PRAGUI – Quick tutorial

PRAGUI (Pipeline for RNAseq Analysis with GUI) is a python3 pipelining tool which allows the processing of fastq files containing read sequences from an RNA-seq experiment in order to ultimately perform differential expression analysis.

PRAGUI makes use of several freely available softwares:

|  |  |  |
| --- | --- | --- |
| Software | Version (>=) | Role |
| TrimGalore! | Cutadapt >=1.14  0.4.4 | Quality and adapter trimming |
| FastQC | v0.11.5 | Quality assessment of trimmed reads |
| STAR | 2.5.3a\_modified | Read alignment allowing for non-contiguous genomic regions |
| Samtools | 1.3.1 | Filtering of low quality reads (MAPQ below threshold) |
| HTSeq | 0.9.1 | Count number of reads per genomic region (e.g., per gene) |
| DESeq2 | R>=3.4.1  1.14.1 | Exploratory analysis to assess replicate quality.  Differential gene expression analysis. |
| Cufflinks | v2.2.1  cummeRbund >= 2.20.0 | De novo alignment to find novel transcripts.  Count number of reads per genomic region (e.g., per gene).  Exploratory analysis to assess replicate quality.  Differential gene expression analysis. |

1) Input

PRAGUI needs to access fastq files, name them and also group them according to their biological condition (e.g., WT vs KO or Treated vs Untreated). To do so, PRAGUI requires a tab separated file (a samples file) such as:

samples read1 read2 condition

AX6130\_1 /path/to/file/AX61301\_r\_1.fq.gz /path/to/file/AX61301\_r\_2.fq.gz AX6130

AX6130\_2 /path/to/file/AX61302\_r\_1.fq.gz /path/to/file/AX61302\_r\_2.fq.gz AX6130

AX6186\_1 /path/to/file/AX61861\_r\_1.fq.gz /path/to/file/AX61861\_r\_2.fq.gz AX6186

AX6186\_2 /path/to/file/AX61862\_r\_1.fq.gz /path/to/file/AX61862\_r\_2.fq.gz AX6186

for paired-end data. Alternatively for single-end data:

samples read1 read2 condition

AX6130\_1 /path/to/file/AX61301\_r\_1.fq.gz NA AX6130

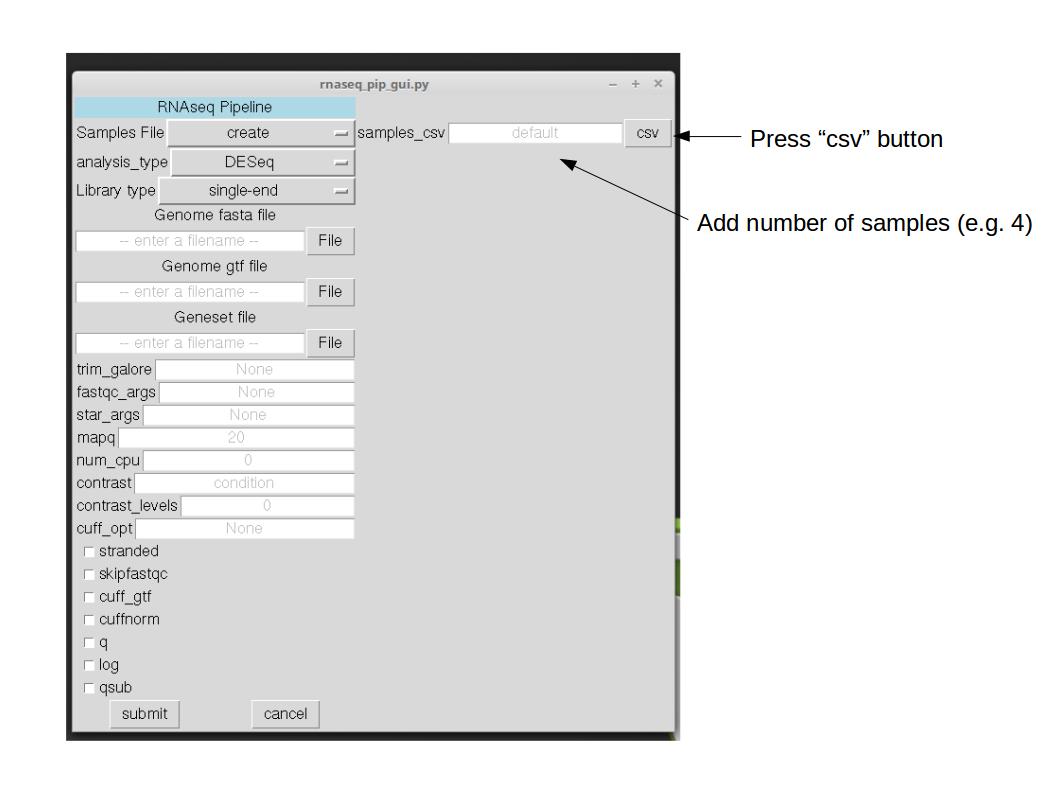
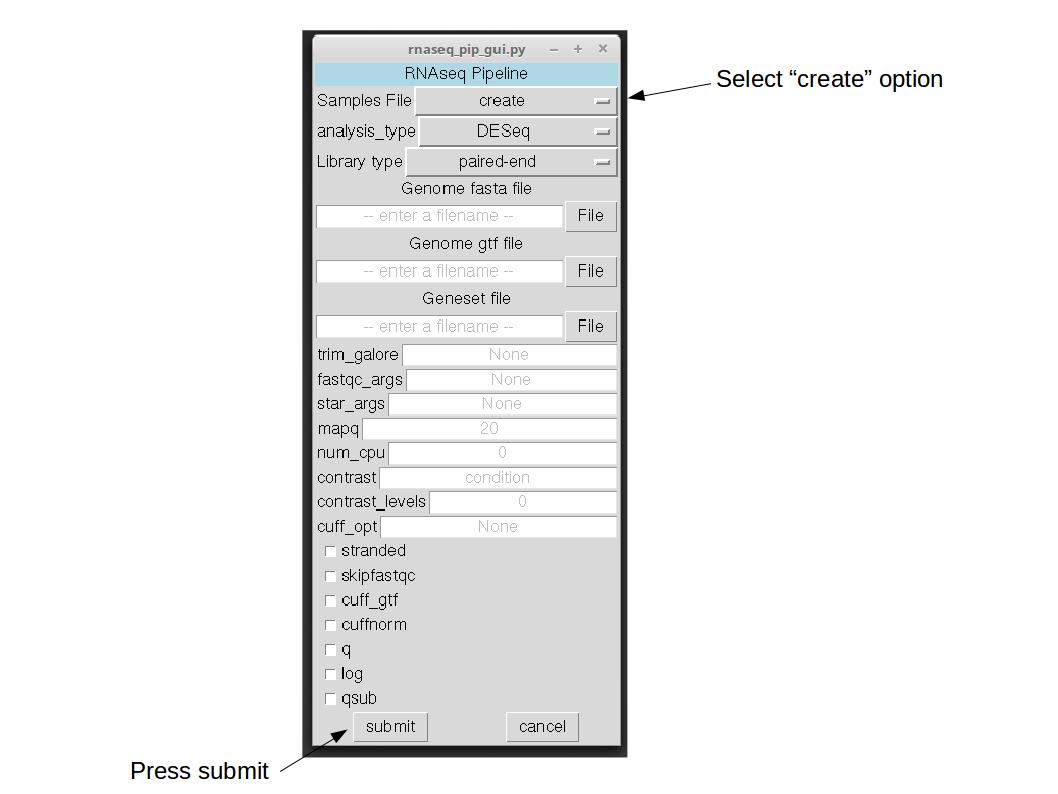
AX6130\_2 /path/to/file/AX61302\_r\_1.fq.gz NA AX6130

AX6186\_1 /path/to/file/AX61861\_r\_1.fq.gz NA AX6186

AX6186\_2 /path/to/file/AX61862\_r\_1.fq.gz NA AX6186

The samples file can be provided by the user or can be created through the GUI.

