The Analysis of Gene Expression Data generated by Microarray and RNA-Seq technologies

Amanda Forde

Supervised by: Dr. Emma Holian

April 2020

NUI Galway
OÉ Gaillimh

Overview

1. Exploration of the multiple statistical methods commonly used in analysis of gene expression data.
2. Identify challenges which arise when analyzing this data
    - e.g. issue of multiple hypothesis tests.
3. Investigate differences between microarray and RNA-Seq data.
4. Experience with real world datasets:
    - `leukemiasEset` dataset and local mouse RNA-Seq dataset.
5. Use of `R` and its packages, in particular `limma` and `edgeR`.
6. Introduction to hierarchical clustering.

Differential Gene Expression Analysis

Gene Expression: process of making proteins from information stored in genes.

- Quantitative description of gene expression needed to investigate activity levels of a gene in particular cell → RNA transcript numbers determined with microarray or RNA-Seq technologies.

Differential Gene Expression Analysis: identifying subset of gene which are expressed to different extents in two contrasting conditions, e.g. treated vs untreated cells.

Statistician's role:

i) Assessment of magnitude and statistical significance of differential expression of each gene.
ii) Overcome issues arising from large size of datasets: many, many genes but very few samples.
iii) Other challenges: measurements greatly affected by noise.

Origin of Data

Microarrays: oligonucleotide ‘probe-set’ arrays

- Each gene represented by 16-20 pairs of perfect match (PM) and mismatch (MM) probes.
- RNA samples hybridized with array and images scanned → intensity values quantify how much hybridization has occurred for each probe.
- Intensities from probe pairs for each gene combined to form summary measurement: *continuous numerical measurements*.

RNA-Seq: next-generation sequencing technology

- Extracted RNA transformed into library of cDNA fragments.
- Sequencing progress produces short reads which are aligned to reference genome.
- No. of sequence reads mapped to each gene recorded: *integer count data*.
- Advantages of RNA-Seq include:
  i) lower background signal
  ii) ability to quantify wide range of expression levels

Introduction to `limma` and Linear Models

`limma`: R/Bioconductor package for differential expression analysis.

i) Differential expression evaluated with linear modelling approach.
ii) Use of empirical Bayes procedures which borrow information across genes to ensure stable inferences.

Linear models:

- Expression levels normally distributed, linear model for each gene:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- β_1 : mean difference in expression levels between two conditions.
- Hypothesis test: $H_0 : \beta_1 = 0$ vs $H_A : \beta_1 \neq 0 \rightarrow$ t-statistic & p-value.
- p-value $< \alpha \rightarrow$ gene differentially expressed.

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Multiple Testing

Multiplicity Problem: thousands of comparison tests → sharp increase in likelihood of obtaining false positives.

- Type I error: rejecting H_0 when it is in fact true.
- For m tests, probability of making *at least one* Type I error: $1 - (1 - \alpha)^m$.
- This grows rapidly → important to make corrections for multiple testing.

Family-Wise Error Rate (FWER): probability that at least one type I error committed:

$$\text{FWER} = P(V \geq 1)$$

Bonferroni Correction: p-values compared with $\frac{\alpha}{m}$.

- Weak control of experiment-level Type I error at level α.
- Over-conservative for large m → many missed findings.

False Discovery Rate (FDR)

False Discovery Rate (FDR): expected proportion of false positive errors among all rejected hypotheses:

$$\text{FDR} = E\left(\frac{V}{R} | R > 0\right) P(R > 0)$$

BH method (Benjamini & Hochberg (1995)):

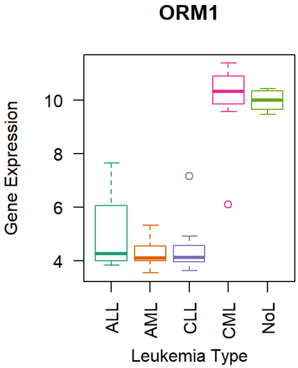
- p-values generated gene-by-gene from m hypothesis tests.
- p-values re-ordered from smallest to largest $\rightarrow p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(g)}$.
- $\text{FDR} = \delta$, test with highest rank, $H_0^{(k)}$, for which $p_{(i)} \leq \frac{i}{m}\delta$ sought out.
- All tests of rank $1, 2, \ldots, k$ are significant as $p_{(i)} \leq \frac{i}{m}\delta$ for $i = 1, 2, \ldots, k$.

Advantage: less conservative than other correction methods.

