**[RNA Quest: Rethinking RNA-Seq reports through reanalysis of previously published neuronal transcriptomes](http://localhost:4000/showcase/rna-quest-report/)**

by Ryan Pevey

**An interesting problem in Omics research**

I’ve noticed something when I work on big omics projects and I know that I’m not the first. Typically there are two teams of people on any given project, or more if it’s a multi-omics project. The bioinformaticists that are performing the statistical analysis, generating output results tables and figure creation (so many figures). As well as the biologists, who may have generated the biological samples that were processed, and will typically handle the downstream interpretation of the omics results and place them in a biological context. Often, but not always, there is a discrepancy between the background and training of these groups where the bioinformaticists tend to come from computer science, statistics or applied math fields. Of course many of them have coursework and training in biology, this isn’t an attack on bioinformaticists or the good work that they do, and it’s becoming more common that they have explicit training in biology or even a degree in biostatics or bioinformatics as these programs proliferate in academics. But few of them have much experience doing wet bench science like their collaborators. Also, many of them work on many projects within many biological models and contexts so they don’t always get the chance to dive deep into a single area of biology.

The opposite is normally true of the biologists on the project. Most of us are generalists in bench science with a single or a small handful of specialty bench skills, and a deep understanding of a typically very narrow field of biology. Few of us have any programming skills to speak of, nor training in how to understand the results of big data projects. Again, this isn’t meant to be an attack on biologists, just stating the reality of the differences in viewpoint and experience between the two groups of people.

This discrepancy in background leads to a second discrepancy. Where some of the people have deep intimate access to the data, and some of the people have a deep intimate knowledge of the biology underlying the data, and those typically are not the same people. This is the very nature of team based projects, but still, a gap in knowledge between groups can lead to a gap in communication between groups. This can have outsized impacts on the outcome of the project.

To put it another way, in large-scale Omics projects, a persistent challenge lies in bridging the communication gap between bioinformaticists who analyze the data and biologists who interpret it. While statistical outputs and processed data files contain a wealth of insight, they often lack the intuitive structure or clarity needed for effective biological interpretation. As a result, important findings may be overlooked or misinterpreted, underscoring the need for more accessible and integrative reporting strategies. I don’t think that this is an easy problem either or caused by a lack of effort. All of the people involved are highly intelligent and highly trained. I think this is just a function of the results being large and complex; and everybody has limited time and attention to give to it. Omics results are a firehose, and the person on the receiving end of it is normally only equipped with a bucket to catch as much water as they can.

**It's about the exploration, not just the destination**

Some people, however, do have skills and expertise in both worlds. I sit at this nexus, with training and experience in bench science and bioinformatics. My degrees are in Biology and Neuroscience and I have over a decade of experience at the bench and the computer terminal (mainly RNA-Seq and scRNA-Seq). Those of us with these dual skills sets are in a relatively unique position where we can act as a liaison between the teams. We can discuss the minutiae of software choices and sharing analysis scripts with the one team, or the particulars of reagent selection and sharing bench protocols with the other team.

I’m proposing a pretty simple solution here for a hard problem, which is a report structure for Omics results that really facilitates the ability of biologists to explore the dataset in an interactive way. The report includes the core results figures and tables that typically are provided to collaborators, but in a single interactive web portal that acts as a clearinghouse for the results of the analysis. It’s the same results, but formatted in a way that really draws in the biologist for a more independent search of the dataset.

To that end I’ve created the below report on an analysis of RNA-Seq data that I performed at home as an example of the kind of reporting strategy that I think could facilitate this kind of exploration. I want to crack open the black box of omics methodologies for biologists so that you can see what is happening behind the scenes when you send samples off to the core for sequencing, or when you send an email to the bioinformaticists to ask for expression plots of a list of genes of interest. Now, they can just… look it up themselves. I believe this approach increases transparency and trust among collaborators, while increasing the biologists ability to engage with the data on a deeper level than is typically possible. Also, given the didactic focus of the forum here, I can afford to be more casual than in a traditional publication. So forgive me the informal tone of a lot of this page.

