Propensity of Endogenous Alternative Splicing to Mediate Mutative Damage

By:

Ashley Lutz

A Thesis
Submitted to the Faculty
Of

Worcester Polytechnic Institute

In partial fulfillment of the requirements for the

Degree of Masters of Science

In

Bioinformatics and Computational Biology

May 2019

APPROVED:

_____
Dr. Dmitry Korkin, Major Advisor

_____
Dr. Amity Manning, Reader

Abstract:

The identification of alternative splicing in the human genome elucidated the potential to several enduring genomic questions. Not only could this phenomenon explain why organism complexity was not at all correlated with the genome size, or explain how an organisms could be affected by experience and environment at the molecular level, but it was perhaps the most flexible and adaptive regulatory mechanism identified to date. While the pathogenic aberrations of this mechanism have generally been readily investigated and identified as potential therapeutic targets, its meditative or advantageous instances have largely not been considered. Initiated exon skipping has been shown to have therapeutic effects in Muscular Dystrophy animal models and even in vitro human muscle cells (Aartsma-Rus, Annemieke, et al, *Human Molecular Genetics* 2003, McClorey, G., et al, *Neuromuscular disorders*, 2006). However, the consideration that this process may be occurring endogenously in human cells and contributing to other complex diseases has remained largely ignored. In this work, we have undertaken the first large-scale statistical examination of alternatively spliced variants between the tissues of diseased and normal patients. We hypothesize that there are endogenous alternative splicing events occurring in these tissues that purposefully mediate mutative damage and contribute to the differentiation between diseased and healthy phenotypes. By integrating data from several different sources and employing statistic and machine learning models, we have identified significant differences in gene characteristics between canonical and spliced variants correlated with changes in clinical outcomes. We conclude that this evidence supports our hypothesis that alternative splicing can be positively driven to mediate genetic damage. Expression of these genetically damaged and canonically spliced variants is clearly implicated in diseased tissue and poor clinical outcomes.

Acknowledgements:

My time at WPI has allowed me the skills and flexibility to learn and explore what I truly want to do in the field of bioinformatics and computational biology. I want to thank all of the faculty I have been fortunate enough to learn from through my coursework and the PI's I have been lucky enough to work under. In particular I would like the thank Dr. Dmitry Korkin for his guidance and mentorship on this thesis and for allowing me the opportunity to explore a field of bioinformatics I had not otherwise been able to. I would also like to thank Dr. Amity Manning for being a reader and providing her expert feedback and Dr. Joseph Duffy, Dr. Liz Ryder and Dr. Jill Rulfs for supporting me through a TA ship for 2 years, allowing me to pursue this degree full time. Finally I would like to thank my fellow BCB students and Korkin lab-mates for their support and feedback throughout the last 2 years. I have very much enjoyed my time at WPI and know my experience here will continue to aid me in my professional career.

Table of Contents:

List of Figures

List of Tables

1. INTRODUCTION

Since the conclusion of the human genome project it has become apparent that the initial perception of the genetic code as a simple blueprint for the resulting organism was woefully inadequate. The genome is a dynamic system, constantly being manipulated and reorganized by various regulatory mechanisms of all forms and etiologies. Of these mechanisms one of the most pervasive yet least well understood is alternative splicing. Initial estimates proposed that 40-60% of genes within the human genome had alternative splice forms (Barmak et al, *Nature* 2002). However more recent investigations using specific mRNA microarrays have reported that as many as 75% of human multi-exon genes exhibit alternative splicing (Johnson et al, *Science* 2003).

Similarly, the effects of alternative splicing are numerous and varied; ranging from complete functional devastation, to only minor implications/functional changes, to acquisition of advantageous functionality (Stamm et al *Gene*, 2005). It has even been proposed that alternative splicing explains the incongruity between complexity of an organism and the size of its genome (Nilsen et al, *Nature* 2010). Despite the clear evidence of the implicit advantageous nature of alternative splicing, most in depth examinations of incidences of alternative splicing focus on its implications in disease. A study examining the range in effects of alternative splicing found that while less frequent than whole domain alterations, within protein alternative splicing events tend to occur significantly more often in functional domains, suggesting that alternative splicing is generally a positively reinforced mechanism (Kriventseva et al *Trends in Genetics*, 2003).

The first experimental evidence of this phenomena was discovered by researchers at the University of Western Australia who were researching muscular dystrophy through a *mdx* mouse model. They noticed the *mdx* mice they were using as a model for muscular dystrophy, a

disorder caused by a single nonsense mutation in exon 23 of the dystrophin gene resulting in a nonfunctional protein, still had some amount of functional dystrophin being produced. They theorized that exon 23 might occasionally be spliced out of transcriptions of the gene resulting in the functional proteins. They tested this theory by exposing myoblasts to oligonucleotides in order to induce this specific exon skipping splice and were able to obtain high concentrations of functional, although shorter, dystrophin coding sequences (*Wilton et al, 1999*). Since then their success in induced exon skipping in muscular dystrophy models has been replicated across other animal models and even in vitro human muscle cells (Aartsma-Rus, Annemieke, et al, *Human Molecular Genetics* 2003, McClorey, G., et al, *Neuromuscular disorders*, 2006). However, investigations into whether or not the same phenomenon is occurring in human subjects has been slow if not nonexistent, essentially ignoring a proven possible therapeutic method for several different types of diseases.

In the following study we will detail a thorough statistical analysis of alternative splicing and mutation events in breast cancer which support the hypothesis that alternative splicing can be used as an advantageous mechanism for mediating mutative damage.

2. METHODS

2.1 Hypothesis and Process

We hypothesize that positively driven alternative splicing is occurring within the human genome. Specifically, we will investigate the propensity of this mechanism to mediate pathogenic mutations. In order to examine this hypothesis, we will compare several characteristics; expression, mutation annotation, and clinical outcome, between canonical and corresponding spliced variants. If our hypothesis is true we would expect to find advantageous differences in

favor of the spliced variants and disadvantageous differences in the corresponding canonical variants.

