

How to spot problems in your sequencing data

Simon Andrews

@simon_andrews

How to spot problems in your sequencing ~~data~~ experiment

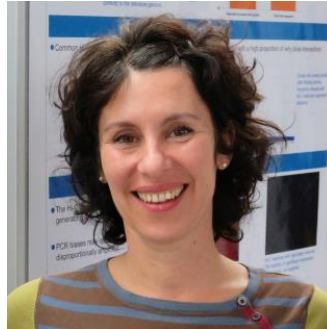
Simon Andrews

@simon_andrews

Babraham Bioinformatics



Simon Andrews
Head of Bioinformatics



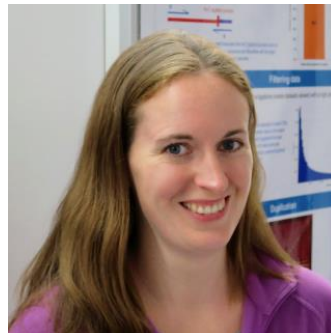
Anne Segonds-Pichon
Biostatistician



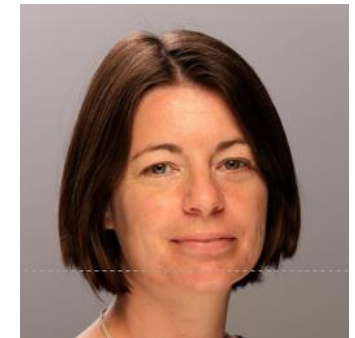
Felix Krueger
Bioinformatician



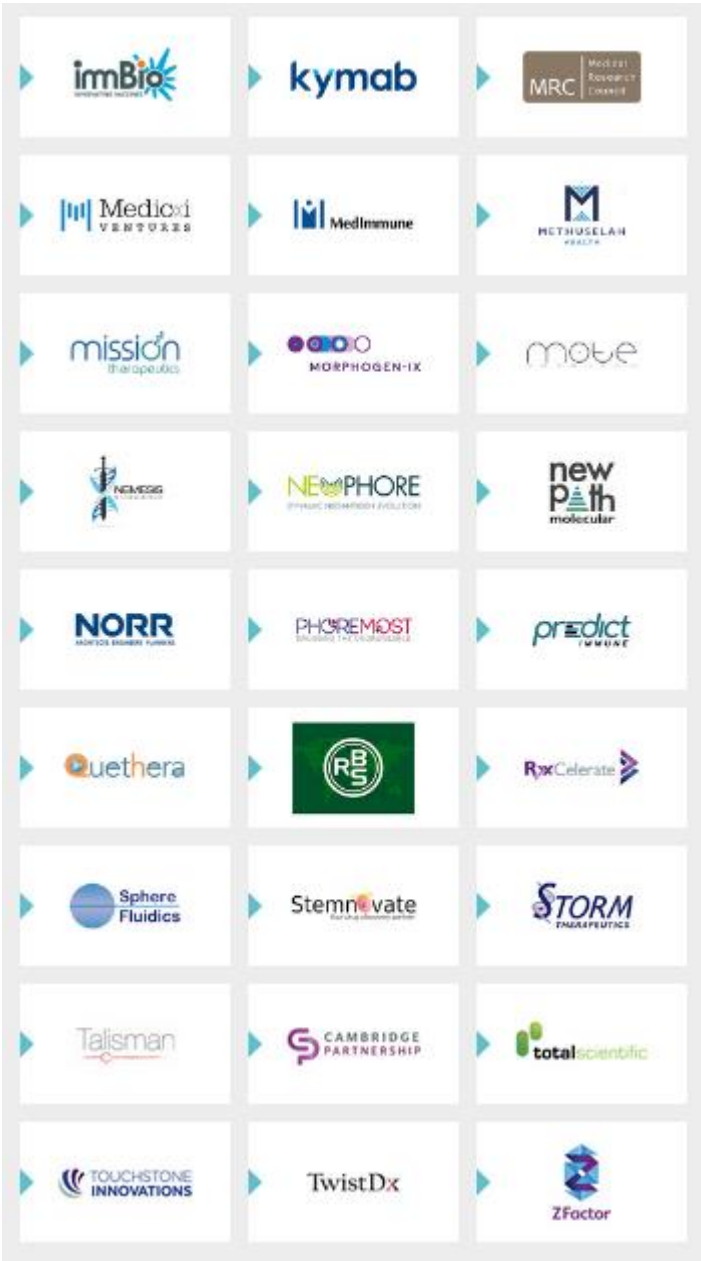
Steven Wingett
Bioinformatician



Laura Biggins
Bioinformatician



Jo Montgomery
Training Developer

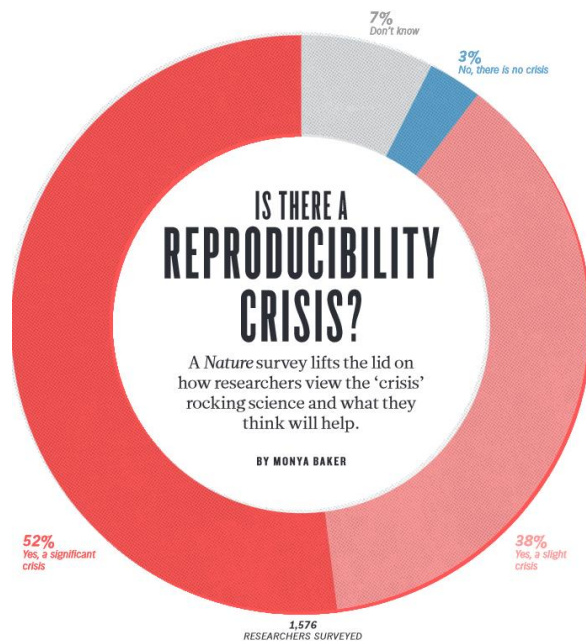


A Crisis of Analysis?

Essay

Why Most Published Research Findings Are False

John P.A. Ioannidis

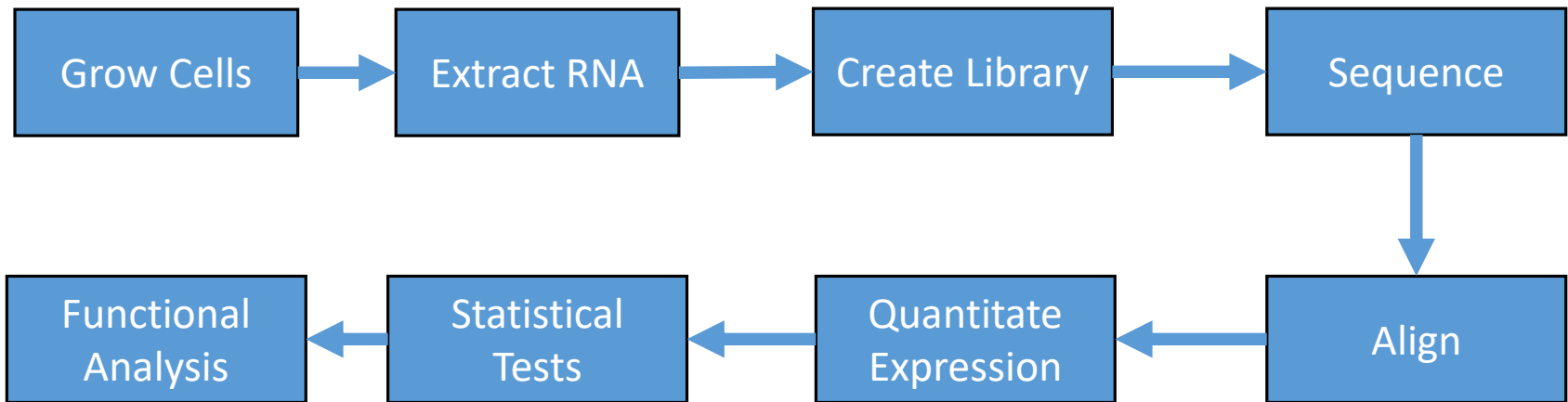


The case against science is straightforward: much of the scientific literature, perhaps half, may simply be untrue. Afflicted by studies with small sample sizes, tiny effects, invalid exploratory analyses, and flagrant conflicts of interest, together with an obsession for pursuing fashionable trends of dubious importance, science has taken a turn towards darkness.

Richard Horton

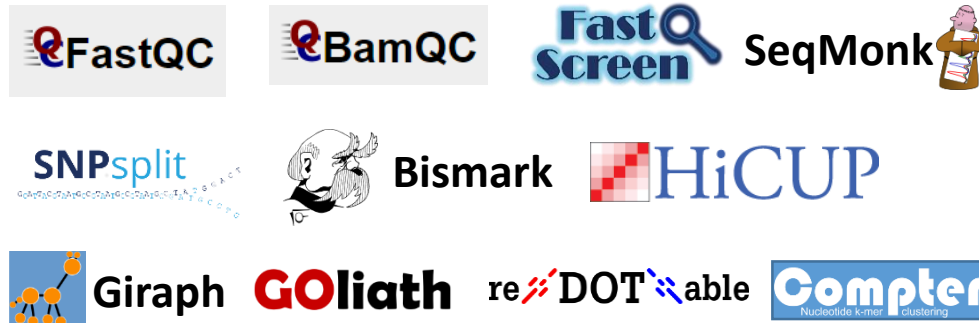
richard.horton@lancet.com

Experiments are fragile



QC at Babraham Bioinformatics

- Software



- Training



In 2018

74 training days

1000 people trained



Articles about common next generation sequencing problems

Search for a topic
FastQC Illumina All Applications SeqMonk

10XQC

Submit your 10X Cell Ranger® report and compare it with data from 112 other reports, contributed by other users from across the globe.

This tool was developed to allow users of 10X Genomics 3'mRNA-seq technology to share their experiences. By providing some simple metadata from your experiment's Cell Ranger report, you can gain access to a pool of knowledge, compare your results against others that have used similar experimental conditions, and determine how well your experiment has worked.



Submit Report

You will not be asked to submit any sensitive information. 10X Cell Ranger reports are not saved on the website.

Vistory DB

A collection of SeqMonk Vistory files

About

Filters

Type search term here

- ngs (4) chip-seq (1)
- peak calling (1) annotation (1)
- promoter (1) rna-seq (2)
- statistics (2) qc (2)
- visualisation (2)

Vistories

Calling Peaks from Replicated ChIP Data

In replicated ChIP-Seq datasets there are a few different ways to call peaks. We go through a few options for how to call peaks, explaining the differences between them and the strengths and weaknesses of each

ngs, chip-seq, peak calling

The latest Illumina sequencing samples

The new Illumina patterned flow cell technology is prone to "index hopping", leading to reads being assigned to the wrong sample in multiplexed sequencing runs. This problem has caused countless research groups to re-assess their experimental protocols from using the new sequencers. However, with careful preparation it may be possible to ameliorate this problem to a negligible level for most applications

June 7, 2017 | Steven Wingett | HiSeq, Illumina

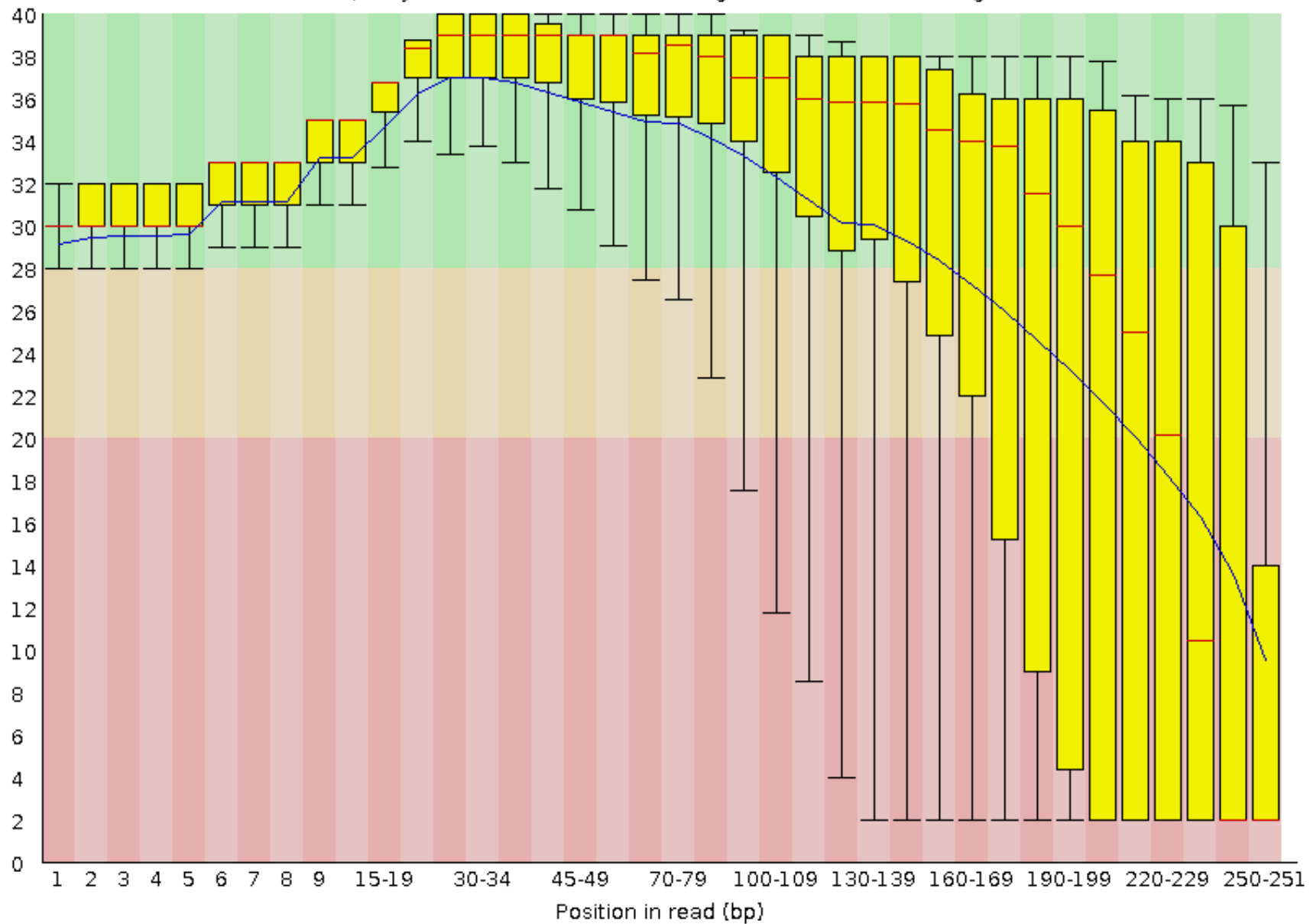


7 short stories...

