Analysis of DRD3 Gene Expression in Essential Tremor Patients with Respect to Gender and Age

Nvolve Fall 2023 Project
Scholars: Olga Drozdovitch, Nora Bui
Coach: Dr. Hermioni Zouridis

Abstract

Background Essential Tremor (ET) is a disorder of the nervous system primarily characterized by rhythmic shaking which may occur in the hands, but can also be present in the head, arms, or legs. Though sharing some similarities with Parkinson's Disease, ET has been classified as a separate disorder. A study by Martuscello et. al (2019) identified 231 'top gene hits' differentially expressed in ET ($n_{ET} = 33$, $n_{control} = 21$) and characterized ET as a family of related disorders caused by various types of biological dysregulation, rather than a single entity. ***Goals and Methods*** We used the dataset from the Martuscello et. al paper for our analysis: 1) Using a template R-script, Principal Component Analysis (PCA), and differential analysis, we explore whether there is genetic variability with respect to age and gender. 2) Using Chi-square and t-tests and focusing on the gene Dopamine D3 Receptor (DRD3), we examined whether DRD3 is expressed more highly in ET vs. control as well as whether there is a pattern in DRD3-expressed samples with respect to age and gender. 3) We utilized the BLAST and SnapGene technologies to determine if the wild-type or mutant variant of the DRD3 gene was present in the samples. ***Results*** 1) PCA revealed that there is some variability in gene expression with respect to age. Differential analysis determined that there's no gender difference in expression of all key genes. 2) DRD3 is expressed slightly higher (but not significantly) in ET when compared to control. DRD3-expressed samples have slightly lower age distribution though still falling within the range of all samples, but there's no difference in gender distribution. 3) One sample of interest aligns to the consensus sequence with good accuracy, but the result is inconclusive due to the short length of fragments.

Essential Tremor (ET) is disorder of the nervous system primarily characterized by rhythmic shaking which may occur in the hands, but can also be present in the head, arms, or legs. Though it has similarities with, and has often been confused as Parkinson's Disease, ET has been classified as a separate disorder. To begin the exploration of ET, we reviewed a study conducted by researchers at the Columbia University Medical Center and the New York Presbyterian Hospital on Gene expression analysis of the cerebellar cortex within ET patients. Utilizing their count RNA-Seq data (Series GSE134878), initial analysis on 32 samples was performed with relation to gender and age to discern if each factor would influence expression levels. A single gene of interest, Dopamine D3 receptor (DRD3), was identified and found to be linked to ET patients of varying nordic and western/eastern european nationalities. Conducting principal component and differential expression analysis in R and performing statistical testing, we determined that there is no correlation between a gender or age group in relation to ET gene expression and DRD3 expression. PCA did yield possible groupings of samples which was corroborated by the cluster dendrogram, leading us to hypothesize there may be additional factors at play not including the patient's age or gender. Additional DRD3 gene alignments were performed to determine whether or not the mutant or wild-type variant of the gene was present within the samples. These alignments were inconclusive due to the short length of fragments and the lack of clean-up in the raw RNA-Seq data used, except

for one sample of interest (Female, 90; GSM3974829). Further cleanup of the dataset would be necessary to yield more conclusive results. As a future consideration, a different gene of interest could be selected, one with more mRNA fragments sequenced. Our findings corroborated that of the original study, where 'gene hits' of interest were found, but they did not exhibit increased expression given any age group or gender.

Introduction

Essential Tremor (ET) is disorder of the nervous system primarily characterized by rhythmic shaking which may occur in the hands, but can also be present in the head, arms, or legs. Though it has similarities with, and has often been confused as Parkinson's Disease, ET has been classified as a separate disorder. To begin the exploration of ET, we reviewed a study conducted by researchers at the Columbia University Medical Center and the New York Presbyterian Hospital on Gene expression analysis of the cerebellar cortex within ET patients. Their goal was to explore the molecular source of ET through transcriptomic analysis. Researchers identified gene transcripts which were differently expressed in control samples when compared to those from ET patients. Using frozen cerebellar cortex tissue, 33 ET patients were compared to 21 normal controls (from donors). Differential analysis revealed 231 differentially expressed gene transcripts (referred to as 'top gene hits')[1]. A notable finding was that genes made up a heterogenous profile, they contribute to diverse categories of dysfunction. Overall, researchers identified four main categories of dysregulation that numerous processes fall into: axonal guidance, microtubule motor activity, ER-Golgi transport, and calcium signaling/synaptic transmission[1]. Their findings supported their hypothesis that ET is not a single entity, it can instead be classified as a family of related disorders caused by cellular reactions from various types of biological dysfunction[1].