The `leukemiasEset` Dataset

- 60 samples from leukemia patients, 5 different groups: `AML`, `ALL`, `NoL`, `CLL`, `CML`.
- $20,172$ genes identified with Gene Ensembl IDs.
- Measurements obtained from high-density oligonucleotide arrays.
- RMA normalisation method removes technical variation from raw data and extracts log-intensity value for each gene and each sample:
  i) transformation
  ii) quantile normalisation
  iii) summarization

Expression measurements for a single gene as they vary across the different conditions.



Application of Linear Models with `limma`

Question of interest: determine genes which are differentially expressed between two groups, `NoL` and `AML`.

- Linear model applied to each gene: $E(y_g) = X\alpha_g$.
- β_g, constants of interest, extracted using $\beta_g = C^T\alpha_g$.
- X is *design* matrix and C is *contrast* matrix.
- Test null hypothesis $H_0 : \beta_{gj} = 0$ → obtain *t*-statistic:

$$t_{gj} = \frac{\hat{\beta}_{gj}}{s_g\sqrt{v_{gj}}}$$

Probability of differential expression detection can vary drastically from gene to gene → `limma` counteracts this with *variance shrinkage*.

Empirical Bayes Method in limma

limma implements empirical Bayes method, Smyth (2004):

- Hyperparameters of hierarchical model, d_0 and s_0^2, estimated.
- Gene-specific and global variability combined → *moderated* t-statistic:

$$\tilde{t}_{gj} = \frac{\hat{\beta}_g}{\tilde{s}_g \sqrt{v_g}}$$

$$\tilde{s}_g^{-2} = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g}$$

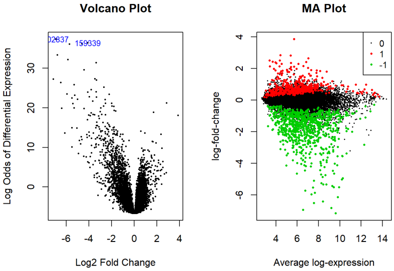
- Gene-wise residual sample variances *shrink* towards common value.
- $\tilde{t}_{gj}$ follows t-distribution with $d_g + d_0$ degrees of freedom → p-values.
- Posterior odds statistic and B-statistic can be reformulated in terms of *moderated* t-statistic.

Results of Microarray Data Analysis

- *p*-values adjusted for multiple testing using BH method.
- FDR controlled at level $\delta < 0.05 \rightarrow$ genes with adjusted *p*-values < 0.05 are significant.
- 3,291 genes differentially expressed
    - 1,518 genes downregulated, 1,773 upregulated.

Top 6 differentially expressed genes for contrast, `AML-NoL`, with summary statistics, shown in table.
Volcano plot and MA plot demonstrate obtained results.

```
            logFC  AveExpr         t      P.Value     adj.P.Val        B
OLFM4   -6.944266 6.547371 -14.74016 3.386891e-21 6.229650e-17 37.37184
PADI4   -4.128657 6.669315 -14.42716 8.993595e-21 6.229650e-17 36.45408
CHIT1   -4.642819 5.957630 -14.36226 1.102905e-20 6.229650e-17 36.26209
ORM1    -5.726453 6.749225 -14.32627 1.235306e-20 6.229650e-17 36.15537
CAMP    -6.818293 9.188380 -13.41648 2.291015e-19 9.242872e-16 33.39741
CHI3L1  -5.873373 7.642469 -13.03379 8.076630e-19 2.715363e-15 32.20243
```



Statistical Methodologies and RNA-Seq Data

Analyzing RNA-Seq data: `edgeR`, `DESeq`/`DESeq2` and `limma-voom`.

- Regression-based models estimate expression changes & hypothesis tests determine significance, but two groups:
  - i) modelling data directly with *negative-binomial* distribution: `edgeR` and `DESeq`/`DESeq2`
  - ii) normal-based approaches implemented upon count transformation: `limma-voom`

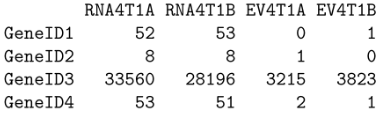
Use of count distributions and limitations:

- Information borrowing overcomes difficulty of parameter estimation
  → `edgeR`: gene-wise dispersion estimates.
- Less tractable: hypothesis tests only asymptotically valid.
- Large sample sizes: `DESeq2` acts conservatively while `edgeR` overly liberal.
- Statistical tools created to perform analyses on microarray data inaccessible to RNA-Seq data modelled with count distributions.

A local RNA-Seq Mouse Dataset

- Local RNA-Seq dataset in which samples extracted from murine cells.
- 4 samples and 1,908 genes but many genes have very low counts across all samples.
- Two conditions: `CellRNA` and `SecretedEV` → data originates from secreted extracellular vesicles and cells.

