*Tissue Classification Using RNA-Seq Transcriptomics with Distribution Analysis and SVM Models**

Dominick DeCanio[1,2], Minah Kim[2], Samuel Haddox[3], Gianluca Guadagni[2,4]

***Abstract*— The human body generates more proteins than it has genes that code for proteins. The diversity of proteins stems from the alternative ways in which RNA can be spliced and reassembled. Each alternative version of RNA produces a different protein, providing a way for our bodies to produce a wide range of proteins with a single gene. Some alternative RNA transcripts, however, have splicing errors and produce faulty proteins involved in genetic diseases. Understanding splicing patterns and profiles has wide implications for our understanding of healthy and diseased tissue states. Currently little is known regarding the splicing profiles of healthy tissue which vary across individuals and within individuals by tissue type. Therefore, this project explored the use of RNA splicing data from the first chromosome to predict the tissue type of non-cancerous samples using distribution analysis and supervised learning methods. The Kolmogorov-Smirnov test was used to classify the samples based on empirical cumulative distribution functions and was not able to reliably distinguish between tissue types. Our SVM model was run using both the quantity of splice junctions observed and their presence, and had a high prediction accuracy for both data sets. The performance between the two SVM model outcomes were not significantly different. Overall, the findings suggest the utility of using splice junction data in biological classification and sets the foundation for future work of mapping splicing patterns with phenotype.**

I. Introduction

According to the central dogma of molecular biology proposed by Francis Crick in 1970 [1], genetic information stored as DNA is transcribed to RNA prior to translation into proteins, which were mainly considered to be the functional drivers of biology at the time. While today we continue to uncover novel biological functions of RNA beyond coding proteins, we have come to appreciate the significance of evaluating total RNA content, or the "transcriptome," as providing one of the most accessible and comprehensive snapshots of a sample's epigenetic status.

The amount and variety of RNA that is expressed differ by cell and tissue types (e.g. liver vs. brain) and using RNA sequencing (RNA-seq) we can quantify the RNA expression levels for different tissue samples [2]. Machine learning algorithms have been used to predict tissue type or disease status using RNA-seq data at the gene level, which groups together the alternative transcripts that are formed from a single gene [3]-[4]. Alternative transcripts are products of different splicing patterns from a single gene, and each transcript can produce a different protein variant or isoform. For example, the Human BCL-2 gene is a regulator of apoptosis; however, there are isoforms of BCL-2 that are pro-apoptotic and isoforms that are anti-apoptotic [5]. These isoforms share much of the same sequence, so only RNA sequencing reads that directly cover the unique regions differentiating these isoforms from one another are useful for interpreting the apoptotic status based on BCL-2 expression.

Previous studies have also found that biological classification can be improved by including transcript-level information [6]-[7]. Our aim is to utilize splice junction data which can capture fuller information about the presence and quantity of alternative transcripts than if we were to evaluate data at just the gene or transcript level. Splice junctions indicate the location coordinates of where the RNA splicing occurred or the location coordinates of the intron boundaries. The coordinates are composed of a 'donor' and 'acceptor' identifier within the genomic sequence. Prediction of tissue type in non-cancerous samples at the splice junction-level has not yet been evaluated to our knowledge. Evaluating the utility of splicing coordinates in predicting tissue type in healthy samples could set the foundation for investigating the splice junction profiles in healthy physiology. Moreover, previous literature has shown that tissue types in the brain have more divergent splicing patterns than anywhere else in the body, suggesting that distribution metrics of splice junctions could be an important predictor in classification [8]. Therefore, this work aims to demonstrate that using splice junction features and distributional metrics in machine learning models can reliably predict tissue types. Specifically, we focus on tissues from the brain, whole blood, and muscle as comparisons.

II. Methodology

The aim of our research is to determine the limits of tissue sample classification based on transcriptome data using traditional machine learning and distribution analysis techniques. We began this analysis with some preliminary data exploration and summary statistics, before moving into analyses with more complex approaches.

