AMSI VACATION RESEARCH SCHOLARSHIPS 2019–20

EXPLORE THE MATHEMATICAL SCIENCES THIS SUMMER

Comparison of Pre-Processing Normalisation Techniques Used in the Bioinformatic Analyses of RNA-Seq Data

Bayan Khalili

Supervised by Dr Alice Johnstone

RMIT University

Vacation Research Scholarships are funded jointly by the Department of Education and the Australian Mathematical Sciences Institute.

Abstract

Gene expression data from high throughput sequencing can be biased by a number of technical effects caused by the experimental procedures themselves when measuring the abundance of transcripts. A number of algorithms have been developed to pre-process expression data in a crucial step called *normalisation*, which aims to reduce the amount of technical bias and allow analysis to focus on the biological factors present in the data. This report evaluates several normalisation techniques by describing how they work, the assumptions they depend on, and their impact on differential expression analysis. This also has implications for the construction of co-expression networks which aim to elucidate relationships amongst transcripts through their measured expression. According to the findings of this report, the choice of normalisation technique has an even stronger bearing on co-expression networks than it does on detecting differentially expressed genes.

Introduction

RNA-sequencing (RNA-seq) (Wang et al., 2009) is a methodology to capture the entire transcriptome (all RNA transcripts) from a biological sample. This data can be used to explore differences in gene expression (Van den Berge et al., 2019), and to build co-expression networks that relate genes to one another (Contreras-López et al., 2018). The raw RNA-seq data is a series of short sequences that need to be aligned to the genome. This currently requires processing on high performance computing systems to generate gene counts, which reflect the gene expression measured in the biological samples.

However, the gene counts produced are comprised of biological, technical and stochastic components (Abbas-Aghababazadeh et al., 2018; Bullard et al., 2010). The biological factors reflect the state of the organism and is the target of interest for research and analysis. We can mitigate the noise by increasing the number of biological replicates. But in order to study the biological factors, we must also control for any technical effects that are an artefact of the experimental procedures. Technical effects can bias the data and make it more difficult to accurately identify the biological factors of interest.

Technical effects can be further categorised into within-sample and between-sample effects. Within-sample effects have a variable impact on different genes within a given

sample. An example of this is the impact of gene length (Li et al., 2015). The number of reads for a given gene is approximately proportional to the length of a gene, so the ratio of genes within a sample is skewed in favour of genes with longer lengths. Within-sample effects are accounted for when processing the alignment of RNA-seq data, but they are not addressed in detail here. For the purposes of this report, the data is assumed to be free from such within-sample technical effects. Between-sample technical effects can have an impact on the counts for a given sample (Abbas-Aghababazadeh et al., 2018; Bullard et al., 2010). An important between-sample factor is the result of an amplification step applied to the biological samples prior to sequencing. To boost the signal, the transcripts are first converted from single RNA strands to cDNA (double-stranded DNA) and then undergo a process called Polymerase Chain Reaction (PCR) (Saiki et al., 1985). PCR repeatedly clones the cDNA fragments, theoretically increasing their abundance at an exponential rate. Since this process is governed by stochastic variables that are difficult to control for, the amplification phase often produces library sizes (total read counts) that differ by multiple orders of magnitude from one sample to another.

To control for the between-sample technical effects, several normalisation techniques have been developed to adjust for difference in sample library sizes to make them comparable (Anders and Huber, 2010; Evans et al., 2018; Robinson and Oshlack, 2010). We briefly summarise six normalisation techniques including their assumptions. The normalisation techniques are applied to an RNA-seq data set from chickpea plants *(Cicer arietinum)*, and a synthetic data set to evaluate their performance.

