Bioinformatics for high-throughput data

Today's Outline

Part A :   Introduction on high-throughput data
           Experimental design
           Platforms for gene expression data

Break – 10 min (questions if any)

Part B :   Exponential and logs
           Low level analysis
   - Gene Expression Estimation
   - Normalisation

Class Activity: debate.

Today we will be learning:

- Characteristics of high-throughput data and how it is produced

- What we intend for OMICS

- How we define high-throughput data

- To define principles of experimental design and pipelines

- Methods for gene expression quantification

- To estimate gene expression levels from data and difference in methods applied

- To normalise the data and what it means

High-throughput data

High-throughput data is a large amount of data collected using automated methods and non-conventional technologies. This is to be able to perform a large number of experiments at the same time to monitor a behavior of a system.

High throughput *cell biology* is the use of automation equipment with classical cell biology techniques. This is to address biological questions that are otherwise unattainable using conventional methods.

High-throughput biology has created a new field of biology called *OMICS*. It is a research filed that interface between large scale biology (genome, proteome, transcriptome), technology and computational methods.

Characteristics of high-throughput data

What are the characteristics of high-throughput data?

- Large in size

- Prone to many false positives (low specificity)

- Capture biological Noise

- Accurate

- Technically noisy

Next Generation Sequencing

OMICS data is produced mainly through modern techniques for sequencing DNA.

It differs from traditional Sanger sequencing for its power of generating fast and efficiently high-throughput data.

Next Generation Sequencing (NGS) it is a technique used to sequence millions of small individual fragments of DNA in parallel. These segments are than mapped onto the reference genome under study to identify:

- Genomic mutations: SNPs, Deletion Insertion and translocation
- Quantification of gene expression
- Quantification and identification of DNA methylation loci
- 3D Genome of single –cell using Hi-C, a molecular method that enables the identification of genomic regions that are in close proximity

OMICS and its hierarchy

To microscopic to macroscopic
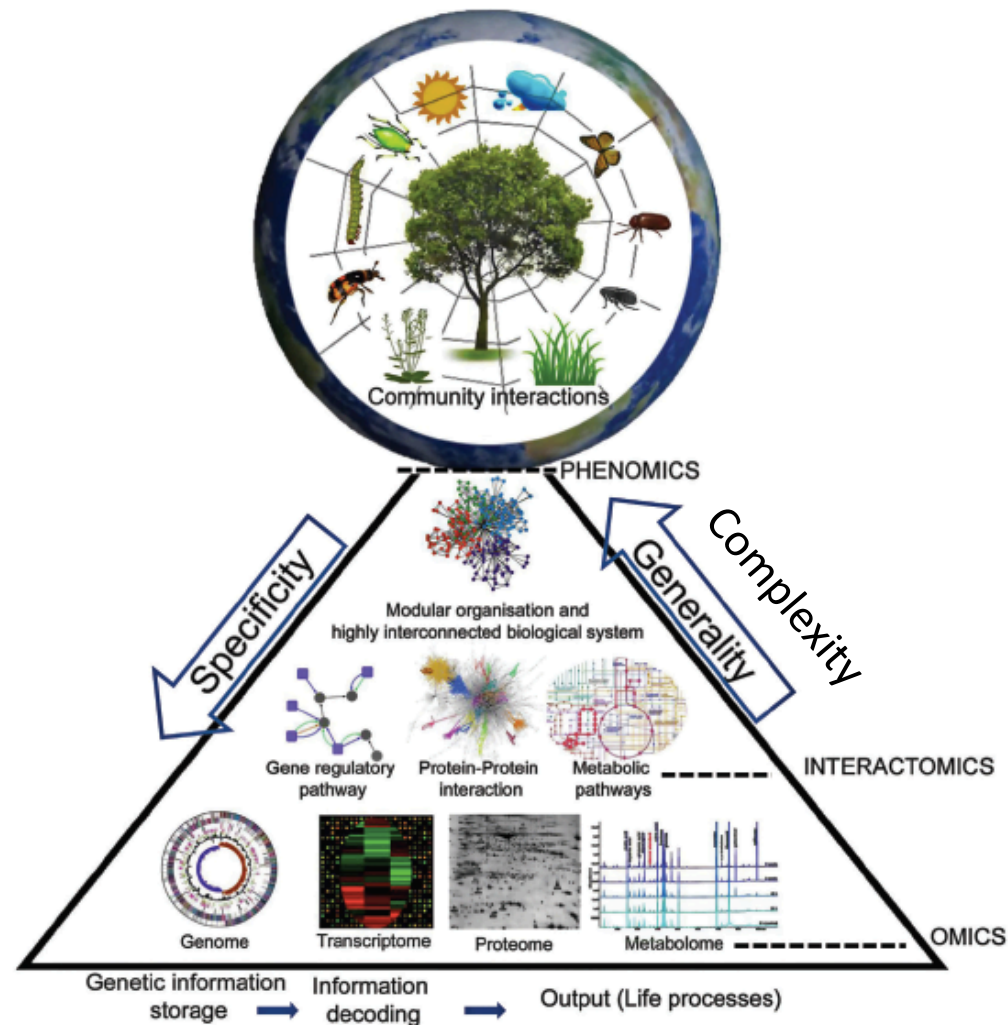
The highly interconnected hierarchical organization and functional complexity need a sophisticated integrated systems approach.

BMS353

Why do we need mathematics to understand biology?

