# Statistical Genomics

Ian Petrus Tan

## 0.1 False Discovery Rate

The False Discovery Rate (FDR) is a method proposed by Benjamini and Hochberg (1994) is a concept used for controlling multiple testing, based on the distribution of p-values, and was argued to be more powerful than the incumbent method at the time: the Family Wise Error Rate (FWER). The definition of the FDR is **the expected number of false positives over the total number of rejected null hypotheses**. This is useful in the context of genomics as it can help to control the rate of false positives in analysing gene expression data and differential analysis (e.g. using t tests to distinguish gene expression in two types of sample, such as treatment and control cells).

Benjamini and Hochberg (1994) argue that traditional methods of FWER such as Bonferroni correction result in a higher rate of false negatives, since it controls more strictly for significance thresholds. The FDR adjusts the p values for multiple testing to narrow down for significance for fewer (subset of) genes. First, the p values are ranked, then a cut off is determined such as 0.05, to determine which of the p-values are selected to represent significant tests. These genes can then be used for further research to be biomarkers, etc, whle minimising the number of false positives, as compared to not using the method at all and taking significance from p values purely.

$$FDR_i = \frac{g \times p_i}{r_i} = \frac{Expected\ number\ of\ false\ positives}{Total\ number\ of\ rejected\ nulls\ at\ \alpha}$$

Hence, FDR gives us a method of adjusting p-values to control for multiple testing, which is common in genomics analysis.

## 0.2 Advanced Regression

The penalised sparse regression is an important method used to find associations (and hence biomarkers) between relatively fewer genes and outcomes such as diseases, in the context of a large number of genes/features from the data. In these cases where g » n, g = number of features, n = sample size, we cannot use ordinary linear models that end up overfitting. Instead, we want a method to find a subset of genes that best explains/predicts an outcome of interest. The main idea is to regularise by constraining the coefficient parameters of the regression model (subset selection or shrinkage), resulting in a sparse model.

In cases of disease prediction, the outcome is binary, and we usually optimise for the loglikelihood function of a bernoulli model for a training set. In the case of sparse regression we add in an additional penalty term that adds a q-norm of the parameter coefficients to the loss function (negative of log likelihood), such that the fitting mechanism favours either a subset of features (subset selection) or shrinkage of parameters. The $\lambda$ hyperparameter controls for the number of features and/or the extent of shrinkage of the parameters.

$$l(\boldsymbol{\beta}) = \frac{1}{n} \times l(\boldsymbol{\beta}) - \lambda \|\boldsymbol{\beta}\|_p$$

, where $\|\boldsymbol{\beta}\|_p$ is the p-norm of the coefficient vector.

## 0.3 Local Network Structures from epiblast data

**0.3.0.0.1 EDA for gene expression data** To get a good sense of the transcription data, I will pick to plot 2 of the highest positive correlation pairs of transcription factors, 2 of the most negative, and one that is close to 0

