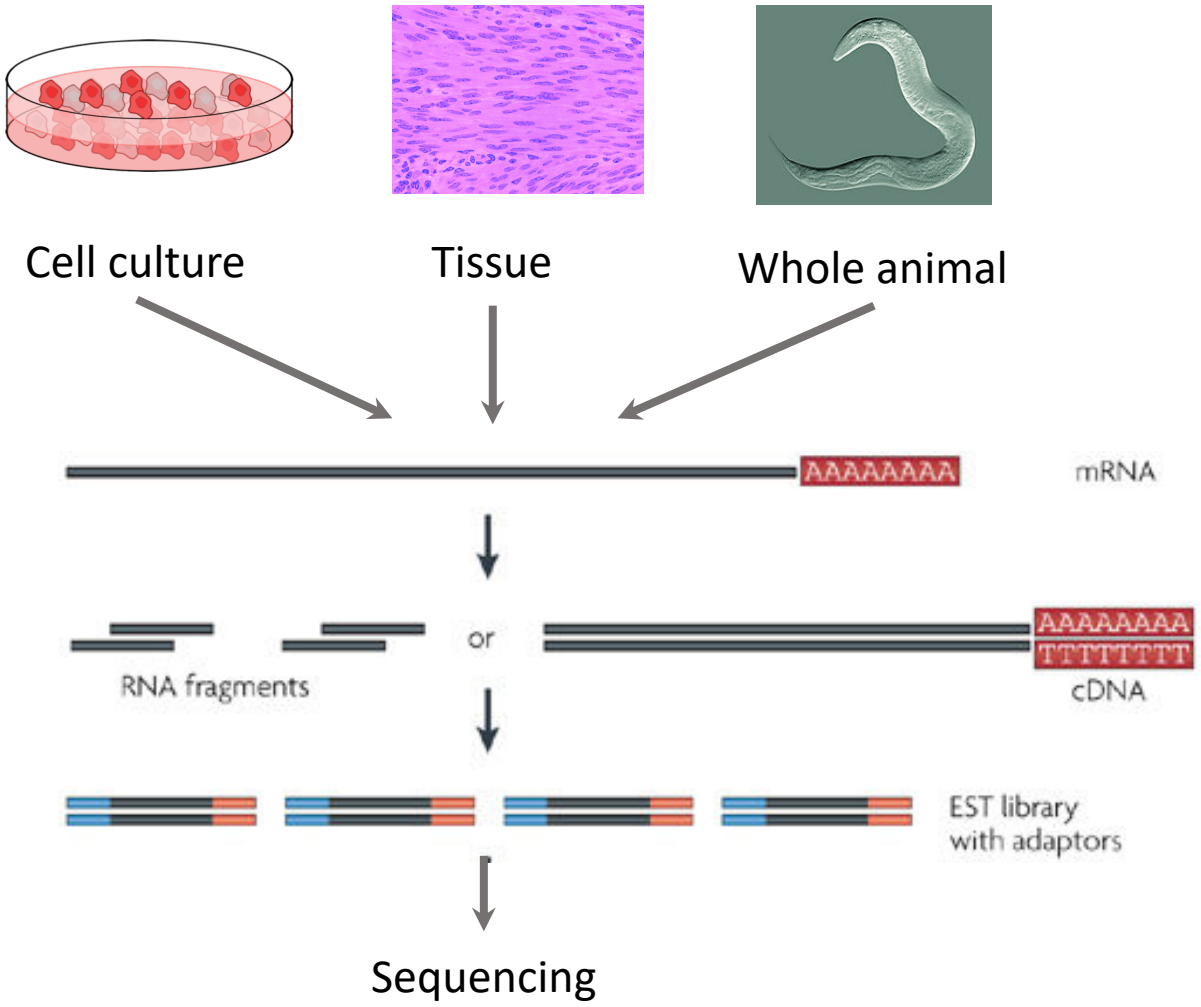
# RNA-seq course – day 1

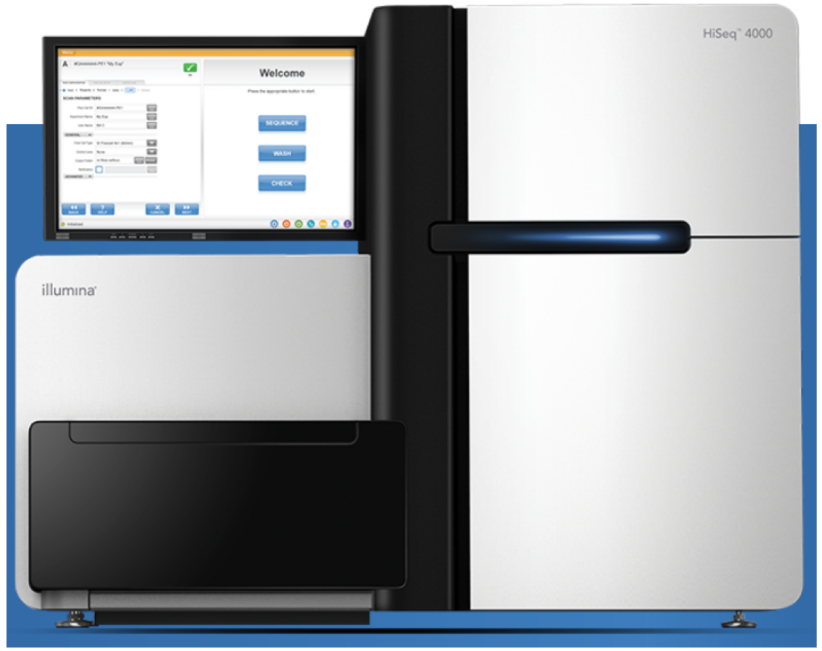
Paula Freire Pritchett
Alastair Crisp

# RNA-seq experimental overview




Adapted from Wang et al 2009

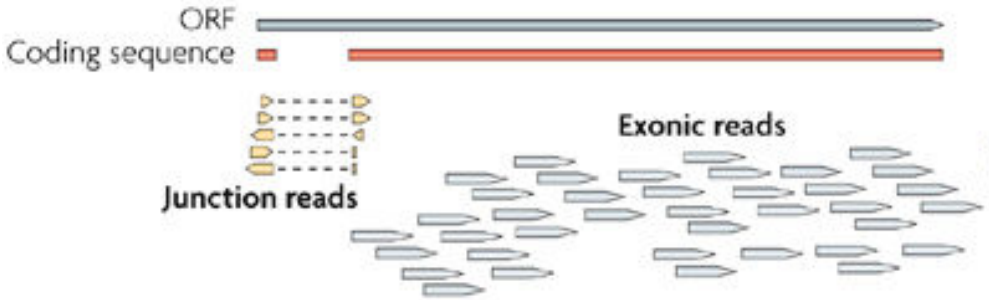
# Processing of RNA-seq data




AGGTCCTGGCTAATGC
GCTACCTGATGCCCTG
ATGCGCTAATGCCCTG
CCTGGCTAAGGTATGC

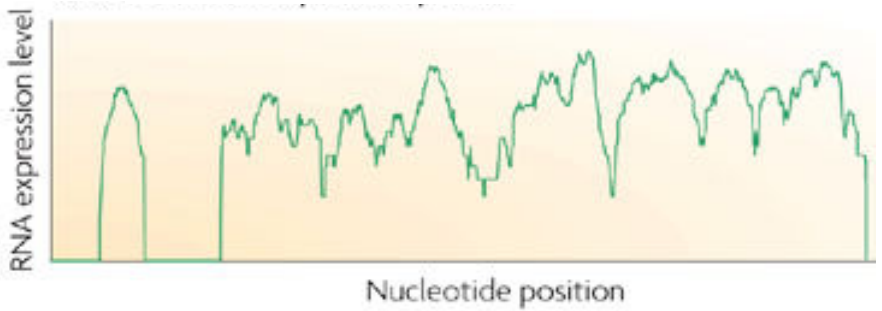
FastQ files

Alignment to the Genome




Bam files

Read count and Normalization




Text Files
Image Files
Pdf Files

Adapted from Wang et al 2009

# Designing an RNA-seq experiment

**Library type**
- Directional libraries
- mRNA
- rRNA depletion
- total RNA

**Number of reads**
- 20 million reads for human/mouse
