Tutorial - RNAseq

Objetivo: Aplicar um pipeline computacional para analisar dados de expressão gênica global (transcritoma).

•Alinhar reads de bibliotecas de RNAseq no genoma

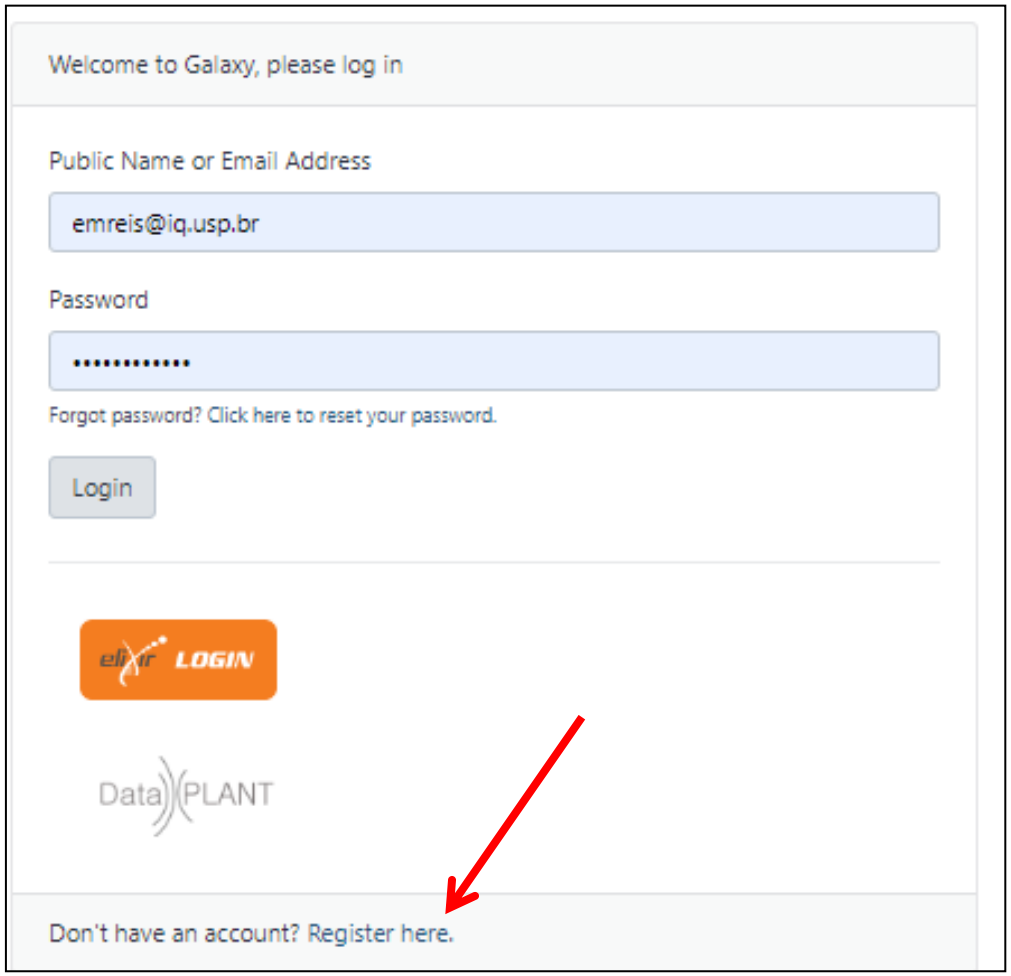
•Visualizar reads alinhados no genoma

•Identificar genes com expressão significativamente alterada em um grupo de amostras

• avaliar a importância do uso de réplicas em estudos de expressão gênica

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

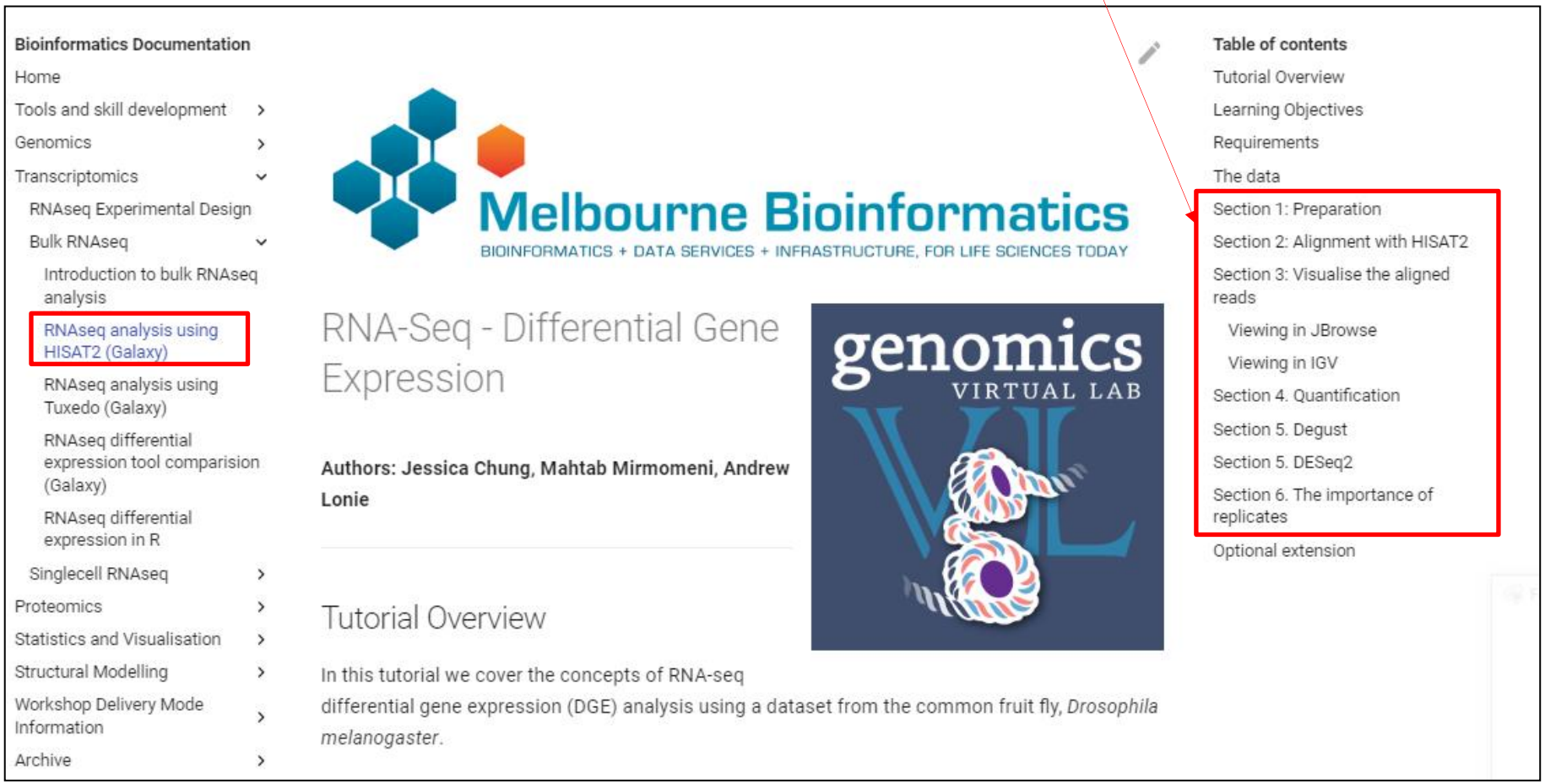
https://usegalaxy.eu/



Link para o tutorial
RNA-Seq Differential Gene Expression: Basic

https://melbournebioinformatics.github.io/MelBioInf_docs/tutorials/rna_seq_dge_basic/rna_seq_basic_tutorial/

Fazer passos das sessões 1 a 6



Estudo original onde foram gerados os dados usados no tutorial

> Genomics. 2016 Oct;108(3-4):158-167. doi: 10.1016/j.ygeno.2016.09.002. Epub 2016 Sep 10.

Transcriptomic response of Drosophila melanogaster pupae developed in hypergravity

Shannon Hateley [1], Ravikumar Hosamani [2], Shilpa R Bhardwaj [3], Lior Pachter [4], Sharmila Bhattacharya [5]

Affiliations + expand

PMID: 27621057 DOI: 10.1016/j.ygeno.2016.09.002

Free article

Abstract

Altered gravity can perturb normal development and induce corresponding changes in gene expression. Understanding this relationship between the physical environment and a biological response is important for NASA's space travel goals. We use RNA-Seq and qRT-PCR techniques to profile changes in early Drosophila melanogaster pupae exposed to chronic hypergravity (3g, or three times Earth's gravity). During the pupal stage, D. melanogaster rely upon gravitational cues for proper development. Assessing gene expression changes in the pupae under altered gravity conditions helps highlight gravity-dependent genetic pathways. A robust transcriptional response was observed in hypergravity-treated pupae compared to controls, with 1513 genes showing a significant ($q<0.05$) difference in gene expression. Five major biological processes were affected: ion transport, redox homeostasis, immune response, proteolysis, and cuticle development. This outlines the underlying molecular and biological changes occurring in Drosophila pupae in response to hypergravity; gravity is important for many biological processes on Earth.

Keywords: Drosophila melanogaster; Hypergravity; Metamorphosis; Pupae; RNA-Seq; Transcriptome

v/27621057/#substances

Tutorial Overview

In this tutorial we cover the concepts of RNA-seq differential gene expression (DGE) analysis using a dataset from the common fruit fly, *Drosophila melanogaster*.

The tutorial is designed to introduce the tools, datatypes and workflows of an RNA-seq DGE analysis. Here, we'll be using a subset of the data from a published experiment by Hateley et. al. in 2016. In practice, full-sized datasets would be much larger and take longer to run.

