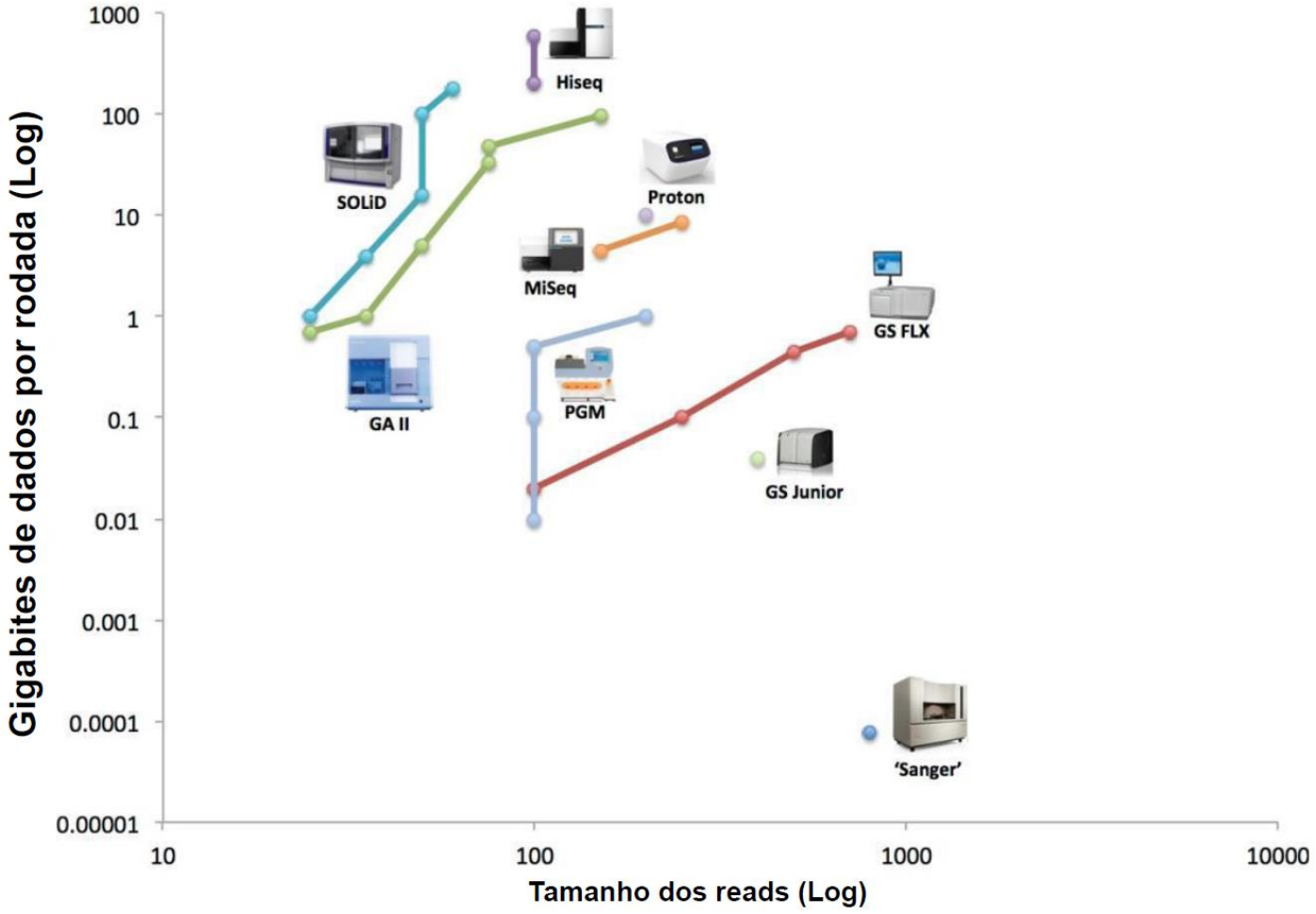
O que é RNA-Seq?

- Nome dado às novas tecnologias de sequenciamento (*Next-generation sequencing*) aplicadas aos transcriptomas, ou seja, às regiões do DNA transcritas em moléculas de RNAs.
- Podemos fazer:
  - Analisar a expressão diferencial em diferentes tecidos ou condições ambientais;
  - Analisar diferentes isoformas (*alternative splicing*);
  - Descobrir novas regiões dos genomas que são transcritas;
  - Identificar moléculas de RNAs que participam de processos regulatórios;
  - etc.

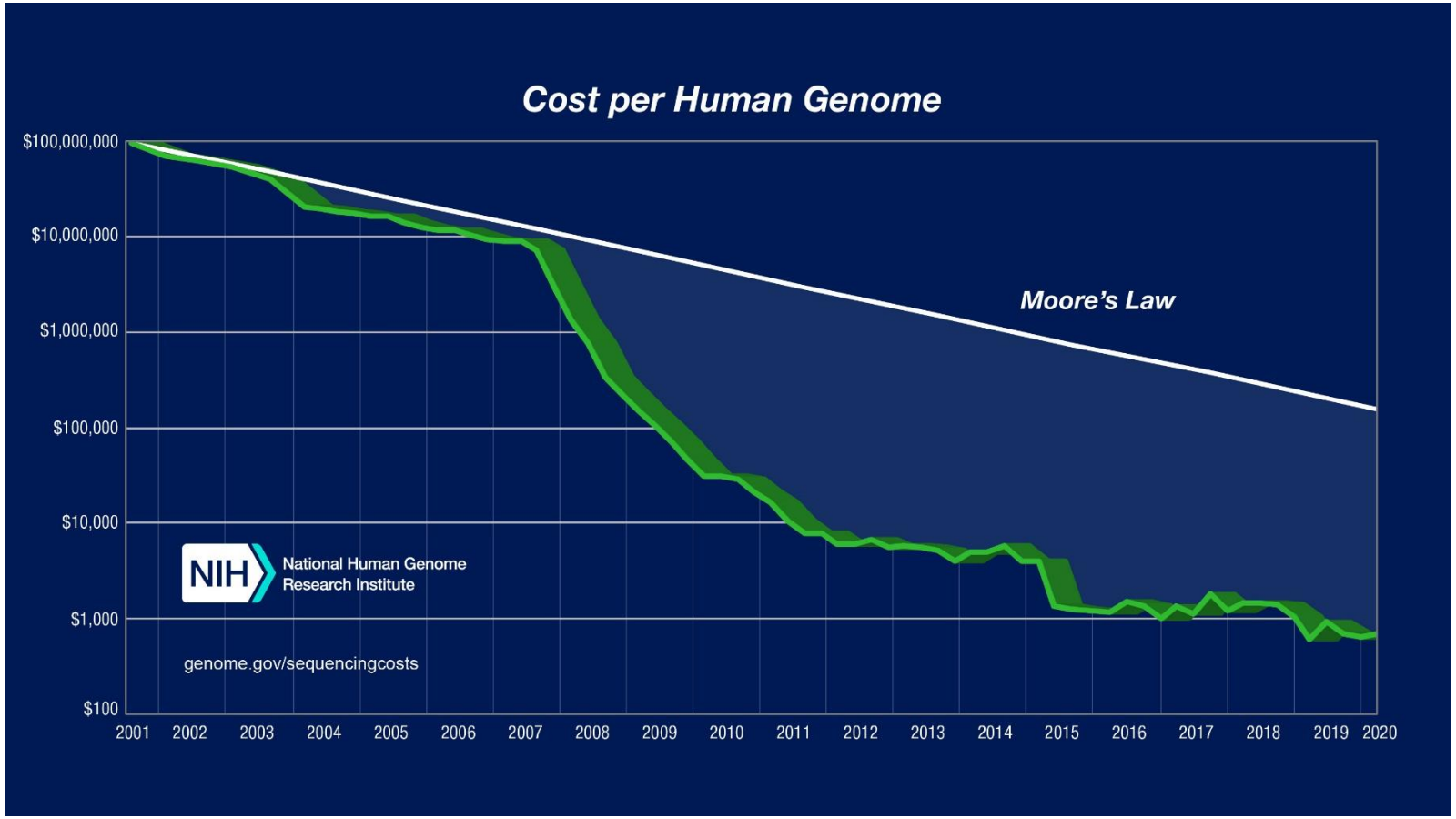
Por que sequenciar?

- Identificar sequências funcionais e possibilitar a caracterização dos componentes moleculares dos sistemas biológicos (ex: genoma/transcriptoma);
- Pesquisar as regiões gênicas;
- Acelerar o processo de anotação genômica;
- Obter a expressão gênica relativa para diferentes células sob diferentes condições;
- Detectar mutações pontuais e/ou estruturais;
- Auxiliar na identificação de eventos de processamento alternativo de transcritos (*Alternative Splicing*) em condições biológicas específicas.