Count data generated for first 4 genes displayed below:



```
        RNA4T1A RNA4T1B EV4T1A EV4T1B
GeneID1      52      53      0      1
GeneID2       8       8      1      0
GeneID3   33560   28196   3215   3823
GeneID4      53      51      2      1
```

NATIONAL BREAST CANCER
RESEARCH INSTITUTE

i) Filtration and normalisation procedures performed first using `edgeR`.

ii) `limma-voom` implemented and normal linear models applied.

Filtration

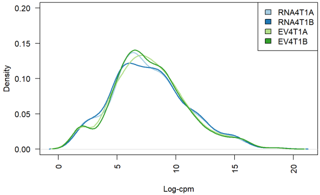
Filtration : first step of analysis in which genes with very low counts across all samples are removed.

Why?

- These genes cannot show evidence for differential expression, negatively impacting statistical approximations which must be made.
- They add to the multiple testing burden, reducing power to detect truly differentially expressed genes.

How?

Compare counts-per-million values, *cpm* as libraries sequenced to different depths & define a *threshold* → 834 genes removed.



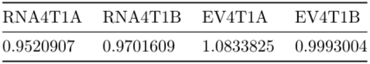
Normalisation of Read Counts

Normalisation ensures all samples have similar distributional patterns & eliminates systematic technical effects → possibility for skewed analyses minimised.

- With RNA-Seq data, no. of mapped reads, r_{gi}, influenced by composition of RNA population from which sample originates.

TMM normalisation:

- Eradicates composition biases across samples → evaluates scaling factors.
- `calcNormFactors()` function in `edgeR` executes TMM normalisation procedure.
- Factors altered so that they multiply to unity.
- TMM doesn't adjust raw data *but* effective library sizes will now be used.
- Effect of TMM-normalisation on this local dataset is very mild:



RNA4T1A	RNA4T1B	EV4T1A	EV4T1B
0.9520907	0.9701609	1.0833825	0.9993004

Linear Modelling and Count Data: `limma-voom`

Count data transformed into normally distributed data by means of *precision weights* - *voom*, Law et. al (2014).

Method:

1. Each observed read count, r_{gi} → log-*cpm* value using effective library size:

$$y_{gi} = \log_2 \left(\frac{r_{gi} + 0.5}{R_i + 1} \times 10^6 \right)$$

2. Linear models fitted to log-*cpm*s → residual standard deviations.
3. Mean-variance trend fitted to standard deviations as a function of average log-count.
4. Standard deviation of each log-*cpm* estimated as a function of fitted count value using mean-variance trend.
5. Squared inverse of estimated standard deviations → precision weights, w_{gi}.
6. Log-*cpm*s, y_{gi} and weights inputted into the `limma`'s linear modelling pipeline.

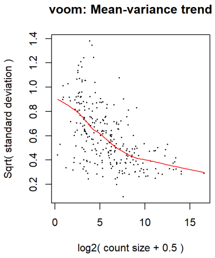
Linear Modelling and Count Data: `limma-voom`

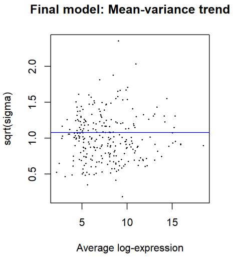
Mean-variance relationship:

- Variance much greater for lower log-*cpm* values → dependence of the variance on the mean:

$$\text{var}(y_{gi}) \approx \text{CV}^2(r_{gi}) = \text{CV}^2_{\text{tech}} + \text{CV}^2_{\text{bio}} = \frac{1}{\mu_g} + \phi_g$$

- Standard deviation of log-*cpm*s is a steadily decreasing function of mean count size, for small to medium counts.
- For greater count sizes, asymptotic behaviour expected at a level dependent on biological variability.





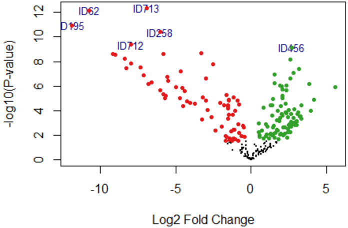
Results of RNA-Seq Data Analysis

164 genes show evidence of differential expression, 74 downregulated and 90 upregulated, contrast of interest: `Cell-SecretedEV`.