In this tutorial we will:

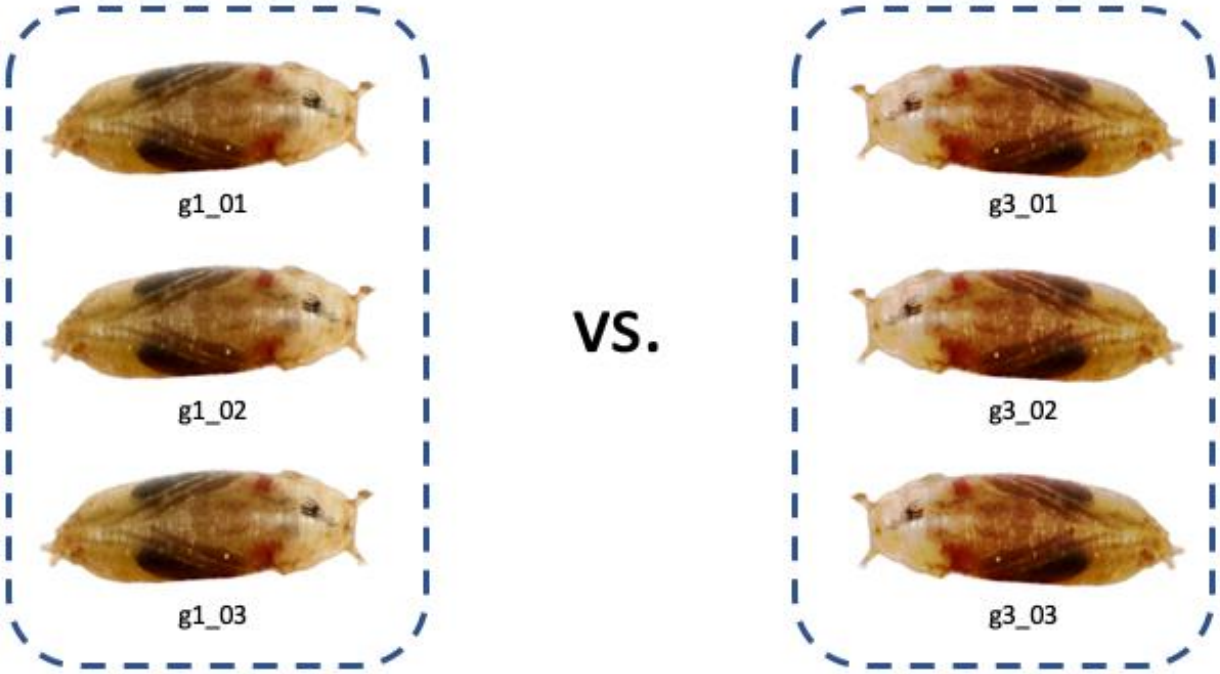
- introduce the types of files typically used in RNA-seq analysis
- align RNA-seq reads with an aligner, HISAT2
- visualise RNA-seq alignment data with IGV or JBrowse
- use a number of different methods to find differentially expressed genes
- understand the importance of replicates for differential expression analysis

This tutorial does not cover the following steps that we might do in a real RNA-seq DGE analysis:

- QC (quality control) of the raw sequence data
- Trimming the reads for quality and for adaptor sequences
- QC of the RNA-seq alignment data

The data

The sequencing data you will be working with is from Drosophila melanogaster pupae from the study, *Transcriptomic response of Drosophila melanogaster pupae developed in hypergravity*. The experiment has two conditions, **g3** where pupae underwent development in three times Earth's gravity (i.e. 3 *g*), and **g1**, the control, where pupae developed in the standard gravitational acceleration felt on the surface of Earth (i.e. 1 *g*). There are three samples in each condition and the sequencing data is paired-end so you will have two files for each of the six samples. Your aim will be to find differentially expressed genes in **g1** vs **g3**.

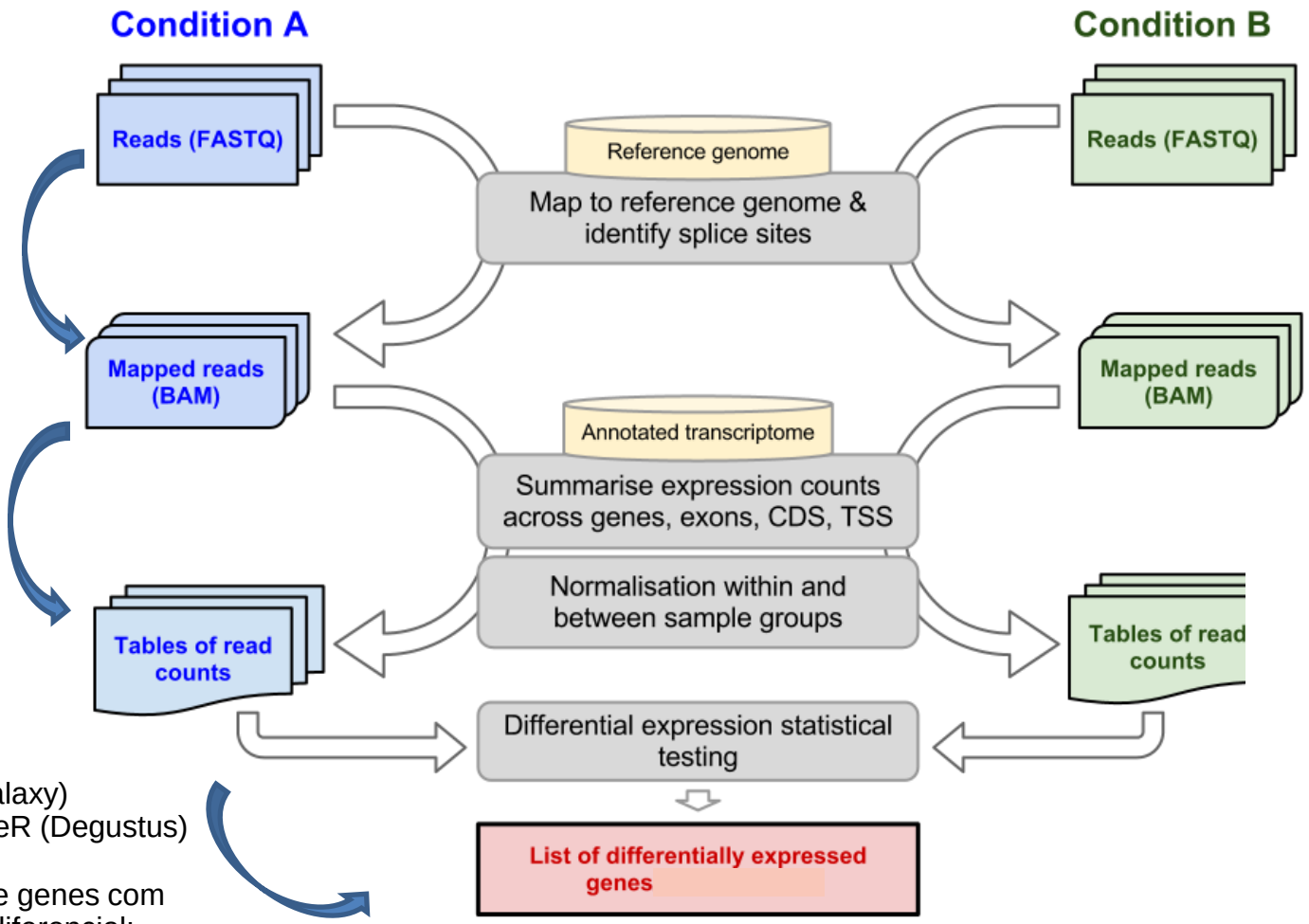


Tutorial (RNAseq Basic): executar sequencia de programas (“pipeline”) para identificar genes com expressão alterada com hipergravidade

1) Mapeamento de reads de bibliotecas de RNA-Seq
2) Obter a contagem de de reads em cada gene
3) Comparar a expressão gênica entre 2 condições experimentais: D.melanogaster exposta a gravidade terrestre (G1) ou hipergravidade (G3), 3 réplicas de cada condição

- HISAT2 alinhamento de sequencias no genoma
- HTseq contagem dos reads alinhados
- DESeq2 (Galaxy) Limma, EdgeR (Degustus)

Identificação de genes com expressão diferencial:



Importante:

algumas informações e links no protocolo online estão desatualizados. Considerar as atualizações abaixo ao executar o tutorial

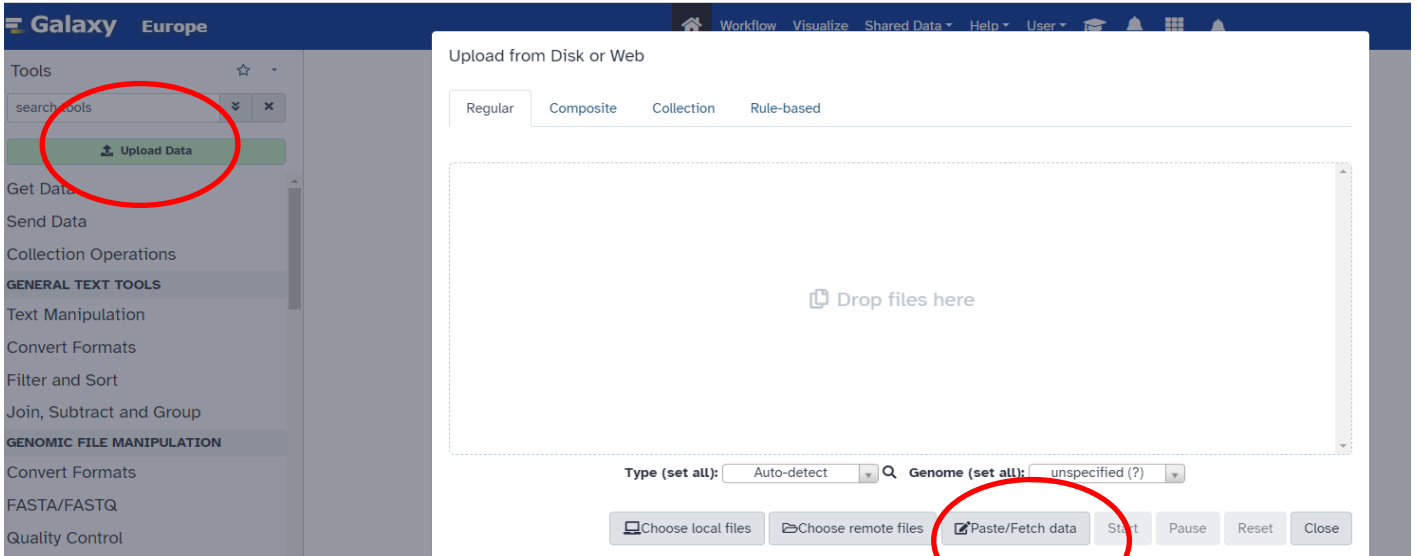
1) Para obter os dados de RNAseq (formato *.fastq), copie e cole os links dos arquivos mostrados abaixo usando a ferramenta “Upload Data” do Galaxy