Evolução do sequenciamento

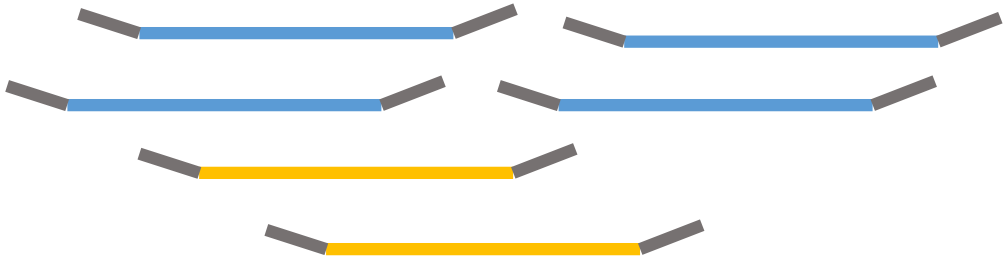


Custo por Genoma



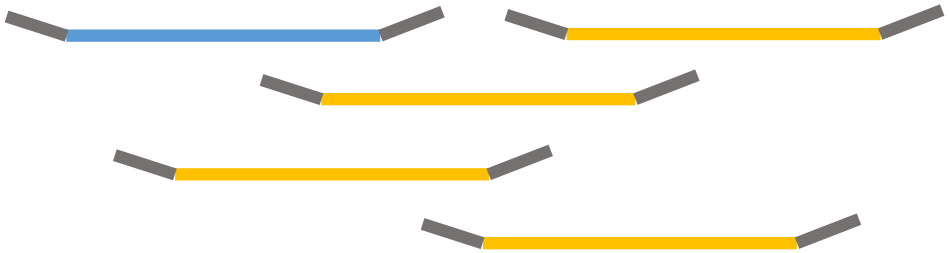
Análise de RNA-Seq

Controle (ex. não tratado)



Reads (leituras) do RNA-Seq

Desafio (ex. tratado com droga)



Genoma de referência

Desafio vs. Controle (Tratado vs. Não tratado)

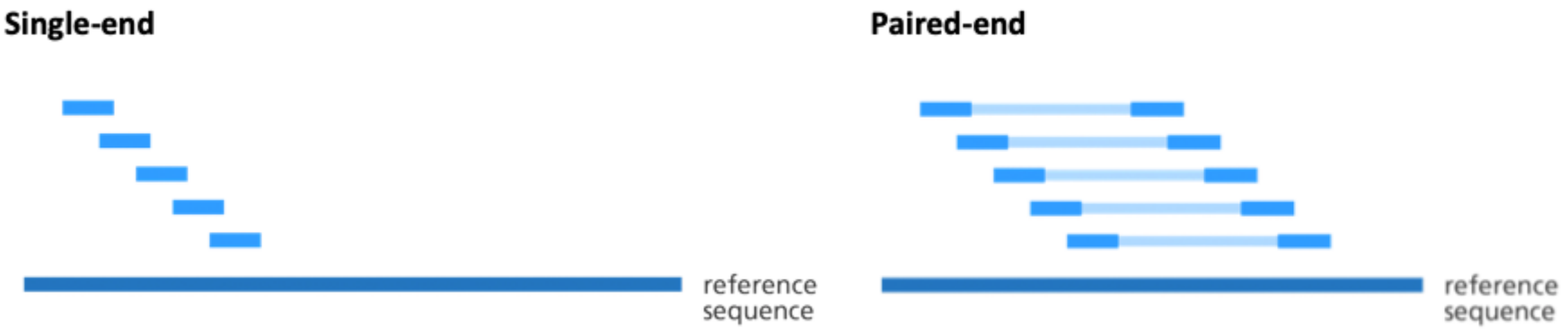
Gene A - UP

Gene B - DOWN

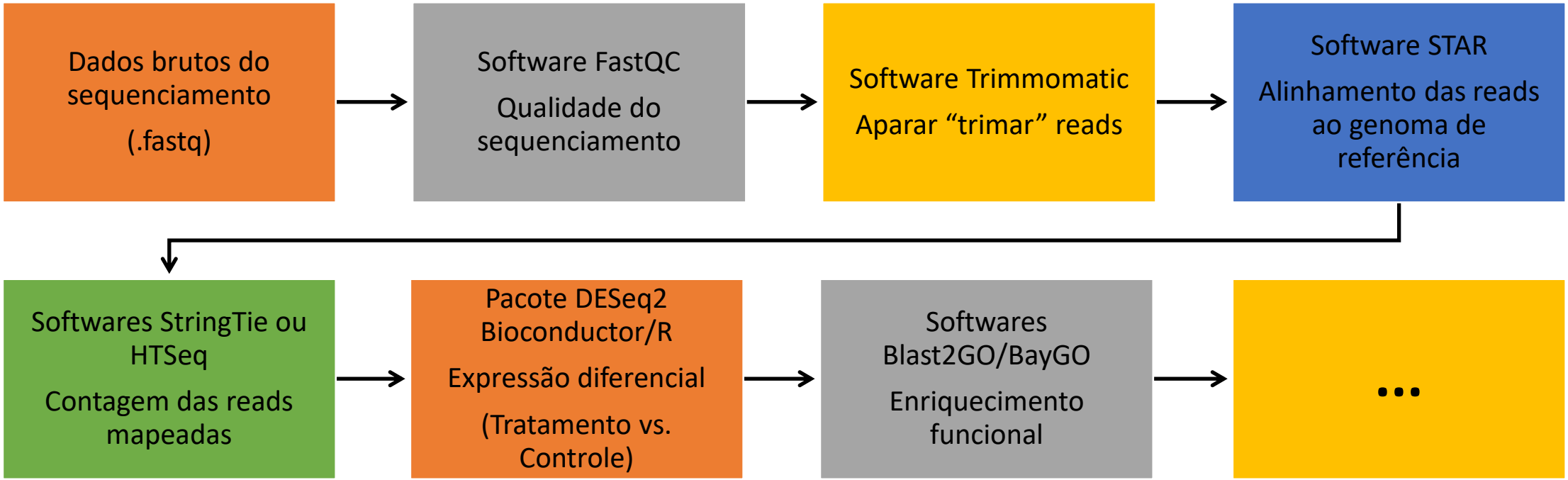
Bibliotecas de sequenciamento (RNA-Seq)

- Bibliotecas de fragmentos (***single-end***)
  - Resultam no sequenciamento de apenas uma das extremidades do fragmento, sendo a metodologia mais simples e barata.

- Bibliotecas de extremidades pareadas (***paired-end***)
  - Resultam em duas leituras para cada fragmento sequenciado, uma referente à fita *forward* e outra à fita *reverse*.



Exemplo de pipeline (RNA-Seq)



Dados brutos do sequenciamento

- Dados gerados pelos sequenciadores automáticos;
- Exemplos de formatos de arquivos gerados:
  - FASTA
  - SFF (ROCHE 454)
  - CSFASTA (ABI SOLiD)
  - FASTQ
  - ...

Formato FASTA

```
>Sequence_1 assembly1
CCCTAAACCCTAAACCCTAAACCCTAAACCTCTGAATCCTTAATCCCTAAATCCCTAAAT
CTTTAAATCCTACATCCATGAATCCCTAAATACCTAATTCCCTAAACCCGAAACCGGTTT
CTCTGGTTGAAAATCATTGTGTATATAATGATAATTTTATCGTTTTTATGTAATTGCTTA
TTGTTGTGTGTAGATTTTTTAAAAATATCATTTGAGGTCAATACAAATCCTATTTCTTGT
GGTTTTCTTTCCTTCACTTAGCTATGGATGGTTTATCTTCATTTGTTATATTGGATACAA
GCTTTGCTACGATCTACATTTGGGAATGTGAGTCTCTTATTGTAACCTTAGGGTTGGTTT
ATCTCAAGAATCTTATTAATTGTTTGGACTGTTTATGTTTGGACATTTATTGTCATTCTT
>Sequence_2
CCCTAAACCCTAAACCCTAAACCCTAAACCTCTGAATCCTTAATCCCTAAATCCCTAAAT
CTTTAAATCCTACATCCATGAATCCCTAAATACCTAATTCCCTAAACCCGAAACCGGTTT
CTCTGGTTGAAAATCATTGTGTATATAATGATAATTTTATCGTTTTTATGTAATTGCTTA
TTGTTGTGTGTAGATTTTTTAAAAATATCATTTGAGGTCAATACAAATCCTATTTCTTGT
GGTTTTCTTTCCTTCACTTAGCTATGGATGGTTTATCTTCATTTGTTATATTGGATACAA
GCTTTGCTACGATCTACATTTGGGAATGTGAGTCTCTTATTGTAACCTTAGGGTTGGTTT
ATCTCAAGAATCTTATTAATTGTTTGGACTGTTTATGTTTGGACATTTATTGTCATTCTT
```

Formato FASTQ

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

Cada read é representada por 4 linhas no arquivo

```
@ read ID
Sequência
+ read ID
Qualidade
```