The study found sources of variation which they could not account for. Using the metadata of their samples (Series GSE134878), we explored the variability with respect to gender and age of the cohort of 32 samples. These samples originated from both patients not diagnosed with ET (control) those diagnosed with ET. Using an R script generated by our sponsor, Dr. Hermioni Zouridis, we utilized unsupervised principal component analysis and hierarchical clustering to generate three groups which seemed to have an association with age, but were not statistically significant. Differential expression analysis was then performed and it was identified that no genes which acted as biomarkers for ET were differentially expressed with relation to gender. Gender as a factor was not explored in the initial paper, but the original researchers did allude to inconclusive findings when considering age as a variable[1].

Pinpointed a particular gene of interest mentioned in the original study, whose role in ET was corroborated by other sources. It was identified that three ET-associated loci (ETM1, ETM2, ETM3) were identified, but no gene causative mutations were identified. A common variation of a DRD3 mutant (rs6280) within the ETM1 locus was suggested to be a susceptibility factor to ET[2.3]. One study by Gorovora et. al found that the Ser/Gly genotype with Ser9Gly polymorphism in the DRD3 gene gene increases the risk of developing ET by 2.35 ($p = 0.02$).[4] Additionally, through a genome wide linkage scan of Icelandic families, a linkage peak marker of the ETM1 locus emerged and was located from 1 to 10 megabases away from the DRD3 gene[3]. Finally, there was a suggested, weak association between the DRD3 mutant present within various western and eastern european populations, but researchers did deem it to not classify as statistically significant between cases and controls, and suggested further studies be conducted[3]. For our own analysis, we performed statistical tests between the four DRD3-presenting samples and the remaining 28

samples. Evolving the scope of the project, our final research aimed to verify whether the mutant or wild-type of the DRD3 gene was present in the original sample. DRD3 fragments were retrieved from the unprocessed mRNA reads and gene alignments were performed to the DRD3 consensus sequence. Results were inconclusive, with one fragment of one DRD3-containing sample aligned to the consensus better than the remaining samples. This may loosely suggest presence of the mutant variant in this sample, smaller fragments non-DRD3 containing control samples also aligned to the consensus sequence suggesting the data needs further cleanup prior to additional alignment.

Methods and Results

PCA and Differential Expression Analysis

The PCA plot with respect to gender shows that there is minor clustering at the higher values of PC1 for females but not as much for male. The PCA plot with respect to age shows that there isn't obvious clusterings, partially due to more ages (15 groups) we are looking at.

For differential expression analysis with respect to gender, we concluded that there are no differentially expressed genes with respect to gender. logFC represents changes in expression of genes between control and treatment conditions. The closer the logFC is to zero, the less of a difference there is between gene expression of control vs. treated group. Since all genes show logFC close to 0, there are no differentially expressed genes with respect to gender. Wald test with $p > 0.05$ so we can conclude that we cannot reject the null hypothesis. This pattern is similar to the PCA. Though there was some clustering for females in the PCA plot, the clustering is perhaps not significant enough to take into account.
Next, we look at sample clustering with respect to age groups (instead of individual age like in the first part). The R script utilized unsupervised hierarchical clustering to identify clusters based on distance metrics, not based on specific commands but based on dissimilarities among the 3 groups. A resultant dendrogram identified three age groups based on PC1 value: Group 1 has lowest PC1 value, group 2 has highest PC1 value, and group 3 has intermediate values. A PCA plot confirms the same pattern, where Group 3 shows the most clustered pattern followed by group 2. Group 1 does not show clustering. Lastly, we looked at box plots to visualize the difference among these three groups. There seemed to be a slight difference between groups with respect to age: group 1 - low, group 2 - higher, group 3 - highest. The variability in ages is lowest in group 1, increases with group 2, and is highest with group 3. However, the result appears to not be statistically significant.

