Bioinformatics 301

Combining laser-assisted microdissection (LAM) and RNA-seq allows to perform a comprehensive transcriptomic analysis of epidermal cells of Arabidopsis embryo

Review by Hadleigh Waldegrave

Intro — Why lasers and ultra-low RNA-seq?

"Genome-wide characterisation of tissue- or cell-specific gene expression is a recurrent bottleneck in biology" because small samples are difficult to isolate and RNA yields are often too small for RNA-seq analysis.[1] Statistical tools and vibrating micro-blades have been used to help prevent neighbouring tissues from contaminating datasets, but they're relatively ineffective.[2] Laser-Assisted Microdissection (LAM) is quite self-explanatory. It involves using a thin laser beam to cut out specific cells under a microscope in preparation for RNA-seq. In this study, the researchers used plant embryos from *Arabidopsis thaliana* to test how effective LAM is when distinguishing the transcriptomes of epidermal and mesophyll cells within the embryos. My review investigates the researcher's claim that as little as 0.02mm^2 of tissue and 0.05ng of RNA can be used to effectively analyse differential gene expression.

Methods

The researcher's dataset was analysed for quality using FastQC v0.11.7[3] as per Babraham Bioinformatics' methodology.[4] 3 out of the 16 available fastq files[5] were randomly selected for quality analysis ('Epidermal 1 Forward', 'Epidermal 3 Forward' and 'Mesophyll 1 Reverse'). Although all 16 fastq files should be checked for quality, the consistent quality across our 3 random analyses warranted further investigation of differential gene expression using RStudio Desktop (open source licence AGPL v3).[6] The packages limma, edgeR, gplots, RColorBrewer and Glimma were used to filter out lowly expressed genes, check the data for quality, normalise the data based on size differences between the libraries, and finally test for genes that were differentially expressed between the epidermal and mesophyll cells. Like the researchers, a p-value of <0.05 was used as the threshold after adjusting for the false discovery rate.

Results

To investigate how much RNA was needed for RNA-seq, the researchers used 5ng, 0.100ng, 0.075ng, 0.050ng, 0.025ng, and 0.010ng (all from the same biological sample). I started by investigating the researchers' claims that only 0.050ng of RNA *"can be used without compromising the number of genes detected*". This wasn't straight forward because the number of genes detected in the 5ng sample was up to 33% lower than the trend line calculated from the other 5 samples (Figure 1). I respectfully disagreed with the researchers, opting to use the 0.100ng sample as the reference instead of the anomalous 5ng sample. However, relative to the 0.100ng sample, the number of genes detected in the 0.050ng sample still only dropped by 4%, keeping the researcher's claim valid.

I found that a low CPM threshold of >0.3 corresponds to counts of 10-15 for the library sizes (Figure 2A). I had initially opted to use a CPM threshold of >0.5 but the researchers used >1.0, so I investigated all three thresholds. The CPM threshold of 1.0 corresponds to counts of 35-40, but this makes very little difference to the differential gene expression analysis (Figure 2B, 5). For each gene, I had initially required all 4 biological replicates to meet the CPM threshold in order for it to be considered as differentially expressed. However, the researchers allowed a gene to be considered as differentially expressed even if only half of the biological replicates met the CPM threshold, so I tested for both. The researchers found a total of 870 DEGs whereas I found a maximum of 312 (Figure 2B).

The AtML1 and AtPDF2 genes were chosen by the researchers to validate their DGE analysis. I have compiled their data in Figure 3 which indeed verifies their expression levels in epidermal and mesophyll cells. I also used limma::topTable in R to curate a list of genes that are most likely to be significantly differentially expressed. I researched a selection of these genes and found that most of their functions are indeed relevant to epidermal cells in Arabidopsis embryos (Figure 4).