Look at the metrics your
instruments / programs give you

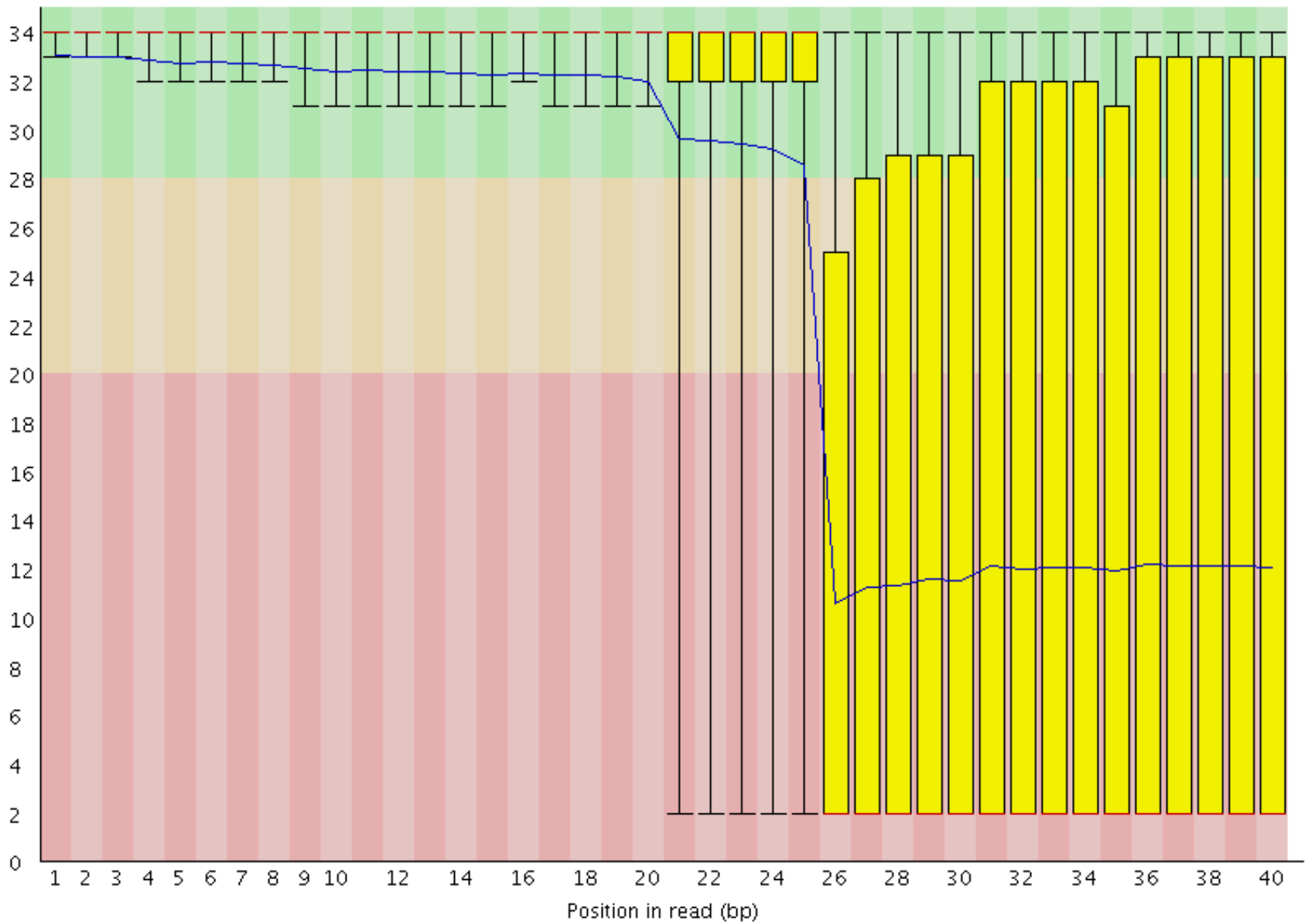


Quality scores across all bases (Sanger / Illumina 1.9 encoding)



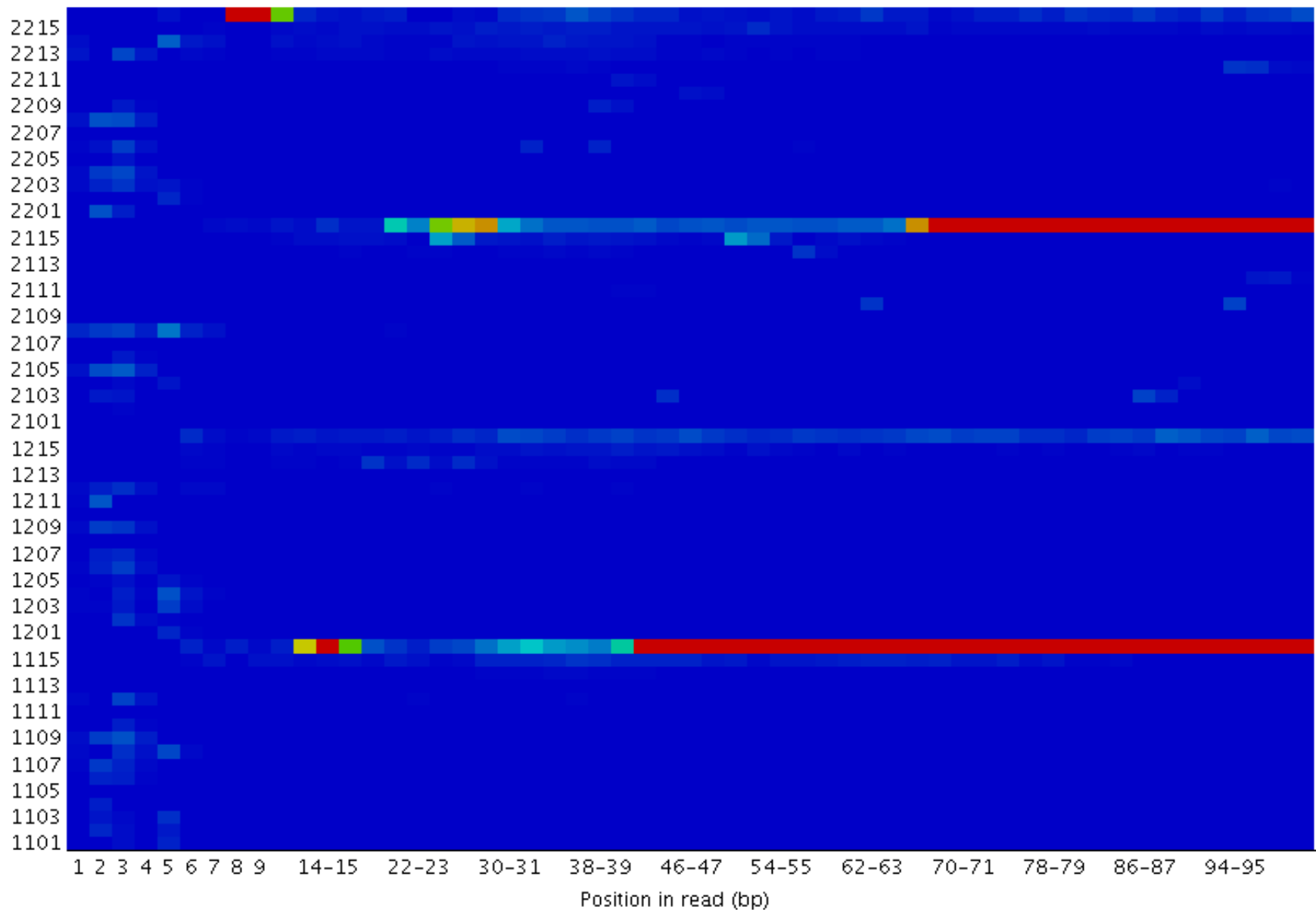
FastQC per base quality plot

Quality scores across all bases (Illumina 1.5 encoding)



