Association analysis of endogenous retrovirus and gene candidate expression in myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS)

Sophie Kogut[1], and Dawei Li[1]
[1]Department of Microbiology and Molecular Genetics, University of Vermont, Burlington, VT

Abstract

Despite the high prevalence and debilitating effects of myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS), no biomarkers or effective treatments have been discovered. As a result, ME/CFS remains difficult to diagnose and manage, and the mechanisms that drive ME/CFS progression remain largely unknown. Our hypothesis is that certain endogenous retroviruses (ERVs) (distinct elements in our genome that are the remnants of ancient retroviral infection) may induce an abnormal immune response that may trigger the development of ME/CFS or otherwise influence the progression of the disease; while many ERVs are silenced by mutations or other genes, others retain the ability to produce viral proteins and potentially impact human health[1]. This project was conducted using RNA-Seq data from ME/CFS patient and healthy control samples to examine the association of ERV expression with ME/CFS diagnosis, symptom severity levels and related genes. We identified ERV candidates that should be considered for further in-depth study with respect to fatigue and post-exertional malaise (PEM) symptoms in ME/CFS patients, and regulatory genes that should be considered in future studies.

Objectives

We aim to identify:

- **top ERVs associated with fatigue and PEM scores**
- **top genes associated with altered ERV expression**

Methods

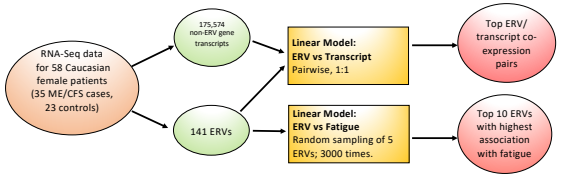

Figure 1 Pipeline for our analysis.

Figure 2 Transformation from raw read counts to ranked counts for each ERV. ERV expression raw read counts were ranked among the 58 samples for each ERV. ERVs with a positive $\log_2$ folder change (FC) value were ranked from the lowest to highest, while ERVs with a negative FC were ranked from the highest to lowest so that the highest score was consistently associated with ME/CFS.

Method and Materials

1. Rstudio[3] was used for all data processing and modelling, including the packages "dplyr"[4], "patchwork"[2], "ggplot2"[5]. Whole transcriptomic RNAseq data was obtained for all patients (231,203 genetic elements). Only 58 Caucasian females were used (35 cases and 23 controls), and only ERVs with a $\log_2$ fold change (FC) of greater than |1| were selected (resulting in 141 ERVs) for analyses (**Figure 1**).
2. RNA expression reads counts were ranked across the patient samples according to the method described in **Figure 2.** Ranks were also weighted by multiplying each rank by the absolute value of the ERV's FC.
3. Linear modelling was used to obtain estimate sizes and *P* values for 141 ERVs relative to fatigue scores by sampling ERVs in groups of 5 (**Figure 3, Table 1**).
4. A list of gene transcripts associated with methylation and ERV regulation were identified and used to determine the top ERV/transcript co-expression pairs via linear modelling (**Figure 4**).
5. Relevant genes were used to construct an expression network using the STRING: Protein-Protein Interaction Network[6] online tool **(Figure 5).**

Results

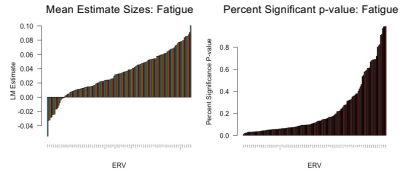

Figure 3 Results of part 1: random selection of 5 ERVs (from list of 141 that met selection criteria) compared against patient fatigue scores using a linear model with 3,000 trials. **3A** illustrates the mean effect sizes for each ERV, ordered from the smallest to largest effect size. **3B** illustrates the percentage of occurrences in which the *P* value for an individual ERV was considered significant (< .05) across all trials.

