MultiQC
v1.7

MultiQC (http://multiqc.info)

A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

This report has been generated by the nf-core/rnaseq (https://github.com/nf-core/rnaseq) analysis pipeline. For information about how to interpret these results, please see the documentation (https://github.com/nf-core/rnaseq/blob/master/docs/output.md).

Report generated on 2019-12-16, 21:31 based on data in: `/scratch/kmddon001/RNAseq_results_Katie/work/90/7faf52dfad1f81e3400272d8fd4e43`

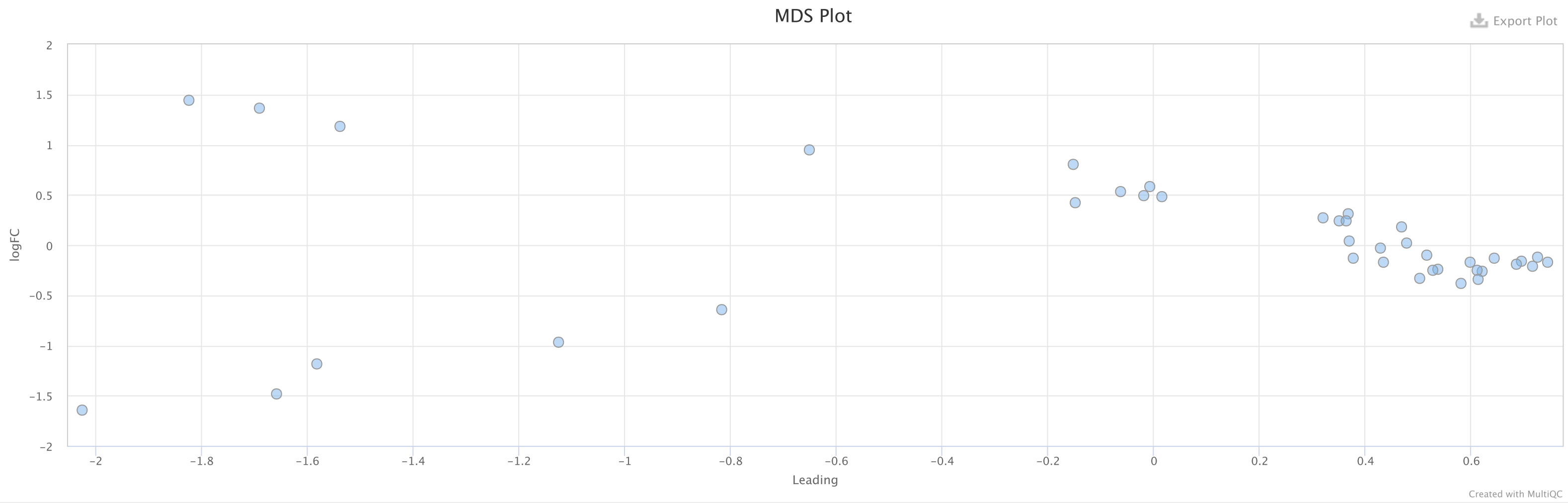
General Statistics

Copy table | Configure Columns | Plot | Showing $^{80}/_{80}$ rows and $^{11}/_{14}$ columns.

Sample Name	dupInt	% rRNA	5'-3' bias	M Aligned	% Assigned	M Assigned	% Aligned	M Aligned	% Trimmed	% GC	M Seqs ▲
JNK_46_S18_L004_R1_001											
JNK_47_S22_L007_R1_001											
JNK_48_S6_L006_R1_001											
JNK_49_S23_L007_R1_001											
JNK_42_S19_L007											
JNK_43_S20_L007											
JNK_44_S21_L007											
JNK_47_S22_L007											
JNK_49_S23_L007											
JNK_30_S25_L008											
JNK_46_S18_L004											
JNK_14_S5_L001											
JNK_19_S14_L004											
JNK_20_S10_L002											
JNK_48_S6_L006											

MDS Plot

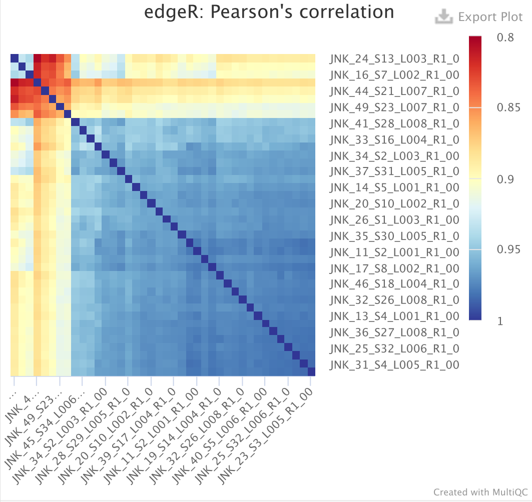
MDS Plot show relatedness between samples in a project. These values are calculated using edgeR (https://bioconductor.org/packages/release/bioc/html/edgeR.html) in the `edgeR_heatmap_MDS.r` (https://github.com/nf-core/rnaseq/blob/master/bin/edgeR_heatmap_MDS.r) script.



edgeR: Sample Similarity

edgeR: Sample Similarity is generated from normalised gene counts through edgeR (https://bioconductor.org/packages/release/bioc/html/edgeR.html). Pearson's correlation between $\log_2$ normalised CPM values are then calculated and clustered.

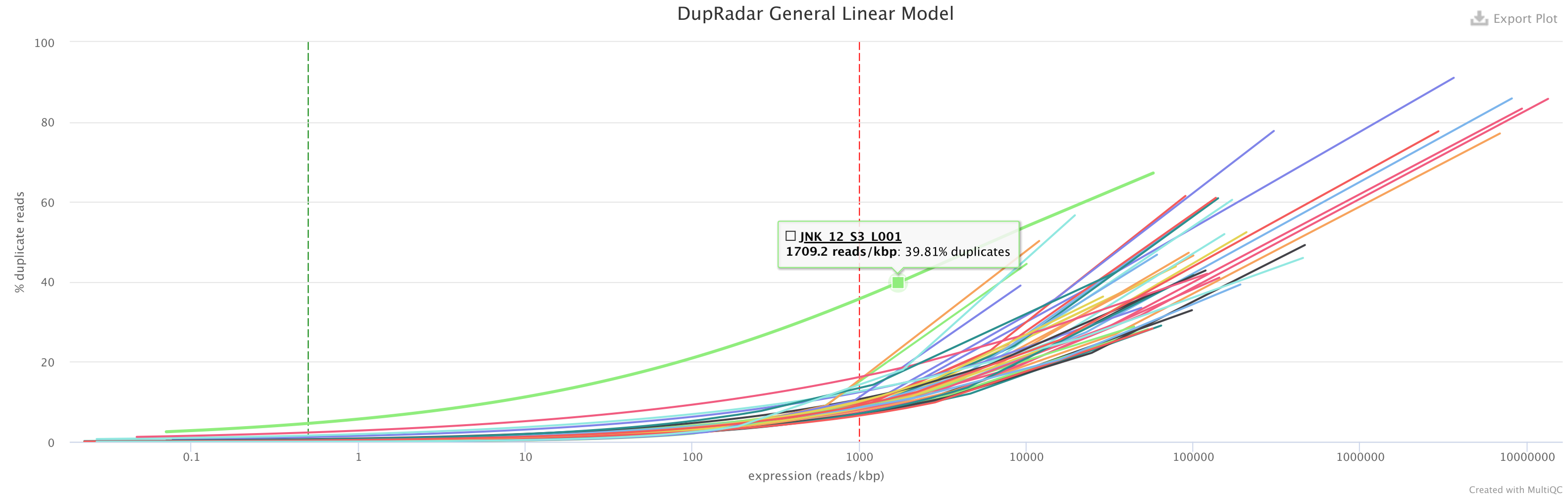
Sort by highlight



DupRadar

DupRadar (bioconductor.org/packages/release/bioc/html/dupRadar.html) provides duplication rate quality control for RNA-Seq datasets. Highly expressed genes can be expected to have a lot of duplicate reads, but high numbers of duplicates at low read counts can indicate low library complexity with technical duplication. This plot shows the general linear models - a summary of the gene duplication distributions.