For the majority of the investigation we examined primary the canonical variants for disadvantageous and potentially pathogenic differences from their corresponding spliced. However, these examinations were corroborated with complementary examinations of the corresponding spliced variants whenever possible.

2.2 Methodology Overview

The data used throughout this study was obtained through open source genetic and medical data portals, namely The Cancer Genome Atlas (TCGA), Ensembl release 96, University of California Santa Clara Genome Browser (UCSC), and the Catalogue of Somatic Mutations in Cancer (COSMIC). Data was merged between all sources via TCGA barcode ID as well as gene name and Ensembl Transcript and UCSC Variant ID's (Figure 1). In the following sections the methodology by which this large data pool was examined and narrowed down to the resulting subset of interest will be detailed. Intermediate data subsets and results were given distinct names for clarity (Table 1).

2.3 Variant Expression Data Extraction

Normalized mRNA expression data for the Breast Invasive Carcinoma (BRCA) cohort was pulled from The Cancer Genome Atlas Firebrowse data portal. The expression data consisted of RSEM normalized transcripts per million (TPM) for approximately 73000 variants from 1212 patient samples. Samples were identified by their unique TCGA barcode. Encoded within this barcode is a numerical division of tissue type and sample condition. All samples were obtained

from the breast tissue of a patient suffering from breast invasive carcinoma. However, there were 3 different sample conditions; Primary Solid Tumor, Solid Tissue Normal, and Metastatic. Specifically, the dataset contained 1093 Primary Solid Tumor samples, 112 Solid Tissue Normal samples, and 7 Metastatic, samples. Due to the significantly imbalanced number of metastatic samples they were disregarded from for future analysis. Further analysis continued with the Primary Solid Tumor and Solid Normal Tissue samples heretofore referred to as tumor samples and normal samples respectively.

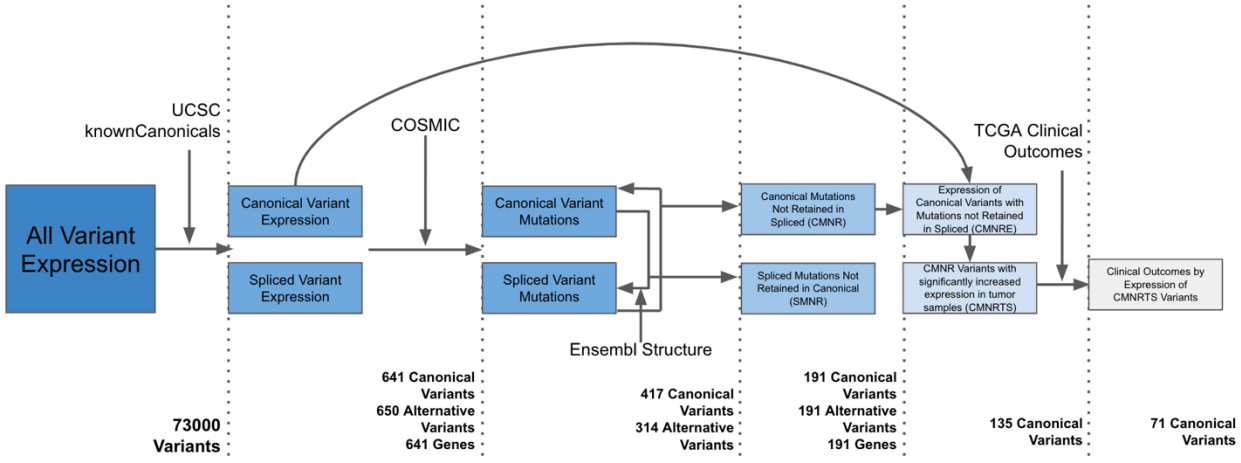


Figure 1. Pipeline of Experiment Methodology: The integration of each data source was performed in tandem with experimentation and resulting in reduction of the size of the data of consideration. Initially the data containing roughly 73000 variants. Following the integration of the mutation database this number was reduced to 1219. After a structural comparison and statistical examination of mutation frequency this number was reduced to 191 canonical and spliced variant pairs. Next a statistical examination of the expression level of these variants reduced this count to only 135 variant pairs with significant differences. Finally, the clinical data was integrated into the. Remaining 135 variants yielding only 71 variants of interest.

Table 1. Data Structures Resulting from the Integration of Original Data Sources: As the data sources were integrated and the subset of data of interest was reduced and evolved through the statistical processes various intermediate data structures were constructed. A general summary of the data structures as well as the names and abbreviations if applicable are provided below. As is the fundamental construction of each data structure. Data structures (DS) are referenced as S: string, DF: DataFrame, L: List.

Term	Composition	DS	Rows	Columns
Canonical Variant	Variant identified by UCSC as the knownCanonical for the corresponding gene	S	N/A	N/A
Spliced Variant	Variant of a gene that is not the corresponding variant. A single gene may have many spliced variants but within this investigation there is generally only 1 spliced variant per gene.	S	N/A	N/A
All Variants with Expression	TPM levels for approximately 73000 variants from obtained from either normal or tumor tissue samples from 1205 patients.	DF	Sample (TCGA barcode)	TPM of Variant (Variant IDs)
Canonical/ Spliced Pairs	650 pairs of canonical and spliced variants of the same gene in which both exist in all variants with expression and therefore have expression data	L	N/A	N/A
Canonical Variant Expression	TPM expression in 1205 samples for all canonical variants of canonical/spliced pairs	DF	Sample (TCGA barcode)	Canonical Variant IDs
Spliced Variant Expression	TPM expression in 1205 samples for all spliced variants of canonical/spliced pairs	DF	Sample (TCGA barcode)	Spliced Variant IDs
Canonical Variant Mutations	Mutation Annotation for every mutation which occurs in a canonical variant of the canonical/spliced pairs	DF	Mutation incidences in Canonical Variants	Variant Mutation Occurs In, Mutation Loci, FATHMM Prediction of Mutation, Corresponding Variant from Canonical/Spliced Pairs, Exon start and stop Loci of corresponding variant
Splice Variant Mutations	Mutation Annotation for every mutation which occurs in a spliced variant of the canonical/spliced pairs	DF	Mutation incidence Spliced Variants	Variant Mutation Occurs In, Mutation Loci, FATHMM Prediction of Mutation, Corresponding Variant from Canonical/Spliced Pairs, Exon start and stop Loci of corresponding variant
Canonical Mutations Not Retained in Spliced (**CMNR**)	Canonical Mutations that were not retained in their corresponding spliced variants	DF	Mutation incidences in Canonical Variants that do not fall within any exon of the corresponding spliced variant and are therefore not retained	Variant Mutation Occurs In, Mutation Loci, FATHMM Prediction of Mutation, Corresponding Variant from Canonical/Spliced Pairs, Exon start and stop Loci of corresponding variant
Spliced Mutations Not Retained in Canonical (**SMNR**)	Spliced Mutations that were not retained in their corresponding canonical variants	DF	Mutation incidences in Spliced Variants that do not fall within any exon of the corresponding canonical variant and are therefore not retained	Variant Mutation Occurs In, Mutation Loci, FATHMM Prediction of Mutation, Corresponding Variant from Canonical/Spliced Pairs, Exon start and stop Loci of corresponding variant