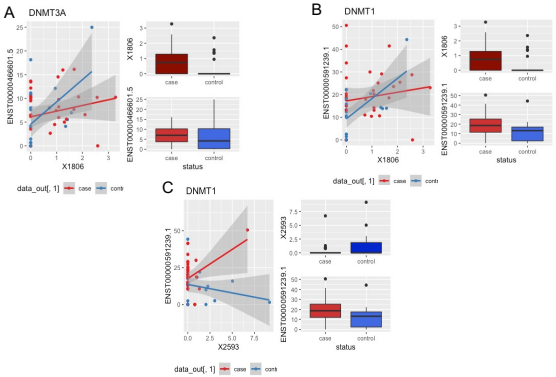

Figure 4 Plots showing co-expression pattern of ERVs and gene transcripts. Each plot shows: **4A** Increased expression of ERV 1806 in cases, higher mean expression of DNMT3A in cases; **4B** Increased expression of ERV 1806 in cases, higher mean expression of DNMT1 in cases. **4C** Decreased expression of ERV 2593 in cases, higher mean expression of DNMT1 in cases.

ERV Sampled	Mean Estimate	Percent Pval
2593	0.10050686	0.97142857
W.117	0.09085905	0.68571429
3184	0.08767367	0.99056604
2393	0.08585614	0.68695652
5874	0.08529644	0.81415929
1806	0.08503485	0.98888889
4017	0.08365421	0.69166667
K.42	0.07990994	0.91752577
5893	0.07911467	0.61538462
1162	0.07812226	0.57657658
UID.142	0.07789061	0.98924731
5110	0.07728484	0.82692308
3723	0.07659009	0.66949153
1424	0.07480841	0.59166667

Table 1 Lists the top 10% of significant ERV's with respect to fatigue scores, ranked (high-low) using estimate size as the ordering criterion.

- A bold designation in **Table 1** indicates that the ERVs appeared in the top 10% based on estimate size and *P* value for both fatigue and PEM scores, which are closely associated with ME/CFS case status. Expression associated with these scores indicates predictive potential of these ERVs.

Network

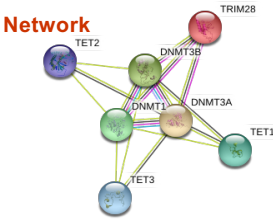

Figure 5 A network of notable protein-coding genes associated with methylation and regulation. Nodes represent all proteins coded by that gene; edges indicate association. Edge colors here represent:

- Light green--- association in literature
- Pink--- experimentally determined association
- Black --- co-expression
- Blue --- associations noted in a database

Conclusions

- The ERVs that were consistently identified in the top 10% based on estimates/*P* value observation for both fatigue and PEM are: **2593, UID.142, 3184, 1806, K.42, 4017 (Table 1):**
  - Future experiments are required to further examine their role.
- Several transcripts from the *DNMT1 and DNMT3A* genes demonstrated the most significant interaction with ERVs (**Figure 4**):
  - DNMT1 is involved in the methylation or silencing of genes.
  - *New Hypothesis*: DNMT1 may be induced by the expression (or lack thereof) of an ERV that is associated with ME/CFS.
- More work is needed to further explore proposed interactions. New methods for interaction identification on a larger scale should be considered.

References

1. Küry, P., Nath, A., Créange, A., Dolei, A., Marche, P., Gold, J., ... & Perron, H. (2018). Human endogenous retroviruses in neurological diseases. *Trends in Molecular Medicine*.
2. Pedersen, T. L. (2019). patchwork: The Composer of Plots. R package version 1.0.0 https://CRAN.R-project.org/package=patchwork
3. R Core Team (2019). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. URL https://www.R-project.org/.
4. Wickham, H., François, R., Henry, L., Müller, K. (2019). dplyr: A Grammar of Data Manipulation. R package version 0.8.3. https://CRAN.R-project.org/package=dplyr
5. Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.
6. Szklarczyk et al. Nucleic acids research 47.D1 (2018): D607-D613.2

Acknowledgments

Thank you to Jason Kost and Dr. Scott Merrill for their help and advice.