Quantification of genome wide gene expression
Gene networks and target predictions
Protein-Protein Interaction Networks

Biological systems

Alignment tools
Mining of large data
Accessing large data from repositories

Large Data

Handling complexity
High dimensions
Patterns in the data

Patterns and structures

Quantification of Uncertainty: Be able to predict
what we cannot measure and have a theoretical
way of quantify accuracy

Modeling

Computational Biology --- Bioinformatics

Tools we need

- Optimal Experimental Design – minimise the noise in the measurement

- Mathematical Models – define rules (functions) to describe processes

- Statistical tools – quantify accuracy in prediction and sensitivity in estimation

- Computational Skills – handling large amount of data in automated way

- Visualisation tools – identify patterns in the data
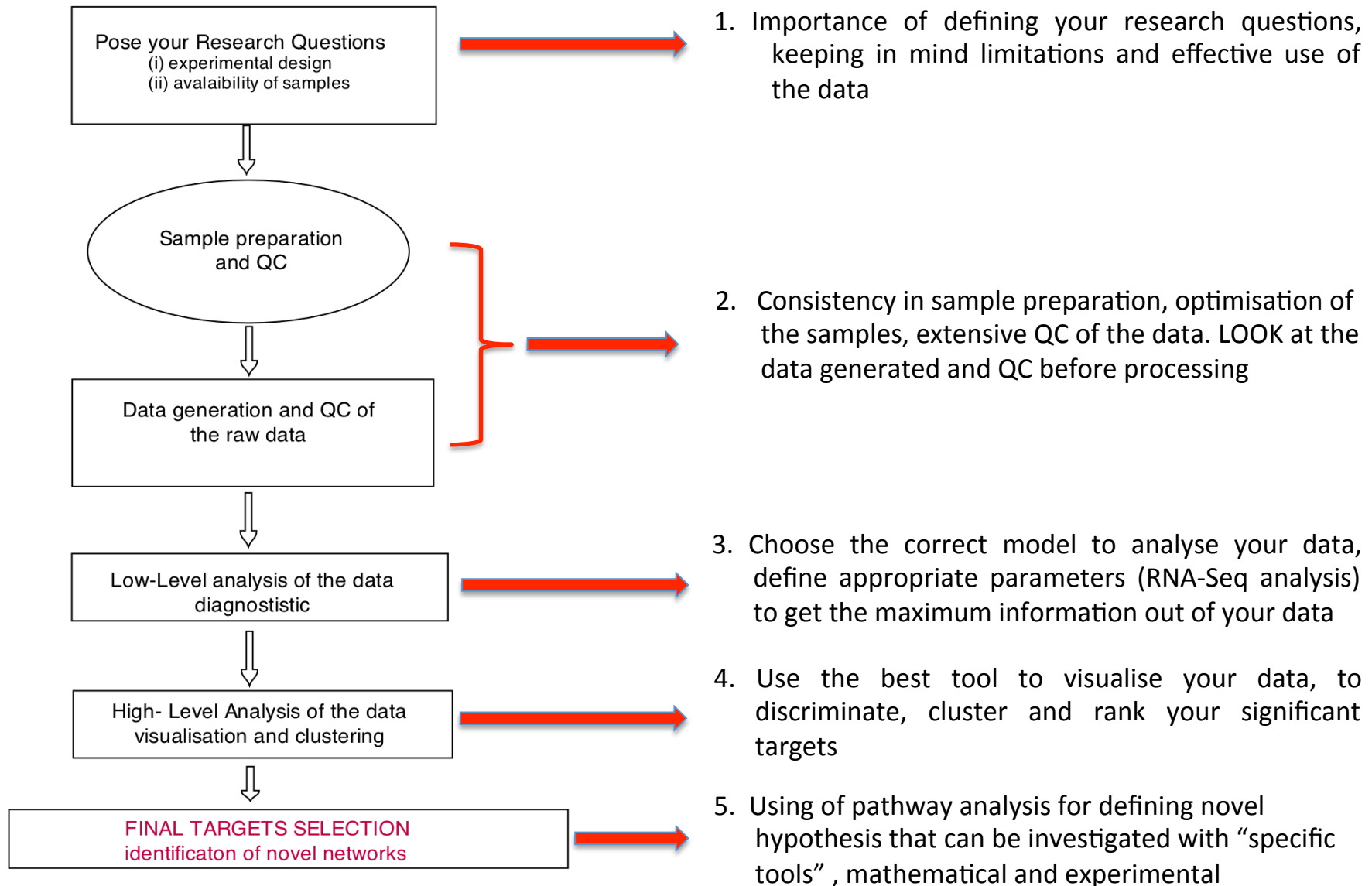
The role of experimental design

Defining your research questions, keeping in mind limitation and effective use of the data

We need to limit the uncertainty of the "unknown" by define very clear questions. This helps to :
- reduce variability in the data.
- reducing the complexity of the data by focusing the search of information on the questions you have in mind.

A proper experimental design **MUST** reflect :
- the biological questions that you are asking,
- the protocols optimised to minimise the variations in the data
- identify the limitations of the data collected.

Pose your Research Questions
(i) experimental design
(ii) avalaibility of samples

Sample preparation and QC

Data generation and QC of the raw data

Low-Level analysis of the data diagnostistic

High- Level Analysis of the data visualisation and clustering

FINAL TARGETS SELECTION
identificaton of novel networks

1. Importance of defining your research questions, keeping in mind limitations and effective use of the data

2. Consistency in sample preparation, optimisation of the samples, extensive QC of the data. LOOK at the data generated and QC before processing

3. Choose the correct model to analyse your data, define appropriate parameters (RNA-Seq analysis) to get the maximum information out of your data

4. Use the best tool to visualise your data, to discriminate, cluster and rank your significant targets

5. Using of pathway analysis for defining novel hypothesis that can be investigated with "specific tools", mathematical and experimental

BMS353

Workflow and Pipelines in our analysis

- They help to define the correct implementation of the data

- Makes us think on the logical steps we need to take

- Our analysis can be reproduced on new data or different data

In the lab we have the task to analyse a real case study :
data collected from qPCR single-cell data

1. Reorganisation of data
   QC data

2. Data transformation

Exercise 9 and 10

3. Data analysis
   normalisation

Pipelines

All the variation and to ensure that the analysis of the data is as reproducible as the experimental collection of samples generate the need to define pipelines for the analysis of the data

PIPELINES: Reproducible and robust protocols for numerical experimentations. In case of biological data they are tailored to the system/organism under study.