https://zenodo.org/record/1409427/files/KO_01_R1.fastq
https://zenodo.org/record/1409427/files/KO_01_R2.fastq
https://zenodo.org/record/1409427/files/KO_02_R1.fastq
https://zenodo.org/record/1409427/files/KO_02_R2.fastq
https://zenodo.org/record/1409427/files/KO_03_R1.fastq
https://zenodo.org/record/1409427/files/KO_03_R2.fastq
https://zenodo.org/record/1409427/files/WT_01_R1.fastq
https://zenodo.org/record/1409427/files/WT_01_R2.fastq
https://zenodo.org/record/1409427/files/WT_02_R1.fastq
https://zenodo.org/record/1409427/files/WT_02_R2.fastq
https://zenodo.org/record/1409427/files/WT_03_R1.fastq
https://zenodo.org/record/1409427/files/WT_03_R2.fastq
https://zenodo.org/record/1409427/files/ensembl_dm3.chr4.gtf

Section 1: Preparation

2) Renomear arquivos:

WT para G1
KO para G3



Prestar atenção nos parâmetros dos programas e na estrutura de dados dos arquivos.

- Formato dos dados de entrada: FASTQ

Verificar se os arquivos de sequencias (“reads”) únicos (R1) ou pareados (“paired-ended”, R1 e R2). Essa informação deve ser informada na etapa de alinhamento.