- >20 million reads for
  de-novo discovery
  lowly expressed transcripts

**Sequencing type**
- 50bp single end
  expression of known genes
- 100bp (or longer) single end
  splice junction usage
- 100bp paired end
  novel transcript discovery

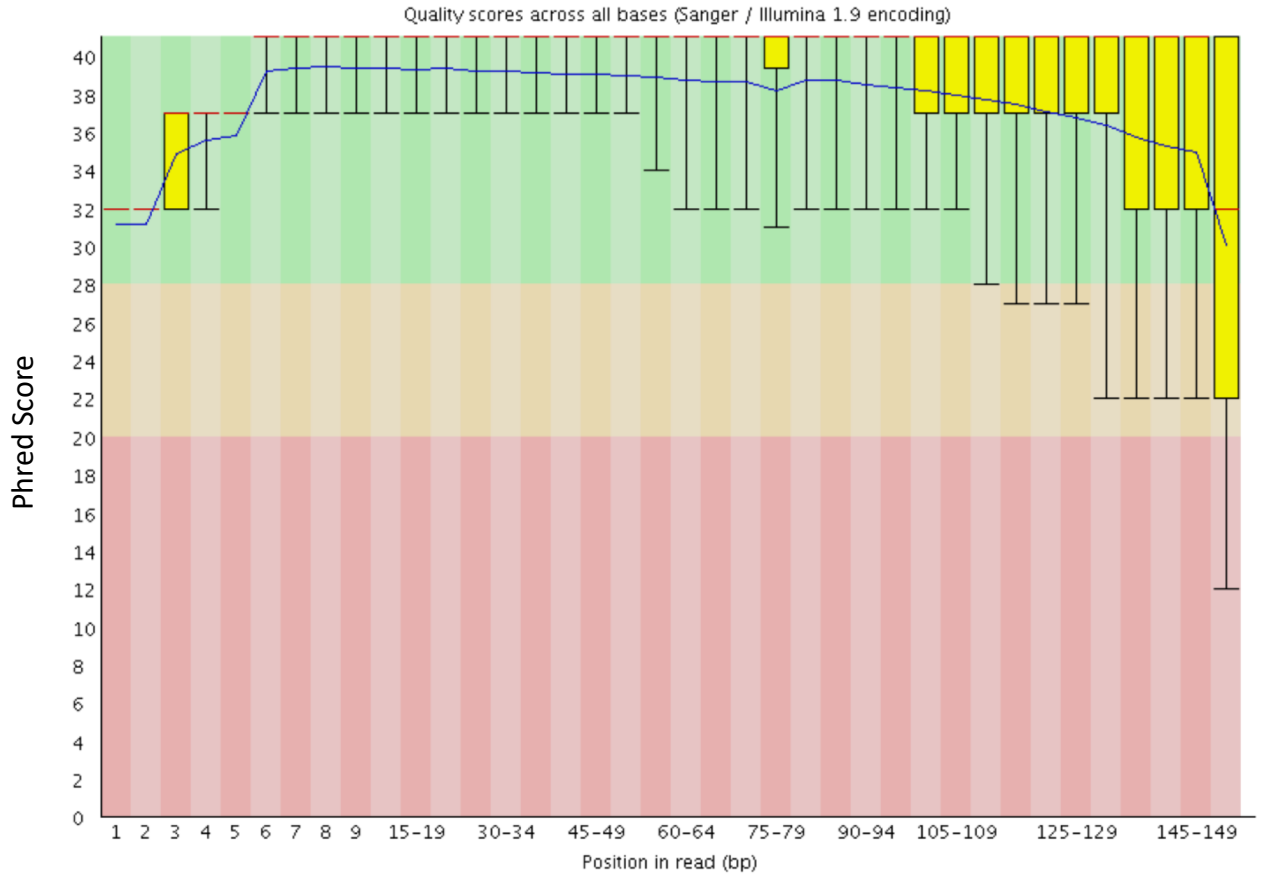
**Number of replicates**
- No need for technical replicates
- 3 biological replicates minimum
- >4 biological replicates when testing multiple conditions
- Always plan for at least one sample to fail
- Randomise across sample groups

# Quality Control

**FastQC**

Quality control tool for high throughput sequence. It provides a modular set of analyses which should help diagnose any abnormalities in the data.