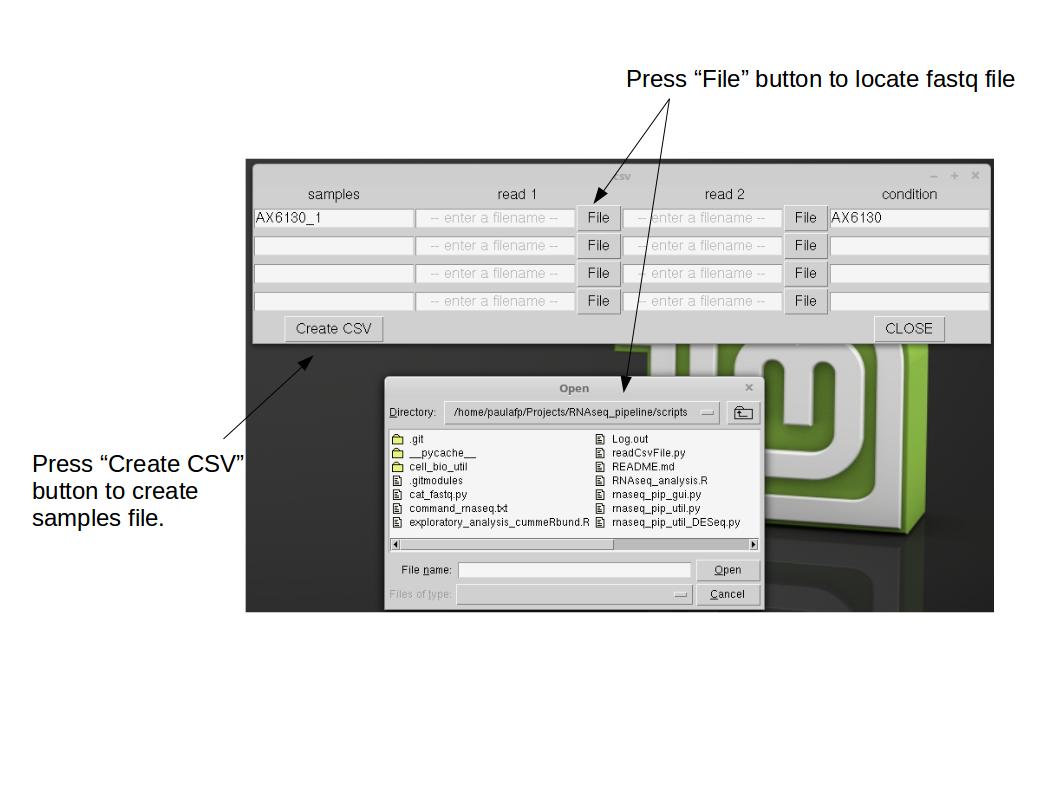
1-a) Create samples file through GUI



A sub window appears where to fill in the information necessary for creating a samples file.



Paths for each fastq file are provided by pressiong the corresponding “File”. A new window appears to allow the selection of file location.