- Formato do arquivo de anotação gênica: GTF

Tipos de arquivos de dados usados nas análises

- Fastq
- SAM/BAM
- GTF/GFF

Formato FASTQ

Arquivos de saída de sequenciadores, entrada em programas de alinhamento

Cada sequencia é representada por 4 linhas no arquivo de entrada:

```
1. @SEQ_ID
2. GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
3. +
4. !''*(((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

1. Caracter “@” seguido de um identificador único da sequência nucleotídica
2. Sequencia de bases obtidas a partir do sequenciamento
3. Caracter “+” seguido do identificador único da sequência (opcional)
4. Qualidade associada a cada base em ASCII

Formato SAM : Sequence Alignment/Map format

- Arquivo de saida dos alinhadores. Contem as informações referentes ao alinhamento dos reads na sequencia de referência (genoma ou transcritoma)

- O formato BAM é uma versão binária (comprimida) do formato SAM

http://samtools.sourceforge.net/SAMv1.pdf

1.4 The alignment section: mandatory fields

In the SAM format, each alignment line typically represents the linear alignment of a segment. Each line has 11 mandatory fields. These fields always appear in the same order and must be present, but their values can be '0' or '*' (depending on the field) if the corresponding information is unavailable. The following table gives an overview of the mandatory fields in the SAM format:

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,255}`	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0,2^{31}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next read
8	PNEXT	Int	$[0,2^{29}-1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{29}+1,2^{29}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

1. QNAME: Query template NAME. Reads/segments having identical QNAME are regarded to come from the same template. A QNAME '*' indicates the information is unavailable. In a SAM file, a read may occupy multiple alignment lines, when its alignment is chimeric or when multiple mappings are given.

2. FLAG: bitwise FLAG. Each bit is explained in the following table:

Bit	Description
0x1	template having multiple segments in sequencing
0x2	each segment properly aligned according to the aligner
0x4	segment unmapped
0x8	next segment in the template unmapped
0x10	SEQ being reverse complemented
0x20	SEQ of the next segment in the template being reversed
0x40	the first segment in the template
0x80	the last segment in the template
0x100	secondary alignment
0x200	not passing quality controls
0x400	PCR or optical duplicate
0x800	supplementary alignment

Formato GTF (Gene Tranfer Format)/GFF (General Featura Format)

Usados para representar as coordenadas genômicas dos genes anotados no organismo

Arquivo com 9 campos (colunas) que lista anotações de genes mapeados no genoma

Seqname	Source	Feature	Start	End	Score	Strand	Frame	Attributes