If you’re a biologist, hopefully you come out of this exercise with a better appreciation for the work that the programmers do and a better understanding of how they produce their datasets and results files. There is a common misunderstanding in biology that “the computer does all the work”. Don’t be that person. This is a toy example with a small sample size and it still took me weeks of work. Hopefully this report highlights just how much effort goes into these kinds of analyses and the format presents a more intuitive way to understand those results.

If you are not a biologist, I think there is a relatively straightforward track that you can take to help bridge this gap that I’ve highlighted. I’m generally in favor of transparency in science, but I think the more you can facilitate your biologist collaborators to engage with the data and results that you’ve created the better. The standard output of results, like excel files and static figures showing top hits or general trends are great, I’m not saying you should stop providing a zip file of results. But I’ve seen many biologists excitedly receive their data from the core, only to be stuck in fear asking the question, “How am I supposed to make any conclusions from a 20,000 row excel file?”. Spending the time to make an engaging and visually appealing clearinghouse of results will also help them explore the data in a way that an excel file will just never provide, hopefully they even find it fun! Not to mention the polish and professional flair this kind of effort can lend to your output. Maybe this project can provide a better framework for reporting results to collaborators, I have a standalone html file ([here](https://github.com/rpevey/rnaQuest/blob/main/bin/rnaQuest_DESeq.Rmd)) that could be provided to a biologist as an entry point to understanding the results, and just bundle it alongside the standard results files that you normally provide. That also provides privacy for unpublished data as nothing is hosted online, people can just download the html file locally and open it in the browser on their computer. If you decide to make the portal public later as part of the publication process, you’ve now just made a fancy web portal for anyone to explore your data with.

Maybe, you’re new to RNA-Seq analysis and you’ve come across this post as you’re doing your first analysis. If that’s the case, I hope I can convince you early in your journey to develop good communication skills. As a data scientist, the results files and reports that you produce is your work product and the better you are at presenting them the more your collaborators will be able to mine from them. Omics datasets are a rich resource, but often they are not fully mined for data purely because of this gap in communication or limited time and resources. That’s a shame in my mind, because the amount of effort that goes into producing any one of these datasets can be enormous and there are over half of a million RNA-Seq datasets publicly available on NCBI resources ([Zeimann, et al. 2019](https://doi.org/10.1093/gigascience/giz022)). The potential of this as a supply of data is immense and it grows every single day. Many groups always do a good job of all of this, but I think as a field we can do better on both sides. And everyone benefits from it.

**Summary of dataset**

This webpage reports the differential expression results of my re-analysis for this really nice RNA-Seq dataset from the neurons of patients with epilepsy ([Pai et al., 2022](https://doi.org/10.1186/s40478-022-01453-1)). The original publication is mostly concerned with transcriptional alterations in glial cells but I’m electing to reanalyze their neuronal samples for a few reasons. First, to highlight the usefulness of previously published data. Second, so that I am not just retreading the same story over again. Third, and most selfishly, my experience lies mainly with neuronal datasets so I’m most familiar with this output. No offense to glia, they are great, but I’m really mostly focused on the reporting of results here!

The dataset consists of bulk RNA-Seq reads from three temporal lobe control post-mortem samples (TL), and three temporal lobe with epilepsy patient samples (TLE). In the original paper, four populations of cells were isolated from the samples via Fluorescence Activated Nuclei Sorting (FANS), one of those populations was the neurons and those are the samples that I’m working with here.

**Differential expression analysis**

Briefly, I started this project by downloading the fastq files for each sample from NCBI’s, Gene Expression Omnibus ([GEO accession: GSE140393](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE140393)), then aligned the reads to the reference genome and produced a set of counts files which contain the number of reads counted for each genetic feature in the reference annotation (ensemble: [GRCh38 release 113](https://ensembl.org/Homo_sapiens/Info/Index)). Those counts files are the input files for this differential expression analysis. Which continues with loading all of the proper software and data files into R for processing with DESeq2.