To assess the ability of traditional distribution analysis techniques we used both the Kolmogorov-Smirnov and Wasserstein distance tests to represent this approach. To assess the abilities of traditional machine learning models we chose to employ a support vector machines (SVM) model. To assess the possibilities of future embedding approaches, we conducted the SVM classification using the quantification data

* Research sponsored by the Li lab
[1] Corresponding author, email: dcd9ce@virginia.edu
[2] School of Data Science, University of Virginia, Charlottesville, VA 22903 USA
[3] Department of Biochemistry and Molecular Genetics, School of Medicine, University of Virginia, Charlottesville, VA, 22903 USA
[4] Department of Engineering and Society, School of Engineering and Applied Sciences, University of Virginia, Charlottesville, VA, 22903 USA

of each splice junction and the presence of each splice junction to simulate an embedding approach and compare the results of the two. Furthermore, we employed the t-SNE algorithm to visualize the high-dimensional space of this quantification and presence data to assess the effectiveness of each data set in capturing the distinctions between each class of tissue.

A. Data

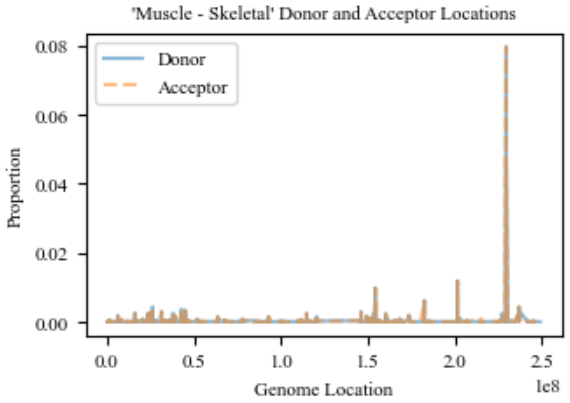
A fastq file contains fragments of nucleotide sequences from every detectable RNA molecule in a sample. These sequences were then aligned to the sample's appropriate genome and compared to gene annotations curated over the last few decades to produce an alignment file that maps each sequencing read to a gene or genomic locus [9]. We used such an alignment file (STAR aligner) with GTEx fastq files to create the SJ.out.tab files containing the splice junction usage info for each sample.

The high dimensionality of this data can create problems for classification models and our workflow in the form of model specification errors and increased computational burdens. To mitigate these issues we chose to only use the transcriptome data corresponding to non-chimeric chromosome 1 splice junction sites in our analyses. In our exploratory data analysis we recognized that the distribution of splice junction observations was heavily skewed. As seen in figure 1, the distribution of splice junction sites observed in a sample are accounted for by the few most frequently occurring junction sites. Because of this, we decided to subset each tissue sample's non-chimeric chromosome 1 transcriptome data to only include the top-500 most frequently occurring splice junction sites to further reduce dimensionality. We also note that for this subset, the distribution of the donor and acceptor sites for a given tissue class are almost identical. This is shown in figure 1. Because of this difference in scale between the deviation of donor and acceptor sites within a tissue class and the deviation between these sites among tissue classes we have chosen to proceed with our distribution analysis considering only the distribution of donor locations.

The data was thus constructed to include the top-500 most frequent splice junctions for each tissue sample. To conduct a distribution analysis to test our ability to classify a tissue sample as a draw from the tissue class's true population we must create a population distribution for each tissue class based on its constituent tissue samples. To do so, we aggregated the number of splice junctions which occurred at each donor site in a sample across equivalent donor sites from all samples within a tissue class.

To conduct the Kolmogorov-Smirnov and Wasserstein distance tests we require a sample from the two distributions in question. We accomplished this by drawing samples of donor sites from the population/sample which were directly proportional to the those found in the population/sample of interest. We conducted further data processing before proceeding with our SVM and t-SNE analyses. In the ideal scenario the top-500 of each tissue sample within the same tissue class will contain the exact same set of donor locations. Because this is not the case in practice, we must subset the top-500 data for each tissue sample so that all observations of the subset contain valid data. Following the methodology of the top-500 selection we will take the top-100 most frequent splice junction donors from the population of each tissue class in the hope that each of these donor locations will be present in the top-500 splice junction donors of each tissue sample in this class.

