Analysing differences in gene expression related to pre-menstrual dysphoric disorder (PMDD) using a mouse model

Eva Burguete-Innocente

Assignment 4

BINF 6210

November 27th, 2025

PMDD affects about 5% of women

- It is characterised by core symptoms:
  - Extreme mood swings, sudden sadness/tearfulness
  - Increased anger and irritability
  - Extremely depressed mood, negative self-image, suicidal thoughts
  - Extreme anxiety and feeling on-edge
- Additional symptoms include:
  - Decreased interest, poor concentration, lethargy
  - Appetite and sleep changes
  - Feeling overwhelmed, out of control
  - Physical symptoms: breast pain, joint pain, bloating
- Symptoms begin 7-10 days before menstruation, and end at the start of menstruation
- **PMDD is both a depressive and gynaecological condition that significantly impacts functioning and quality of life**

(Cary and Simpson, 2023)

PMDD is an “abnormal brain response to normal menstrual hormonal fluctuations.”

- Exact causes are somewhat understudied but may include:
  - Genetic heritability: a SNP in the *Esr1* gene confers hormone sensitivity (Huo et al., 2007).
  - Decreased availability of serotonin receptors during the late luteal phase (Jovanovic et al., 2006).
  - Allopregnanolone modulation of GABA-A receptors producing an unexpected effect (Bäckström et al., 2014).
  - Environmental factors like stress and inflammation (Raffi and Freeman, 2017).
  - Brain-derived neurotrophic factor polymorphisms (Hantsoo and Epperson, 2015).
- This is an ongoing area of research.

Marrocco et al. (2018) evaluated the intersection of estradiol response and BDNF Val66Met genotype

- Estradiol is a neurosteroid that fluctuates during the menstrual cycle.
- *Val66Met* is a SNP in the brain-derived neurotropic factor gene at position 66, causing a change from valine to methionine.
  - This has been associated with negative behavioural traits in reaction to ovarian hormone fluctuations.
- They administered either estradiol (E2) or vehicle in drinking water to mice that were either WT and or had the BDNF *Val66Met* SNP (Het-Met).
- They conducted behavioural tests on the mice, and found that E2 causes anxiety and depression behaviours in Het-Met mice, but not WT.
- RNA sequencing revealed gene expression differences between WT and Het-Met mice in response to E2.
- They compared these results to a human dataset and found similarities.

(Marrocco et al., 2018)

Objective: Analyse the mouse RNA-seq data from Marrocco et al. (2018) to determine differences in gene expression between genotypes and treatments, while exploring the effects of different methodological choices in the analysis pipeline

I will attain this by…

- Downloading the publicly available dataset from NCBI GEO (accession: GSE121412).
- Using the skills I learned in class and following the Bioconductor vignette by Law et al. (2018) to perform RNA-seq analysis of differential gene expression.
  - These methods differ from Marrocco et al.: they used DEseq2 but did not provide detailed or reproducible methods.
- Explore how varying methodological choices affect the data as per Tong et al., 2020:
  - I decided to look at normalisation with trimmed mean of M values (TMM) vs. upper quartile vs. no normalisation.
- Link to my github: https://github.com/evainnocente/PMDD_RNA-seq_analyses.git

Description of dataset

- This dataset is available on NCBI GEO under the accession GSE121412.
- It is the raw count values of gene expression levels generated by whole-genome RNA-sequencing of the mouse ventral hippocampus.
  - Data was extracted by decapitation after six weeks of treatment.
- It contains ten samples: four WT mice (of which two received no treatment (vehicle), and two received E2), and six Het-Met mice (of which three received vehicle and three received E2).
- Gene expression analysis was performed at the exon level.
  - There are multiple exons for each transcript for each gene in the dataset.
- Sample names in the dataset begin with either VV (WT) or VM (Het-Met) followed by an underscore, then Veh (treated with vehicle) or E2 (estradiol), followed by a number (or period followed by number) from 1-3 (the number of samples that belong to each level).