In conclusion, we did not identify any significant clustering pattern in the PCA plots with respect to age and gender nor did we identify differentially expressed genes with respect to age and gender.

DRD3 Gene Expression Analysis

We performed an analysis of the DRD3 gene using a RNA sequencing gene expression dataset provided by our sponsor (the rows represent different genes, the column represents different samples, and the number represents the number of molecules expressed.) Genecard.org indicates that here are the different names for our gene of interest: DRD3, D3DR, ETM1, FET, and the dataset does contain information on our gene of interest. We used a sample info dataset to correspond the sample expression data of our gene of interest to the donor’s age, gender, and diagnosis. We decided to perform three analyses with Excel: comparison of gene expression level between control samples vs. essential tremor samples, age distribution in DRD3-expressed vs.

non-DRD3-expressed samples, and gender distribution in DRD3-expressed vs. non-DRD3-expressed samples

Comparing the mean gene expression level (number of DRD3 molecules expressed) between ET samples (n = 32) and control samples (n = 20), a student t-test yields 0.57 (> 0.05). Therefore we cannot reject the null hypothesis that there is a statistically significant difference between the gene expression level of DRD3 between ET and control groups.

Since DRD3 is expressed at a low level in the samples (either 0 or 1), we wanted to explore if there is a pattern with respect to age or gender in the samples that do express DRD3, regardless of whether the samples have essential tremor. To visualize the age distribution in DRD3-expressed vs non-expressed sample, we created a box plot which shows that DRD3-expresed samples have age distribution on the lower end, though still fitting within the Quartile 1- Quartile 3 range of the non-expressed samples. To analyze the gender distribution in DRD3-expressed vs non-expressed sample, we conducted a goodness-of-fit Chi-square test ($X^2 = 0.0625$, p-value = 0.80 > 0.05) comparing the DRD3-expressed gender ratio (female:male) to the total ratio (female:male) to see if there is a deviation from expected. We concluded that we cannot reject the null hypothesis that there is a difference between gender ratio in DRD3-expressed sample when compared to the total sample. Due to the statistically insignificant findings of these analyses, we were unable to conclude that gender or age played a role in gene expression of DRD3.

Gene Sequence Alignments

The final step of exploration evolved beyond the initial scope of the project, out of a desire to investigate the gene further. It was determined that the primary variant of DRD3 within various ET loci was found to be a point mutant version (rs6280) where a serine residue at position 9 of the N terminal part of the receptor is replaced by a glycine[3]. Of the 32 present, four DRD3 expressing samples were identified, each with only a single sample of mRNA identified during sequencing. To discover if the mutant or wild-type variant of DRD3 was present, BLAST was utilized to extract fragments of the DRD3 gene from the raw RNA-Seq reads of the original sample. There was a range of 3-11 fragments per each DRD3 expressing sample, with their length staying consistently around 75 base pairs. Within SnapGene, all mRNA fragments were aligned to isoform a of DRD3 (NM_000796.6) acting as the consensus sequence. Using the Muscle alignment algorithm, these alignments proved to be inconclusive, but one sample of interest (DRD3-presenting, Female, 90; GSM3974829) aligned to the consensus sequence with greater accuracy when compared to the other fragments. This result suggests that this sample may contain the mutant variant of the DRD3 gene, but further investigation is required as the alignment of the other samples to the consensus sequence suggests this may be a characteristic of PCR amplification and not of the samples. Even the control samples without identified DRD3 expression had fragments which aligned to the consensus sequence, suggesting these samples require further treatment.

To achieve more conclusive results with sequencing, we would need to proceed with the process of obtaining a true final mRNA sequence, following subsequent steps of RNA-Seq cleanup. The dataset used lacked this cleanup, however the researchers created a completed RNA-Seq library following the treatment and purification of these samples for the original study which was reviewed. There are analytical tools present to allow for the creation of a single sequence for each sample from the fragments initially aligned to the consensus of the complete DRD3 gene, and this could be a possible next step to retrieve better alignment data. An additional future consideration would be the selection of a different gene of interest from the ‘top gene hits’ presented in the

original study. Each DRD3-presenting sample only had a single mRNA hit identified, due to the small sample size there could be additional errors involved despite the high thru-put accuracy of RNA-Seq technology. Choosing a different gene could potentially provide more conclusive results. Overall, our findings corroborated that of the original study, where a 'hit genes' of interest were analyzed, but they did not exhibit differential expression given any age group or gender.