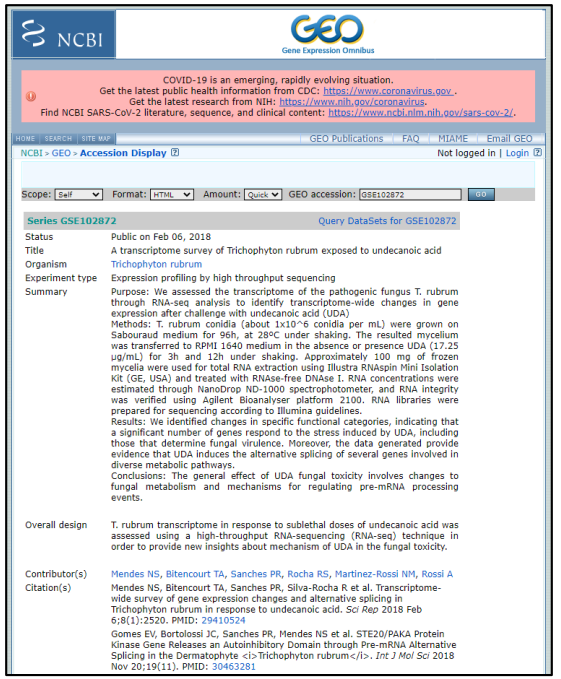
Dados brutos disponíveis na web

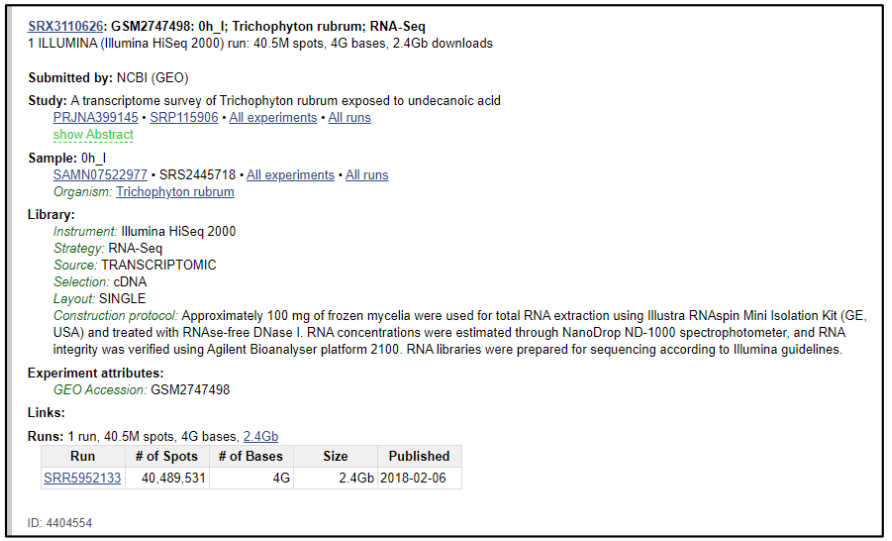
- GEO Gene Expression Omnibus
  - http://www.ncbi.nlm.nih.gov/geo/

- Exemplo:
  - *A transcriptome survey of Trichophyton rubrum exposed to undecanoic acid*
  - https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE102872

- Após download converter arquivos SRA para FASTQ

```
$ fastq-dump –B arquivo.sra
```





Qualidade do sequenciamento

- Avaliar a qualidade das reads
  - Identificar contaminantes;
  - Identificar amostras com baixa performance de sequenciamento.
  - Softwares: **FastQC**, SAMStat, ...

- Download do software FastQC
  - http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

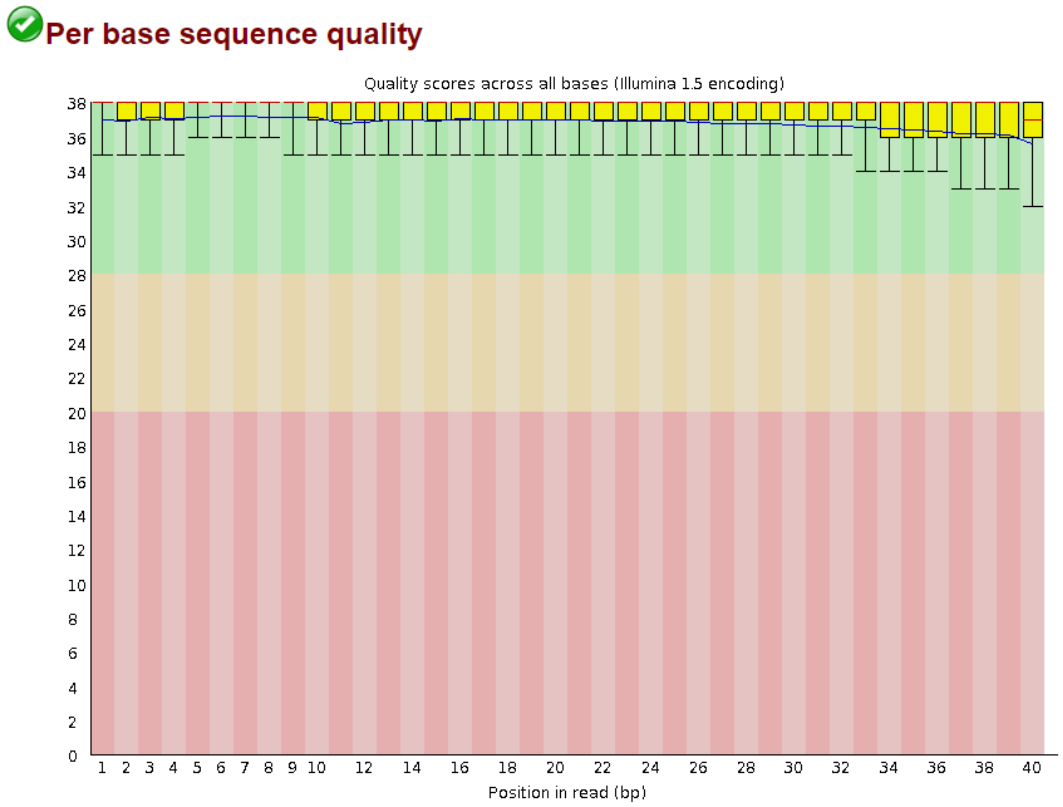
- Exemplo de uso do software FastQC

```
$ fastqc arquivo.fastq
```

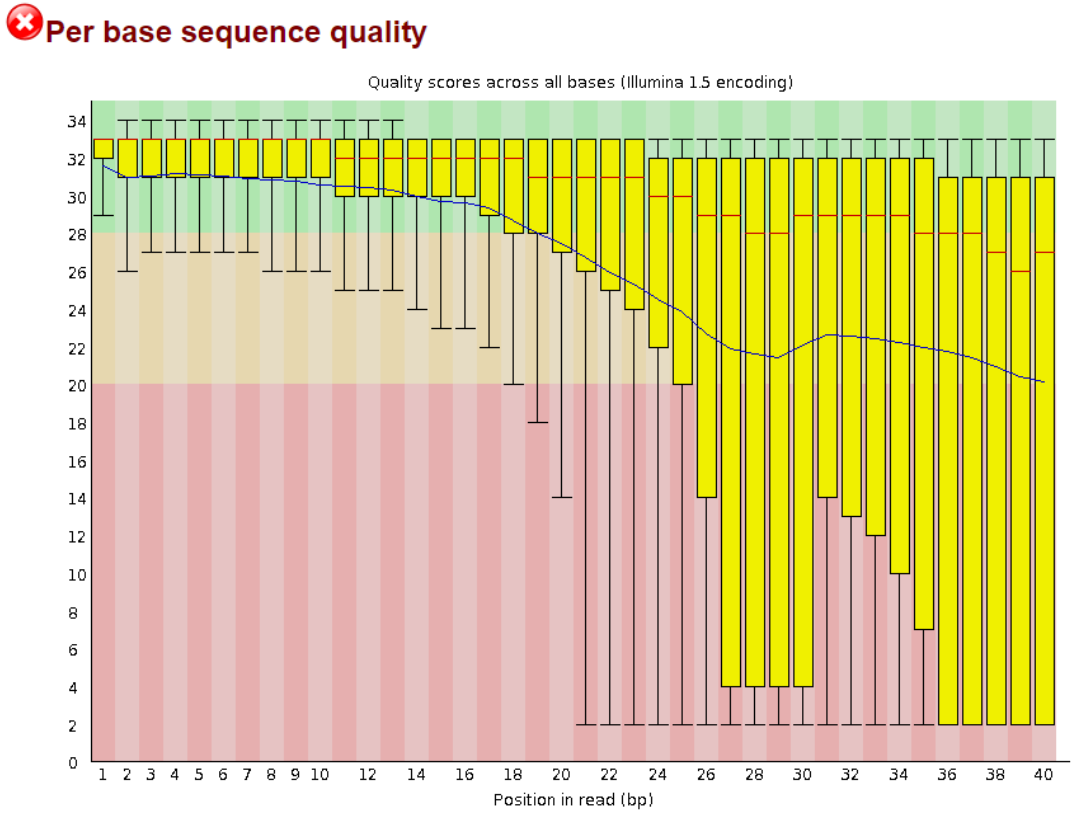
- Resultado do software FastQC
  - Serão gerados arquivos formato .html para visualização em navegador web

