RNA-seq workshop

Differential expression analysis

Mik Black & Ngoni Faya
Genomics Aotearoa

Fold changes

- Differential expression (i.e., a change in gene activitation level) 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 ($log_2(2) = 1$) less than 1/2 ($log_2(\frac{1}{2}) = -1$) were considered to have undergc differential expression.

Detecting changes in expression

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

Determining differential expression

- We are investigating differences in gene expression between t strains of yeast (WT and MT)
- We have three replicates of the WT samples and three replicates 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 the normalization process, we conclude that any remain differences in intensity are result of differences in expression leve
- To test this, we can conduct a formal hypothesis test (for each ge to determine whether the mean intensity changed between the and MT samples.
- Since most basic statistical tests are set up to provide answers the additive scale, and fold changes are on the multiplicative sca we generally take logs of the data.

Hypothesis testing

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

Hypothesis testing

- In order to conduct the hypothesis test, we need a test statistic. T most simple approach is to utilize the test statistic of the standa 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, a $SE(\hat{\mu}_1 - \hat{\mu}_2)$ is some appropriate measure of variability (in t case the standard error).

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

P-values

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

P-values

- We have to decide how small a p-value needs to be for us to th that the difference we are observing cannot be explained solely noise.
- When we test a single hypothesis, it is common to fix a Type I er rate of α= 0.05 or α = 0.01.
- **Type I error:** reject null hypothesis when it is true (i.e., say a gen differentially expressed when it really isn't).
- **Type II error:** fail to reject the null hypothesis when it is false (i 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 make a Type I error in 5% of our hypothesis tests (i.e., 5% of t 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 exp 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 suita when testing large numbers of hypotheses.
- To get around this problem we use Multiple Comparis Procedures (MCPs).
- MCPs provide error rate control, allowing us to keep a lid on h many Type I errors we make.

Family-wise error rate control

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

What's so great about FWER control?

- Advantage: FWER controlling procedures provide a high level certainty in your result. The null hypotheses rejected by the procedures are very unlikely to be true (i.e., all of the rejected r hypotheses are likely to be correct rejections).
- Disadvantages: This level of control is very conservative - it is lik that some genes undergo differential expression, but their r hypotheses are not rejected. As the number of hypotheses be tested becomes very large, the significance threshold becom 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 a Hochberg (1995 - JRSS(B)).
- Provides a less conservative approach to error rate control th FWER controlling procedures.
- Greater power comes at the cost of an increased likelihood of Ty I errors.
- Has become very popular in genomic analysis, plus astronor brain imaging, and genetics (all test large numbers of hypotheses

Error rate control

- FWER control is concerned with making sure that the probability a single testing error is small.
- FDR control in concerned with keeping the proportion of Typ 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 th not adjusting at all, but are a lot more likely to make Type I err than FWER controlling procedures.
- The flip side is that FDR controlling methods are more likely reject false null hypotheses (i.e., they achieve greater power).

Comparing approaches

- Instead of adjusting the signifiance threshold, we can adjust t p-values themselves.
- The table below contains unadjusted p-values ("P-value"), a 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

Modification to t-test procedure

- One problem with the t-statistic approach to determin significance is that some genes with small, but consistent f changes can end up with very large t-statistics.
- This is especially common in genomics experiments involving onl few samples.
- Generally feel that genes with small fold changes shouldn't considered as having undergone significant differential expressio
- 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 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 derived by taking a Bayesian approach to analysis (and varic other ways).
- The Bayesian derivation relies on the assumption that the standa errors for each gene have an underlying common distribution. T s_0 parameter then contains information from this underly distribution.

Significance Analysis for Microarrays (SAM)

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

Significance Analysis for Microarrays (SAM)

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

Detecting differential expression with `limma`

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

Empirical Bayes analysis

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

Shrinkage estimation

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

Why is it empirical Bayes?

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

Back to limma

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

Determining differential expression

- Because of the mathematics underpinning the empirical Ba approach (conjugacy for the STATS students), **the moderat t-statistics still follow a standard t-distribution** (unlike the S approach), with degrees of freedom based on both the number observations for each gene, and the parameters of the underly prior distribution.
- This allows the calculation of parametric p-values, to wh 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 ha
  y (gene expression) as continuous, but x (group) is discrete, so
  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 (t
    difference between our predicted $\hat{y}_i$ and observed y_i val

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 determ the value of the coefficients.

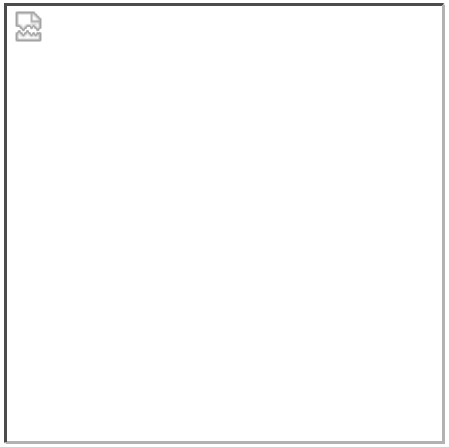

Image from: https://onlinecourses.science.psu.edu/stat501/node/382

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 mo
    per gene and estimate its probability of differential expressic

The design matrix

- For our simple two-group differential expression analysis, t design matrix has two columns:
  - the first is all ones, and relates to the intercept coefficient: i the average level of log-expression for the gene (remember t linear model is fit to each gene, so we have an intercept an "slope" term per gene),
  - the second has zeroes for one group, and ones for the oth and relates to the coefficient for group ("slope"): it is t average difference in log-expression between the groups that gene). This is what we are interested in.
- The residuals (the ϵ_i's) for each gene are used to determ whether the observed expression difference is statisic significant.