chr4	protein_coding	CDS	24053	24477	.	+	0	exon_number "1"; gene_id "FBgn0040037"; gene_name "JYalpha"; p_
chr4	protein_coding	exon	24053	24477	.	+	.	exon_number "1"; gene_id "FBgn0040037"; gene_name "JYalpha"; p_
chr4	protein_coding	CDS	24979	25153	.	+	1	exon_number "2"; gene_id "FBgn0040037"; gene_name "JYalpha"; p_
chr4	protein_coding	exon	24979	25153	.	+	.	exon_number "2"; gene_id "FBgn0040037"; gene_name "JYalpha"; p_
chr4	protein_coding	CDS	25218	25450	.	+	0	exon_number "3"; gene_id "FBgn0040037"; gene_name "JYalpha"; p_
chr4	protein_coding	exon	25218	25450	.	+	.	exon_number "3"; gene_id "FBgn0040037"; gene_name "JYalpha"; p_
chr4	protein_coding	CDS	25501	25618	.	+	1	exon_number "4"; gene_id "FBgn0040037"; gene_name "JYalpha"; p_
chr4	protein_coding	exon	25501	25621	.	+	.	exon_number "4"; gene_id "FBgn0040037"; gene_name "JYalpha"; p_
chr4	protein_coding	stop_codon	25619	25621	.	+	0	exon_number "4"; gene_id "FBgn0040037"; gene_name "JYalpha"; p_
chr4	pseudogene	exon	26994	27101	.	-	.	exon_number "7"; gene_id "FBgn0052011"; gene_name "CR32011";
chr4	pseudogene	exon	27167	27349	.	-	.	exon_number "6"; gene_id "FBgn0052011"; gene_name "CR32011";
chr4	pseudogene	exon	28371	28609	.	-	.	exon_number "5"; gene_id "FBgn0052011"; gene_name "CR32011";

1. cromossomo
2. programa ou fonte de dados
3. anotação
4. coordenada inicial no genoma
5. coordenada final no genoma
6. score
7. fita
8. Fase de leitura
9. Atributos (separados por virgula)

Section 2: Alignment with HISAT2

In this section we map the reads in our FASTQ files to a reference genome. As these reads originate from mRNA, we expect some of them will cross exon/intron boundaries when we align them to the reference genome. We will use HISAT2 to perform our alignment. HISAT2 is a fast, splice-aware, alignment program that is a successor to TopHat2. More information on HISAT2 can be found here.

1. Align the RNA-seq reads to a reference genome.

2. Examine the alignment stats

HISAT2 outputs one bam file for each set of paired-end read files. Rename the 6 files into a more meaningful name (e.g. 'HISAT on data 2 and data 1' to 'g1_01.bam') by using the **pen icon** next to the file.

Section 3: Visualise the aligned reads

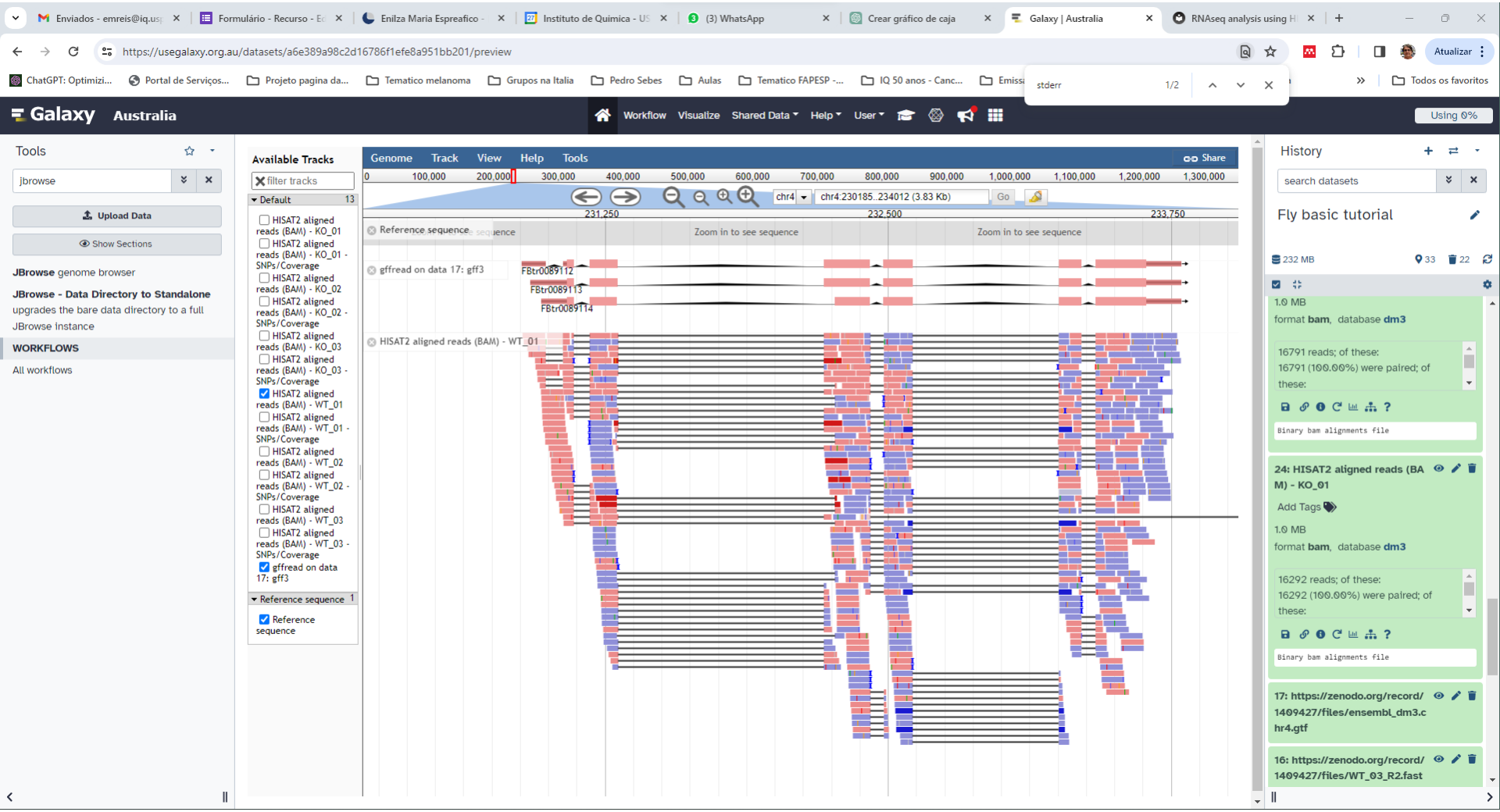
The purpose of this step is to :

- visualise the quantitative, exon-based nature of RNA-seq data
- visualise the expression differences between samples represented by the quantity of reads, and
- become familiar with interactive visualisation tools such as JBrowse and IGV.

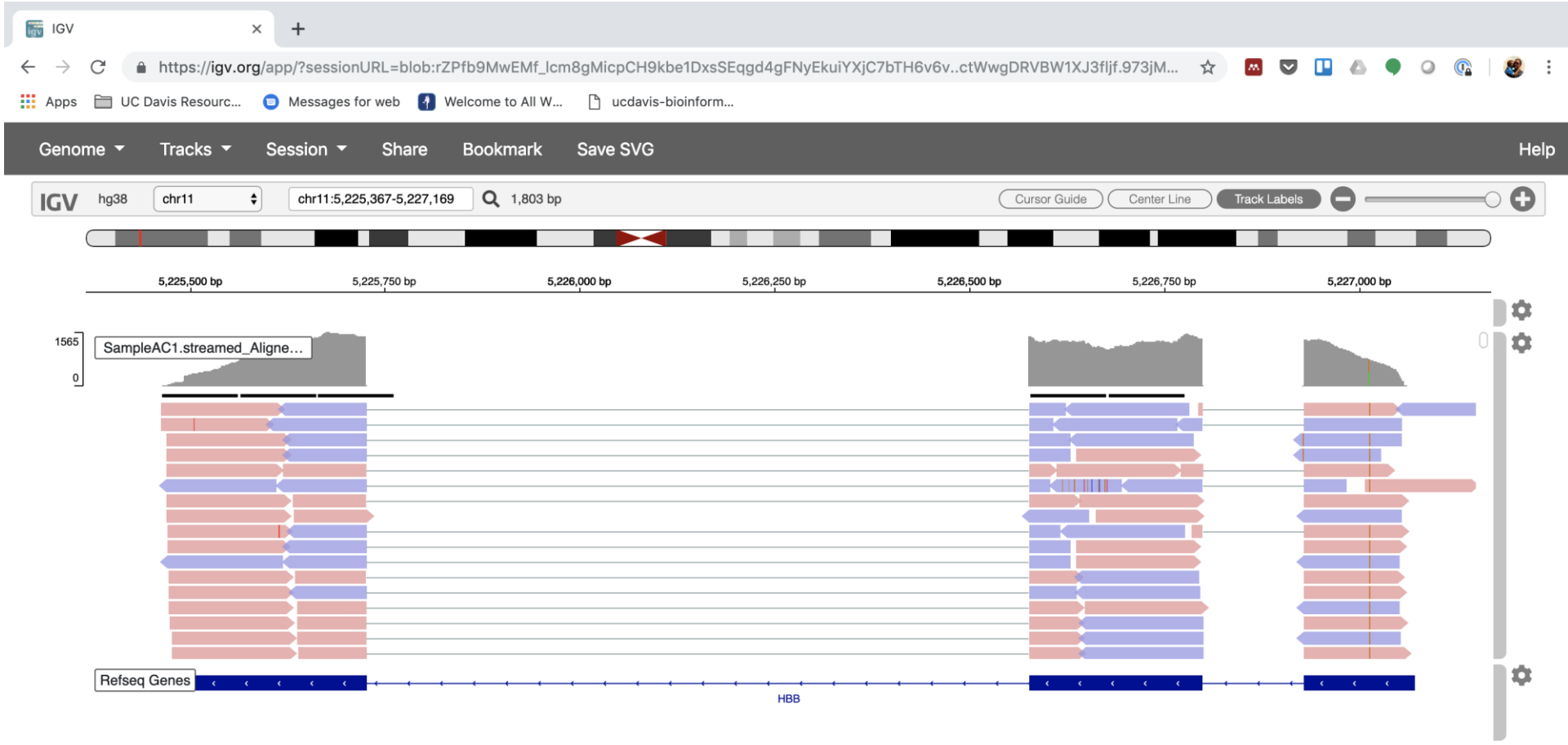
JBrowse and IGV are both interactive tools that can visualise BAM files. You can pick either one to use in this section. JBrowse is run on Galaxy which means you can view your BAM file in your browser, but it takes a while to run the job (~30 mins). IGV is a separate application you'll need to download to your computer and run locally.