Statement of Authorship

My work was guided and developed under the direction of my academic supervisor, Dr Alice Johnstone. The chickpea data was acquired from Mayank Kaashyapp. Joel Robertson contributed an early version of the co-expression network code. I made extensive use of R libraries, especially those made available from Bioconductor. I developed all other code in R, including my own implementations of the *Total Counts*, *Median* and *Upper Quartile* normalisation techniques in R. Graphs were all generated using R, while the tables were formatted using Microsoft Excel

Normalisation Techniques

The normalisation techniques reviewed in this report are described briefly below. Each normalisation technique is designed to adjust the library sizes of different samples so that they can be compared within an experiment. The techniques selected differ in both their algorithms and their assumptions.

Total Counts Normalisation

Description

A very simple way to scale each sample is by dividing their gene counts with their library size (Evans et al., 2018). This transforms the counts into proportions per sample, which are unitless and directly comparable with one another. The implementation used for this report also multiplies gene counts with the average library size, which preserves the relative proportions, but converts them from fractions into counts that are closer in scale to the original raw values.

Assumptions

The total variation across samples is expected to be balanced. This means that an increase in the number of transcripts for some genes will be offset by an equivalent decrease in others. Each proportion is a function of all gene expression counts, which implies that the counts are not independent from one another.

Mathematical formulation

$$y_{gs} = \frac{r_{gs} \sum_{i=1}^{N} R_i}{R_s N} \quad (1)$$

Where y_{gs} is the normalised count for gene g in sample s, r_{gs} is the original raw count for gene g in sample s, N is the number of samples, and R_i is the library size for sample i.

Median Normalisation

Description

This method is identical to the *Total Counts* method, except that instead of scaling each sample by their library size, they are scaled according to their medians (Evans et al., 2018).

Assumptions

Most genes are expected to have a consistent expression level across samples, and the median expression level in particular is the same.

Mathematical formulation

$$y_{gs} = \frac{r_{gs} \sum_{i=1}^{N} M_i}{M_s N} \quad (2)$$

Where y_{gs} is the normalised count for gene g in sample s, r_{gs} is the original raw count for gene g in sample s, N is the number of samples, and M_i is the median expression count for sample i.

Upper Quartile Normalisation

Description

This method is identical to the median normalisation, except that it uses the upper quartile value to scale each sample (Evans et al., 2018). This is done because in many data sets, there are a large number of zero counts, which can potentially give a low or even zero value for the scaling factor.

Assumptions

Most genes are not differentially expressed between conditions, there are a lot of zero values for the expression counts, and the upper quartile is the same for all samples.

Mathematical formulation

$$y_{gs} = \frac{r_{gs} \sum_{i=1}^{N} U_i}{U_s N} \quad (3)$$

Where y_{gs} is the normalised count for gene g in sample s, r_{gs} is the original raw count for gene g in sample s, N is the number of samples, and U_i is the upper quartile of the expression count for sample i.

Quantile Normalisation

Description

Quantile normalisation equalises the distribution of each sample by ranking genes by their counts and replacing their values with the average of all counts at the same rank across samples (Evans et al., 2018). The implementation used here was from the *limma* R package from Bioconductor (Ritchie et al., 2015).

Assumptions

The assumption here is that the biological samples have identical distributions. This has the potential of either removing all differences between samples in some cases, and overexaggerating their differences in others.

Mathematical formulation

The genes are ranked according to their expression counts.

$$w_{rs} = \frac{\sum_{i=1}^{N} k_{ri}}{N} \quad (4)$$

Where w_{rs} is the normalised count for the gene at rank r in sample s, k_{ri} is the original raw count for the gene at rank r in sample i, and N is the number of samples.

Trimmed Mean of M-values (TMM)

Description

The trimmed mean of M-values (Robinson and Oshlack, 2010) is the default method provided by the *edgeR* R package from Bioconductor (Robinson et al., 2010; McCarthy, Chen and Smyth, 2012).

Assumptions

Over 50% of the genes are expected to have a stable expression level across samples.