Canonical Variants with Mutations Not Retained in Spliced with Expression (**CMNRE**)	Expression of canonical variants that contain mutations not retained in corresponding spliced	DF	Sample (TCGA barcode)	TPM of CMNR variants (Variant IDs)
Canonical Variants with Mutations Not Retained in Spliced with Significantly Increased Expression in Tumor Samples (**CMNRTS**)	CMNR variants with significantly increased expression in tumor samples compared to normal	DF	Sample (TCGA barcode)	TPM of CMNRTS variants (Variant IDs)
Canonical Variants with Mutations Not Retained in Spliced with Significantly Reduced Survival (**CMNRTSRS**)	CMNRTS variants for which expression greater than or equal to the third quartile of all samples for that variant is related to significantly reduced survival	DF	Sample (TCGA barcode)	'High' or 'Low' depending on whether the TPM of the specific variant is greater than or equal to or less than the third quartile of expression for that variant overall.

2.4 Identification of Canonical/Spliced Pairs

All variants with expression were identified by their unique UCSC IDs. The UCSC genome browser conveniently contains a "KnownCanonicals" table, accessible through their publicly available Table Browser. This table was queried for the IDs of all the variants with expression yielding a two-dimensional table of the submitted variants, heretofore referred to as spliced variants, and their corresponding canonical variant for that gene. Rows in which the submitted variant was also the canonical variant were eliminated from the table. Additionally, each obtained canonical variant was also compared to a list of the original 73,000 variants in the expression data. If the canonical variant was not found in the list there was no expression data for this variant and all rows containing this canonical variant were removed. Following these reductions 650 canonical/spliced pairs remained in the table. These pairs included 1,291 individual variants; 641 unique canonical variants and unique 650 corresponding spliced variants. The expression file was then reduced to include only expression levels from all samples for these 1,291 variants and then split by variant type creating the canonical variant expression and spliced variant expression tables. The canonical/spliced pairs were then converted to a

dictionary (canonical/spliced dictionary) as well as two distinct lists containing only the canonical variants or spliced variants respectively (canonical variants and spliced variants).

2.5 Association of Annotated Mutations with Variants

Annotated mutation data was obtained via the Catalogue of Somatic Mutations in Cancer (COSMIC) breast carcinoma dataset (Simon A. Forbes, et al *Nucleic Acids Resea*rch 2017). The dataset contained 260,303 mutations between 19,203 genes and 26,402 variants. This dataset was then compared to the canonical variants and spliced variants lists. Only mutations associated with a variant contained in either of these lists was retained, resulting a dataframe of 8,236 mutations each with various annotated features including mutation loci and Functional Analysis Through Hidden Markov Models (FATHMM) prediction. Of the 641 canonical and 650 corresponding spliced variants with expression data, 417 canonical variants also had annotated COSMIC mutations as did 314 of the corresponding spliced variants. The mutations were grouped by their corresponding variant type creating the canonical variant mutations and spliced variant mutations tables.

For this investigation we were particularly interested in how mutations were retained or removed by splicing events. In order to assess whether or not mutations could exist in variants they were not annotated in, in this case the corresponding variant of the canonical/splice pair, the loci of all exon start and stops for each variant were obtained through the Ensembl Biomart, resulting in a structural data table consisting of each variant ID and all exon start and stop loci for that variant.

Both of the canonical variant mutations and spliced variant mutations files were merged with the structural data for their corresponding variants. For each mutation it was determined

whether or not that mutation loci was within an exon of the corresponding variant. If the mutation existed in the corresponding variant it was designated as retained, and if it was not present in any exons the mutation was designated as not retained. For the purposes of this investigation the mutations that were not retained in the corresponding spliced variant were saved for further analysis in the canonical mutations not retained (CMNR) and spliced mutations not retained (SMNR) tables.

2.6 Integration of Clinical Outcomes

Like the expression data, the clinical data was also obtained through the TCGA Firebrowse portal (Broad Institute TCGA Genome Data Analysis Center, 2016) Each sample of the original TCGA mRNAseq dataset was assigned a unique TCGA barcode ID. This was the same barcode that was used to differentiate between the sample conditions in section 2.2. The ID was composed of seven parts, each indicating some aspect of the sample's nature or processing. A four-digit unique identifier was used to designate the patient ID. This patient ID was extracted from all samples within the expression data and was matched to the patient ID of each patient within the clinical data file. All patients whose samples were in the expression data had corresponding canonical data.

The clinical data contained several features of interest. This investigation focused primarily on survival analysis and thusly isolated only a few columns from the clinical datafile including, days to death, days to last follow up, and survival status. There are other columns of interest within the data file that may be used in future examinations such as stage and days since diagnosis.