(Marrocco et al., 2018)

Methods

1. Data preprocessing

Packages: tidyverse, edgeR → Downloaded raw count data and metadata file → Data wrangling and filtering NAs → Conversion to matrix and DGElist object

2. Preparation for DE analysis

Packages: tidyverse, edgeR, limma, RColorBrewer → Transforming data by log counts per million (log-CPM) → Filtering out lowly-expressed genes, visualising results of filtering → Normalising gene expression distribution by TMM, upperquartile, or no method → Multi-dimensional scaling plot for unsupervised sample clustering

3. DE analysis

Packages: edgeR, limma, Glimma → Create a design matrix and contrasts → Remove heteroscedascity, fit linear models to the data → Examine the number and direction of exons in DE, visualise results → Output exact exon ID and plot data interactively

Figure 1. The effects of different types of normalisation on the data.

This figure shows the distribution of log-CPM values for each sample with A) no normalisation applied, B) TMM normalisation applied, and C) Upper quartile normalisation applied. Slight changes in the median value are evident with normalisation, particularly in the sample VM_Veh_1 (green box).

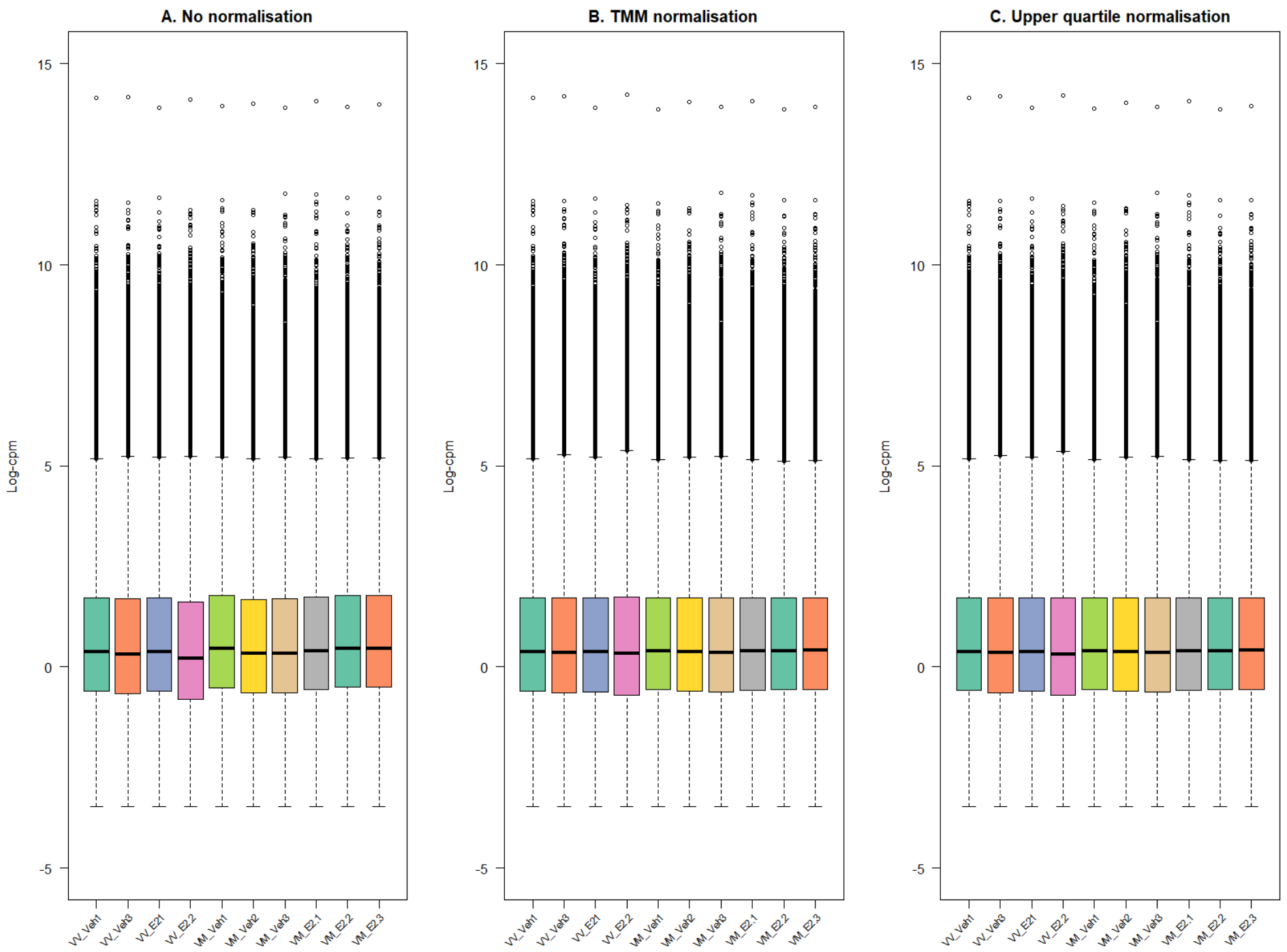
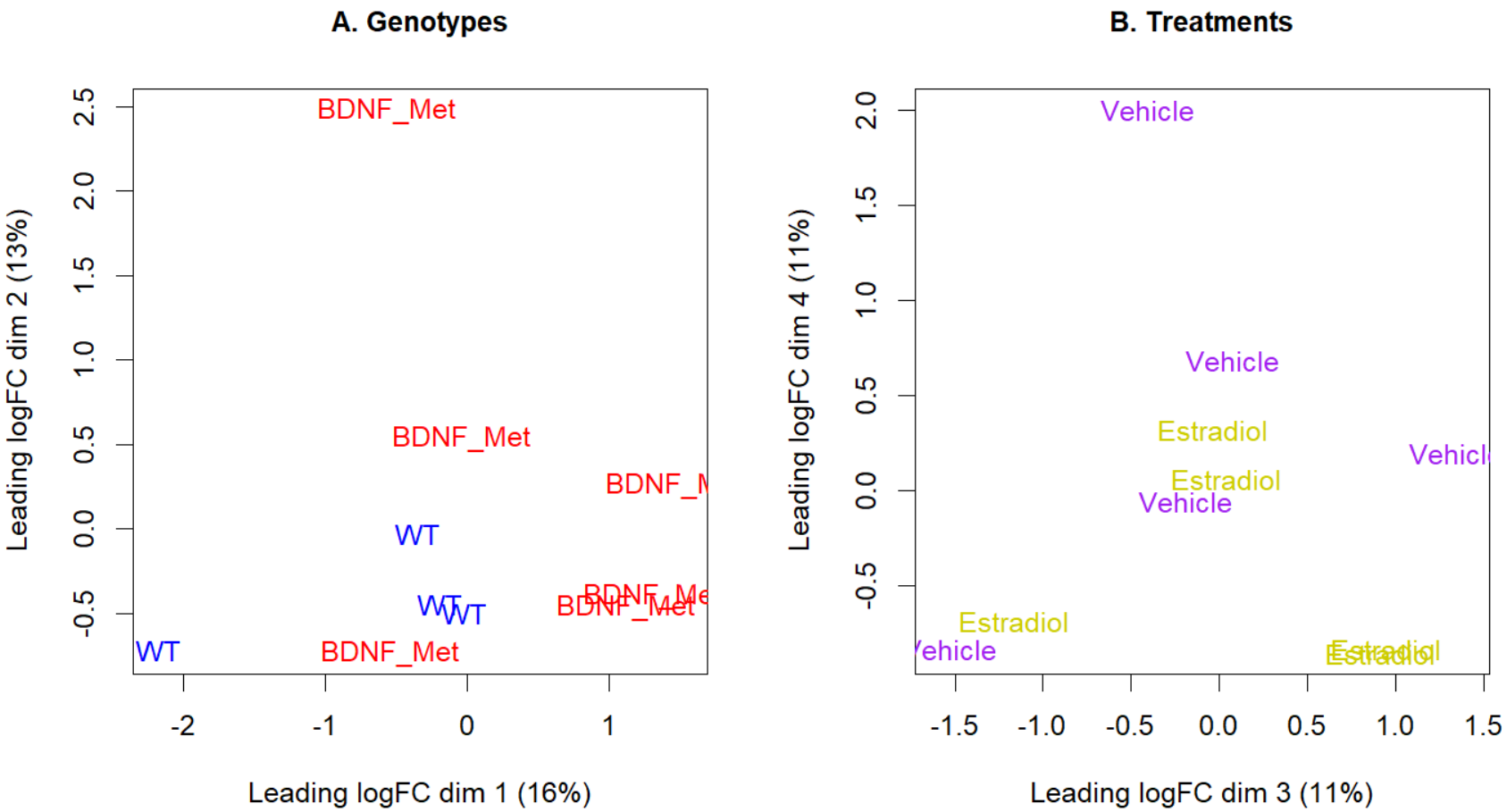
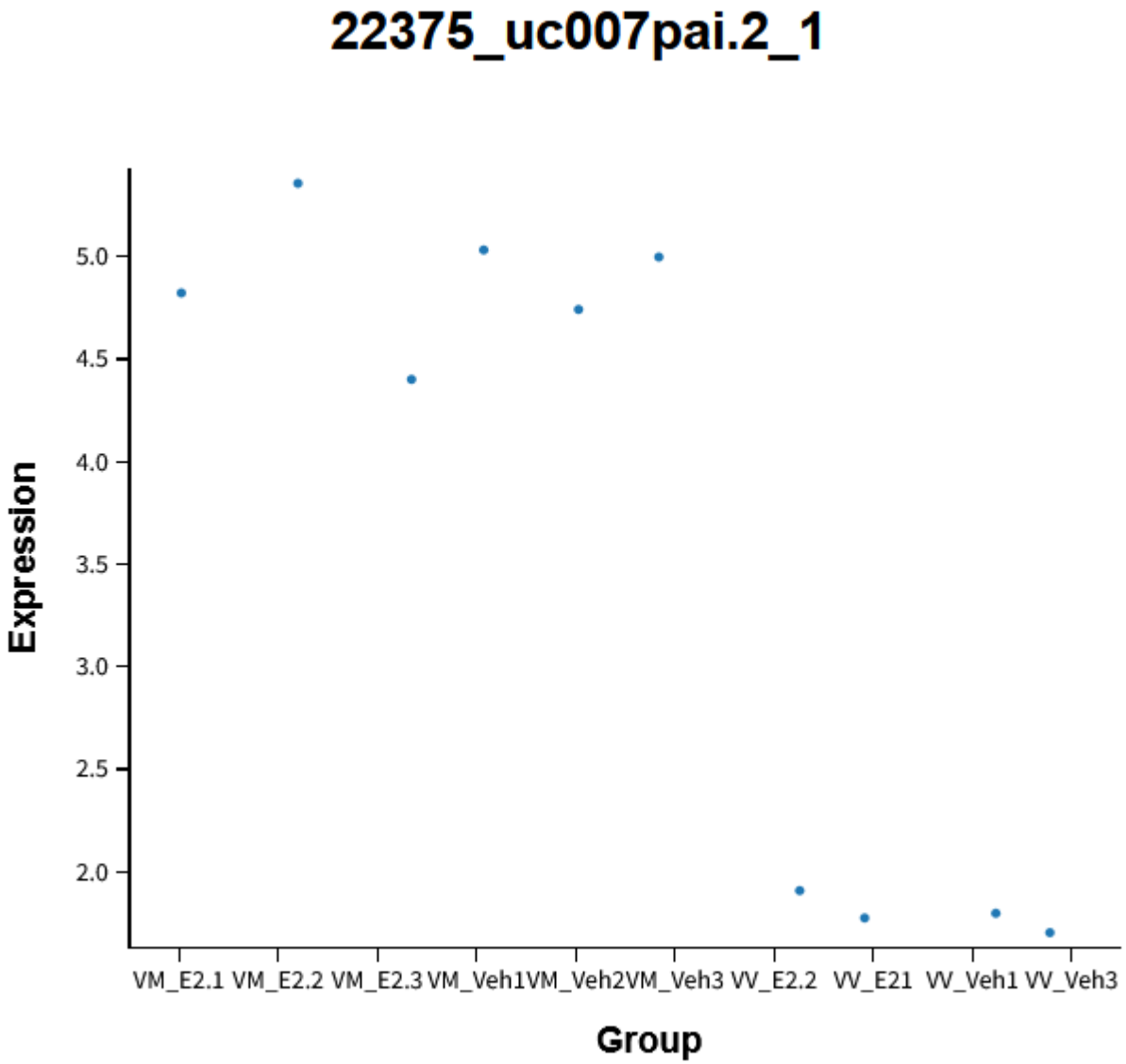