Mathematical formulation

One of the samples (denoted as r) is selected as a reference sample. Let Y_{gk} be the observed count for gene g in sample k, and N_k be library size for sample k. We calculate the absolute intensity A_g^r and M-value M_{gk}^r for every gene g in each sample k:

$$A_g^r = \frac{1}{2} log_2\left(Y_{gk}/N_k \bullet Y_{gr}/N_r\right) \qquad M_{gk}^r = \frac{log_2\left({}^{Y_{gk}}/_{N_k}\right)}{log_2\left({}^{Y_{gr}}/_{N_r}\right)} \quad (5)$$

A subset of genes (G^*) are selected for each sample are selected by excluding those that have the highest and lowest absolute intensities and M-values. The $\log_2$ of the scaling factor $TMM_k^{(r)}$ for sample k is then calculated using:

$$log_2\left(TMM_k^{(r)}\right) = \frac{\sum_{g \in G^*} w_{gk}^r M_{gk}^r}{\sum_{g \in G^*} w_{gk}^r} \quad \text{where} \quad w_{gk}^r = \frac{N_k - Y_{gk}}{N_k Y_{gk}} + \frac{N_r - Y_{gr}}{N_r Y_{gr}} \quad (6)$$

Median Ratio Normalisation

Description

This method (Anders and Huber, 2010) finds the median of the ratios for all genes in a sample when compared with the geometric mean of the counts across samples. It is implemented in both the EBSeq (Leng and Kendziorski, 2019) and the Deseq2 (Love et al., 2014) R packages from Bioconductor.

Assumptions

Over 50% of the genes are expected to have a stable expression level across samples.

Mathematical formulation

Let m and n be the number of genes and samples respectively, and k_{ij} be the count for gene i in sample j. Each sample is normalised by dividing by a scaling factor s_j:

$$s_j = median_i \frac{k_{ij}}{(\prod_{v=1}^{m} k_{iv})^{1/m}}$$

Methods

The RNA-seq Data

Gene expression counts for two genotypes of chickpea plants (*Cicer arietinum*) (ICCV2 and JG11) were generously provided by Mayank Kaashyap. (Kaashyap et al., 2018) identified genes that were differentially expressed in species that showed tolerance and sensitivity to salinity. Each genotype had six replicates, half of which were in the control groups and half of which were in the treatments, as shown in Figure 1. When no normalisation techniques have been applied, differences in distribution across samples, treatments and species are visually apparent.

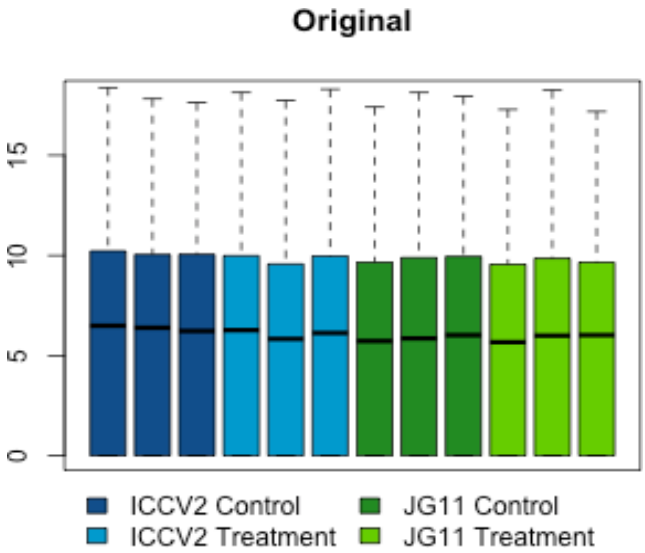

Figure 1 – Box plot of the raw expression counts from chickpea (C. arietinum) after applying a log2 transformation

In Figure 2, we have excluded genes with low expression counts (less than $2^6 = 64$) to reduce noise and applied a $\log_2$ transformation. The bulk of the data appears to follow a

normal distribution (overlaid in blue). However, the tail ends deviate from the Q-Q line as shown in Figure 3, indicating that the raw $\log_2$ transformed data does not follow a normal distribution. This lack of normality is also confirmed by an Anderson-Darling test (AD = 33223 with $p < 0.001$).