Figure 1. 'Muscle – Skeletal' population empirical pdf



After drawing the top-100 from each tissue class population and ensuring that all tissue samples within the tissue class have values for this top-100, we combine all the top-100 population vectors to allow for overlap of observed tissue sample quantification of a top-500 donor location from a different tissue class than the observation without requiring this to exist. We will term this set of donor locations that are present in at least one top-100 subset drawn from one of the tissue classes as the "donor variables". We then compare this set of donor variables to every tissue sample to achieve a set of vectors containing the number of splice junctions observed for each of the donor variables for every sample tissue. The dataset is thereby constructed such that each row represents an observed tissue sample and each column, a donor location. Each value of this table indicates the number of splice junctions observed in each sample (row) associated with a specific donor location (column). We refer to this table as the "quantification" data. The "presence" data is simply a version of the quantification data that was overwritten to include one as the cell value if the quantification value was greater than zero.

B. Preliminary Analysis

To inform our distribution analysis, we first conducted a preliminary analysis to derive descriptive statistics of each tissue class based on the samples therein. These descriptive statistics include a naïve measure of the expected value and standard deviation of each tissue class's distribution of donor sites, and a novel method of calculating intraclass deviation. For calculation of the naïve expected value of these distributions we used equation 1.

$$\bar{x} = \sum_{i=1}^{n} x_i p(x_i) \qquad (1)$$

Where x is the location of a given donor site i and $p(x_i)$ is the proportion of the population's unique splice junction donor sites accounted for by location i. Similarly, for the calculation

of the naïve standard deviation of these distributions we used equation 2.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}[x_i - \bar{x}]^2 \quad (2)$$

In equation 2 the variable x_i retains its interpretation, $\bar{x}$ is the expected value of the tissue class, and n is the number of samples observed in the class. When defining our measurement of deviation we chose to focus on the relative presence of splice junction donors between the population of a tissue class and each sample therein. Thus, we use equation 3 to derive a deviation statistic for each tissue class.

$$\sum_{i=1}^{p}\left|\frac{c_{z_i}}{\sum_{j=1}^{m} c_{y_j}} - \frac{c_{z_i}}{\sum_{k=1}^{n} c_{x_k}}\right| \quad (3)$$

In this equation, x_k and y_j represent a donor location observed in the sample and population of a tissue class respectively, where the sample has n elements and the population contains m elements. The set Z, containing z_i with p elements, is the intersection between the donor locations contained in the sample and population. The variable c_{a_b} indicates the count of splice junctions observed in association with donor location a_b. This measurement compares each of the samples to the population by finding the difference between the percentage of splice junction locations found at a given donor site in the sample and population. From this we view a measurement of similarity between each of the samples and the population in terms of the proportional differences between observed splice junction sites. In the ideal situation, each sample of the population will contain the same set of donor locations, and these donor locations will be present in the same proportion across all samples. The deviation metric of this ideal population will be zero.

C. Distribution Analysis

For this analysis we utilized two hypothesis tests to investigate the classification power of distribution-based approaches within the confines of our problem. We chose to utilize the Kolmogorov-Smirnov and Wasserstein distance tests to compare the empirical cumulative distribution functions (ECDFs) of our samples. The Kolmogorov-Smirnov test works by calculating a cumulative distribution function (CDF) from the difference of two input ECDFs, and then taking the maximum absolute value of this resulting function as the test statistic [10]. This test statistic is given in equation 4.

$$\mathrm{KS} = \max_{x\in\mathbb{R}} |\hat{F}(x) - \hat{E}(x)| \quad (4)$$

The Wasserstein distance test works in a similar way, although it compares the difference between two ECDFs by integrating between both ECDFs to find the area between them and using this area as the test statistic. The formula for the Wasserstein distance statistic is given in equation 5.

$$\mathrm{Wass} = \int_{-\infty}^{\infty} |\hat{F}(x) - \hat{E}(x)|dx \quad (5)$$

Both tests take as inputs two ECDF functions. In practice the inputs to these functions are two vectors of draws, where the aim of the test is to determine if the draws were pulled from the same distribution. To gauge the effectiveness of each of these methods in tissue classification we developed an evaluation statistic to measure the effectiveness of each test in classifying true positives across all samples within a tissue class. We derived a draw from the population of each tissue class and used both the Kolmogorov-Smirnov and Wasserstein distance methods to test if this population sample was drawn from the same distribution as each tissue sample of this class. Using a p-value of 0.05, we recorded whether the result of each of these tests correctly failed to reject the null hypothesis or incorrectly rejected the null hypothesis (that the two samples were drawn from the same distribution). Finally, we aggregated the number of correct results of each test and divided by the total number of tests conducted to find the proportion of the samples for which the null hypothesis was correctly not rejected.

