Preprocessing Methods

Justin Chumbley

Extended Methods Abstract

Sequencing: FASTQ sequencing data was transformed into counts using the STAR aligner[1]. ENSG gene identifiers were then mapped to a Gene Symbol[2] and counts for the same gene symbol were summed (i.e., summing over alternative transcripts/versions of the same gene). Note that in many cases multiple distinct ENSG identifiers may be ascribed to a single gene symbol, and in other cases, there may be no gene symbol to map to. In the latter case, the ENSG identifier is retained. **Filtering:** Starting from 59068 gene ids, we removed haemoglobin genes and genes with no HUGO ID, leaving 55772. We then removed genes with insufficiently large counts to be retained in a statistical analysis, resulting in 8484. The latter was performed by filterByExpr function in edgeR. The latter was performed in edgeR, using filterByExp, using a filtering strategy described by Chen & Smyth (2016). Roughly speaking, this strategy keeps genes that have at least 10 reads in a worthwhile number samples. **Normalizations**: We explored three normalization procedures (Evans, Hardin, and Stoebel 2017). Reference-gene normalized, log Transcripts Per Million (TPM). This aims to exploit expert biological knowledge to normalize, and involves dividing by the mean count over the 11 housekeeping genes identified by Eisenberg and Levanon (2013) (2013, Table 1): C1orf43, CHMP2A, EMC7, GPI, PSMB2, PSMB4, RAB7A, REEP5, SNRPD3, VCP, VPS29. It assumes these reference genes are invariant a priori over subjects, so seeks to equalize their value in the normalized data, adjusting all other genes accordingly. Trimmed Mean of M-values (TMM) normalization using the 'voom' method of the 'limma' package (Law et al 2014), which performs the mean-variance transformation and returns a new dataset with values in logCPM (log2 counts per million). It assumes most genes are not differentially expressed, and seeks to reduce the disproportionate effect of outlying genes. Quantile normalization, also implemented in 'voom' is a global adjustment method that assumes the statistical distribution of each sample is the same. This is perhaps the most invasive of the methods considered. **Batch effects:** as described in Table 1, our linear models adjusted for Assay Plate, Average Sample Profile Correlation, Pregnancy at wave 5, Any illness in the past 4 weeks, Any illness in the past 2 weeks, Smoker at the time of interview, Kit condition, Tube condition, Fasting hours, Travelling in the past one month outside the United States, Interview month and Interview hour. **Covariates:** gender, race, age. **Statistical methods:** All analyses were done in R version 3.6.1 (R Core Team 2019). Filtering, TMM and quantile normalization done using edgeR version 3.28.0 (Robinson, McCarthy, and Smyth 2010) and limma version 3.42.0 (Law et al. 2014). Reference gene normalization was performed using a simple in-house script.

Please also see supplementary material below.

[1] https://academic.oup.com/bioinformatics/article/29/1/15/272537
[2] http://www.ensembl.org/biomart/martview/

	Overall (N=1121)
Assay Plate	
Year1Plate01	88 (7.9%)
Year1Plate02	95 (8.5%)
Year1Plate03	93 (8.3%)
Year1Plate04	96 (8.6%)
Year1Plate05	94 (8.4%)
Year1Plate06	92 (8.2%)
Year1Plate07	91 (8.1%)
Year1Plate08	92 (8.2%)
Year1Plate09	95 (8.5%)
Year1Plate10	96 (8.6%)
Year1Plate11	96 (8.6%)
Year1Plate12	93 (8.3%)
AvgCorrelogram100 (Average Sample Profile Correlation)	
Mean (SD)	0.940 (0.023)
Range	0.800 - 0.970
Pregnancy at wave 5	
No	1107 (98.8%)
Yes	14 (1.2%)
Any illness in the past 4 weeks	
No	786 (70.1%)
Yes	335 (29.9%)
Any illness in the past 2 weeks	
No	859 (76.6%)
Yes	262 (23.4%)
Smoking at the time of interview	
No	887 (79.1%)
Yes	234 (20.9%)
Kit condition	
Normal	963 (85.9%)
Some Problems	158 (14.1%)
Tube condition	
Normal	1073 (95.7%)
Some Problems	48 (4.3%)
Fasting hours	
Mean (SD)	9.899 (5.499)
Range	0.000 - 29.400
Travelling in the past one month outside the United States	
No	1082 (96.5%)
Yes	39 (3.5%)
Interview month	
1	48 (4.3%)
2	0 (0.0%)
3	7 (0.6%)
4	14 (1.2%)
5	15 (1.3%)
6	31 (2.8%)
7	132 (11.8%)
8	261 (23.3%)
9	204 (18.2%)
10	182 (16.2%)
11	128 (11.4%)

