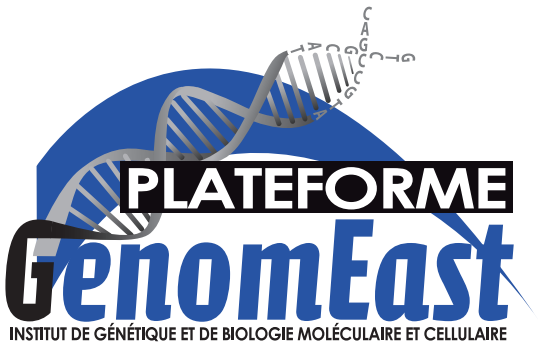
Primary analyses of sequencing data

Stéphanie Le Gras

GenomEast Platform, Illkirch, France



Program

1. Introduction
    1. Data analyzed during this course
2. Analysis of RNA-seq data
    1. Introduction
    2. Experimental design
    3. Library preparation
    4. Sequencing
    5. Data analysis

Neuronal identity genes regulated by super-enhancers are preferentially down-regulated in the striatum of Huntington's disease mice

Mayada Achour, Stéphanie Le Gras, Céline Keime, Frédéric Parmentier, François-Xavier Lejeune, Anne-Laurence Boutillier, Christian Néri, Irwin Davidson, Karine Merienne ✉ Author Notes

Human Molecular Genetics, Volume 24, Issue 12, 15 June 2015, Pages 3481–3496,
https://doi.org/10.1093/hmg/ddv099
Published: 17 March 2015 **Article history** ▾

Achour et al, 2015

Achour et al, 2015

- Striatum of HD (R6/1) and Control (WT) mice:
  - RNA-seq data:
    - At a late pathological stage (30 weeks)
    - 2 replicates for HD mice
    - 3 replicates for Control mice

All examples (RNA-seq) are based on data analyzed in this publication

Analysis of RNA-seq data

Introduction

Transcriptome

- The transcriptome is the collection of all transcripts produced during the biological process of transcription.
- Unlike genome, transcriptome is dynamic and varies depending on cell types, time and environmental conditions.

Types of RNAs

- 2 main classes of RNA molecules:
  - RNAs translated into proteins *i.e.* messenger RNAs (mRNAs)
  - non-coding RNAs (ncRNAs)
    - Long non-coding RNAs (lncRNAs)
    - Small non-coding RNAs

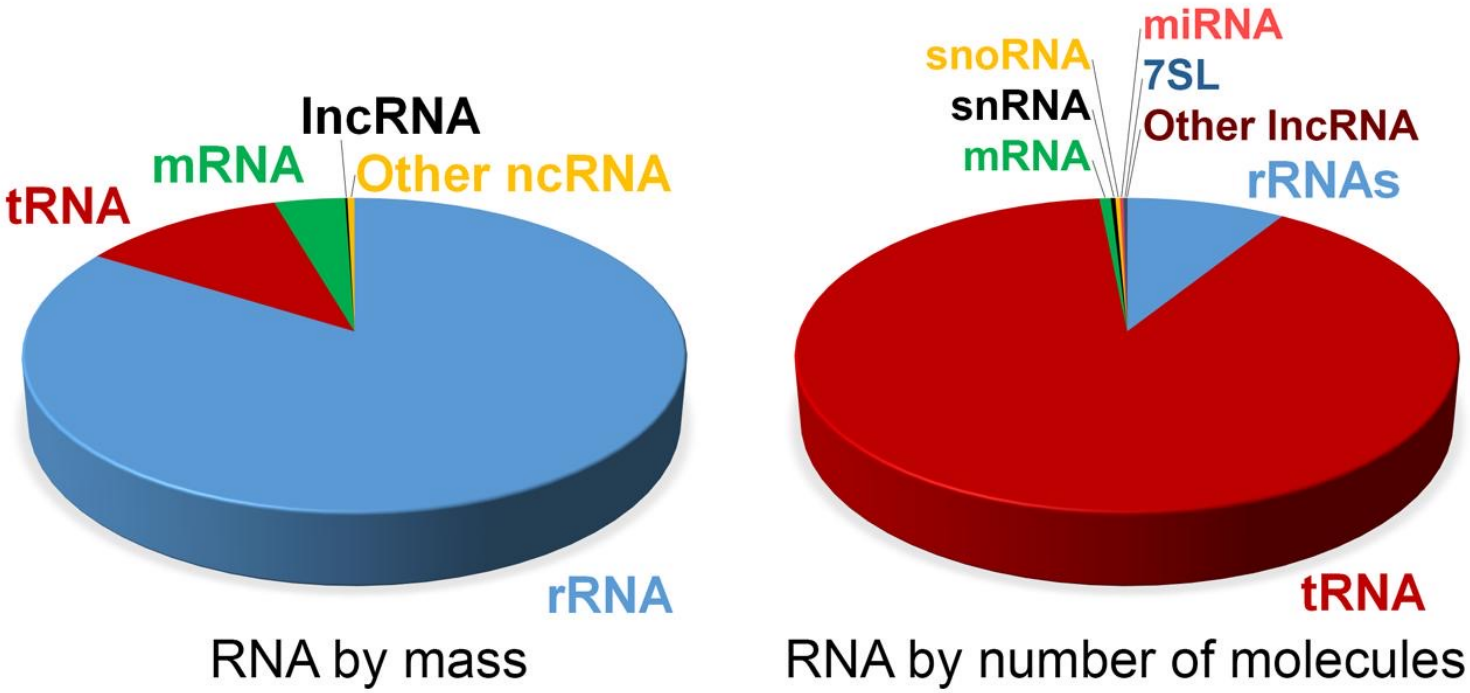
Table 1 - Different types, main characteristics and functions of non-coding RNAs.

		Mean size	Function
Long ncRNA	Ribosomal RNA	~1.9 kb	Essential for protein synthesis
	XIST RNA	~17 kb	Chromosome X inactivation
	Other lncRNA	> 200 nt	Involved in epigenetic modification, post-transcriptional processing, modulation of chromatin structure, etc.
Small ncRNA	miRNAs	18-21 nt	Gene regulation
	siRNA	~21 nt	Gene regulation; defense against viruses and transposon activity
	rasiRNA	24-27 nt	Orientation of heterochromatin in the formation of centromeres
	snoRNA	60-300 nt	Methylation and pseudo uridylation of other RNAs
	snRNA	100-300 nt	Involved in spliceosome complex
	piRNA	26-30 nt	Regulation of transposon activity and chromatin state

Dogini et al, 2014

Types of RNAs

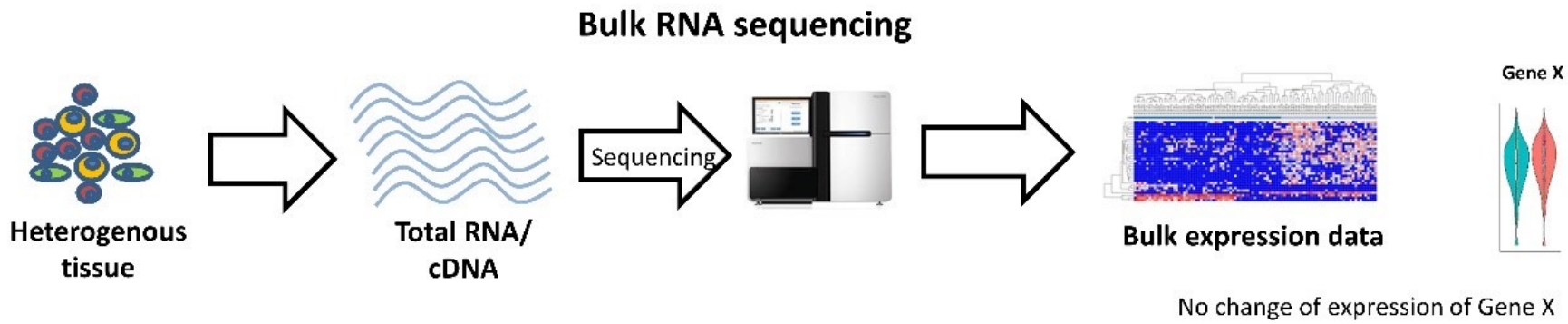
- Estimate of RNA levels in a typical mammalian cell.