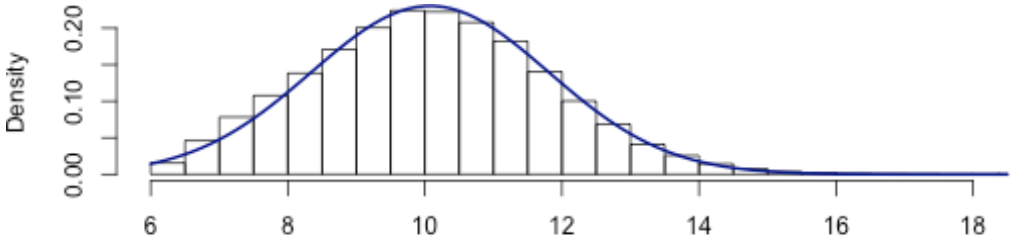

Figure 2 - Histogram of $\log_2$ transformed count distribution for chickpea RNA-seq data. A normal distribution is overlaid in blue.

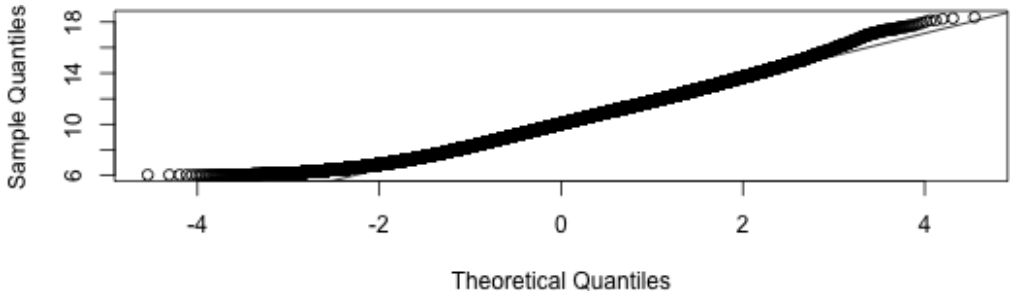

Figure 3 - Q-Q plot of $\log_2$ transformed counts for chickpea RNA-seq data

Synthetic Data

Synthetic data was generated using the ICCV 2 counts based on the techniques described in (Roca et al., 2017). The techniques used a normalised data set from which to generate new samples. New sample values were generated based on the mean and variance of the normalised data set. In each simulation, three control samples were created to mimic the original data, while differentially expressed genes for the three treatment samples were created by shifting the mean up or down by twice the variance. A total of four synthetic data sets were produced, with either 10%, 20%, 30% or 40% of their genes randomly selected to be differentially expressed. Each treatment had an equivalent number of up and down regulated genes to give the corresponding total percentages specified above. Scaling factors derived from the original data were also applied to each sample to simulate differences in library size.

Evaluation Methods

Two methods of analysis were employed for this report: differential expression detection (Van den Berge et al., 2019) and co-expression networks (Ballouz et al., 2015).

Differential Expression

Differentially expressed genes were detected using the instructions provided in (Contreras-López et al., 2018). Specifically, a gene is presumed to be differentially expressed if there is a $\log_2$ fold change of at least 1, with an FDR adjusted p-value less than 0.01.

Co-expression networks

RNA-seq data can be used to construct *co-expression networks*. A co-expression network is an undirected graph where the genes correspond to nodes, and relationships between genes correspond to edges. Co-expression networks are built from similarity matrices. In this report, the metric used for similarity was the Pearson's correlation coefficient. The constructed networks were all unweighted and were built from adjacency matrices. Edges were defined to exist whenever the corresponding magnitude of the correlation coefficient in the similarity matrix was greater than 0.9.