D. Support Vector Machines

The support vector machine approach defines a linear boundary in a transformed version of the high dimensional input space [11] to make classification decisions. SVM is a generalization and extension of methods used to create an optimal separating hyperplane that separates distinct classes to cases where the classes may not be fully separable. SVM uses the data nearest to the class differences to determine the separation boundary, giving the method advantageous properties over methods like linear discriminant analysis (LDA) which use centroids of the classes for this purpose. SVM uses data nearest to class differences by tuning the proportional number of predictions which fall on the wrong side of the separation hyperplane [11].

We conducted two experiments using SVM to classify tissue samples based on transcriptome data. In the first we utilized the quantification data and in the second we utilized the presence data. In both experiments we used the sci-kit learn function for fitting our model and making predictions, and we left the default parameters alone; keeping the penalty term (C) equal to one and the kernel set to a radial kernel ('rbf').

E. t-SNE Visualizations

To reduce the complexity of the high-dimensional data, t-distributed Stochastic Neighbor Embedding (t-SNE) was applied to visualize the splice junction profiles across tissue types for both quantification and presence data. t-SNE uses a nonlinear dimensionality-reduction method [12] to produce a 2-dimensional visualization of samples that are mapped more closely together if there's greater pairwise similarity while dissimilar samples are placed farther apart from one another. To generate the lower-dimensional points t-SNE minimizes the Kulback-Leibler (KL) divergence between the joint probability of pairwise similarities of data in the high-dimensional space–with similarity being measured using Euclidean distance–and the joint probability of pairwise similarities of low-dimensional points. evaluated using a normalized Student's t-distribution with one degree of freedom.

III. Results & Discussion

A. Data

There are many ways to decrease the dimensionality of our data through filtering. One method is to select only splice junction sites which are observed enough times to exceed some threshold [13]. Another is to select the top-x most relevant junction sites, as has been done in other studies using genomic characteristics to classify cell type [14]. In our analysis we developed a third method of filtering. In this scheme, a splice junction is considered "relevant" if the observations of this splice junction exceed a threshold-percent of all observed splice junctions using this donor. In this way we sought to remove all splice junctions that could be considered "noise" insofar as they comprised a small percentage of all junction sites with that donor. Interestingly, this third method of filtering did not reduce the dimensionality of the data significantly enough for our purposes. This is elaborated in the conclusions section.

B. Preliminary Analysis

The results of our preliminary analyses can be viewed in table 1. The figures for naïve expected value and standard deviation refer to the distributions of donor sites in the populations of each tissue class, whereas the figures for deviance utilize the relative proportions of donor locations between each sample and the population of a given tissue class.

We can see from these descriptive statistics that tissues from similar regions share characteristics of their distributions of donor sites. Each of the brain tissue classes has a similar naïve standard deviation, which are approximately an order of magnitude smaller than those of "Muscle - Skeletal" and "Whole Blood". We can see the similarity of tissue also in the two closest tissue classes-"Brain - Caudate (basal ganglia)" and "Brain - Nucleus accumbens (basal ganglia)"-which share a similar naïve standard deviation and expected value of their distributions of donor locations. We can see the deviance metric vary widely between tissue classes, and the interesting observation that the two closest tissue classes-"Brain - Caudate (basal ganglia)" and "Brain - Nucleus accumbens (basal ganglia)"-have the same value for deviance.

C. Distribution Analysis

The results of the Kolmogorov-Smirnov test can be seen in table 1. The results of the Wasserstein distance test were not significantly different from those of the Kolmogorov-Smirnov test, so we have not reported them separately.