All of the scripts for my full analysis pipeline, including downloading and preprocessing of the data before the differential expression analysis, are on my GitHub repository for this project here: [RNA Quest at GitHub](https://github.com/rpevey/rnaQuest). If you’re not a programmer and you’ve ever been curious about the steps that go into the differential expression analysis, then click on the buttons labelled ‘Click to expand code’ and it will reveal the code that goes into each step. Just remember that this script is only the very last stage of the analysis. I’ve also included extensive comments in the code so hopefully you can understand something of it even if you have no programming experience. Note, the control is the reference group, that’s hopefully obvious but I want to be explicit about it. All statistical analysis results are epileptic neurons referenced in comparison to the neurons of control patients.

Temporal lobe neurons of Epilepsy patients (TLE) vs. Temporal lobe neurons of Control post-mortem tissue (ref.) Differential Expression Analysis Results

Alright, the meat of the report, the results. Keep in mind that there is a fair amount of exploratory analysis and quality control that goes into the analysis at this stage to ensure that the data is processed properly (e.g. normalization, regularization). Those steps were performed here but are presented further down the page so that we can jump right to the exciting part: the results.

There are 1576 significant differentially expressed genes (FDR < 0.05) when comparing Epilepsy to Control as the reference. 758 are up-regulated, 818 are down-regulated. The results of the DE analysis are summarized in the table below. Overall, roughly the same number of genes(~3%) are down-regulated compared to up-regulated genes.

Summary of results

## out of 26557 with nonzero total read count

## adjusted p-value < 0.05

## LFC > 0 (up) : 758, 2.9%

## LFC < 0 (down) : 818, 3.1%

## outliers [1] : 36, 0.14%

## low counts [2] : 7209, 27%

## (mean count < 22)

**Gene expression dashboard**

I’ve made this interactive gene expression dashboard below to help you explore the dataset. Mouse over the genes on the volcano plot to explore the highlighted significant genes. If you want to see the expression profile for that specific gene then enter it, or your favorite gene, into the gene expression plot on the right. If the gene is statistically significant, than an FDR value and bracket indicating significance are automatically populated. If not then the words “No DE result” will appear instead. If you want to restrict the gene expression plot to only display significant genes than toggle the checkbox beneath the plot. You can see the data used to create the plot with the table beneath it or download it by clicking the download graph data button to capture it as a csv.

I have the gene expression plot hosted by a [Render](https://render.com/) server then piped into this page through an iframe. So if the plot isn’t loading correctly it has probably just gone inactive and you can find it here (<https://geneexprdash.onrender.com/>). Actually, if you visit there then first or just reload this page, it should reactivate and load correctly.

Gene expression dashboard

**Heatmap**

Heatmaps provide a visual representation of gene expression across multiple samples, using color gradients to indicate expression levels. Clustering patterns can reveal relationships between genes and conditions, helping to identify co-regulated genes and distinct expression profiles. The most down-regulated genes in epilepsy patient samples all cluster in the top half of this heatmap together, these genes correspond to the left side of the volcano plot above. Conversely, the most up-regulated genes in the neurons of epilepsy patients cluster to the bottom of the heatmap and show up on the right side of the volcano plot.

**Up-regulated and Down-regulated genes: Epilepsy vs. Control (ref.)**

Heatmap

**Results table: Epilepsy vs. Control (ref.)**

**Table 1:** These are the results of the differential expression analysis when comparing, *Epilepsy* and *Control* (ref.) neurons. The columns are ordered by log2FoldChange with up-regulated genes at the top and down-regulated genes at the bottom. Numeric values listed to 3 significant digits. An FDR is not calculated for non-significant results. You can search for a gene of interest by Ensembl ID or gene name using the search box at the top.

Results table

**Gene expression plots**

These are the exact same type of plot as the right side of the dashboard above, except these are static instead of interactive. Gene expression plots visualize how a gene’s expression varies across conditions or samples, helping identify trends, variability, and potential differential expression. Higher expression levels indicate greater transcript abundance, while statistical comparisons (e.g., boxplots, dot plots) help assess significance between groups.

Let’s take a moment to discuss interpretting these figures. It’s important to interpret these boxplots carefully, especially with the small sample size here. The boxplot includes an in-built outlier test where any point that is more than 1.5x the interquartile range (IQR, the top and bottom of the box) above or below the median value will be plotted as an individual datapoint outside of the range of the “whiskers”. Generally, it would be preferrable to have a larger sample size than 3 for each group, but sometimes you have to work with what samples you have access to and this post is for educational purposes remember.