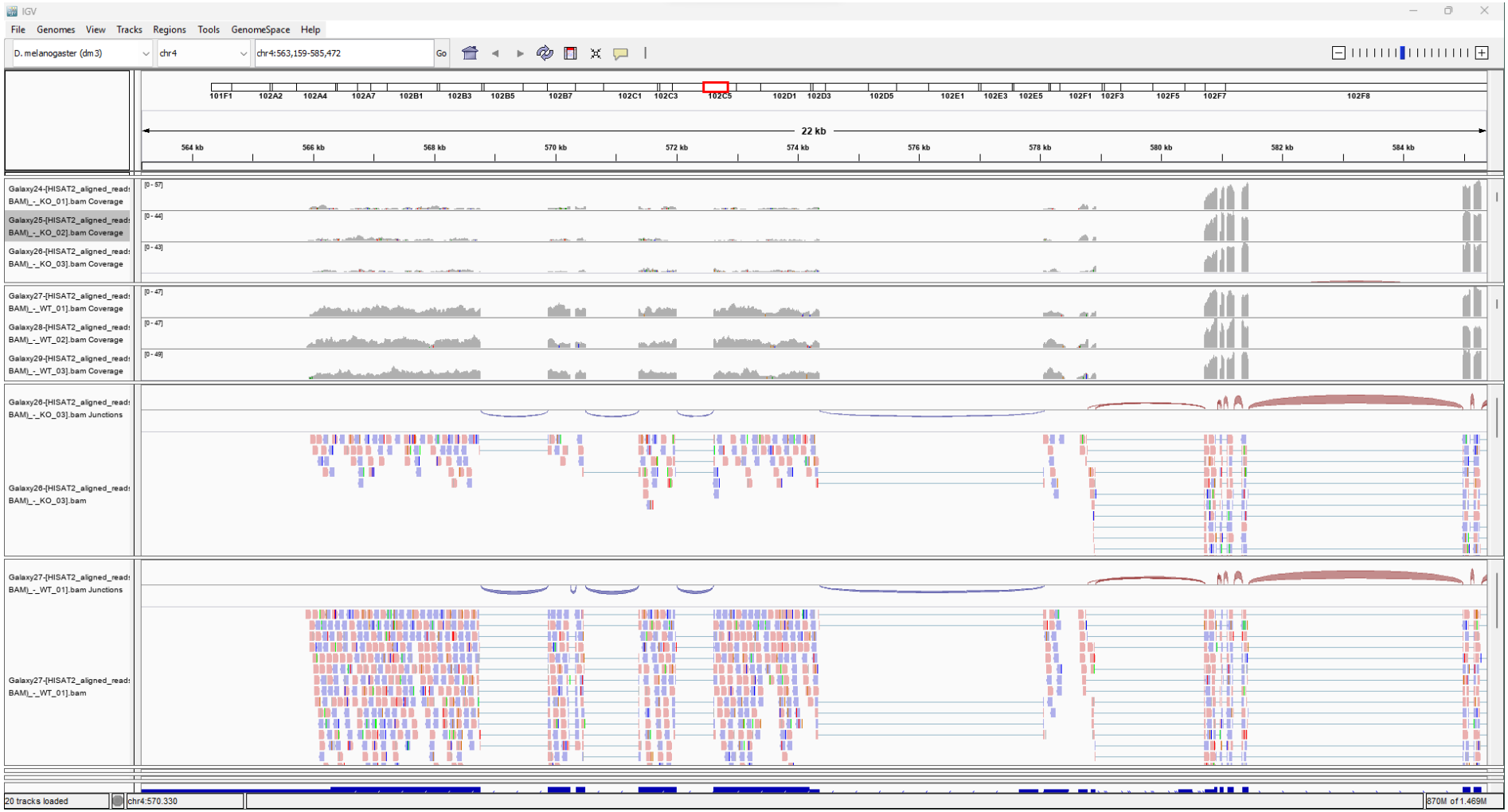
- As informações para uso do IGV estão desatualizadas no tutorial.
- Para visualizar os dados utilizando o browser IGV, utilize a versão online do programa (https://igv.org/app/). Os arquivos *.bam com os alinhamentos precisam ser baixados localmente no seu computador e carregados no IGV
- Pode usar a visualização no Jbrowse,do Galaxy como alternativa

Visualização dos reads alinhados no genoma no Jbrowse (Galaxy)

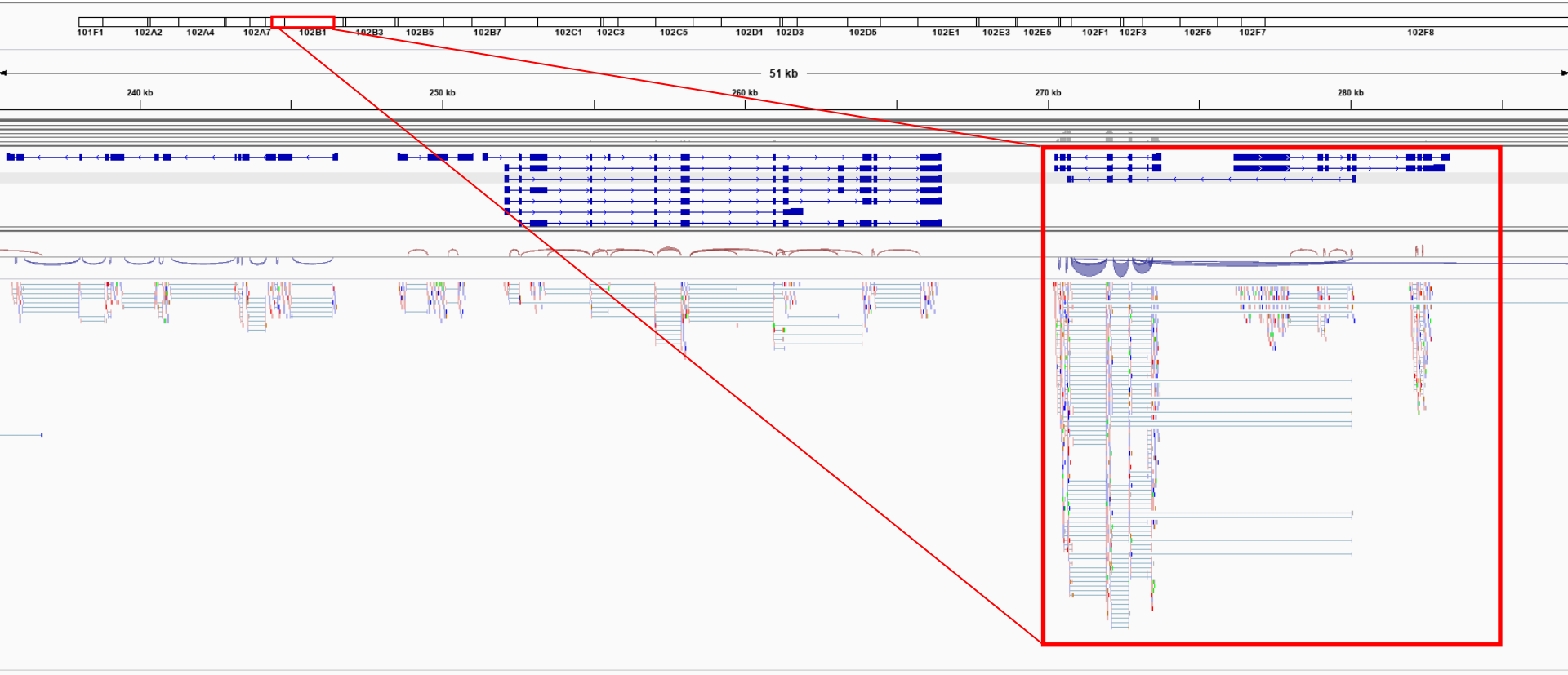


Integrative Genome Viewer
https://igv.org/app/



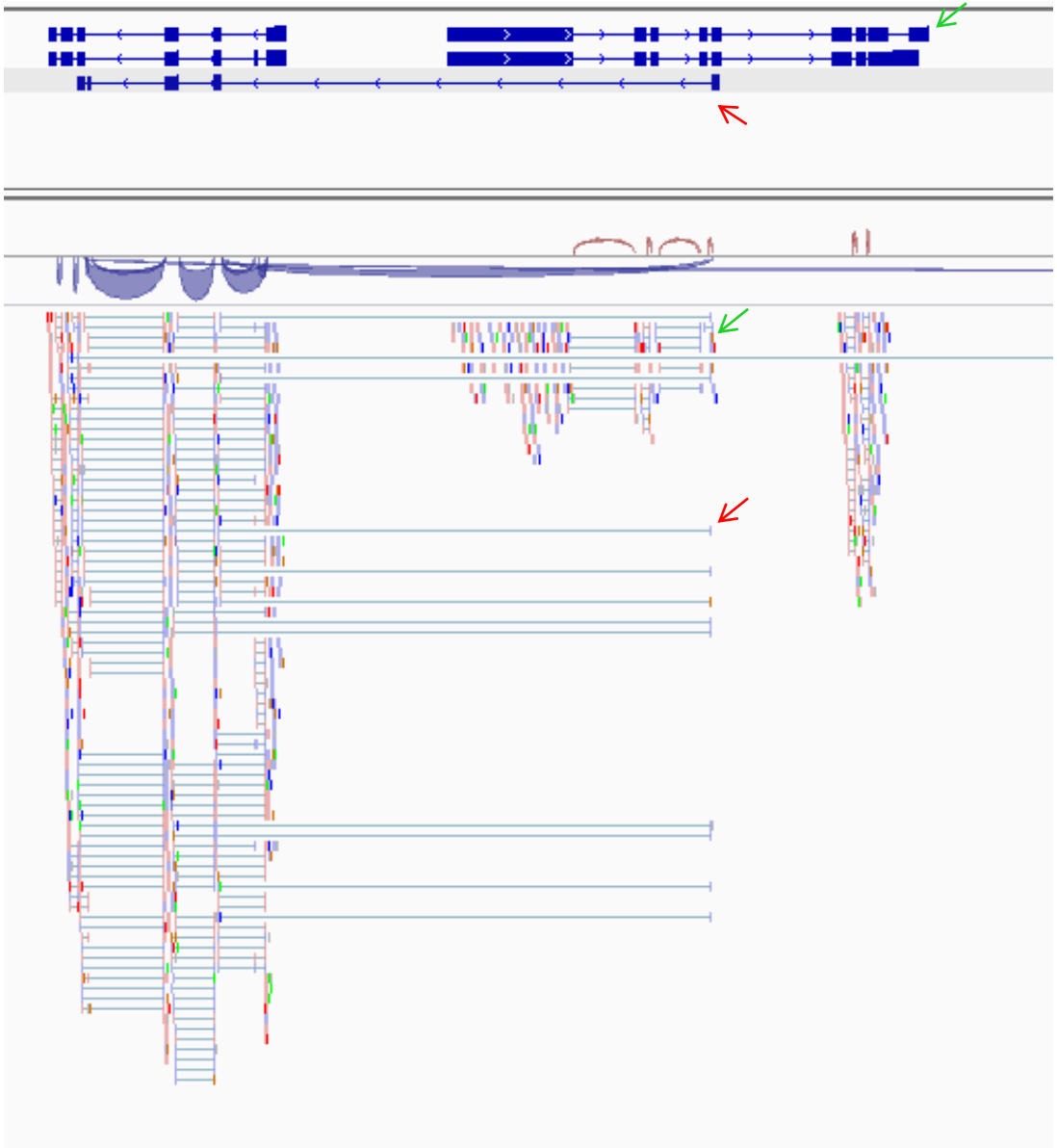

IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
D. melanogaster (dm3)
chr4
chr4:563,159-585,472
Go
101F1 102A2 102A4 102A7 102B1 102B3 102B5 102B7 102C1 102C3 102C5 102D1 102D3 102D5 102E1 102E3 102E5 102F1 102F3 102F5 102F7 102F8
22 kb
564 kb 566 kb 568 kb 570 kb 572 kb 574 kb 576 kb 578 kb 580 kb 582 kb 584 kb
Galaxy24-[HISAT2_aligned_reads BAM)_-_KO_01].bam Coverage
Galaxy25-[HISAT2_aligned_reads BAM)_-_KO_02].bam Coverage
Galaxy26-[HISAT2_aligned_reads BAM)_-_KO_03].bam Coverage
Galaxy27-[HISAT2_aligned_reads BAM)_-_WT_01].bam Coverage
Galaxy28-[HISAT2_aligned_reads BAM)_-_WT_02].bam Coverage
Galaxy29-[HISAT2_aligned_reads BAM)_-_WT_03].bam Coverage
Galaxy26-[HISAT2_aligned_reads BAM)_-_KO_03].bam Junctions
Galaxy26-[HISAT2_aligned_reads BAM)_-_KO_03].bam
Galaxy27-[HISAT2_aligned_reads BAM)_-_WT_01].bam Junctions
Galaxy27-[HISAT2_aligned_reads BAM)_-_WT_01].bam
20 tracks loaded
chr4:570,330
870M of 1,469M

Visualização de reads alinhados no genoma (IGV)



Visualização de reads alinhados no genoma (IGV)

Expressão relativa de variantes de splicing em amostras/tecidos distintos



Section 4. Quantification

HTSeq-count counts the number of the reads from each bam file that map to the genomic features in the provided annotation file. For each feature (a gene for example) we will obtain a numerical value associated with the expression of that feature in our sample (i.e. the number of reads that were aligned to that gene).

1. Examine the GTF file

Click on the **eye icon** to display the ensembl_dm3.chr4.gtf file in Galaxy.

This GTF file is essentially a list of chromosomal features which together define genes. Each feature is in turn defined by a chromosomal start and end point, feature type (CDS, gene, exon etc), and parent gene and transcript. Importantly, a gene may have many features, but one feature will belong to only one gene. More information on the GTF format can be found here.

Seqname	Source	Feature	Start	End	Score	Strand	Frame	Attributes
chr4	protein_coding	CDS	24053	24477	.	+	0	exon_number "1"; gene_id "FBgn0040037"; gene_name "JYalpha"; p
chr4	protein_coding	exon	24053	24477	.	+	.	exon_number "1"; gene_id "FBgn0040037"; gene_name "JYalpha"; p