To evaluate the Kolmogorov-Smirnov test as a metric for tissue classification, each value of the table for the column "KS test" represents the percentage of Kolmogorov-Smirnov tests between the ECDF of the tissue class's population and each sample of this tissue class where the null hypothesis was correctly not rejected (the Kolmogorov-Smirnov test correctly recognizes that the two ECDFs are from the same distribution). We can see that the highest achieved value of this statistic is found for the tissue class "Brain - Spinal cord (cervical c-1)," where 70.59% of the tissue samples whose true class was "Brain - Spinal cord (cervical c-1)" were correctly classified to be a sample of the population of this class by the Kolmogorov-Smirnov test. Because the Kolmogorov-Smirnov and Wasserstein distance tests have poor true positive classification we can determine that using distribution comparisons methods are insufficient for transcriptome-based tissue classification.

Another interesting feature of the distribution analysis results is the corroboration between the KS test figure and the preliminary analysis descriptive statistics. We can see from the preliminary analysis descriptive statistics that the standard deviation of the brain tissue classes are approximately one order of magnitude smaller than those of the "Muscle - Skeletal" and "Whole Blood" tissue classes. Likewise, the deviance of the two brain tissue classes "Brain - Caudate (basal ganglia)," "Brain - Nucleus accumbens (basal ganglia)" and "Muscle - Skeletal" are very similar (4.23, 4.23, 4.06) and higher than those for the tissue classes "Whole Blood" and "Brain - Spinal cord (cervical c-1)" (1.00, 1.37). We see that where the relative deviance is high and the standard deviation is high, we have the worst KS test value (tissue class "Muscle - Skeletal", 12.86%) and where the relative deviance is low and the standard deviation is low we have the highest KS test value (tissue class "Brain - Spinal cord (cervical c-1)", 70.54%). This predictable and interpretable variation of the computed KS test increases our confidence that the distribution analysis techniques were developed and implemented to the extent of their capabilities within this formulation of the tissue classification problem.

D. Support Vector Machines

The results of the support vector machine approaches were surprisingly good and corroborated by the results found in [15] where the SVM approach was used to great effect on the GTEx dataset. We can see from other work toward classification of tissue samples using transcriptome data found in the GTEx dataset that great classification results can be achieved [16] and, in comparison to [16], we have achieved similar results using a smaller sample space and without using deep learning methods.

We can see the results of the SVM model using both quantification and presence data summarized by the confusion matrices in Fig 3, whose labels correspond to those in Table 1. We can see from these plots that the performance of both SVM approaches is the same, with both approaches resulting in a

Table 1. Preliminary and distribution analysis results

Descriptive Statistics by Tissue Class	**Statistic**			
	Expected Value	***Standard Deviation***	***Deviance***	***KS Test***
Muscle – Skeletal (1)	1.74 e8	1.19 e19	4.06	12.86 %
Whole Blood (2)	1.13 e8	1.35 e19	1.00	25.54 %
Brain - Spinal cord (cervical c-1) (3)	1.15 e8	8.28 e18	1.37	70.59 %
Brain - Caudate (basal ganglia) (4)	1.05 e8	9.29 e18	4.23	43.22 %
Brain - Nucleus accumbens (basal ganglia) (5)	1.03 e8	9.37 e18	4.23	37.72 %

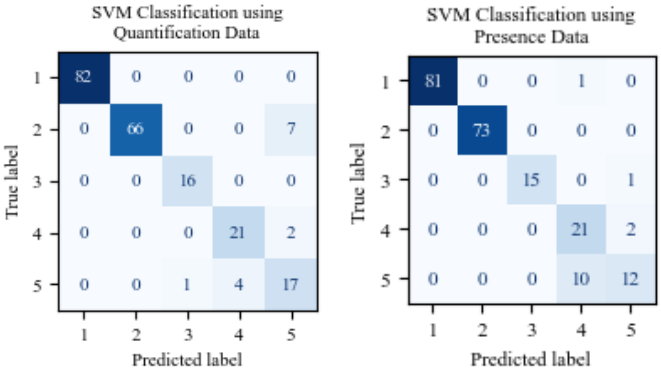
prediction accuracy of 93.5%. It is interesting to note that the misclassifications of these two models do not follow the same pattern. Although the misclassification rate is higher in the SVM case using presence data, the misclassifications mimic underlying biological intuition. We expect tissue classes with similar biology to be "closer" to one another in the latent space of the decision than tissue classes with dissimilar biology, thereby causing the most misclassifications between tissue classes that are the most similar biologically. We see this expectation realized in the SVM experiment using presence data where 85.7% of misclassifications are between tissue types within the same region of the brain (between tissue classes "Brain - Caudate (basal ganglia)" and "Brain - Nucleus accumbens (basal ganglia)"). For the SVM experiment using quantification data many misclassifications occur between dissimilar tissue types, indicating that the latent space of the SVM decision boundary may not resemble the true decision boundary of the data i.e. this SVM model uses features of the quantification data other than those which determine biological similarity to make its decisions.