**Per base sequence quality**




Phred Score: $Q = -10 \log_{10} P$

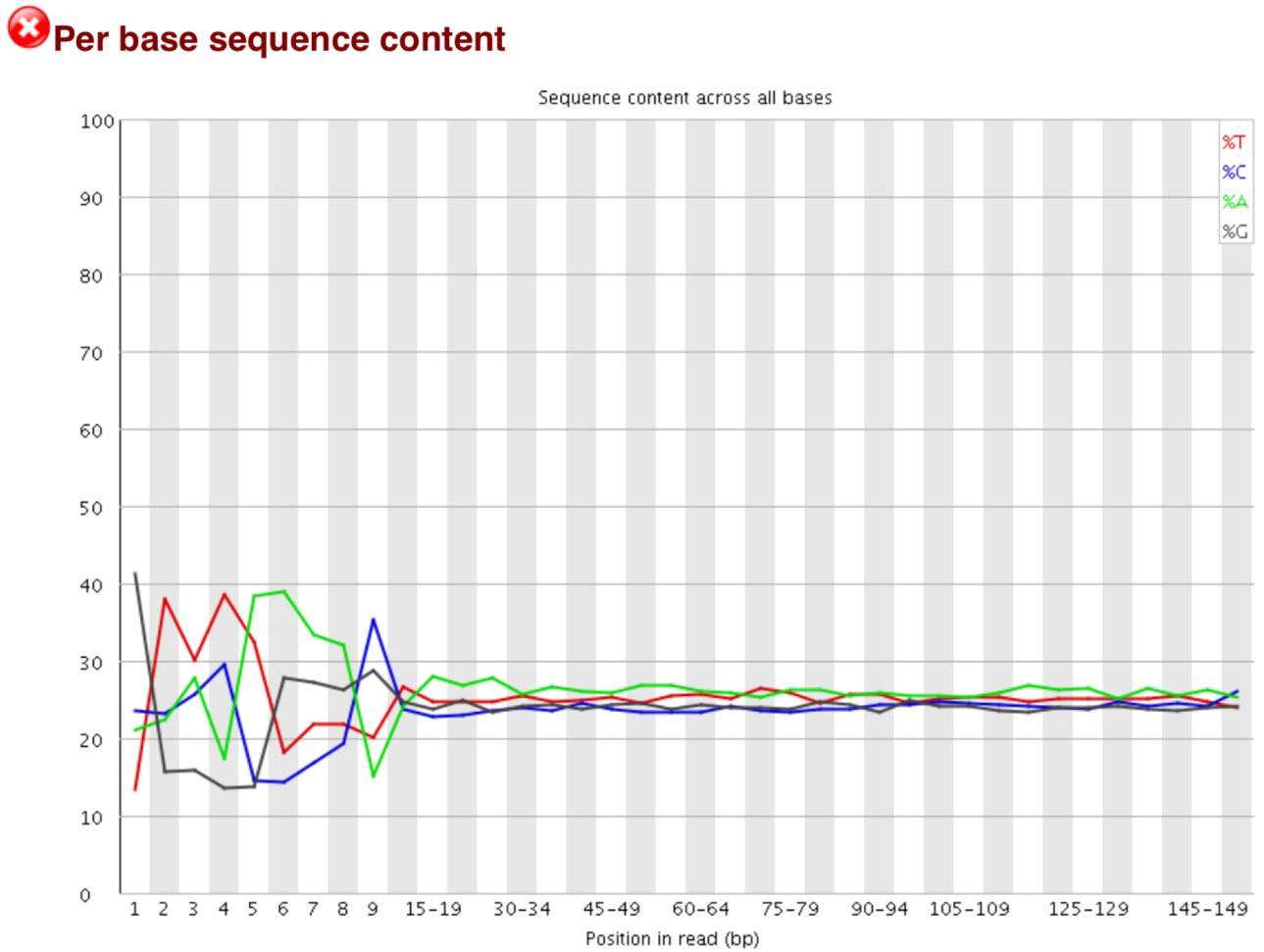
**Phred quality scores are logarithmically linked to error probabilities**

| 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% |
| 50 | 1 in 100,000 | 99.999% |
| 60 | 1 in 1,000,000 | 99.9999% |

# Quality Control

**FastQC**

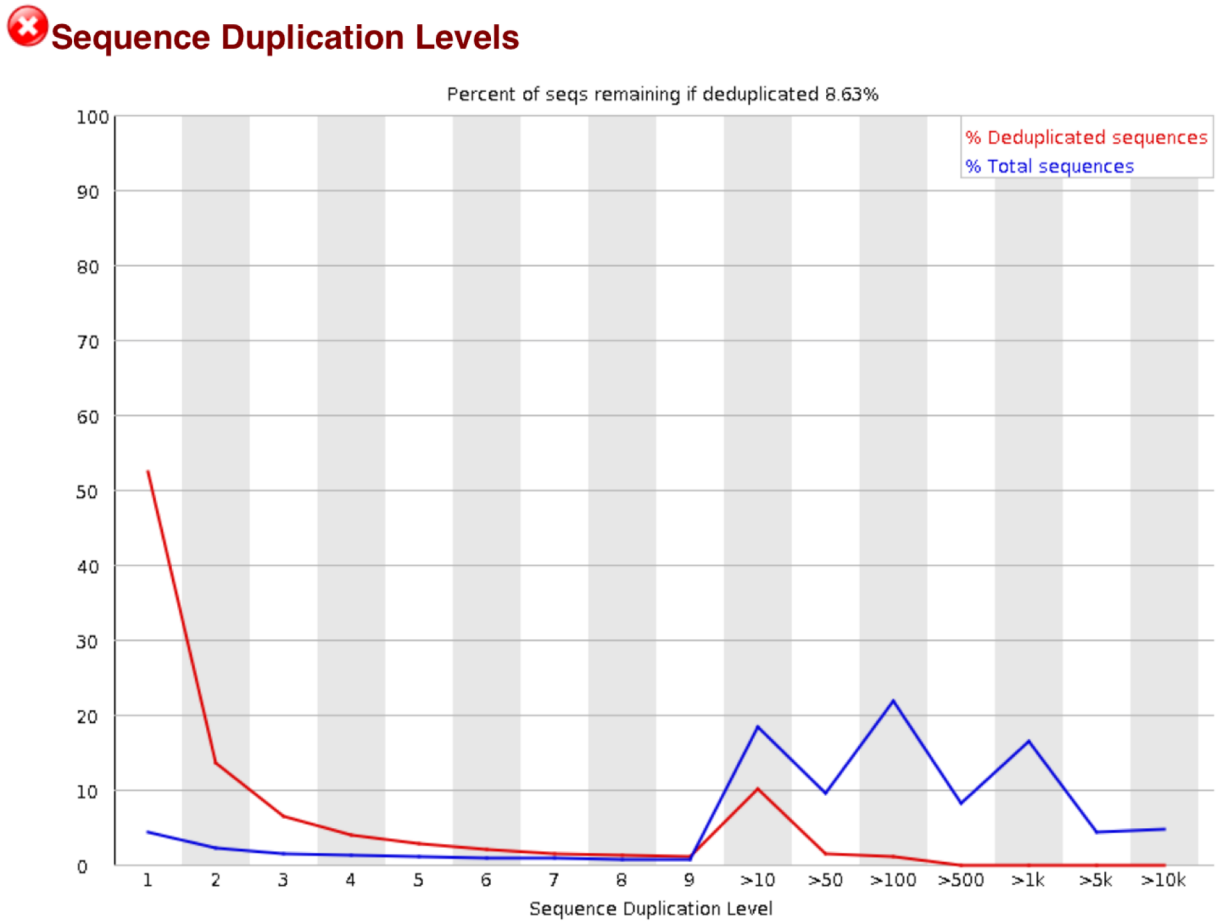
Quality control tool for high throughput sequence. It provides a modular set of analyses which should help diagnose any abnormalities in the data.