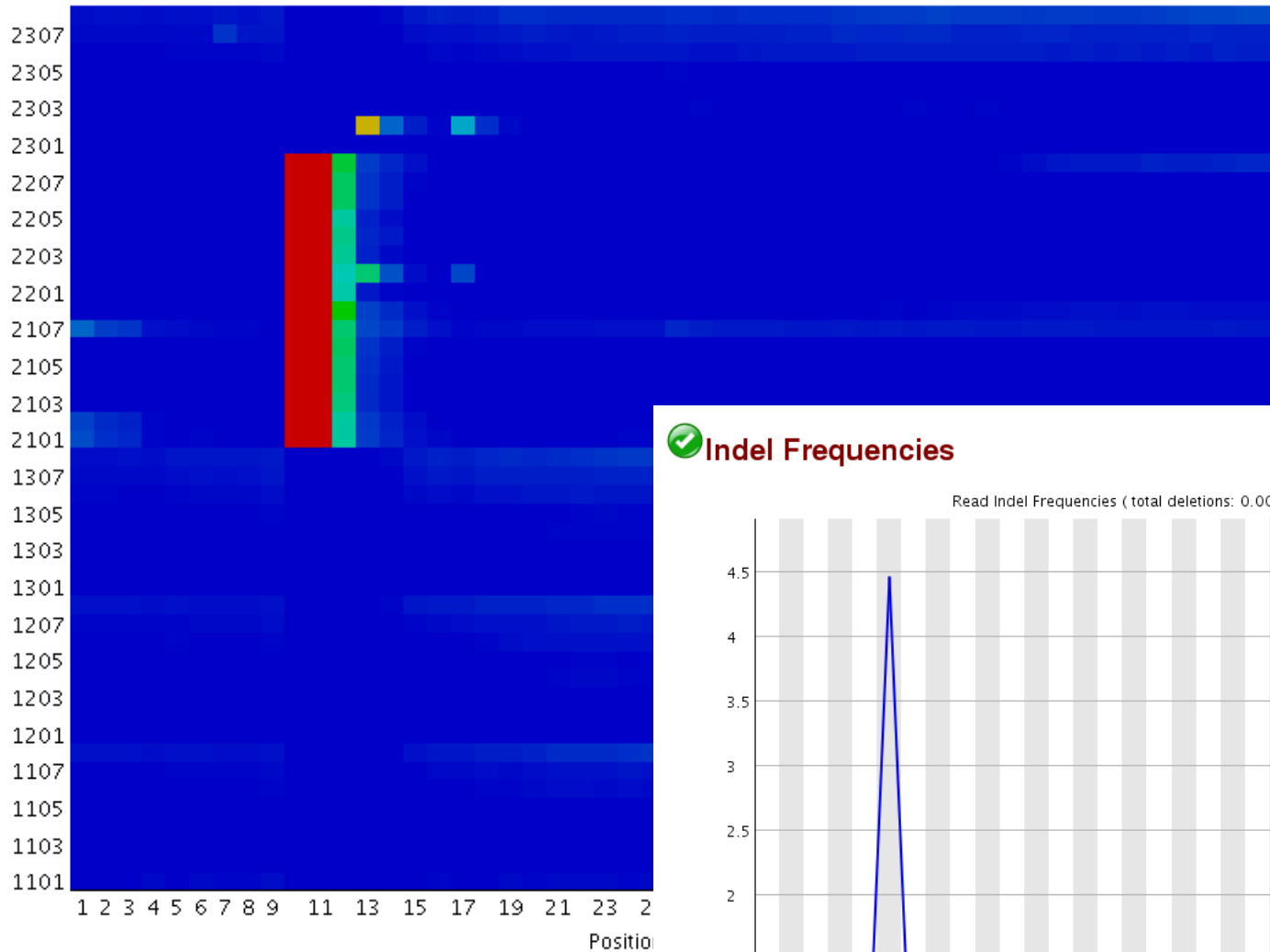
FastQC per base quality plot

Quality per tile



FastQC per tile quality plot

Quality per tile

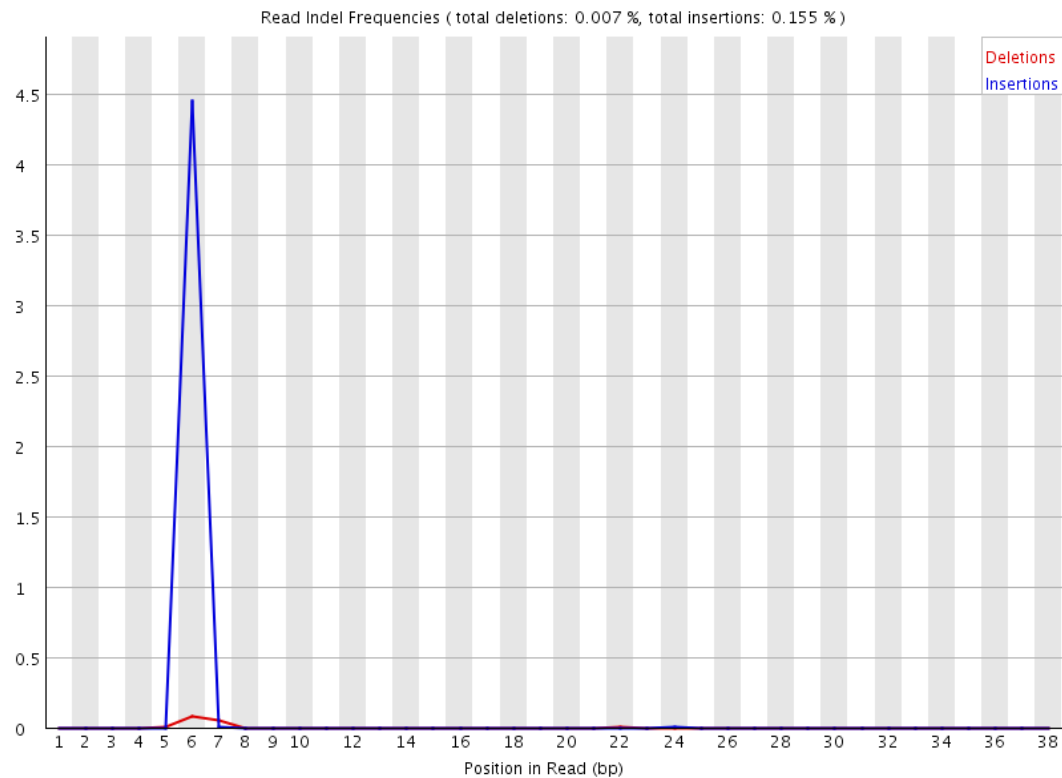


FastQC per tile quality plot

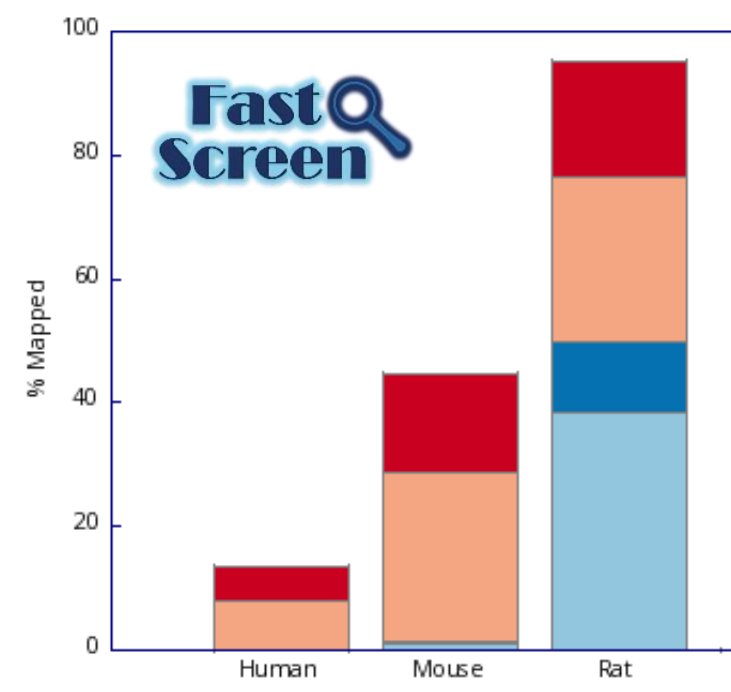
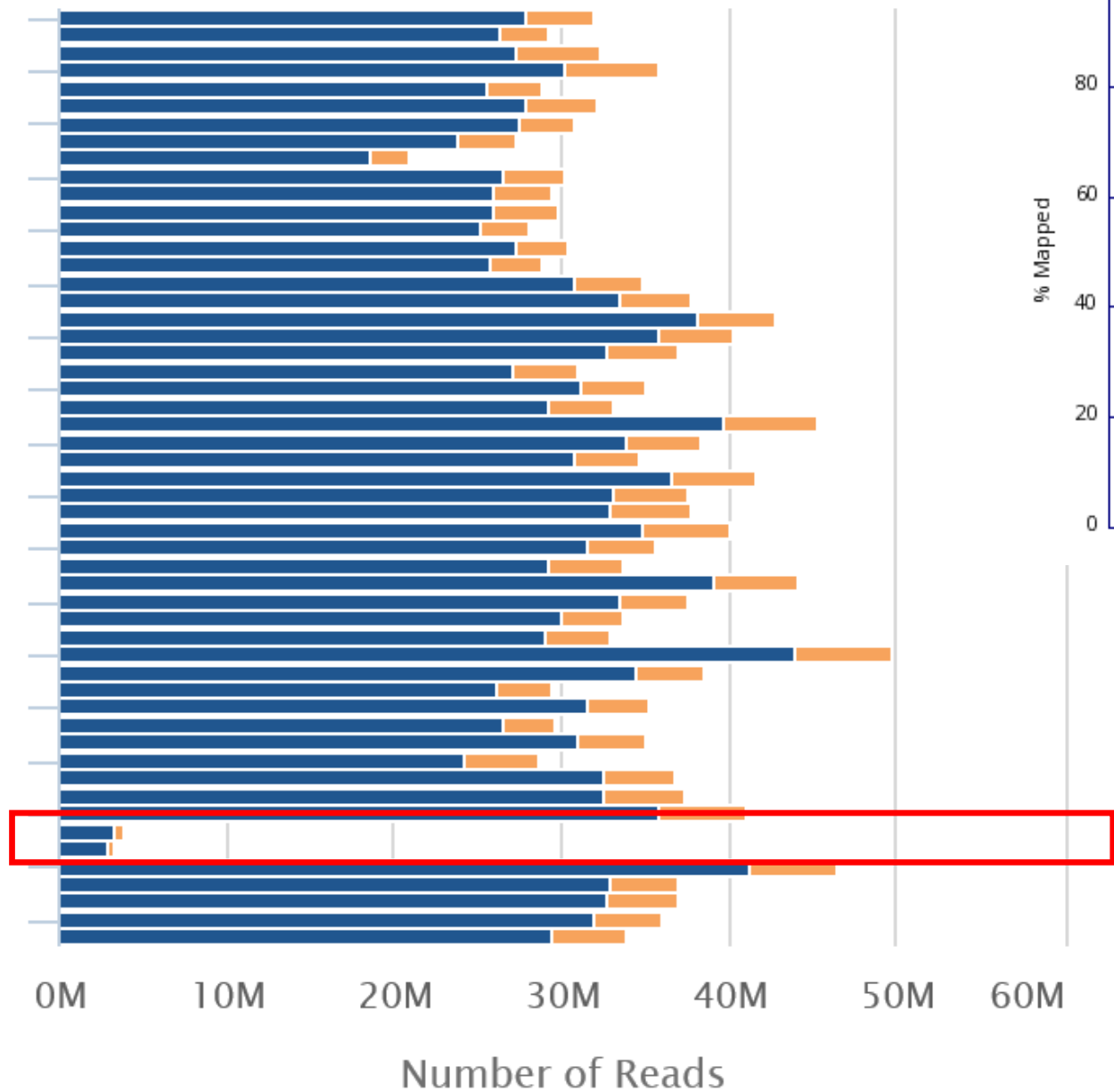
BamQC indel plot



Indel Frequencies




Time loading forward index: 00:01:10
Time loading reference: 00:00:05
Multiseed full-index search: 00:20:47
24548251 reads; of these:
 24548251 (100.00%) were paired; of these:
 1472534 (6.00%) aligned concordantly 0 times
 21491188 (87.55%) aligned concordantly exactly 1 time
 1584529 (6.45%) aligned concordantly >1 times
94.00% overall alignment rate
Time searching: 00:20:52
Overall time: 00:22:02




Take note of flags, warnings and errors


The analysis detected some issues with your sequencing run. [Details »](#)

Alert	Value	Detail
 Low Fraction Reads Confidently Mapped To Transcriptome	51.5%	Ideal > 60%. This can indicate use of the wrong reference transcriptome, poor library quality, or poor sequencing quality. Application performance may be affected.

the design formula contains a numeric variable with integer values, specifying a model with increasing fold change for higher values.







did you mean for this to be a factor? if so, first convert this variable to a factor using the `factor()` function

 Request Generated Warnings... ×


 There were 54227 warnings when processing your request - showing the first 2386

```
[571 times] Reading position 102488814 was 497625bp beyond the end of chr15 (101991189)
[554 times] Reading position 102484240 was 493051bp beyond the end of chr15 (101991189)
[545 times] Reading position 87434055 was 4176614bp beyond the end of chr17 (83257441)
[528 times] Reading position 103242298 was 1251109bp beyond the end of chr15 (101991189)
[519 times] Reading position 102487129 was 495940bp beyond the end of chr15 (101991189)
[493 times] Reading position 103241196 was 1250007bp beyond the end of chr15 (101991189)
[482 times] Reading position 103243780 was 1252591bp beyond the end of chr15 (101991189)
[480 times] Reading position 102474691 was 483502bp beyond the end of chr1
[457 times] Reading position 102488277 was 497088bp beyond the end of chr1
[447 times] Reading position 103241592 was 1250403bp beyond the end of chr
[433 times] Reading position 120944561 was 13900843bp beyond the end of ch
```

Close

-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per tile sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per sequence GC content](#)

×

 This test requires a representative set of all probes to be valid. Be careful running it on a biased subset of probes

OK

1: In `fitNbinomGLMs(objectNZ, maxit = maxit, useOptim = useOptim, useQR = useQR, : 1rows had non-positive estimates of variance for coefficients`

Look at your data

Google: “Simple RNA-Seq analysis”

SOFTWARE TOOL ARTICLE

REVISED RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR [version 3; peer review: 3 approved]



RNA-Seq Differential Expression

Illustration by
Illumina, Inc.

[Bookmark this app](#) [? Help](#)

The RNA-Seq Differential Expression Analysis workflow performs the following functions

- Differential expression analysis of reference genes with DESeq2

A.I.R.

ARTIFICIAL INTELLIGENCE RNA-SEQ

THE SIMPLEST, FASTEST AND MOST ACCURATE SOFTWARE ON THE MARKET FOR RNA-SEQ ANALYSIS

Galaxy

Analyze Data Workflow Visualize Shared Data Help Login or Register

RNA-seq Analysis Exercise

Galaxy provides the tools necessary to creating and executing a complete RNA-seq analysis pipeline. This exercise introduces these tools and guides you through a simple pipeline using some example datasets. Familiarity with Galaxy and the general concepts of RNA-seq analysis are useful for understanding this exercise. This exercise should take 1-2 hours. You can check your work by looking at the history and visualization at the bottom of this page, which contain the datasets for the completed exercise.

The START App: a web-based RNaseq analysis and visualization resource FREE

Step 1.
Upload or Fetch RNA-seq Data

- Upload your raw or processed RNA-seq data
- Fetch >8,000 public RNA-seq datasets published in the Gene Expression Omnibus

Step 2.
Select Data Analysis Tools

- Select from multiple state-of-the-art RNA-seq data analysis tools
- Contribute your computational tool as a plugin

Step 3.
Generate Your Notebook

- Access and share your results through a permanent URL
- Download, rerun and customize your notebook using Docker

BioJupies Automatically Generates RNA-seq Data Analysis Notebooks

With BioJupies you can produce in seconds a customized, reusable, and interactive report from your own raw or processed RNA-seq data through a simple user interface

[Get Started](#)

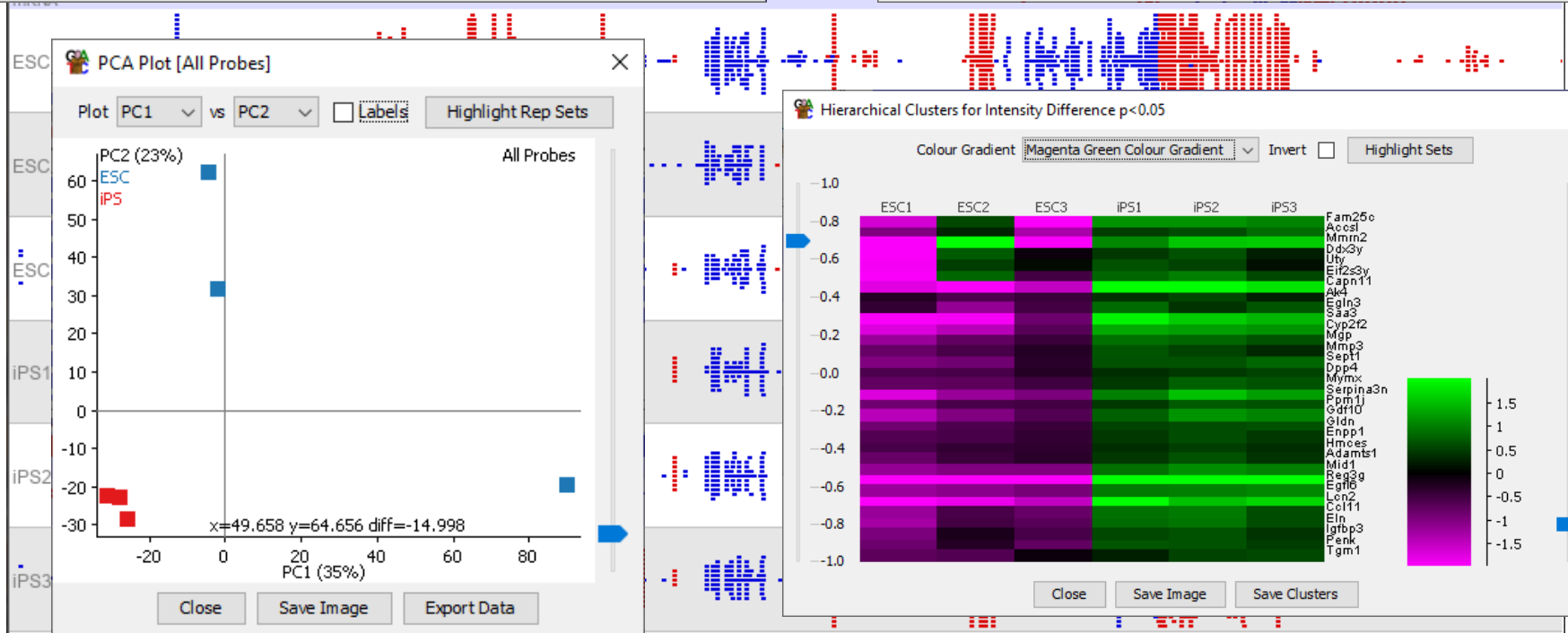
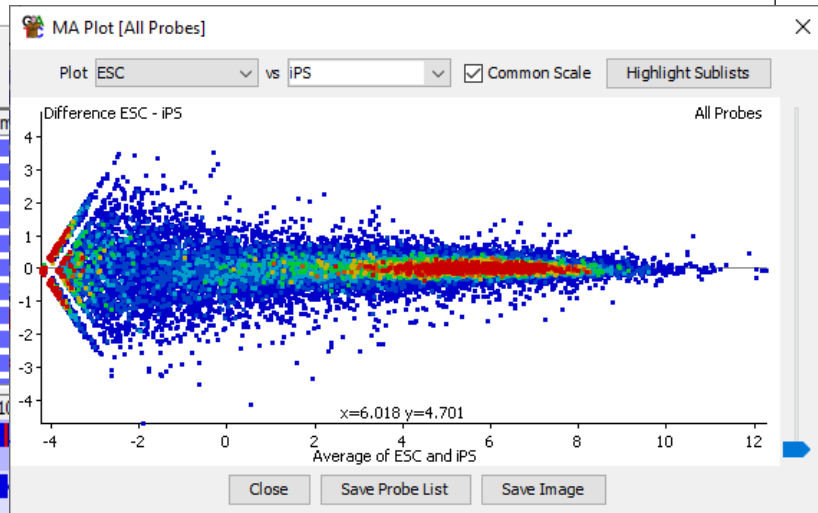
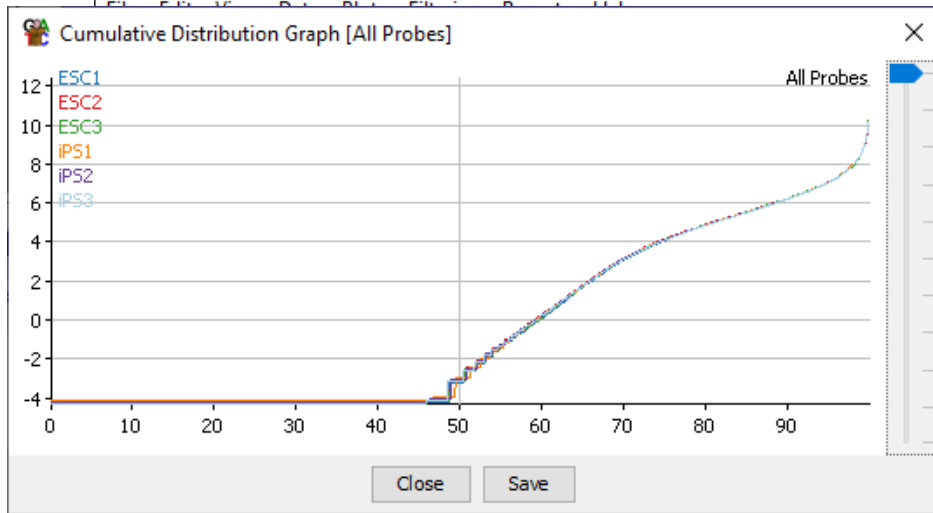
QLUCORE