Discussion

The researchers took some merit away from their claim when they opted to collect 0.100ng samples for their DGE analysis. Why spend twice as long collecting 0.100ng of RNA while claiming that 0.050ng is enough "*to perform a comprehensive analysis of the expressed genome*"? The researchers also claim that their results support using LAM for DGE analysis of other plant tissues, but they did not address how total RNA yield varies between cells of different types and sizes.[7]

Because using LAM for ultra-low RNA-seq is a relatively new method, there is not much relevant data in the literature, however, Illumina's user guide suggests RNA samples as low as 0.010ng can be used if the RNA quality is high enough.[8] The Arabidopsis eFP Browser has data for expression in the cotyledons of torpedo-stage embryos, but it is not yet specific enough to distinguish expression in the epidermal cells from the mesophyll cells. Not only is LAM being used more for cell-specific RNA-seq, it is also being used to isolate samples from plants that are hard to transform,[15] and to accurately automate tumour sampling for personalised medicine.[16]

For the differential expression analysis, I used a bayesian approach to adjust p-values to control the false discovery rate, whereas the researchers used a Benjamini–Hochberg approach. *"In typical QTL or microarray experiments, where m is large and some rejections typically occur, the difference between [Bayesian] FDR and [Benjamini–Hochberg] FDR is usually very small"*[17] so this can not explain the large difference between the researcher's 870 DEGs and my 312 DEGs. By applying a logFC > 1 threshold, the researchers' number of DEGs is limited to 571. However, I have not yet found a reason that explains this near 2-fold difference, and I welcome feedback.

As referenced before, Figure 4 shows that most of the DEGs I researched encode proteins whose functions are relevant to epidermal differentiation, germination, seed dormancy, cuticle wax biosynthesis, anthocyanin biosynthesis, trichome formation and more. As such, this paper has shown that sequencing just 0.100ng of RNA from laser-dissected samples is an effective method for exploring differential expression at a microscopic level. The researchers spent a lot of effort exploring the cutin and anthocyanin biosynthesis pathways. However, given climate change's increasing threat on agricultural industries,[18] it is a shame that the researchers did not investigate the seed dormancy and trehalose genes which have shown potential to increase crop yields during drought.[19]

Finally, an obvious limitation to this study, like most RNA-seq studies, are the small number of replicates used. Having only 4 replicates limits the number of significant DEGs that can be found. However, 2 comparisons with 4 biological replicates each is relatively generous given the current costs of Next Generation Sequencing (NGS). As the cost per nucleotide sequenced continues to decrease,[20] it will be interesting to see how physical limits affect server storage growth and encourage data compression in coming years.[21]

Figures

Figure 1 — The relationships between RNA sample size, library size and genes detected

