The University Of Sheffield.

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

The University Of Sheffield.

Normalisation

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

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.

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.

The University Of Sheffield.

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

The University Of Sheffield.

MAS5: characteristics

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

The University Of Sheffield.

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

The University Of Sheffield.

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.

The University Of Sheffield.

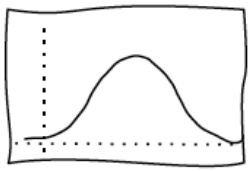
Probabilistic Models

A *probability model* is a mathematical representation of a random phenomenon. It is defined by its sample space, events within the sample space, and probabilities associated with each event.

These are models that represents unknowns in terms of probability distributions instead of values and a confidence interval.

For example if we assume that the data is generated by:

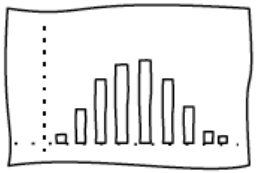
$$\theta \sim N(\mu|\sigma)$$



I

we can measure few samples of it, but we don't know the true distribution:

$$\theta \sim \mathbf{S} = \{s_1, \ldots, s_n\}$$



The University Of Sheffield.

Probabilistic Models (cont.)

We can use a probability theory to manipulate those functions (probabilities) and make inference on the unknown parameters as well as evaluate the *uncertainty* that is associated to their estimates. In other words:

- We describes data that one could observe from a system
- We use the mathematics of probability theory to express all forms of uncertainty and noise associated with our model...
- We use inverse probability (i.e. Bayes rule) to infer unknown quantities, adapt our models, make predictions and **learn** from data.

Why are they good?

- They Faithfully represent uncertainty in our model structure and parameters and noise in our data
- They are automated, adaptive and robust
- They scale well to large data sets

The University Of Sheffield.

Bayes' Rule

$$P(hypothesis \mid data) = \frac{P(data \mid hypothesis)P(hypothesis)}{P(data)}$$

Rev'd Thomas Bayes (1702-1761)

- Bayes rule tells us how to do predict outcomes of hypotheses from data. We can do inference about hypotheses once we have observed the data

- Learning and prediction can be seen as forms of inference. If we use Bayes' Rule we call it *Bayesian Inference*

P(hypothesis) = prior
P(hypothesis|data) = posterior
P(data|hypothesis) = likelihood
P(data) = marginal likelihood

Not always possible to be computational efficient and likelihood are estimated using sampling methods.

The University Of Sheffield.

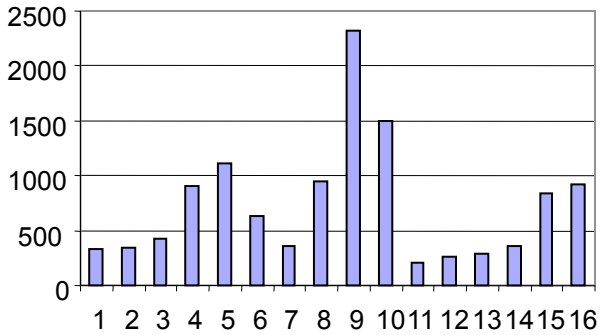
How do probabilistic model help in gene expression analysis?

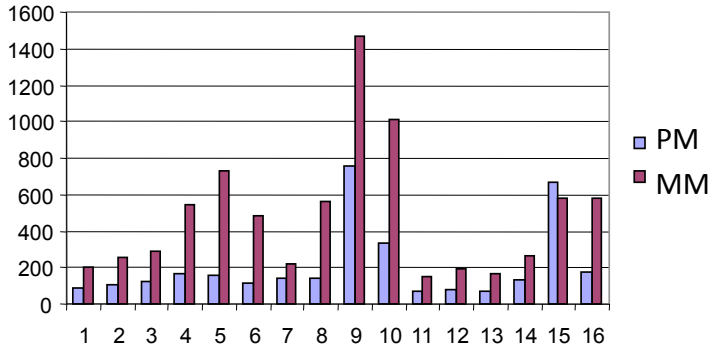
We know that:

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



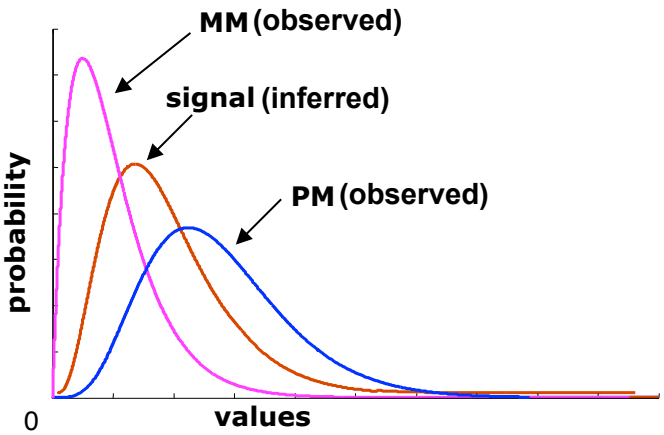


4. The signal needs to be scaled before comparing data from different arrays

They can help to define a measure that best represent the absolute expression level of each gene on the chip?

The University Of Sheffield.

puma
Propagating Uncertainty in Microarray Analysis



Milo M *et al*, Biochem transcation 2003
Liu X *et al*, Bioinformatics 2005
***Pearson R et al*, BMC Bioinformatics, 2009**

Estimate the distribution of the data and we learn the parameters to define it from the data (gamma distributions)

We built priors on the hypothesis (our belief is that the true signal is gamma distributed

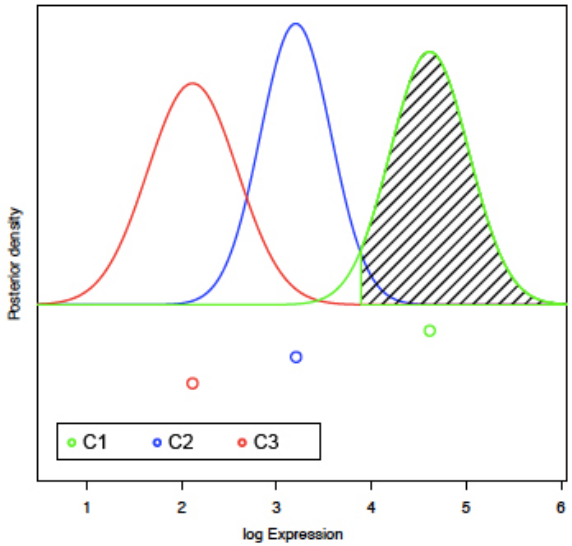
We then calculate the likelihood using the model defined by Affymetrix

Signal= PM-MM

We apply Bayes' rule to calculate the signal distribution (posterior)

Computational Efficient --- we don't need to use sampling methods.

Differential Expression: pumaDE and PPLR



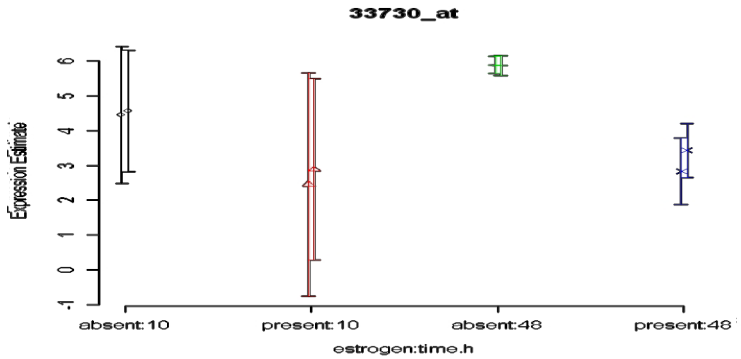
Probability of Positive Log ratio: PPLR

Gives the probability of the log ratio (log FC) to be positive.

$$FC = \log(\frac{x}{y})$$

Example:
In differential Expression analysis the goal is to estimate genes that change across conditions.

What happens then if we do not evaluate uncertainty?



Data from Choe et al, *Genome Biology*(2005)

The University Of Sheffield.

High Level Analysis

The University Of Sheffield.