HOW DO WE GET THEM?

Gene expression estimation

Interpret and analyse the data by first understanding
where the data is coming from

Different platforms that generate gene expression:

- Two or One color spotted cDNA arrays

- Affymetrix - new Human/Mouse whole transcriptome arrays

- Illumina Arrays

- Reads – RNA-Seq and other NGS assays

BMS353

GeneChip® Array

Single stranded, labeled RNA sample

Oligonucleotide element

20μm

1.28cm

Image of Hybridised Array

Millions of copies of a specific oligonucleotide sequence element

~ 1,000,000 different complementary oligonucleotides
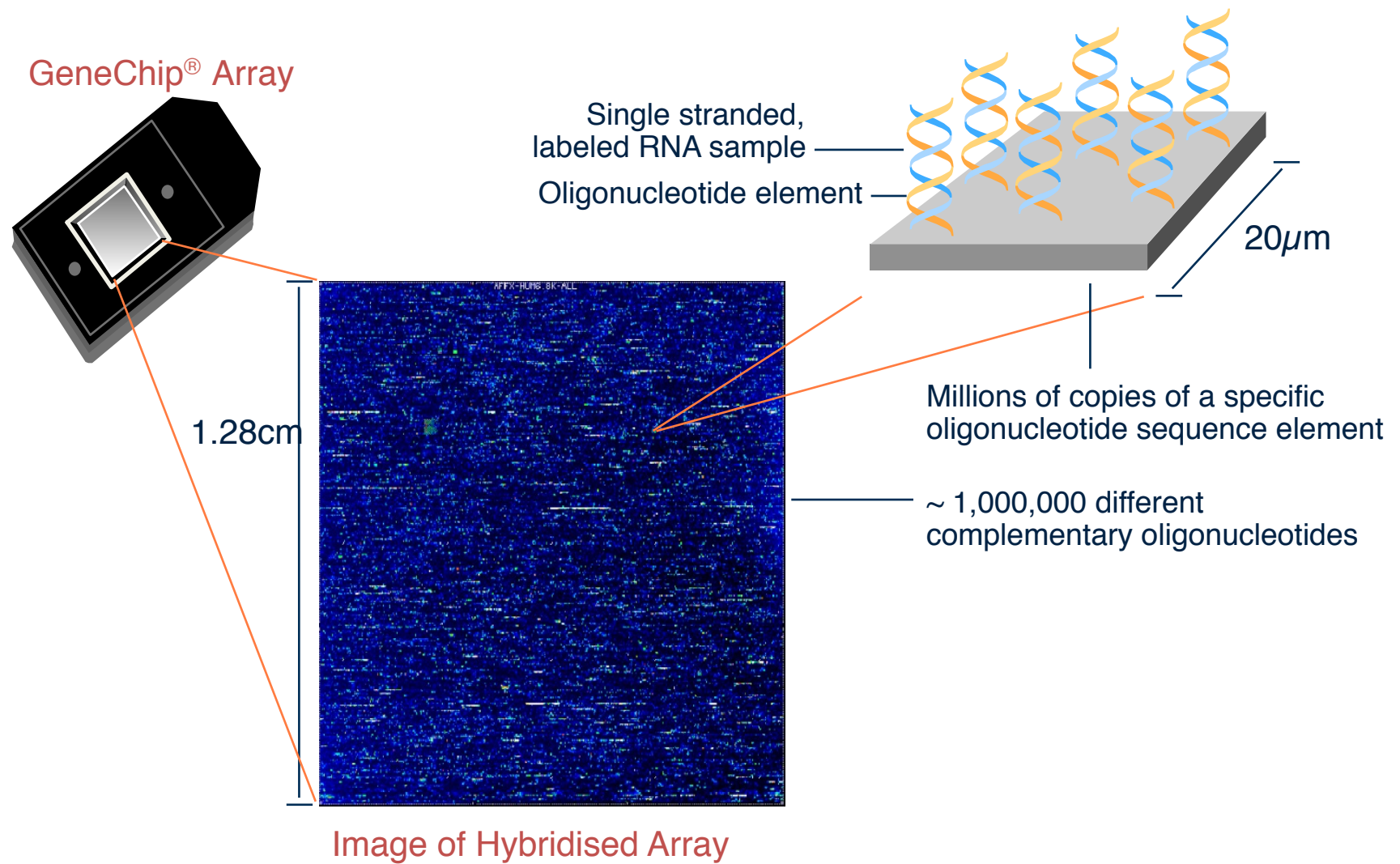
Example: HG_U133 Plus v2 Affymetrix geneChip

The sequences from which these probe sets were derived were selected from GenBank®, dbEST, and RefSeq.

A single array contains with more than 54,000 probe sets representing approximately 38,500 genes (estimated by UniGene coverage).