**Interesting genes: Epilepsy vs. Control (ref.)**

GOI panel, find one more down-reg gene

Most down-regulated gene: *ARL6IP1*. This gene has known functions in ER and Mitochondrial organelle homeostasis ([Foronda et. al., 2023](https://doi.org/10.1038/s41586-023-06090-9)), and specifically in neurons it is a modulator of glutamate transport ([Akiduki, 2008](https://pubmed.ncbi.nlm.nih.gov/18684713/)). As the most down-regulated gene in the dataset is such an important gene for excitatory neuron function, it’s safe to say that these neurons are not doing great.

Gene of interest down-regulated: *CALB2*. The *CALB2* gene (Calbindin 2), encoding the calcium-binding protein Calretinin, plays critical roles in neuronal excitability ([Camp and Wijesinghe, 2009](https://doi.org/10.1016/j.biocel.2009.05.007)). Abundant in hippocampus, amygdala, and cerebral cortex, with modified expression linked to temporal lobe epilepsy ([Dixit et. al., 2015](https://doi.org/10.1093/cercor/bht254)).

Gene of interest up-regulated: *LGR6*. Is a Wnt-signaling associated stem cell marker with links to cancer ([Huang et. al., 2017](https://doi.org/10.1038/ng.3957)). Generally it is not associated with expression in neurons and has very low expression in our control neurons, but sees a huge increase in epilepsy patient neurons in this dataset.

Most up-regulated gene: *MIR376C*. Is a microRNA involved in post-transcriptional gene regulation normally associated with roles in development ([Fu et. al., 2013](https://doi.org/10.1161/hypertensionaha.111.203489)) and homeostasis through interacting with SOX6-mediated Wnt signalling ([Cao et. al., 2021](https://doi.org/10.2147/IJGM.S278538)). So, we’re already seeing Wnt signalling showing up as a potential dysregulated pathway.

**Discussion**

Congrats programmers, you can now hand your report off to collaborators! I’m not going to do a full in-depth review of the results here. The point of this post is to highlight how this interactive report format can facilitate a deeper more intuitive exploration of the dataset than scrolling static excel files. Having said that! We can see some trends that jump right out at us. SNORA genes, are small non-coding nucleolar RNAs that are generally associated with RNA regulation ([Huang et. al., 2022](https://doi.org/10.1038/s41420-022-01056-8)) and cellular stress response ([Cheng et. al., 2024](https://doi.org/10.1016/j.cell.2024.06.019)). Indicating a big increase in transcriptomic dysregulation and cellular stress in neurons. Our Wnt signaling associated genes are a pretty interesting hit. A cursory submission of the results into [GOrilla](https://cbl-gorilla.cs.technion.ac.il) (results not presented here) reveal similarities to the enrichment analysis results in the original paper. Up-regulated results: Semaphorin-Plexin signaling pathway, gene expression regulation, lipid metabolism pathways. My analysis seems to show a bigger inclusion of RNA metabolism than their analysis did. Down-regulated results: Protein transport, cellular secretion, GABAergic synaptic transmission, axon formation. What is clear, is that there are a lot of results to dig through here. Happy hunting!

**Methods**

I am performing a reanalysis of previously published data ([Pai et al., 2022](https://doi.org/10.1186/s40478-022-01453-1)). Their study, briefly, collected temporal lobe tissue from patients with medically refractory epilepsy and control donors, preparing samples for immunofluorescence, single-cell dissociation, or Fluorescence-Activated Nuclei Sorting (FANS). Using an enhanced FANS protocol, they isolated nuclei from NEUN+, OLIG2+, and PAX6+ populations and preserved RNA from sorted nuclei for downstream analysis. Total nuclear RNA was extracted, depleted of rRNA, and used to generate libraries for paired-end sequencing on an Illumina HiSeq 2500 (50 bp pair-end sequencing, 38–50 million paired-end reads/sample). They uploaded their data as fastq files to NCBI’s Gene Expression Omnibus ([GEO accession: GSE140393](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE140393)), which is were I accessed the data on 3rd, March 2025. Sequencing reads were aligned to the human genome (ensemble: [GRCh38 release 113](https://ensembl.org/Homo_sapiens/Info/Index)) using STAR with default settings, and quantified gene-level expression using STAR’s ‘--quantMode GeneCounts’ option. Sorted Bam files were produced with Samtools.