Palazzo et al, 2015

Assays and bioinformatics protocols change upon RNAs of interest

Bulk RNA-seq

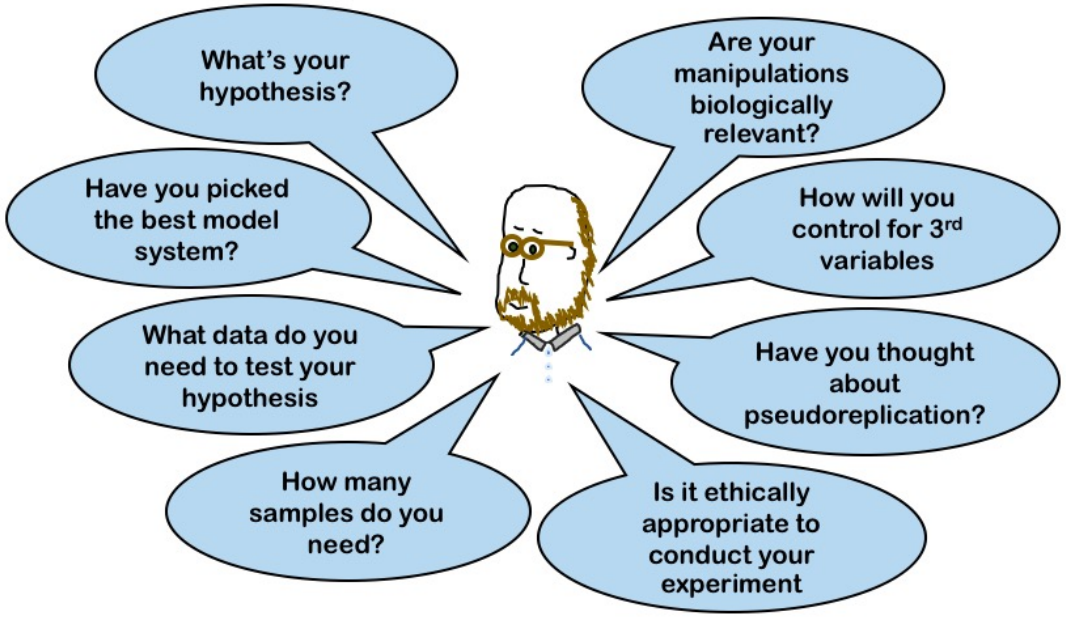


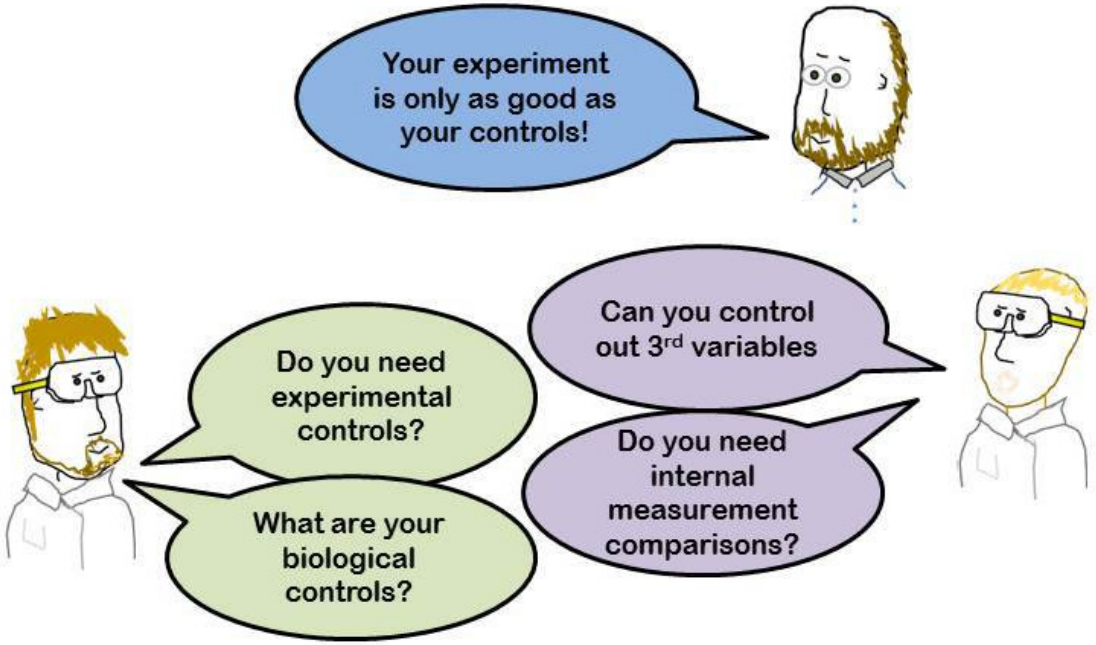
https://www.technologynetworks.com/genomics/articles/recent-advances-in-single-cell-genomics-techniques-324695

Analysis of RNA-seq data

Experimental design

Experimental design

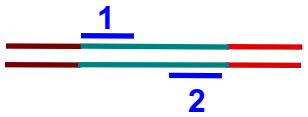




Which sequencing strategy ?

- It depends on the transcriptome analysis of interest:
  - Expression quantification on annotated transcripts
    - **Single-end** sequencing provides good results

  - Alternative splicing analysis, fusion transcript detection, mapping over repetitive regions, de novo transcriptome assembly
    - **Paired-end** sequencing is needed



How many reads are needed ?

- Transcriptome coverage as a function of sequencing depth:
  highly dependant on transcriptome complexity
- Sequencing depth should be determined by the goals of the experiment
- General recommendations for typical mammalian tissues
  - > 30 million reads with polyA+ protocols
  - > 50 million reads with total protocols
  - … if the goal is to quantify expression of annotated genes
- Higher sequencing depth needed if
  - the sensitivity of detection is important
  - the purpose is to discover novel transcripts
  - the purpose is to precisely quantify transcript isoforms

Examples
on our NextSeq 2000

Application	Number of million reads per sample for standard experiments on mammalian genomes	Sequencing length
scRNA-seq	100 (~5,000 cells)	Custom paired-end
small RNA-seq	10	1 x 50 bp
mRNA-seq with polyA selection		
➢ for gene expression quantification	30	1 x 50 bp
➢ for alternative splicing analysis	60	2 x 50 bp
RNA-seq with ribodepletion		
➢ for gene expression quantification	50	1 x 50 bp

How many replicates are needed ?

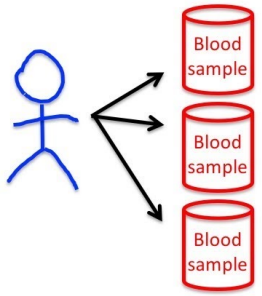
- Make sure that results are reliable and valid
- Low technical variability
  and technical variability << biological variability

  (Marioni et al. Genome Research 2008. Bullard et al. BMC Bioinformatics 2010)

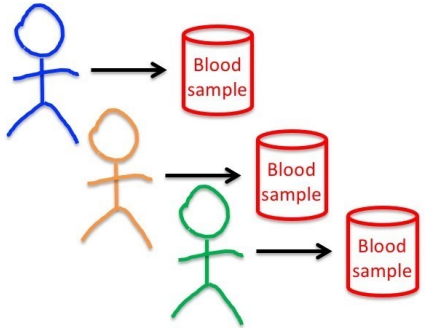
  -> Technical replicates not required
- But “sequencing technology does not eliminate biological variability”

  (Hansen et al. Nat Biotechnol. 2011)
  - **Biological replicates are fundamental !**