70 percent of the probe sets represent subcluster assemblies containing one or more non- EST sequences. Of the 16,737 EST-based probe sets, approximately 9,000 probe sets can now be associated with an mRNA or other non-EST sequence.
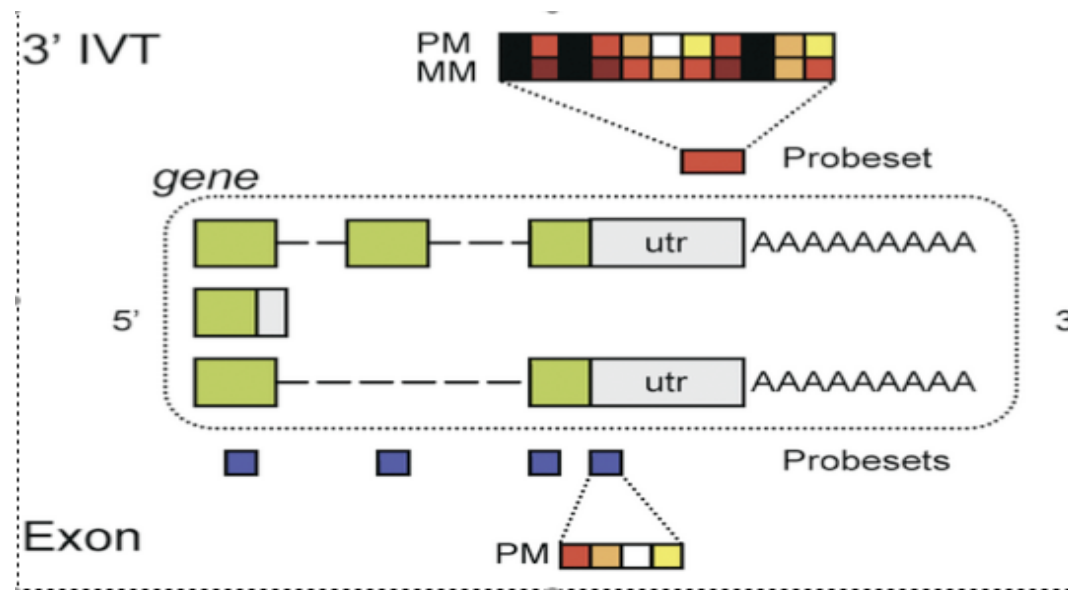
Now with new arrays HJAY…..

Probe Set Notation
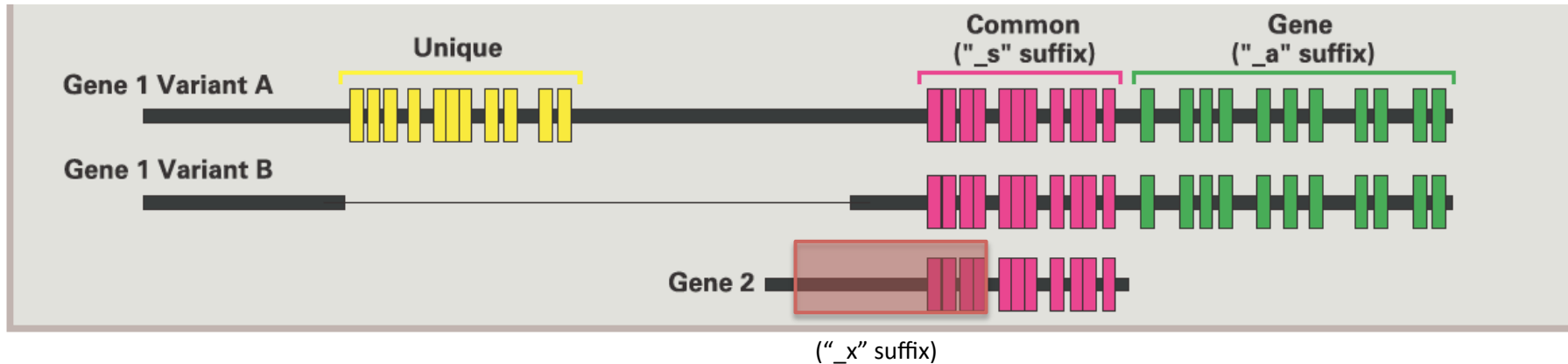
_at : probe sets are predicted to perfectly match only a single transcript

_s_at : are predicted to perfectly match multiple transcripts, which may be from different genes

a_at : all probes in the probe set hit alternate transcript of the same gene

_x_at : probe sets will contain some probes that are identical or highly similar to other sequences from different gene.

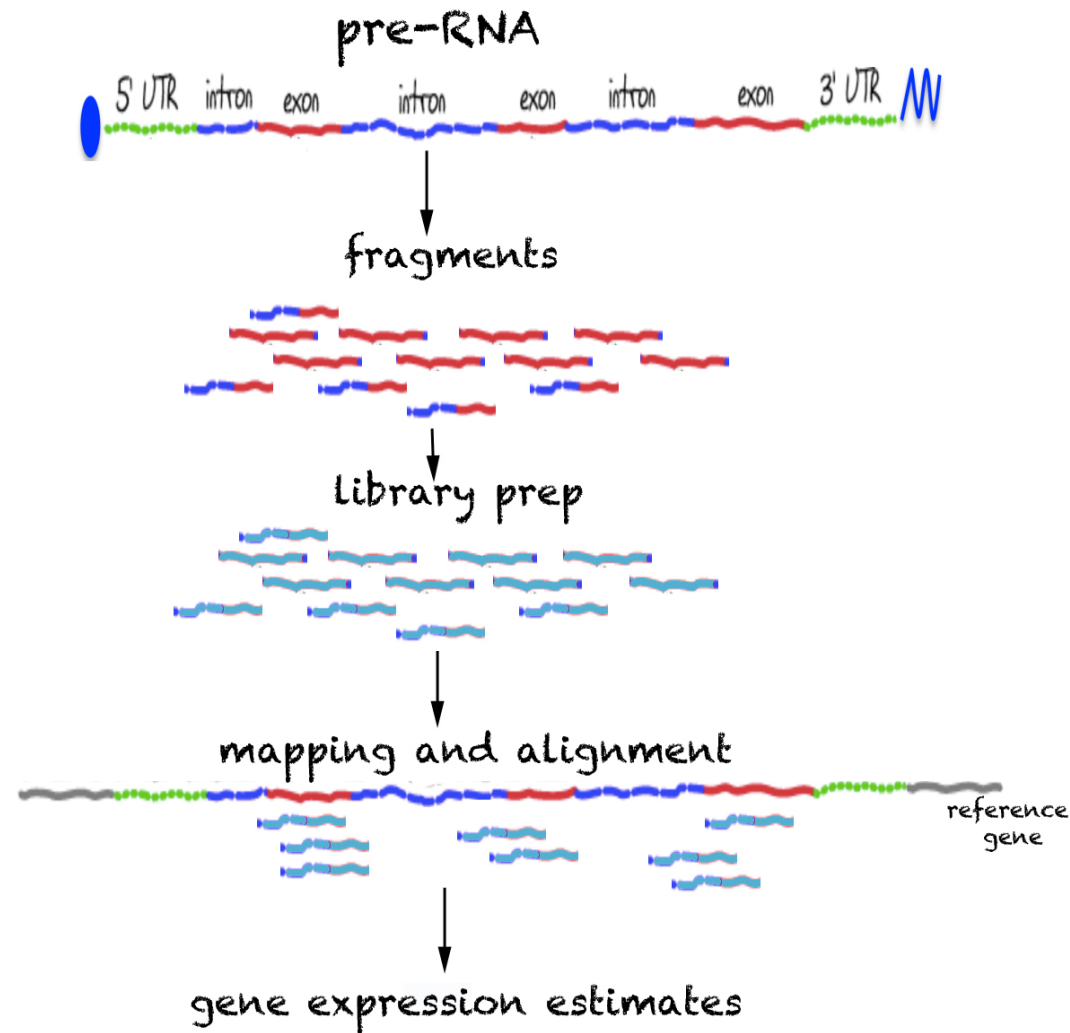Hybridize uniformly across probe pairs to the intended target