3. Results

3.1 Comparative Examination of Mutations in Canonical and Spliced Variants

Of the 417 canonical variants and 314 spliced variants of consideration there were proportionately more annotated mutations associated with canonical variants than spliced with 4,972 and 3,264 mutations, respectively. Similarly, the proportion of canonical variant mutations retained in the corresponding spliced variants is significantly greater than the proportion of spliced variant mutations maintained in the corresponding canonical when examined through a one tailed Fisher's exact test with an alpha of 0.05 ($p = 0.0013$).

Of the mutations which were removed in the corresponding variant, canonical mutations removed in spliced variants had a significantly higher proportion of pathogenic mutations than the spliced mutations removed in canonical mutations, according to their FATHMM prediction when examined through a Fisher's Exact test with an alpha of 0.05 ($p= 0.0009$) (Figure 2.A). Interestingly, the distribution of the various types of mutations was not significantly different between the canonical mutations removed in splice and spliced mutation removed in canonical (Figure 2.B,C) . However, upon further examination, the SMNR variants did have an increased proportion of frameshift insertions and deletions, also known as the loss-of-function mutations (Figure 2.D). This suggests that the main determining factor for a mutation to be removed from the canonical in the spliced version is the degree of damage. In order to investigate this hypothesis further, all proceeding analysis was done with regard to the CMNR subset of variants, which have been shown to contain significantly more pathogenic mutations. This subset contains 218 different genes, however only 191 of these genes had expression data for both the canonical and alternative variant, heretofore referred to as the Canonical Mutation not Retained with

Expression data set (CMNRE). These genes range in family and function from the widely implicated BRCA1 gene to the generally innocuous homeobox gene POU2F1.

3.2 Statistical Examination of Expression Levels

3.2.1 Statistical Examination of Expression Levels of Genes

The expression levels for each of the CMNRE variants and their corresponding spliced variants was extracted from the original variant expression file. The CMNRE variants were first expressed on the gene level. For each gene within the subset there were 2 comparison conditions to examine, the expression of the canonical variant versus the spliced variant and the expression level in tumor vs. normal samples. In order to examine these differential expressions efficiently a two-way ANOVA was performed for each gene, with the variant type as the between group factor and sample type as the within group factor. 184 of the 191 genes examined reported having significant difference through the ANOVA.

Table 2. Difference in Proportion of Pathogenic Variants within CMNR and SMNR Subsets: CMNR variants have not only a larger number of mutations than the corresponding spliced dataset but of these mutations there is a significant difference in the proportion of pathogenic mutations between these subsets when examined by a Fisher's exact test.

	Pathogenic (FATHMM Prediction)	Neutral (FATHMM Prediction)
Canonical Mutations not in Spliced Variant (CMNR)	391	328
Spliced Mutations not in Canonical Variant (SMNR)	174	222
		p = 0.009

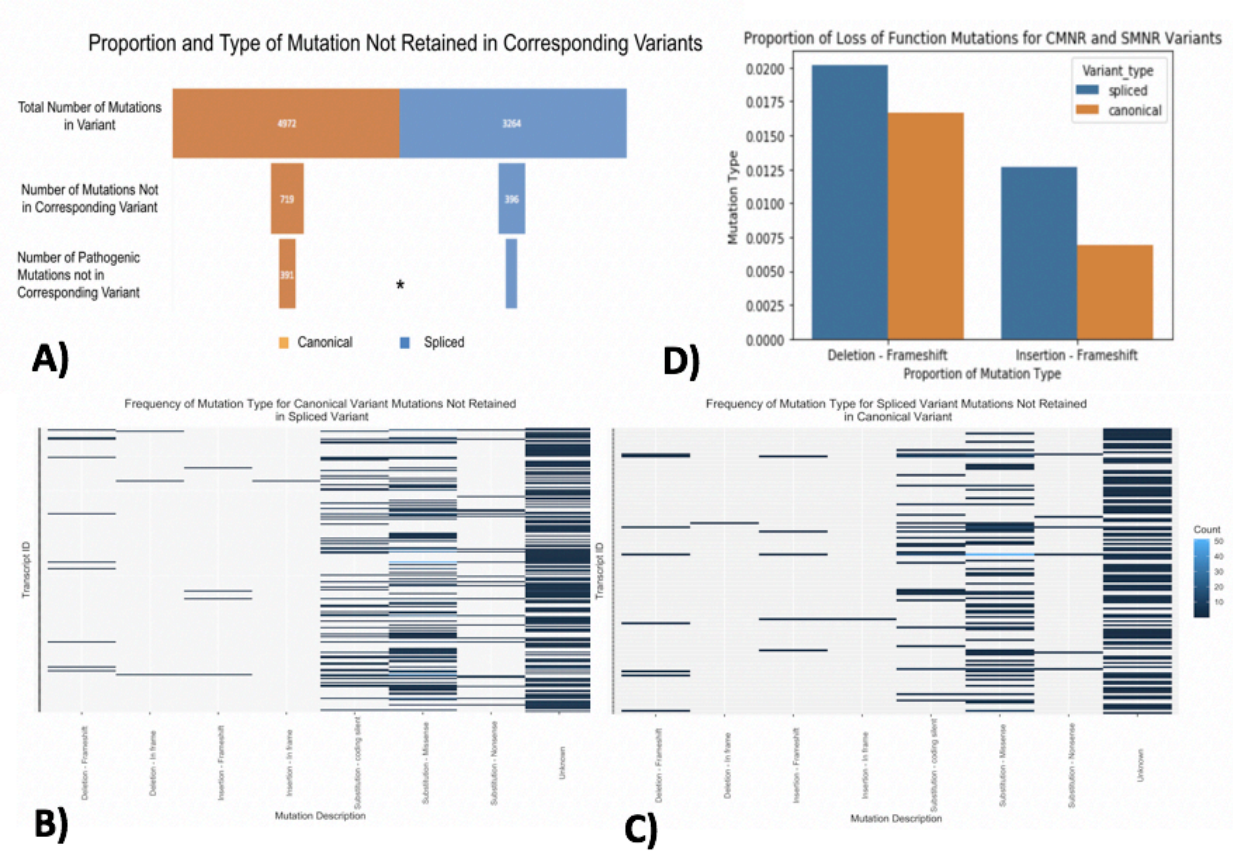


Figure 2. Stratification of mutation retention between variant types and the composition of mutations for each variant. A) The number of total mutations associated with each Variant type is shown first followed by proportion of these mutation which are not retained in the corresponding variant and therefore associated with the CMNR and SMNR datasets. Lastly the proportion of the mutations not retained in the corresponding variant which are pathogenic according to their FATHMM prediction. **B)** The frequency of each mutation type among the CMNR mutations. **C)** The frequency off each mutation type among the SMNR mutations. **D)** The amount of loss of function mutations, that is mutations which result in a nonfunctional protein contained in the CMNR and SMNR datasets.