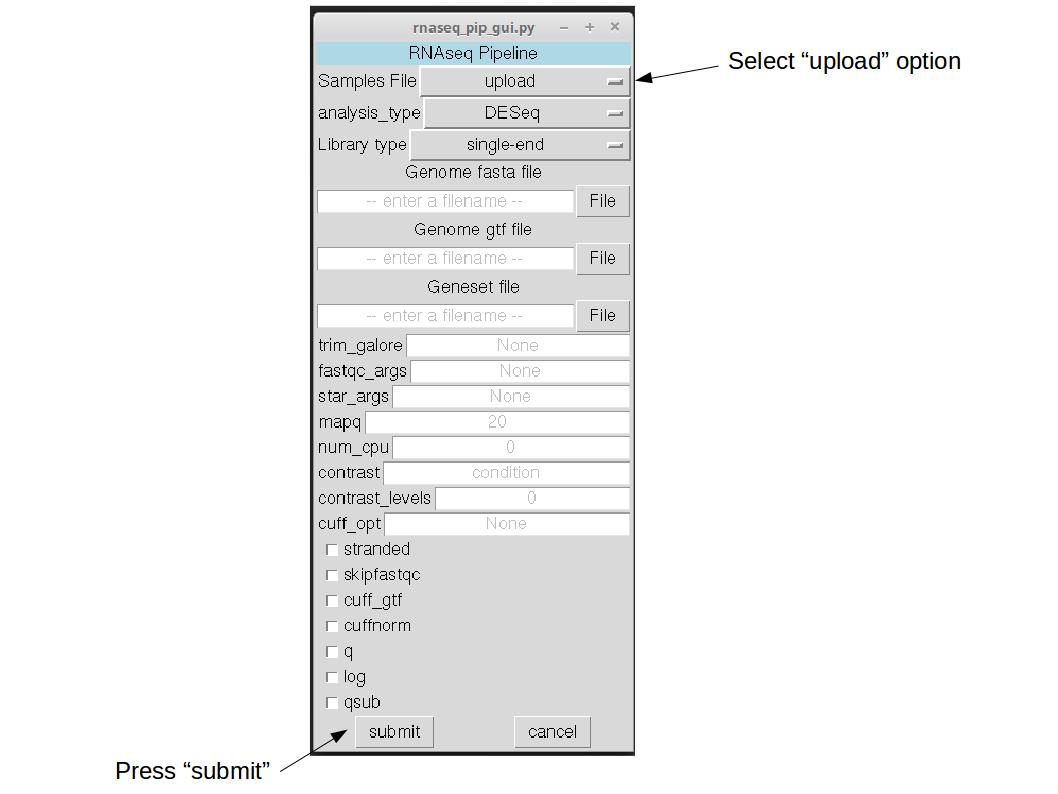


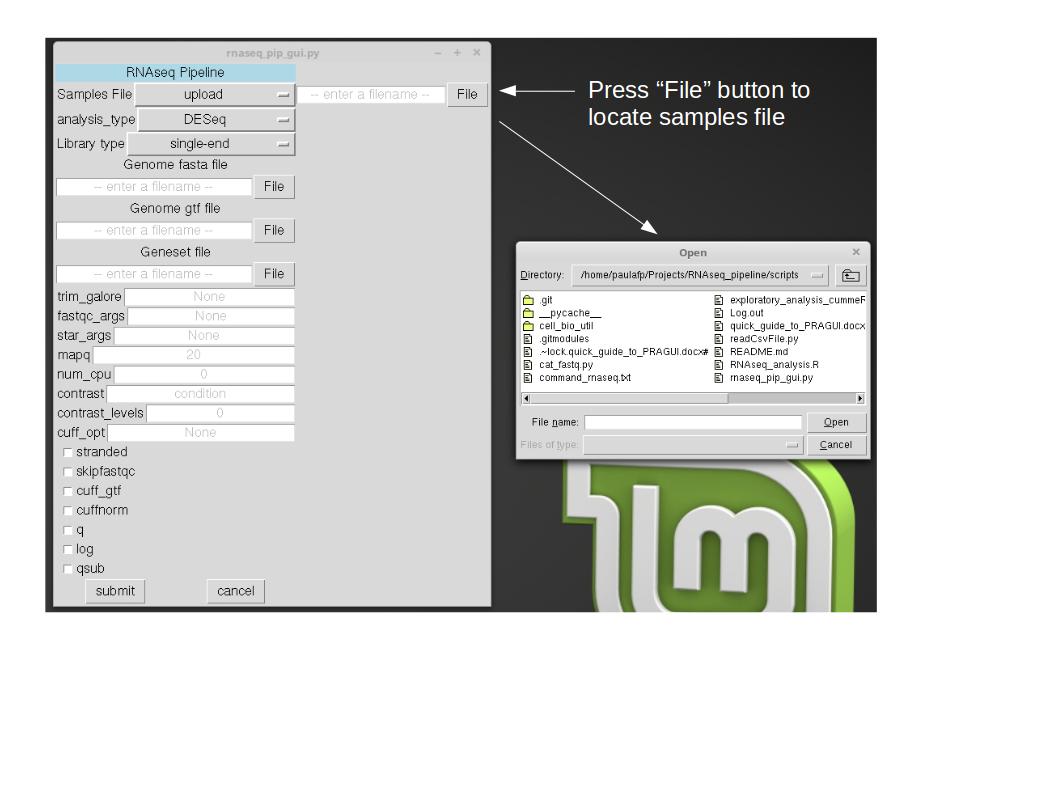
Once all sections have been filled, press “Create CSV”. If all entries get cleaned (window goes back to initial empty state), the samples file has been created. Press “Close” to close this window, it is no longer needed.

NOTE: If your library is single end, please leave “read2” column empty.

1-b) Upload samples file to GUI

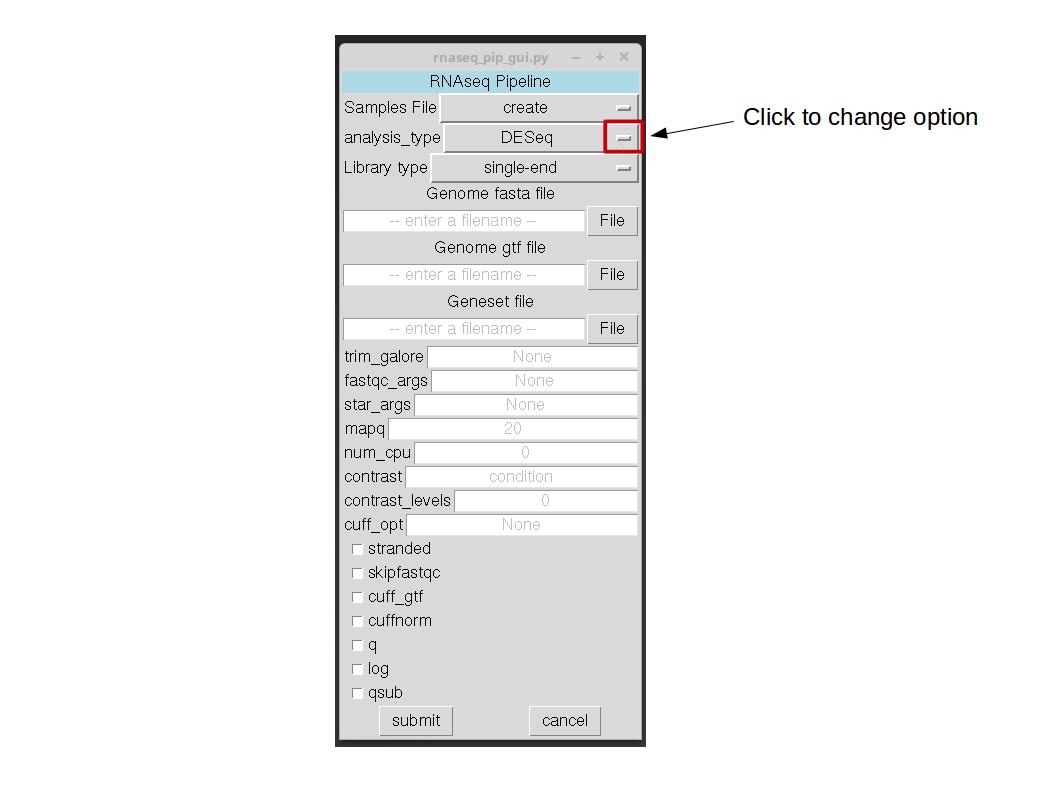
There is also the option to upload a samples file. Sometimes it might be easier to create a file using a text editor or to re-use a file that was generated by a previous run of PRAGUI.





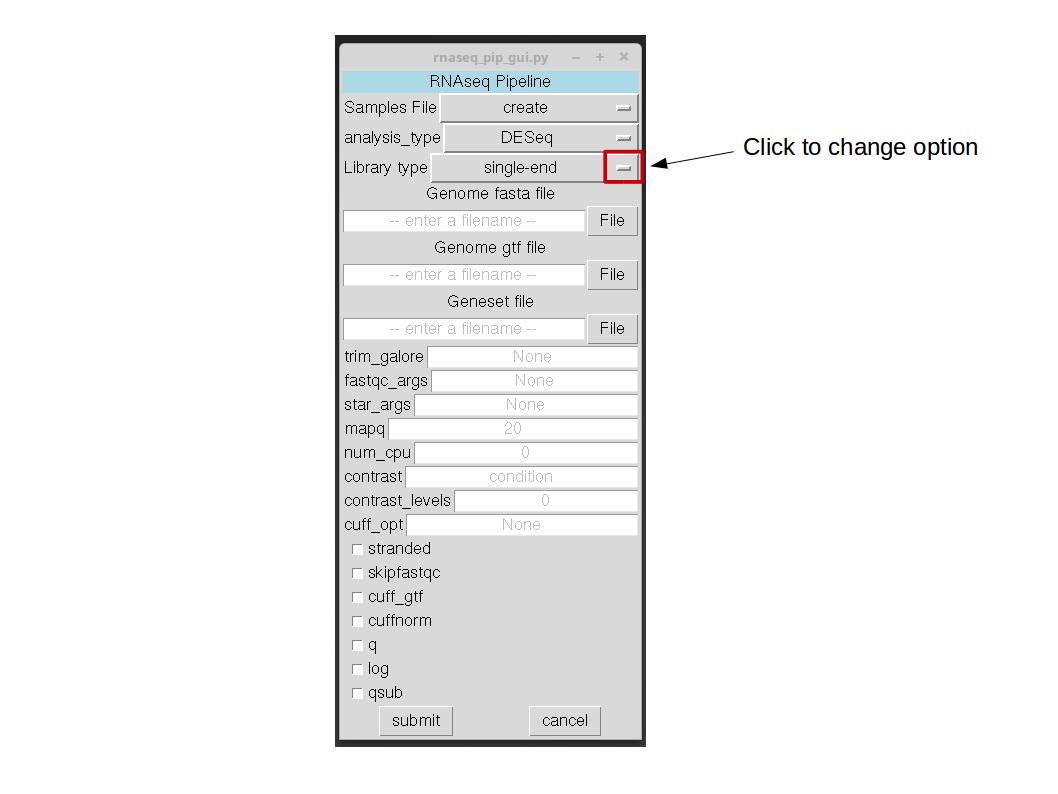
2 – Differential Expression Analysis

PRAGUI can perform differential expression analysis using one of two different statistical tools: DESeq2 or Cufflinks. By default, PRAGUI will run DESeq2 however the user can change this setting by clicking the button “DESeq2” and chosing option cufflinks:



3 – Specify library type

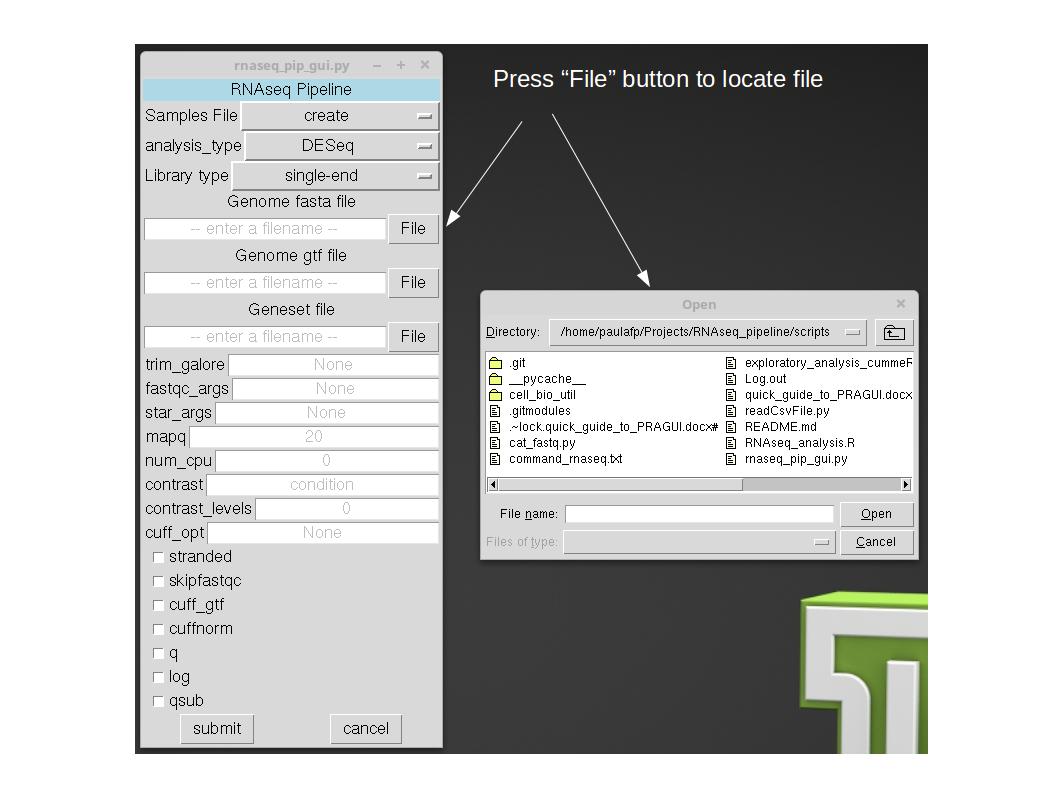
Libraries can be single-end or paired-end (fragments are sequenced from only one end or both). This will affect file processing and therefore needs to be specified. The user can do this by clicking the right-side button in line “Library type”:



4 – Provide Genome Information

4-a) Species reference genome in Fasta format

The reference genome file (in fasta format) is required at the alignment step. It should be specified under the “Genome fasta file” entry:



4-b) Species features file in gtf/gff format

To compute raw read counts over features (e.g. genes or transcripts), PRAGUI needs information about their location in the genome. This must be specified in a gtf/gff file.

If running Cufflinks, this file might be omitted. Cufflinks will then perform a de novo genome assembly and compute read counts over the result.

This file should be provided under the “Genome gtf file” entry, as shown for step 4-a) .

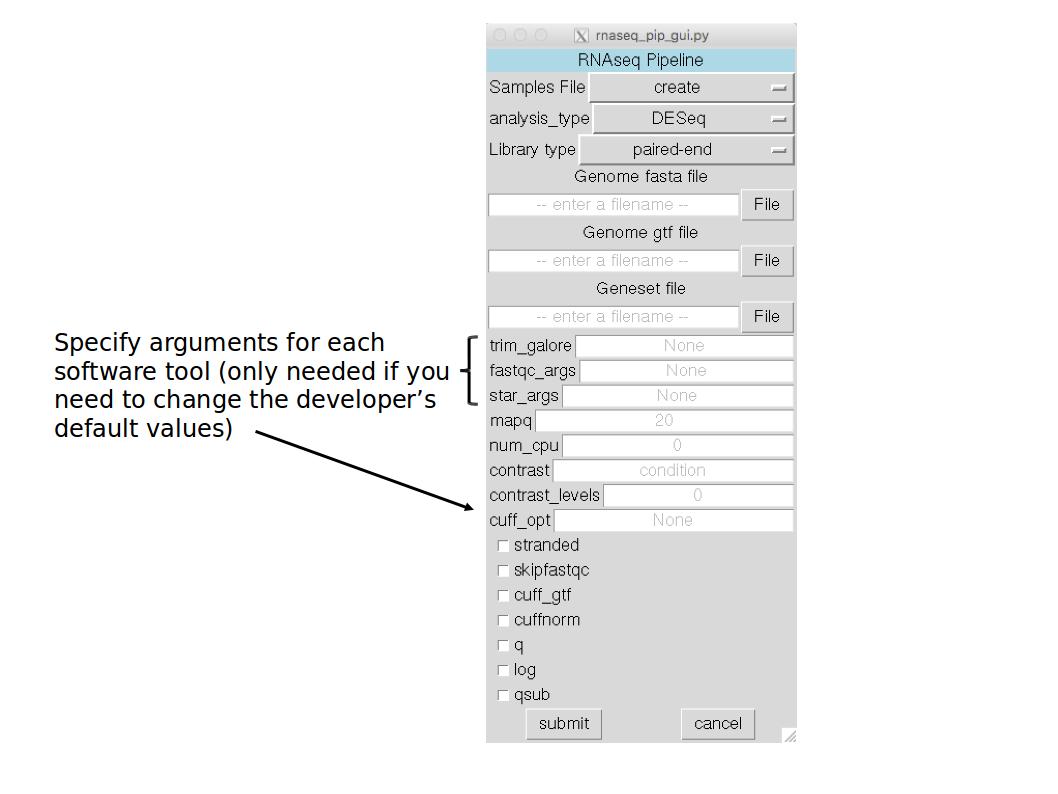
4-c) Alternative Species features file in gtf/gff format

To compute Transcripts Per Million (TPMs) over features, PRAGUI needs information about their location in the genome. This must be specified in a gtf/gff file. By default, PRAGUI will use the file provided in the “Genome gtf file”. However, another file might be provided for this step under the “Geneset file” entry (this can be done as shown for step 4-a).