Gene expression quantification using RNA-Seq
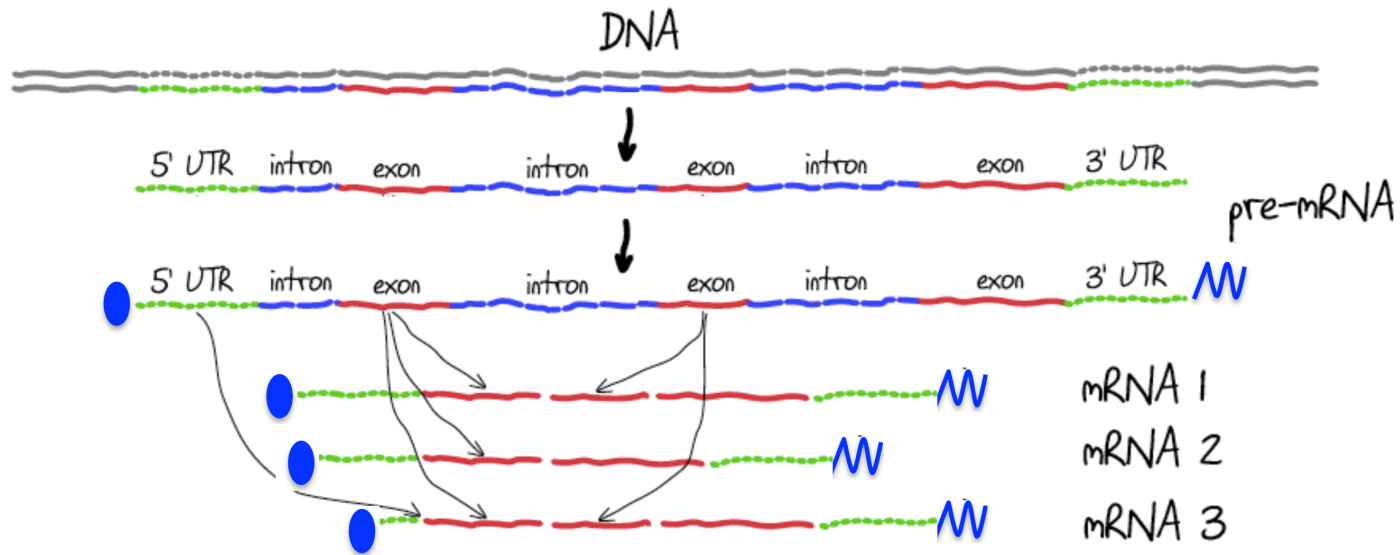
High-throughput sequencing of cDNA:

- Shared exons
- Biological variance of fragments
- Splicing variations

Example of RNA-Seq workflow

Illumina sequencing – HiSeq


Illumina HiSeq sequencer

What are we able to detect?

- Mapped read count proportional to abundance of fragments
- Abundance of fragments ≈ (gene expression) x (length)
- Problems which length? which transcripts?
- Other difficulties: mismatches, varying quality of reads, non-uniform read distribution

Part B

What's in the GeneChip data?