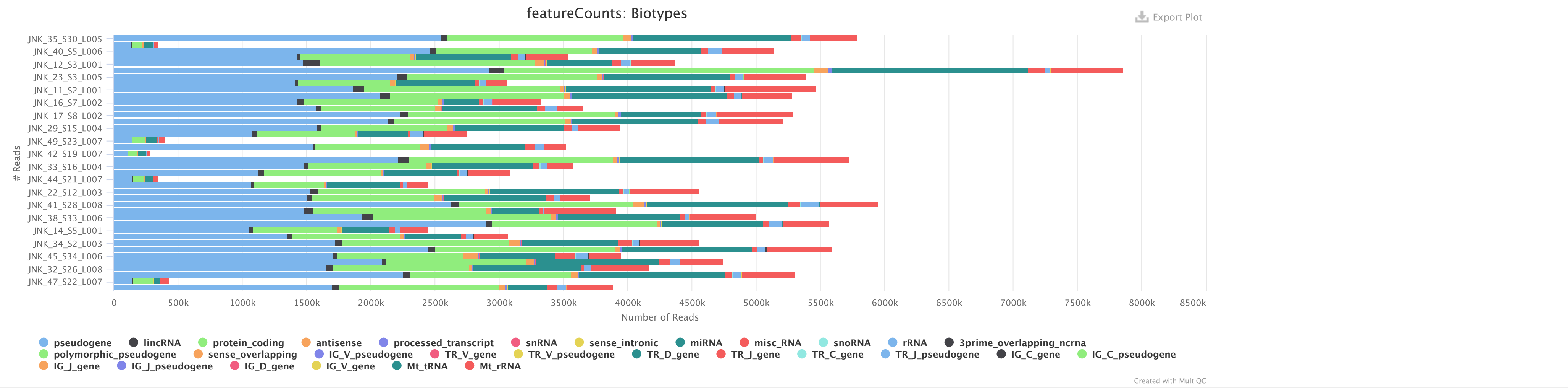
Y-Limits: on



Biotype Counts

Biotype Counts shows reads overlapping genomic features of different biotypes, counted by featureCounts (http://bioinf.wehi.edu.au/featureCounts).

Number of Reads | Percentages



QualiMap

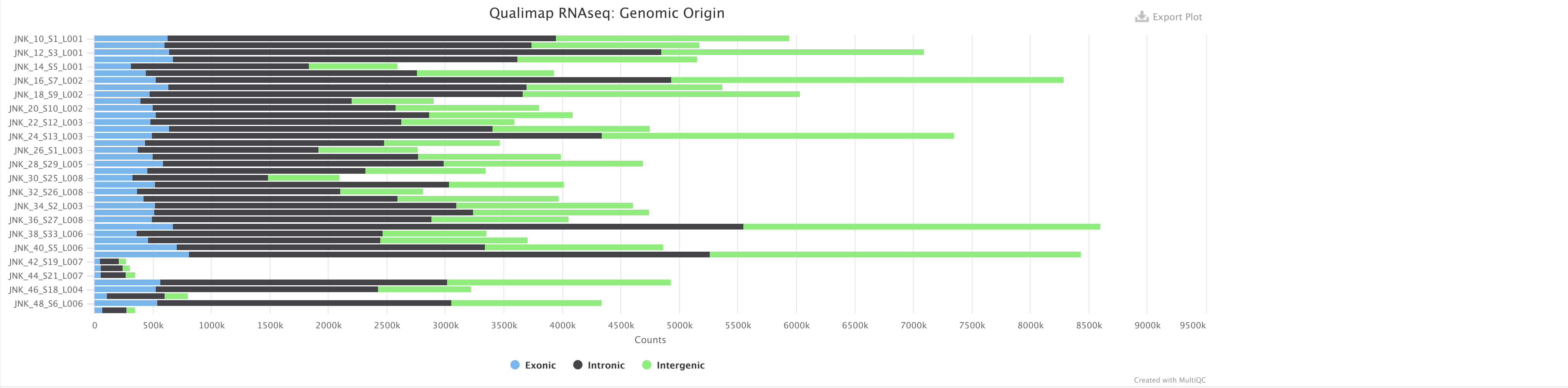
QualiMap (http://qualimap.bioinfo.cipf.es/) is a platform-independent application to facilitate the quality control of alignment sequencing data and its derivatives like feature counts.

Genomic origin of reads

❓ Help

Classification of mapped reads as originating in exonic, intronic or intergenic regions. These can be displayed as either the number or percentage of mapped reads.

Counts | Percentages



Gene Coverage Profile

❓ Help

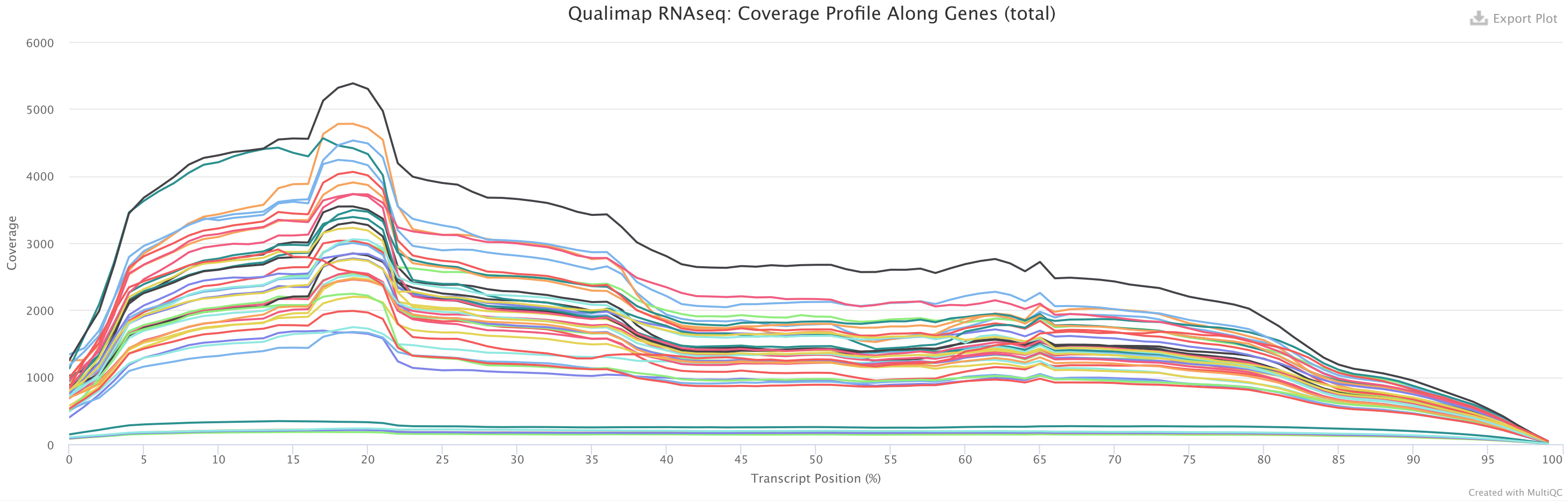
Mean distribution of coverage depth across the length of all mapped transcripts.