Glossary

Unsupervised method: A machine learning data sorting technique where sorting occurs into predefined categories which an algorithm is trained to recognize.

PCA: Principal Component Analysis. An unsupervised technique where, given a large unsorted dataset, principal components are identified via level of variability with PC1 being the highest and PC2 being the second highest. Both are linear combinations of expression levels of genes and are collections of components; with a variety of genes contributing to the variability of a variety of principal components. Primary purpose is to dimensionally reduce a dataset and collapse expressed genes on the basis of variance, ignoring other multidimensional components.

Clustering: Also referred to as hierarchical cluster analysis. An unsupervised method where data is presented to an algorithm and the algorithm builds various clusters by measuring dissimilarities between data, without a target variable. This data can be partitioned in regard to the grouping of similar output data points.

RNA-Seq: RNA sequencing. A sequencing technique which uses next-generation sequencing (NGS), a model of sequencing which offers ultra-high throughput sequencing, the parallel sequencing of multiple nucleotide molecules, enabling a far faster rate of processing at a time. This can be utilized even with small trace amounts of a gene in the original sample, to potentially detect trace amounts of a single copy of mRNA.

FASTA/FASTA Sequence: An abbreviation of 'Fast-All', FASTA is a sequence alignment tool which uses nucleotide or protein sequences as inputs and compares these sequences to existing databases. The FASTA Sequence is a text based format for representing nucleotides or single-letter codes of amino acids in an easily accessible file type.

Gene Alignment: The process of aligning RNA or DNA nucleotide fragments to a consensus sequence which is usually the sequence of the entire gene. The format used most often to input these sequences is typically FASTA.

Differential Analysis: A method used to identify genes that have differential expression pattern between two groups due to biological condition through calculating log fold change in expression of genes between treated and untreated

Random forest: a machine learning algorithm where one takes out each sample then reclusters (repeated for 1000's times) to test how stable the clustering is. The result of the random forest is summarized by a confusion matrix

Batch effect: variability with respect to different institutions and sources that the samples come from rather than meaningful differences

Works Cited

1. Martuscello, R. T., Kerridge, C. A., Chatterjee, D., Hartstone, W. G., Kuo, S. H., Sims, P. A., Louis, E. D., & Faust, P. L. (2020). Gene expression analysis of the cerebellar cortex in essential tremor. Neuroscience letters, 721, 134540. https://doi.org/10.1016/j.neulet.2019.134540
2. Merner, N. D., Girard, S. L., Catoire, H., Bourassa, C. V., Belzil, V. V., Rivière, J. B., Hince, P., Levert, A., Dionne-Laporte, A., Spiegelman, D., Noreau, A., Diab, S., Szuto, A., Fournier, H., Raelson, J., Belouchi, M., Panisset, M., Cossette, P., Dupré, N., Bernard, G., … Rouleau, G. A. (2012). Exome sequencing identifies FUS mutations as a cause of essential tremor. American journal of human genetics, 91(2), 313–319. https://doi.org/10.1016/j.ajhg.2012.07.002
3. Siokas, V., Aloizou, A. M., Tsouris, Z., Liampas, I., Aslanidou, P., Dastamani, M., Brotis, A. G., Bogdanos, D. P., Hadjigeorgiou, G. M., & Dardiotis, E. (2020). Genetic Risk Factors for Essential Tremor: A Review. Tremor and other hyperkinetic movements (New York, N.Y.), 10, 4. https://doi.org/10.5334/tohm.67
4. Govorova, T. G., Popova, T. E., Tappakhov, A. A., Golikova, P. I., Danilova, A. L., Antipina, U. D., Samorseva, V. N., Petrova, A. Y., Andreev, M. E., & Lyasheeva, N. N. (2020). The impact of DRD3, HS1-BP3, and LINGO1 gene mutations on the development and clinical heterogeneity of essential tremor in the Sakha Republic (Yakutia). Annals Of Clinical And Experimental Neurology, 14(1), 55-61. doi: 10.25692/ACEN.2020.1.6