5) Software options

Each software called by PRAGUI has its own set of user-options. PRAGUI allows the user to specify those options, so that the processing can be costumized according to the user’s needs.

To do so, the user must type their prefered options in the box dedicated to each software: “trim\_galore”, “fastqc\_args”, “star\_args”, “cuff\_opt”.



The options must be specified according to the format required by each software. If unsure, please follow the links:

<https://github.com/FelixKrueger/TrimGalore/blob/master/Docs/Trim_Galore_User_Guide.md>

<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

<https://github.com/alexdobin/STAR/blob/master/doc/STARmanual.pdf>

5) Read quality filter

Poor quality reads can be filtered out based on the MAPping Quality (MAPQ) values generated by STAR. By default, PRAGUI filters out reads with MAPQ<20. This value can be changed by the user in the entry “mapq”.

6) Number of CPUs used by PRAGUI

If you have installed PRAGUI onto your computer, you can specify how many cores PRAGUI can use to run. If this is not specified, PRAGUI will try to use all cores avaliable to speed processing.

To change the default behaviour, the user can specify a number using the entry “num\_cpu”.

NOTE: Please leave this entry empty if running PRAGUI on the cluster.

7) Contrast and Contrast levels

Please leave entries “contrast” and “contrast\_levels” empty. These option are still being implemented and have not been tested.

8) RNA-seq protocol

Depending on the RNA-seq protocol, results may retain or not the information pertaining to strand origin of reads. When the information is kept, the protocol is known as stranded RNAseq.

The user should specify this information by ticking (or not) the “stranded” box.

9) FastQC

PRAGUI authomatically runs FastQC after trimming reads using TrimGalore! . Though this step is recommended, it is not essential for the analysis itself. If the user decides to skip to this step, he/she should tick the “skipfastqc” box.

10) Cufflinks additional options

10-a) Species features file in gtf/gff format

If the user is not interested in finding novel features (i.e. new transcripts), Cufflinks will perform faster if provided with a features file. By ticking the “cuff\_gtf” box, the programm will use the file specified under “Geneset file”, if that file is empty, it will use the one from “Genome gtf file” instead.

NOTE: This option should not be set if "-cuff\_gtf" box already incorporates a gtf file to be used.

10-b) Cuffnorm

By default, PRAGUI will not run Cufflinks tool “Cuffnorm”. The user can run that option by ticking the “cuffnorm” box.

11) Logging

By default, PRAGUI will print the process of each run on the terminal. The user can prevent this from happening by ticking the “q” box.

To save the pipeline’s progress in a log file, the user should tick the “log” box. Though this is not a default option, it is strongly recommended to always tick this option. Otherwise crucial information on how the analysis was generated will be lost.

12) Output

PRAGUI’s output varies depending on which tool was used to perform the differential gene expression analysis. However, PRAGUI always returns three different types of output:

- plots from exploratory analysis

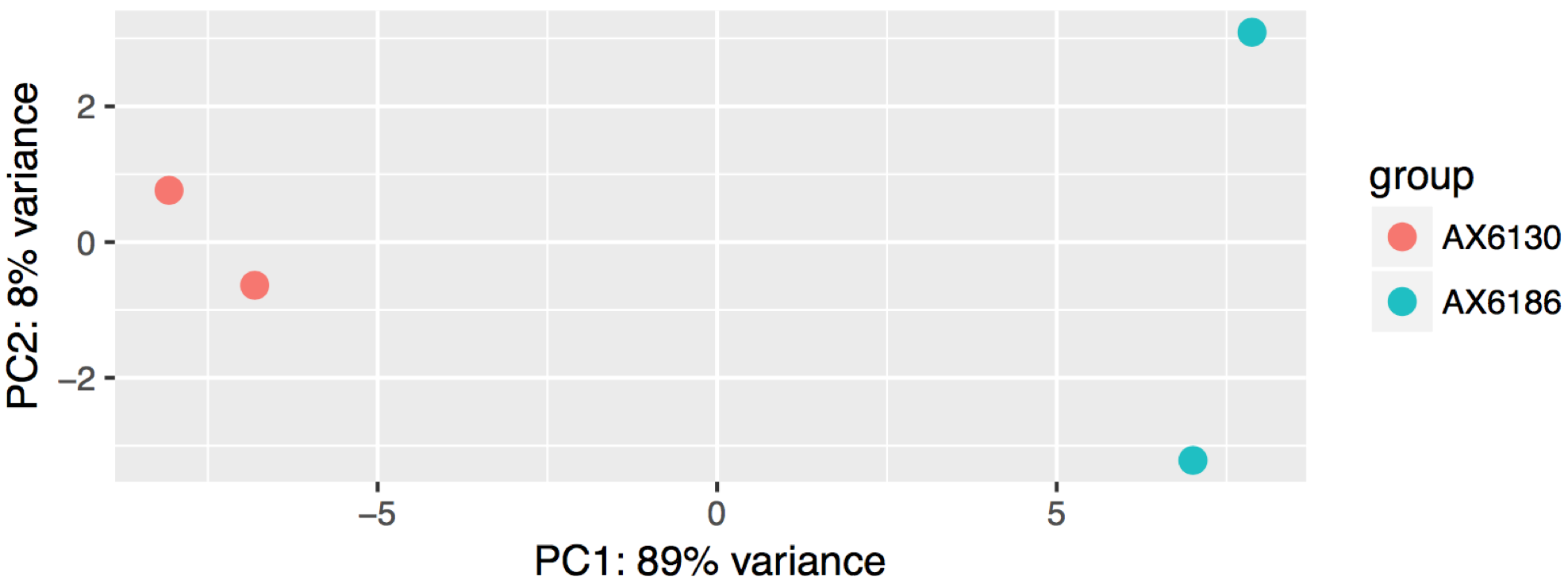
- files with read counts

- files with results from differential expression analysis.