There are currently three main approaches to map reads to transcripts in an RNA-seq experiment: mapping reads to a reference genome to identify expressed transcripts that are annotated (and discover those that are unknown), mapping reads to a reference transcriptome, and *de novo* assembly of transcript sequences (Conesa et al. 2016 (https://doi.org/10.1186/s13059-016-0881-8)).

For RNA-seq QC analysis, QualiMap can be used to assess alignments produced by the first of these approaches. For input, it requires a GTF annotation file along with a reference genome, which can be used to reconstruct the exon structure of known transcripts. QualiMap uses this information to calculate the depth of coverage along the length of each annotated transcript. For a set of reads mapped to a transcript, the depth of coverage at a given base position is the number of high-quality reads that map to the transcript at that position (Sims et al. 2014 (https://doi.org/10.1038/nrg3642)).

QualiMap calculates coverage depth at every base position of each annotated transcript. To enable meaningful comparison between transcripts, base positions are rescaled to relative positions expressed as percentage distance along each transcript (*0%, 1%, …, 99%*). For the set of transcripts with at least one mapped read, QualiMap plots the cumulative mapped-read depth (y-axis) at each relative transcript position (x-axis). This plot shows the gene coverage profile across all mapped transcripts for each read dataset. It provides a visual way to assess positional biases, such as an accumulation of mapped reads at the 3′ end of transcripts, which may indicate poor RNA quality in the original sample (Conesa et al. 2016 (https://doi.org/10.1186/s13059-016-0881-8)).

Y-Limits: on



Preseq

Preseq (http://smithlabresearch.org/software/preseq/) estimates the complexity of a library, showing how many additional unique reads are sequenced for increasing total read count. A shallow curve indicates complexity saturation. The dashed line shows a perfectly complex library where total reads = unique reads.

Complexity curve

Note that the x axis is trimmed at the point where all the datasets show 80% of their maximum y-value, to avoid ridiculous scales.

Y-Limits: on

Preseq: Complexity curve

Export Plot

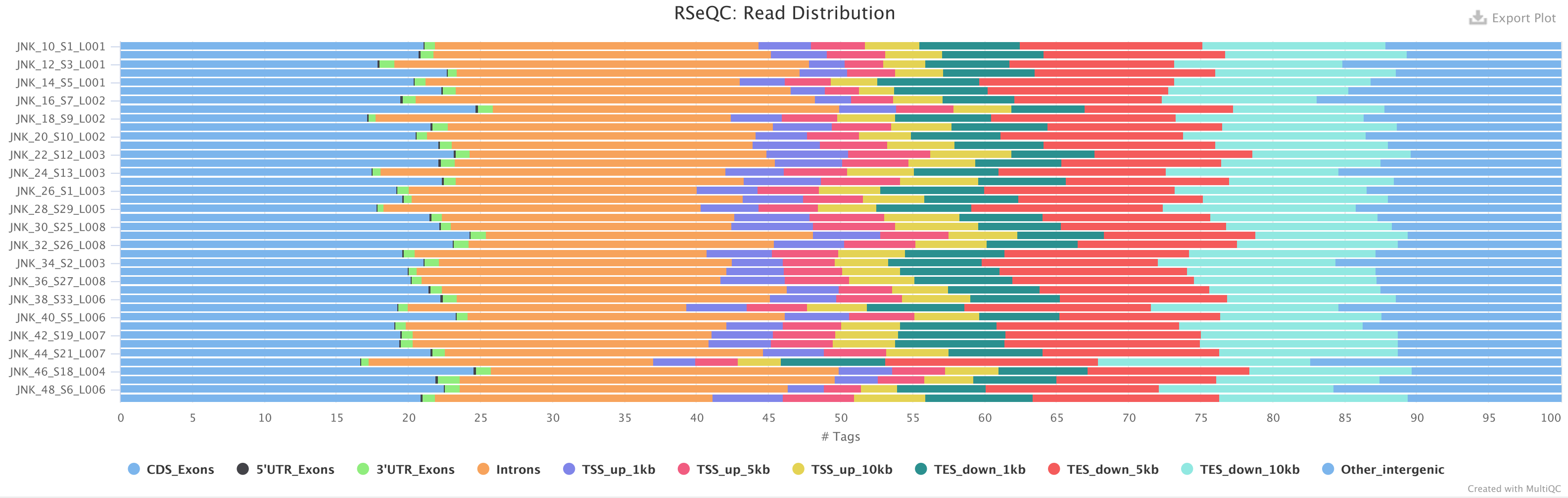
RSeQC

RSeQC (http://rseqc.sourceforge.net/) package provides a number of useful modules that can comprehensively evaluate high throughput RNA-seq data.

Read Distribution

Read Distribution (http://rseqc.sourceforge.net/#read-distribution-py) calculates how mapped reads are distributed over genome features.

Number of Tags | Percentages



Inner Distance

Inner Distance (http://rseqc.sourceforge.net/#inner-distance-py) calculates the inner distance (or insert size) between two paired RNA reads. Note that this can be negative if fragments overlap.