PRODUCT DOWNLOADS ABOUT US SUPPORT CONTACT

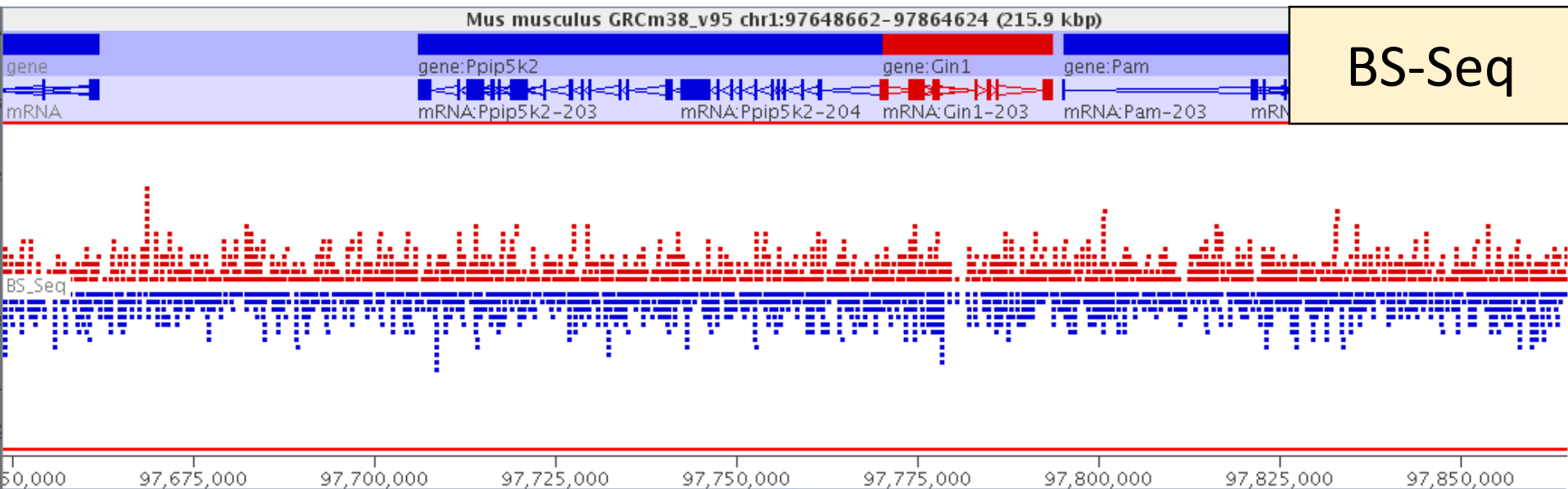
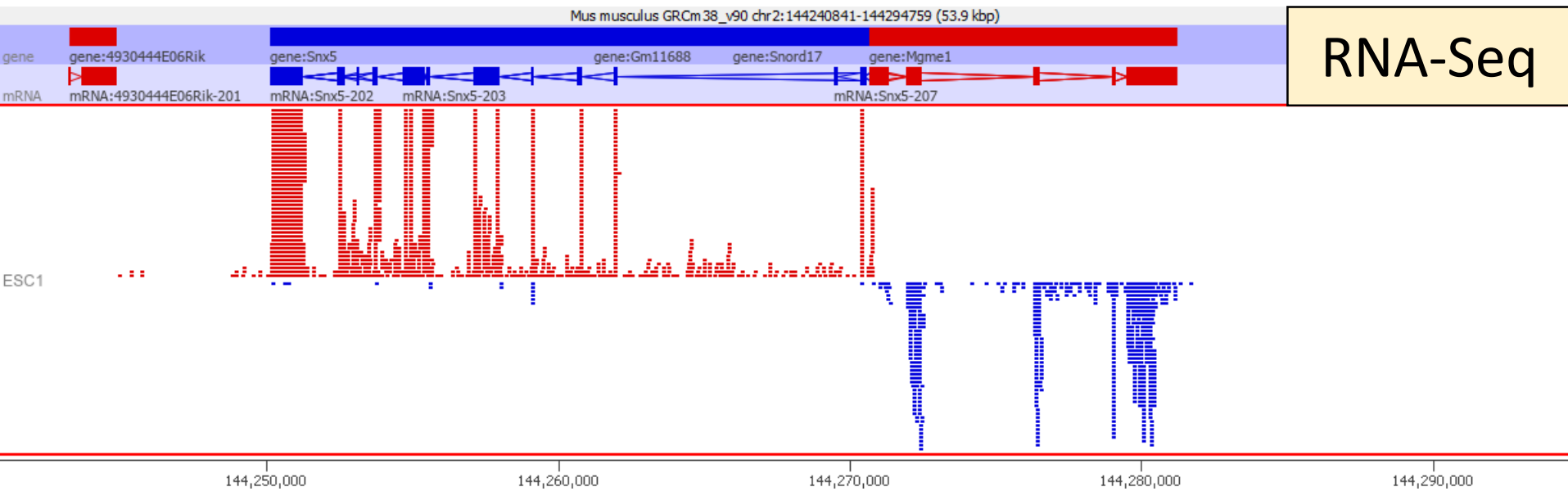
RNA-seq Analysis

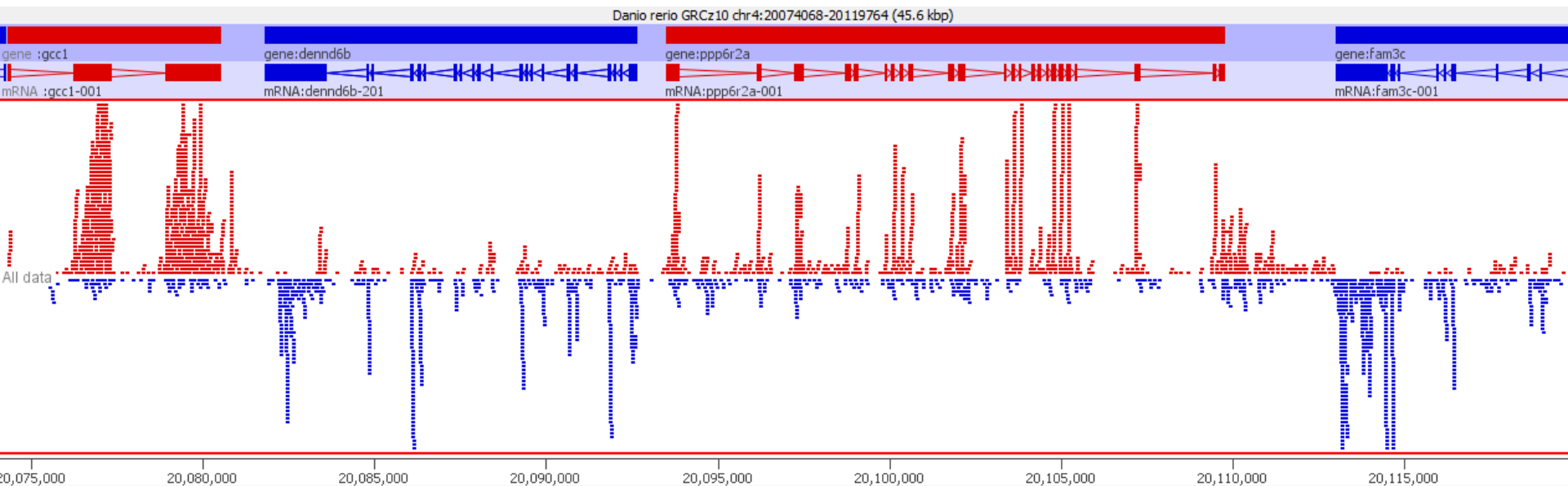
With a few mouse clicks aligned BAM files are imported (including normalization) and the discriminating genes are identified and visualized.

VIPER: Visualization Pipeline for RNA-seq, a Snakemake workflow for efficient and complete RNA-seq analysis



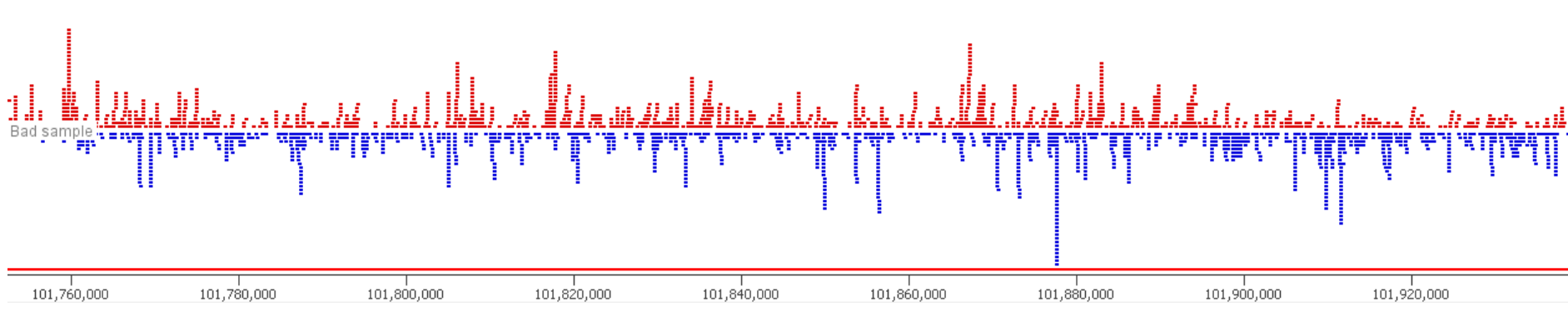
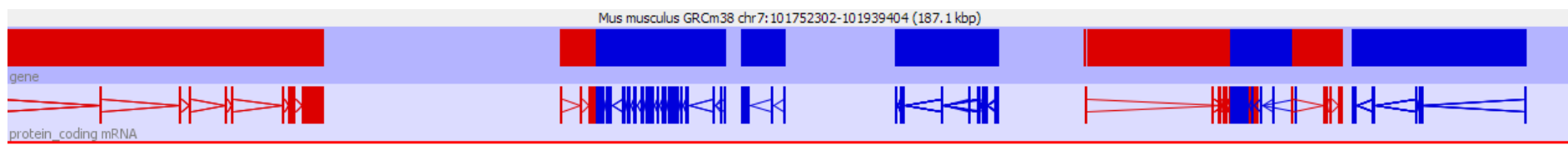
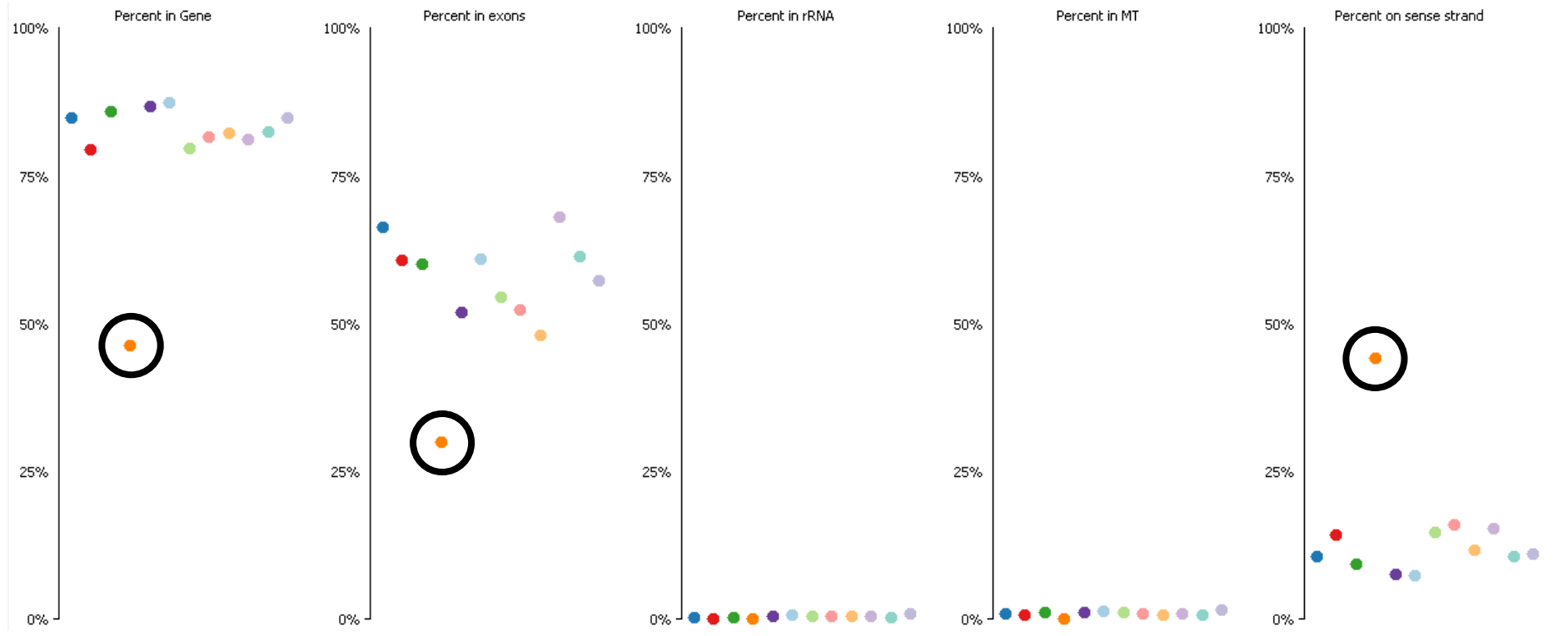






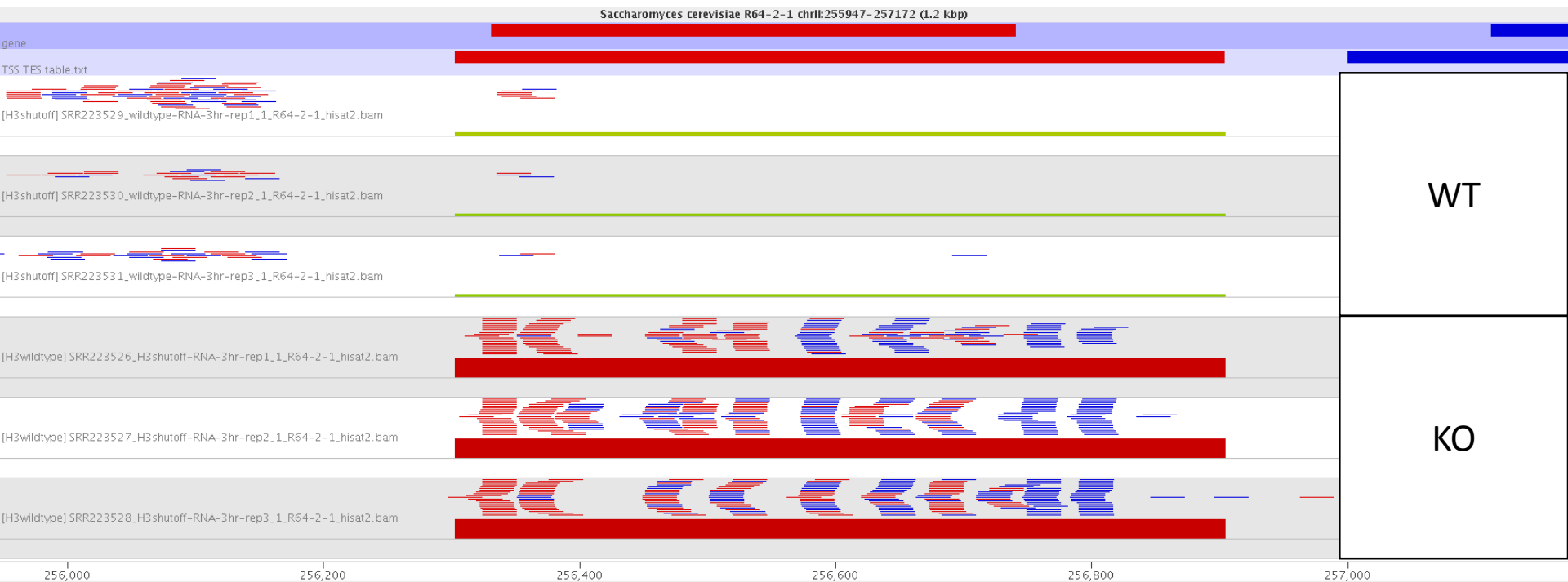
Tris(1,3-dichloro-2-propyl)phosphate Induces Genome-Wide Hypomethylation within Early Zebrafish Embryos

“Moreover, TDCIPP exposure predominantly resulted in **hypomethylation** of positions outside of CpG islands and **within intragenic (exon) regions** of the zebrafish genome.”