E. t-SNE Visualizations

We can observe the t-SNE clusters for both the quantification and presence data in figure 4. In these plots we observe innocuous curvature in the t-SNE plot for quantification data which we can interpret as a component of "distances between clusters might not mean anything" detailed in [17]. Despite difference in cluster shapes between plots, it is evident from both that the two classes which are most similar biologically ("Brain - Caudate (basal ganglia)" and "Brain - Nucleus accumbens (basal ganglia)") cannot be accurately split into different clusters by the t-SNE algorithm in this setting.

Another interesting feature of these results is that the tissue class with small deviance and standard deviation from our preliminary analysis ("Brain - Spinal cord (cervical c-1)") exhibited the most compact cluster in the t-SNE visualization. Similarly, the class with high deviance and standard deviation from our preliminary analysis ("Muscle - Skeletal") exhibits the most spread in its cluster in the t-SNE visualization.

Figure 2. SVM Prediction results

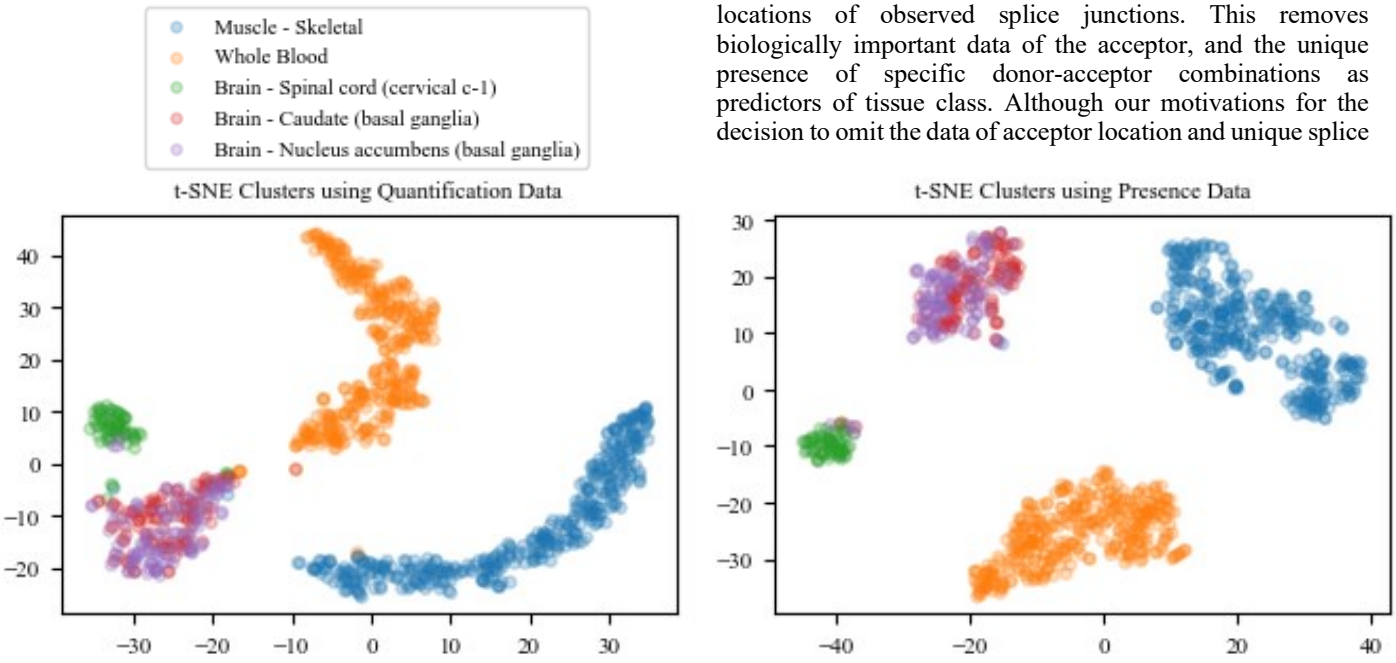


F. Limitations

The common limitation of the specified approaches lies in the reduced dimensionality of the input space when compared to the source data. In the distribution analysis section, we might have specified larger sample sizes or chosen the top-x number of most frequent splice junction sites to analyze, thereby increasing the dimensionality of the data used in our analysis. We might have also used data other than non-chimeric chromosome 1 splice junctions. All of these changes have the potential to increase the accuracy of our distribution analysis methods, though they may simply increase the noise evident in the data. An increase of the dimensionality of the input space might have similarly helped the SVM and t-SNE approaches, with the same tradeoff being increased noise in the data used.

Also, our analyses were conducted using the donor locations of observed splice junctions. This removes biologically important data of the acceptor, and the unique presence of specific donor-acceptor combinations as predictors of tissue class. Although our motivations for the decision to omit the data of acceptor location and unique splice

Figure 3. t-SNE clustering results



junction pair were founded in the preliminary analysis of the paper, we recognize that in methods that specialize in high dimensional feature recognition the addition of this data will likely increase model performance.

As a result of both choices we have considered only a small subset of the available data for this classification problem, thereby limiting the classification accuracy of both the SVM and t-SNE models described in this paper.

IV. Conclusion

A. Data