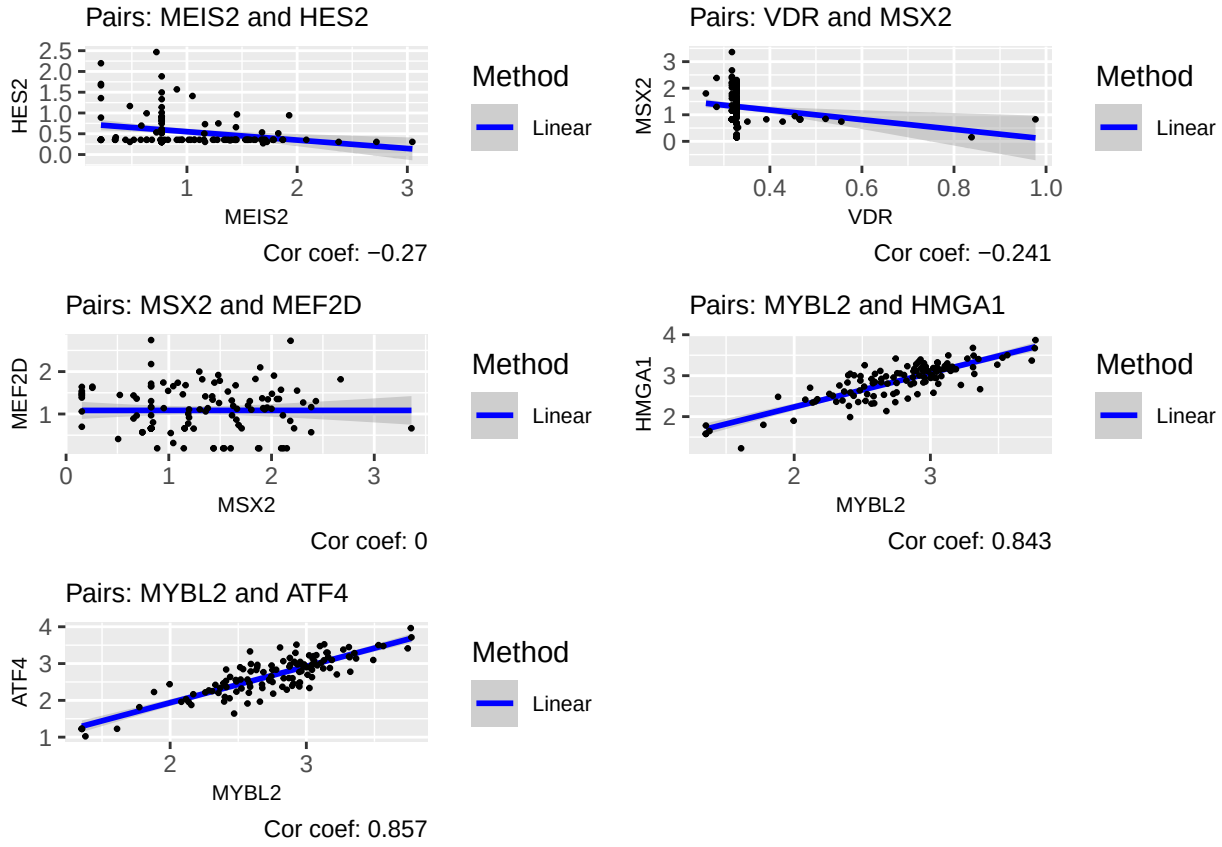


Figure 1: Exploratory Analysis: 5 Pairwise Plots

From Fig 1, the first two plots reveal pairplots that have negative correlation, the middle plot show lack of correlation, and the final two show positive correlation.

These plots reveal that some pairs of genes may be antagonistic or inversely regulated expression levels. They could have opposite functions where one expression affects the other. On the other hand, some genes, like the `MYBL2` and `HMGA1` have high correlation (>0.8), indicating that their expression levels are positively related and that the genes are co-regulated or are both used in similar biological pathways. Furthermore, the two highest positive correlation pairs involve `MYBL2`, which could mean that it is a general-purpose or possibly a housekeeping gene, although more research or expertise would be needed in this area. Finally, there are genes that are not correlated and possibly independently regulated, such as `MSX2` and `MEF2D`.

From these plots, linearity does not seem to be violated (non-linearity) and outliers do not look like a big issue. Apart from the 2 negative correlation plots, where there seem to many data points that lie on or around the same X or Y line, the data seem to be somewhat normally spread. Log function is typically used to make gene expression data more Gaussian. Hence, Pearson may overall be the better measure of correlation.

**0.3.0.0.2 Local Network Structure via Pearson Correlation** From Fig 2, we can see that the MYBL2 TF that was selected from the high correlation pairwise plot (EDA) is highly connected to many other TF/genes. This could mean that MYBL2 has a more central role in gene regulation pathways and/or involved in more biological processes, while HES2 has less regulatory influence on other TFs/genes. MSX2 is in the middle, with connections to 2 others. It should be noted that we used a high threshold of 0.99th percentile of the correlation coefficients, and the graphs would change if we had set a different percentile.

**0.3.0.0.3 Local Network Structure via Advanced Regression** The advanced regression method produces different plots in Fig 3, and is able to pick out connections for the 'low connection' HES2 gene, identifying 3 other nodes now, such as STAT3. Notably, MSX2 is now very much connected to other TFs/genes, showing that it has involvement with other genes in regulatory pathways. The comparison with the previous method using Pearson correlation may not be clear, since previously we had to choose a specific threshold, while here we simply identified genes with non-0 (1e-10) coefficients from the regression, which is a low threshold.

## 0.4 Synthetic data and AUC

**0.4.0.0.1 4.3 AUC Plots** On the left of Fig 4, holding `n = 100` and `n = 1000` constant, we can see that increasing the number of genes (and therefore the number of dimensions) generally decreases the AUC (despite there being a slight increase at the end for both).