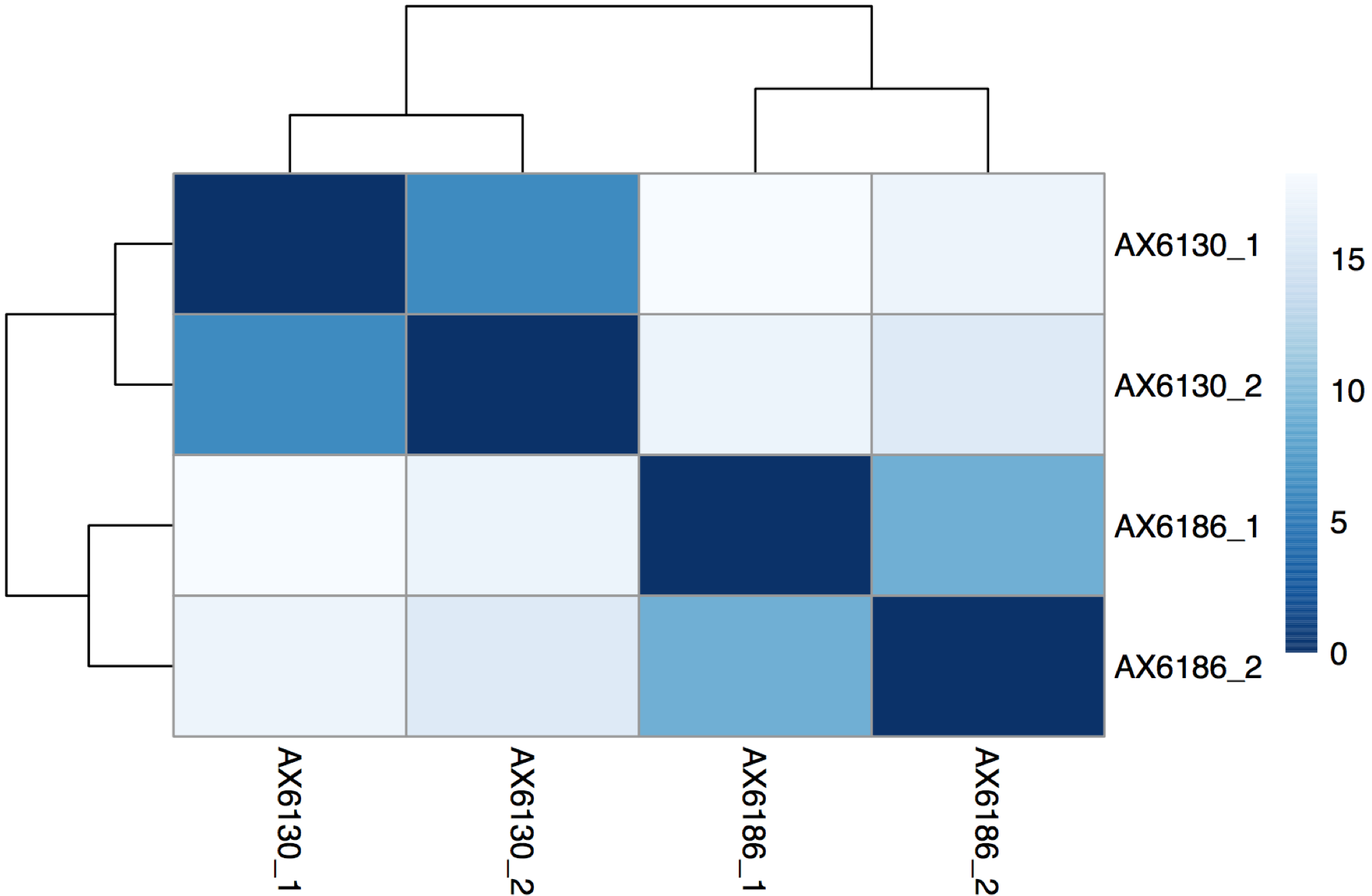
12-a) DESeq2 branch

In order to assess replicate similarity (ie reproducibility), PRAGUI generates a PCA plot and a eucledian sample distance heatmap. Please see examples below:

PCA plot:



Distance Heatmap and dendrogram:



This plots can be found in file XXX\_sclust.pdf

This branch also ouputs a series of files:

i) XXX\_tpm.txt - normalised read counts for each sample in TPMs (Transcript Per Million)

geneName Brain\_I Brain\_III SA\_I SA\_II

ENSMUSG00000000001 105.668741393666 96.1845052420974 96.6721294736383 81.1300724501742

ENSMUSG00000000003 0 0 0 0

ENSMUSG00000000028 9.27276962363166 7.86120331674685 6.8383971237575 10.8227901913994

ENSMUSG00000000031 47.5736143717538 42.7858259493294 99.5364316503101 78.291368530754

ENSMUSG00000000037 2.68646350558992 2.45088437697655 3.62382623242496 2.36372949863057

ENSMUSG00000000049 0.503867319573847 0 0.703801397416308 0.342322751985497

ENSMUSG00000000056 55.0650655823002 56.0200497712557 62.8772648512837 52.4922902804961

ENSMUSG00000000058 27.293946278485 28.5390841778243 31.651206417976 21.4338180309144

ENSMUSG00000000078 88.5291782240821 81.4334381834138 68.4482667602795 35.2922205722676

ii) XXX\_DESeq\_summary.txt – Overall information of the statistical tests performed and overall result statistics

iii) XXX\_DESeq\_results\_4\_peat.txt

test\_stat – Wald statistic

p\_value – Wald test p-value

q\_value – p-value adjusted for multiple testing following the Benjamini-Hochberg method

test\_id gene\_id gene locus sample\_1 sample\_2 status value\_1 value\_2

ENSMUSG00000021848 ENSMUSG00000021848 Otx2 14:48657679\_48667644 Deep Superficial OK 706.85 2592.39

ENSMUSG00000058099 ENSMUSG00000058099 Nfam1 15:82997721\_83033306 Deep Superficial OK 82.59 326.47

ENSMUSG00000032128 ENSMUSG00000032128 Robo3 9:37415669\_37433246 Deep Superficial OK 47.07 384.33

ENSMUSG00000028280 ENSMUSG00000028280 Gabrr1 4:33132521\_33163588 Deep Superficial OK 30.93 225.29

ENSMUSG00000045991 ENSMUSG00000045991 Onecut2 18:64340364\_64398488 Deep Superficial OK 212.45 43.35

ENSMUSG00000030337 ENSMUSG00000030337 Vamp1 6:125215551\_125245964 Deep Superficial OK 4844.90 2091.72

ENSMUSG00000039714 ENSMUSG00000039714 Cplx3 9:57599992\_57606281 Deep Superficial OK 32.25 235.98

ENSMUSG00000045518 ENSMUSG00000045518 Onecut3 10:80494835\_80517276 Deep Superficial OK 158.69 19.43

ENSMUSG00000022054 ENSMUSG00000022054 Nefm 14:68082590\_68124846 Deep Superficial OK 6543.20 1718.54

log2(fold\_change) test\_stat p\_value q\_value significant

-1.88 -15.49 4.05E-54 8.54E-50 yes

-1.99 -14.41 4.71E-47 4.97E-43 yes

-3.02 -14.35 1.00E-46 7.04E-43 yes

-2.88 -14.00 1.54E-44 8.10E-41 yes

2.29 13.91 5.14E-44 2.17E-40 yes

1.21 13.69 1.11E-42 3.92E-39 yes

-2.91 -13.46 2.79E-41 8.41E-38 yes

3.01 13.34 1.45E-40 3.81E-37 yes

1.93 12.81 1.40E-37 3.28E-34 yes

iv) XXX\_DESeq\_norm\_read\_counts.txt - normalised read counts for each sample using DESeq scaling factors