In our development of a donor-level relevance threshold for filtering out irrelevant splice junctions we found that this method did not significantly decrease the dimensionality of our problem. Although there is some skew to the tissue empirical distributions, we can understand this small reduction in the dimensionality to be caused by many unique splice junction sites in its tails. Because these splice junction sites have a relatively small number of total observations pertaining to a unique donor, it is easier for the unique variants that use the same donor to pass the relevance threshold. Thus, we can infer that there are many unique and infrequently occurring splice junction sites that comprise most of the empirical distribution. A deep analysis of what constitutes "noise" in the tails of this distribution is a promising research direction that would probe the heart of transcriptome-based connections to biological phenotype.

B. Extensions

For the SVM approach, the limitation of acceptor inclusion may-to some extent-be avoided by decreasing dimensionality through considering fewer tissue classes and tuning the cost and kernel parameters to increase model performance on a smaller subset of the data. Likewise, the t-SNE method evaluated in this paper might be extended by increasing dimensionality to include a larger top-x number of donor locations, including acceptor locations, or including some amount of data from other chromosomes, so that the t-SNE algorithm has more high dimensional data to work with.

C. Embedding

As we observed from the results of the SVM approaches, embedding techniques that utilize only the presence of a donor location are sufficient for the accurate classification of tissues. Our t-SNE visualizations show that while there are clean clusters across histological types (i.e. brain vs. muscle vs. whole blood), there is little separability among the different subcategories of brain tissue (spinal cord vs. caudate vs. nucleus accumbens). This result may be due to our filtering method that used donor site frequency as the criteria for inclusion. By filtering our data at the donor site level, we selected the splice junctions that correspond to the primary or most expressed transcript. Primary transcripts are cell-type-specific for non-brain tissues, but the cellular specificity for brain tissues is typically driven by alternative splicing patterns [8]. This helps to explain why our t-SNE plots do not show distinction among the brain tissue types and a different filtering method may better capture the slicing variability within the brain.

References

[1] CRICK, F. Central Dogma of Molecular Biology. Nature 227, 561–563 (1970). doi: doi.org/10.1038/227561a0

[2] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-seq," Nature Methods, vol. 5, no. 7, pp. 621–628, 2008.

[3] I. H. Wei, Y. Shi, H. Jiang, C. Kumar-Sinha, and A. M. Chinnaiyan, "RNA-seq accurately identifies cancer biomarker signatures to distinguish tissue of origin," Neoplasia, vol. 16, no. 11, pp. 918–927, 2014.