Technical replicates



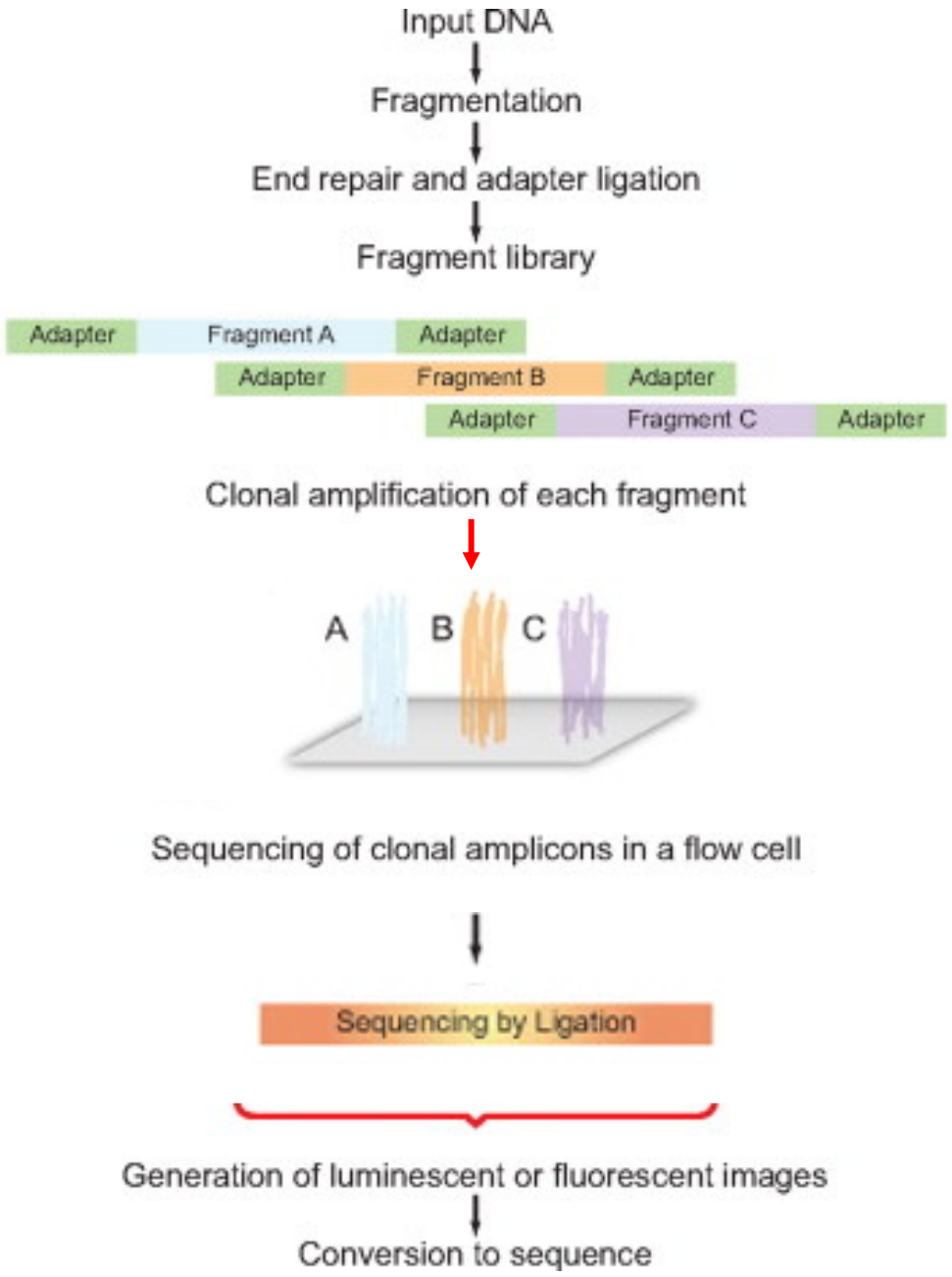
Biological replicates



Analysis of RNA-seq data

Library prep

Illumina's sequencing process



1. Library prep

2. Amplification (cluster growth)

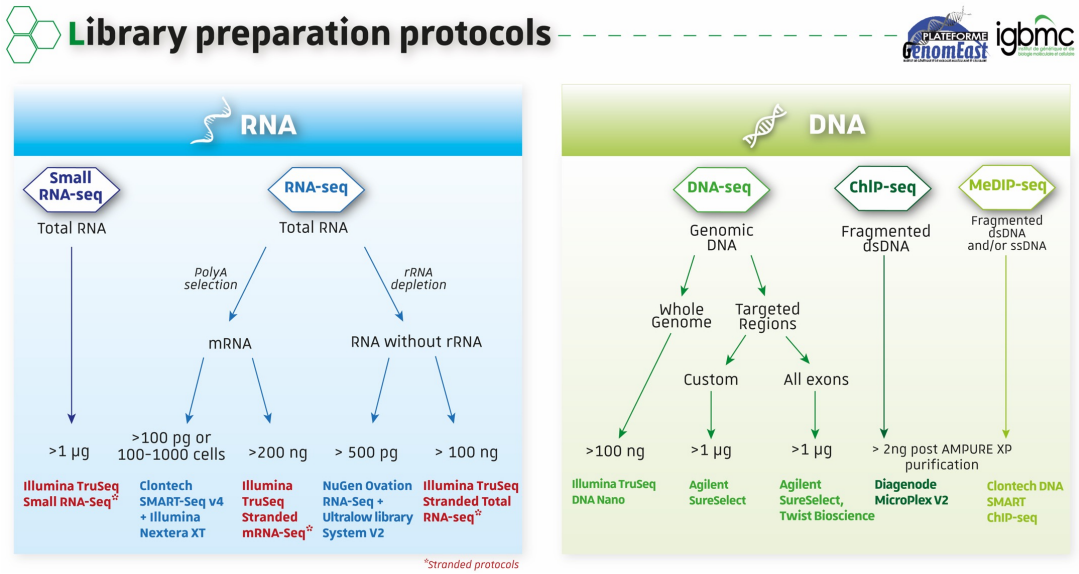
3. Sequencing

4. Generation of images / sequence calling

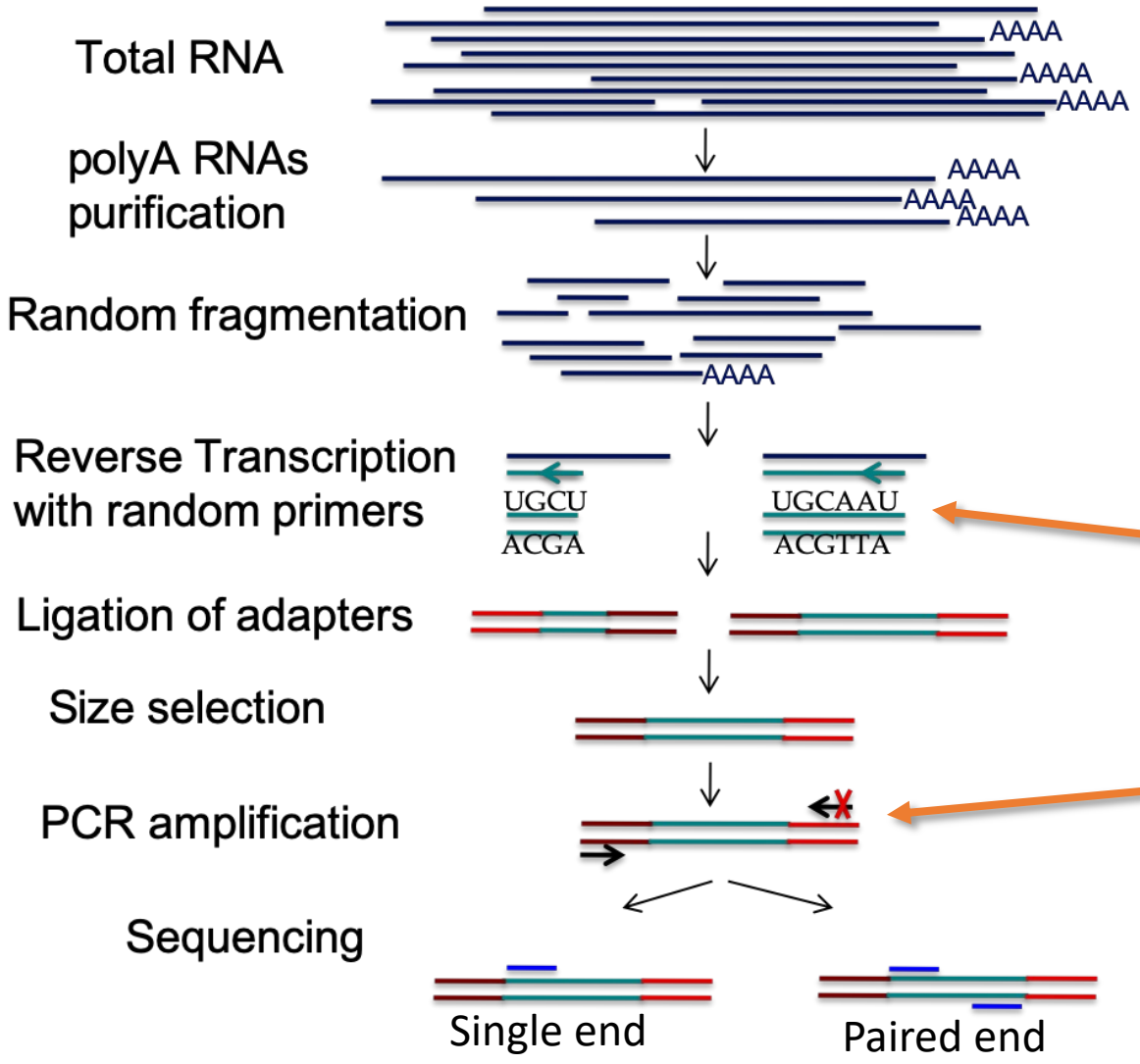
Modified from Voelkerding KV, at al. 2010

RNA-seq library preparation protocols

- A lot of RNA-seq protocols exists depending of the RNA type of interest and the goals of the experiment



RNA-seq library preparation (stranded polyA+)



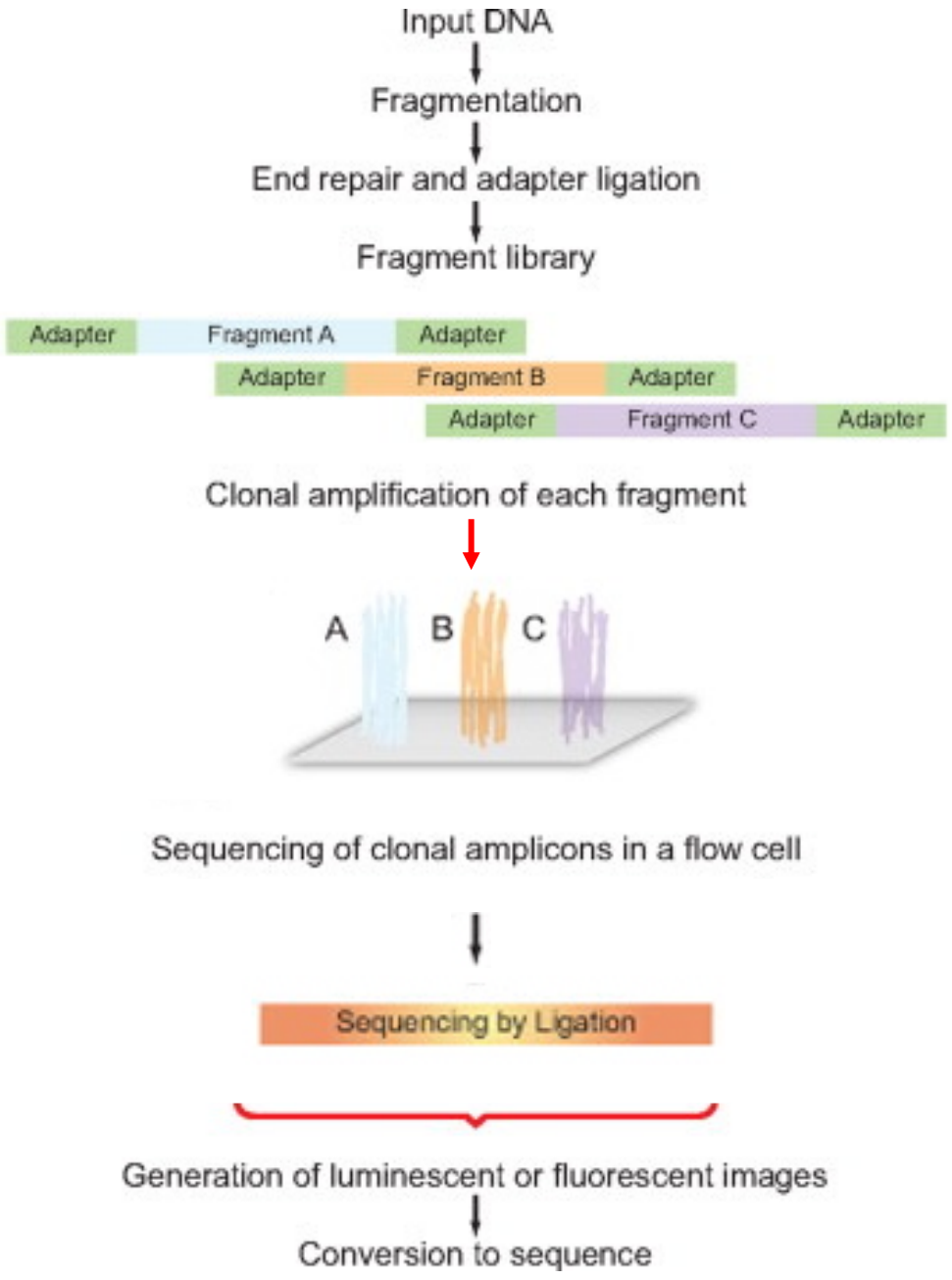
Incorporation of dUTP instead of dTTP in the second strand cDNA synthesis