To accurately determine which group of factors was responsible for the significant difference two Tukey HSDs were also performed. When applied to the between group variables, canonical variant expression compared to spliced variant expression, of the 191 genes 178 were found to have significantly different expression *(alpha = 0.05).* Additionally, when examining the within group variables, tumor sample vs normal sample, 123 genes had significantly different

expression between the two groups *(alpha =0.05)*. Finally, 116 genes had a significant difference between both factors *(alpha= 0.05)*.

This finding is particularly interesting as it suggests that just as the CMNRE subset contained significantly different amount of pathogenic mutations than their spliced correspondents, they also have significantly different expression levels between tumor and normal samples. This suggests that there is an intrinsic relationship between the presence of either variant and the health of the tissue and prognosis of the patient.

3.2.2 Statistical Examination of Expression Levels of Variants

In order to further elucidate the nature of the relationship between expression of the CMNRE variants, which have already been found to have evidence of damaging characteristics, a one tailed t-test was performed on the expression level of the canonical variants. Of the 191 canonical variants within the subset of interest, 135 had significantly greater expression in tumor samples than in normal samples *(alpha = 0.05)* (Figure 3). The majority of CMNRE variants had significantly higher TPM levels in tumor samples, which suggests that there is something inherently reparative in the way their corresponding spliced variants are being transcribed.

In order to validate these findings, the opposite one-tailed T test was run on the expression levels of the corresponding spliced variants of the CMNRE subset. Of the 191 spliced variants, 133 had significantly decreased expression levels in tumor versus normal samples *(alpha = 0.05)*. Additionally, of the 191 CMNRE canonical/spliced variant pairs examined, 105 had significantly increased canonical expression and significantly reduced spliced expression in tumors. These results reinforce the previous findings, suggesting that there is a strong relationship between expression of CMNRE spliced isoforms and health. This indicates that the

endogenous signaling responsible for the revision of the canonical variant into the corresponding spliced variant is a form of mediating or repairing the pathogenic mutative damage identified previously. In order to investigate this assertion in more detail, the subset of 135 CMNRE genes that were found to have significantly increased expression in tumor samples was isolated for further analysis and will heretofore be referred to as the Canonical Mutation Not Retained Tumor Significant (CMNRTS) subset and table.

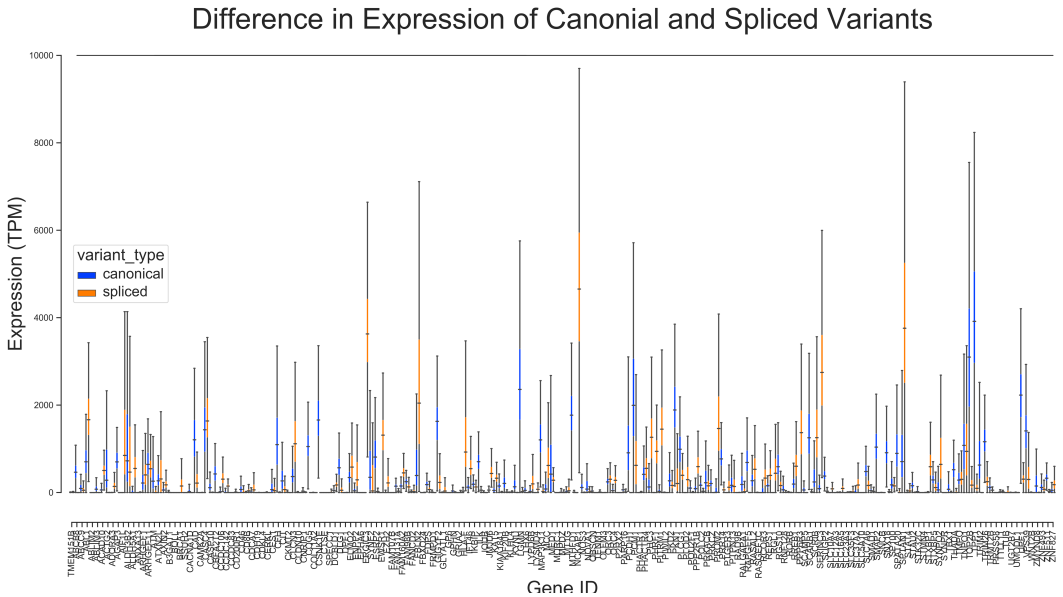


Figure 3. Comparison in Expression of CMNRE variants in Tumor and Normal Samples: Of the 191 CMNRE variants 135 had significantly reduced increased expression in tumor samples compared to normal samples when examined via a one tailed T-test *(alpha = 0.05).* These finding further suggest that splicing mechanisms are positively selecting for healthier variants. The 135 CMNRE variants with significantly increased expression are referred to as the CMNRTS variants.

3.3 Survival Examination of individuals with these Expression Characteristics:

To examine the clinical manifestations of these statistical findings, the canonical variants composing the subset of consideration were then evaluated in terms of survival analysis for patients with higher expression levels of these variants versus those with lower expression levels. The upper third quartile of expression level TPM for each CMNRTS variant was identified.