# Quality Control

**FastQC**

Quality control tool for high throughput sequence. It provides a modular set of analyses which should help diagnose any abnormalities in the data.



# Expression Levels

**Read Count and Normalisation**

- Raw read counts are non-informative due to different sources of bias: gene length, GC-content and sequencing depth;

- Most common normalization metrics:
  - RPKMs (reads per kilobase million)

    pm = [total number of reads] / 1 000 000
    [gene A RPM] = [gene A read counts] / pm
    [gene A RPKM] = [gene A RPM] / [length of gene A]

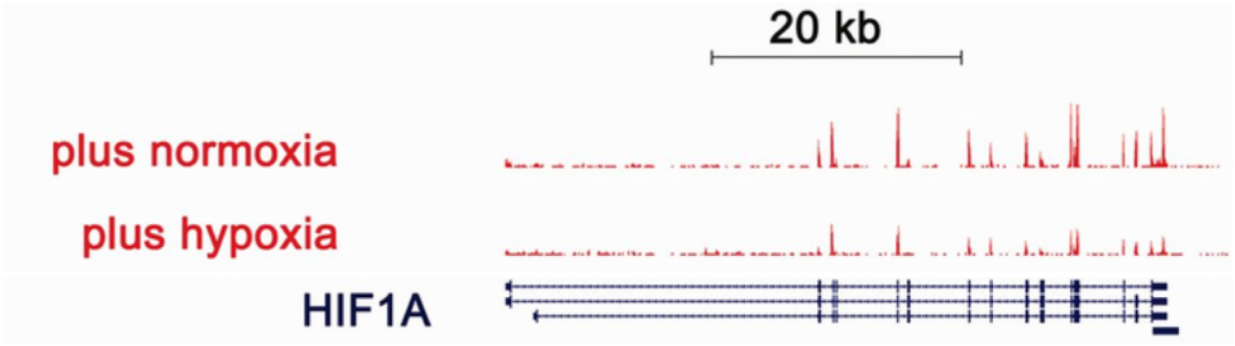
  - FPKMs (fragments per kilobase million)

    pm = [total number of read pairs] / 1 000 000
    [gene A FPM] = [gene A read pairs] / pm
    [gene A FPKM] = [gene A FPM] / [length of gene A]

  - TPM (transcripts per million).

    [gene A RPK]= [gene A raw read counts] / [length of gene A]
    pm = RPK/ 1 000 000
    [gene A TPM] = gene A RPK/ pm

# Differential Gene Expression Analysis




Adapted from Choudhry et al 2015

Statistical methods:

- **DESeq2** (Love, Huber, and Anders 2014)
- **Cufflinks** (Trapnell et al. 2010)
- edgeR (M. D. Robinson, McCarthy, and Smyth 2009)
- limma (Law et al. 2014)

Results across these methods are generally very consistent!

**DESeq2**
Experiments with complex design (multiple factors included)

**Cufflinks**
De novo transcript assembly

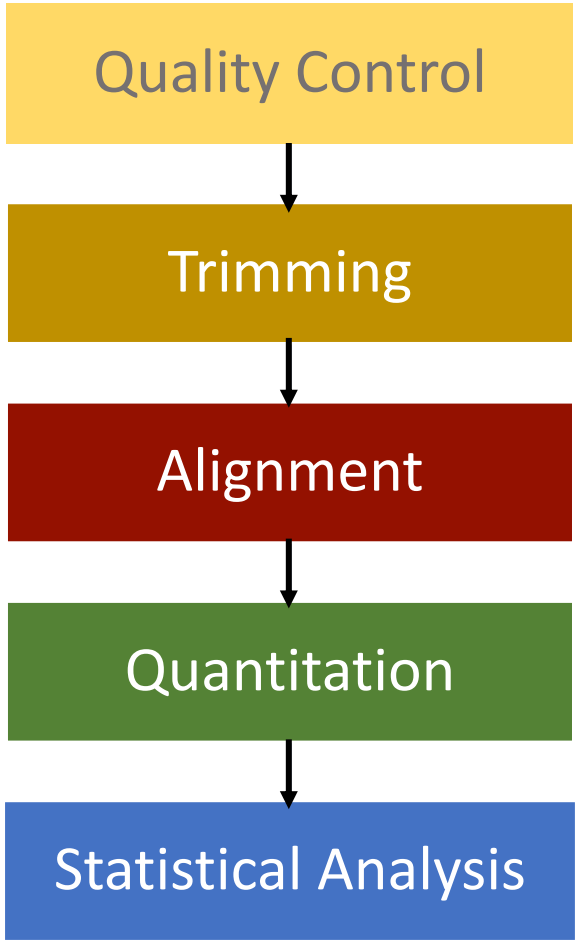
(Good review paper: Hong et al 2014)