**Scripts**

This is the R script that I used for this analysis, it has been ported into this report file. The following code is not executed, but is presented for transparency and reproducibility. Again, this is only the script that I used for the differential expression analysis. If you want all of the scripts for the full analysis including the downloading and preprocessing of the data before the differential expression, you can find it on my github repository for this project here: [RNA Quest at GitHub](https://github.com/rpevey/rnaQuest).

**Exploratory data analysis**

Exploratory Data Analysis (EDA) is the process of visually and statistically examining data to identify patterns, trends, and potential issues before formal analysis. It helps ensure data quality, detect outliers, and guide analytical decisions. If you aren’t doing the analyses yourself, you might normally not even be presented these figures. But they’re important sanity checks to ensure data quality. These steps were performed before the results presented above, but I’m presenting them here for the sake of completeness.

**All samples total gene counts**

Gene counts across all samples are shown both before and after normalization. The first plot illustrates that raw counts are already within a comparable range across samples. Normalization further refines the data, ensuring consistency for downstream analysis, as seen in the second plot.

Gene normalization

**Plot dispersion estimates**

A diagnostic plot of how well the data fits the model created by DESeq2 for differential expression testing. Good fit of the data to the model produces a scatter of points (black and blue) around the fitted curve (red). Low expressing genes, with counts below 10 reads across all samples, have been filtered out for computational efficiency.

Dispersion est. plot

**Dimensional reduction analyses**

In a way, a lot of omics analysis in really a game of dimensional reduction. You start off with a huge dataset where you have gene expression data for roughly 20,000 genes, each of them it’s own dimension. But some of the analysis techniques, such as clustering, and especially visualizations must be done on a lower number of dimensions, a small handful or maybe even just two.

In comes Principal Component Analysis (PCA) and Multidimensional Scaling (MDS) plots, they serve similar functions to each other and are interpretted in roughly the same way for our purposes here. PCA and MDS are dimensionality reduction techniques used in RNA-Seq analysis to visualize patterns of similarity or difference between samples. They transform high-dimensional gene expression data into a smaller number of dimensions, typically two. Thereby making it easier to detect clustering by condition, batch effects, or outliers. These plots provide an intuitive overview of global transcriptional variation across the dataset.

They differ in how they reduce dimensions. Critically, MDS is plotted in a way such that distances between samples are preserved as much as possible in the projection to lower dimensions. This is similar to the difference between T-SNE plots and UMAP plots, where the distance between clusters is more meaningful in UMAPs while it lacks interpretability in T-SNE plots. However, since MDS is based on preserving pair-wise distance you don’t natively get a measure of how much ‘information’ each dimension carries. That’s in contrast to the measure of variance explained by each dimension that you get out of PCA plots. In that way they both have benefits and are a good compliment to each other.

PCA and MDS plots

**Up-regulated vs. Down-regulated genes**

We can see here that there are roughly the same number of up-regulated as down regulated genes as was noted in the results summary and could be seen in the volcano plot. 758 are up-regulated, 818 are down-regulated.

Up-regulated and down-regulated genes

**Gene Biotype treemap for all genes**

The treemap below shows the relative proportion of each gene biotype for all genes in the dataset. The area of the rectangle is proportional to the percentage of each biotype. The area of each of the rectangles is proportional to the percentage of each biotype when comparing **Epilepsy** and **Control** (ref.). The majority are protein coding which is normal, but protein coding genes make up a significantly higher proportion of differentially expressed genes (right figure).

Treemaps

***X*2 Goodness of fit table**

Treemap table

There is a statistically significant difference in the proportion of gene types when comparing the full dataset to the subset of differentially expressed genes (*X*2 = 117.371, df = 3, *p*-value = 2.842e-25).

**Session info**

Session Info provides details about the computational environment used for this analysis, including the versions of R, loaded packages, and system settings. This ensures that the analysis can be reproduced accurately in the future, even if software updates change how certain functions behave.