Figure 2. Unsupervised clustering of samples normalised with the TMM method



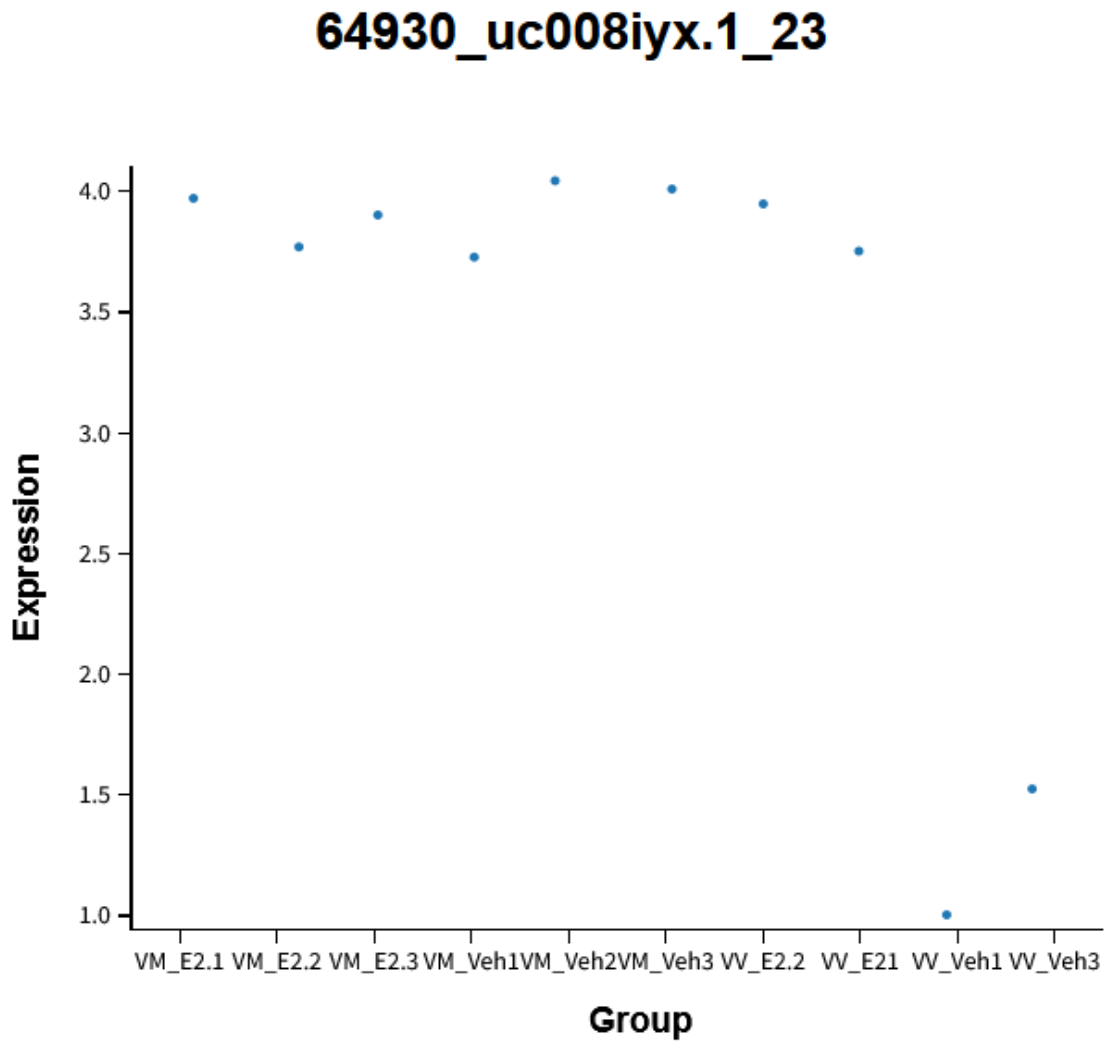
These multi-dimensional scaling plots show how samples cluster based on A). their genotype, and B). the treatment they received. Dimension 1 shows the leading log-fold change that explains the most variation in the data and separates the samples the best, with subsequent dimensions explaining smaller proportions of variation in the data. I examined both genotype and treatment over multiple dimensions, and each clustered best over dimensions 1 and 2 or 3 and 4, respectively.