# PRAGUI

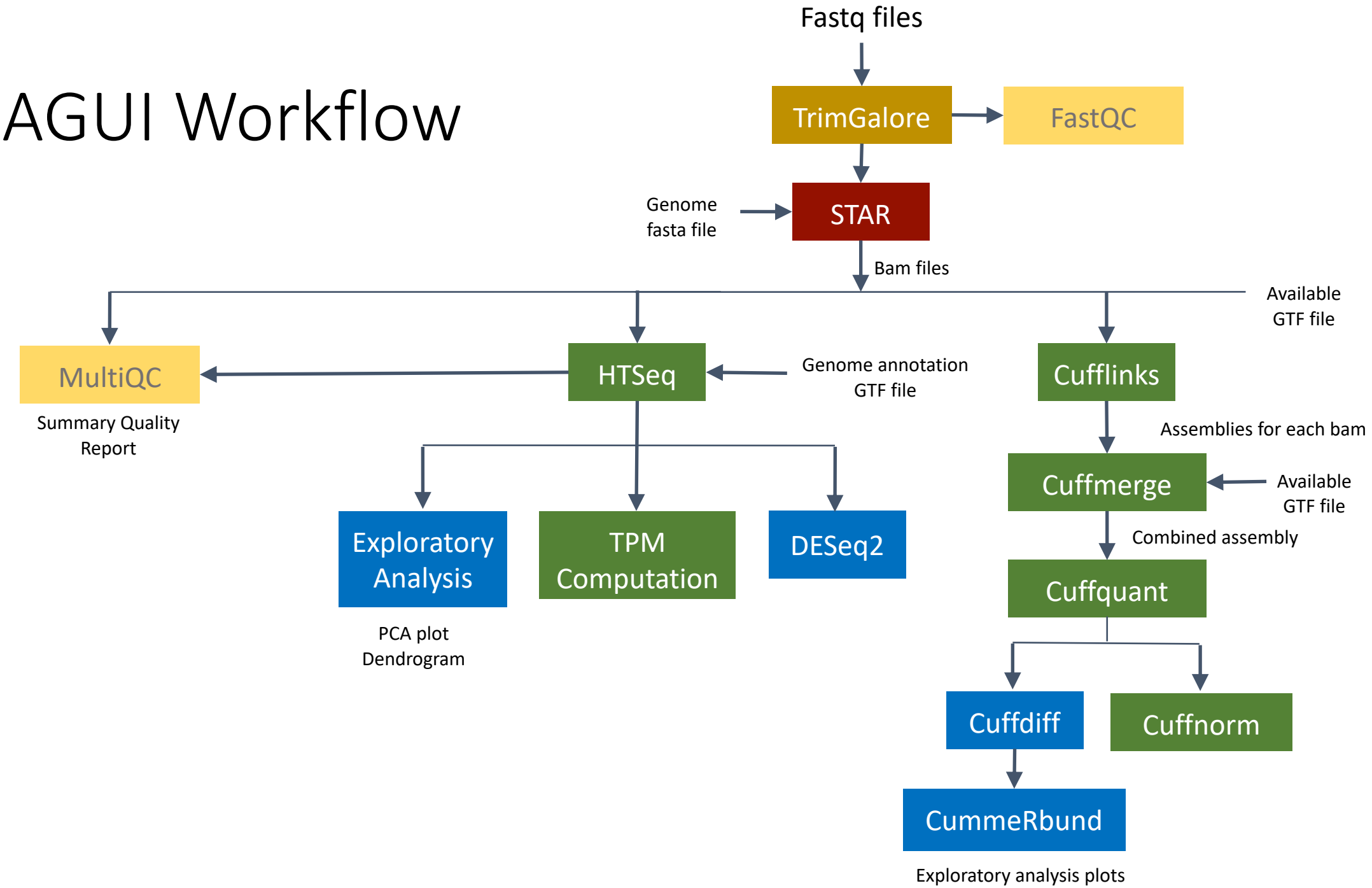
**P**ipeline for **R**NA-seq **A**nalyses with **GUI**

https://github.com/lmb-seq/PRAGUI.git

# RNA-seq data processing



# PRAGUI Workflow



# PRAGUI Input

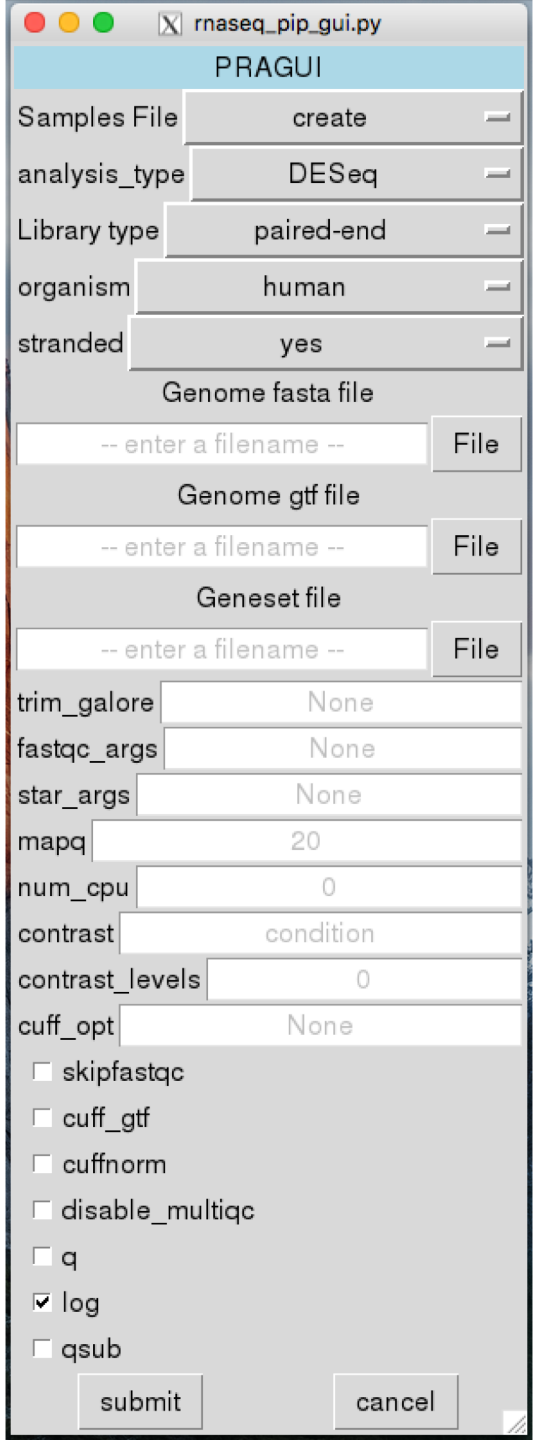
Specify if protocol was strand specific

Gene annotations file to compute TPMs by HTSeq

Specify arguments for each software tool (only needed if you need to change the developer's default values)

Specify whether to run cuffnorm

Option to do a qsub run on the cluster




Input your sample names, biological conditions and file paths

Genome assembly file needed for aligner

Optional gene annotations file to be used by Cufflinks

Threshold used to remove reads mapping multiply to the genome.