Exemplos de resultados - FASTQC

Qualidade boa



Qualidade ruim



Exemplos de resultados - FASTQC

Qualidade boa

Overrepresented sequences

No overrepresented sequences

Qualidade ruim

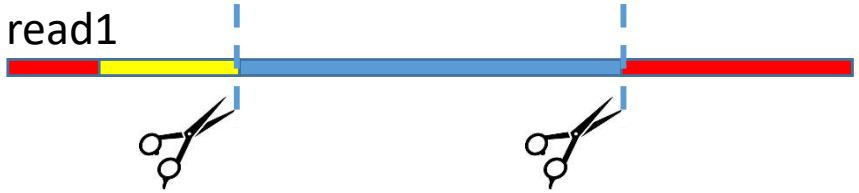
Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTTC	2065	0.5224039181558763	No Hit
GATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATG	2047	0.5178502762542754	No Hit
ATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATGA	2014	0.5095019327680071	No Hit
CGATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTTAT	1913	0.4839509420979134	No Hit
GTATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAGA	1879	0.47534961850600066	No Hit
AAAAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCT	1846	0.4670012750197325	No Hit
TGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCAT	1841	0.46573637449150995	No Hit
AACCTGCAGAGTTTTATCGCTTCCATGACGCAGAAGTTAA	1836	0.46447147396328753	No Hit
GATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTTATC	1831	0.4632065734350651	No Hit
AAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTC	1779	0.45005160794155147	No Hit
ATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCA	1779	0.45005160794155147	No Hit
AATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCC	1760	0.4452449859343061	No Hit
AAAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTT	1729	0.4374026026593269	No Hit
CGTATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAG	1713	0.43335492096901496	No Hit
ATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAGAAG	1708	0.43209002044079253	No Hit
CAGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTT	1684	0.42601849790532476	No Hit
TGCAGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACT	1668	0.4219708162150128	No Hit
CAACCTGCAGAGTTTTATCGCTTCCATGACGCAGAAGTTA	1668	0.4219708162150128	No Hit
TATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAGAA	1630	0.4123575722005221	No Hit

Aparar “trimar” *reads*

- Processo que permite fazer os cortes e ajustes necessários nas *reads* (leituras)
  - Retirada das sequências de adaptadores;
  - Manutenção de sequências com escore mínimo de qualidade e tamanho mínimo;
  - Softwares: **Trimmomatic**, Prinseq, FASTX-Toolkit, ...

- Download do software Trimmomatic
  - http://www.usadellab.org/cms/?page=trimmomatic



- Exemplo de uso do software Trimmomatic para biblioteca *paired-end*

```
$ java -jar trimmomatic-0.36.jar PE -threads 8 -phred33 ARQ_R1.fastq.gz
ARQ_R2.fastq.gz ARQ_R1.paired.fastq.gz ARQ_R1.unpaired.fastq.gz ARQ_R2.paired.fastq.gz
ARQ_R2.unpaired.fastq.gz ILLUMINACLIP:/adapters/TruSeq3-PE-2.fa:2:30:10 LEADING:3
TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
```

- Resultado do software Trimmomatic
  - Serão gerados arquivos formato .fastq com *reads* “trimadas”

Exemplos de resultados - Trimmomatic

Arquivo de resultado

```
Input Read Pairs: 28987947
Both Surviving: 27213240 (93.88%)
Forward Only Surviving: 784150 (2.71%)
Reverse Only Surviving: 863068 (2.98%)
Dropped: 127489 (0.44%)
```

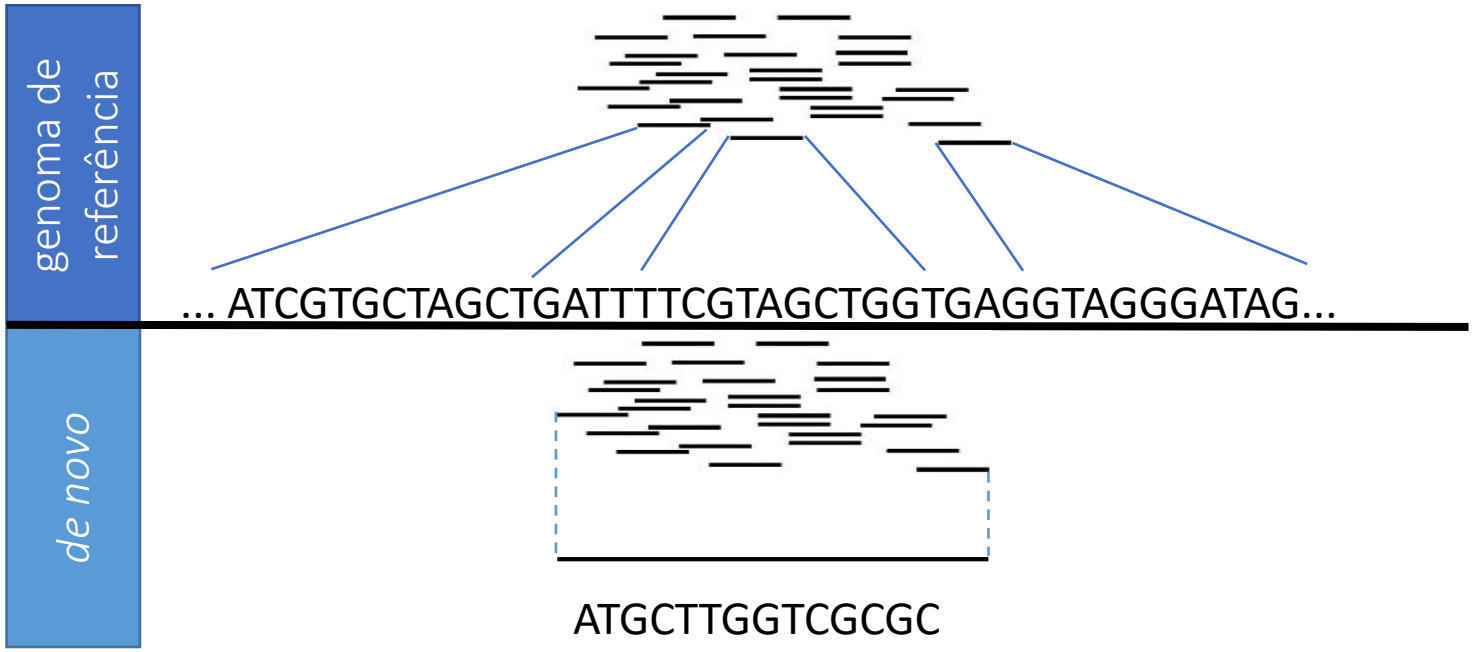
Outros resultados

Sample	Raw reads	High-quality reads
0 hour I (SR)	40,489,531	40,149,007
0 hour II (PE)	28,987,947	27,213,240
0 hour III (PE)	28,893,017	27,136,542
3 hours I (SR)	60,724,079	60,235,207
3 hours II (PE)	67,859,806	63,440,961
3 hours III (PE)	28,182,638	26,438,636
12 hours I (SR)	45,617,478	44,174,487
12 hours II (PE)	30,746,489	28,524,813
12 hours III (PE)	12,463,078	10,635,512

SR=Single-read; PE=Paired-end

Alinhamento das *reads*

- Processo de alinhamento das *reads* sequenciadas (já “trimadas”) no genoma de referência.
  - Quando não houver um genoma de referência, utilizar o método de Alinhamento *de novo.*
  - Softwares: **STAR**, BWA, Bowtie, Bowtie2, Tophat, Tophat2, HISAT, HISAT2, Velvet, ...



Exemplo de uso do software STAR

- Passo 1 - Criando o arquivo de índice do genoma de referência

```
$ STAR  --runThreadN 8  --runMode genomeGenerate --genomeDir
dir_do_genoma --genomeFastaFiles arquivo_supercontigs.fasta --
sjdbGTFfile arquivo_transcripts.gtf --sjdbOverhang 99
```

GTF

```
Supercontig21   TR_CBS118892_V2_FINAL_CALLGENES_5   start_codon   14925   14927   .   -   0   gene_id "TERG_00002"; transcript_id "TERG_00002T0";
Supercontig21   TR_CBS118892_V2_FINAL_CALLGENES_5   stop_codon    12841   12843   .   -   0   gene_id "TERG_00002"; transcript_id "TERG_00002T0";
Supercontig21   TR_CBS118892_V2_FINAL_CALLGENES_5   exon    14609   14927   .   -   .   gene_id "TERG_00002"; transcript_id "TERG_00002T0";
Supercontig21   TR_CBS118892_V2_FINAL_CALLGENES_5   CDS     14609   14927   .   -   0   gene_id "TERG_00002"; transcript_id "TERG_00002T0";
Supercontig21   TR_CBS118892_V2_FINAL_CALLGENES_5   exon    13933   14529   .   -   .   gene_id "TERG_00002"; transcript_id "TERG_00002T0";
Supercontig21   TR_CBS118892_V2_FINAL_CALLGENES_5   CDS     13933   14529   .   -   2   gene_id "TERG_00002"; transcript_id "TERG_00002T0";
Supercontig21   TR_CBS118892_V2_FINAL_CALLGENES_5   exon    13702   13869   .   -   .   gene_id "TERG_00002"; transcript_id "TERG_00002T0";
Supercontig21   TR_CBS118892_V2_FINAL_CALLGENES_5   CDS     13702   13869   .   -   2   gene_id "TERG_00002"; transcript_id "TERG_00002T0";
Supercontig21   TR_CBS118892_V2_FINAL_CALLGENES_5   exon    13470   13638   .   -   .   gene_id "TERG_00002"; transcript_id "TERG_00002T0";
Supercontig21   TR_CBS118892_V2_FINAL_CALLGENES_5   CDS     13470   13638   .   -   2   gene_id "TERG_00002"; transcript_id "TERG_00002T0";
Supercontig21   TR_CBS118892_V2_FINAL_CALLGENES_5   exon    12841   13414   .   -   .   gene_id "TERG_00002"; transcript_id "TERG_00002T0";
Supercontig21   TR_CBS118892_V2_FINAL_CALLGENES_5   CDS     12844   13414   .   -   1   gene_id "TERG_00002"; transcript_id "TERG_00002T0";
```

