RNA-seq workshop

Differential expression analysis

Mik Black & Ngoni Faya
Genomics Aotearoa

Fold changes

- Differential expression (i.e., a change in gene activitation level) is often reported as a fold change in activity.
- Often the log_2 scale is used (i.e., log fold change).
- Initially, genes with fold changes greater than 2 ($log_2(2) = 1$) or less than 1/2 ($log_2(\frac{1}{2}) = -1$) were considered to have undergone differential expression.

Detecting changes in expression

- In order to determine whether a gene has undergone differential expression between two conditions, multiple observations are generally required. Note: In reality, all we can determine is whether the probes which represent a gene, exhibit consistent changes in intensity.
- Assuming that we have multiple intensity measurements for a gene under each condition, basic statistical methods can be used to answer this question.

Determining differential expression

- We are investigating differences in gene expression between two strains of yeast (WT and MT)
- We have three replicates of the WT samples and three replicates of the MT samples (6 RNA-seq samples in total).
- For each gene this gives:
  - Treatment 1 (WT): x_{11}, x_{12}, x_{13}
  - Treatment 2 (MT): x_{21}, x_{22}, x_{23}

Determining differential expression

- If we assume that all experimental artifacts have been removed by the normalization process, we conclude that any remaining differences in intensity are result of differences in expression level.
- To test this, we can conduct a formal hypothesis test (for each gene) to determine whether the mean intensity changed between the WT and MT samples.
- Since most basic statistical tests are set up to provide answers on the additive scale, and fold changes are on the multiplicative scale, we generally take logs of the data.

Hypothesis testing

- In statistics, we think of our sample means as providing estimates of the underlying (true) population means for each gene, μ_1 and μ_2.
- For each gene, we want to test the following null hypothesis: $H_0 : \mu_1 = \mu_2$ against the alternative hypothesis: $H_A : \mu_1 \neq \mu_2$
- If we reject the null hypothesis for a particular gene, we think that gene is likely to be differentially expressed.

Hypothesis testing

- In order to conduct the hypothesis test, we need a test statistic. The most simple approach is to utilize the test statistic of the standard t-test:

$$T = \frac{\hat{\mu}_1 - \hat{\mu}_2}{SE(\hat{\mu}_1 - \hat{\mu}_2)}$$

  where $\hat{\mu}_1$ and $\hat{\mu}_2$ are the sample means of the data, and $SE(\hat{\mu}_1 - \hat{\mu}_2)$ is some appropriate measure of variability (in this case the standard error).

- Various choices are possible for the denominator depending on the structure of the data.

P-values

- Once we have calculated a gene-specific test statistic (e.g., a t-statistic in our simple example), we calculate a p-value for each gene, p_k.
- The p-value represents the probability of observing this (or a more extreme) result, if no differential expression occurred. (i.e., what is the chance we are just observing noise?)
- We reject H_{0k} (i.e., say gene k is differentially expressed) if p_k is small.
- Question: what does small mean?

P-values

- We have to decide how small a p-value needs to be for us to think that the difference we are observing cannot be explained solely by noise.
- When we test a single hypothesis, it is common to fix a Type I error rate of α= 0.05 or α = 0.01.
- **Type I error:** reject null hypothesis when it is true (i.e., say a gene is differentially expressed when it really isn't).
- **Type II error:** fail to reject the null hypothesis when it is false (i.e., say a gene is not differentially expressed when it really is).

Type I errors

- Using a Type I error rate of α= 0.05 means that we are willing to make a Type I error in 5% of our hypothesis tests (i.e., 5% of the time that the null hypothesis is true, we will say that it's false).
- So for every 20 hypothesis tests we perform, on average we expect 1 Type I error.
- What if we are performing 20,000 hypothesis tests?

1000 TYPE I ERRORS!

Adjusting the α level

- Obviously using an α level of 0.05 (or even 0.01) is not suitable when testing large numbers of hypotheses.
- To get around this problem we use Multiple Comparison Procedures (MCPs).
- MCPs provide error rate control, allowing us to keep a lid on how many Type I errors we make.

Family-wise error rate control

- Control of the family-wise error rate (FWER) is very common in multiple testing problems.
- MCPs which control the FWER guarantee that the FWER $< \alpha$, where a "family-wise error" is defined to be the occurrence of a single Type I error in the entire family (set) of hypotheses being tested.
- In an RNA-seq experiment we test each gene for differential expression, so there are as many hypothesis tests as there are genes.
- The Bonferroni and Holm procedures both provide control of the FWER.