Arguments for DESeq2 analysis (leave blank, work in progress)

# PRAGUI output

Summary Report generated by MultiQC

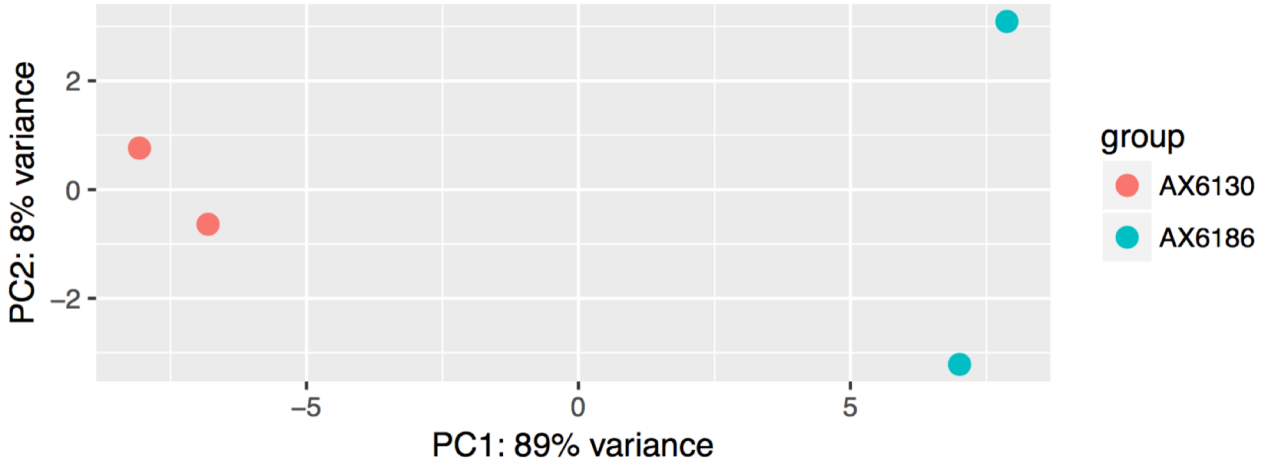
**3 Types of results files:**

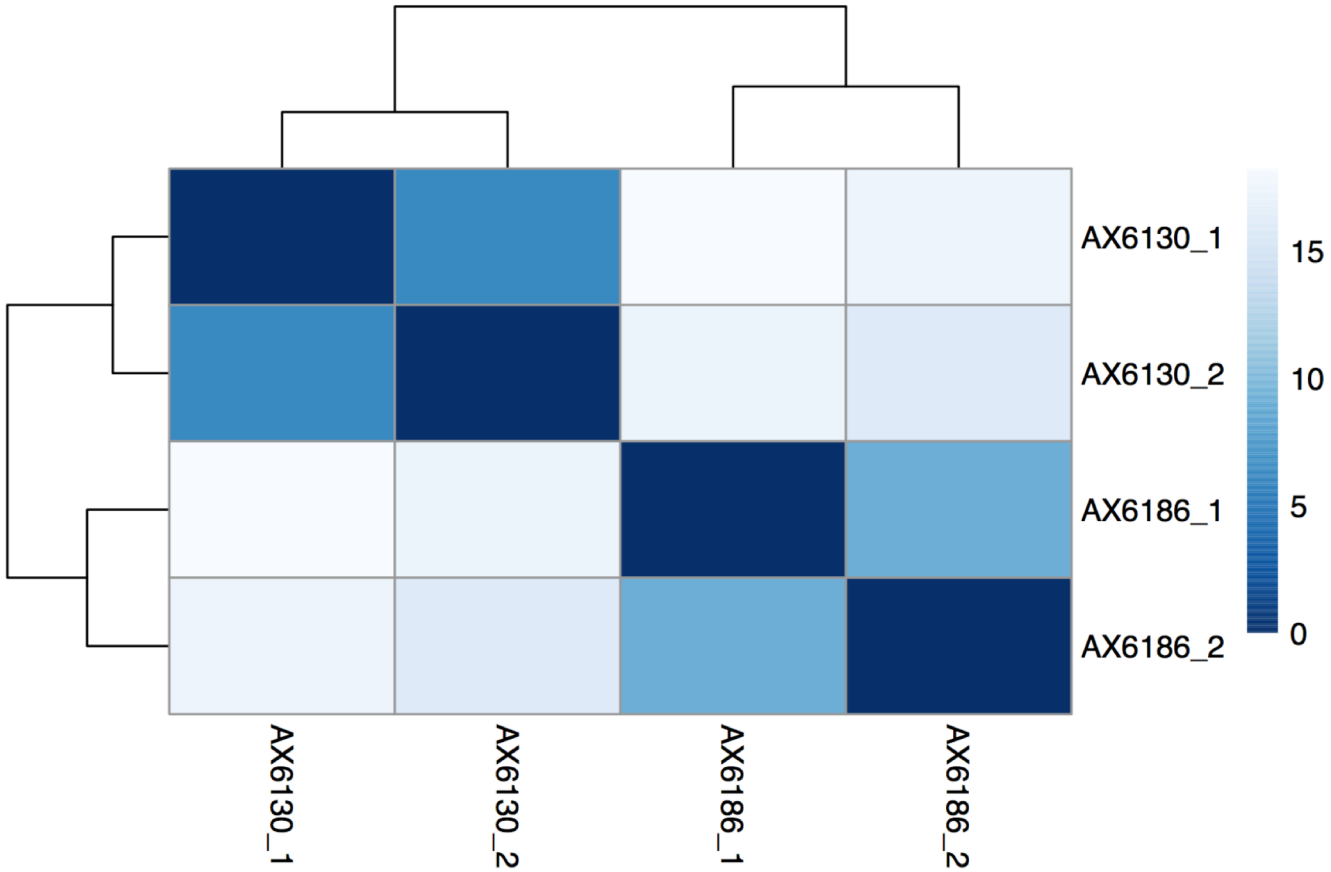
i. Exploratory analysis plots

ii. Files with read counts

iii. Files with results from differential expression analysis.