FASTA

```
>Supercontig21 of Trichophyton rubrum CBS 118892
ATAGTATTACTTACTACTACTATAAGTCCCTATTATAGAAGGTATGCTCTACTATTAGTA
TCTAATCTAGATAACTATAAAACTATATTAAAATATCTTTAAATACTTAGATATATAGCG
GTAGTTAAACTACTAGTATACCTCTATTTAGAAGGCGTAGATAGCTTATTTAACTCTCTA
GATAATACTAAAGTATAGAGATTTAATATTAAGTACTACCTATATTAGCCTTTTTATAAT
ATTAATAATAATCTTCTATAAAGTAGTAATACCTATAAAAAAAGCTATTACTATTCCTATA
GAGTGTAATAATAATATTATATCTAATTCTAATAGTACTTTAGAGGATATTAATATAAAA
GATAACTCCACTACTATTTTAATAGCCCCTACTATATCTAGACCCCCTAAGTACCGTAAG
TATAAGGACGCGTCCTCTAATAGAGCTACTATAGCTAAGTAGCTAAAAGTAGATAGGCTA
ATATCTTATATTATATTAAGGACTAATATAGTACTAAATATAACTAATAAATCTAAAAAA
GATAGTAAAGGTTTTATATTTAGATTTTAGAATATATTAATATCTAGAGTATTATCTACT
```

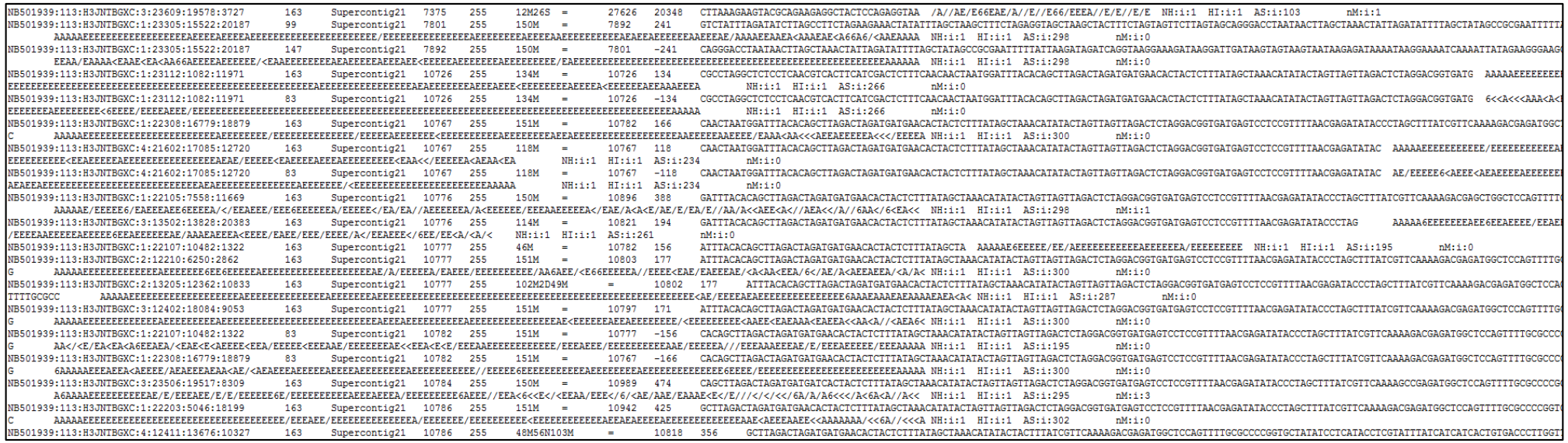
Exemplo de uso do software STAR

- Passo 2 – Alinhando as *reads* ao genoma de referência

```
$ STAR --runThreadN 8 --genomeDir dir_do_genoma --readFilesIn
ARQ_R1_TRIMMED.paired.fastq.gz ARQ_R2_TRIMMED.paired.fastq.gz --
readFilesCommand zcat --outSAMtype BAM SortedByCoordinate --
outReadsUnmapped Fastx --outSJfilterReads Unique --twopassMode
Basic --outFilterMultimapNmax 1 --outFilterType BySJout --
alignSJoverhangMin 15 --alignSJDBoverhangMin 3 --
outFilterMismatchNoverLmax 0.06 --quantMode TranscriptomeSAM
GeneCounts --outFileNamePrefix dir_de_saída
```

- Resultado do software STAR
  - Serão gerados arquivos formato .BAM com *reads* alinhadas;
  - Caso utilize o parâmetro GeneCounts, será gerado o arquivo de contagem de *reads.*

Exemplos de resultados - STAR



SAM/BAM

```
                 Number of input reads |       46661059
           Average input read length |       271
                         UNIQUE READS:
        Uniquely mapped reads number |       45099456
             Uniquely mapped reads % |       96.65%
```

LOG

```
TERG_00002      2441
TERG_00003      7014
TERG_00004      12719
TERG_00008      9056
TERG_00009      1960
TERG_00010      35
TERG_00011      483
TERG_00012      4716
TERG_00013      691
TERG_00014      83
TERG_00015      3444
```

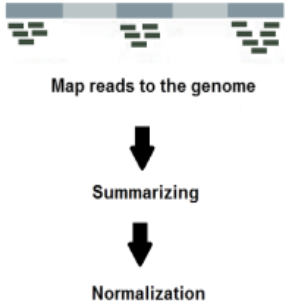
COUNT

Sample	High-quality reads	Mapped reads STAR	Total mapped reads (%)
Amostra I	46,661,059	45,099,456	96.65
Amostra II	31,721,176	22,577,724	71.18
Amostra III	42,817,030	42,286,629	98.76

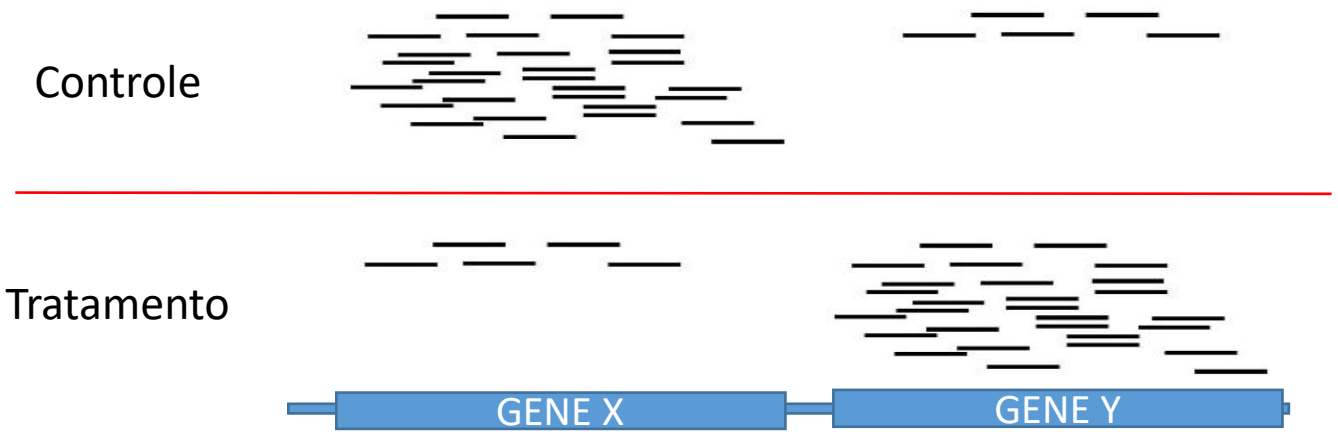
Outros resultados

Expressão diferencial

- Ideal que se tenha 3 ou mais sequenciamentos independentes para cada tratamento e para cada grupo controle;
- Podemos usar softwares como: **DESeq2**, EdgeR para normalizar e extrair o diferencial de expressão (estatísticas);
- Ainda outras opções: cuffdiff, limma, baySeq, ...