chr4	protein_coding	CDS	24979	25153	.	+	1	exon_number "2"; gene_id "FBgn0040037"; gene_name "JYalpha"; p
chr4	protein_coding	exon	24979	25153	.	+	.	exon_number "2"; gene_id "FBgn0040037"; gene_name "JYalpha"; p
chr4	protein_coding	CDS	25218	25450	.	+	0	exon_number "3"; gene_id "FBgn0040037"; gene_name "JYalpha"; p
chr4	protein_coding	exon	25218	25450	.	+	.	exon_number "3"; gene_id "FBgn0040037"; gene_name "JYalpha"; p
chr4	protein_coding	CDS	25501	25618	.	+	1	exon_number "4"; gene_id "FBgn0040037"; gene_name "JYalpha"; p
chr4	protein_coding	exon	25501	25621	.	+	.	exon_number "4"; gene_id "FBgn0040037"; gene_name "JYalpha"; p
chr4	protein_coding	stop_codon	25619	25621	.	+	0	exon_number "4"; gene_id "FBgn0040037"; gene_name "JYalpha"; p
chr4	pseudogene	exon	26994	27101	.	-	.	exon_number "7"; gene_id "FBgn0052011"; gene_name "CR32011";
chr4	pseudogene	exon	27167	27349	.	-	.	exon_number "6"; gene_id "FBgn0052011"; gene_name "CR32011";
chr4	pseudogene	exon	28371	28609	.	-	.	exon_number "5"; gene_id "FBgn0052011"; gene_name "CR32011";

The ensembl_dm3.chr4.gtf file contains ~4900 features which together define the 92 known genes on chromosome 4 of Drosophila melanogaster.

2. Run HTSeq-count

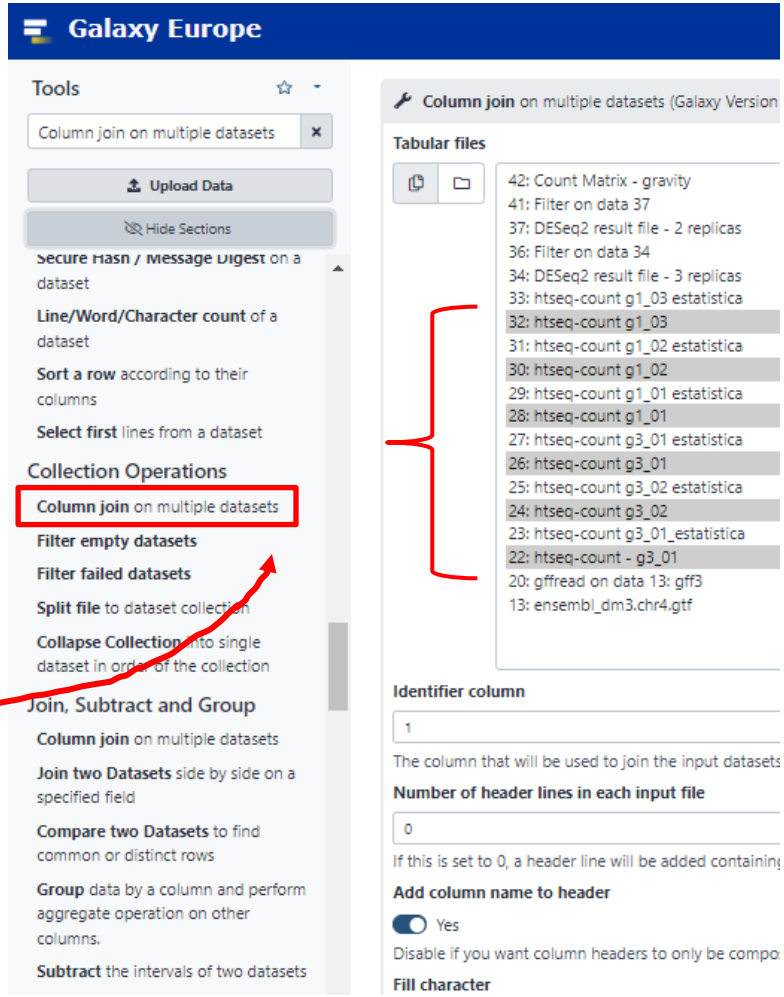
Section 4 – Quantification

“3. Generate a count matrix”.

Para gerar a matriz de contagem, no painél da esquerda, selecione a ferramenta: **“Column join on multiple datasets”.**

Selecione os arquivos com as contagens de reads em cada amostra (G1 ou G3, 3 réplicas cada) obtidas com o HTSeq

Baixar a matriz e contagem gerada e Carregar na ferramenta “Degust”: https://degust.erc.monash.edu/



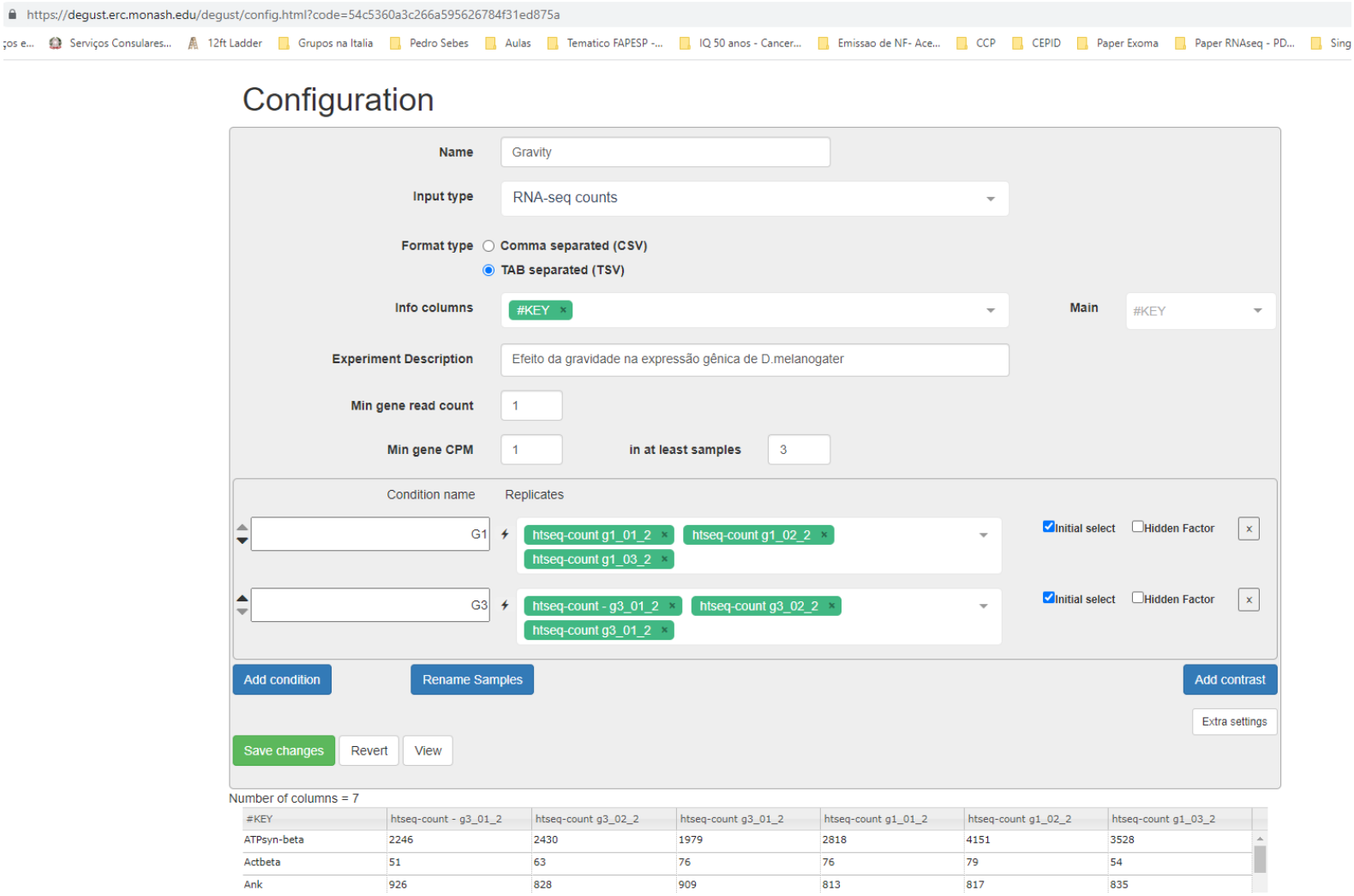
Analise de expressão diferencial

Controle de qualidade com metodos não supervisionados (ex. PCA)

Aplicar métodos de visualização e analise estatística de dados para Identificar genes com expressão (abundancia) alterada entre condições.

Ex. gravidade normal (G1) vs. hipergravidade (G3)

- Salvar localmente a matriz de contagem
- Carregar o arquivo em https://degust.erc.monash.edu/, e selecionar as replicas de cada condição



Section 5. DESeq2 ¶

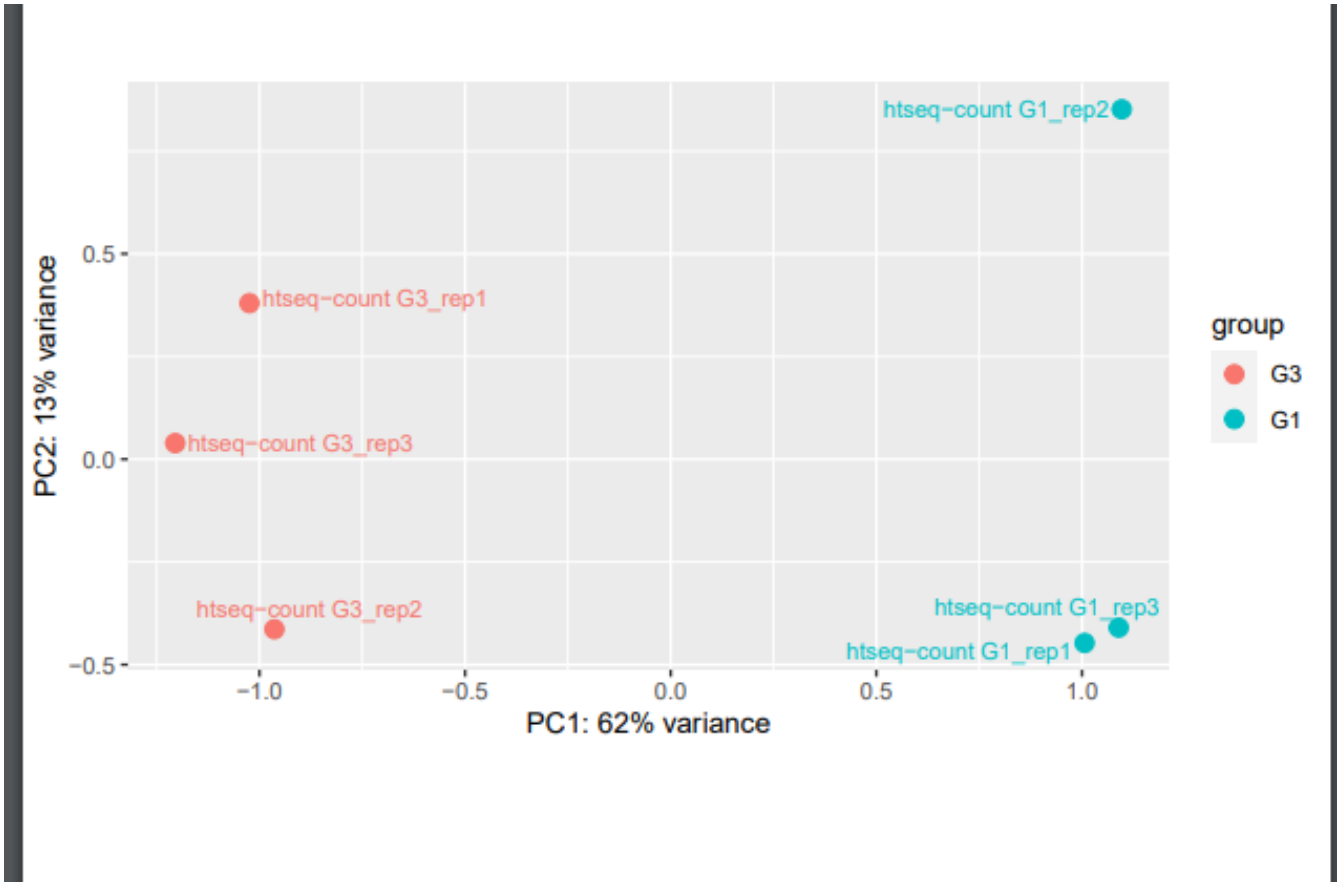
In this section we'll use the "DESeq2" tool in Galaxy to do our differential gene analysis. This tool uses the separate HTSeq files we generated in section 4.

Similar to Voom/Limma or edgeR that was used in Degust to statistically test our data, DESeq2 will:

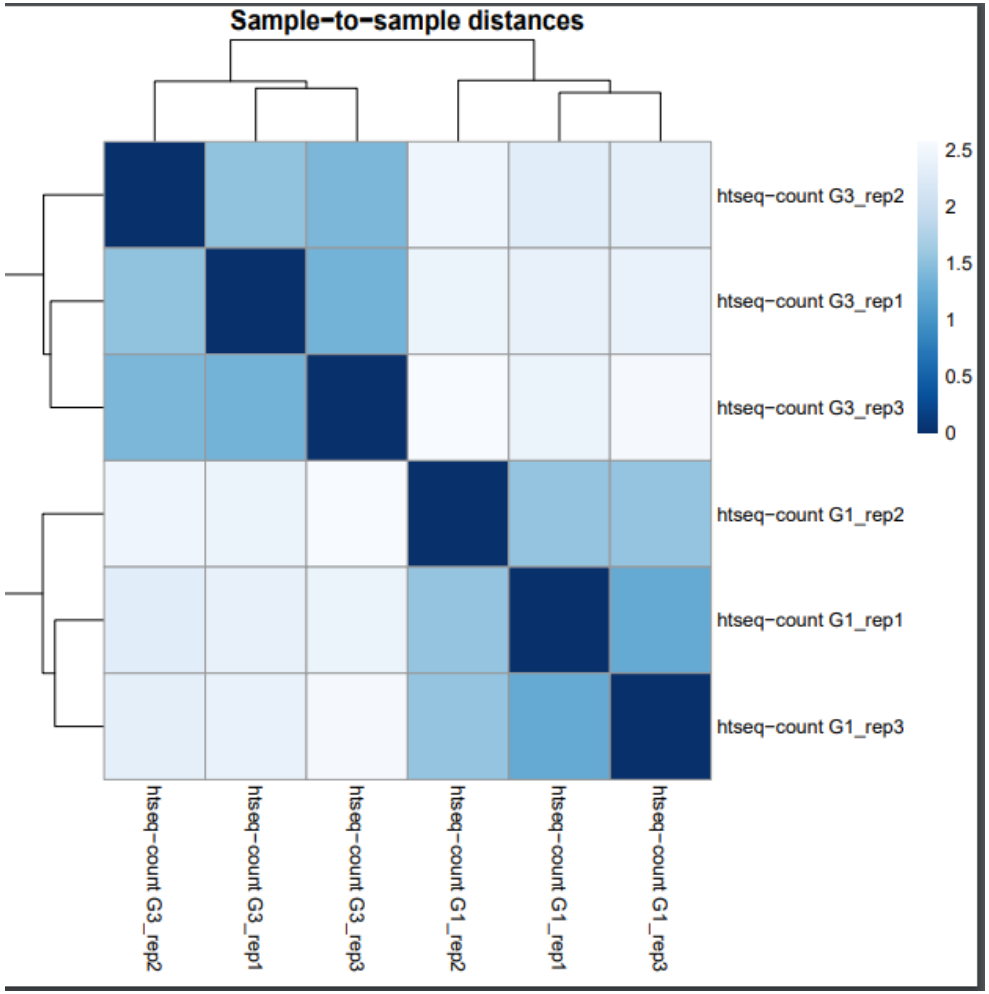
- statistically test for expression differences in normalised read counts for each gene, taking into account the variance observed between samples,
- for each gene, calculate the p-value of the gene being differentially expressed– this is the probability of seeing the data or something more extreme given the null hypothesis (that the gene is not differentially expressed between the two conditions),