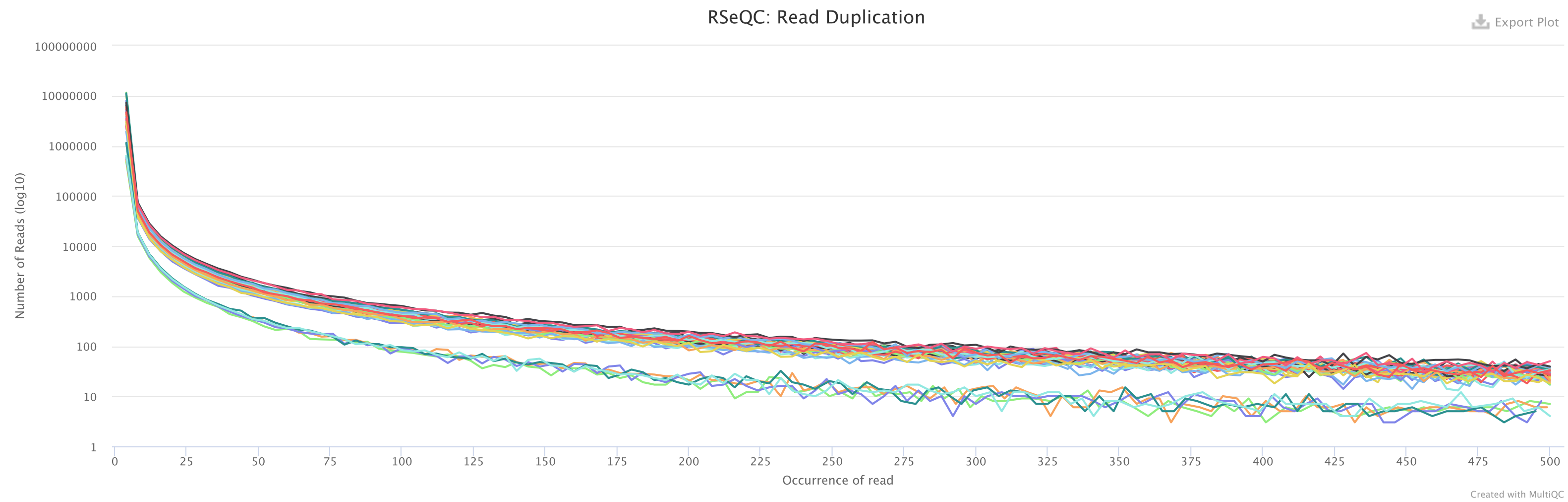
Counts | Percentages

RSeQC: Inner Distance

Export Plot

Read Duplication

read_duplication.py (http://rseqc.sourceforge.net/#read-duplication-py) calculates how many alignment positions have a certain number of exact duplicates. Note - plot truncated at 500 occurrences and binned.



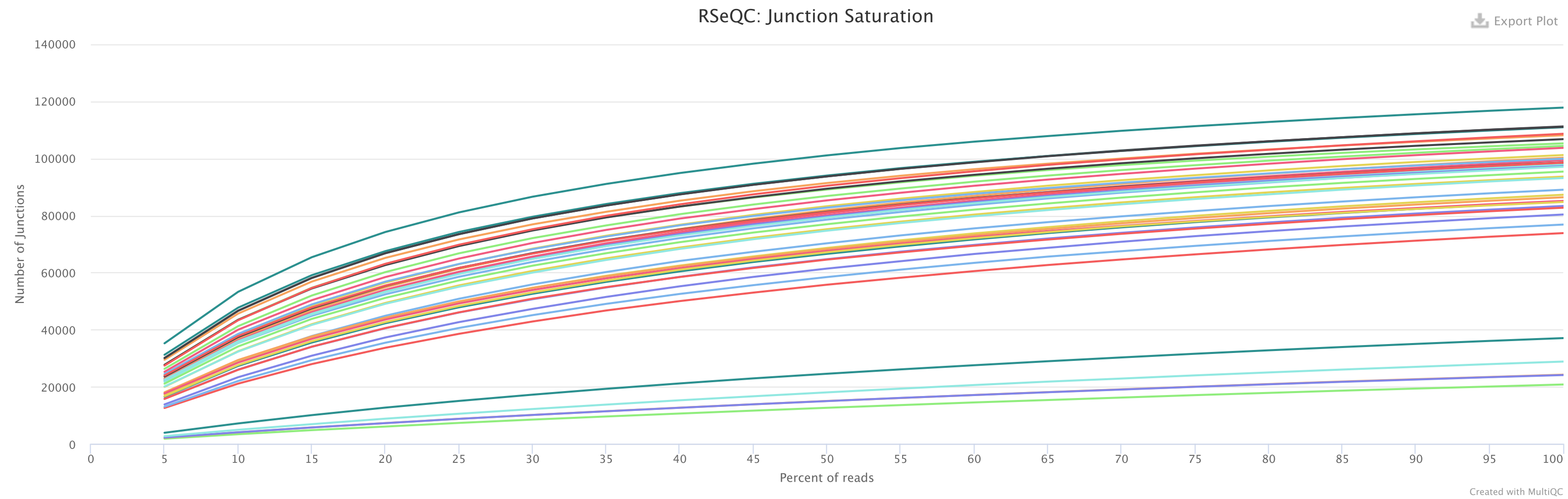
Junction Saturation

Junction Saturation (http://rseqc.sourceforge.net/#junction-saturation-py) counts the number of known splicing junctions that are observed in each dataset. If sequencing depth is sufficient, all (annotated) splice junctions should be rediscovered, resulting in a curve that reaches a plateau. Missing low abundance splice junctions can affect downstream analysis.

Click a line to see the data side by side (as in the original RSeQC plot).