1. Summarise to a single expression level the probe intensities for each probe set

2. Estimate the variations introduced by
   - background effect
   - probe affinity effect

3. Some PM/MM pairs are more reliable than others

4. The signal needs to be scaled before comparing data from different arrays

How we define a measure that best represent the absolute expression level of each gene on the chip?
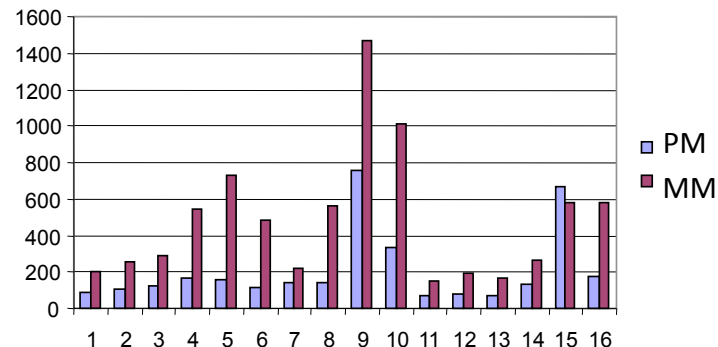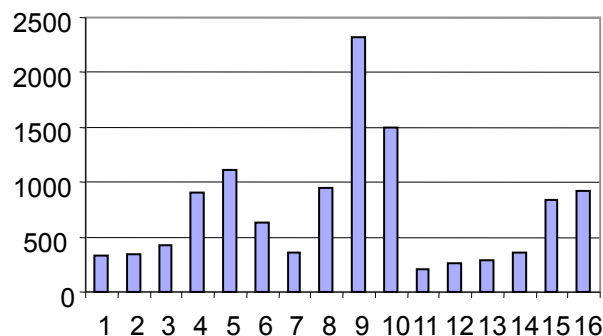
The approaches

Use **Statistical methods (single point statistics)**

make use of the information we have to define values that estimate gene expression

MAS 5.
RMA – GCRMA
PLIER

Use a **probabilistic approach (in Week 10- probability functions)**

make use of the observed data to estimate function that have generated that data

Estimates of gene expression will be the most probable value that summarises the probe set

Data transformation

Data transformation converts a set of data values from the format of its original source into a new format that better suits the destination system where the data is then manipulated/used/ mined etc...

In statistics, data transformation refers to the modification of every point in a data set by a mathematical function.

Data transformation is most often used to change data to the appropriate form for a particular statistical test or method.

In practice:
- calculate the gene expression as the 2^{-deltaCt}
  deltaCt<-sweep(Ct_full,MARGIN=2,Ct_full[1,],FUN="-")
  Gene_expression=2^-deltaCt

Data Transformation: Exponential functions

$f(x) = 2^x$
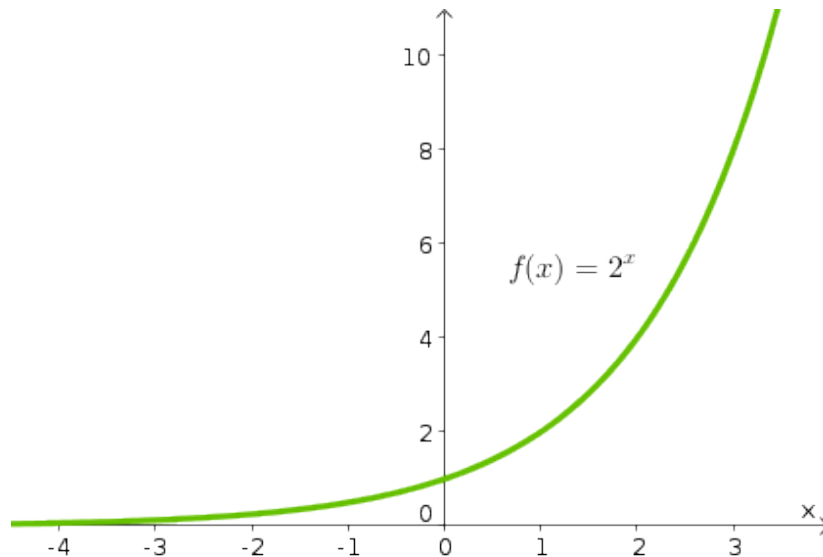
$$y = f(x) = e^x = \exp(x)$$

$$1 + x + x^2 + \ldots + x^n$$

$$x^n * x^m = x^{n-m}$$

$$x^n / x^m = x^{n-m}$$

The number **e** is the Euler's number, an irrational number
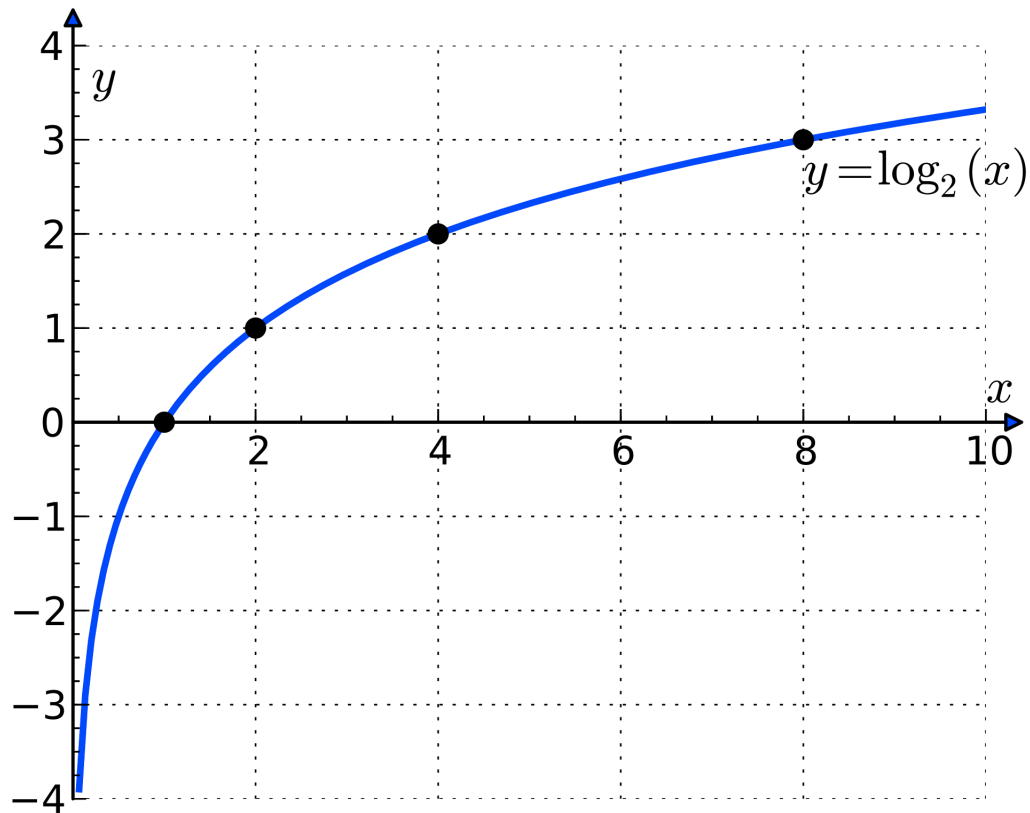The first few digits are:
2.71828182845904523536028747713527 (and more ...)