gene\_id SA\_I SA\_II SA\_III DA\_I DA\_II DA\_III

ENSMUSG00000000001 1431.24 1191.69 1525.41 1322.00 1052.32 1215.52

ENSMUSG00000000028 69.90 109.75 77.44 51.10 58.98 75.97

ENSMUSG00000000031 1071.58 836.23 871.96 623.66 272.14 198.56

ENSMUSG00000000037 99.98 64.70 102.90 104.82 94.74 165.75

ENSMUSG00000000049 5.09 2.46 1.06 11.79 0.93 6.91

ENSMUSG00000000056 1371.53 1136.00 1215.66 1145.12 1221.36 1174.08

ENSMUSG00000000058 479.09 321.88 399.92 412.72 277.24 464.45

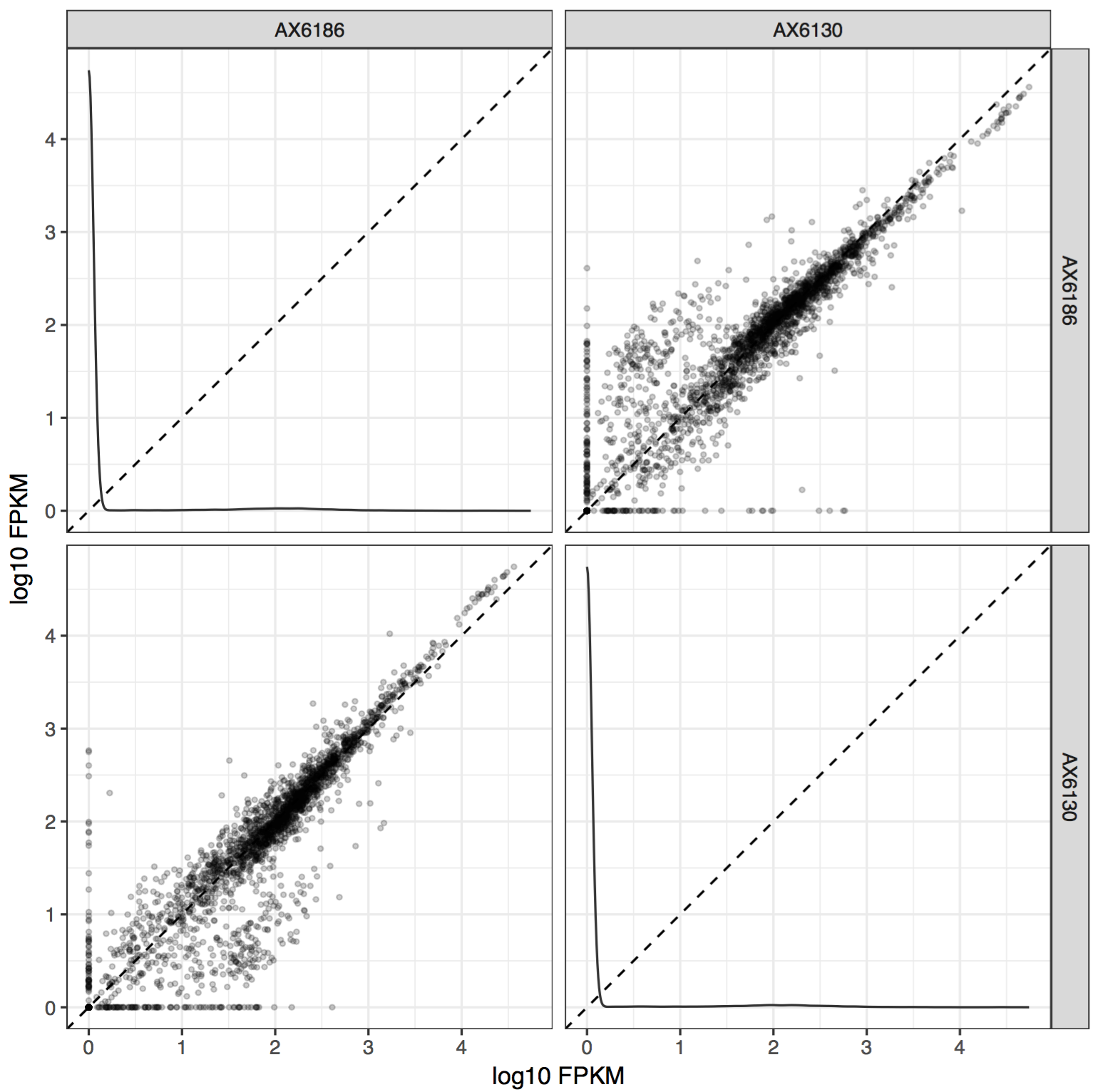
ENSMUSG00000000078 1364.12 697.82 844.38 860.81 819.19 904.74