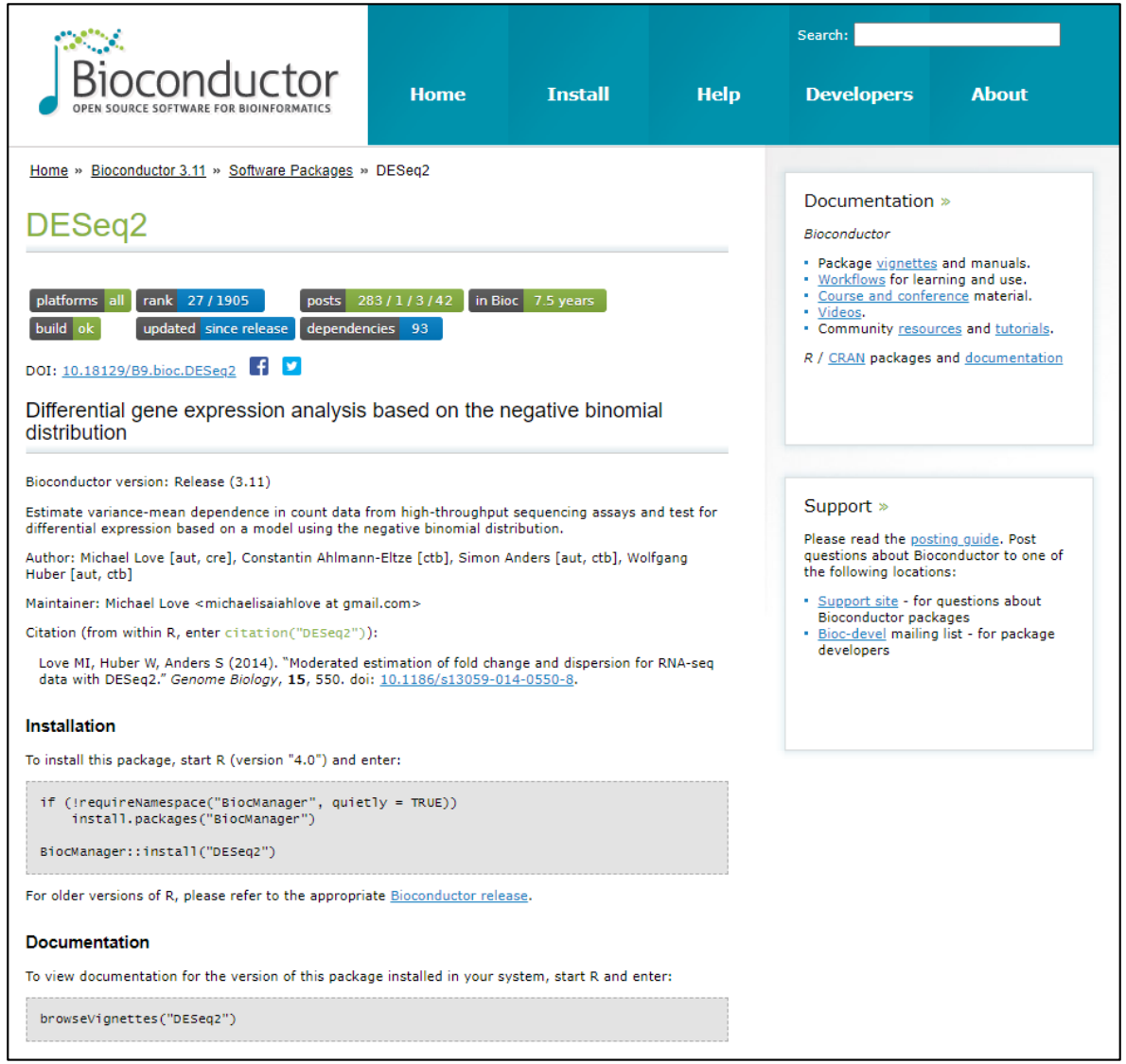
Representação hipotética



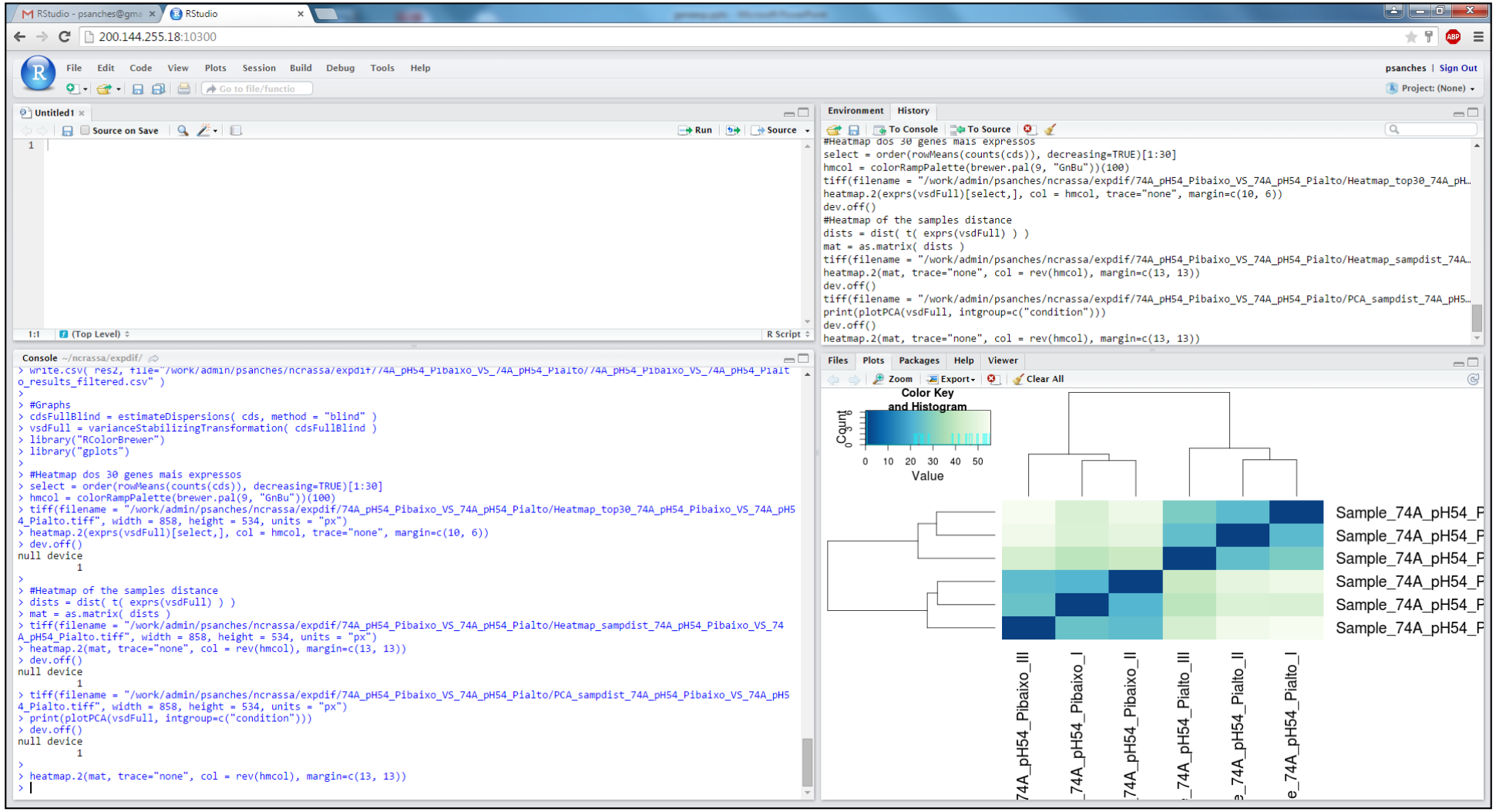
	Controle	Tratamento	FoldChange	log_2FC
Gene X	28	5	0,18	-2,49
Gene Y	5	28	5,60	2,49

DESeq2

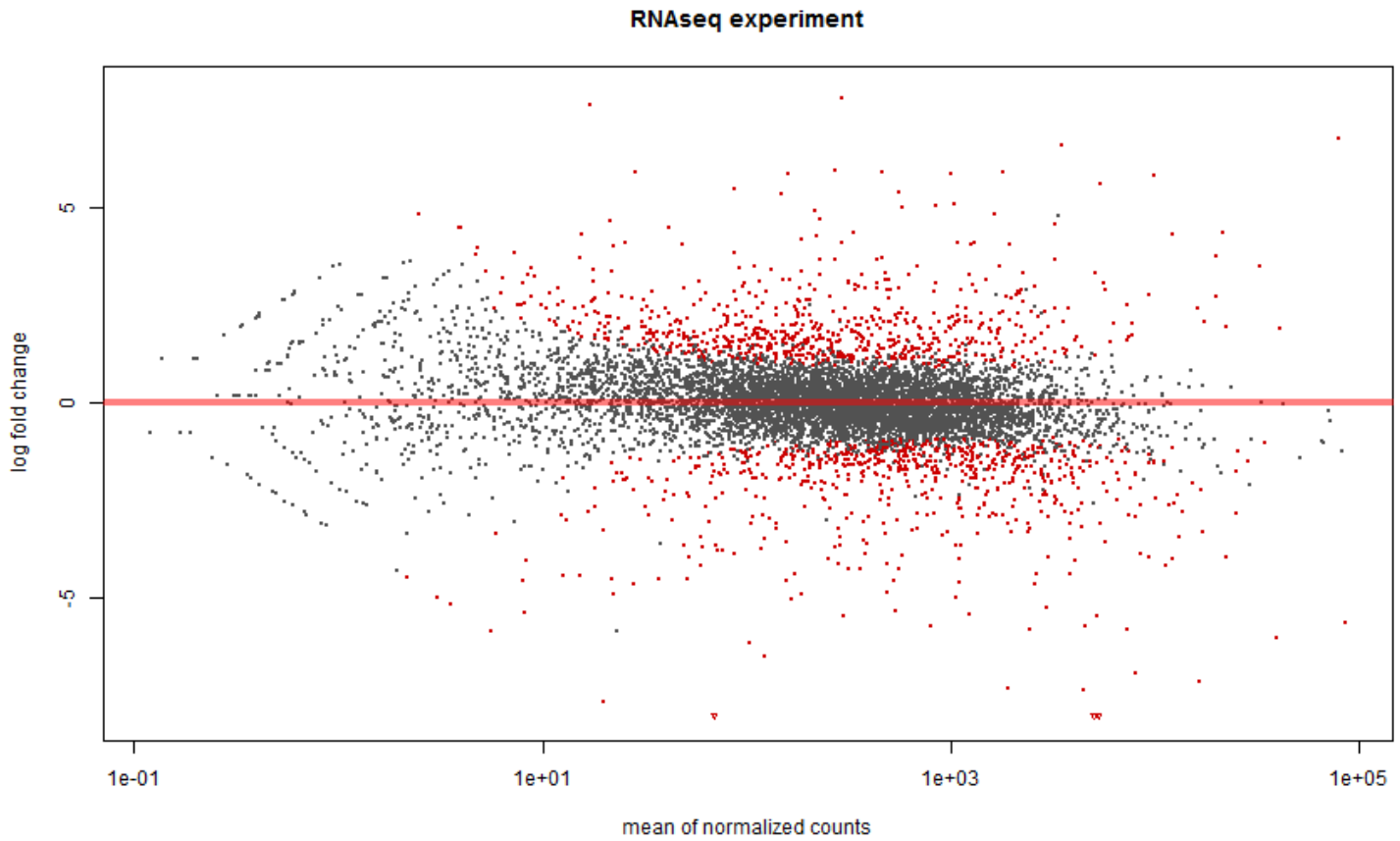
- Baseada em *read counts*
- Pacote do Bioconductor/R
  - https://www.bioconductor.org/
- R
  - https://www.r-project.org/