Amplification of reverse strand only (high fidelity Taq polymerase)

Analysis of RNA-seq data

Sequencing

Illumina's sequencing process



1. Library prep

2. Amplification (cluster growth)

3. Sequencing

4. Generation of images / sequence calling

Modified from Voelkerding KV, at al. 2010

Analysis of RNA-seq data

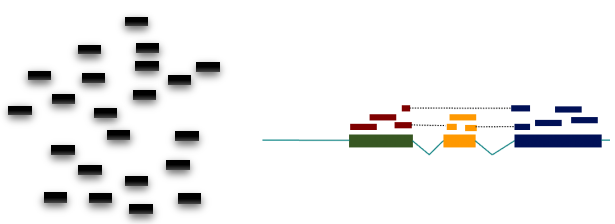
Data analysis

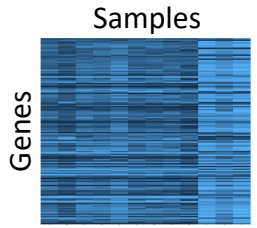
Analysis of RNA-seq data

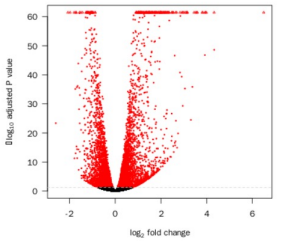
Example of a workflow for analyzing RNA-seq data:

- The goal of the analysis described here is to quantify gene expression and measure differential expression profiles between 2 or more conditions
- A sequenced reference genome exists

From nucleotide sequence to count matrix

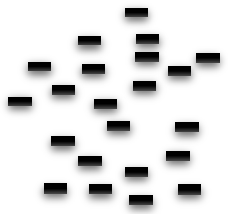


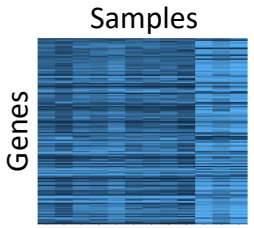


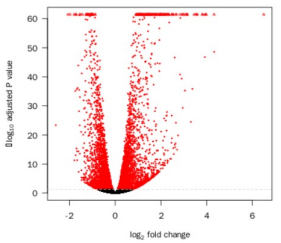


Analysis of RNA-seq data

From nucleotide sequence to count matrix







FASTQC

- Allows quality control of NGS data
  - FASTQ, gzip compressed FASTQ (base or colorspace)
  - SAM, BAM alignment files
- Can be used *via* a graphical interface, in command line or in Galaxy
- Generates graphs and tables with several quality control analyses
  - Allows a global quality assessment of NGS data and rapid identification of possible problems

Basic Statistics

Measure	Value
Filename	R6_1_387_St_chr19_fastq_gz.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	2215292
Sequences flagged as poor quality	0
Sequence length	50
%GC	46

The Basic Statistics module generates some simple composition statistics for the file analysed.

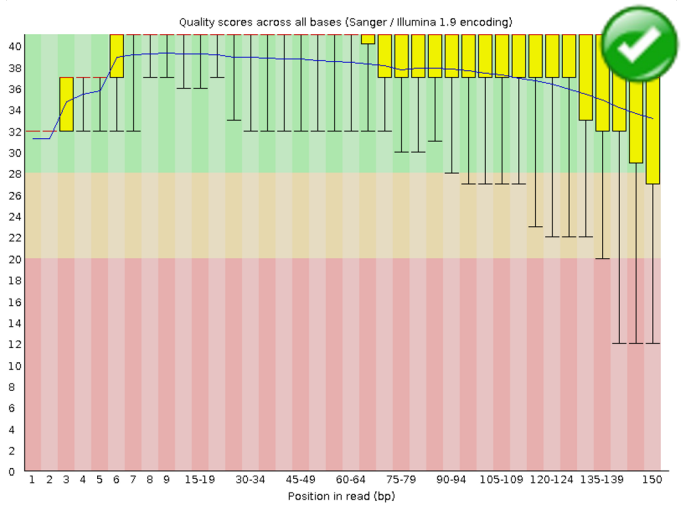
- Filename
- File type
- Encoding: Says which ASCII encoding of quality values was found in this file.
- Total Sequences: A count of the total number of sequences processed.
- Filtered Sequences:
- Sequence Length
- %GC

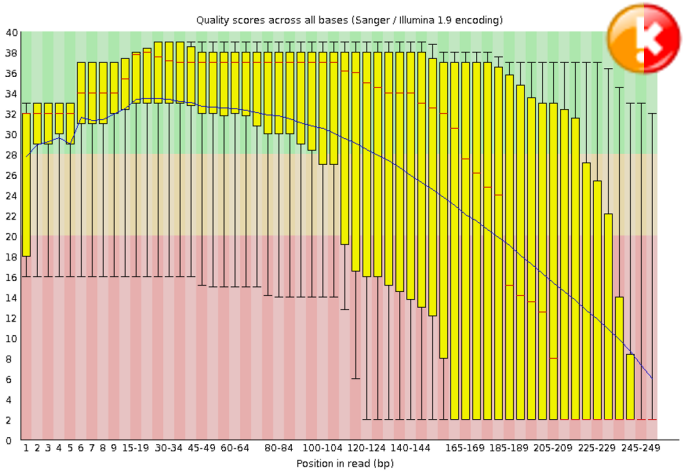
Basic Statistics never raises a warning.

Basic Statistics never raises an error.

Source: EBAII2022

Per Base Sequence Quality





This view shows an overview of the range of quality values across all bases at each position in the FastQ file.

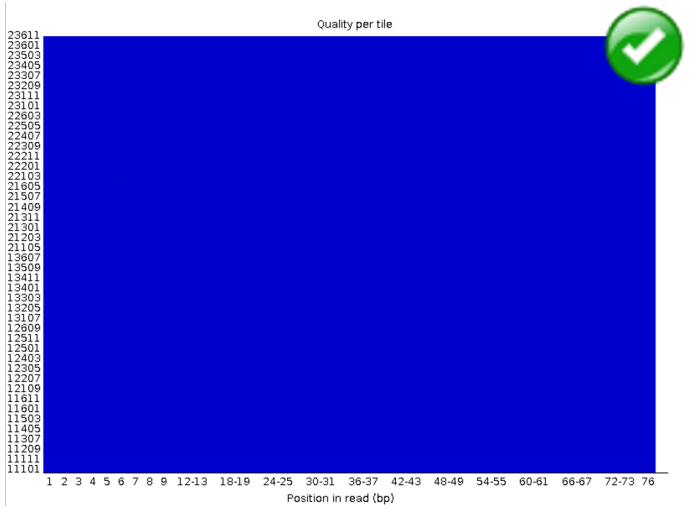
For each position a BoxWhisker type plot is drawn. The elements of the plot are as follows: The central red line is the median value The yellow box represents the inter-quartile range (25-75%) The upper and lower whiskers represent the 10% and 90% points The blue line represents the mean quality The y-axis on the graph shows the quality scores. The higher the score the better the base call. The background of the graph divides the y axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red). The quality of calls on most platforms will degrade as the run progresses, so it is common to see base calls falling into the orange area towards the end of a read.

A warning will be issued if the lower quartile for any base is less than 10, or if the median for any base is less than 25.

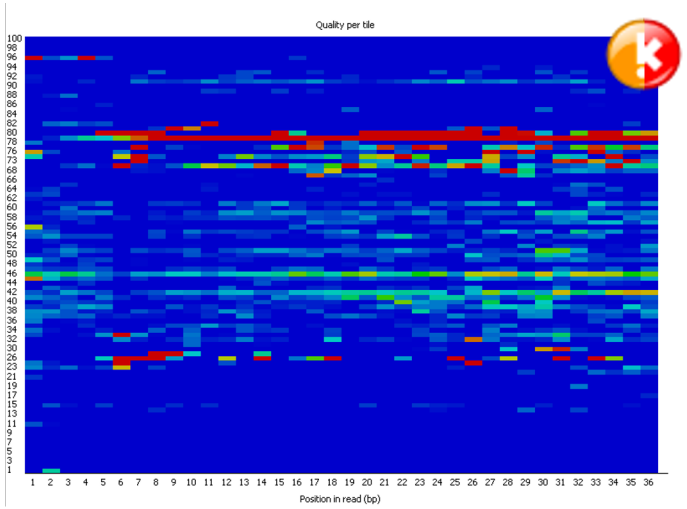
This module will raise a failure if the lower quartile for any base is less than 5 or if the median for any base is less than 20.