Results

Impact on Count Distributions

Figure 4 shows the resulting box plots after applying each of the between normalisation techniques analysed in this report. It is clear from the median and upper quartile plots that these normalisation methods ensure the 50^{th} and 75^{th} percentiles respectively match across samples. The quantile method takes this further by making the entire distribution of each sample identical. All methods shift the distributions to make the samples more comparable.

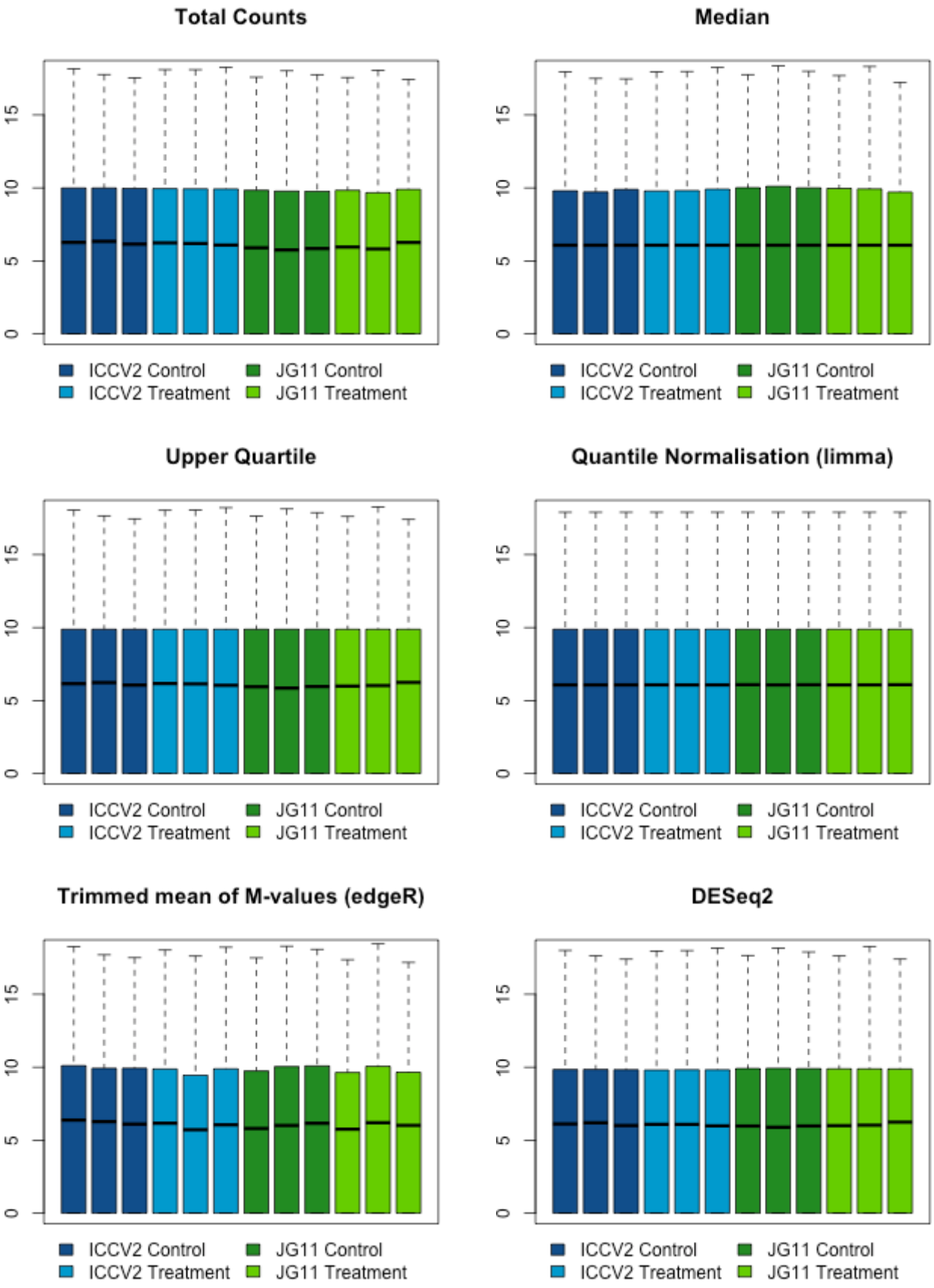

Figure 4 – Log_2 transformed box plots for chickpea data preprocessed with 6 different normalisation techniques.

Impact on Differential Expression Analysis

Table 1 summarises the differences in the number of differentially expressed genes found after each normalisation technique was applied to the chickpea count data. Each species was analysed separately with two conditions each: controls and salinated treatments. Deseq2 normalisation finds about 30% more differentially expressed genes in the ICCV2 species than the rest, while the quantile technique finds slightly less for the JG11 species. The rest of the results are very similar regardless of which technique is used.

	ICCV2			JG11		
	Up	Down	Total	Up	Down	Total
Total counts	59	41	100	71	17	88
Median	59	44	103	72	17	89
Upper quartile	59	43	102	70	17	87
Quantile	63	41	104	57	19	76
TMM	58	39	97	71	17	88
DESeq2	83	50	133	71	17	88

Table 1 - Number of differentially expressed genes identified by each normalisation technique