Patient data was merged into the database using the unique TCGA barcode identifier given to each sample. The expression level of each variant for each patient was then compared to the variant's upper quartile. Patients with expression levels greater than or equal to this amount were classified as having high expression, while patients with expression levels below this amount were classified as having low expression. The right censored survival for each patient was determined using the number of days from when the sample was taken to either the patient's death or to the patient's last follow up appointment if they are still alive. The survival was examined using a Kaplan-Meier estimate. From these charts it is qualitatively clear that for a considerable portion of the variants patients with higher expression had reduced longevity. In order to examine this difference quantitatively, a log rank test was performed comparing patients with high and low expression for each variant. Of the 135 variants examined, 71 had significantly different survival rates between patients with high and low expression.

In addition to examining the effect on survival of the expression level of each CMNRTS variants independently an investigation into the sum effect of the expression of all variants was conducted. In order to examine the overall effect, the samples were divided depending on whether the majority of the CMNRTS variants for the sample were high or low. The samples with majority high CMNRTS expression were classified as high overall whereas the samples with minority high CMNRTS expression were classified as low overall. A survival analysis was then conducted comparing these two groups. The overall high expression group had noticeably reduced survival which was then determined to be significant through a log rank test ($p = 1.4E\text{-}10$, Figure 4.A).

Motivated by these findings we also sought to examine what proportion of high expression variants was needed to have a significant effect on survive. We iteratively calculated

the significance of the reduction in survival for increasing number of high variants. Beginning with 0 high variants we iteratively increased that amount to 90, an amount slightly above the majority measure used previously.

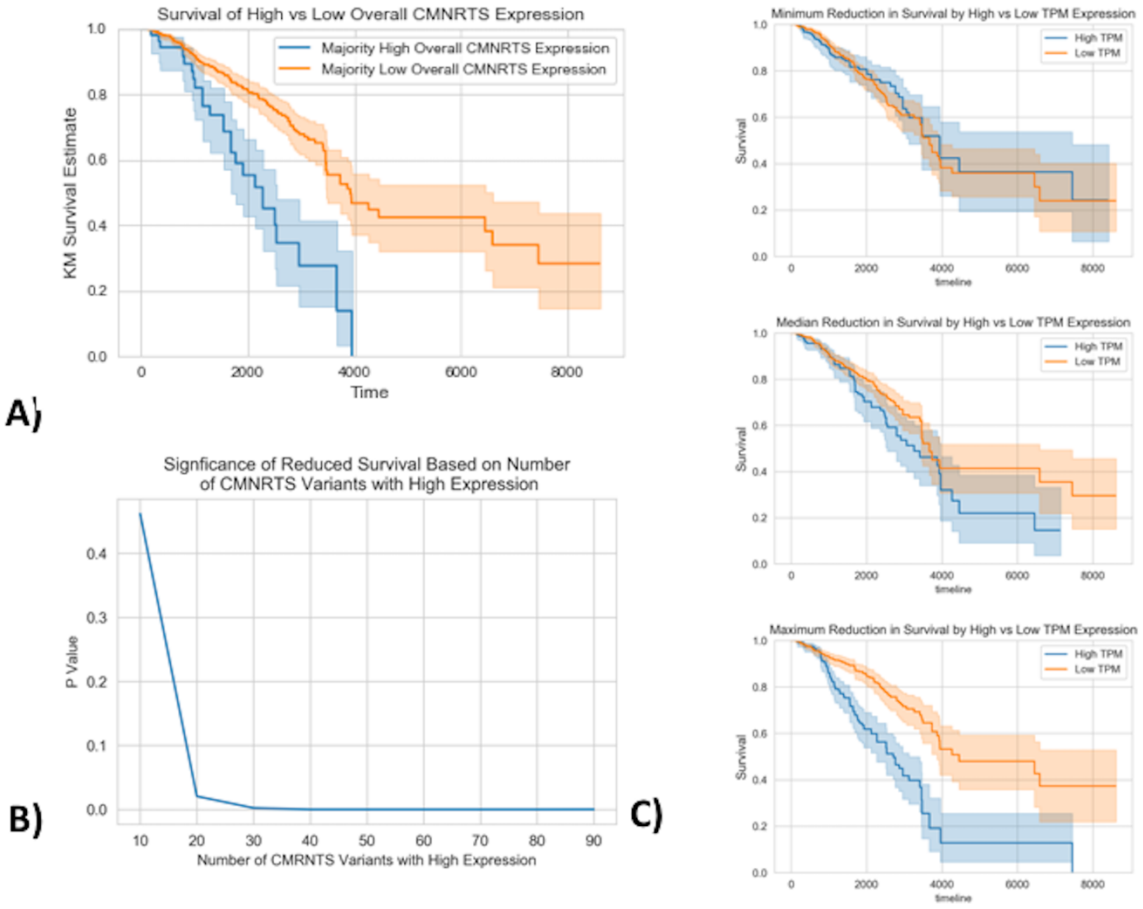


Figure 4. Survival analysis of patients with regard to expression level of CMNRTS variants. A) Survival of patients for whom the majority of the CMNRTS variants was high compared to those for whom it was not. Individuals with overall high expression of CMNRTS variants had significantly reduced survival compared to individuals with non-majority high expression. **B)** The significance in reduced survival by number of CMNRTS variants with high expression. There is a clear trend in that the more CMNRTS variants with high expression the more significant the reduction in survival. **C)** Survival curves corresponding to the variants with the maximum, median and minimum p value for reduction in survival. As the majority of variants did have a significant reduction even the curve corresponding to the median p values has visible qualitative separation between the two populations.

We found that there was noticeable increase in significance of reduced survival at 30 variants. This suggests that although high expression of 71 variants were correlated with reduced survival, high expression of less than half of was enough to dramatically affect clinical outcomes.