1

Figure 1: Batch effects.

	mRNA_sample (N=1122)	Total wave 5 sample (N=3872)
Sex		
Male	445 (39.7%)	1611 (41.6%)
Femal	677 (60.3%)	2261 (58.4%)
Race/ethnicity (%)		
White (Nonhispanic)	742 (66.1%)	2390 (61.7%)
Black (Nonhispanic)	183 (16.3%)	648 (16.7%)
Asian	42 (3.7%)	219 (5.7%)
Other (Nonhispanic)	10 (0.9%)	48 (1.2%)
Hispanic	145 (12.9%)	567 (14.6%)
Age at wave 1 (yrs)		
N-Miss	2	12
Mean (SD)	15.500 (1.743)	15.602 (1.719)
Range	12.000 - 20.000	12.000 - 21.000
Birth Year		
1974	1 (0.1%)	3 (0.1%)
1975	4 (0.4%)	17 (0.4%)
1976	57 (5.1%)	216 (5.6%)
1977	184 (16.4%)	690 (17.8%)
1978	220 (19.6%)	743 (19.2%)
1979	215 (19.2%)	739 (19.1%)
1980	179 (16.0%)	616 (15.9%)
1981	148 (13.2%)	504 (13.0%)
1982	112 (10.0%)	341 (8.8%)
1983	2 (0.2%)	3 (0.1%)
Binge drinking in the past 12 months (> 4 (female) and 5 (males) drinks in a row)**		
N-Miss	2	10
No	574 (51.2%)	2108 (54.6%)
Yes	546 (48.8%)	1754 (45.4%)
Current smoking		
N-Miss	4	21
No	881 (78.8%)	3023 (78.5%)
Yes	237 (21.2%)	828 (21.5%)
Region of residence at wave 5		
N-Miss	19	58
Northeast	202 (18.3%)	845 (22.2%)
Midwest	354 (32.1%)	1053 (27.6%)
South	392 (35.5%)	1358 (35.6%)
West	155 (14.1%)	558 (14.6%)
Born preterm		
N-Miss	10	50
No	990 (89.0%)	3443 (90.1%)
Yes	122 (11.0%)	379 (9.9%)

1

Figure 2: Covariates.

Supplementary information

Normalization

1. Reference gene normalization

The total "forward read" counts were transformed to the reference-gene adjusted, log Transcripts Per Million (TPM) metric used in linear model RNAseq statistical analyses as follows. Let $x_{i,j}$ be the count for any (non-housekeeping) gene i in subject j, then

$y_{i,j} = log_2(1 \vee 10^6 \times \frac{x_{i,j}}{\bar{h}_j}))$

Here $\vee$ is the max operator and the reference-gene adjustment for subject j, denoted $\bar{h}_j$, is defined to be the mean count over the 11 housekeeping genes identified by Eisenberg and Levanon (2013) (Table 1): C1orf43, CHMP2A, EMC7, GPI, PSMB2, PSMB4, RAB7A, REEP5, SNRPD3, VCP, VPS29.

In this case the TPM values are floored at 1, so any TPM value below 1 copy per million is raised up to 1 per million. Once log2 transformed, those count values go to 0. Note that this squashes the low end of the distribution a bit, but that turns out to be empirically accurate because values below 1 per million are not reliable enough to represent to downstream analyses as meaningful variation. And this tends to reduce total variance and thereby generally yield slightly conservative estimates of the magnitude of differential expression if some samples have values in that low range.

Some authors have used 5 counts per million as the lower cut-off, which involves even more flooring/squashing. However empirical studies suggest that one can find reliable values below 5/million when using our particular assay and sequencing depth - which usually yields around 10 million total reads, with each read corresponding to a unique RNA transcript, so 1 per million comes out to around 10 absolute sequence counts, and replicate variability on the same sample is usually around 5 counts or less, so >10 is a reasonably reliable number.