ENSMUSG00000000085 1333.57 1505.38 1510.56 1472.68 1515.32 1534.94

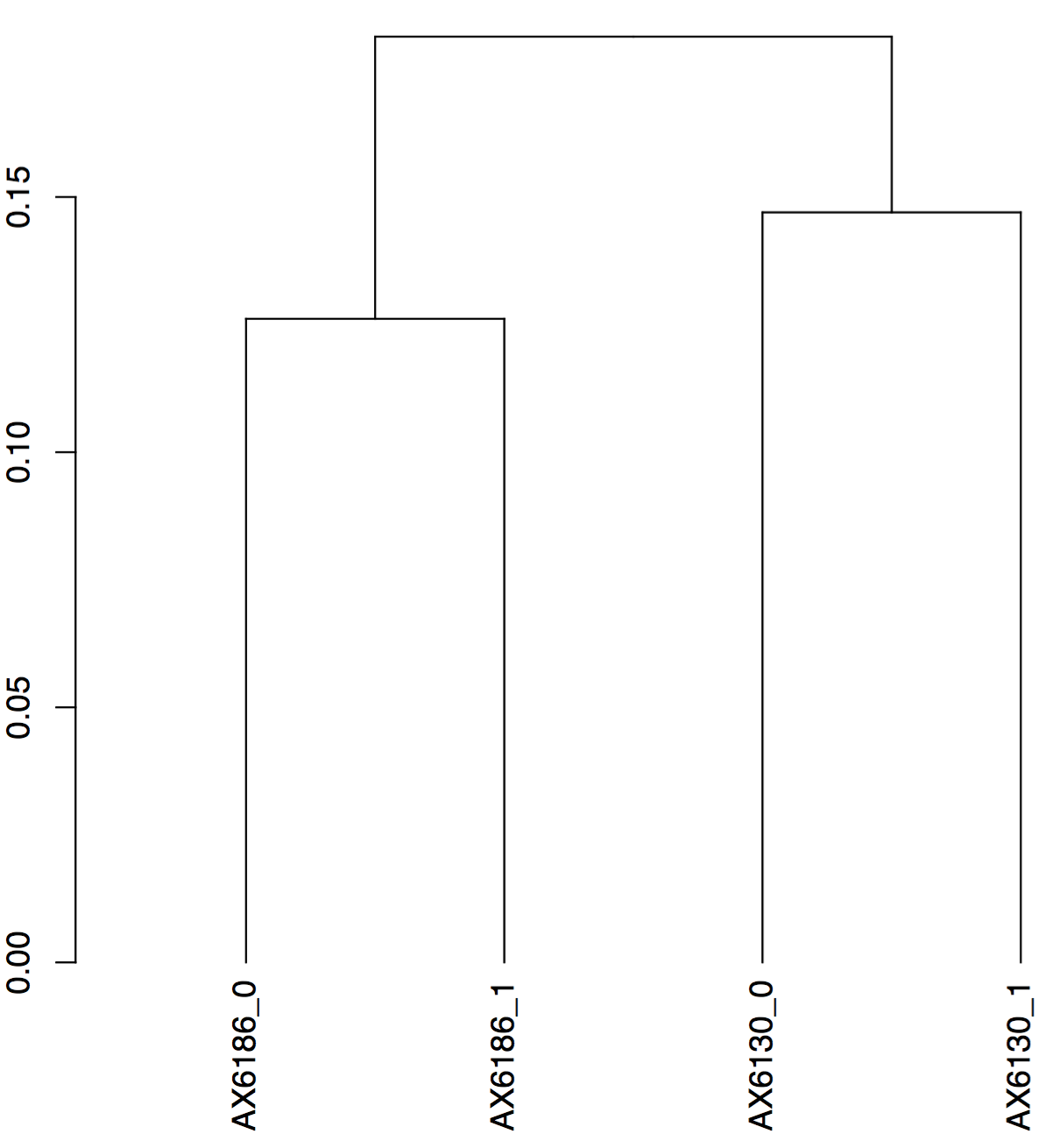
12-b) Cufflinks

Similarly to the analysis with DESeq2, PRAGUI generates different plots to assess replicate similarity.

Scatterplot:

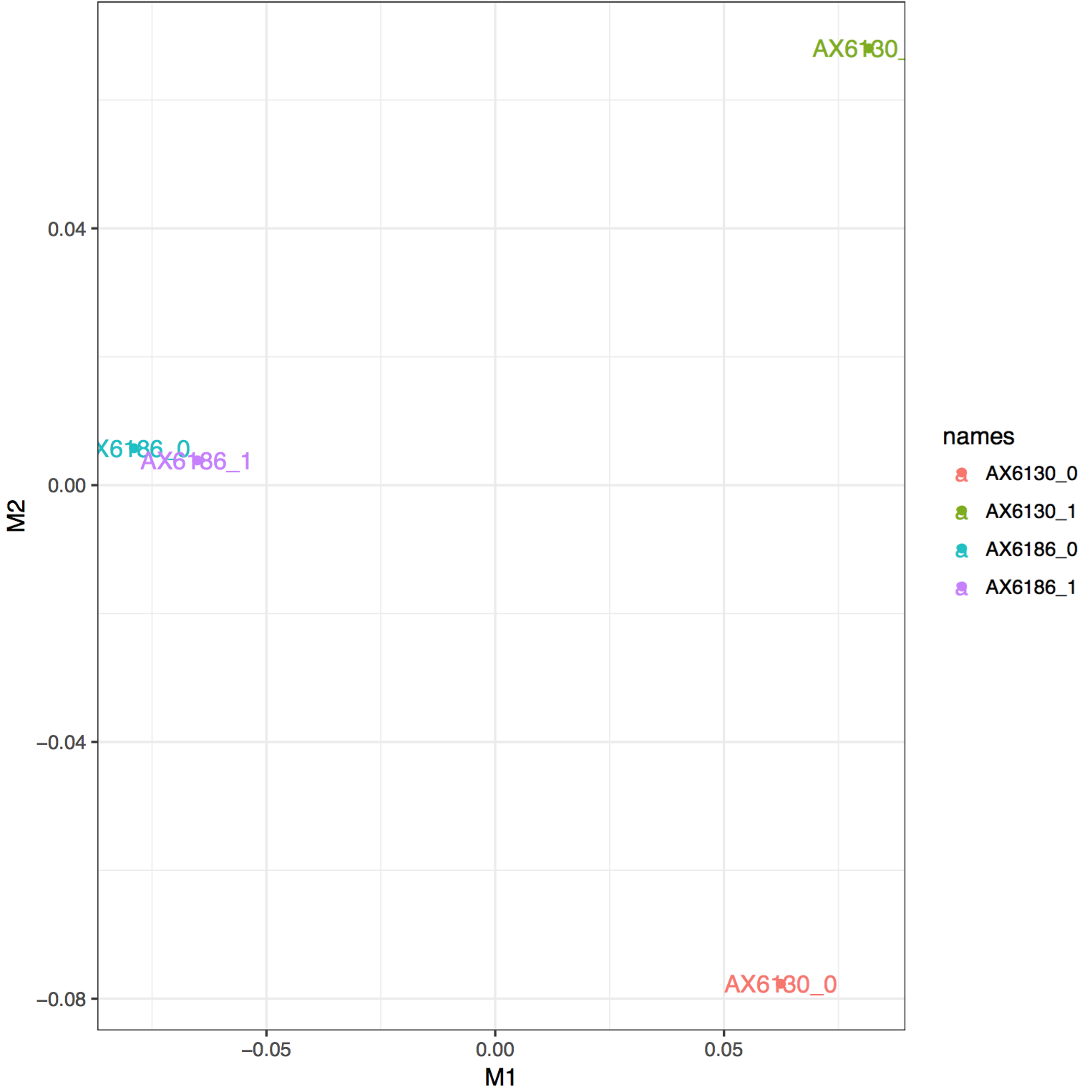


Dendrogram:



Multi-dimensional scaling (MDS) plot:

(The MDS algorithm performs a non-linear dimensionality reduction such that the between-object distances are preserved as well as possible.)



Cufflinks outputs a series of files but the most relevant one for downstream analysis is XXX\_exp.diff:

test\_id gene\_id gene locus sample\_1 sample\_2 status value\_1 value\_2 log2(fold\_change) test\_stat

XLOC\_000001 XLOC\_000001 - III:16041-17305 AX6186 AX6130 OK 141.692 153.87 0.11895 0.256734

XLOC\_000002 XLOC\_000002 - III:41089-42019 AX6186 AX6130 OK 1644.62 107.245 -3.93877 -10.9919

XLOC\_000003 XLOC\_000003 - III:52507-53145 AX6186 AX6130 OK 60.6312 282.131 2.21823 1.74852

XLOC\_000004 XLOC\_000004 - III:87591-89243 AX6186 AX6130 OK 401.666 416.504 0.0523329 0.148819

XLOC\_000005 XLOC\_000005 - III:120660-124730 AX6186 AX6130 OK 132.296 254.977 0.946597 2.40236

XLOC\_000006 XLOC\_000006 - III:124949-126530 AX6186 AX6130 OK 115.5 214.812 0.89518 2.27517

XLOC\_000007 XLOC\_000007 - III:130955-131752 AX6186 AX6130 OK 205.194 358.335 0.804321 1.8577

XLOC\_000008 XLOC\_000008 - III:134190-134314 AX6186 AX6130 OK 1185.81 0 -inf -nan

XLOC\_000009 XLOC\_000009 - III:136160-136309 AX6186 AX6130 OK 2499.8 431.347 -2.53489 -2.18465

p\_value q\_value significant

0.7171 0.871856 no

5e-05 0.00129273 yes

0.05305 0.268938 no

0.83445 0.9292 no

0.0012 0.0162514 yes

0.00215 0.0262429 yes

0.00965 0.0844449 no

0.00015 0.00304714 yes

0.15475 0.452787 no

q-value – p-value adjusted for multiple testing following the Benjamini-Hochberg method