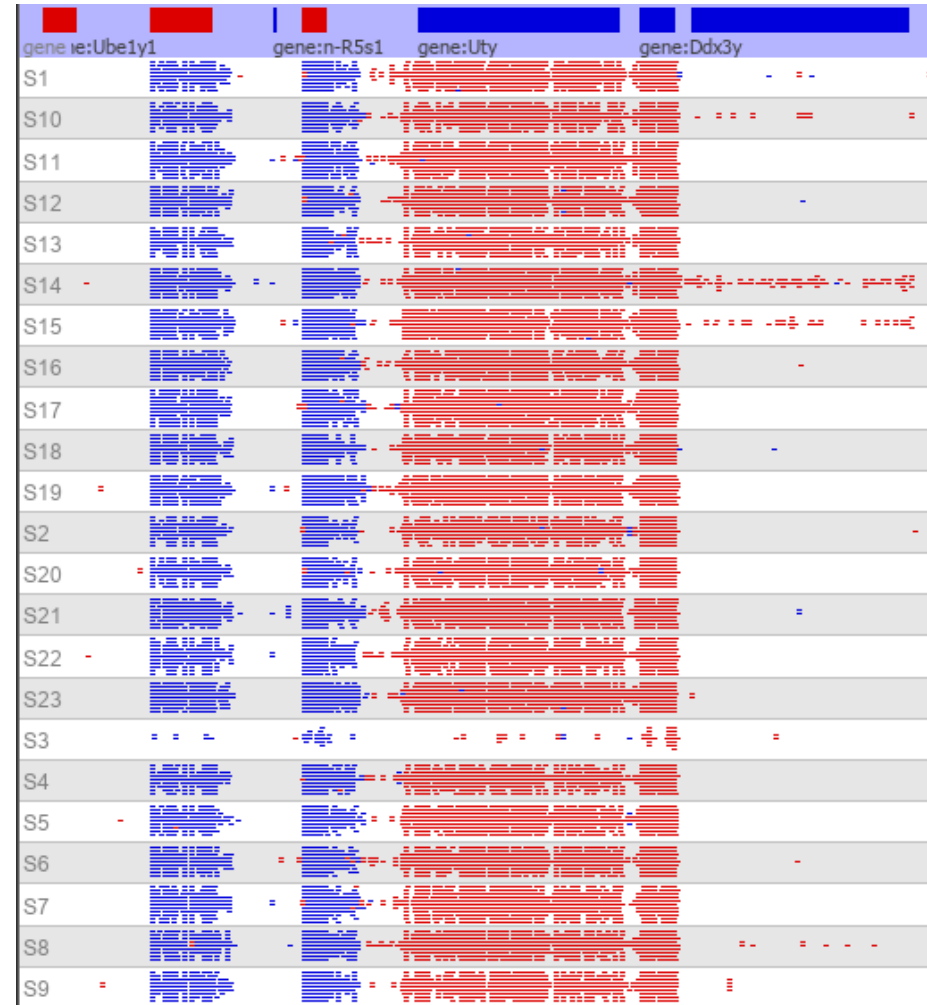
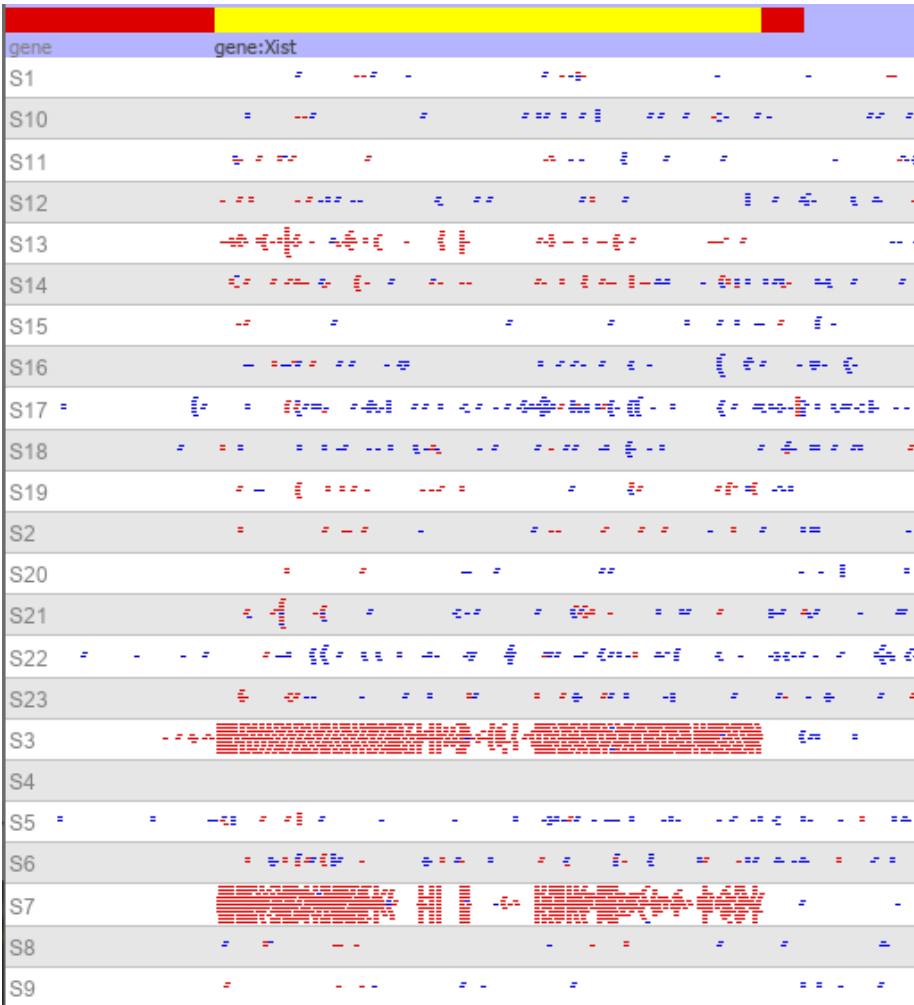