Reference gene normalization fits into one of the three classes of normalization descerned by Evan's et al (2017). It has the advantage of exploiting prior biological information to determine non-DE genes - rather than attempting to algorithmically identify non-DE genes from the data itself as is common to many normalization pipelines.

2. TMM normalization

Trimmed Mean of M-values (TMM) normalization is based on the hypothesis that most genes are not differentially expressed (DE).

A trimmed mean is the average after removing the upper and lower x% of the data. The TMM can eliminate effect of few genes which have very high counts and which would dominate the library size calculation and have an effect on simple normalization by using only library size. using the 'voom' method of the 'limma' package, it performs the mean-variance transformation and returns a new dataset with values in logCPM (log2 counts per million).

```
dge <- DGEList(counts = counts) # creates a DGEList object from edgeR using raw counts data

dge <- calcNormFactors(dge, method="TMM") # calculates normalization factors using TMM

v <- voom(dge, design = design_matrix) # applies voom transformation to count data
```

3. Quantile normalization

Quantile normalization is a global adjustment method that assumes the statistical distribution of each sample is the same. Normalization is achieved by forcing the observed distributions to be the same for each sample by replacing each quantile with the average of that quantile across all samples.

```
v <- voom(counts, design = ..., plot = TRUE, normalize = "quantile")
```

Batch effects and covariates

References

Eisenberg, E., & Levanon, E. Y. (2013). Human housekeeping genes, revisited. TRENDS in Genetics, 29(10), 569-574.

Evans, C., Hardin, J., & Stoebel, D. M. (2017). Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. Briefings in bioinformatics, 19(5), 776-792.

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome biology, 15(2), R29.

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.

Socratic dialogue

Michael: in the normalisation procedure with house-keeping genes, did you use the raw counts or the library size log transformed count?

Steve: We library-size normalize first, and then carry out an additional round of reference gene normalization. However, I don't believe there's really a difference between the two, because a/b -> b/c = a/c (or in log2 metric, a-b -> b-c = a-c) In other words, the first library size normalization is moot because it cancels out after the subsequent reference gene normalization. So you should be able to normalize to reference genes directly, and get quite similar results (barring any minor rounding difference due to the initial data flooring occurring in the reference gene-normalized metric rather than the initial library size-normalized metric)

Cecilia: I have a question about how to carry out consistency checks after normalisation. We are in the process of constructing internal consistency benchmarks to validate other normalisation procedures. Could you offer some guidelines on how to conduct these checks after any normalisation has taken place?

Steve: We generally just check which normalization method produces the clearest (most significant) findings for several effects that are generally known to be true based on previous research, including 1 X and Y-linked genes should be differentially expressed as a function of subject-reported sex 2 inflammatory genes should be upregulated in proportion to BMI (and possibly also smoking, although this is a more subtle and variable effect) 3 interferon-related gene expression should be upregulated in people of African ancestry relative to Caucasians The word document "UCLA RNAseq QC summary - v1.docx" is a QC document recently prepared for Add Health data administrators that provides a little more information about this (but not too much...) at the end of the document, in the endpoint QC section In my analyses of the Add Health data last summer, all 3 normalizations passed check 1 above (which is not uncommon because the effect size is so big, and so this check is generally not useful for deciding among different normalizations), whereas only the reference gene normalization yielded positive results for checks 2 and 3.

References

Eisenberg, Eli, and Erez Y Levanon. 2013. "Human Housekeeping Genes, Revisited." *TRENDS in Genetics* 29 (10). Elsevier: 569–74.

Evans, Ciaran, Johanna Hardin, and Daniel M Stoebel. 2017. "Selecting Between-Sample Rna-Seq Normalization Methods from the Perspective of Their Assumptions." *Briefings in Bioinformatics* 19 (5). Oxford University Press: 776–92.

Law, Charity W, Yunshun Chen, Wei Shi, and Gordon K Smyth. 2014. "Voom: Precision Weights Unlock Linear Model Analysis Tools for Rna-Seq Read Counts." *Genome Biology* 15 (2). BioMed Central: R29.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson, Mark D, Davis J McCarthy, and Gordon K Smyth. 2010. "EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data." *Bioinformatics* 26 (1). Oxford University Press: 139–40.