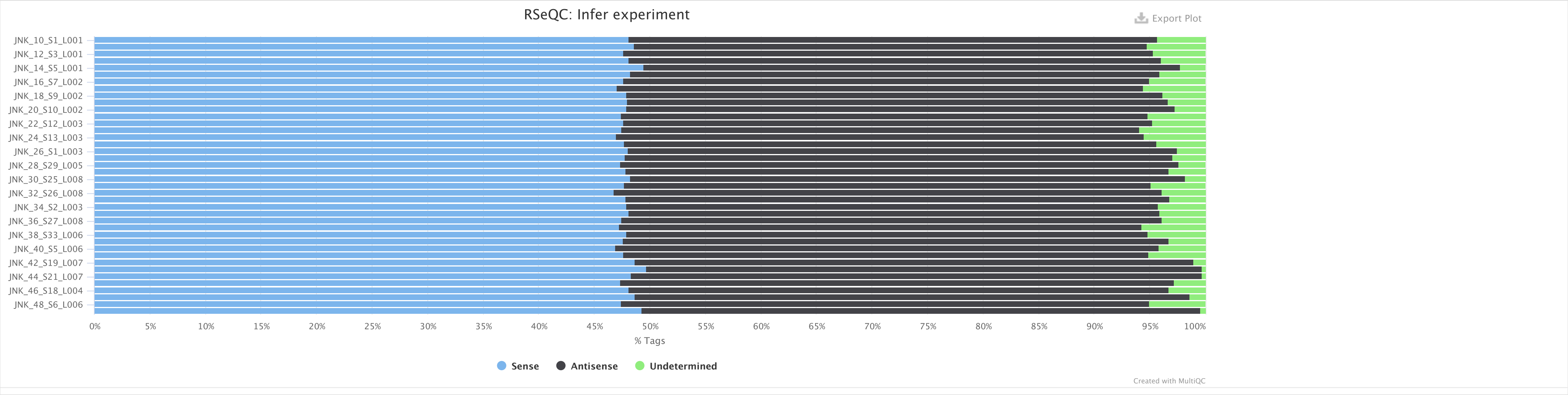
Y-Limits: on

Known Junctions | Novel Junctions | All Junctions



Infer experiment

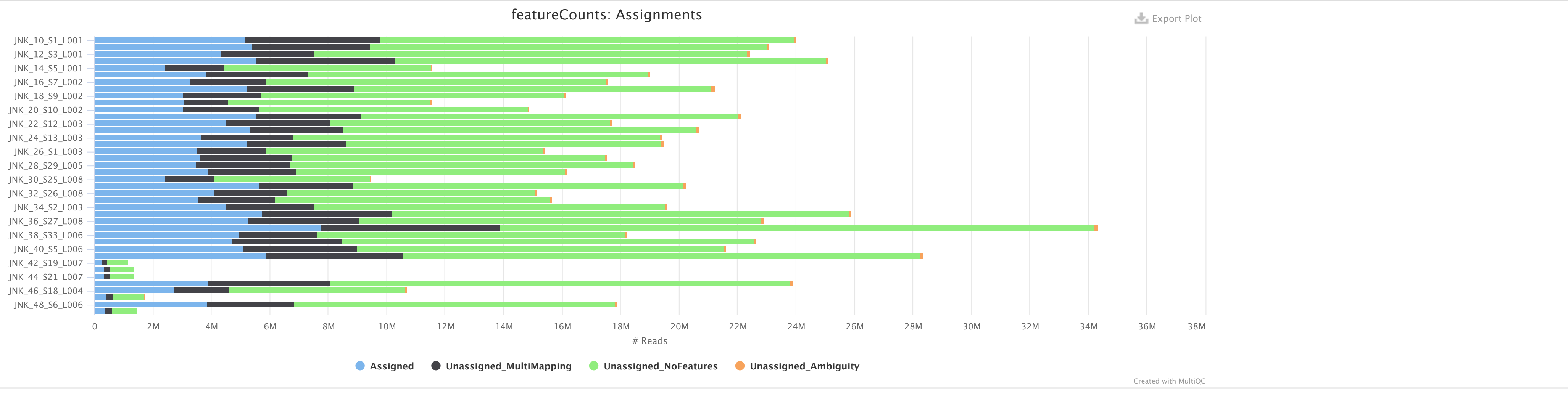
Infer experiment (http://rseqc.sourceforge.net/#infer-experiment-py) counts the percentage of reads and read pairs that match the strandedness of overlapping transcripts. It can be used to infer whether RNA-seq library preps are stranded (sense or antisense).



featureCounts

Subread featureCounts (http://bioinf.wehi.edu.au/featureCounts/) is a highly efficient general-purpose read summarization program that counts mapped reads for genomic features such as genes, exons, promoter, gene bodies, genomic bins and chromosomal locations.

Number of Reads | Percentages

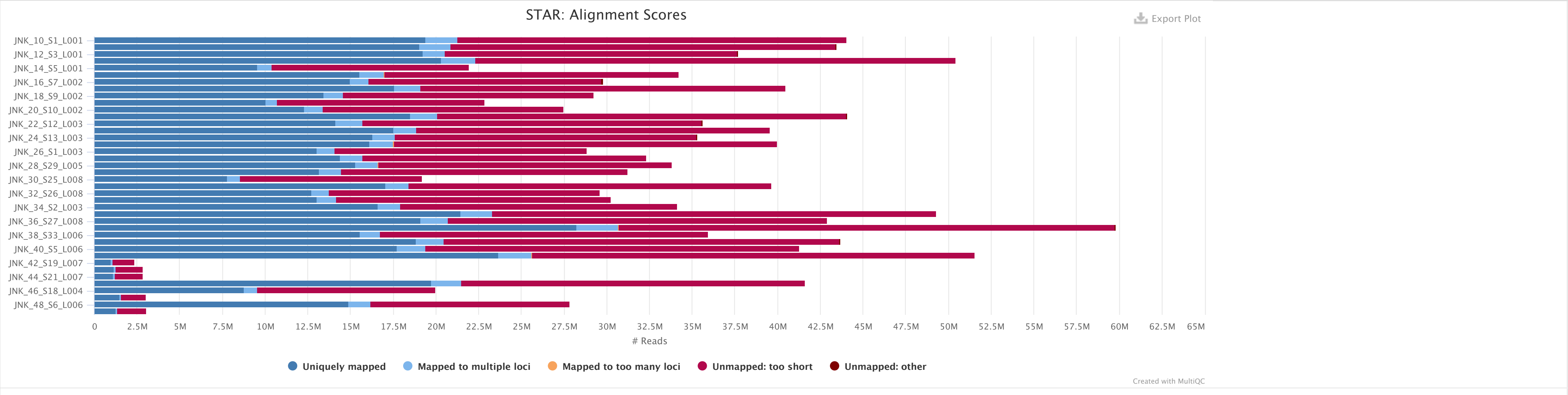


STAR

STAR (https://github.com/alexdobin/STAR) is an ultrafast universal RNA-seq aligner.

Alignment Scores

Number of Reads | Percentages



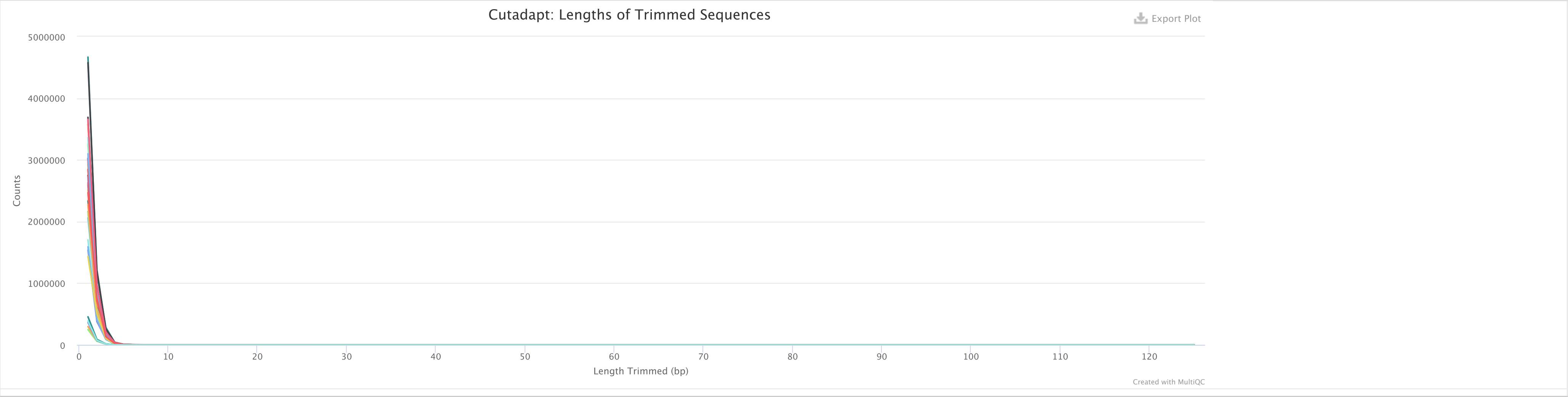
Cutadapt

Cutadapt (https://cutadapt.readthedocs.io/) is a tool to find and remove adapter sequences, primers, poly-Atails and other types of unwanted sequence from your high-throughput sequencing reads.

This plot shows the number of reads with certain lengths of adapter trimmed. Obs/Exp shows the raw counts divided by the number expected due to sequencing errors. A defined peak may be related to adapter length. See the cutadapt documentation (http://cutadapt.readthedocs.org/en/latest/guide.html#how-to-read-the-report) for more information on how these numbers are generated.