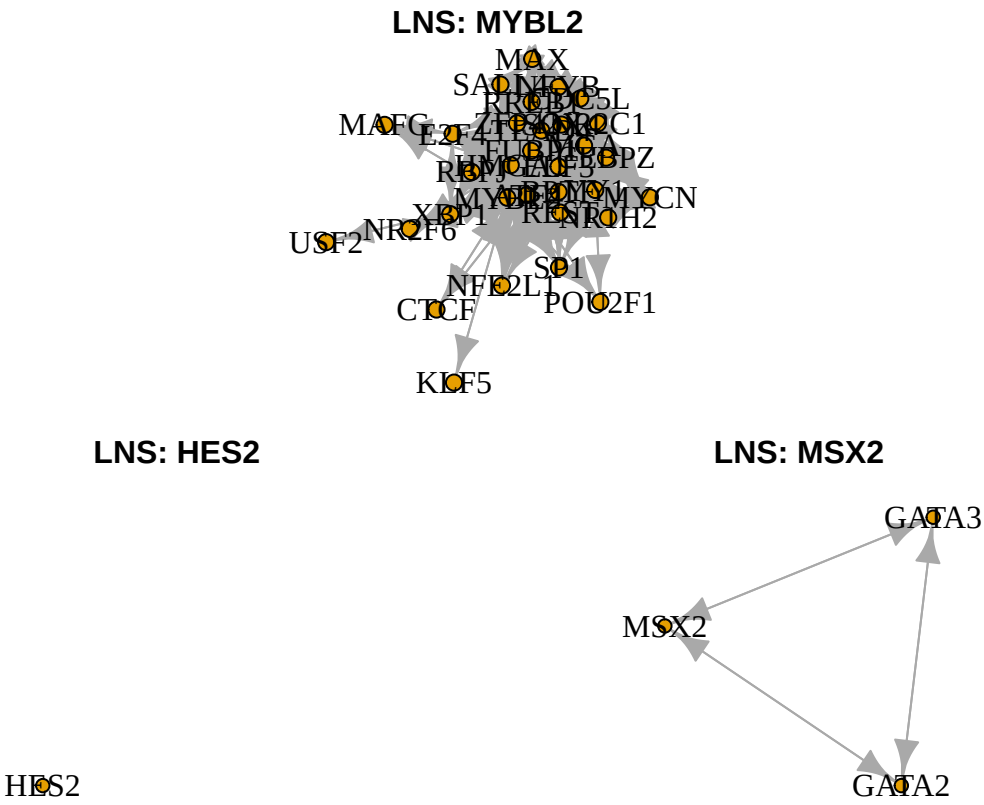


Figure 2: Correlation Method: Local Network Structures (LNS), 99-th percentile correlation as threshold

This may be due to the high dimensionality making the model too complex and overfitted (high p relative to n), hence worsening the models ability to detect connections between TFs.

On the right, we can see that generally, AUC increases as the number of samples increase. Meanwhile, while holding `p = 10` constant, we get a more noisy trend line which fluctuates up and down as n increases, but arguably has an overall increasing general trend. This could be because when there is more data for the same number of TFs, the regression coefficients estimates are more precise and better able to predict a presence of a connection between genes when there is one. However, for `p = 50` which represents a higher dimension, this effect of increasing sample size does not hold.

#### 0.4.0.0.2 4.4 Difficulty in obtaining AUC in real genomics data, how to validate equivalent findings from real data

It might be difficult to get AUC data due to small number of valid cell samples from real data. This may be due to quality control, such as poor quality samples due to cells being damaged or sick before mRNA data could have been measured, or if the cell has been contaminated. Few cells might meet the requirement in terms of minimum threshold of transcripts in a cell.

Furthermore, as we have seen from the data, when there is high dimensionality (number of genes) especially when sample size is small, the AUC suffers and the model is less able to detect the presence of relationships between genes/ TFs. If we are able to extrapolate the case between `p = 10` vs `p = 50`, increasing the dimension negates the benefit of increasing sample size.

Other reasons could include natural biological variability between cell samples, blurring the connections between TFs. There is a possibility of the models finding transitive edges as well, causing false positives.