The synthetic data sets were constructed with *a priori* knowledge about which genes were differentially expressed to evaluate the impact on accurately identifying differentially expressed genes. Tables 2 to 4 summarise the accuracy, sensitivity and specificity when different proportions of genes were simulated to be differentially expressed in the treatments. Interestingly, increasing the percentage of differentially expressed genes degrades the performance of all techniques, including the original synthetic data, where no normalisation is needed.

	% Differentially Expressed			
	10%	20%	30%	40%
Total Count	96.9%	94.3%	91.6%	88.8%
Median	97.0%	94.5%	91.7%	88.9%
Upper Quartile	97.0%	94.5%	91.6%	88.7%
Quantile	91.0%	80.0%	71.3%	65.1%
TMM	97.1%	94.7%	91.9%	89.1%
Deseq2	97.0%	94.5%	91.7%	89.0%
Synthetic	97.0%	94.5%	91.8%	89.1%

Table 2 - Accuracy results for synthetic data

	% Differentially Expressed			
	10%	20%	30%	40%
Total Count	71.8%	73.1%	72.8%	72.5%
Median	72.5%	74.0%	73.2%	72.7%
Upper Quartile	72.4%	73.9%	73.0%	72.3%
Quantile	73.3%	76.7%	79.2%	80.7%
TMM	73.6%	74.8%	73.7%	73.2%
Deseq2	72.5%	74.1%	73.2%	73.0%
Synthetic	72.1%	73.8%	73.3%	73.4%

Table 3 - Sensitivity results for synthetic data

	% Differentially Expressed			
	10%	20%	30%	40%
Total Count	99.7%	99.6%	99.7%	99.6%
Median	99.7%	99.6%	99.6%	99.7%
Upper Quartile	99.7%	99.6%	99.6%	99.7%
Quantile	93.0%	80.9%	68.1%	55.3%
TMM	99.7%	99.6%	99.7%	99.7%
Deseq2	99.7%	99.6%	99.6%	99.7%
Synthetic	99.7%	99.6%	99.7%	99.6%

Table 4 - Specificity results for synthetic data

According to the accuracy and specificity results in tables 2 and 4 respectively, the quantile normalisation method performs significantly worse than all the other techniques. In fact, it performed significantly worse than applying no normalisation at all. The other methods performed much better overall, with good sensitivity, and very high accuracy and specificity values.