Y-Limits: on

Counts | Obs/Exp



FastQC

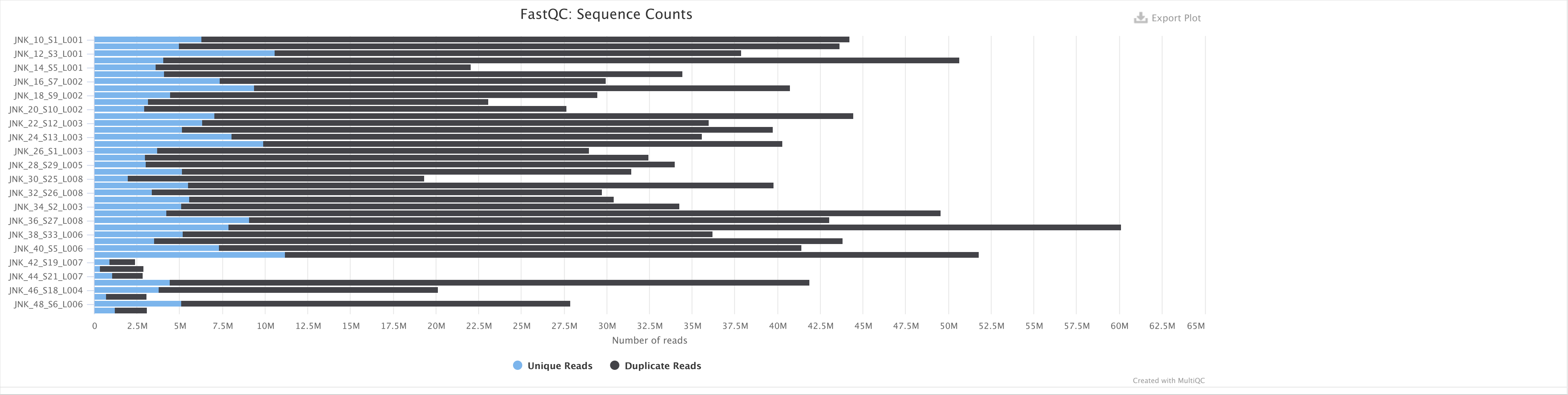
FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) is a quality control tool for high throughput sequence data, written by Simon Andrews at the Babraham Institute in Cambridge.

Sequence Counts

❓ Help

Sequence counts for each sample. Duplicate read counts are an estimate only.

Number of reads | Percentages

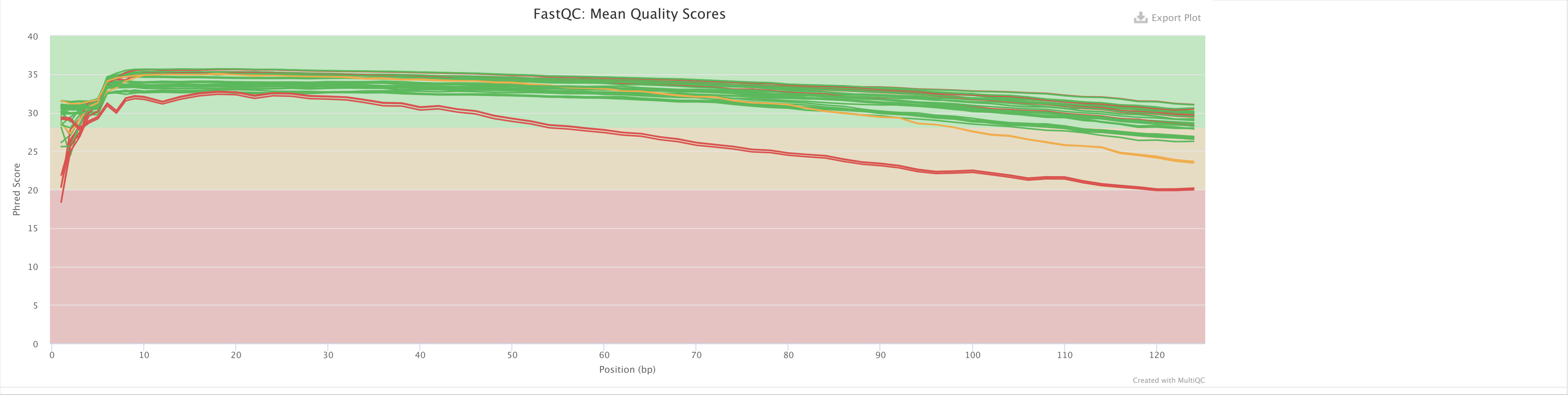


Sequence Quality Histograms 29 2 9

❓ Help

The mean quality value across each base position in the read.

Y-Limits: on

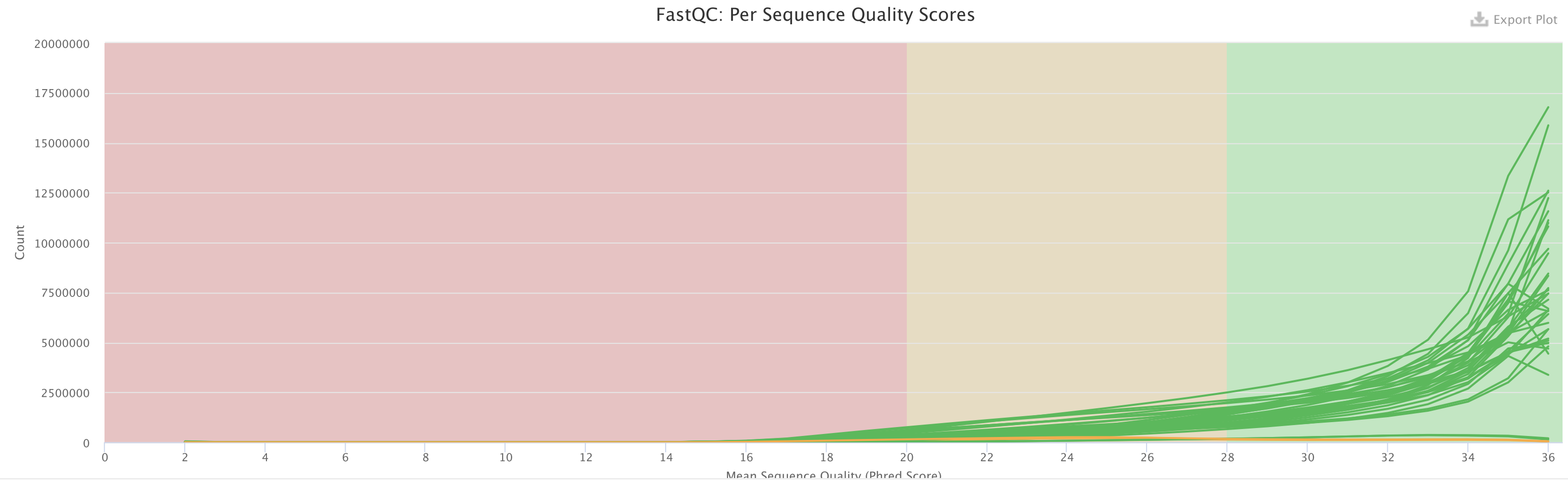


Per Sequence Quality Scores 37 3

❓ Help

The number of reads with average quality scores. Shows if a subset of reads has poor quality.

Y-Limits: on



Per Base Sequence Content 0 28 12

Help

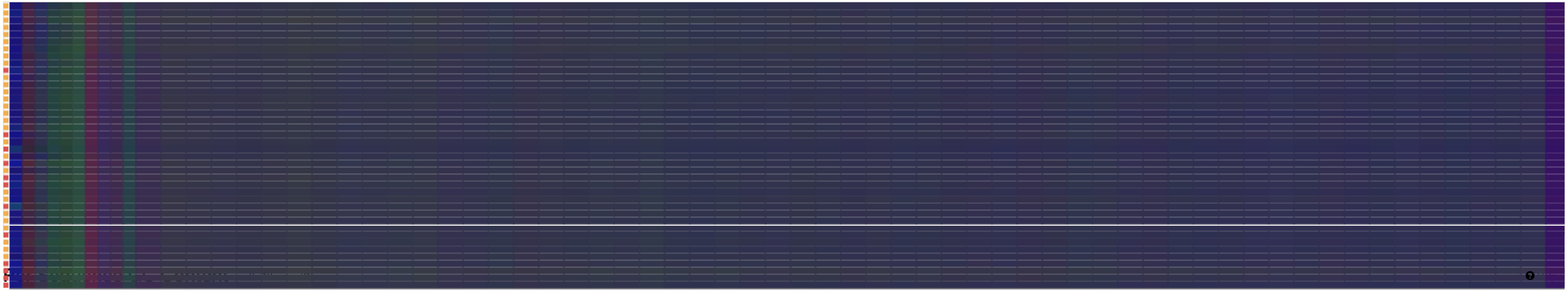
The proportion of each base position for which each of the four normal DNA bases has been called.