Figure 3. Plot generated with Glimma showing the most significantly differentially expressed exon between WT and BDNF_Met, normalised with UQ. Each point on the graph shows the expression level in each sample on the x axis. ** LINK DOESN'T WORK- copoy from web browser instead



This plot screenshot shows the expression levels of the exon 22375_uc007pai.2_1 in each sample, which was the most significantly differentially expressed exon between WT and Het-Met samples. Het-Met samples (beginning with VM) have much higher expression levels compared to WT samples (WT). This exon was significantly down-regulated with a log-fold change in expression of -3.123 (p.adj =0.007153). Click **here** to view the interactive plot. On the web page, the left-hand plot shows the log-fold-change vs average expression of all exons. The right plot shows the expression levels of a particular exon. At the bottom is a list of all exons with their log-fold change and associated p-values. Click each one to display its plot.

Figure 4. Plot generated with Glimma showing the most significantly differentially expressed exon between WT and BDNF_Met genotypes treated with Vehicle, normalised with UQ. Each point on the graph shows the expression level in each sample on the x axis.



This plot screenshot shows the expression levels of the exon 64930_uc008iyx.1_23 in each sample, which was one of the five significantly differentially expressed exon between WT and Het-Met genotypes treated with Vehicles. Het-Met samples (beginning with VM) and WT (VV) samples treated with estradiol have much higher expression levels compared to WT samples treated with vehicle. This exon was significantly down-regulated with a log-fold change in expression of -2.694 (p.adj =0.02475). Click **here** to view the interactive plot. On the web page, the left-hand plot shows the log-fold-change vs average expression of all exons. The right plot shows the expression levels of a particular exon. At the bottom is a list of all exons with their log-fold change and associated p-values. Click each one to display its plot.

Table 1. Differentially expressed exons between genotypes and treatment based on normalisation method

Normalisation method	**Contrast**	**Exon name**	**Log-fold change**	**Adjusted p-value**
TMM	WT vs Het-Met	22375_uc007pai.2_1	-3.119	0.0063
		22375_uc007pai.2_3	-3.170	0.0255
	WT vs Het-Met with Vehicle treatment	64930_uc008iyx.1_23	-2.6945	0.0278
		207728_uc009iot.3_28	-2.4716	0.0396
		22375_uc007pai.2_1	-3.2012	0.0396
		64930_uc008iyz.1_23	1.3469	0.0396
		330119_uc012dyf.1_1	1.8548	0.0396
Upper quartile	WT vs Het-Met	22375_uc007pai.2_1	-3.1234	0.0072
		22375_uc007pai.2_3	-3.1746	0.0277
	WT vs Het-Met with Vehicle treatment	64930_uc008iyx.1_23	-2.6938	0.0247
		207728_uc009iot.3_28	-2.477	0.0402
		64930_uc008iyz.1_23	1.3470	0.0402
		22375_uc007pai.2_1	-3.2010	0.0402
		330119_uc012dyf.1_1	1.8547	0.0402

Normalisation method	Contrast	Exon name	Log-fold change	Adjusted p-value
None	WT vs Het-Met	**22375_uc007pai.2_1**	**-3.1796**	**0.0143**
		64930_uc008iyx.1_23	**-1.4178**	**0.0252**
		22375_uc007pai.2_3	-3.2308	0.0272
	Vehicle vs E2 treatment for WT genotype	64930_uc008iyx.1_23	-2.5878	0.0274
		64930_uc008iyz.1_23	1.4647	0.0325
	WT vs Het-Met with Vehicle treatment	64930_uc008iyx.1_23	-2.7148	0.0099
		64930_uc008iyz.1_23	1.32439	0.02467

- As seen in the tables, normalisation method greatly affects the data.
- Different methods found different significant exons, and if they did find the same significant exons, log-FC and adjusted p-values were different
- No normalisation resulted in drastically different results, highlighting the importance of normalising the data in order to uncover biologically true results

More in depth on annotations…

- I chose to investigate the annotation of significant exons found by both TMM and upper quartile normalisation.
  - For simplicity, I will focus on two: the most significant exon between genotypes, **22375_uc007pai.2_1,** and the most significant exon between genotypes with Vehicle treatment, **64930_uc008iyx.1_23.**
- The authors were not transparent about how they annotated their data, which made it difficult for me to determine exon identity. However, I was able to use the UCSC genome tool to search transcript IDs, and used NCBI Gene to search gene IDs
- **22375_uc007pai.2_1:** *Wars1* gene, involved in tryptophanyl-tRNA aminoacylation and tryptophan-tRNA ligase activity (*Wars1*, NCBI Gene).
- **64930_uc008iyx.1_23:** *Tsc1* gene, involved in ATPase inhibitor activity, associative learning, negative regulation of ATP-dependent activity, and protein stabilization. Implicated in several disorders and tumour suppression *(Tsc1*, NCBI Gene*)*.