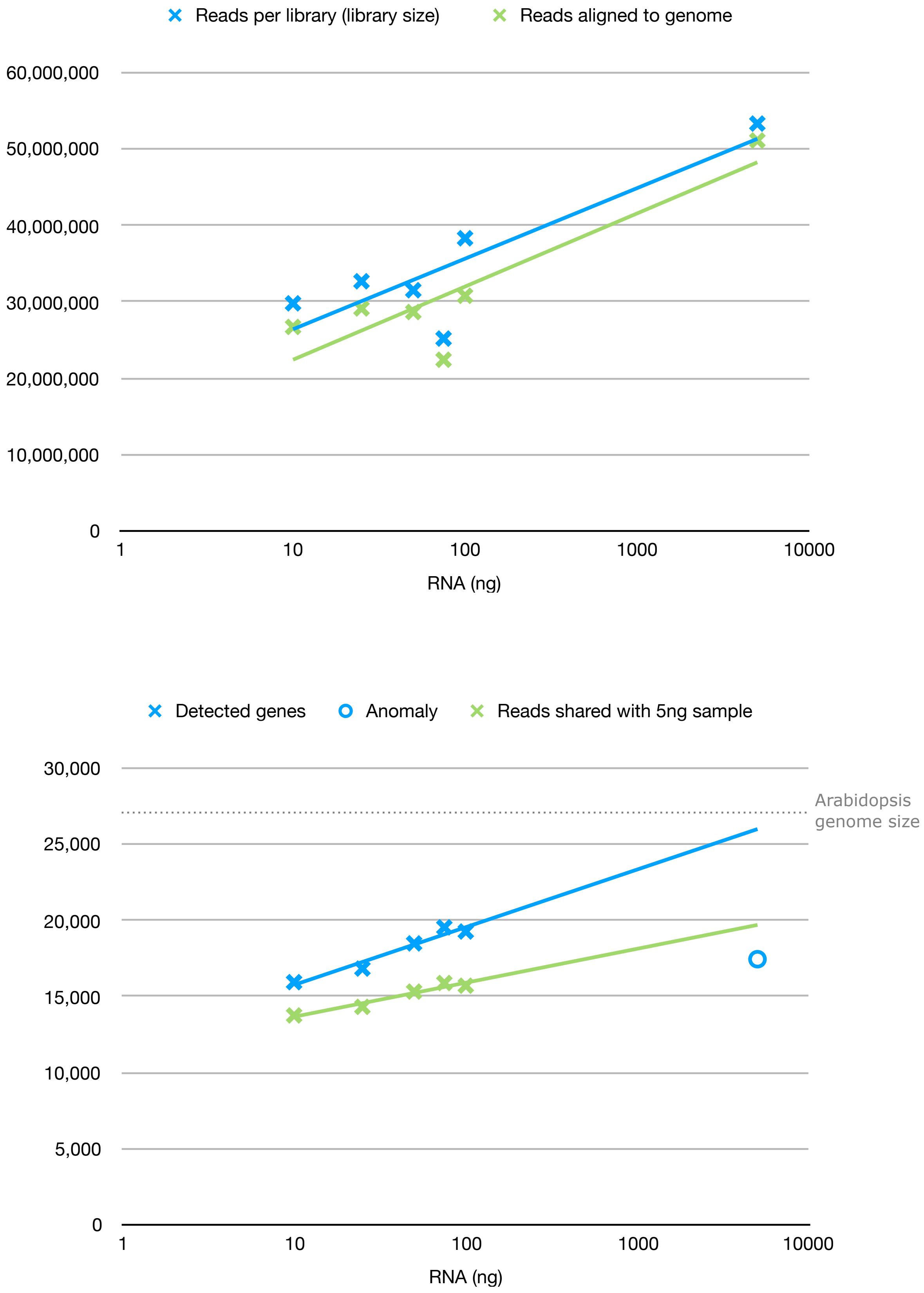

Figure 2A — Plot of CPM thresholds and corresponding read counts

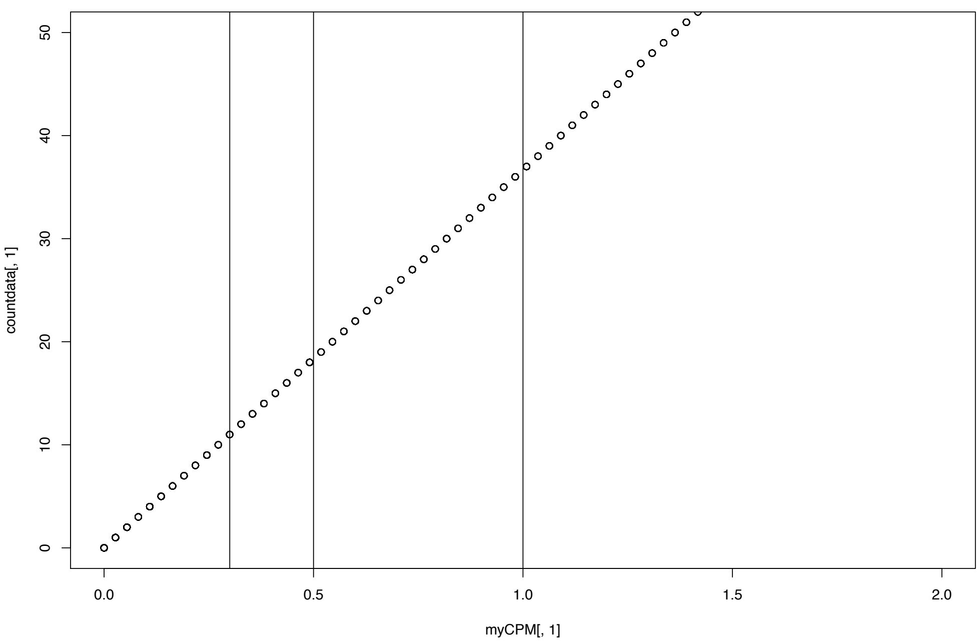

Figure 2B — Differentially expressed genes found at varying CPM thresholds across all 4 biological replicates, versus in at least 2 out of 4 biological replicates

CPM across 4 of 4 replicates	No Significant Difference	Up	Down	Total DEGs
0.3	16439	121	84	205
0.5	16137	121	84	205
1.0	15598	114	82	196

CPM in ≥2 of 4 replicates	No Significant Difference	Up	Down	Total DEGs
0.3	17817	201	111	312
0.5	17492	170	97	267
1.0	16943	144	94	238

Figure 3 — Expression levels of validation genes from study vs from Arabidopsis eFP Browser*

Gene	logFC	Adjusted p-value	Mean CPM Epidermal	Mean CPM Mesophyll	CPM Range*
AtML1	3.132	0.015	132	25	17–27
AtPDF2	3.142	0.003	231	31	29–31

Figure 4 — Selection of differentially expressed genes for functional analysis

Gene	logFC	Adjusted p-value	Function	Relevance
TPS2	10.52	0.0433	Dehydration tolerance, Prevent ice formation	Needed for seed dormancy[11]
MFH8.2	10.13	0.0433	Oxidoreductase for ROS reduction	Removes ROS from germination, Dormancy signalling[12]
SVB	3.03	0.0003	Controls length and number of trichomes	Controls epidermal cell differentiation into trichomes[13]
SBT	2.97	0.0003	Serine protease for proteolysis	Mobilising precursor proteins for seed maturation[14]