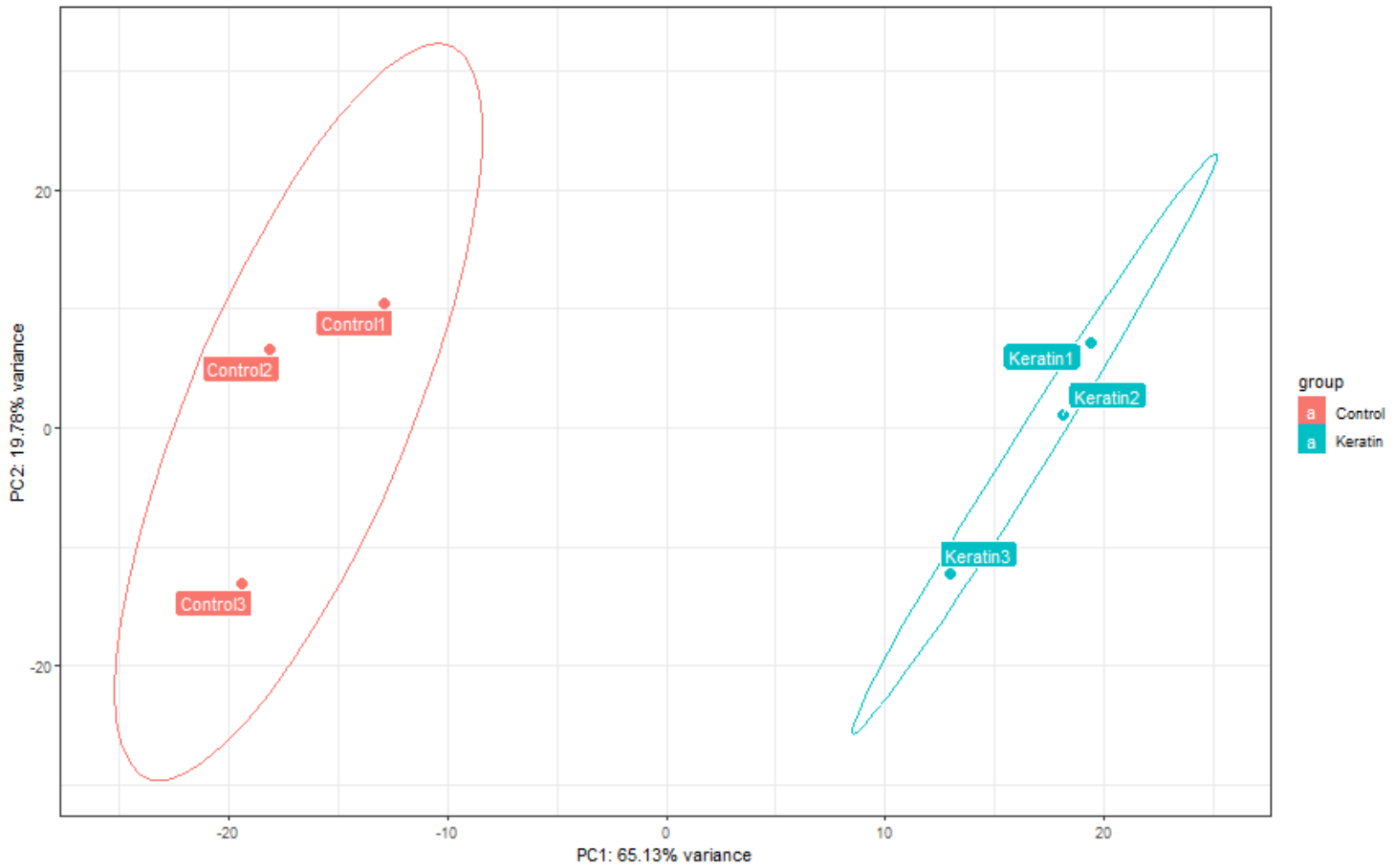
RStudio - https://rstudio.com/



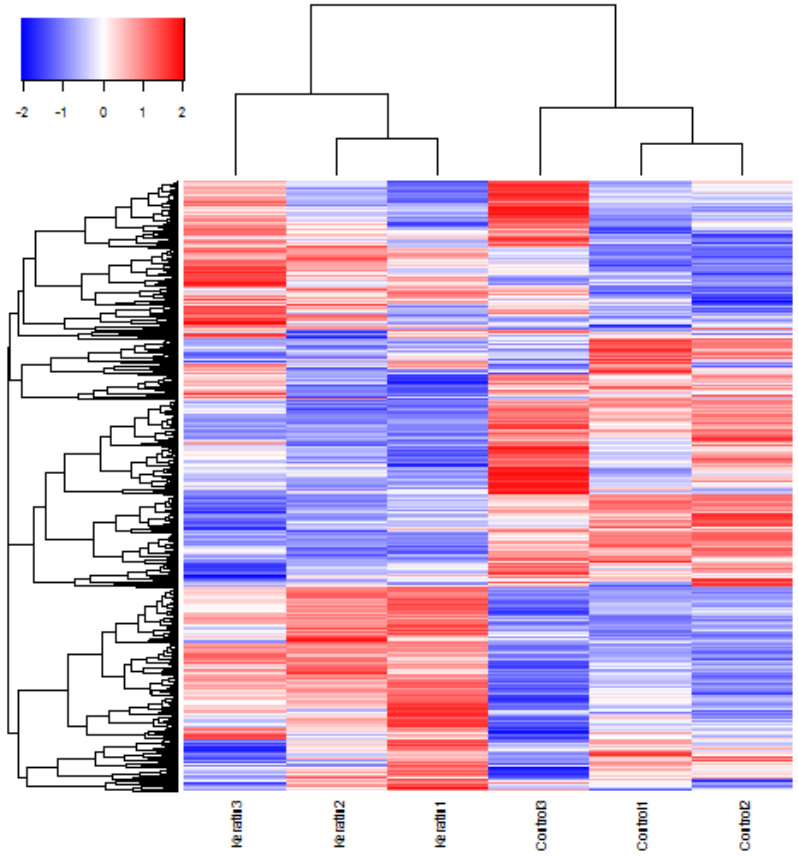
Exemplos de resultados – DESeq2



Exemplos de resultados – DESeq2



Exemplos de resultados – DESeq2



Exemplos de resultados – DESeq2

	gene	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	Control1	Control2	Control3	Keratin1	Keratin2	Keratin3
1	TERG_08353	5263,696589	-9,002452266	0,457685141	-19,66953142	3,93E-86	3,35E-82	12972,30597	8400,824323	10148,10033	25,66658515	23,4192922	11,86303686
2	TERG_00867	8034,057675	-6,924413959	0,449552077	-15,40291839	1,56E-53	6,65E-50	21224,51933	16861,695	9725,389828	177,182233	117,096461	98,46320595
3	TERG_03274	39068,16881	-6,010855869	0,410694252	-14,63584122	1,66E-48	4,70E-45	75212,20447	101602,9569	54013,45959	911,57775	1348,458193	1320,356003
4	TERG_06242	9819,287482	5,837124878	0,403135796	14,47930184	1,64E-47	3,48E-44	323,0069501	284,3704525	405,4362701	19229,24	24895,9402	13777,73101
5	TERG_01435	1004,302718	5,852167669	0,406323546	14,40272839	4,97E-47	8,46E-44	36,85314196	33,9885003	31,5			
6	TERG_02842	4997,657133	-8,213330898	0,751223985	-10,93326499	7,99E-28	1,13E-24	10492,3063	699,0301561	1869			
7	TERG_05652	1815,001137	5,907089754	0,547900611	10,78131623	4,22E-27	5,12E-24	96,10721335	40,78620035	40,6			
8	TERG_02974	5210,880922	-5,450595199	0,510534625	-10,67624982	1,31E-26	1,40E-23	5386,339611	18580,38016	6599			
9	TERG_04580	85237,93713	-5,637341532	0,533843746	-10,5599093	4,57E-26	4,32E-23	238209,3157	163247,8999	9989			
10	TERG_00856	1096,023656	-4,605000909	0,44760248	-10,28814877	7,97E-25	6,78E-22	2089,067322	3030,641276	1197			
11	TERG_06585	4478,365775	-7,364525546	0,723264717	-10,18233764	2,38E-24	1,84E-21	1282,633862	20739,78288	4685			
12	TERG_04942	3811,654367	-4,396265658	0,437335943	-10,05237672	8,97E-24	6,36E-21	4557,505223	9213,14948	8061			
13	TERG_03223	3494,277166	6,577117485	0,657796482	9,998711852	1,54E-23	1,01E-20	21,6783188	49,84980043	146,			
14	TERG_01841	160,0771112	5,859269453	0,586463982	9,990842806	1,67E-23	1,02E-20	3,613053134	10,19655009	3,04			
15	TERG_03226	5420,83863	5,624314883	0,564531946	9,962792936	2,22E-23	1,26E-20	76,59672643	140,4858012	429,			
16	TERG_02844	16356,73185	-7,115805721	0,721553954	-9,861779122	6,10E-23	3,24E-20	30338,80716	3611,844631	6348			
17	TERG_06548	804,7925492	-5,693311316	0,586368898	-9,709436045	2,75E-22	1,38E-19	1157,622224	2977,392626	602,			
18	TERG_11610	293,7058974	4,113211732	0,427687852	9,617321872	6,76E-22	3,19E-19	26,01398256	35,12145031	35,			
19	TERG_12038	4566,96921	-5,693308736	0,592763145	-9,604694199	7,64E-22	3,42E-19	10717,76082	13017,59561	3147			
20	TERG_05627	465,3771355	5,890375985	0,618818924	9,518739255	1,75E-21	7,45E-19	24,56876131	5,664750049	15,2			
21	TERG_06509	7364,750494	-5,777097642	0,607896269	-9,503426708	2,03E-21	8,23E-19	14262,16594	21570,23524	7566			