Are these genes linked to PMDD?

- Although I could not find direct evidence of Wars or Tsc1 being linked to PMDD, I used NCBI Gene to view information about these genes.
- Both genes are highly important:
  - *Tsc1* associated most significantly with tuberous sclerosis, which manifests as a variety of diseases (Henske et al., 2016).
  - *Wars1* is associated with gastric cancer (Oshima et al., 2024) and sepsis (Kim et al., 2023), amongst other diseases.
- Further investigation into the roles of these genes in response to hormone fluctuation is warranted.

Discussion and conclusion

- I found that normalisation of data, and method of normalisation, affects results of differential gene expression analysis
  - TMM vs. upper quartile: top DE exons had different log-FC and adjusted p-values
  - No normalisation: DE exons in different contrasts, different log-FC and adjusted p-values
- The authors found many more differentially expressed genes than I did:
  - They found hundreds of significant genes when comparing genotypes and treatments, including specific genes like *Gria2* and *Gabarap* (to name only a few), none of which match the genes that I found.
  - Their lack of reproducible code or detailed computational methods meant I could not reproduce their analyses. They only mentioned that they used DEseq2 to do their analyses.
  - This highlights how different analytical pipelines can produce different results, as per Tong et al. (2020). However, such drastically different results are surprising, and motivate me to investigate further. Perhaps the authors used methods or tools that are better suited to the data, leading to more accurate results.
- The genes that I found are not explicitly linked to PMDD
  - Given more time, a full investigation into these genes could shed light on their functions or potential relation to PMDD
- The biggest limitation of this project is lack of time to investigate methods further (i.e., why my results were so different) or conduct a full literature review of the genes that I did find. Additionally, the format of the data (not being able to merge annotations to the data as in Law et al. (2018)) limited the analyses I could do.
- Another limitation is the data type cannot be plotted in ggplot, which I am used to- I found the functionality and flexibility of plotting greatly reduced.

Reflection

- I chose this project so that I could learn about RNA-seq analysis in preparation for my BINF 6999 project. I chose this topic as it is something I knew almost nothing about, and I will always welcome the opportunity to explore something new. Through completing this project, I definitely learned a lot about differential gene expression analysis and advanced my skills with R and certain packages like edgeR. Although, as is the case with almost all topics that I learn in biology and bioinformatics, I feel as though I have learned more about what I don't know instead, as there is a staggering amount of knowledge related to these types of analyses, methodological choices, statistical knowledge, etc.
- Particularly with regards to this project, PMDD is a disorder with debilitating symptoms, yet we still do not know the exact cause or causes. Examining data of this nature motivates me to pursue a career in bioinformatics, as there are so many unanswered questions. This is particularly relevant to women's health, wherein many health conditions are severely understudied or misunderstood.
- In my future coursework, BINF 6999 project, and career, I will take forward my motivation to answer pressing questions in women's health, as well as the ability to recognise gaps in my knowledge and take the time to fill them and acquire new skills in the process.

Acknowledgements

Thank you to Dr. Karl Cottenie for assistance with figuring out the analyses for this project.

References

Bäckström, T., Bixo, M., Johansson, M., Nyberg, S., Ossewaarde, L., Ragagnin, G., Savic, I., Strömberg, J., Timby, E., van Broekhoven, F., & van Wingen, G. (2014). Allopregnanolone and mood disorders. *Progress in Neurobiology*, *113*, 88–94. https://doi.org/10.1016/j.pneurobio.2013.07.005

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*(1), 94. https://doi.org/10.1186/1471-2105-11-94

Cary, E., & Simpson, P. (2024). Premenstrual disorders and PMDD - a review. *Best Practice & Research. Clinical Endocrinology & Metabolism*, *38*(1), 101858. https://doi.org/10.1016/j.beem.2023.101858

Hantsoo, L., & Epperson, C. N. (2015). Premenstrual dysphoric disorder: Epidemiology and treatment. *Current Psychiatry Reports*, *17*(11), 87. https://doi.org/10.1007/s11920-015-0628-3