Source: EBAII2022

Per Tile Sequence Quality



The plot shows the deviation from the average quality for each tile. The colours are on a cold to hot scale, with cold colours being positions where the quality was at or above the average for that base in the run, and hotter colours indicate that a tile had worse qualities than other tiles for that base. In the example below you can see that certain tiles show consistently poor quality. A good plot should be blue all over.

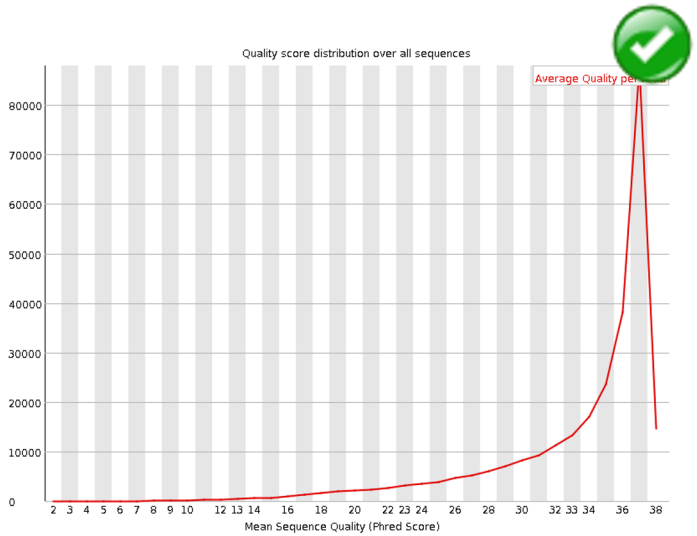


This module will issue a warning if any tile shows a mean Phred score more than 2 less than the mean for that base across all tiles.

This module will issue a warning if any tile shows a mean Phred score more than 5 less than the mean for that base across all tiles.

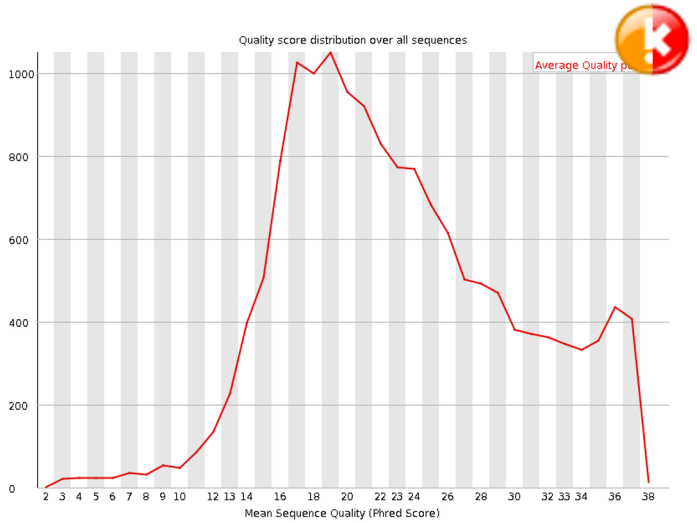
Source: EBAII2022

Per Sequence Quality Scores



The per sequence quality score report allows you to see if a subset of your sequences have universally low quality values. It is often the case that a subset of sequences will have universally poor quality, often because they are poorly imaged (on the edge of the field of view etc), however these should represent only a small percentage of the total sequences.

If a significant proportion of the sequences in a run have overall low quality then this could indicate some kind of systematic problem - possibly with just part of the run (for example one end of a flowcell).

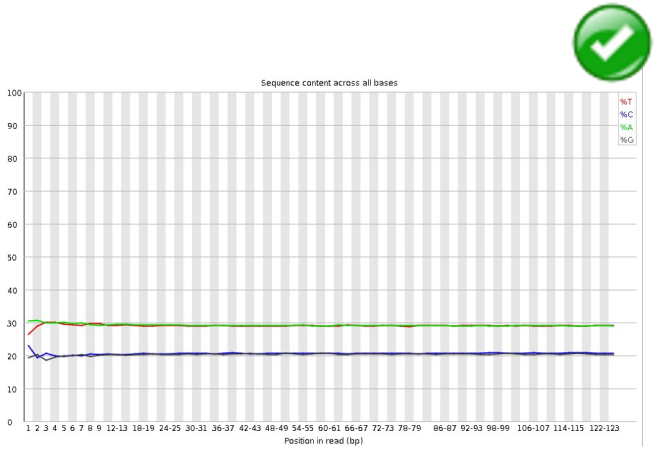


A warning is raised if the most frequently observed mean quality is below 27 - this equates to a 0.2% error rate.

An error is raised if the most frequently observed mean quality is below 20 - this equates to a 1% error rate.

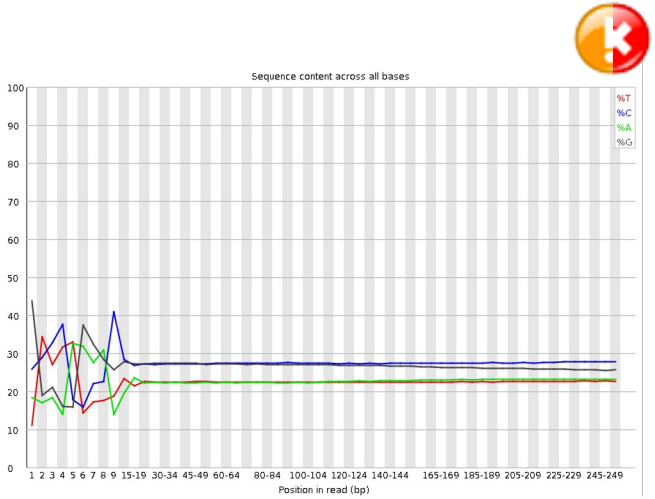
Source: EBAII2022

Per Base Sequence Content



Per Base Sequence Content plots out the proportion of each base position in a file for which each of the four normal DNA bases has been called.

In a random library you would expect that there would be little to no difference between the different bases of a sequence run, so the lines in this plot should run parallel with each other.



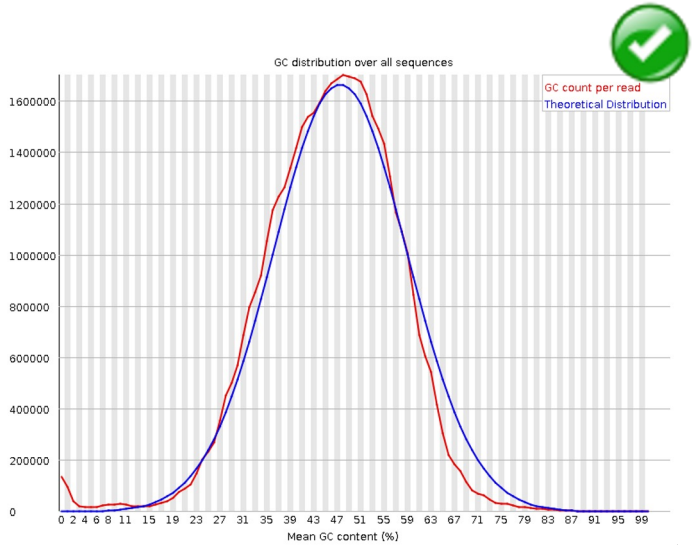
If you see strong biases which change in different bases then this usually indicates an overrepresented sequence which is **contaminating** your library. A bias which is consistent across all bases either indicates that the **original library was sequence biased**, or that there was **a systematic problem during the sequencing of the library**.

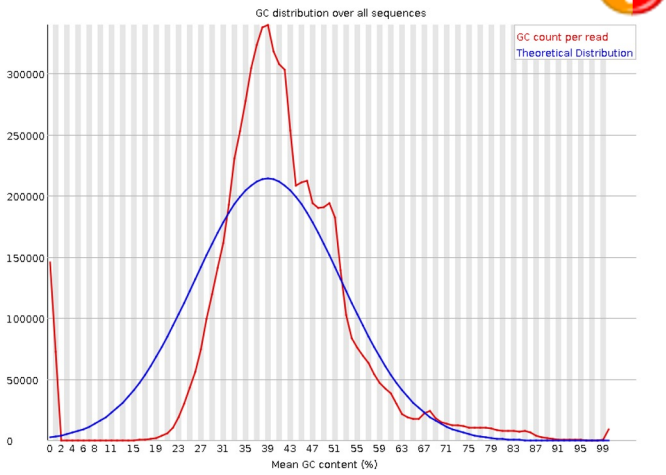
This module issues a warning if the difference between A and T, or G and C is greater than 10% in any position.

This module will fail if the difference between A and T, or G and C is greater than 20% in any position.