[4] A. Jabeen, N. Ahmad, and K. Raza, "Machine learning-based state-of-the-art methods for the classification of RNA-Seq Data," 2017.

[5] Warren, C.F.A., Wong-Brown, M.W. & Bowden, N.A. BCL-2 family isoforms in apoptosis and cancer. Cell Death Dis 10, 177 (2019). doi: 10.1038/s41419-019-1407-6

[6] N. T. Johnson, A. Dhroso, K. J. Hughes, and D. Korkin, "Biological classification with RNA-seq data: Can alternatively spliced transcript expression enhance machine learning classifiers?," RNA, vol. 24, no. 9, pp. 1119–1132, 2018.

[7] C. J. Labuzzetta, M. L. Antonio, P. M. Watson, R. C. Wilson, L. A. Laboissonniere, J. M. Trimarchi, B. Genc, P. H. Ozdinler, D. K. Watson, and P. E. Anderson, "Complementary feature selection from alternative splicing events and gene expression for phenotype prediction," Bioinformatics, vol. 32, no. 17, pp. i421–i429, 2016.

[8] M. Melé, P. G. Ferreira, F. Reverter, D. S. DeLuca, J. Monlong, M. Sammeth, T. R. Young, J. M. Goldmann, D. D. Pervouchine, T. J. Sullivan, R. Johnson, A. V. Segrè, S. Djebali, A. Niarchou, T. G. T. E. Consortium, F. A. Wright, T. Lappalainen, M. Calvo, G. Getz, E. T. Dermitzakis, K. G. Ardlie, and R. Guigó, "The human transcriptome across tissues and individuals," Science, vol. 348, no. 6235, pp. 660–665, 2015.

[9] Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics (Oxford, England), 29(1), 15–21. doi: 10.1093/bioinformatics/bts635

[10] C. Dowd, "A New ECDF Two-Sample Test Statistic." arXiv, Jul. 02, 2020. Accessed: Apr. 12, 2023. [Online]. Available: http://arxiv.org/abs/2007.01360

[11] T. Hastie, R. Tibshirani, J. Friedman, "Support Vector Machines and Flexible Discriminants," in The Elements of Statistical Learning, 2nd Ed., New York, NY USA: Springer, 2009, ch. 12, pp. 417–425

[12] G. Hinton and S. Roweis, "Stochastic Neighbor Embedding," in NeurIPS, 2002, pp. Xxx-xxx, [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2002/file/6150ccc6069bea6b5716254057a194ef-Paper.pdf

[13] M. Farlik, N. C. Sheffield, A. Nuzzo, P. Datlinger, A. Schönegger, J. Klughammer, C. Bock, "Single-Cell DNA Methylome Sequencing and Bioinformatic Inference of Epigenomic Cell-State Dynamics," Cell Rep., vol. 10, no. 8, pp. 1386–1397, Mar. 2015, doi: 10.1016/j.celrep.2015.02.001.

[14] A. G. Martini, J. P. Smith, S. Medrano, N. C. Sheffield, M. L. S. Sequeira-Lopez, and R. A. Gomez, "Determinants of renin cell differentiation: a single cell epi-transcriptomics approach," Developmental Biology, preprint, Jan. 2023. doi: 10.1101/2023.01.18.524595.

[15] J. Li, L. Chen, Y.-H. Zhang, X. Kong, T. Huang, and Y.-D. Cai, "A Computational Method for Classifying Different Human Tissues with Quantitatively Tissue-Specific Expressed Genes," Genes, vol. 9, no. 9, p. 449, Sep. 2018, doi: 10.3390/genes9090449.

[16] M. Yap, R. L. Johnston, H. Foley, S. MacDonald, O. Kondrashova, K. A. Tran, K. Nones, L. T. Koufariotis, C. Bean, J. V. Pearson, M. Trzaskowski, N. Waddell, "Verifying explainability of a deep learning tissue classifier trained on RNA-seq data," Sci. Rep., vol. 11, no. 1, p. 2641, Jan. 2021, doi: 10.1038/s41598-021-81773-9.

[17] W. Martin, V. Fernanda, and J. Ian, "How to Use t-SNE Effectively," Distill, 2016, doi: 10.23915/distill.00002