Principal Component Analysis

It is one of the most commonly used technique to visualise and interpret high dimensional data

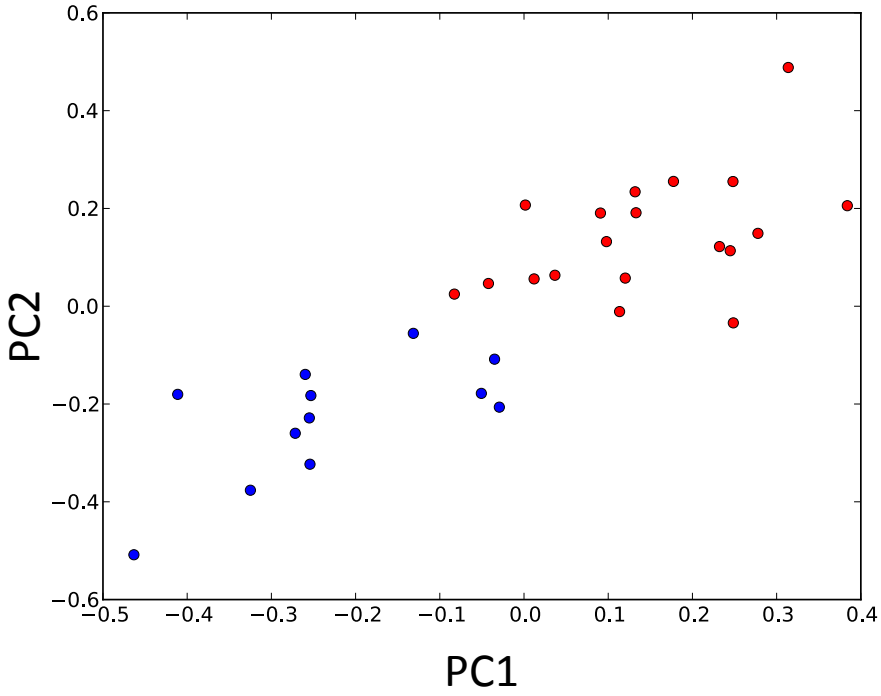
It identifies the maximum spread of the data maximising the variance by rotating the space where the data lives.

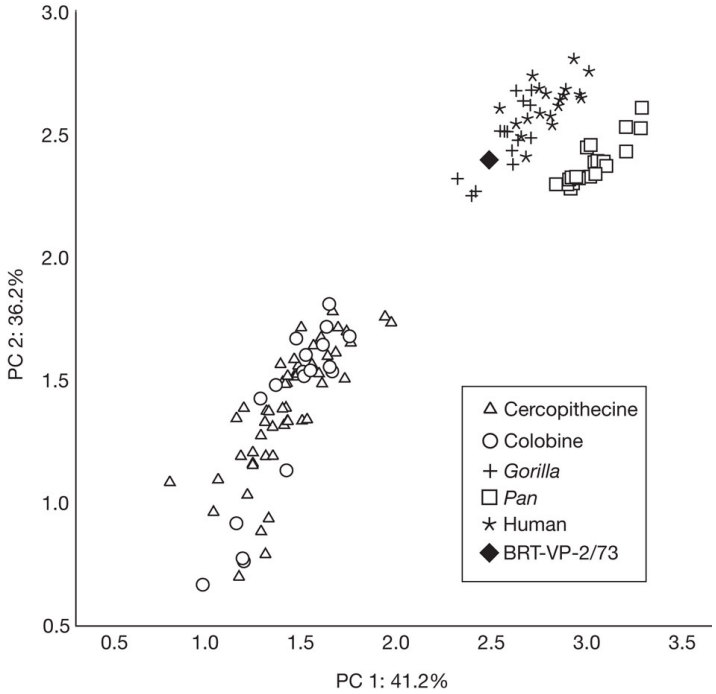
It uses a set of variables that are hidden to the user and are implicitly explained by the data (latent variables)

Every direction found that extract informative features from the “noisy” cloud of data points is called a principal component

Dimensionality reduction

Principal Component Analysis (*cont…)*





Y Haile-Selassie *et al*. *Nature* **483**, 565-569 (2012)
doi:10.1038/nature10922

usually reasonable, but it assumes that the uncertainty associated to each gene is constant

non-linear transformation of gene expression (Huber et al. 2002), PUMA PCA (Sanguinetti et al., 2005)

```
pca_estrogen <- prcomp(t(exprs(eset_estrogen)))
pumapca_estrogen<-pumaPCA(eset_estrogen_puma)
```

The University Of Sheffield.

Clustering

- basic idea: group together genes that have similar pattern of expression across conditions or across time

- what do we mean by similar?

- different measures of similarity: Euclidean distance, angle
- between vectors, correlation coefficient, . . .

- Shared pattern of expression might be associate to similar functions

The University Of Sheffield.

Similarity measures

A *similarity measure* (*function*) is a function that quantifies the similarity between two objects.

