An AI-driven platform to classify and subtype CNS disease using small RNAs

David W. Salzman*, Terran Melconian, Neal Foster, Nathan S. Ray

sRNAlytics Inc. AstraZeneca BioHub Incubator, 35 Gatehouse Drive, Waltham MA, 02451
*Corresponding author: david.Salzman@srnalytics.com

sRNAlytics

Abstract

Drug developers targeting CNS diseases have a unique challenge due to the complexity and heterogeneity of the pathology they aim to treat (generally at a relatively advanced stage by the time of diagnosis). This complexity and heterogeneity has also impeded the development of preclinical systems that fully model these diseases. Therefore, there is an urgent need to accurately classify and subtype the molecular pathways driving CNS diseases. Understanding the true etiology of these diseases will (i) permit the establishment of more faithful model systems, (ii) lead to the identification of diagnostic biomarkers to enable the detection of early-onset disease, and (iii) accelerate drug development by defining homogeneous disease subpopulations with shared molecular biology.

Because CNS diseases differentially affect different populations of neurons (or overlapping neuronal populations), we hypothesized that taking a multi-disease classification approach would allow us to leverage data from one disease against others to uncover genes and molecular pathways that are uniquely dysregulated in the brains of one disease type compared to all the others. We further hypothesized that these molecular markers would correlate with additional clinical features such as clinical stage and pathological grade.

To test our hypothesis, we focused on analyzing small RNA (sRNA) sequencing data because they are master regulators of gene expression controlling every biological pathway and process in every cell of our body, and gene dysregulation drives many diseases. We amassed a database of sRNA-sequencing data from over 15,000 brain, CSF, and serum samples from neuropathologically-defined, postmortem samples spanning 8 CNS diseases (AD, PD, ALS, SMA, DLB, PSP, MS, and HD). Using a proprietary suite of algorithms, we discovered over 22 million unique sRNAs and created an annotated dataset of their expression patterns across multiple iPSC-derived cell types (e.g. neural progenitors, oligodendrocytes, astrocytes, motor neurons, etc.), using iPSC lines generated from patients with a variety of diseases.

Methods and Results

Figure 1: sRNA-TRIM and sRNA-MAP are proprietary algorithms developed by sRNAlytics that allow us to capture and annotate more sequencing data with better fidelity than traditional approaches. sRNA-TRIM is an adaptive string search algorithm that processes more sequencing data with better fidelity compared to open-source algorithms (left). sRNA-MAP is a short read alignment algorithm that allowed us to annotate over 4.2 million functional small RNAs. To do this sRNA-TRIM utilizes a library of over 5 billion (14 – 44 nucleotide) sequence tags containing SNPs and INDELs from the dbSNP 151 build, canonical and non-canonical RNA edits for the RefSeq non-coding RNA gene list.

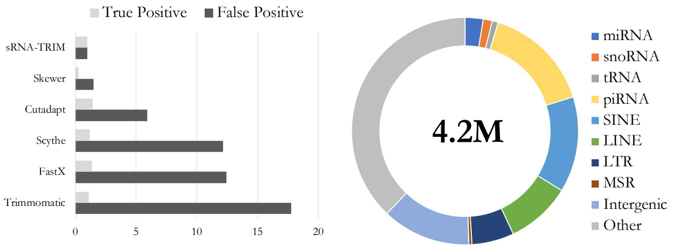
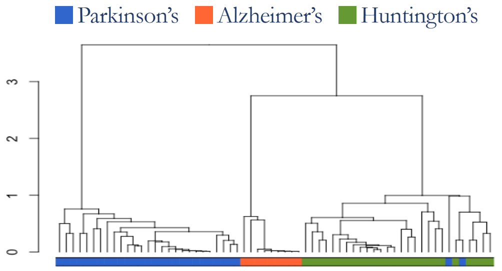
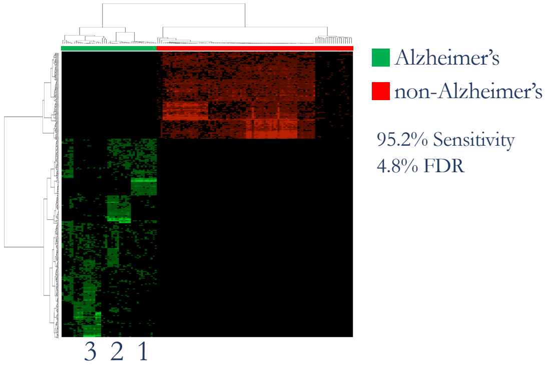


Figure 2: sRNA-FIND classifies 3 CNS diseases with 90.5% accuracy. 175 neuropathologically verified frontal cortex (BA9) samples with small RNA sequencing data were randomized into Training (n=110) and Test (n=65) sets. A total of 60 small RNAs (10 per disease pair) were used to train the sRNA-FIND AI-powered classifier.



85.7% Sensitivity
92.9% Specificity
85.7% PPV
92.9% NPV
Accuracy: 90.5%
F1-score: 85.7%

Figure 3: sRNA-FIND classifies and subtypes Alzheimer's Disease. Unbiased feature selection uncovered novel subtypes of disease, and identified just 8 small RNAs across frontal cortex, cerebrospinal fluid, serum and whole blood that could classify AD vs other CNS diseases. Semi-supervised hierarchical clustering of AD patient samples using these 8 sequences stratified samples into 3 molecular subtypes defined by 2-3 small RNAs per subtype. Investigation of these small RNAs showed that they converge on distinct biological pathways including BACE1, APP and neuroinflammation.



95.2% Sensitivity
4.8% FDR

Conclusion

Utilizing small RNA sequencing and the sRNA-FIND discovery platform, we can classify and stratify CNS diseases into distinct molecular subtypes. This approach has the potential to accelerate both basic mechanistic understanding of disease, as well as targeted drug development. Over the next 36 months sRNAlytics will incorporate over 10,000 new (unique) patient into our database. Each patient will have matched brain tissue (6 regions per patient), CSF, and serum, and each sample will have small RNA, DNA and RNA sequencing, as well as neuropathological metadata attributes. Using this data, we will validate and refine our existing markers, as well as discover new biomarkers that classify and subtype CNS disease based on molecular pathways and neuropathologic data features.