Impact on Co-expression Networks

Co-expression networks depend on variability in their data. This tends to be over-corrected with quantile normalisation, so it was excluded from the analysis in this section. To compare the agreement between the remaining five methods, edge lists for the co-expression networks for the chickpea data were created and are shown in the form of a Venn diagram in Figure 5. The values signify the number of correlated pairs that are in common with one another.

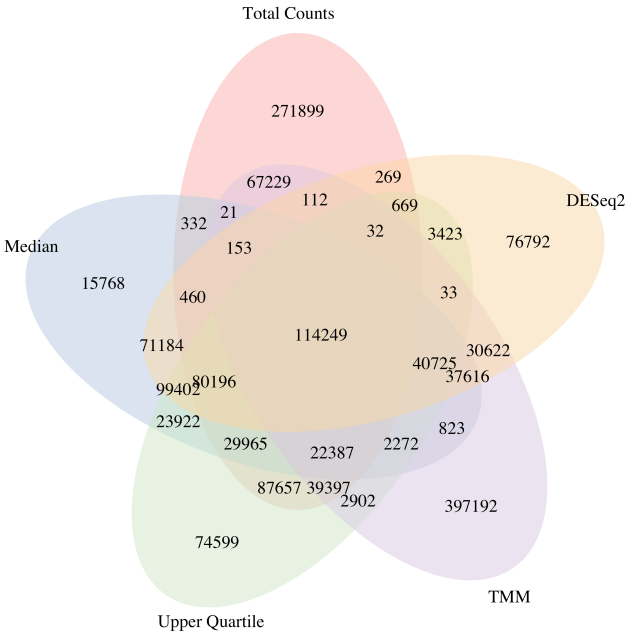

Figure 5 - Venn diagram of overlap in correlations

From a total of 1,592,302 correlations, only 114,249 of them are in common with all five normalisation techniques. This represents less than 7.2% of the total set of significant correlations. A further 836,250 correlations were found using a single normalisation technique. This includes almost 670,000 correlations that are only found using either TMM or Total Counts normalisation.

Discussion and Conclusion

Between-sample normalisation techniques alter the distribution of gene expression to make samples more comparable with one another. For the most part, the findings in this report indicated that the choice of normalisation did not make a significant difference with respect to differential expression analysis. The exception to this was quantile normalisation, which underperformed when compared with the five other methods evaluated in the synthetic simulations.

In contrast, co-expression networks were found to be highly dependent on which technique was used. Despite some overlap, most of the relationships that were identified were not common to all, or even most techniques. A potential cause for this diversity of results may be the way the adjacency matrix that feeds into the network was constructed. Perhaps the minimum cut-off value to identify relationships (Pearson's correlation coefficient of 0.9) was too high. Since the data is not normally distributed, perhaps a different similarity metric should be used, such as the Spearman's rank-order correlation coefficient. Or maybe further transformations need to be applied first to make the data comply with the normality requirements.

A key assumption in all methods evaluated here was the stability of expression counts for the majority of genes across conditions. This means that the techniques were not designed to cater for situations such as the global increase or decrease in expression levels. The procedures that used synthetic data for this report could be extended to test the normalisation techniques under these conditions, and also assess alternative normalisation techniques that do not have this assumption.

Although Synthetic data was generated as a means of evaluating differential expression, it is limited by the approximations and assumptions used to generate the data. Generating synthetic co-expression networks represents a much more challenging problem as it requires building a correlation matrix with predetermined values. This represents a significant area for future work and could be an interesting avenue for further research.

Acknowledgements

This project was supported by the AMSI Vacation Research Scholarship program.
I would like to thank the support and advice I received from Dr Jess Hopf and Associate Professor Stephen Davis. I would especially like to thank Joel Robertson for all of his guidance and assistance with this project, and for helping me get the resources and data I needed. Most of all, I have to give my biggest thanks to Dr Alice Johnstone, who was always supportive, helpful and kind; and for being so generous with her valuable time, spending countless hours directing me, discussing concepts, correcting my path, demanding evidence to back up my claims, keeping me honest and sceptical, and answering all of my questions.