You can think of it as the inverse of distance metrics:

- large values for similar objects;
- zero or a negative value for very dissimilar objects.

E.g., in the context of cluster analysis we can use the following similarity measure:

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$ squared Euclidean distance

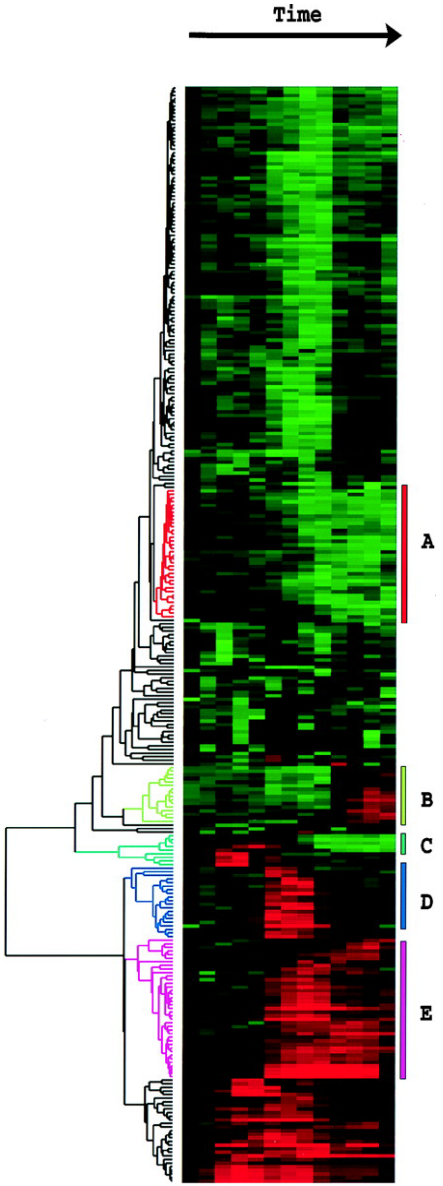
$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$ Pearson Correlation

A similarity matrix is a matrix of scores that represent the similarity between a number of data points. Each element of the similarity matrix contains a measure of similarity between two of the data points. (from *Wikipedia*)

The University Of Sheffield.

Hierarchical Clustering

- builds a hierarchy of clusters
- bottom up (merging clusters) or top down (splitting clusters)
- Eisen et al. (1998). The genes that are most correlated are joined together, the expression value for the resulting node is the average expression of the two (or more) genes. The similarity matrix is then updated with the new node.
- Different similarity measures lead to different interpretations



The University Of Sheffield.

Functional Analysis

The University Of Sheffield.

Questions we will address

- What is functional analysis?
- Open source packages which enable to explore the pathways

Using methods like pathways analysis we can build:

- Gene Network – not in this module

The University Of Sheffield.

Gene Ontologies

The **Gene Ontology (GO)** is a controlled vocabulary of terms to describe gene product characteristics in the domains of localization and function.

The aims of the GOs are:

- Maintain and develop its controlled vocabulary of gene and gene product attributes
- Annotate genes and gene products
- Acquire and disseminate annotation data
- Provide tools for easy access to the databases
- enable functional interpretation of experimental data using the GO

The ontologies covers these domains:

1. cellular component: the parts of a cell or its extracellular environment;
2. molecular function: the elemental activities of a gene product at the molecular level, for example as binding or catalysis;
3. biological process: operations or sets of molecular events

The Gene Ontology Project is what promoted and created this concept:

http://geneontology.org/

The University Of Sheffield.

Pathway analysis

To identify interaction between genes based on literature knowledge: description of annotation and pattern of expression, association to disease etc.

There are two ways of approaching this type of analysis:

Top down or bottom up

Top down:
Look at the whole organism and abstract large portions of it

Bottom up:
Try to understand each small piece and assemble into the whole

Both are used, valid and interconnect.

The University Of Sheffield.

Pathway analysis (*cont…*)

Biological annotations have started to include descriptions of gene interactions in the form of gene signaling networks, such as KEGG (Ogata et al.,1999), BioCarta (www.biocarta.com) this makes the pathway analysis *in silico* more informative than it was in the past.

There is a variety of open source software available for this.
There are also some very good commercial packages

We will explore the two main pathways navigators based on annotations and protein interactions.

The University Of Sheffield.