# Results from DESeq run



# Results from DESeq run

- Normalised read counts for each sample in TPMs (**XXX_tpm.txt** )

| 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 |

- Normalised read counts for each sample in DESeq normalised read counts (**XXX_DESeq_norm_read_counts.txt** )

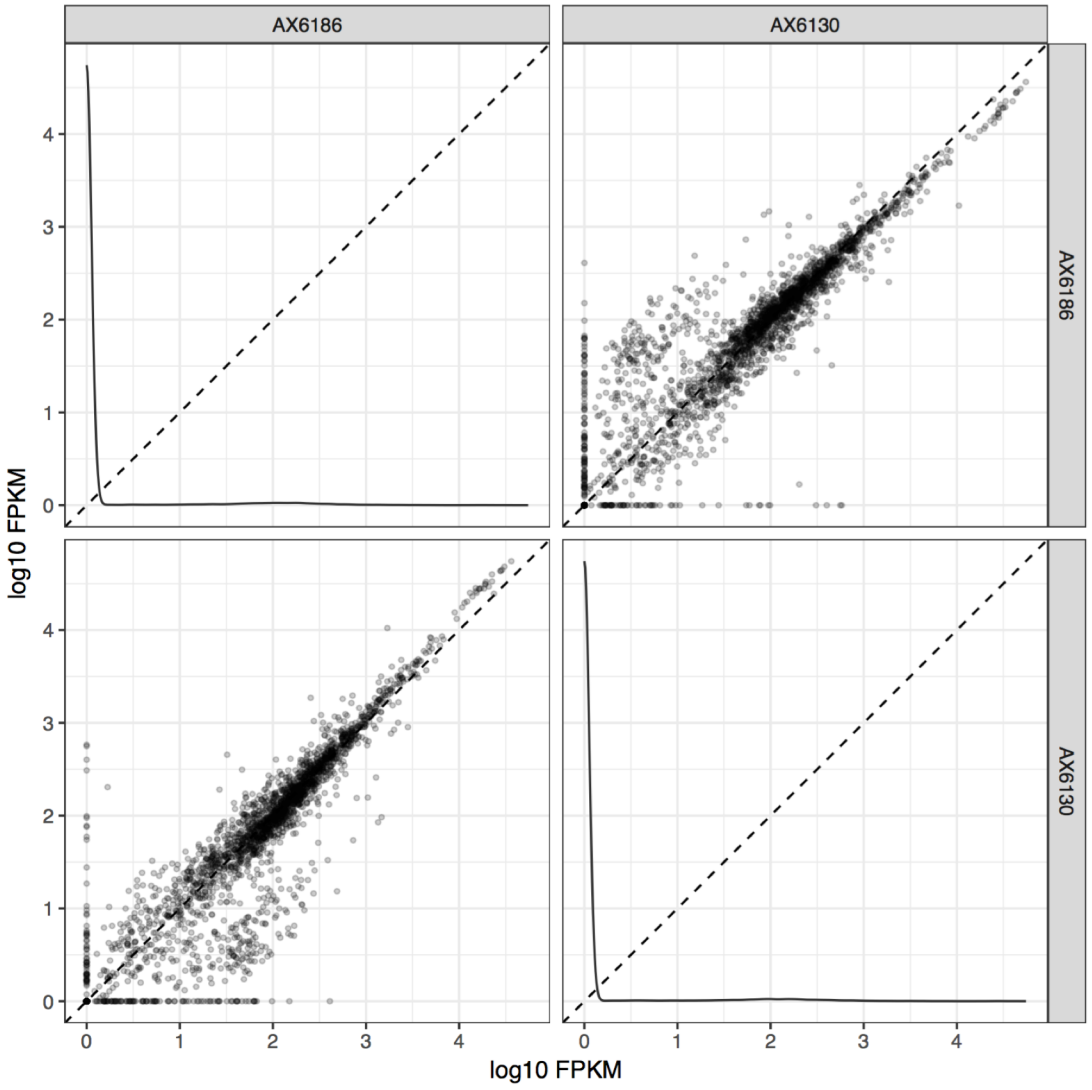
| gene_id | SA_I | SA_II | SA_III | DA_I | DA_II | DA_III |
|---|---|---|---|---|---|---|
| ENSMUSG00000000001 | 1431.24 | 1191.69 | 1525.41 | 1322.00 | 1052.32 | 1215.52 |
| ENSMUSG00000000028 | 69.90 | 109.75 | 77.44 | 51.10 | 58.98 | 75.97 |
| ENSMUSG00000000031 | 1071.58 | 836.23 | 871.96 | 623.66 | 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 |