To validate equivalent findings from real data, we could use other sources of data such as epigeomic data, to infer causality between genes and proteins. For example, chromatin accessibility data can help us figure out DNA locations that are accessible for TF, and this can help us infer causality between this TF and a specific gene down the DNA, and whether it is likely to be activated or repressed. We can better identify regulatory pathways that are not confounded (e.g. transitive edges)

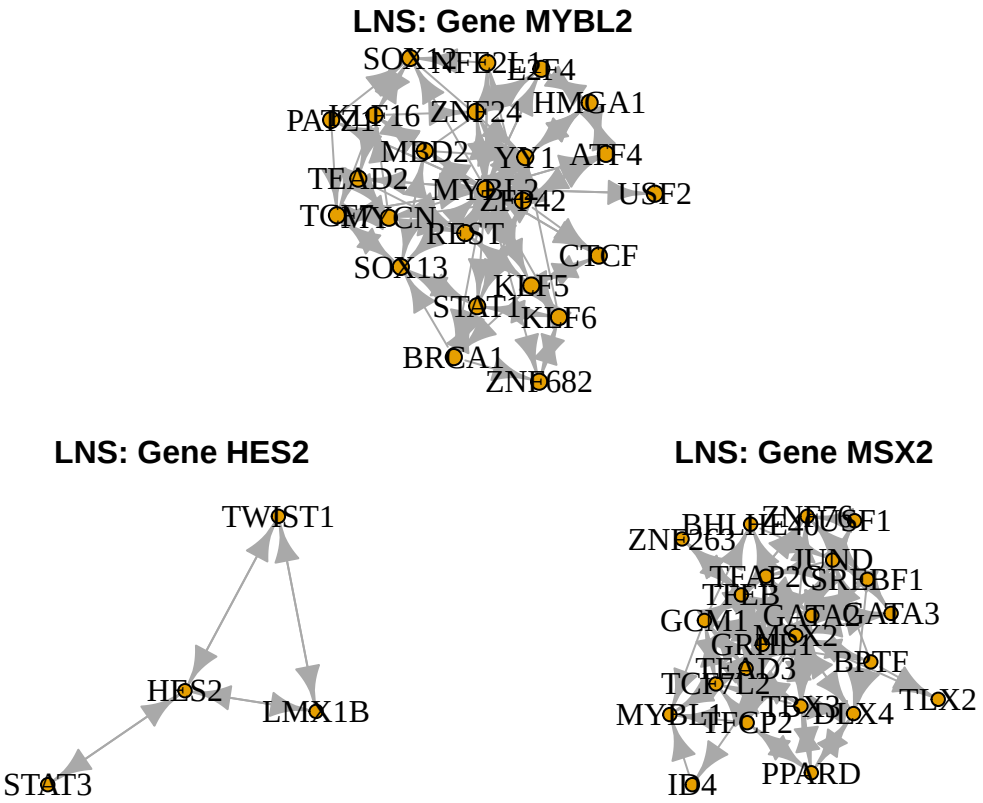


Figure 3: Regression Method: Local Network Structures (LNS)

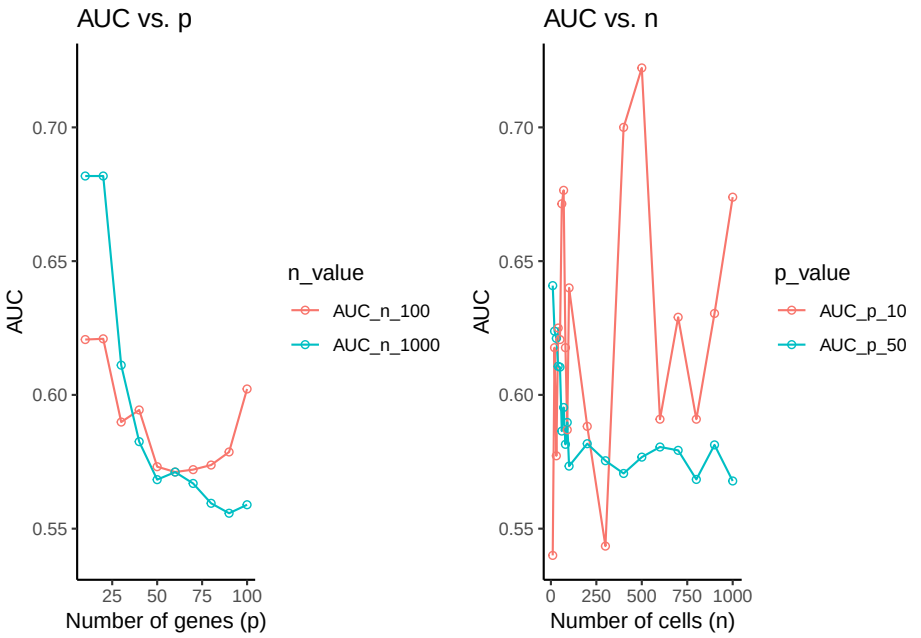


Figure 4: AUC when varying p (n = 100,1000), when varying n (p = 10,50)