DAVID

The Database for Annotation, Visualization and Integrated Discovery (**DAVID**)

http://david.abcc.ncifcrf.gov/

Entirely based on ontologies and annotations relies on KEGG and BioCarta

The University Of Sheffield.

PANTHER

The **PANTHER** (Protein ANalysis THrough Evolutionary Relationships)
It is a Classification System designed to classify proteins (and their genes)

Proteins have been classified according to:
Family and subfamily: families are groups of evolutionarily related proteins; subfamilies are related proteins that also have the same function

Molecular function: the function of the protein by itself or with directly interacting proteins at a biochemical level, e.g. a protein kinase

Biological process: the function of the protein in the context of a larger network of proteins that interact to accomplish a process at the level of the cell or organism, e.g. mitosis.

Pathway: similar to biological process, but a pathway also explicitly specifies the relationships between the interacting molecules.

http://www.pantherdb.org/about.jsp

The University Of Sheffield.

What else …

PANTHER, to define a list of enriched pathways that are represented by the list of genes that you are interested in.

You can also look at more systems biology oriented approaches, where probabilistic models are used to manage the annotations and the gene interactions, but this is not part of the module

The University Of Sheffield.

Project Allocation and Discussion

The University Of Sheffield.

Learning with projects

- What bioinformatics can do for biology: the projects
- The importance of knowing your data in the context of what you are studying
- What are the project that will be allocated to you for the final assignment
- Critical discussion of the module material

The University Of Sheffield.

BMS353 assessment

The exam for this module will be split in two parts:

Part A – A Multiple Choice Question test for the duration of 1hr, that will count 30% of the final grade

Part B – A notebook with the implementation of allocated projects that will count for 70% of the final grade.

The project will be a collection of all the tools experienced in the practical labs implemented on a set of real data. It will be developed in groups of three students, but notebook will have to be handed individually.

MCQ assessment:

Each question will have 4 possible responses A, B, C or D. **ONLY ONE RESPONSE IS CORRECT IN EACH CASE.** Each question is worth one mark, correct answer will count as 1, an incorrect answer will count as 0. **Not answered questions will count as 0**.

Assessment criteria

The grading for each sub-session follows the scale below:

1. Fail
2. Pass
3. Lower Second
4. Upper Second
5. First

1. **Pipelines and experimental design (20 points)**

   Structure of the pipeline (5 point)
   Overall clarity (5 points)
   Use of details (5 points)
   Exhaustive cover of required analysis (5 points)

2. **Use of methods (20 points)**

   Use of data visualization methods (5 points)
   Use of analysis methods (5 points)
   Use of annotation (5 points)
   Use of functional annotation tools (5 points)

The University Of Sheffield.

Assessment criteria (cont…)

3. Use of programming tools (20 points)

Clarity of algorithm (5points)
Correctness of the code (5points)
Efficient programming, speed of code (5 points)
Use of innovative tools/code (5 points)

4. Use of basic statistics (20 points)

Appropriateness of the statistics (5points)
Correct implementation of the methods (5 points)
Novel statistics used (5 points)
Interpretation of the results (5 points)

5. Overall impression and interpretation of the results (20 points)

Overall clarity of the notebook (5 points)
Clarity of the code documentation (5 points)
Innovation (5points)
Biological interpretation of the results (5 points)

Feedback

The University Of Sheffield.

Some examples of projects

The University Of Sheffield.

Project A

This study is to explain the effect of the transcription factor SP1 in colon cells. To elucidate this effect a colon cell line was used and a silencing of the transcription factor SP1 was obtained using RNAi techniques *in vitro*. A gene expression profile of the cells with SP1 silencing and without silencing was done after 48hrs in culture.

The expression profiles were quantified using Affymetrix GeneChip HGU133 PLUS 2. The files containing the data are as follow:

* M48-1.CEL control at 48hrs in culture - sample 1
* M48-2.CEL control at 48hrs in culture - sample 2
* S48-1.CEL SP1 silenced at 48hrs in culture - sample 1
* S48-2.CEL SP1 silenced at 48hrs in culture - sample 2

After estimating gene expression, visualise the data and describe the findings. Identify which genes are changing between conditions and define any potential pathway that the silencing of SP1 might have altered.

The University Of Sheffield.