3.4 Support Vector Machine Classification of Sample Condition

The general inference obtained from the statistical analysis done so far is that there is a constitutional difference in the transcriptome of cancerous and normal tissues. In order to examine this inference more directly, a support vector machine (SVM) machine learning algorithm was employed. SVM algorithms are designed to identify a linear separator between two classes of data points in the multidimensional space. This is done by considering only a few points closest to the class boundary and maximizing the distance of the support vectors for both classes. Support vectors refer to vectors originating on these points and terminating at the boundary. SVMs can be altered to identify nonlinear boundaries as well however for this investigation is was determined that a straight forward linear boundary was most accurate.

The initial dataset contained 1093 data instances from tumor samples and only 112 from normal samples. When using SVM and many other machine learning algorithms this degree of imbalance can often result in a propensity of type 1 errors. In order to eliminate this bias the data was balanced using SMOTE resampling. This technique creates additional data instances of the minority class by finding feature values, in this case variant expression levels, in between the values of data instances of the same class. By applying this technique to the dataset a balanced dataset of 1093 tumor and 1093 normal data instances. Using this balanced data set the algorithm was trained on 400 data instances at a time and tested on 100, this process was repeated five times, each time with a randomly selected subset for both training and testing, a process known

as five fold cross validation. The average accuracy for this algorithm was 0.9679 (Figure 4.B). Additionally when the dataset was projected down to two dimensions there was a viable visible linear boundary between the tumor and normal points (Figure 4.A). These results reaffirm that there is constitutive difference in expression of the CMNR variants in tumor and normal samples.

3.5 Correlation Between Derived Features

Combining the results from the previously described statistical tests, a data subset containing 135 different genes with mutation, expression level, and survival characteristics was obtained. In order to better understand the relationship between all these derived characteristics (number of mutations, number of pathogenic mutations, statistical significance of increased canonical variant expression in tumor samples, statistical significance of decreases spliced variant expression in normal samples, and statistical significance of reduced survival) the correlation between all characteristics was examined through a facet plot. In order to better interpret the relationships between the p-value based features and the numerical features, the p-values were transformed by the absolute value of the log. It follows that a higher value is indicative of a lower p-value and therefore a more significant result.

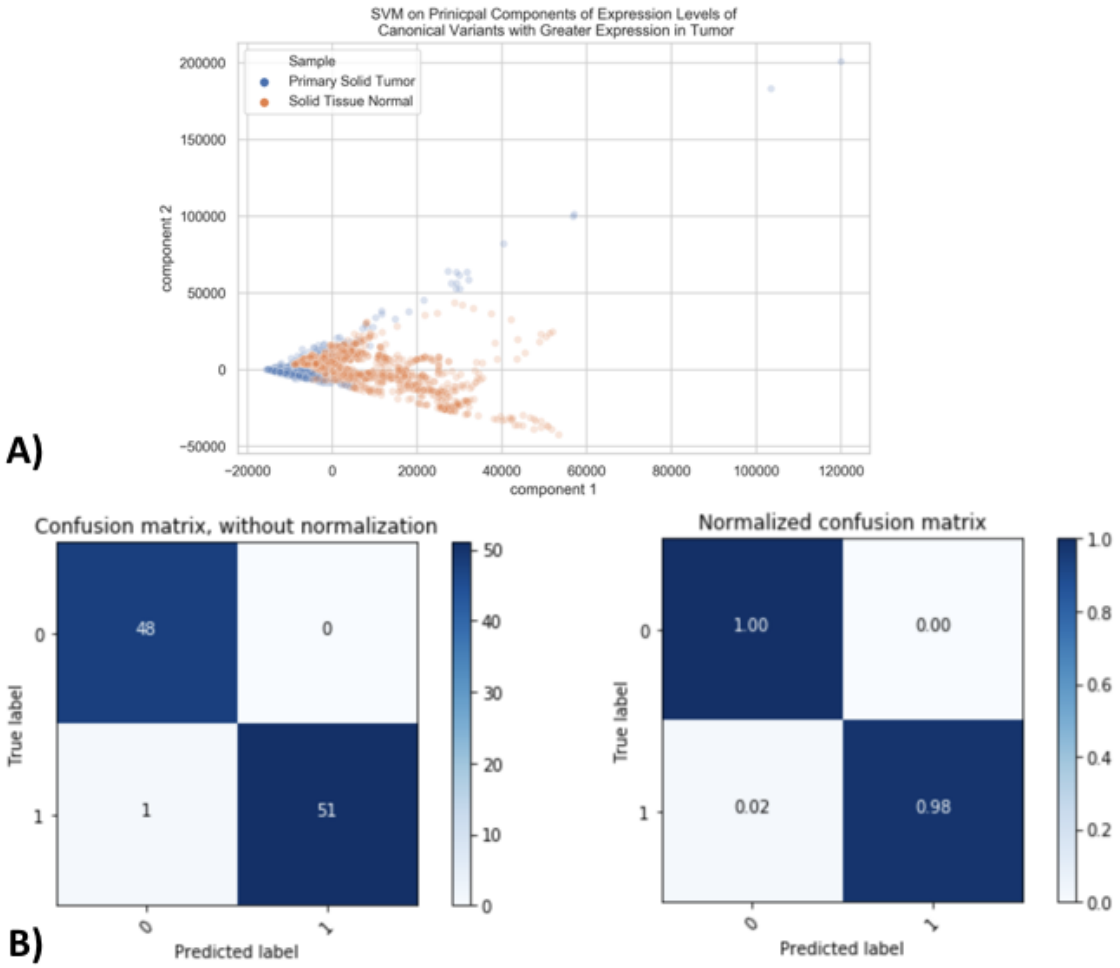


Figure 5. SVM Classification of CMNRTS variants: A) Using principal component analysis (PCA) the expression of all 71 variants was projected down to 2 dimensions. In this two dimensional plane there is a noticeable separation between the tumor and normal tissue classes. **B)** The overall accuracy of the SVM was about 96.79% and as can be seen by the confusion matrix there is no bias towards type 1 errors, illustrating the effectiveness of the SMOTE resampling towards generating reliable results.

The most correlated features were the total number of mutations in the canonical variant and not in the spliced variant and the number of pathological mutations in the canonical variant and not in the spliced variant. This is to be expected as the number of pathological mutations is essentially a subset of the number of total mutations and, as was shown previously, a large portion of the total mutations of this type are also pathogenic.