Click a sample row to see a line plot for that dataset.

Rollover for sample name

Position: - %T: - %C: - %A: - %G: -

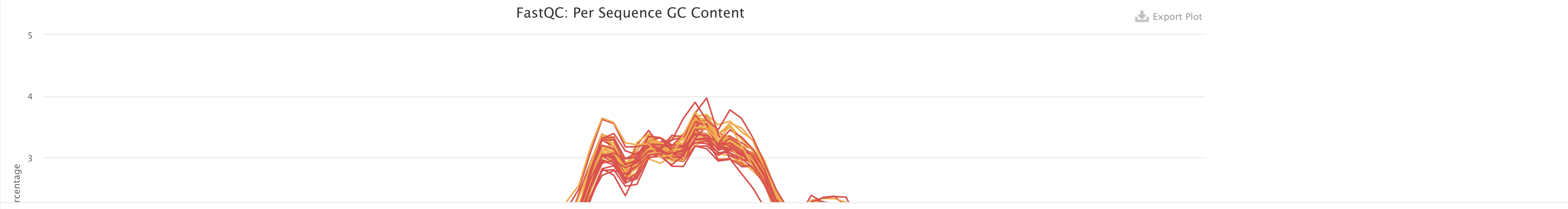
Export Plot



The average GC content of reads. Normal random library typically have a roughly normal distribution of GC content.

Y-Limits: on

Percentages Counts



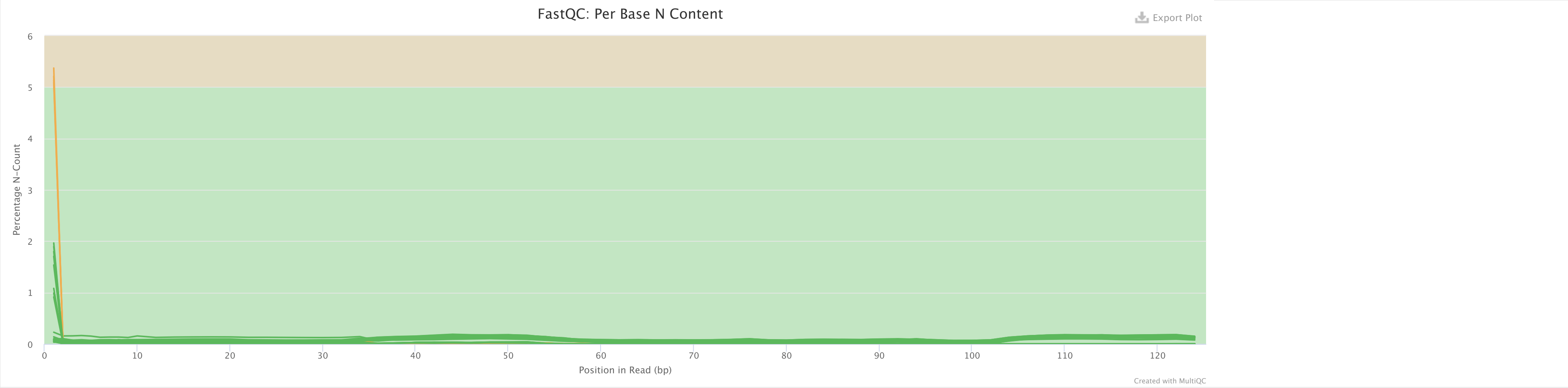
Per Base N Content

36 4

The percentage of base calls at each position for which an `N` was called.

Help

Y-Limits: on

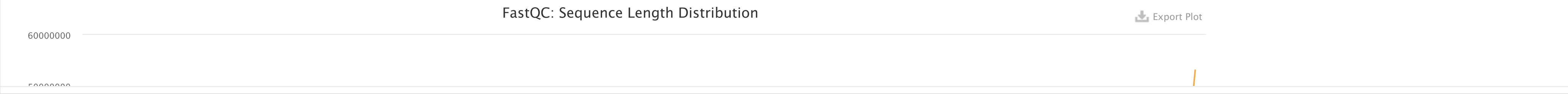


Sequence Length Distribution

0 40

The distribution of fragment sizes (read lengths) found. See the FastQC help (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/7%20Sequence%20Length%20Distribution.html)

Y-Limits: on



Sequence Duplication Levels 0 40

Help

The relative level of duplication found for every sequence.

From the FastQC Help (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/8%20Duplicate%20Sequences.html):

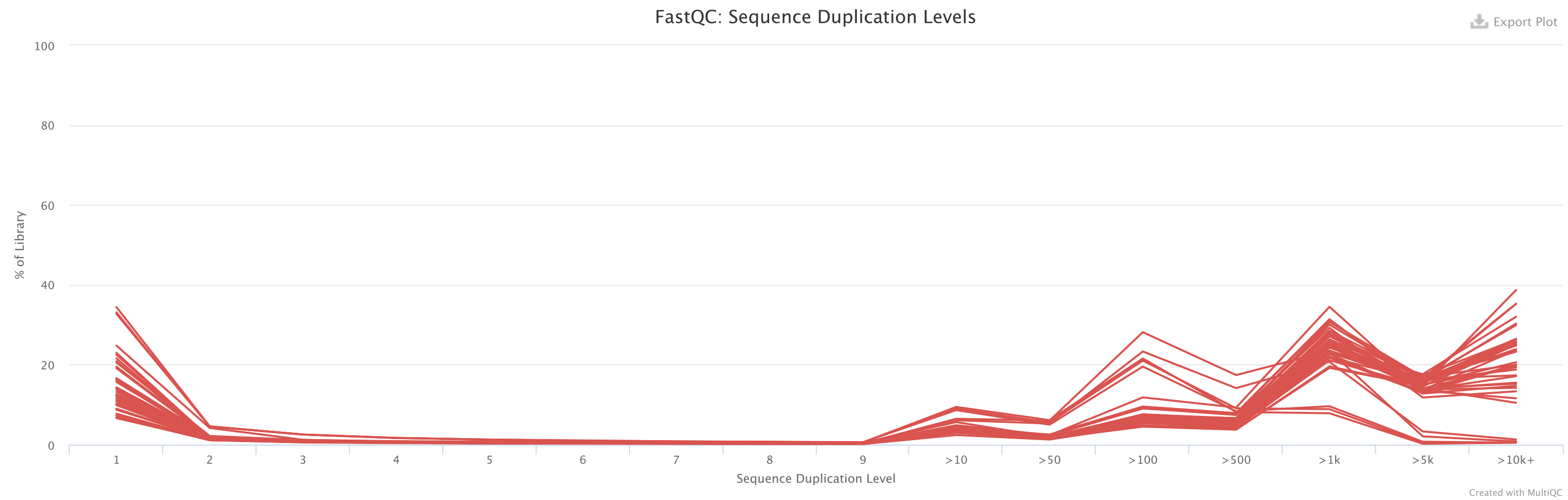
In a diverse library most sequences will occur only once in the final set. A low level of duplication may indicate a very high level of coverage of the target sequence, but a high level of duplication is more likely to indicate some kind of enrichment bias (eg PCR over amplification). This graph shows the degree of duplication for every sequence in a library: the relative number of sequences with different degrees of duplication.