Validate what you know about
your samples

Gene Knockout

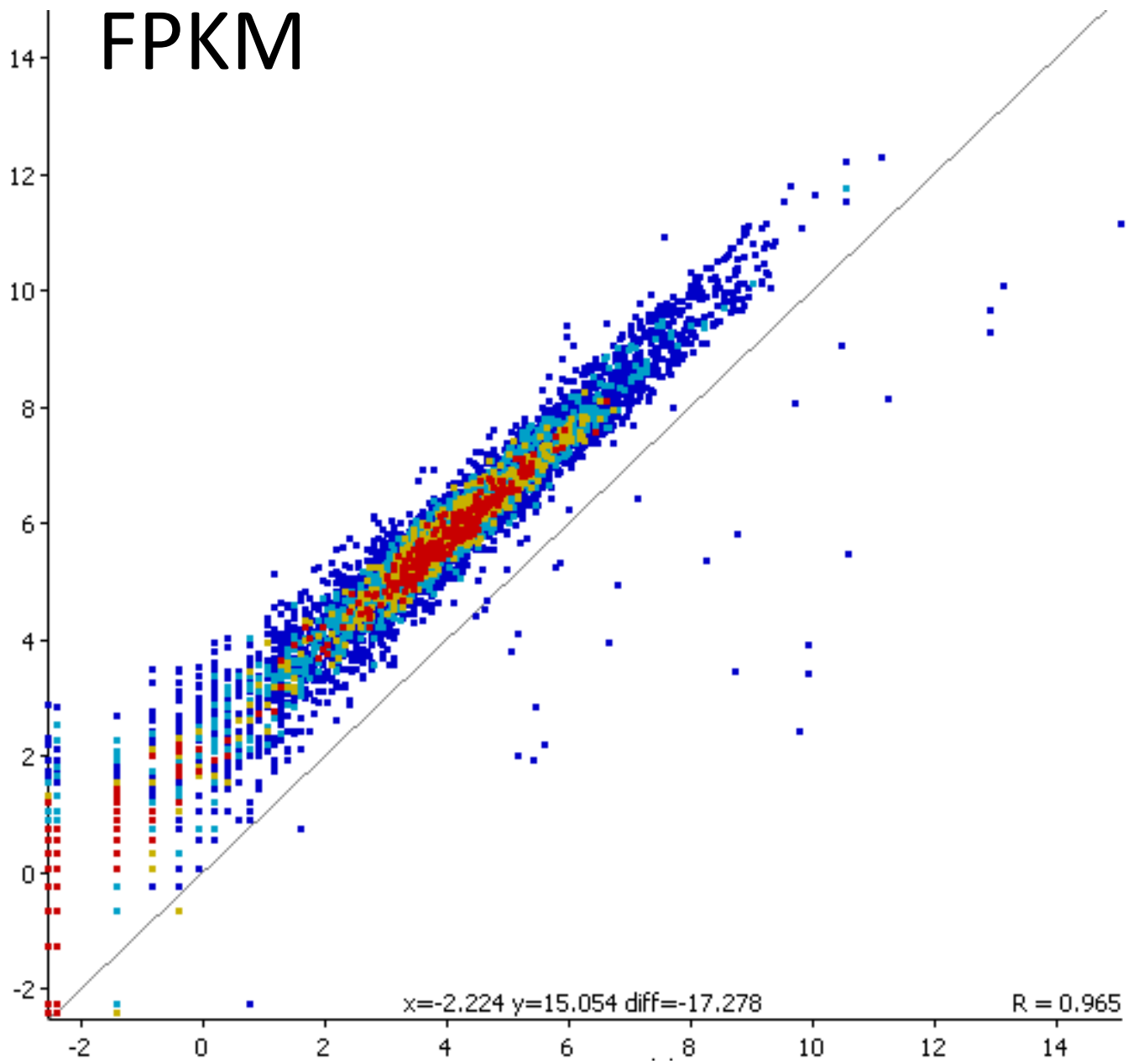


Sample sex

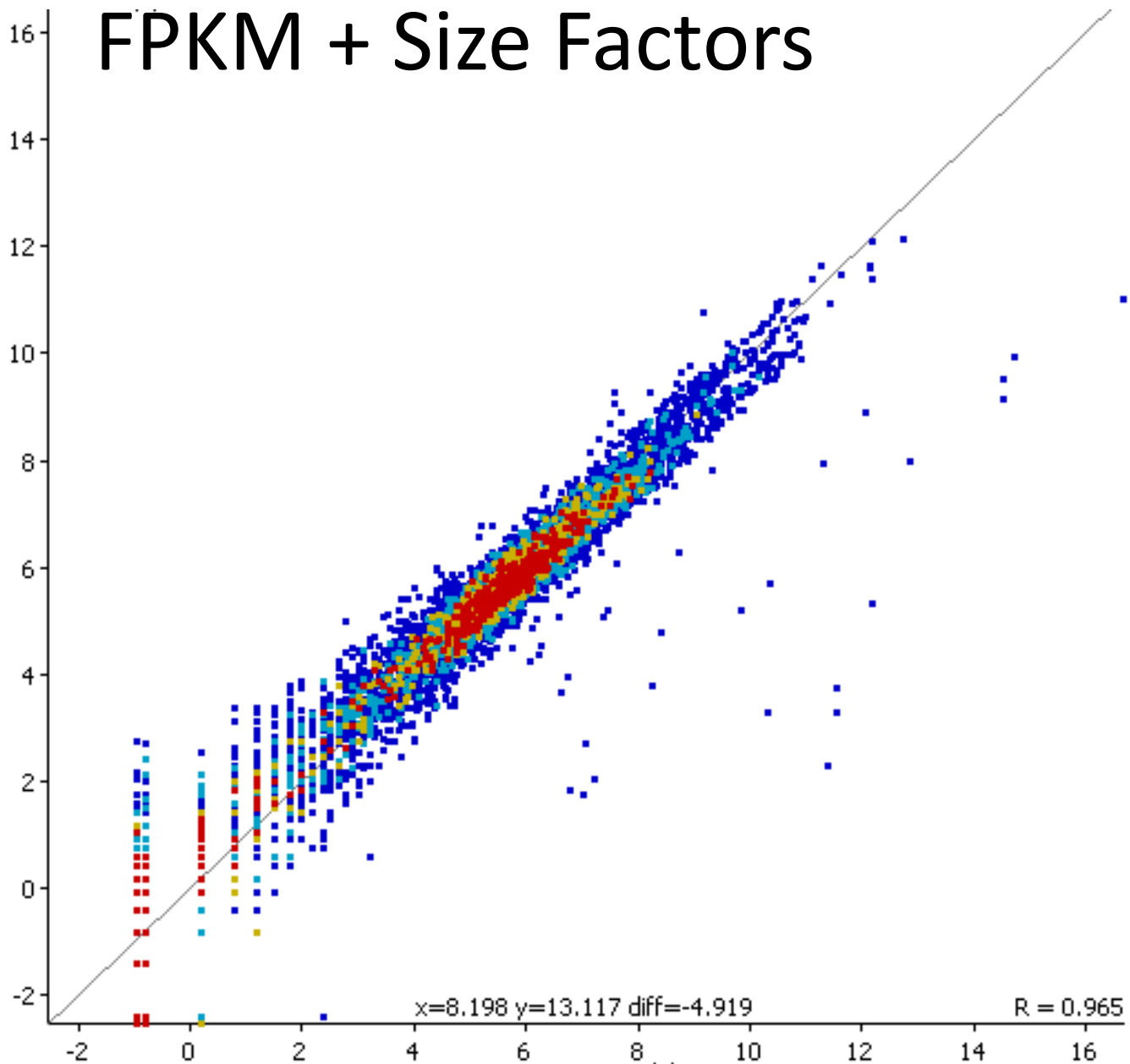


Check your quantitations

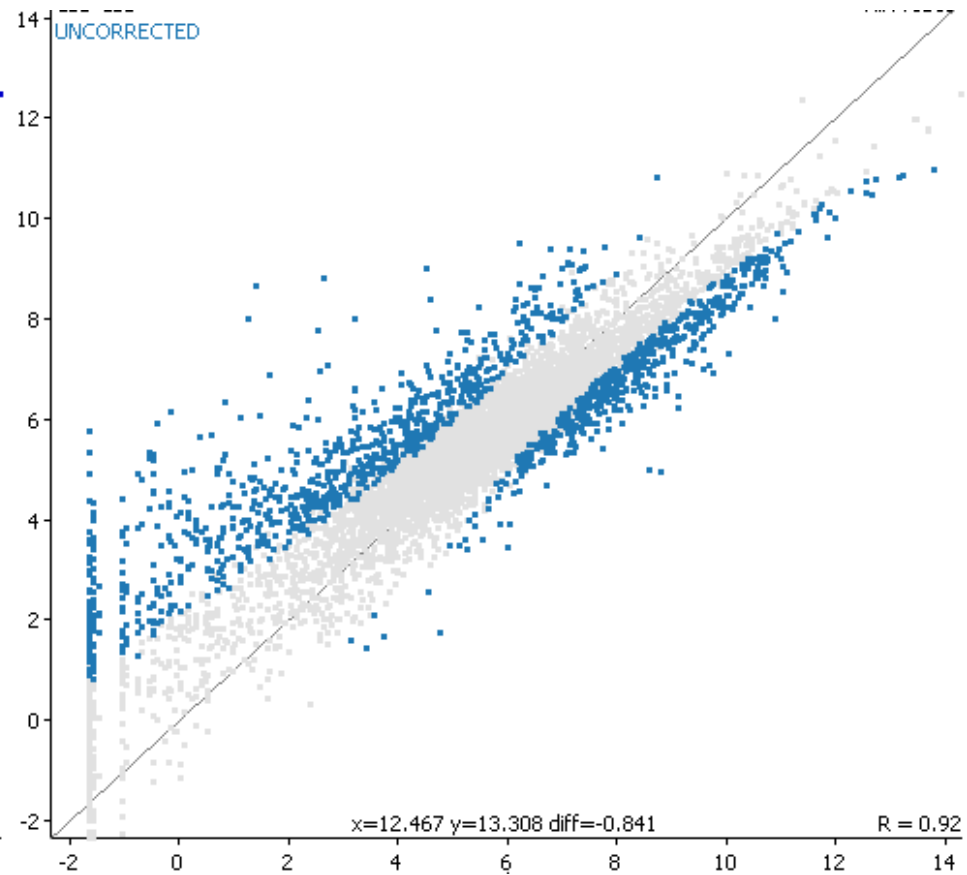
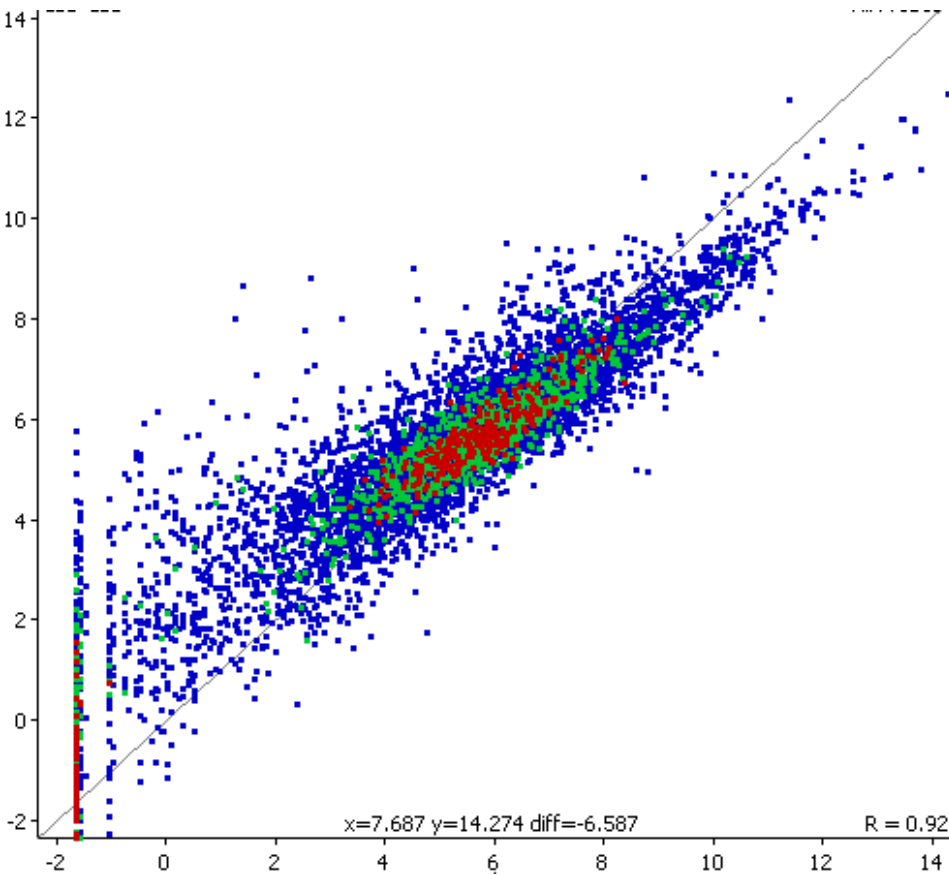
FPKM



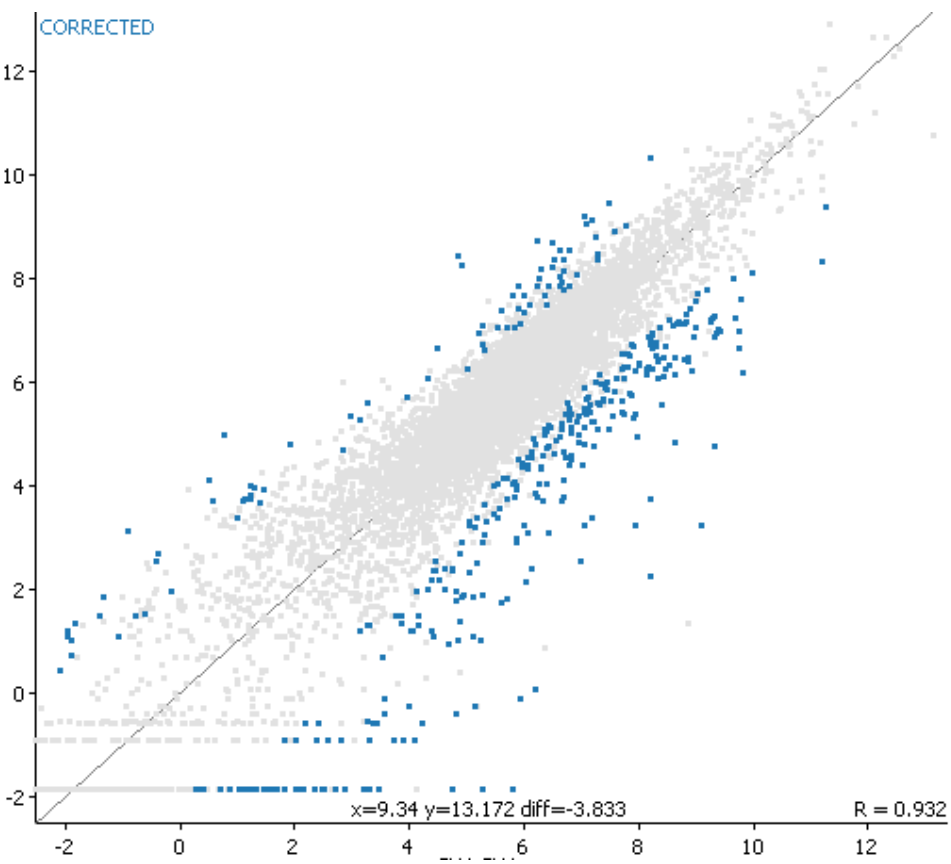
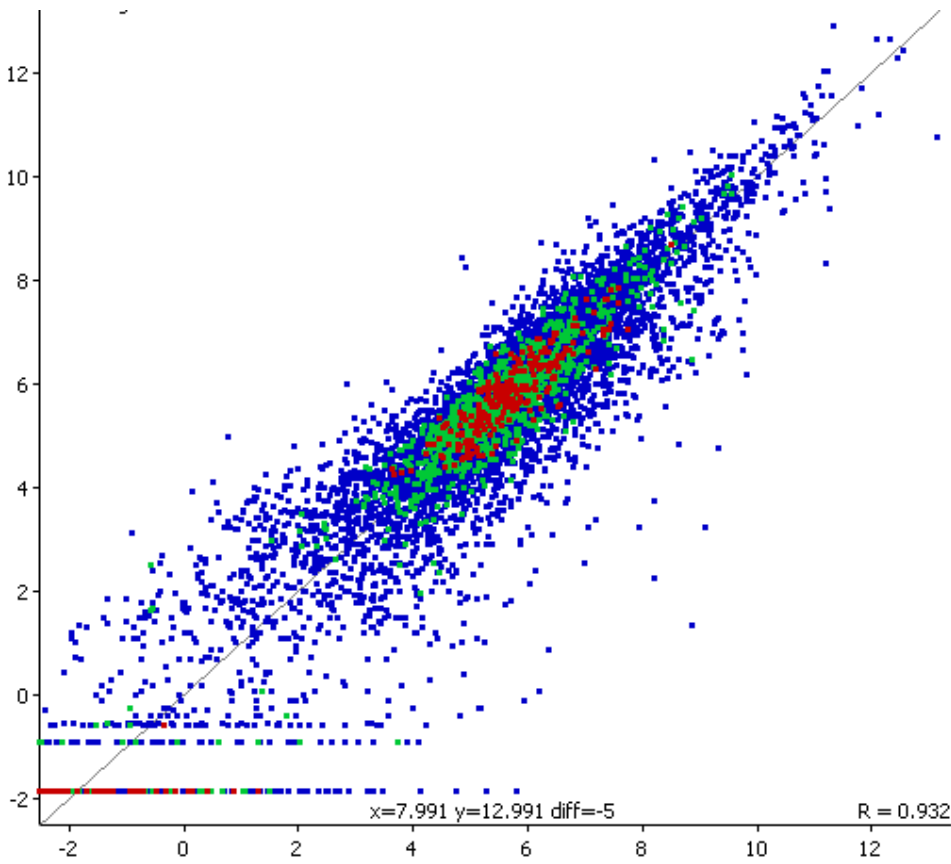
FPKM + Size Factors