22	TERG_04228	12204,28639	-3,970785992	0,42008175	-9,452412526	3,31E-21	1,28E-18	18508,94859	28842,64135	2148			
23	TERG_06807	457,7852227	3,709139395	0,402693409	9,210827178	3,24E-20	1,20E-17	50,58274387	66,84405058	78,2			
24	TERG_02169	561,5525446	5,384599423	0,588892347	9,143605702	6,04E-20	2,14E-17	33,24008883	24,92490022	20,3			

Exemplo de aplicação de filtros

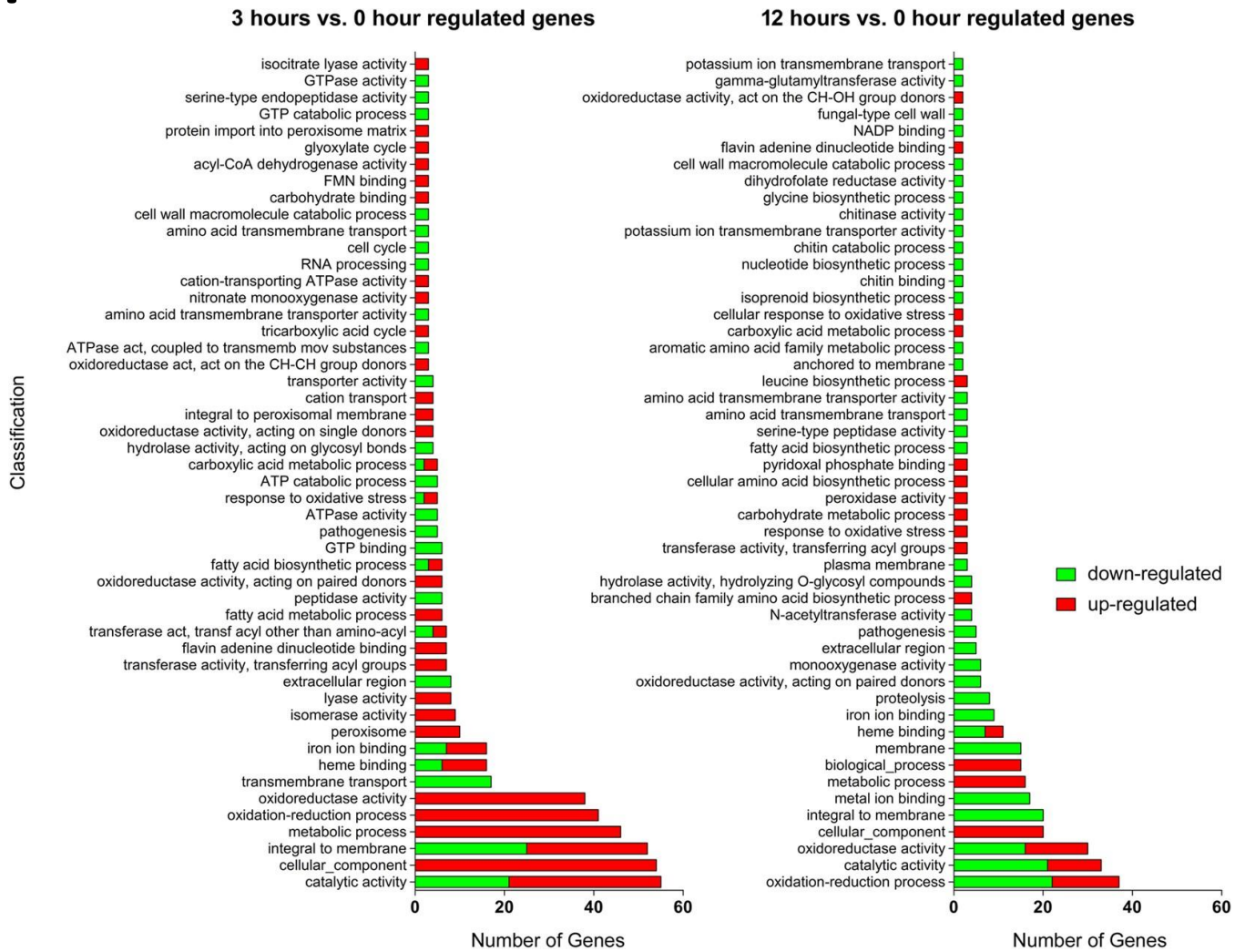
- pvalue < 0,05
- log2FoldChange >= 1,5 (*up-regulated*)
- log2FoldChange <= -1,5 (*down-regulated*)

ID	Gene Product Name	log_2(Fold change)
24 hours		
	Up-regulated	
TERG_01599	hypothetical protein	7.81
TERG_05652	leucine aminopeptidase 1	6.23
TERG_03223	N-acetylglucosamine-6-phosphate deacetylase	6.07
TERG_05627	LysM domain-containing protein *(M. canis)*	6.04
TERG_06242	glucanase, putative *(T. verrucosum)*	5.67
TERG_12035	NB-ARC and TPR domain protein *(A. benhamiae)*	5.64
TERG_07909	isochorismatase family hydrolase, putative *(A. benhamiae)*	5.11
TERG_01841	hypothetical protein	5.06
TERG_03226	glucosamine-6-phosphate deaminase	4.92
TERG_01435	flavin containing polyamine oxidase, putative *(A. benhamiae)*	4.78
	Down-regulated	
TERG_02842	6-hydroxy-D-nicotine oxidase *(T. equinum)*	-9.33
TERG_08353	cytochrome P450 55A3 *(T. tonsurans)*	-8.87
TERG_02746	hypothetical protein	-8.29
TERG_06049	dimethylallyl tryptophan synthase, putative *(T. verrucosum)*	-8.11
TERG_06054	hypothetical protein	-8.05
TERG_02844	major facilitator superfamily transporter *(T. tonsurans)*	-7.69
TERG_02412	HHE domain protein *(T. verrucosum)*	-7.52
TERG_02024	hypothetical protein	-6.95
TERG_11792	hypothetical protein	-6.73
TERG_00867	DUF1212 domain membrane protein *(A. benhamiae)*	-6.67

Enriquecimento funcional

- Identificar classes de genes que estão super-representadas em um grande conjunto de genes;
- Uso de Bancos de Dados e Ferramentas como:
  - Gene Ontology
  - Blast2GO
  - BayGO
  - FunRich
  - ...

Exemplos de resultados – Enriquecimento funcional



Habilidades essenciais - Bioinformata

- Conhecimento na área de Biologia Molecular, Computação e Estatística;
- Conhecimento no uso de ferramentas e pacotes de Bioinformática;
- Desejável conhecimento em linguagens de programação;
- Não ter “medo” da interface de linha de comandos (ex. Linux).

Principais softwares utilizados - Bioinformata

- Sistemas operacionais
  - Linux, MacOS e Windows
- Linguagens de programação
  - R, Python, Perl, Java, C/C++, etc.
- Sistemas Gerenciadores de Bancos de Dados
  - MySQL, PostgreSQL, MariaDB, MongoDB, etc.
- Outros
  - MySQL Workbench, Microsoft Excel, Power BI, etc.







Considerações finais

Resultados ruins também são resultados.

O software auxilia o processo de tomada de decisão, mas quem toma a decisão final é você.

"*Garbage in, garbage out* (GIGO)" -> "lixo entra, lixo sai".
George Fuechsel (Técnico da IBM)

"se devidamente torturados, os dados contam qualquer coisa".
Darrel Huff - Como Mentir com Estatísticas.


Obrigado Pessoal!