Only sequences which first appear in the first 100,000 sequences in each file are analysed. This should be enough to get a good impression for the duplication levels in the whole file. Each sequence is tracked to the end of the file to give a representative count of the overall duplication level.

The duplication detection requires an exact sequence match over the whole length of the sequence. Any reads over 75bp in length are truncated to 50bp for this analysis.

In a properly diverse library most sequences should fall into the far left of the plot in both the red and blue lines. A general level of enrichment, indicating broad oversequencing in the library will tend to flatten the lines, lowering the low end and generally raising other categories. More specific enrichments of subsets, or the presence of low complexity contaminants will tend to produce spikes towards the right of the plot.

Y-Limits: on



Overrepresented sequences 0 40

Help

The total amount of overrepresented sequences found in each library.

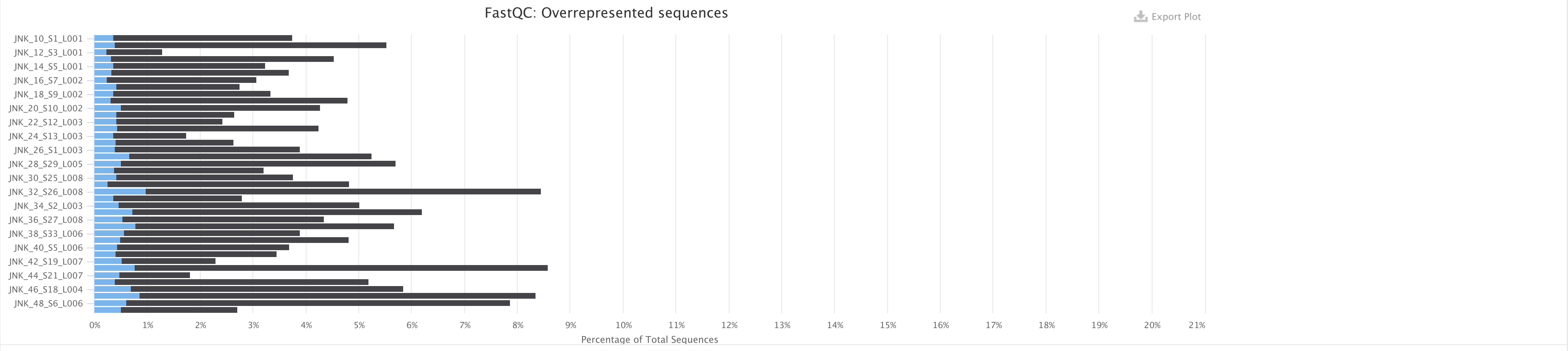
FastQC calculates and lists overrepresented sequences in FastQ files. It would not be possible to show this for all samples in a MultiQC report, so instead this plot shows the *number of sequences* categorized as over represented.

Sometimes, a single sequence may account for a large number of reads in a dataset. To show this, the bars are split into two: the first shows the overrepresented reads that come from the single most common sequence. The second shows the total count from all remaining overrepresented sequences.

From the FastQC Help (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/9%20Overrepresented%20Sequences.html):

A normal high-throughput library will contain a diverse set of sequences, with no individual sequence making up a tiny fraction of the whole. Finding that a single sequence is very overrepresented in the set either means that it is highly biologically significant, or indicates that the library is contaminated, or not as diverse as you expected.

FastQC lists all of the sequences which make up more than 0.1% of the total. To conserve memory only sequences which appear in the first 100,000 sequences are tracked to the end of the file. It is therefore possible that a sequence which is overrepresented but doesn't appear at the start of the file for some reason could be missed by this module.



Adapter Content 40

Help

The cumulative percentage count of the proportion of your library which has seen each of the adapter sequences at each position.

No samples found with any adapter contamination > 0.1%

nf-core/rnaseq Software Versions

nf-core/rnaseq Software Versions (https://github.com/nf-core/rnaseq) are collected at run time from the software output.

nf-core/rnaseq	v1.4.2
Nextflow	v19.10.0
FastQC	v0.11.8
Cutadapt	v2.5
Trim Galore!	v0.6.4
SortMeRNA	v2.1b
STAR	vSTAR_2.6.1d
HISAT2	v2.1.0
Picard MarkDuplicates	v2.21.1
Samtools	v1.9
featureCounts	v1.6.4
Salmon	v0.14.1
StringTie	v2.0
Preseq	v2.0.3
deepTools	v3.3.1
RSeQC	v3.0.1
dupRadar	v1.14.0
edgeR	v3.26.5
Qualimap	v.2.2.2-dev
MultiQC	v1.7

nf-core/rnaseq Workflow Summary

nf-core/rnaseq Workflow Summary (https://github.com/nf-core/rnaseq) - this information is collected when the pipeline is started.

Pipeline Release	master
Run Name	confident_gautier
Reads	/scratch/kmddon001/rnaseq_raw_files/*_R{1,2}_001.fastq
Data Type	Paired-End
Genome	GRCh37
Strandedness	None
Trimming	5'R1: 0 / 5'R2: 0 / 3'R1: 0 / 3'R2: 0 / NextSeq Trim: 0
Aligner	STAR

STAR Index	/scratch/kmddon001/Katie_results//Homo_sapiens/Ensembl/GRCh37/Sequence/STARIndex/
GTF Annotation	/scratch/kmddon001/Katie_results//Homo_sapiens/Ensembl/GRCh37/Annotation/Genes/genes.gtf
BED Annotation	/scratch/kmddon001/Katie_results//Homo_sapiens/Ensembl/GRCh37/Annotation/Genes/genes.bed
Remove Ribosomal R…	N/A
Biotype GTF field	gene_biotype
Save prefs	Ref Genome: No / Trimmed FastQ: No / Alignment intermediates: No
Max Resources	384 GB memory, 40 cpus, 24d 20h 31m 24s time per job
Container	singularity - nfcore/rnaseq:1.4.2
Output dir	/scratch/kmddon001/RNAseq_results_Katie/
Launch dir	/scratch/kmddon001/RNAseq_results_Katie
Working dir	/scratch/kmddon001/RNAseq_results_Katie/work
Script dir	/home/kmddon001/.nextflow/assets/kviljoen/RNAseq
User	kmddon001
Config Profile	uct_hpc

MultiQC v1.7 (http://multiqc.info) - Written by Phil Ewels (http://phil.ewels.co.uk), available on GitHub (https://github.com/ewels/MultiQC).
This report uses HighCharts (http://www.highcharts.com/), jQuery (https://jquery.com/), jQuery UI (https://jqueryui.com/), Bootstrap (http://getbootstrap.com/), FileSaver.js (https://github.com/eligrey/FileSaver.js) and clipboard.js (https://clipboardjs.com/).

SciLifeLab (http://www.scilifelab.se/)