FPKM + Size Factors

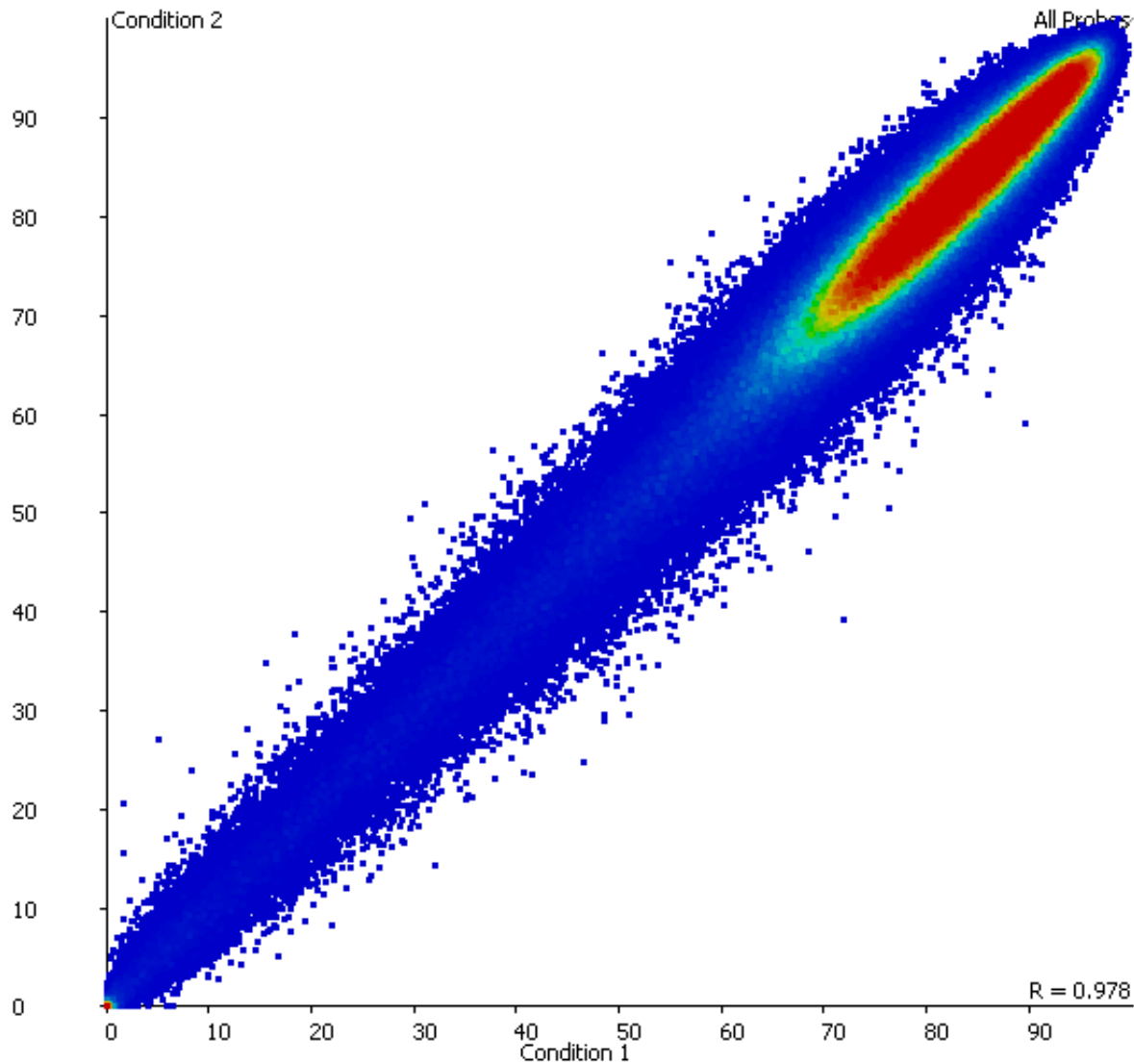


FPKM + Size Factors + Quantile

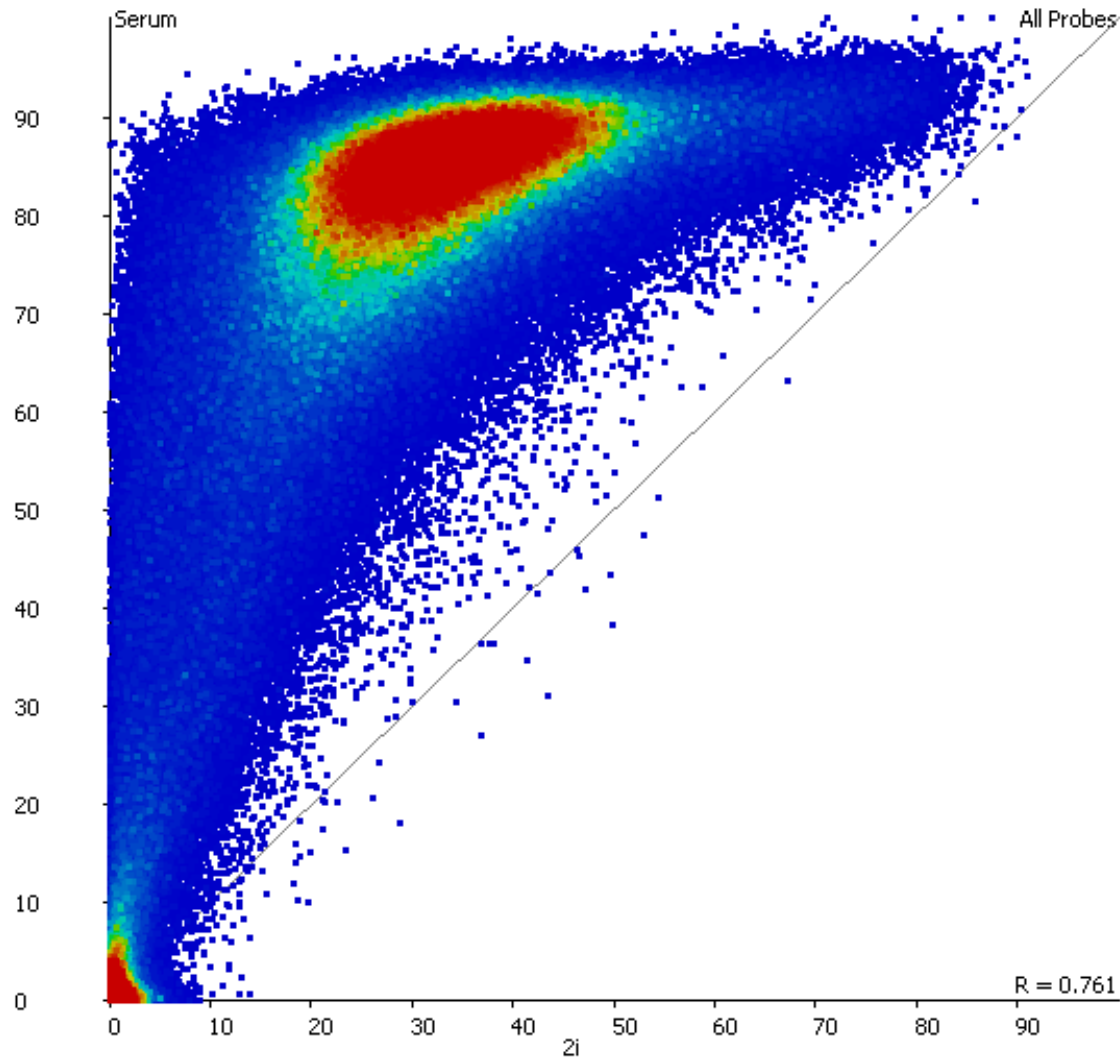


Look for global explanations
before local ones

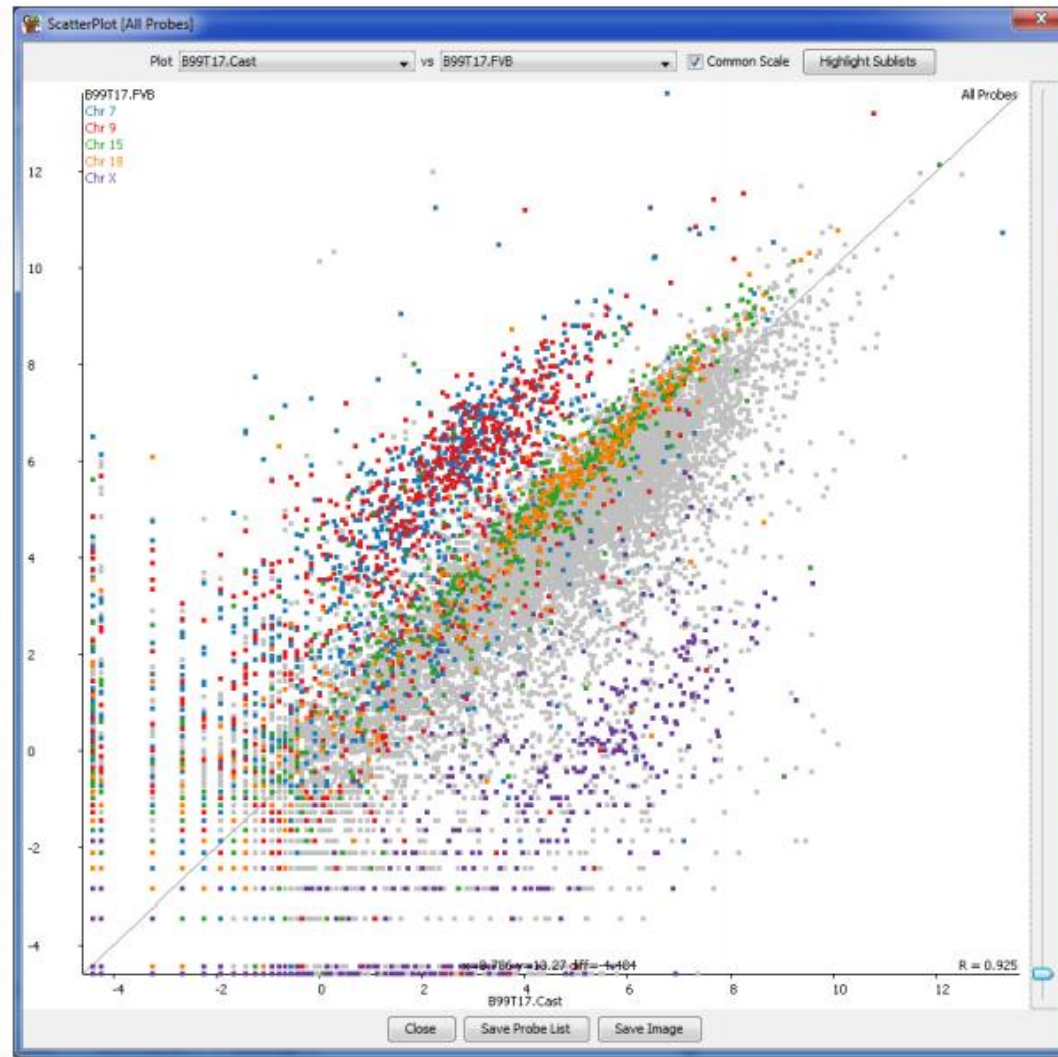
A 'local' explanation makes sense



A 'global' explanation is most important

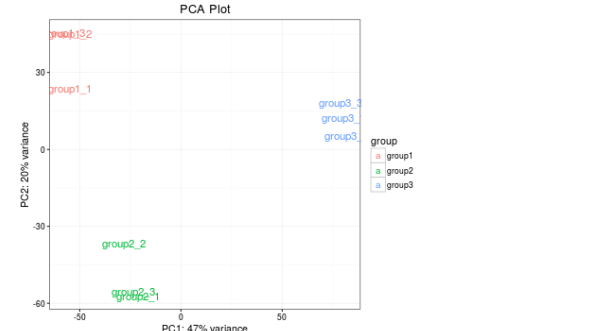
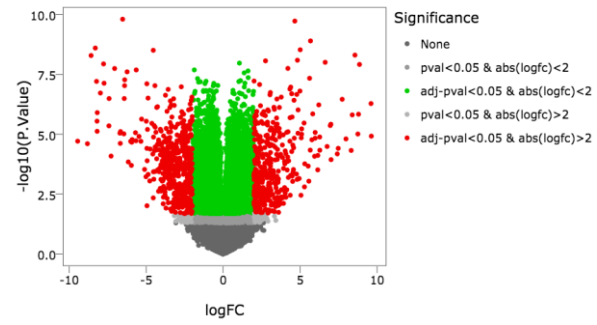
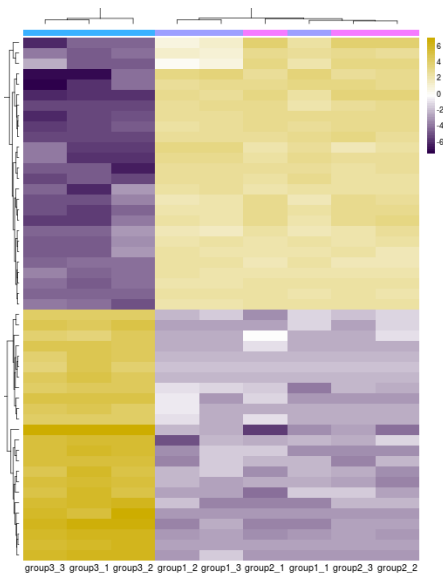
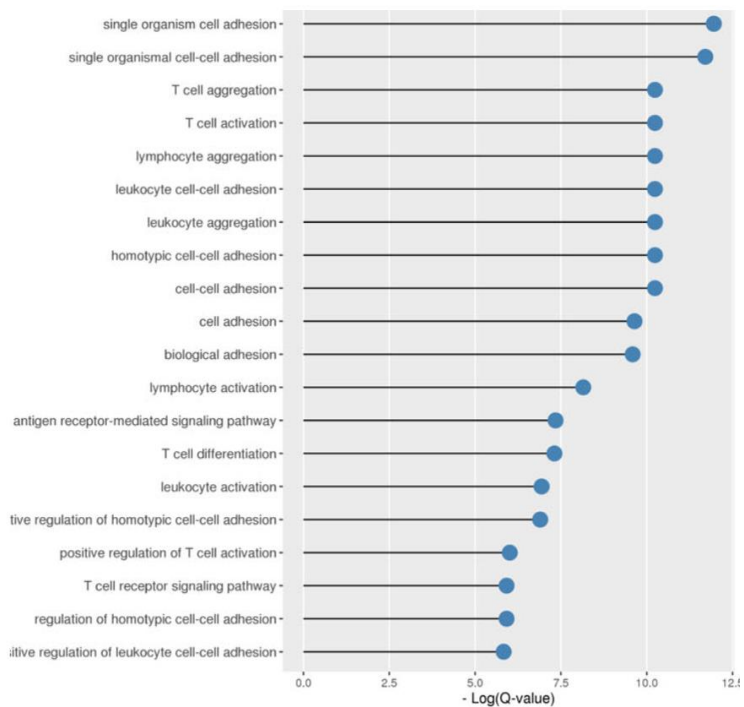


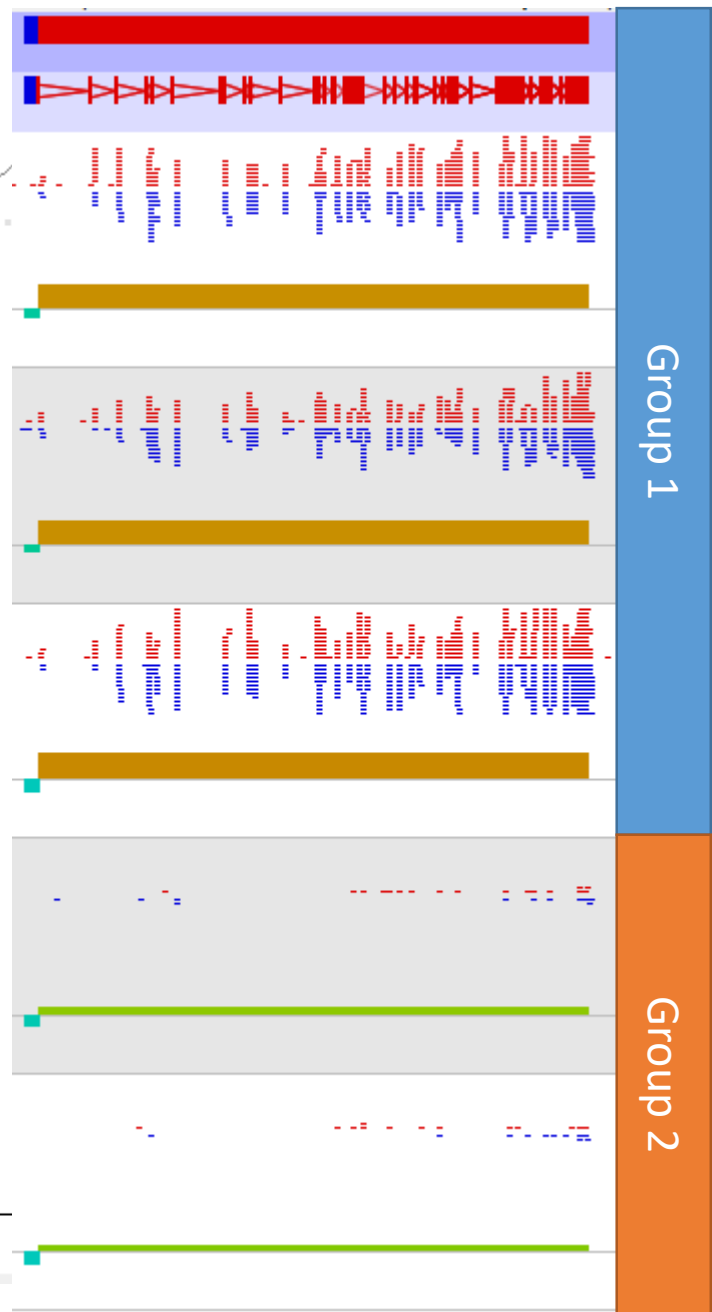
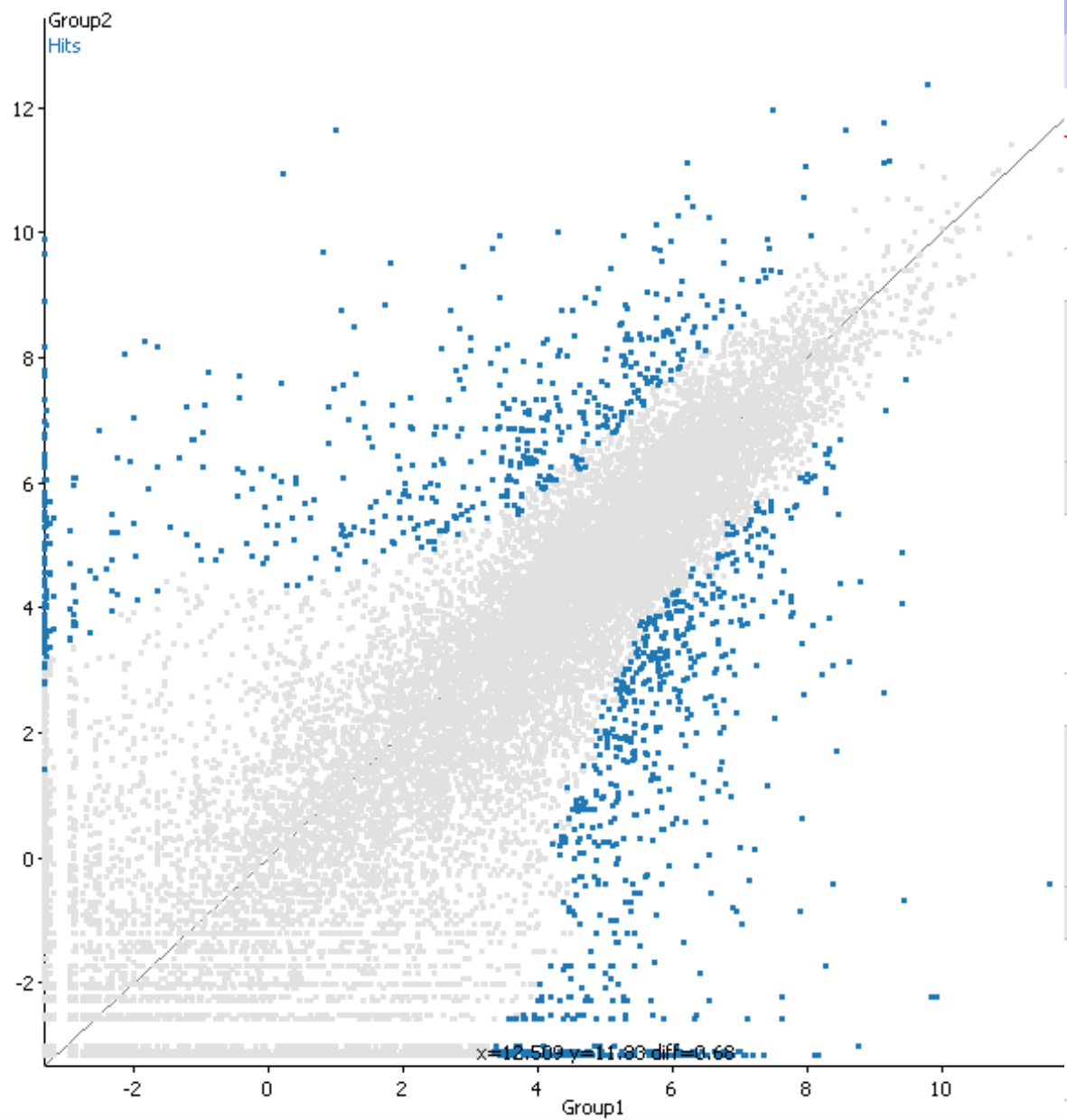
There is obvious structure in the hits