Source: EBAII2022

Per Sequence GC Content





Source: EBAII2022

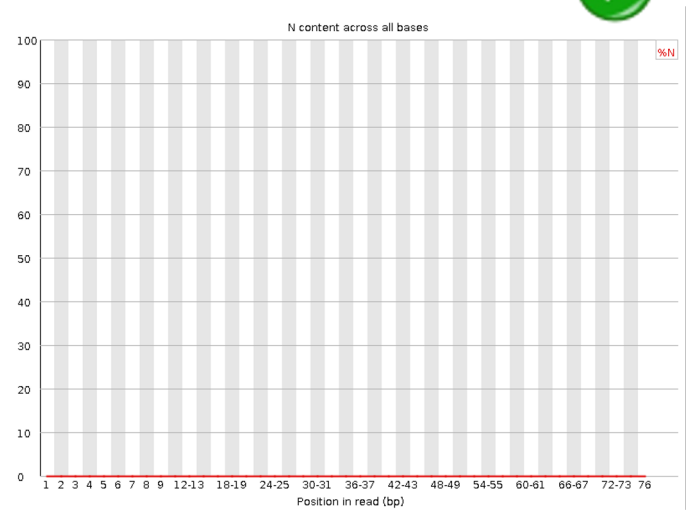
This module measures the GC content across the whole length of each sequence in a file and compares it to a modelled normal distribution of GC content.

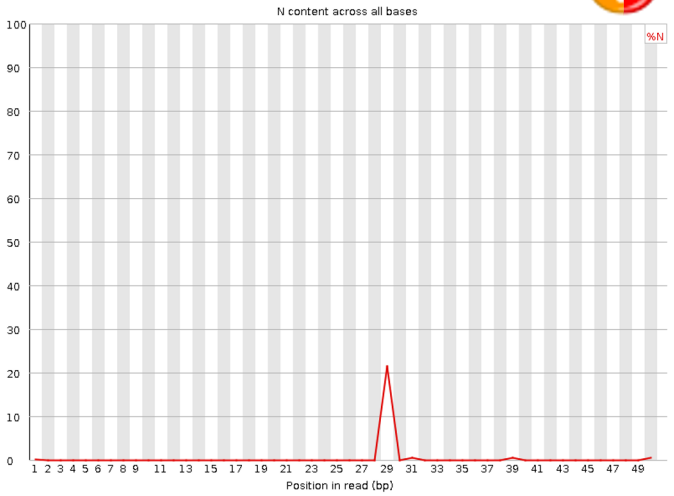
In a normal random library you would expect to see a roughly normal distribution of GC content where the central peak corresponds to the overall GC content of the underlying genome. Since we don't know the the GC content of the genome the modal GC content is calculated from the observed data and used to build a reference distribution. An unusually shaped distribution could indicate a **contaminated library** or some other kinds of biased subset. A normal distribution which is shifted indicates some systematic bias which is independent of base position. If there is a systematic bias which creates a shifted normal distribution then this won't be flagged as an

A warning is raised if the sum of the deviations from the normal distribution represents more than 15% of the reads.

This module will indicate a failure if the sum of the deviations from the normal distribution represents more than 30% of the reads.

Per Base N Content





If a sequencer is unable to make a base call with sufficient confidence then it will normally substitute an N rather than a conventional base call This module plots out the percentage of base calls at each position for which an N was called.

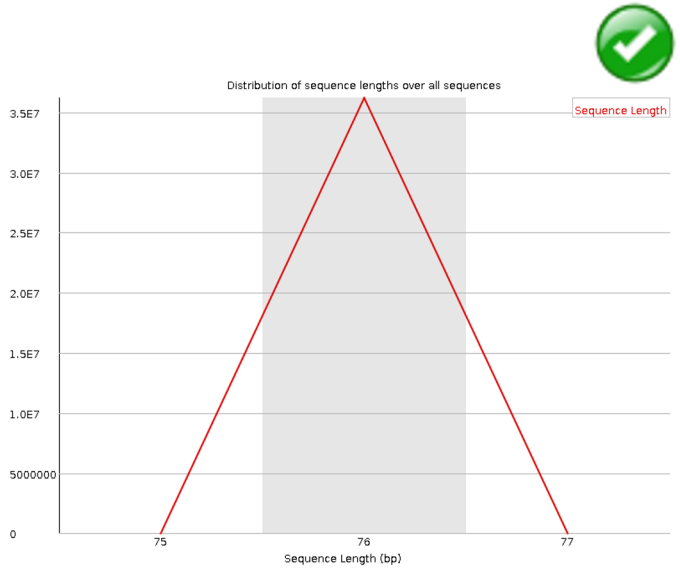
It's not unusual to see a very low proportion of Ns appearing in a sequence, especially nearer the end of a sequence. However, if this proportion rises above a few percent it suggests that the analysis pipeline was unable to interpret the data well enough to make valid base calls.

This module raises a warning if any position shows an N content of >5%.

This module will raise an error if any position shows an N content of >20%.

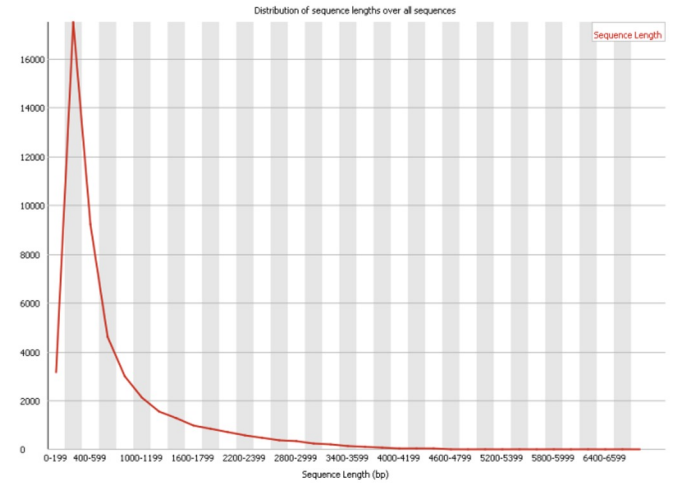
Source: EBAII2022

Sequence Length Distribution



Some high throughput sequencers generate sequence fragments of uniform length, but others can contain reads of wildly varying lengths. Even within uniform length libraries some pipelines will trim sequences to remove poor quality base calls from the end. This module generates a graph showing the distribution of fragment sizes in the file which was analysed.

In many cases this will produce a simple graph showing a peak only at one size, but for variable length FastQ files this will show the relative amounts of each different size of sequence fragment.

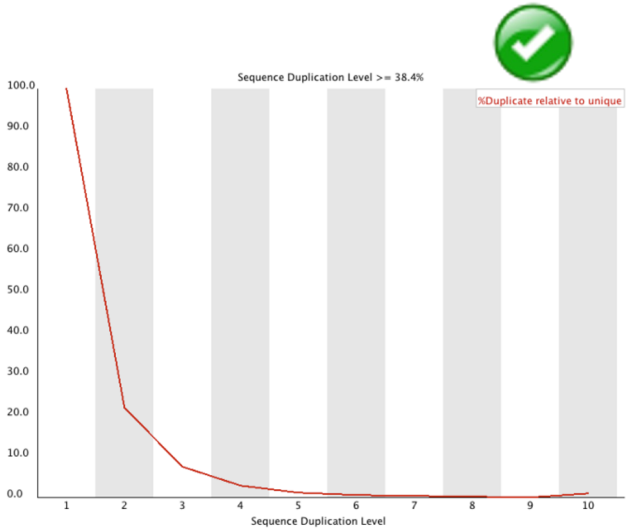


This module will raise a warning if all sequences are not the same length.

This module will raise an error if any of the sequences have zero length.

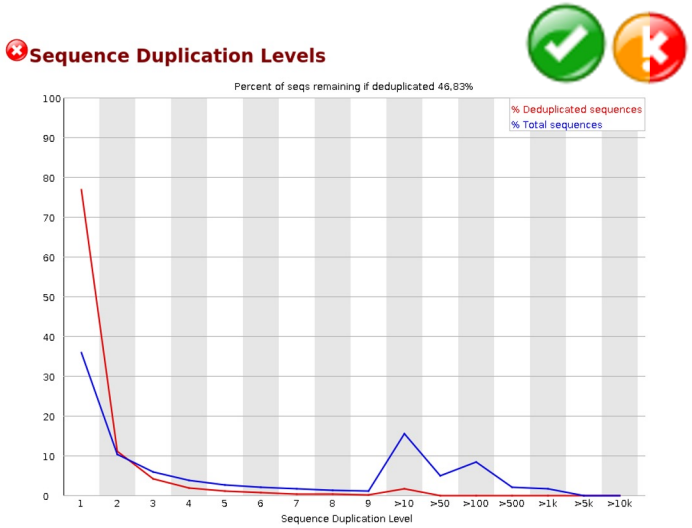
Source: EBAII2022

Duplicate Sequences



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 module counts the degree of duplication for every sequence in the set and creates a plot showing the relative number of sequences with different degrees of duplication.



This module will issue a warning if non-unique sequences make up more than 20% of the total.

This module will issue a error if non-unique sequences make up more than 50% of the total.