Figure 5 — R Script for DGE analysis

```
library(edgeR)
library(limma)
library(Glimma)
library(gplots)
library(RColorBrewer)
seqdata <- read.csv("GSM2588913_counts.csv", stringsAsFactors = FALSE)
sampleinfo <- read.delim("SampleInfo.txt")
countdata <- seqdata[,-(1:1)]
rownames(countdata) <- seqdata[,1]
myCPM <- cpm(countdata)
thresh3 <- myCPM > 1
keep3 <- rowSums(thresh3) >= 2      # I initially used >=4
counts.keep3 <- countdata[keep3,]
y3 <- DGEList(counts.keep3)
logcounts3 <- cpm(y3,log=TRUE)
group <- paste(sampleinfo$CellType)
group <- factor(group)
y3 <- calcNormFactors(y3)
design <- model.matrix(~ 0 + group)
colnames(design) <- levels(group)
v3 <- voom(y3,design)
fit3 <- lmFit(v3)
cont.matrix <- makeContrasts(EpidermalVsMesophyll=epiderm -
mesophyll,levels=design)
fit3.cont <- contrasts.fit(fit3, cont.matrix)
fit3.cont <- eBayes(fit3.cont)
limma.res3 <-
topTable(fit3.cont,coef="EpidermalVsMesophyll",sort.by="p",n="Inf")
write.csv(limma.res3,file=“EpidermalVsMesophyll-3.csv",row.names=TRUE)
```

References

1. Sakai, K., Taconnat, L., Borrega, N., Yansouni, J., Brunaud, V., Paysant-Le Roux, C., Delannoy, E., Martin Magniette, M., Lepiniec, L., Faure, J., Balzergue, S. and Dubreucq, B. (2018). Combining laser-assisted microdissection (LAM) and RNA-seq allows to perform a comprehensive transcriptomic analysis of epidermal cells of Arabidopsis embryo. *Plant Methods*, 14(1).