Gaussian Function.

$$y = f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

BMS353
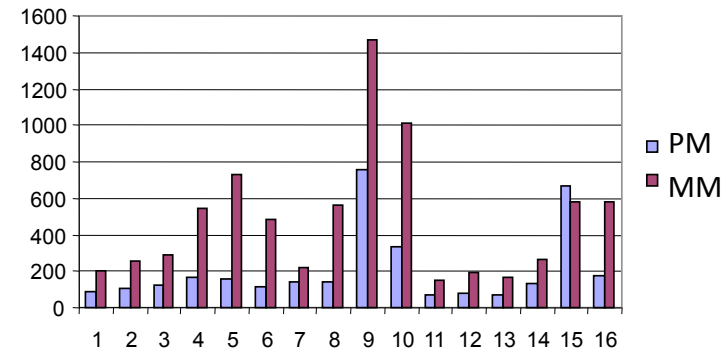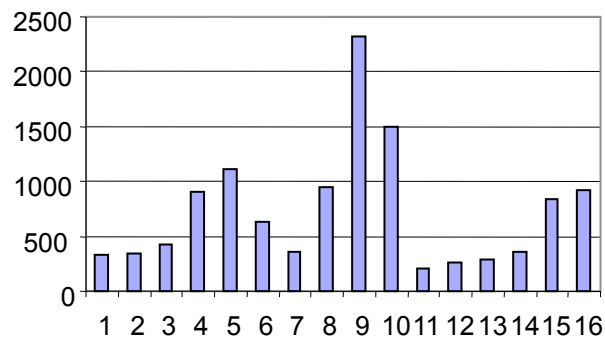
Data Transformation: Logarithmic functions

$y = \log_2(x)$

$log_b(xy) = log_b x + log_b y$
$log_b(x/y) = log_b x - log_b y$
$log_b(xn) = n \, log_b x$
$log_b x = log_a x \, / \, log_a b$

y = log$_b$x if and only if b^y = x,
where x > 0, b > 0, and b ≠1.

Estimation of Gene expression

1. How we define a measure that best represent the absolute expression level of each gene on the chip?

2. How we ensure that comparison across chips in unbiased?

Normalisation : removing bias

Why do we need to normalise the data?

    1. we want to compare across chips
    2. we need to ensure that all the data is equally compared across baseline
       within the chip

Most methods will have normalisation step incorporated, some other will need to perform it after gene expression estimation

Scaling – Mean and Median
Quantile

Loess (not relevant for affymetrix data and sequencing data)

Normalisation: scaling

The assumption that normalising using **quantiles or scaling** is reasonable, is based on the assumption that "most genes don't change"

If this underlying assumption is doubtful, then using the above methods is not advisable.

Scaling normalisation, linearly scale the gene expression values so that the overall mean (or median) are the same.

The median is more scale-invariant, but for the most part there is little practical difference.

Normalisation: quantile

In statistics, *quantile normalization* is a technique for making two distributions identical in statistical properties.

When we quantile-normalise a sample distribution to a reference distribution of the same length, we align the sample distribution to the reference so to make them the same.

Assume that the distributions of probe intensities should be completely the same across chips.

Start with n arrays, and p probes, and form a *[p,n]* matrix X.

Rank first:
Sort the columns of X, so that the entries in a given row correspond to a fixed quantile (Q2= the median==mean).

Then align:
Replace all entries in that row with their mean value.

Microarray Suite (MAS5.0)

- Signal = Smoothed average over PM,MM pairs representing a gene

- Signal is always positive: Absent - Present Call

$$Signal \sim TukeyBiweight(log_2(PM_j - IM_j))$$

Correction for global background.- based on 11 sectors on each array

Ideal mismatch (IM) intensity calculated from MM value and subtracted from PM.
- if MM < PM then IM = MM
- if MM > PM then IM = PM − correction value

MAS5: characteristics

- Not very precise
- accurate only when many replicates are available.
- Dependent strongly on MM
- Uses linear scaling normalisation

Robust Multi-array Average (RMA)

Signal = regression-based average over PM pair representing a gene

$$Signal \sim Tukey\ (log_2(PM_j - bkgd_j))$$

- Subtract background for each array from PM

- Intensity- dependent normalisation of PM-Bkgd

    - Quantile normalisation :Fit all the chips to the same distribution. Scale the chips so that they have the same mean.