Source: EBAII2022

Overrepresented Sequences

Sequence	Count	Percentage	Possible Source
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	125085	0.34543496720376876	No Hit

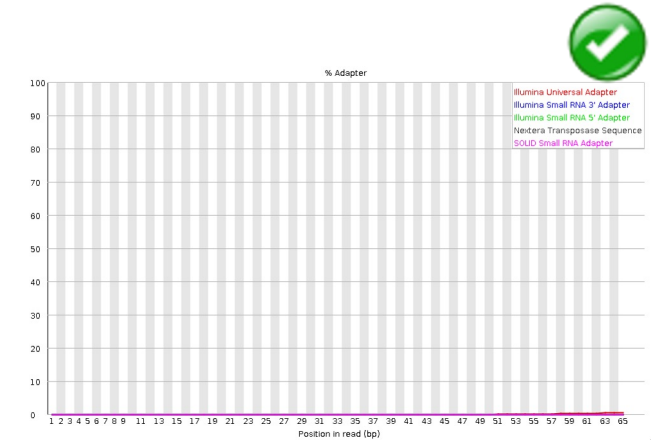
A normal high-throughput library will contain a diverse set of sequences, with no individual sequence making up a tiny fraction of the whole. Finding that a single sequence is very overrepresented in the set either means that it is highly biologically significant, or indicates that the library is contaminated, or not as diverse as you expected. This module lists all of the sequence which make up more than 0.1% of the total. To conserve memory only sequences which appear in the first 200,000 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.

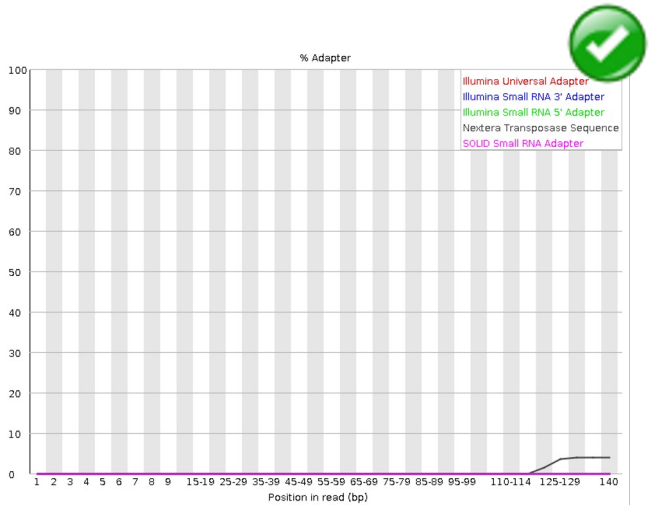
This module will issue a warning if any sequence is found to represent more than 0.1% of the total.

This module will issue an error if any sequence is found to represent more than 1% of the total.

Source: EBAII2022

Adapter Content





The Kmer Content module will do a generic analysis of all of the Kmers in your library to find those which do not have even coverage through the length of your reads. This can find a number of different sources of bias in the library which can include the presence of read-through adapter sequences building up on the end of your sequences.

You can however find that the presence of any overrepresented sequences in your library (such as adapter dimers) will cause the Kmer plot to be dominated by the Kmers these sequences contain, and that it's not always easy to see if there are other biases present in which you might be interested.

This module will issue a warning if any sequence is present in more than 5% of all reads.

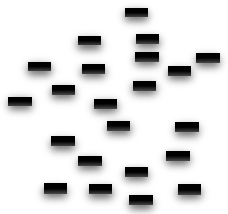
This module will issue a warning if any sequence is present in more than 10% of all reads.

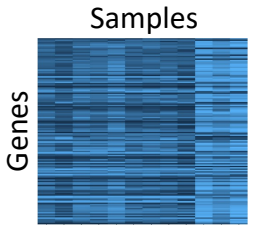
Source: EBAII2022

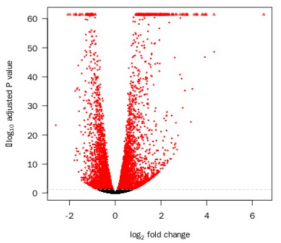
Hands-on

Analysis of RNA-seq data

From nucleotide sequence to count matrix

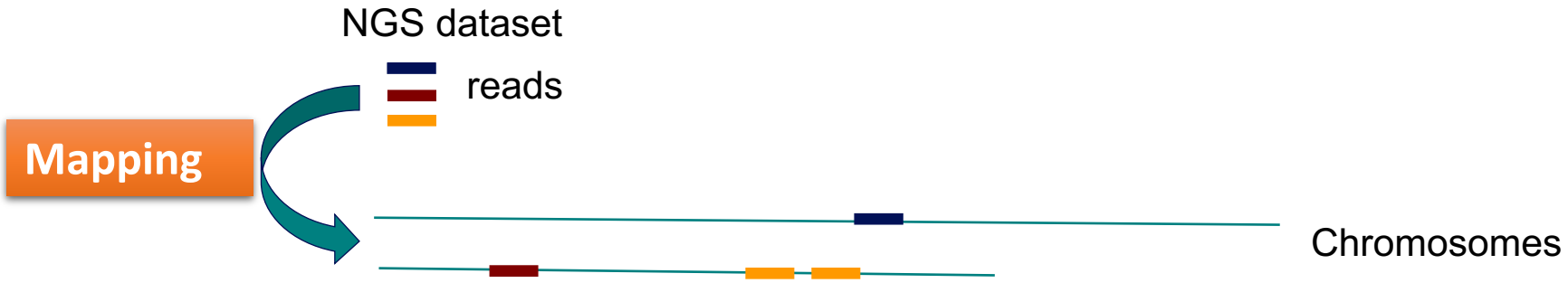






What is mapping ?

- Map reads against a reference genome
  = Predict the locus from which a read originates
  ➔ Find the loci with sufficient similarity



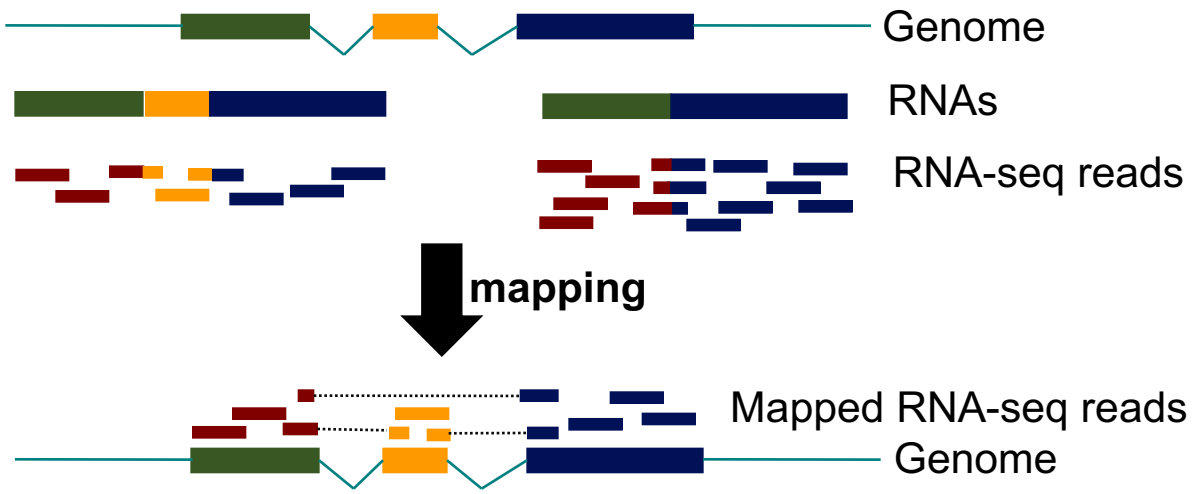
- Sufficient similarity
  ➔ Less mismatches / indels

reference genome	CACGTACC	CACGTA_CC	CACGTACC
reads	CACGTTCC	CACGTATCC	CACGT_CC
Alignment	mismatch	indels (insertion/deletion)	

Spliced mapping

Specificity of RNA-seq data mapping

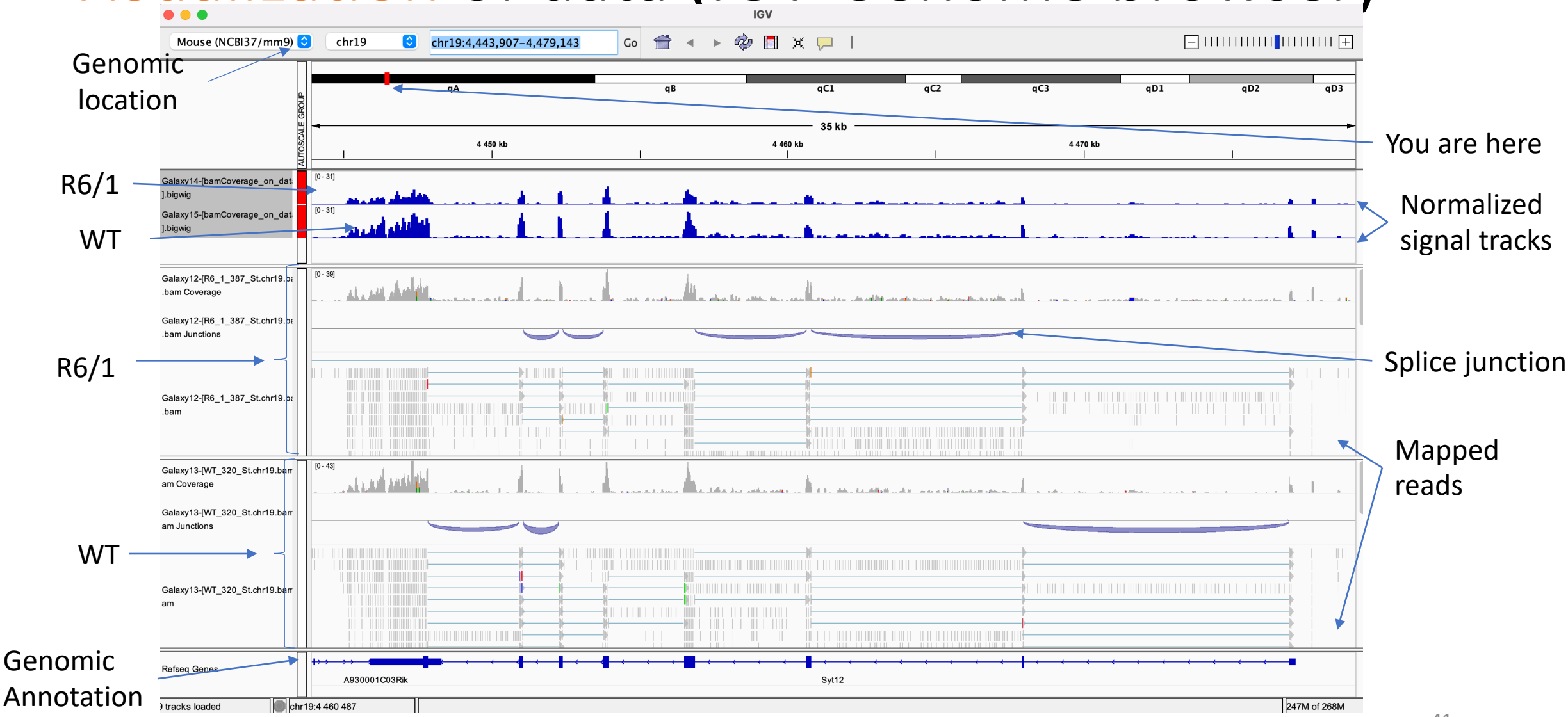
- Allows mapping of reads across splice junctions