- for each gene, estimate the fold change in expression between the two conditions.

Saida do DESeq2
Análise de Componentes Principais

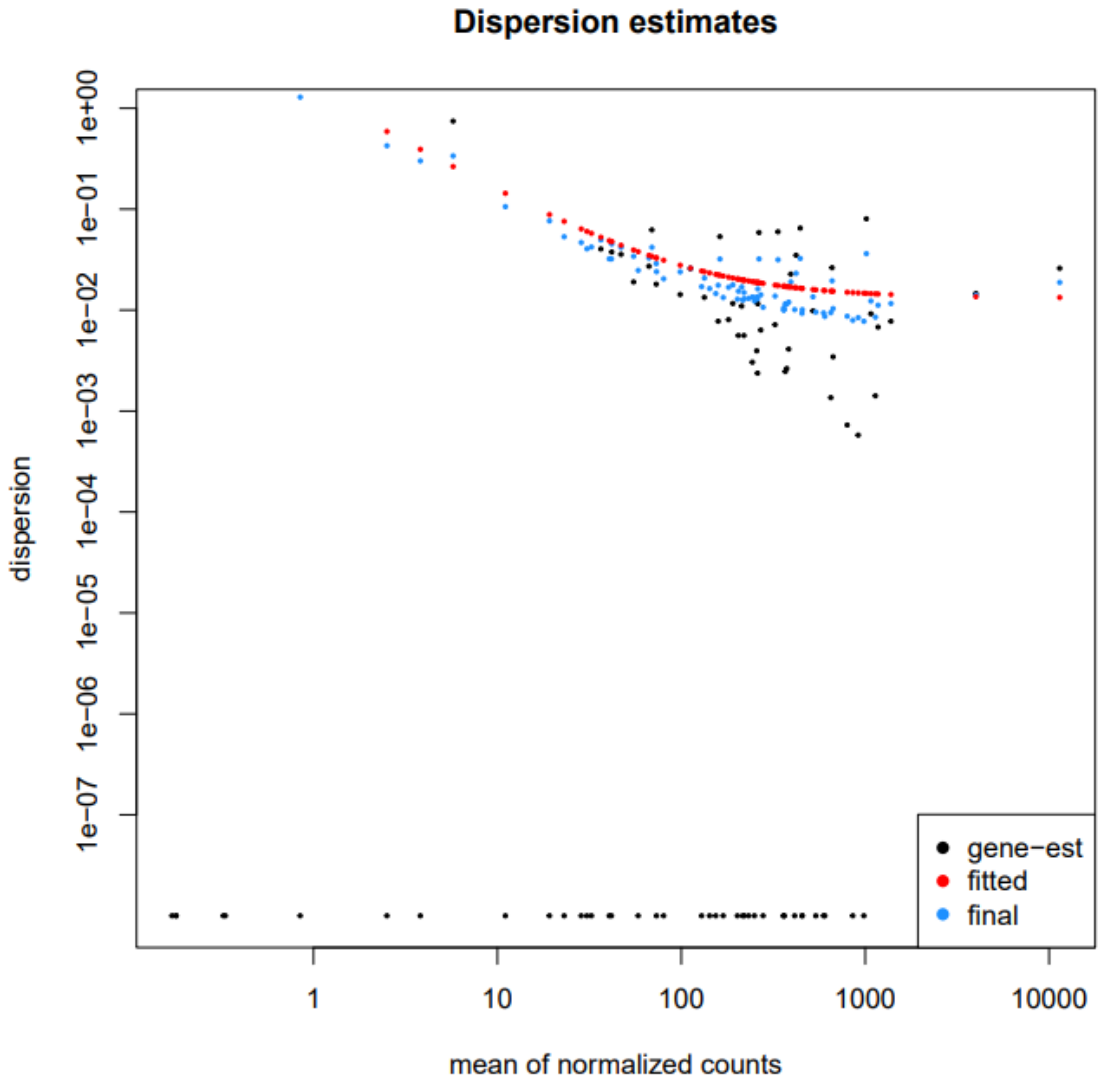
Técnica de redução de dimensionalidade dos dados de expressão gênica.
A expressão de todos os genes detectados (N vetores) é projetada em nas duas dimensões que melhor representam a variabilidade dos dados (PC1 e PC2, eixos X e y) .



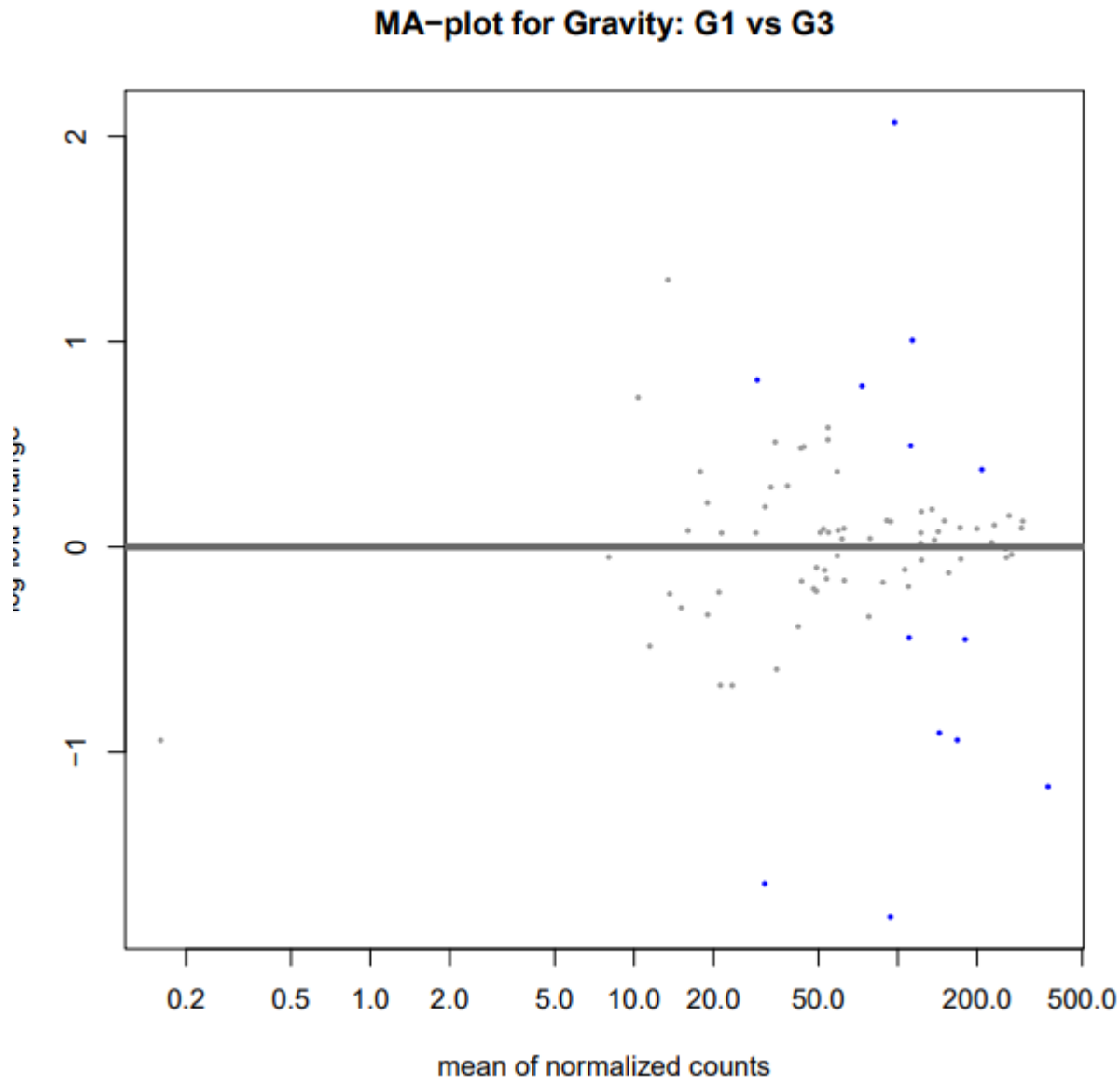
DESEq2: Correlação entre expressão gênica nas amostras



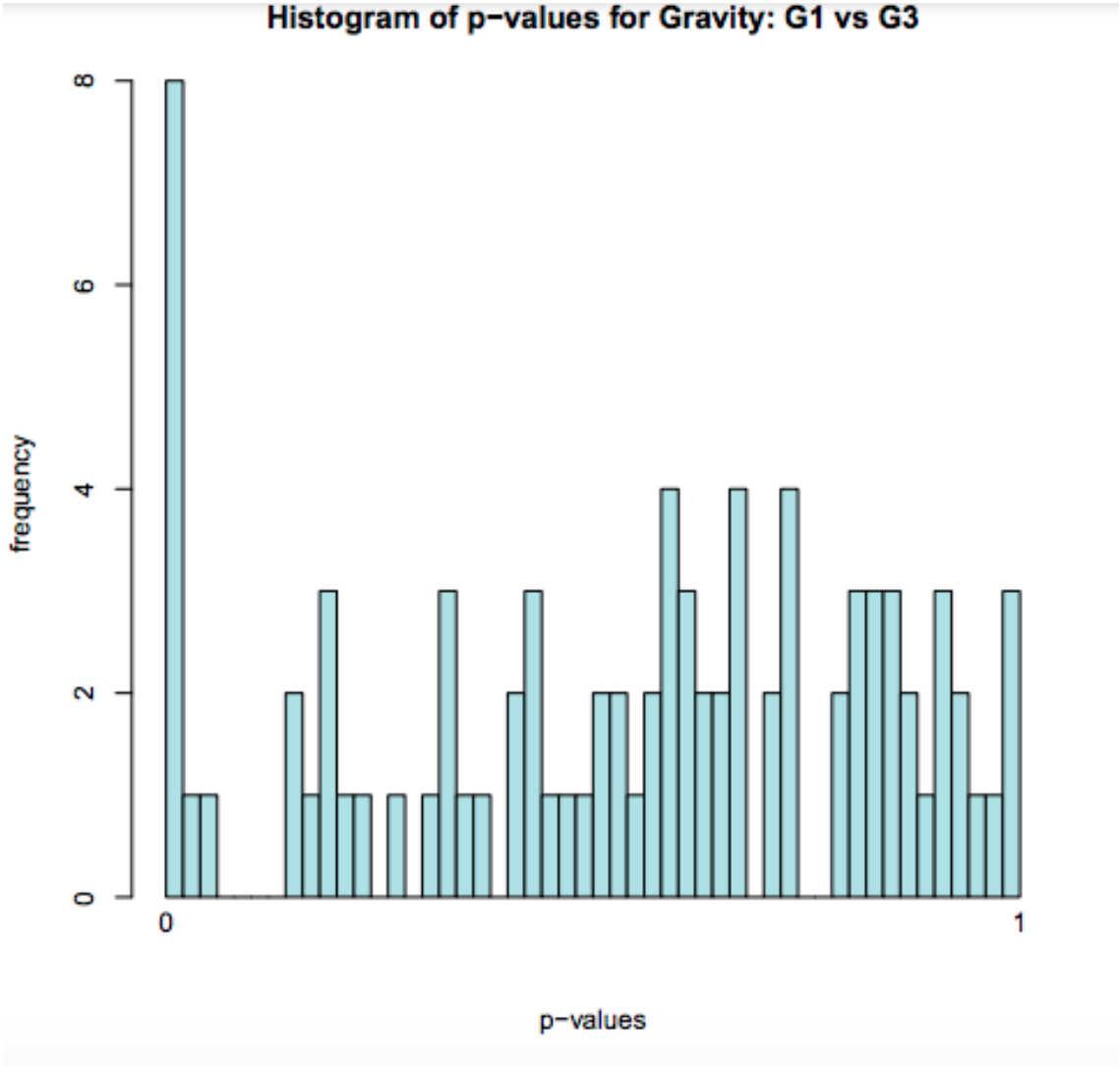
Precisão da medida da expressão gênica proporcional a contagem de reads



Gráficos M (log da razão entre a expressão nas duas condições) vs. A (intensidade média da expressão nas amostras)



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



Tutorial avançado

4. Explore the demo data

Degust also provides an example dataset with 4 conditions and more genes. You can play with the demo dataset by clicking on the “Try the demo” button on the Degust homepage. The demo dataset includes a column with an EC number for each gene. This means genes can be displayed on Kegg pathways using the module on the right.

5. Explore the full dataset

The FASTQ files we started with is only a small proportion of the full dataset. If you wish, you can download the full count matrix here, upload it to Degust, and explore the results.

Instruções para confecção do relatório:

- Descrever sucintamente os passos realizados em cada etapa do tutorial , indicando: o objetivo da etapa de análise, o tipo de dado de entrada, o tipo de dado de saída.
- Responder as perguntas abaixo. Podem usar nas respostas as imagens geradas na analise.
- O relatório deve ser entregue através do e-disciplinas.

1. A partir da inspeção do alinhamento dos reads no genoma usando o IGV ou JBrowse, selecione um gene com evidencia de splicing alternativo. Faça um print da imagem e explique porque selecionou esse gene.
2. Quantos genes diferencialmente expressos (DEGs) foram encontrados nas moscas submetidas a hipergravidade (G3) em comparação as mantidas em gravidade normal (G1) considerando um pvalor ajustado < 0.05 ? Quantos genes estão aumentados e quantos estão diminuídos nas larvas mantidas em hipergravidade?
3. O número de DEGs é o mesmo usando diferentes métodos estatísticos (Deseq2, Limma, EdgeR) ? Existe sobreposição entre as listas de genes?
4. Quantos genes se mantem significativamente alterados se a análise for realizada utilizando apenas 2 replicas experimentais de cada condição (escolha duas replicas ao acaso) ?
5. Compare os gráficos das analises geradas pelo programa DEseq2 usando 3 replicas ou 2 replicas em cada grupo. O que se observa em relação a distribuição dos pvalores encontrados para a expressão diferencial dos genes entre as duas condições?