What's so great about FWER control?

- Advantage: FWER controlling procedures provide a high level of certainty in your result. The null hypotheses rejected by these procedures are very unlikely to be true (i.e., all of the rejected null hypotheses are likely to be correct rejections).
- Disadvantages: This level of control is very conservative - it is likely that some genes undergo differential expression, but their null hypotheses are not rejected. As the number of hypotheses being tested becomes very large, the significance threshold becomes extremely small.

What is the alternative?

- Continue to control the FWER, but use a larger value?
- Switch to a different error rate?
- What other error rates exist? (not many...)

False Discovery Rate control

- The False Discovery Rate was introduced by Benjamini and Hochberg (1995 - JRSS(B)).
- Provides a less conservative approach to error rate control than FWER controlling procedures.
- Greater power comes at the cost of an increased likelihood of Type I errors.
- Has become very popular in genomic analysis, plus astronomy, brain imaging, and genetics (all test large numbers of hypotheses).

Error rate control

- FWER control is concerned with making sure that the probability of a single testing error is small.
- FDR control in concerned with keeping the proportion of Type I errors out of the total number of rejected hypotheses small.
    - This value can be anything from 0 to 1.

FDR control versus FWER control

- FDR controlling procedures provide more error rate protection than not adjusting at all, but are a lot more likely to make Type I errors than FWER controlling procedures.
- The flip side is that FDR controlling methods are more likely to reject false null hypotheses (i.e., they achieve greater power).

Comparing approaches

- Instead of adjusting the signifiance threshold, we can adjust the p-values themselves.
- The table below contains unadjusted p-values ("P-value"), and p-values adjusted using the Bonferroni, Holm, and FDR methods.
- For α=0.05, the four approaches find 7, 2, 4, or 6 tests significant.

TEST NUMBER	P-VALUE	BONFERRONI	HOLM	FDR
1	0.002	0.016	0.016	0.0160
2	0.004	0.032	0.028	0.0160
3	0.007	0.056	0.042	0.0187
4	0.010	0.080	0.050	0.0200
5	0.020	0.160	0.080	0.0320
6	0.030	0.240	0.090	0.0400
7	0.050	0.400	0.100	0.0571
8	0.080	0.640	0.100	0.0800

Modification to t-test procedure

- One problem with the t-statistic approach to determining significance is that some genes with small, but consistent fold changes can end up with very large t-statistics.
- This is especially common in genomics experiments involving only a few samples.
- Generally feel that genes with small fold changes shouldn't be considered as having undergone significant differential expression.
- Need a way to prevent these genes showing up as significant.

Significance Analysis for Microarrays (SAM)

- Tusher et al. (2001) proposed a modification to the denominator of the t-statistic to reduce the influence of tiny standard deviations.

$$T = \frac{\hat{\mu}_1 - \hat{\mu}_2}{SE(\hat{\mu}_1 - \hat{\mu}_2) + s_0}$$

- Although this modification looks somewhat arbitrary, it can be derived by taking a Bayesian approach to analysis (and various other ways).
- The Bayesian derivation relies on the assumption that the standard errors for each gene have an underlying common distribution. The s_0 parameter then contains information from this underlying distribution.

Significance Analysis for Microarrays (SAM)

- Although simple, this approach is highly effective, and has become a popular method for detecting genes undergoing significant differential expression.
- Various methods can be employed for estimating the s_0 parameter.
- Tusher et al. (2001) chose s_0 to minimize the coefficient of variation.
- Other authors have suggested using quantiles of the underlying empirical (observed) distribution of standard errors (much easier).

Significance Analysis for Microarrays (SAM)

- Has the effect of restricting significant genes to those exhibiting large fold changes.
- Although the distribution of T is unknown, resampling methods (e.g., bootstrapping) can be used to approximate the null distribution, allowing calculation of p-values.
- Multiple comparison procedures can then be used to provide control of the Type I error rate (FWER or FDR).

Detecting differential expression with `limma`

- The `limma` package takes a linear models approach to detecting genes which have undergone differential expression.
- After the data have been normalized, a linear model is fit to the expression values to determine which genes underwent significant changes.
- Although a standard t statistic can be used to assess differential expression, `limma` goes a little bit further...