- Transcript structures is needed (prediction or from public databases)

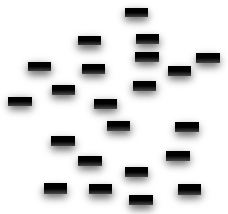
Hands-on

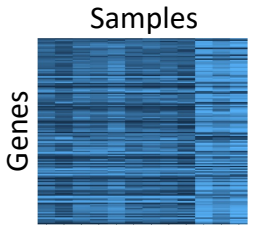
Visualization of data (IGV Genome browser)

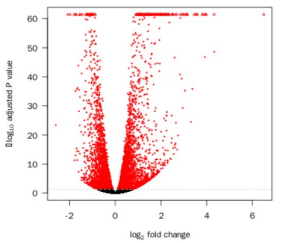


Analysis of RNA-seq data

From nucleotide sequence to count matrix

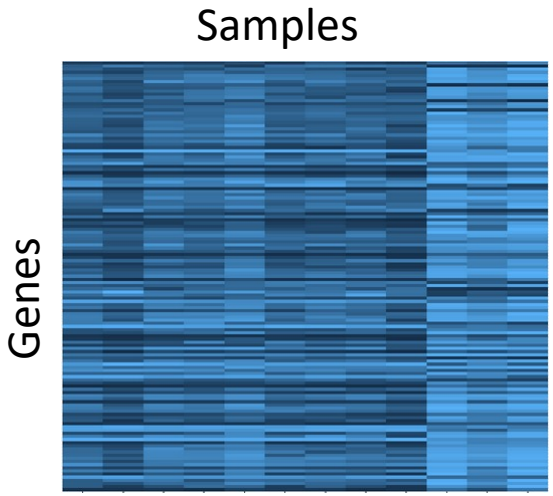




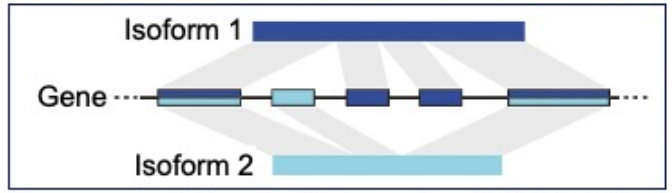


Quantification

- Quantify number of reads on each gene (gene level quantification)



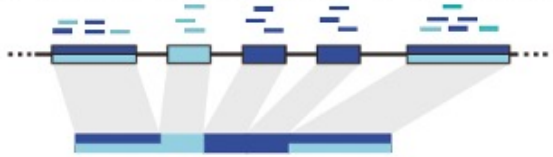
How to summarize expression level of genes with several isoforms ?



Garber et al., Nature methods 2011; 8(6):469-77

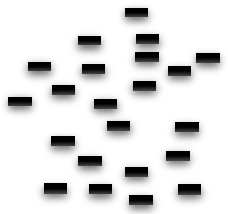
- Exon-union method

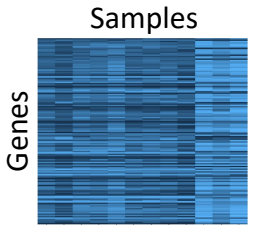
  Count reads mapped to all exons from all isoforms of the gene

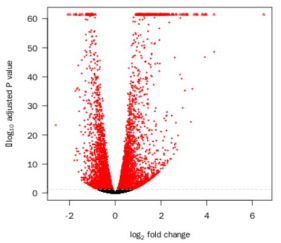


Analysis of RNA-seq data

From Nucleotide sequence to count matrix





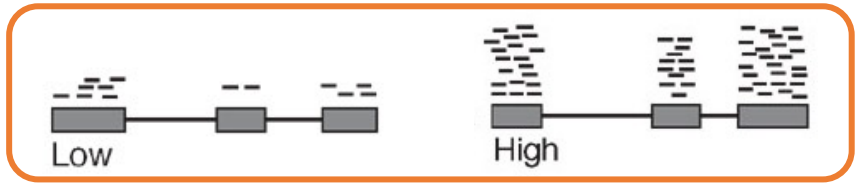


Normalization: why ?

- To compare RNA-seq libraries
- To compare the expression level of several genes within a library

- Indeed read counts depend on
  - Expression level
  - Gene length
  - Library size

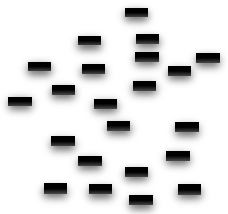
- Method used for RNA-seq is based on the “effective library size” concept
  - Assumption
    - Most genes are not differentially expressed
  - 2 methods
    - Trimmed Mean of M values (Robinson et al. Genome Biol. 2010;11:R25)
    - DESeq normalization (Anders et al. Genome Biol. 2010;11:R106)

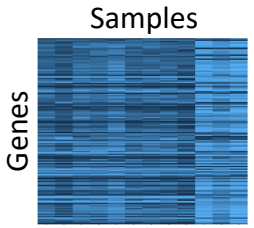


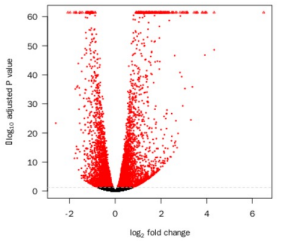
Sample name	Total number of reads
Cond1 sample1	43,672,265
Cond1 sample2	46,565,834
Cond2 sample1	43,985,979
Cond2 sample2	51,348,313

Analysis of RNA-seq data

From Nucleotide sequence to count matrix







Statistic to search for
significantly differentially expressed genes

- What is significant differential expression ?
  - The observed difference between conditions is statistically significant
    i.e. greater than expected just due to random variation

- Microarray vs RNA-seq
  - Microarray
    Fluorescence proportional to expression ➔ continuous data
  - RNA-seq
    Number of reads assigned to a feature (gene, transcript) proportional to expression ➔ count data

- Here we focus on count-based measures of gene expression

Statistic to search for
significantly differentially expressed genes

- Hypothesis testing
  - For each gene
    - H0 : No gene expression difference between the compared conditions
    - H1 : There is a gene expression difference between the compared conditions
- Steps
  - Choose a statistic (Biological replicates ~ Negative binomial distribution)
  - Define a decision rule
    - Define a threshold below which we will reject H0

Hands-on

Significantly differentially expressed genes (results)

- For all genes
  - Log FC (Fold Change): ≈ log (mean normalized values WT / mean normalized values NK-EOMES$^{-/-}$)
  - P-value: is a measure of how likely you are to get this gene if no real difference existed. Therefore, a small p-value indicates that there is a small chance of getting this gene if no real difference existed.
  - FDR: p-value are corrected for multiple testing with the False Discovery Rate method. When you are working with hundreds and thousands of genes you are going to get false positives just by chance, this is inherent to the definition of p-value. Setting a threshold on the adjusted p-value is used to control the number of false positives. For instance with an adjusted p-value < 0.05, we expect 5% of false positives in the gene list

5. Differential Expression Table

Gene expression signatures are alterations in the patterns of gene expression that occur as a result of cellular perturbations such as drug treatments, gene knock-downs or diseases. They can be quantified using differential gene expression (DGE) methods, which compare gene expression between two groups of samples to identify genes whose expression is significantly altered in the perturbation. The signature table is used to interactively display the results of such analyses.

Gene	logFC	AveExpr	P-value	FDR
*Ryr1	-3.94	5.49	1.746138e-09	0.000029
*Gpx6	-7.30	0.62	1.371891e-08	0.000113
*Scn4b	-3.88	8.97	2.539462e-08	0.000139
*Adora2a	-2.43	6.58	5.225166e-08	0.000181
*Abi3bp	-3.83	2.85	6.714640e-08	0.000181

Download Signature

Table 3 | Differential Expression Table. The figure displays a browsable table containing the gene expression signature generated from a differential gene expression analysis. Every row of the table represents a gene; the columns display the estimated measures of differential expression. Links to external resources containing additional information for each gene are also provided

Significantly differentially expressed genes (results)