Henske, E. P., Jóźwiak, S., Kingswood, J. C., Sampson, J. R., & Thiele, E. A. (2016). Tuberous sclerosis complex. *Nature Reviews Disease Primers*, *2*(1), 16035. https://doi.org/10.1038/nrdp.2016.35

Kim, Y. T., Huh, J. W., Choi, Y. H., Yoon, H. K., Nguyen, T. T., Chun, E., Jeong, G., Park, S., Ahn, S., Lee, W.-K., Noh, Y.-W., Lee, K. S., Ahn, H.-S., Lee, C., Lee, S. M., Kim, K. S., Suh, G. J., Jeon, K., Kim, S., & Jin, M. (2023). Highly secreted tryptophanyl tRNA synthetase 1 as a potential theranostic target for hypercytokinemic severe sepsis. *EMBO Molecular Medicine*, *16*(1), 40–63. https://doi.org/10.1038/s44321-023-00004-y

Law, C. W., Alhamdoosh, M., Su, S., Dong, X., Tian, L., Smyth, G. K., & Ritchie, M. E. (2018). *RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR*. https://www.bioconductor.org/packages/devel/workflows/vignettes/RNAseq123/inst/doc/limmaWorkflow.html

Law, C., Zeglinski, K., Dong, X., Alhamdoosh, M., Smyth, G. K., & Ritchie, M. E. (2020). *A guide to creating design matrices for gene expression experiments*. https://bioconductor.org/packages/release/workflows/vignettes/RNAseq123/inst/doc/designmatrices.html

References

Marrocco, J., Einhorn, N. R., Petty, G. H., Li, H., Dubey, N., Hoffman, J., Berman, K. F., Goldman, D., Lee, F. S., Schmidt, P. J., & McEwen, B. S. (2020). Epigenetic intersection of BDNF Val66Met genotype with premenstrual dysphoric disorder transcriptome in a cross-species model of estradiol add-back. *Molecular Psychiatry*, *25*(3), 572–583. https://doi.org/10.1038/s41380-018-0274-3

Oshima, T., Hashimoto, I., Hiroshima, Y., Kimura, Y., Tanabe, M., Onuma, S., Morita, J., Nagasawa, S., Kanematsu, K., Aoyama, T., Yamada, T., Ogata, T., Rino, Y., Saito, A., & Miyagi, Y. (2024). Clinical significance of tryptophanyl-trna synthetase 1 gene expression in patients with locally advanced gastric cancer. *Anticancer Research*, *44*(2), 673–678. https://doi.org/10.21873/anticanres.16857

Raffi, E. R., & Freeman, M. P. (2017). The etiology of premenstrual dysphoric disorder: 5 interwoven pieces: a better understanding of the causes of PMDD can lead to improved diagnosis and treatment. *Current Psychiatry*, *16*(9), 21-30.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47. https://doi.org/10.1093/nar/gkv007

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). Edger: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. https://doi.org/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*(3), R25. https://doi.org/10.1186/gb-2010-11-3-r25

Su, S., Law, C. W., Ah-Cann, C., Asselin-Labat, M.-L., Blewitt, M. E., & Ritchie, M. E. (2017). Glimma: Interactive graphics for gene expression analysis. *Bioinformatics (Oxford, England)*, *33*(13), 2050–2052. https://doi.org/10.1093/bioinformatics/btx094

Tong, L., Wu, P.-Y., Phan, J. H., Hassazadeh, H. R., Tong, W., & Wang, M. D. (2020). Impact of RNA-seq data analysis algorithms on gene expression estimation and downstream prediction. *Scientific Reports*, *10*(1), 17925. https://doi.org/10.1038/s41598-020-74567-y

Tsc1 TSC complex subunit 1 [Mus musculus (House mouse)]—Gene—NCBI. (n.d.). Retrieved December 1, 2025, from https://www.ncbi.nlm.nih.gov/gene/64930

Wars1 tryptophanyl-tRNA synthetase1 [Mus musculus (House mouse)]—Gene—NCBI. (n.d.). Retrieved December 1, 2025, from https://www.ncbi.nlm.nih.gov/gene/22375

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." *Journal of Open Source Software*, **4**(43), 1686. doi:10.21105/joss.01686.