- Log transform

Robust Multi-array Average (RMA)

- Precise
- Only works if there are replicates
- accurate only when many replicates are available.
- Quantile normalisation flattens the tails. Only strong signals are detected.
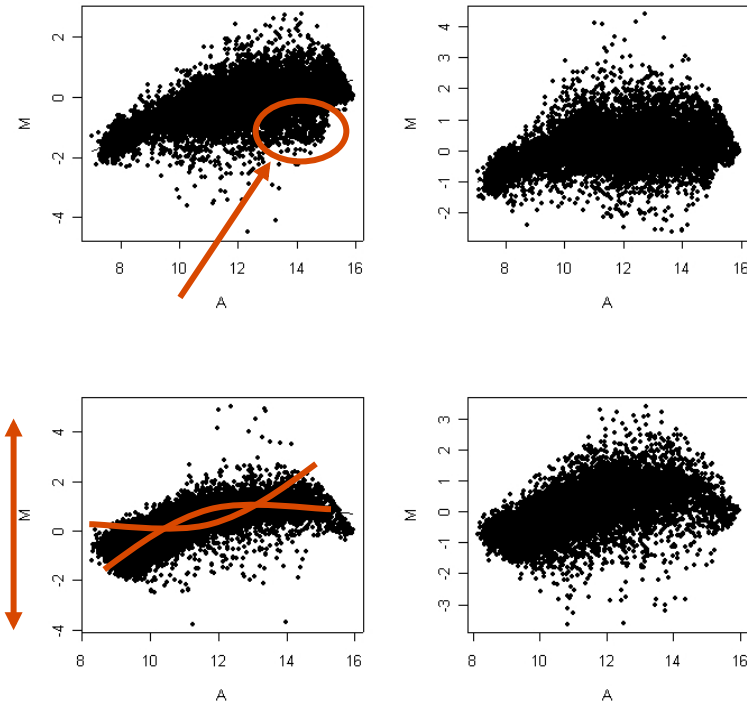
- Visualise the effect: M-A plot

$$M = \log_2\left(\frac{R}{G}\right)$$

$$A = \log_2\left(RG\right)$$

- Correction of the intensity dependant variations:

$$\log_2\left(\text{ratio}\right) = \log_2\left(\frac{R}{G}\right) - c(A)$$
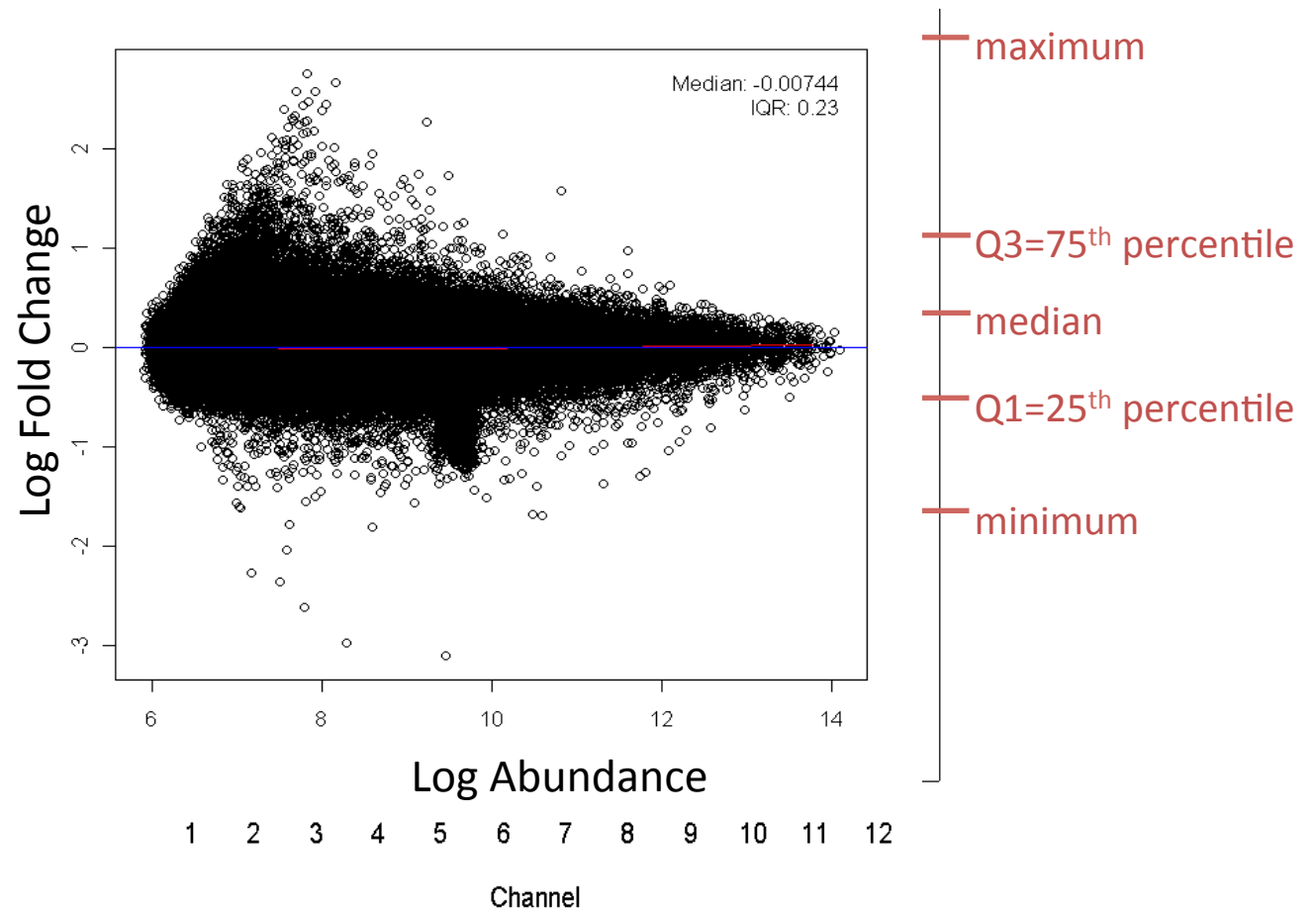
$$c(A) = \log_2 k(A)$$

Data Visualisation

Scatter Plot

Box Plot

MA Plot

BMS353

Class activity: debate the following questions:

1. How much do I need to know about the system that I am studying?

2. How much the technologies that are available for data collection need to to sensitive for my system?

3. What is sensitivity and specificity?

4. In sequencing what is a reference genome and how I get it?

5. When the high-throughput approach is the correct approach for my research question?