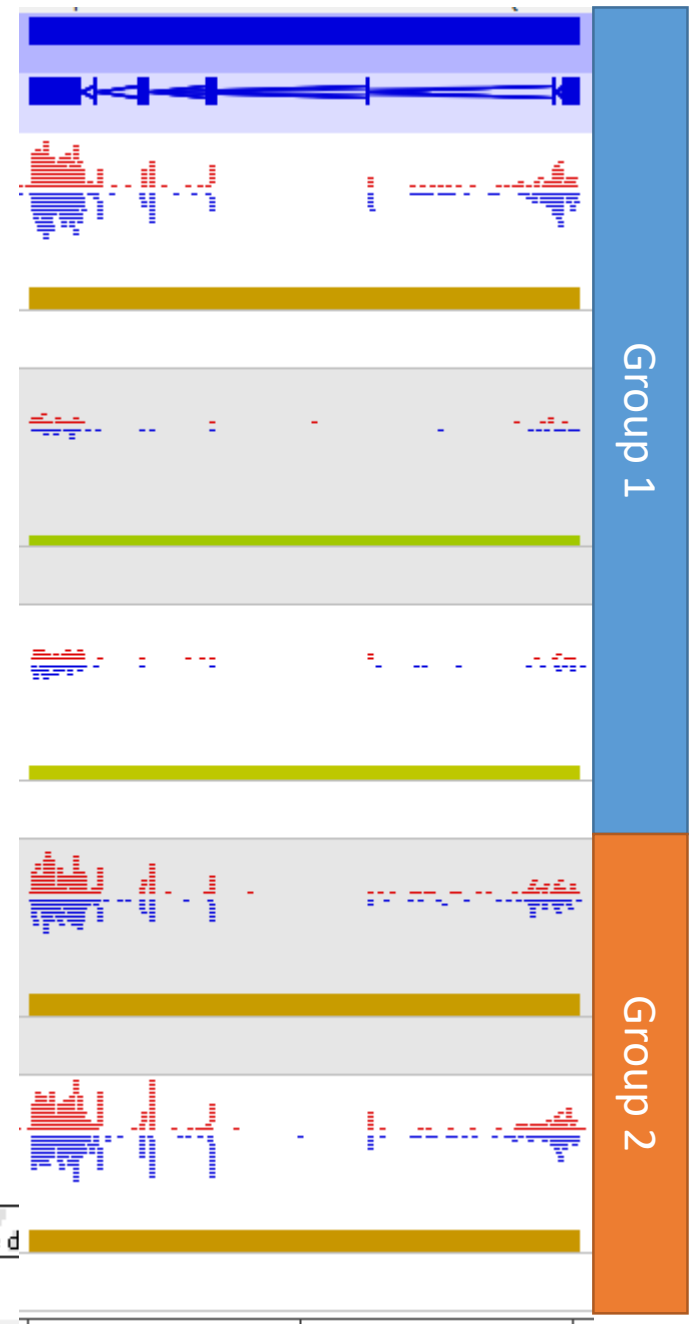
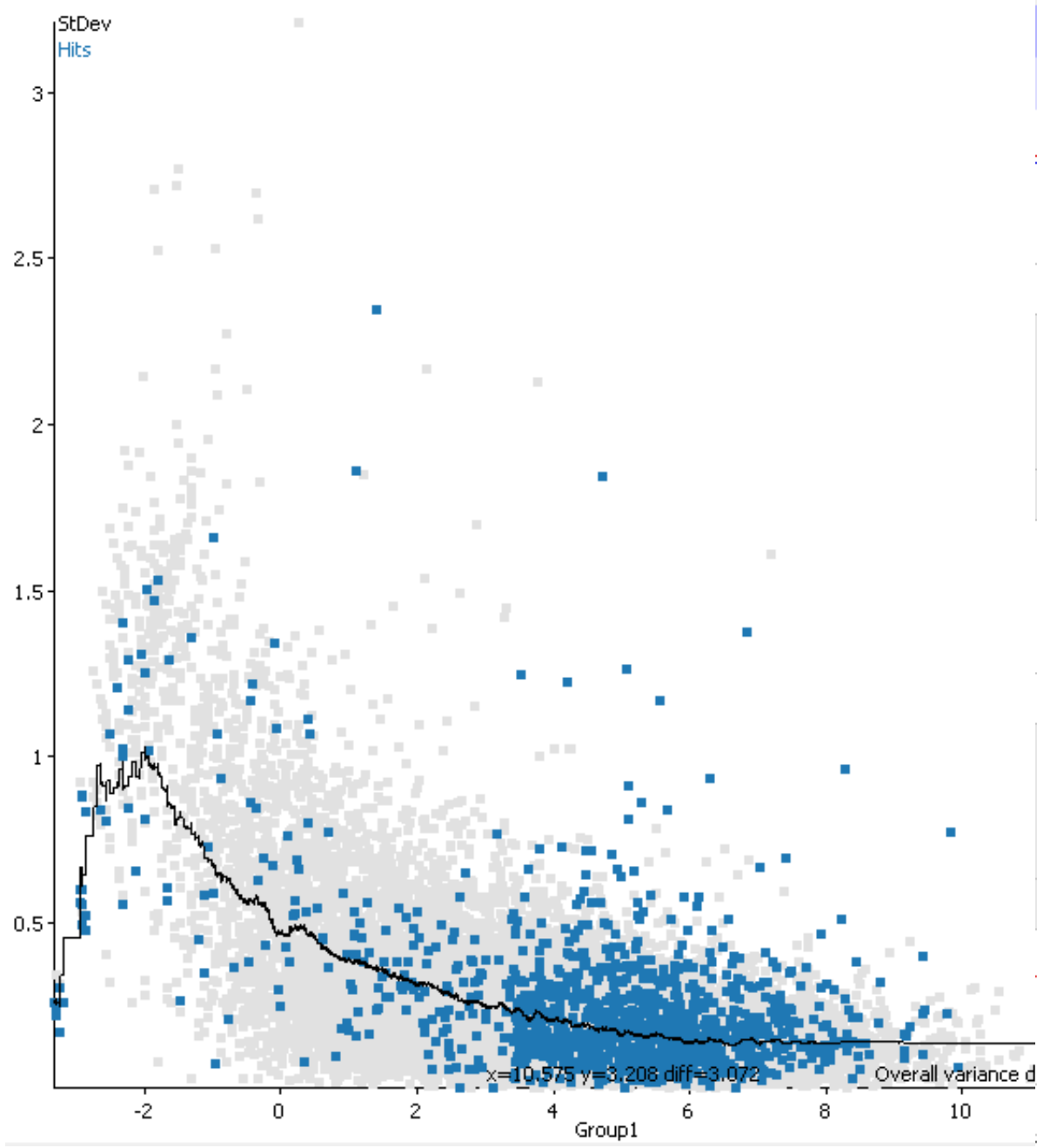


Work backwards through your hits

Gene	ID	Description	P-Value	FDR	Log2 FC
FUT11	ENSG00000196968	fucosyltransferase 11	3.07E-04	0.0010	0.6677
RHOF	ENSG00000139725	ras homolog gene family, member F	3.08E-04	0.0010	0.5691
STAB1	ENSG00000010327	stabilin 1	3.09E-04	0.0010	2.2114
CTNNA1	ENSG00000044115	catenin	3.10E-04	0.0010	0.4730
RAB19	ENSG00000146955	member RAS oncogene family	3.10E-04	0.0010	-2.2223
PPWD1	ENSG00000113593	peptidylprolyl isomerase domain and WD repeat containing 1	3.11E-04	0.0011	0.5757
KCNC3	ENSG00000131398	potassium voltage-gated channel, member 3	3.15E-04	0.0011	-1.0448
CERKL	ENSG00000188452	ceramide kinase-like	3.16E-04	0.0011	1.5089
FBXL8	ENSG00000135722	F-box and leucine-rich repeat protein 8	3.17E-04	0.0011	-1.1472
ZNF488	ENSG00000165388	zinc finger protein 488	3.17E-04	0.0011	-1.4103
FAM82A2	ENSG00000137824	family with sequence similarity 82, member A2	3.17E-04	0.0011	-0.5956
NIT1	ENSG00000158793	nitrilase 1	3.19E-04	0.0011	0.6283







Homo sapiens GRCh37 chr1:206258752-206308625 (49.8 kbp)

Hit1

Hit2

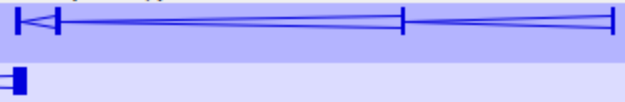
[Group1] Group1_B

[Group1] Group1_A

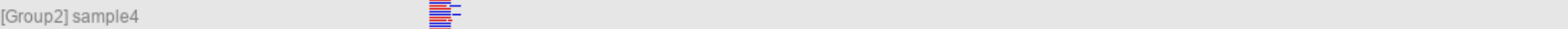
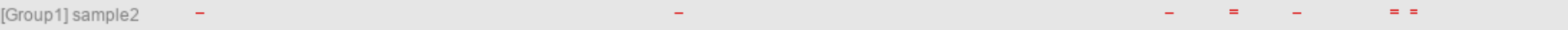
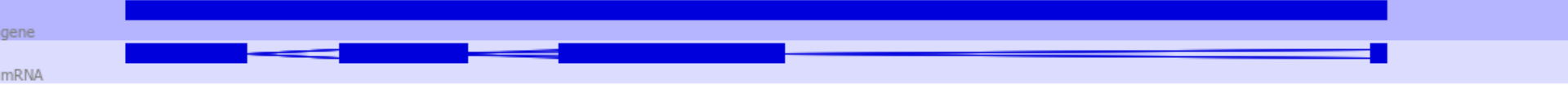
[Group1] Group1_C

[Group2] Group2_A

[Group2] Group2_B



Mus musculus GRCm38 chr11:120447007-120454351 (7.3 kbp)



120,448,000

120,449,000

120,450,000

120,451,000

120,452,000

120,453,000

120,454,000

Summary

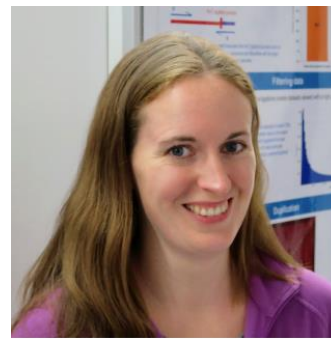
1. Look at your metrics
2. Take notes of errors/warnings
3. Look at your data
4. Validate what you know
5. Check your quantitation
6. Look globally before locally
7. Work backwards through your hits



Anne Segonds-Pichon



Felix Krueger



Laura Biggins



Christel Krueger



Steven Wingett



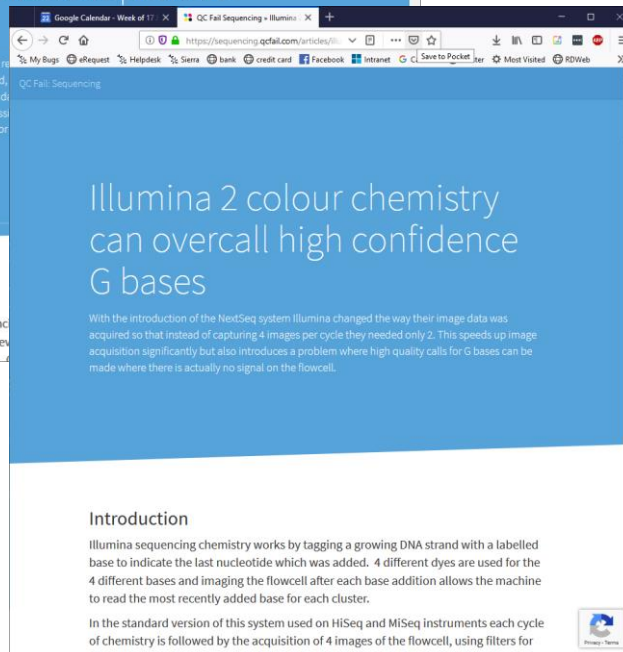
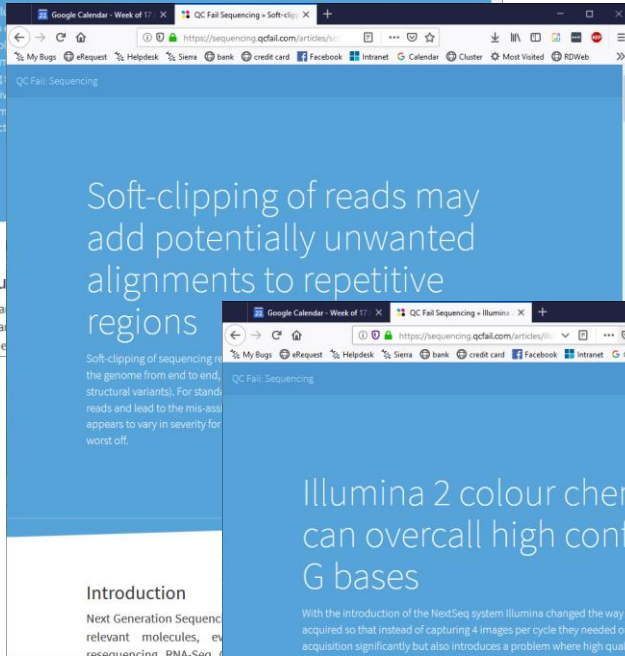
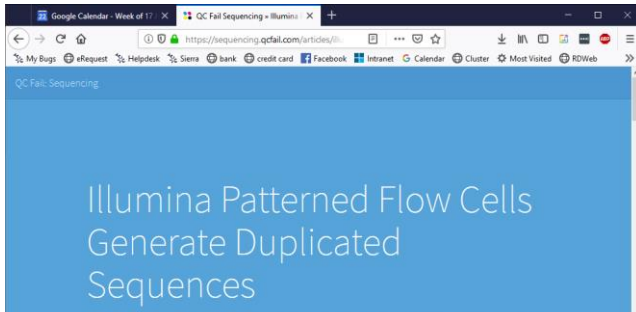
Phil Ewels



www.bioinformatics.babraham.ac.uk

10Xqc.com

qcfail.com



[Sequencing.qcfail.com](https://sequencing.qcfail.com)

[Statistics.qcfail.com](https://statistics.qcfail.com)

[Imaging.qcfail.com](https://imaging.qcfail.com)

[Proteomics.qcfail.com](https://proteomics.qcfail.com)

[Genomics.qcfail.com](https://genomics.qcfail.com)

[Flowcytometry.qcfail.com](https://flowcytometry.qcfail.com)