Project B

This study is to explain the effect of the transcription factor SP1 in colon cells. To elucidate this effect a colon cell line was used and a silencing of the transcription factor SP1 was obtained using RNAi techniques in *vitro*. A gene expression profile of the cells with SP1 silencing and without silencing was done after 72hrs in culture.

The expression profiles were quantified using Affymetrix GeneChip HGU133 PLUS 2. The files containing the data are as follow:

* M72-1.CEL control at 72hrs in culture - sample 1
* M72-2.CEL control at 72hrs in culture - sample 2
* S72-1.CEL SP1 silenced at 72hrs in culture - sample 1
* S72-2.CEL SP1 silenced at 72hrs in culture - sample 2

After estimating gene expression, visualise the data and describe the findings. Identify which genes are changing between conditions and define any potential pathway that the silencing of SP1 might have altered.

The University Of Sheffield.

Project C

This study is to explain the effect of Hypoxia on human Neutrophils to identify possible involvement of inflammatory response in adverse prognosis of hypoxia-related disease, i.e. pulmonary hypertension, myocardial infarction. To elucidate this effect primary cultures of human neutrophils were studied at normal condition and in a hypoxia condition. A gene expression profile of the neutrophil in normal and hypoxia condition was done after certain amount of hrs in culture.

The expression profiles were quantified using Affymetrix GeneChip HGU133 PLUS 2. The files containing the data are as follow:

* LPGMa.CEL neutrophils at normal condition in culture - sample 1
* LPGMb.CEL neutrophils at normal condition in culture - sample 2
* LPHa.CEL neutrophils with hypoxia induced in culture - sample 1
* LPHb.CEL neutrophils with hypoxia induced in culture - sample 2

After estimating gene expression levels, visualise the data and describe the findings. Identify which genes are changing between conditions and define any potential pathway that the hypoxia might have altered in neutrophils.

The University Of Sheffield.

Project D

This study is to explain how Stem cells use their potential to generate different lineages, particularly on how they can provide a solution for replacing damaged or lost cells within the inner ear. It is known that human embryonic stem cells can be induced to differentiate into otic progenitors and then into hair cell-like cells and neurons that display expected electrophysiological properties. Once these otic progenitors are transplanted into animals with induced hearing loss, they differentiate and elicit a significant recovery of auditory function. The generation of otic progenitors is triggered by FGF signalling and this data aims to analyse the global gene expression profile of undifferentiated hESCs and compared with cultures that have been treated with FGF3 and 10.

The expression profiles were quantified using Affymetrix GeneChip HGU133 PLUS 2. The files containing the data are as follow:

H14_hES.CEL H14 embryonic cell line at normal condition in culture - sample 1
Shef1_hES.CEL Shef1 embryonic cell line at normal condition in culture - sample 2
H14_FGF.CEL H14 embryonic cell line cultured with FGF3 and 10 growth factors - sample 1
Shef1_FGF.CEL Shef1 embryonic cell line cultured with FGF3 and 10 growth factors - sample 2

After estimating gene expression levels, visualise the data and describe the findings. Identify which genes are changing between conditions and define any potential pathway that FGF3 and 10 might have altered in human Embryonic Stem Cells.

The University Of Sheffield.

Project E

This study is to explain how Stem cells use their potential to generate different lineages, particularly on how they can provide a solution for replacing damaged or lost cells within the inner ear. It is known that human embryonic stem cells can be induced to differentiate into otic progenitors and then into hair cell-like cells and neurons that display expected electrophysiological properties. Once these otic progenitors are transplanted into animals with induced hearing loss, they differentiate and elicit a significant recovery of auditory function. The generation of otic progenitors is triggered by FGF signalling and this data aims to analyse the global gene expression profile of undifferentiated hESCs and compared with cultures that are only culture in DFNB reach medium and not treated with FGF3 and 10.

The expression profiles were quantified using Affymetrix GeneChip HGU133 PLUS 2. The files containing the data are as follow:

H14_hES.CEL H14 embryonic cell line at normal condition in culture - sample 1
Shef3_hES.CEL Shef1 embryonic cell line at normal condition in culture - sample 2
H14_DFBN.CEL H14 embryonic cell line cultured in DFNB - sample 1
Shef3_DFBB.CEL Shef1 embryonic cell line cultured in DFNB - sample 2

After estimating gene expression levels, visualise the data and describe the findings. Identify which genes are changing between conditions and define any potential pathway that the DFNB medium alter in human Embryonic Stem Cells.