2. Jung, J. & Jung, H. Methods to analyze cell type-specific gene expression profiles from heterogeneous cell populations. *Animal Cells and Systems* 20, 113-117 (2016).

3. FastQC A Quality Control tool for High Throughput Sequence Data. *Bioinformatics.babraham.ac.uk* (2018). at <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

4. Babraham Bioinformatics. Using FastQC to check the quality of high throughput sequence. *YouTube* (2018). at <https://www.youtube.com/watch?v=bz93ReOv87Y>

5. DRA Search - SRP105190. *Trace.ddbj.nig.ac.jp* (2018). at <https://trace.ddbj.nig.ac.jp/DRASearch/study?acc=SRP105190>

6. Take control of your R code. *RStudio* (2018). at <https://www.rstudio.com/products/rstudio/#Desktop>

7. Brennecke, P. et al. Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods* 10, 1093-1095 (2013).

8. Low-Input & Single-Cell RNA-Seq | Single-cell sequencing benefits. *Illumina.com*(2018). at <https://www.illumina.com/techniques/sequencing/rna-sequencing/ultra-low-input-single-cell-rna-seq.html>

9. Abe, M. Regulation of shoot epidermal cell differentiation by a pair of homeodomain proteins in Arabidopsis. *Development* 130, 635-643 (2003).

10. Sessions, A., Weigel, D. & Yanofsky, M. The Arabidopsis thaliana MERISTEM LAYER 1 promoter specifies epidermal expression in meristems and young primordia. *The Plant Journal* 20, 259-263 (1999).

11. Fait, A. et al. Arabidopsis Seed Development and Germination Is Associated with Temporally Distinct Metabolic Switches. *PLANT PHYSIOLOGY* 142, 839-854 (2006).

12. El-Maarouf-Bouteau, H. & Bailly, C. Oxidative signaling in seed germination and dormancy. *Plant Signaling & Behavior* 3, 175-182 (2008).

13. Marks, M., Wenger, J., Gilding, E., Jilk, R. & Dixon, R. Transcriptome Analysis of Arabidopsis Wild-Type and gl3–sst sim Trichomes Identifies Four Additional Genes Required for Trichome Development. *Molecular Plant* 2, 803-822 (2009).

14. Srivastava, R., Liu, J. & Howell, S. Proteolytic processing of a precursor protein for a growth-promoting peptide by a subtilisin serine protease in Arabidopsis. *The Plant Journal* 56, 219-227 (2008).

15. Wuest, S. & Grossniklaus, U. Laser-Assisted Microdissection Applied to Floral Tissues. *Methods in Molecular Biology* 329-344 (2013). doi:10.1007/978-1-4614-9408-9_19

16. Zakowski, M. & Bibbo, M. Lung Carcinoma in the Era of Personalized Medicine: Th e Role of Cytology. *Karger.com* (2018). at <https://www.karger.com/Article/PDF/356235>

17. Bogdan, M., Ghosh, J. & Tokdar, S. A comparison of the Benjamini-Hochberg procedure with some Bayesian rules for multiple testing. *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen* 211-230 (2008). doi:10.1214/193940307000000158

18. Williams, B. et al. Trehalose Accumulation Triggers Autophagy during Plant Desiccation. *PLOS Genetics* 11, e1005705 (2015).

19. Mathewson, S. Australian 'Resurrection Plants' Shed Light On Ways To Create Drought-Tolerant Crops. *Nature World News* (2015). at <https://www.natureworldnews.com/articles/18602/20151207/australian-resurrection-plants-shed-light-ways-create-drought-tolerant-crops.htm>

20. DNA Sequencing Costs: Data. *National Human Genome Research Institute (NHGRI)* (2018). at <https://www.genome.gov/27541954/dna-sequencing-costs-data/>

21. Muir, P. et al. The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biology* 17, (2016).