Summary statistics for top 4 differentially expressed genes, ordered according to p-values:

```
                logFC  AveExpr         t      P.Value    adj.P.Val        B
GeneID713  -6.912208 8.765715 -21.53126 4.472199e-13 8.748242e-11 19.87161
GeneID62  -10.708523 8.039360 -20.98119 6.627456e-13 8.748242e-11 18.15124
GeneID195 -11.887197 5.092344 -17.33674 1.177454e-11 1.036160e-09 14.99547
GeneID258  -6.012078 8.061566 -15.94870 4.082689e-11 2.694575e-09 15.69905
```

- Note: ranking with respect to B-statistics different → unscaled standard deviations, v_{gj} no longer constant.



Introduction to Hierarchical Clustering

Hierarchical Clustering: unsupervised method → discovery of co-regulated genes and identify related samples.

i) distance between two samples:
   - Use *correlation coefficient*, r, distance measure is $d = 1 - |r|$.
   - Exactly 0 when $r = \pm 1$ → perfectly correlated genes have no distance between them.

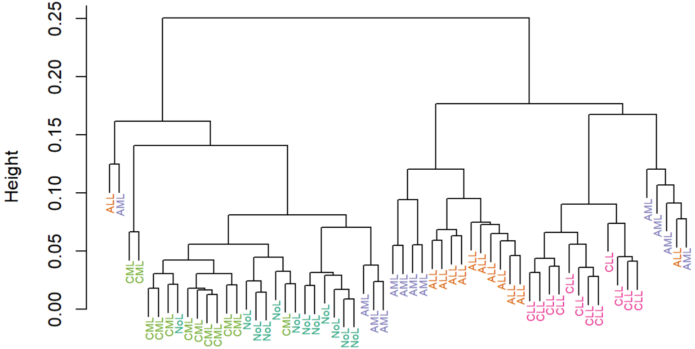
ii) inter-cluster distance:
   - *complete-linkage*: distance between two clusters of samples = largest distance between any two neighbours.

Construction of dendrogram using *agglomerative* approach:

1. Begin with n clusters, each consists of single sample.
2. At each time point, distances between clusters computed based on distance definition.
3. Two nearest clusters repeatedly merged into new single super-cluster until entire tree constructed.

Introduction to Hierarchical Clustering

Dendrogram obtained when samples from `leukemiasEset` *dataset clustered according to correlation coefficients and complete-linkage used to define inter-cluster distance.*

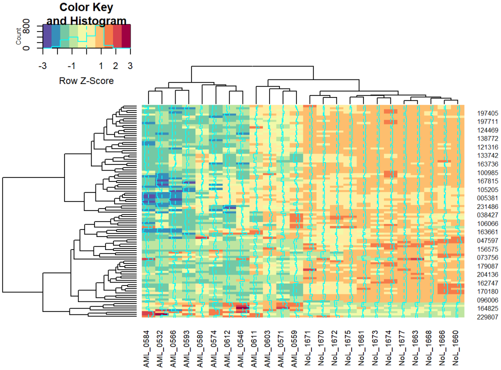


Heatmaps

Heatmap: popular form of visualisation of gene expression data.

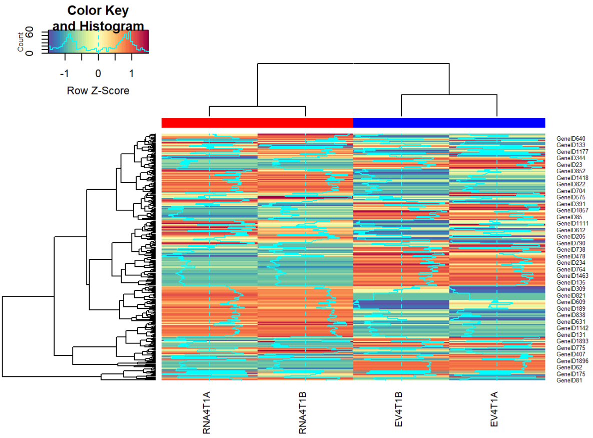
- Two separate dendrograms included - one for samples and one for genes.
- Different colours display degree of expression changes.

Heatmap produced using `NoL` *and* `AML` *samples and 100 most variable genes across samples from* `leukemiasEset` *dataset.*



Heatmaps

Heatmap produced using the genes retained after filtration and all four samples in local mouse RNA-Seq dataset.



Conclusion

1. Challenge of multiple testing and variance estimation when few samples available → BH method and empirical Bayes method.
2. Requirement of normalisation techniques: RMA and TMM normalisation.
3. Modelling mean-variance relationship with *voom* to make normal-based approaches accessible for count data (*precision weights*).
4. Implementation of methods on datasets and visualising results.

Further work:

- Investigate more modern methods which control false discoveries.
- Application of mean-variance modelling approach to other areas of research concerning tricky count data.
- Comparing results obtained when RNA-Seq data modelled with count distributions.
- Exploration of other forms of principal component analyses.