References

Abbas-Aghababazadeh, F., Li, Q., Fridley, B.L., 2018. Comparison of normalization approaches for gene expression studies completed with high-throughput sequencing. PLoS ONE 13, e0206312. https://doi.org/10.1371/journal.pone.0206312

Anders, S., Huber, W., 2010. Differential expression analysis for sequence count data. Nature Precedings 1–1. https://doi.org/10.1038/npre.2010.4282.2

Ballouz, S., Verleyen, W., Gillis, J., 2015. Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. Bioinformatics 31, 2123–2130. https://doi.org/10.1093/bioinformatics/btv118

Bullard, J.H., Purdom, E., Hansen, K.D., Dudoit, S., 2010. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics 11, 94. https://doi.org/10.1186/1471-2105-11-94

Contreras-López, O., Moyano, T.C., Soto, D.C., Gutiérrez, R.A., 2018. Step-by-Step Construction of Gene Co-expression Networks from High-Throughput Arabidopsis RNA Sequencing Data. Methods in molecular biology 1761, 275–301. https://doi.org/10.1007/978-1-4939-7747-5_21

Evans, C., Hardin, J., Stoebel, D.M., 2018. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. Brief Bioinform 19, 776–792. https://doi.org/10.1093/bib/bbx008

Kaashyap, M., Ford, R., Kudapa, H., Jain, M., Edwards, D., Varshney, R., Mantri, N., 2018. Differential Regulation of Genes Involved in Root Morphogenesis and Cell Wall Modification is Associated with Salinity Tolerance in Chickpea. Scientific Reports 8, 1–19. https://doi.org/10.1038/s41598-018-23116-9

Leng N, Kendziorski C (2019). EBSeq: An R package for gene and isoform differential expression analysis of RNA-seq data. R package version 1.26.0.

Li, P., Piao, Y., Shon, H.S., Ryu, K., 2015. Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data. BMC Bioinformatics 16. https://doi.org/10.1186/s12859-015-0778-7

Love MI, Huber W, Anders S (2014). “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.” Genome Biology, 15, 550. doi: 10.1186/s13059-014-0550-8.

McCarthy DJ, Chen Y, Smyth GK (2012). “Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation.” Nucleic Acids Research, 40(10), 4288-4297. doi: 10.1093/nar/gks042.

Robinson MD, McCarthy DJ, Smyth GK (2010). “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.” Bioinformatics, 26(1), 139-140. doi: 10.1093/bioinformatics/btp616.

Robinson, M.D., Oshlack, A., 2010. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology 11, R25. https://doi.org/10.1186/gb-2010-11-3-r25

Roca, C.P., Gomes, S.I.L., Amorim, M.J.B., Scott-Fordsmand, J.J., 2017. Variation-preserving normalization unveils blind spots in gene expression profiling. Scientific Reports 7, 1–19. https://doi.org/10.1038/srep42460

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015). “limma powers differential expression analyses for RNA-sequencing and microarray studies.” Nucleic Acids Research, 43(7), e47. doi: 10.1093/nar/gkv007.

Saiki, R.K., Scharf, S., Faloona, F., Mullis, K.B., Horn, G.T., Erlich, H.A., Arnheim, N., 1985. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science 230, 1350–1354. https://doi.org/10.1126/science.2999980

Van den Berge, K., Hembach, K., Soneson, C., Tiberi, S., Clement, L., Love, M., Patro, R., Robinson, M., 2019. RNA Sequencing Data: Hitchhiker’s Guide to Expression Analysis. Annual Review of Biomedical Data Science 2. https://doi.org/10.1146/annurev-biodatasci-072018-021255

Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: A revolutionary tool for transcriptomics. Nature Reviews Genetics 10, 57–63. https://doi.org/10.1038/nrg2484