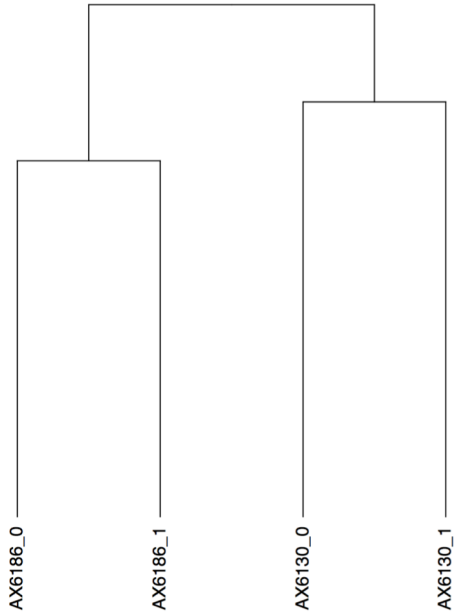
# Results from DESeq run

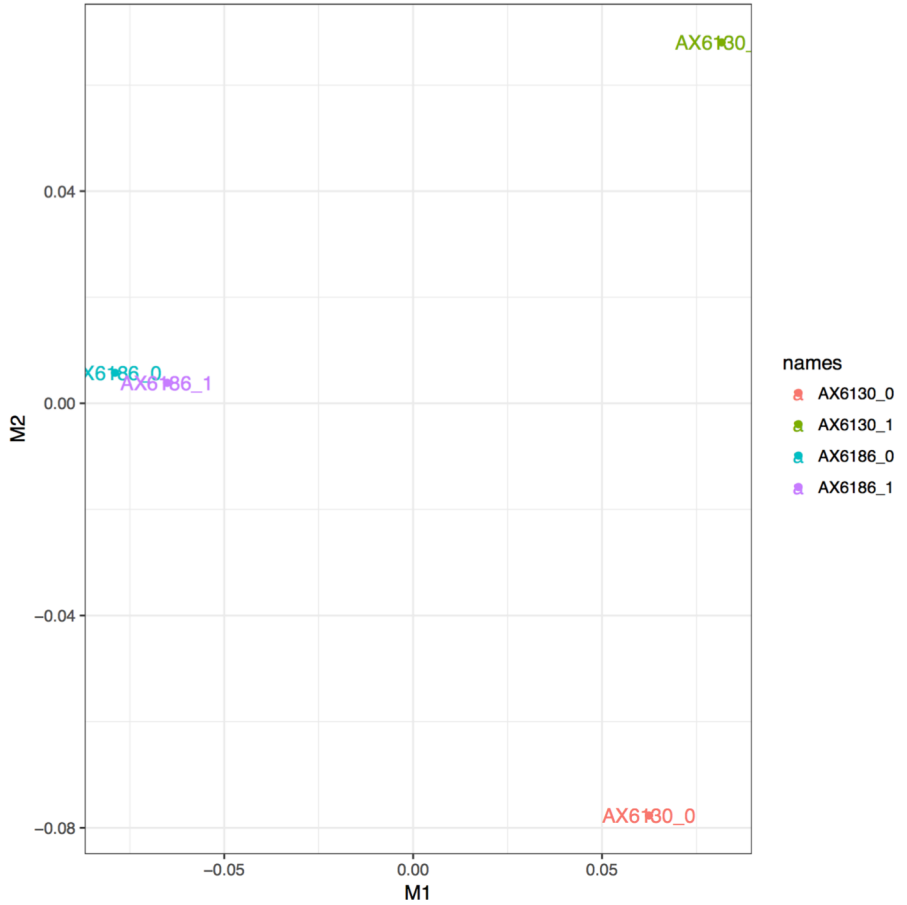
- Differential expression results from DESeq2 (**XXX_DESeq_results_4_peat.txt**)

| test_id | gene_id | gene | locus | sample_1 | sample_2 | status | value_1 | value_2 | log2(fold_change) | test_stat | p_value | q_value | significant |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ENSMUSG00000021848 | ENSMUSG00000021848 | Otx2 | 14:48657679_48667644 | Deep | Superficial | OK | 706.85 | 2592.39 | -1.88 | -15.49 | 4.05E-54 | 8.54E-50 | yes |
| ENSMUSG00000058099 | ENSMUSG00000058099 | Nfam1 | 15:82997721_83033306 | Deep | Superficial | OK | 82.59 | 326.47 | -1.99 | -14.41 | 4.71E-47 | 4.97E-43 | yes |
| ENSMUSG00000032128 | ENSMUSG00000032128 | Robo3 | 9:37415669_37433246 | Deep | Superficial | OK | 47.07 | 384.33 | -3.02 | -14.35 | 1.00E-46 | 7.04E-43 | yes |
| ENSMUSG00000028280 | ENSMUSG00000028280 | Gabrr1 | 4:33132521_33163588 | Deep | Superficial | OK | 30.93 | 225.29 | -2.88 | -14.00 | 1.54E-44 | 8.10E-41 | yes |
| ENSMUSG00000045991 | ENSMUSG00000045991 | Onecut2 | 18:64340364_64398488 | Deep | Superficial | OK | 212.45 | 43.35 | 2.29 | 13.91 | 5.14E-44 | 2.17E-40 | yes |
| ENSMUSG00000030337 | ENSMUSG00000030337 | Vamp1 | 6:125215551_125245964 | Deep | Superficial | OK | 4844.90 | 2091.72 | 1.21 | 13.69 | 1.11E-42 | 3.92E-39 | yes |
| ENSMUSG00000039714 | ENSMUSG00000039714 | Cplx3 | 9:57599992_57606281 | Deep | Superficial | OK | 32.25 | 235.98 | -2.91 | -13.46 | 2.79E-41 | 8.41E-38 | yes |
| ENSMUSG00000045518 | ENSMUSG00000045518 | Onecut3 | 10:80494835_80517276 | Deep | Superficial | OK | 158.69 | 19.43 | 3.01 | 13.34 | 1.45E-40 | 3.81E-37 | yes |
| ENSMUSG00000022054 | ENSMUSG00000022054 | Nefm | 14:68082590_68124846 | Deep | Superficial | OK | 6543.20 | 1718.54 | 1.93 | 12.81 | 1.40E-37 | 3.28E-34 | yes |

# Results from Cufflinks run



# Results from Cufflinks run

- Differential expression results from Cufflinks (**XXX_exp.diff** )

| test_id | gene_id | gene | locus | sample_1 | sample_2 | status | value_1 | value_2 | log2(fold_change) | test_stat | p_value | q_value | significant |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| XLOC_000001 | XLOC_000001 | - | III:16041-17305 | AX6186 | AX6130 | OK | 141.692 | 153.87 | 0.11895 | 0.256734 | 0.7171 | 0.871856 | no |
| XLOC_000002 | XLOC_000002 | - | III:41089-42019 | AX6186 | AX6130 | OK | 1644.62 | 107.245 | -3.93877 | -10.9919 | 5e-05 | 0.00129273 | yes |
| XLOC_000003 | XLOC_000003 | - | III:52507-53145 | AX6186 | AX6130 | OK | 60.6312 | 282.131 | 2.21823 | 1.74852 | 0.05305 | 0.268938 | no |
| XLOC_000004 | XLOC_000004 | - | III:87591-89243 | AX6186 | AX6130 | OK | 401.666 | 416.504 | 0.0523329 | 0.148819 | 0.83445 | 0.9292 | no |
| XLOC_000005 | XLOC_000005 | - | III:120660-124730 | AX6186 | AX6130 | OK | 132.296 | 254.977 | 0.946597 | 2.40236 | 0.0012 | 0.0162514 | yes |
| XLOC_000006 | XLOC_000006 | - | III:124949-126530 | AX6186 | AX6130 | OK | 115.5 | 214.812 | 0.89518 | 2.27517 | 0.00215 | 0.0262429 | yes |
| XLOC_000007 | XLOC_000007 | - | III:130955-131752 | AX6186 | AX6130 | OK | 205.194 | 358.335 | 0.804321 | 1.8577 | 0.00965 | 0.0844449 | no |
| XLOC_000008 | XLOC_000008 | - | III:134190-134314 | AX6186 | AX6130 | OK | 1185.81 | 0 | -inf | -nan | 0.00015 | 0.00304714 | yes |
| XLOC_000009 | XLOC_000009 | - | III:136160-136309 | AX6186 | AX6130 | OK | 2499.8 | 431.347 | -2.53489 | -2.18465 | 0.15475 | 0.452787 | no |