Empirical Bayes analysis

- Limma uses Empirical Bayes methods to produce a modified test statistic.
- The idea is similar to that employed by the SAM procedure, but is more sophisticated, and has more solid mathematical foundations.
- The goal is to modify the denominator of a standard t test statistic, by making large standard errors smaller, and small standard errors larger.
- This is known as shrinkage estimation.

Shrinkage estimation

- The underlying assumption is that the gene-specific variances follow a standard distribution (e.g., a gamma distribution) with some fixed parameters.
- This provides us which information about the underlying spread of the gene-specific variances.
- When we see extreme values from this distribution, we would like to moderate them, so that they don't have a major effect on our results (i.e., want to make large standard errors smaller, and small standard errors larger).
- To accomplish this, a weighted variance is calculated, based on the observed gene-specific variance, and the characteristics of the underlying distribution.
- This has the effect of pulling the extreme value towards the centre of the observed (empirical) distribution of gene-specific variances.

Why is it empirical Bayes?

- The procedure is considered Bayesian because by assuming an underlying distribution, we are effectively adding a priori knowledge to our problem by imposing a prior distribution on the gene-specific variances.
- This particular approach is empirical Bayes because it uses the data from the empirical (observed) distribution of gene-specific variances to estimate the parameters of the prior distribution.
- ASIDE (STATS students): if we specified the parameters using hyperpriors (additional prior distributions on the prior parameters) we would be in the hierarchical Bayes setting.

Back to limma

- Once limma has fit a linear model to the normalized data (using `lmFit`), a second function (`eBayes`) is used to calculate moderated t-statistics based on shrunken estimates of the per-gene variances.
- The moderated t-statistics can be quite different than the standard t-statistics, especially for small sample sizes.
- In general, the moderated t-statistics make it more likely that significant genes will have a large fold change, and a small variance, rather than a small fold-change and a tiny variance.

Determining differential expression

- Because of the mathematics underpinning the empirical Bayes approach (conjugacy for the STATS students), **the moderated t-statistics still follow a standard t-distribution** (unlike the SAM approach), with degrees of freedom based on both the number of observations for each gene, and the parameters of the underlying prior distribution.
- This allows the calculation of parametric p-values, to which standard multiple comparisons procedures can be applied.

Background: linear models

- Simple linear regression: $y = mx + b$
- Linear model equivalent: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
- In linear regression, x and y are continuous variables. Here we have y (gene expression) as continuous, but x (group) is discrete, so our linear model is actually equivalent to ANOVA (analysis of variance).
- For a single gene:
  - y_i are our gene expression values
  - x_i is the group (GFP or MYC) for the i^{th} sample
  - β_0 and β_1 are the intercept and slope coefficients
  - ϵ_i is the residual (or error) associated with obsevation y_i (the difference between our predicted, $\hat{y}_i$ and observed, y_i, values that cannot be explained by the model).

Background: linear algebra

- In practice, we represent our linear model in matrix form:

$$Y = X\beta + \epsilon$$

and use basic linear algebra to solve the equation and determine the value of the coefficients.

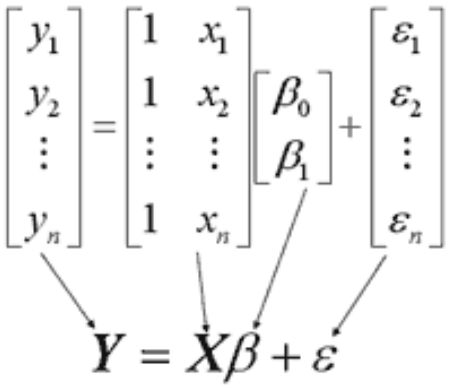

Image from: https://onlinecourses.science.psu.edu/stat501/node/382

Background: linear algebra

- The solution that minimises the "sums of squared error":

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

  is given by:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

- Why do we care?
  - Because `limma` requires the design matrix, X, to fit this model per gene and estimate its probability of differential expression.

The design matrix

- For our simple two-group differential expression analysis, the design matrix has two columns:
  - the first is all ones, and relates to the intercept coefficient: it is the average level of log-expression for the gene (remember the linear model is fit to each gene, so we have an intercept and a "slope" term per gene),
  - the second has zeroes for one group, and ones for the other, and relates to the coefficient for group ("slope"): it is the average difference in log-expression between the groups for that gene). This is what we are interested in.
- The residuals (the ϵ_i's) for each gene are used to determine whether the observed expression difference is statisically significant.