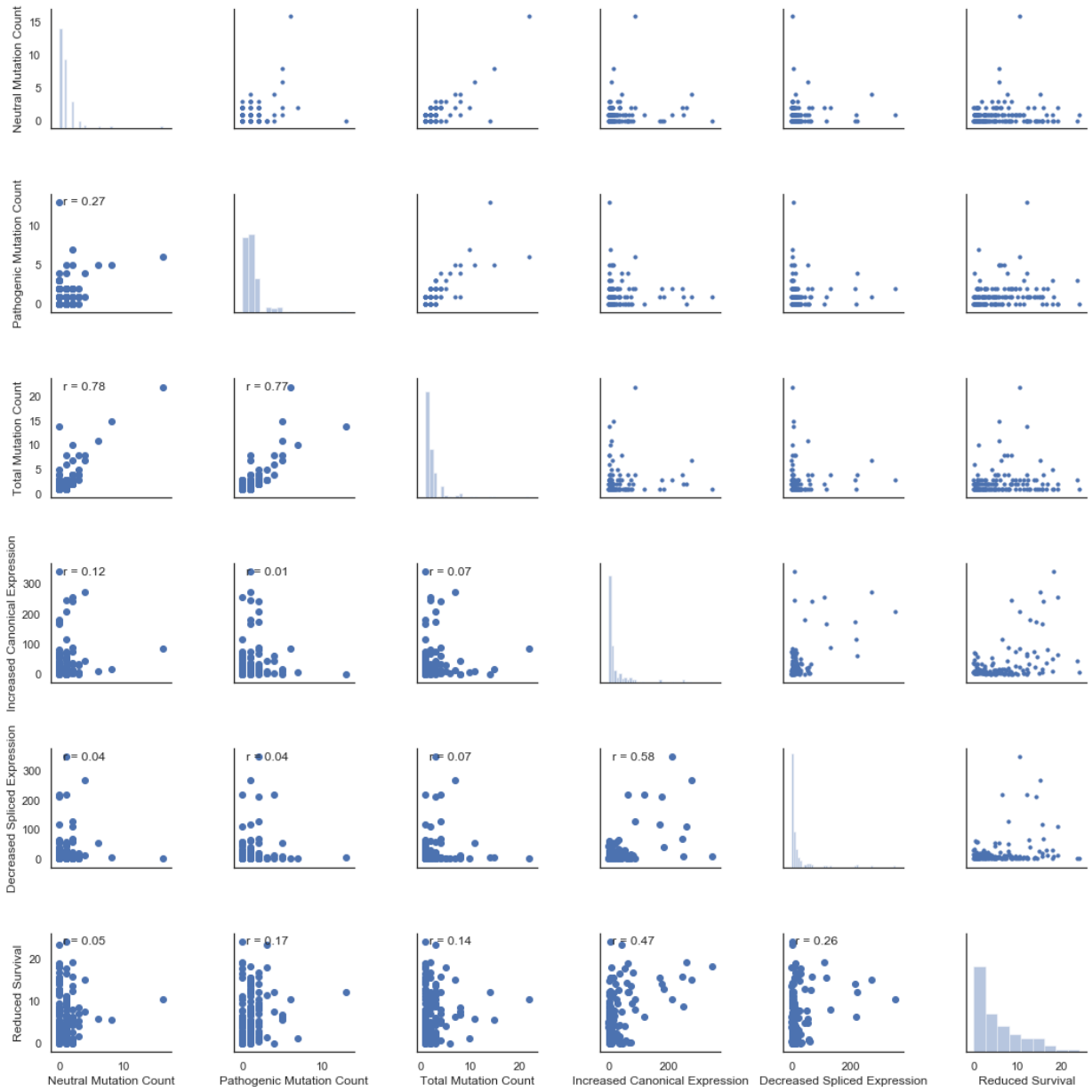


Figure 6. Correlation between all derived features: For each of the 135 CMNRTS variants the number of total mutations, number of pathogenic mutations, number of neutral mutations, p value for increased expression in tumor samples compared to normal, p value for reduced expression of the corresponding spliced variant in tumor compared to normal and p value of reduced survival were compared. The diagonal portrays a histogram for each feature while the upper and lower triangles show the pairwise correlation between all features. Discounting the features that have implicit relationships such as number of total mutations and number of pathogenic mutations, the highest correlation between features was the p value of increased expression in tumor samples compared to normal and the p value of reduced survival.

The next most correlated features, and perhaps the most interesting were the level of significance of reduced survival and the level of significance of increased canonical expression in tumors ($r = 0.47$). This is particularly interesting because it provides further evidence to support our hypothesis that having these affected canonical isoforms instead of the spliced alternatives is linked to a less desirable clinical prognosis. Overall there was at least a slight positive correlation among all features (Figure 6).

4. DISCUSSION

4.1 Overview of Results

We completed the first large scale statistical investigation into the occurrence of mediative alternative splicing in the human genome. Specifically, we provide evidence for the contribution of this mechanism to preventing complex diseases such as breast cancer. There is a clear endogenous system at work which repairs or negates the effect of potentially disease contributing mutations by excluding these mutations from transcription, preserving functionality of the protein. Therefore, contrary to common conception it appears a considerable contributor to the expression of disease phenotypes is not aberrant splicing mechanisms but rather the absence of splicing mechanisms. The identification of this phenomena may have been stunted by its inherent advantageous aspects in that the result of this process is a healthy or normal phenotype.

Perhaps the most intriguing consequence of these findings are the implications for the degree of dexterity and influence alternative splicing has on gene regulations. Alternative splicing can occur on a tissue specific and even cell specific basis, meaning that these mechanisms may be dynamically mediating mutations differently across tissues, potentially preventing the occurrence of various diseases at once. This degree of control over the genome

would surpass any existing considerations of the implications of these mechanisms and potentially irrevocably alter the way researchers view the relationship between mutations, genetics, and complex diseases.

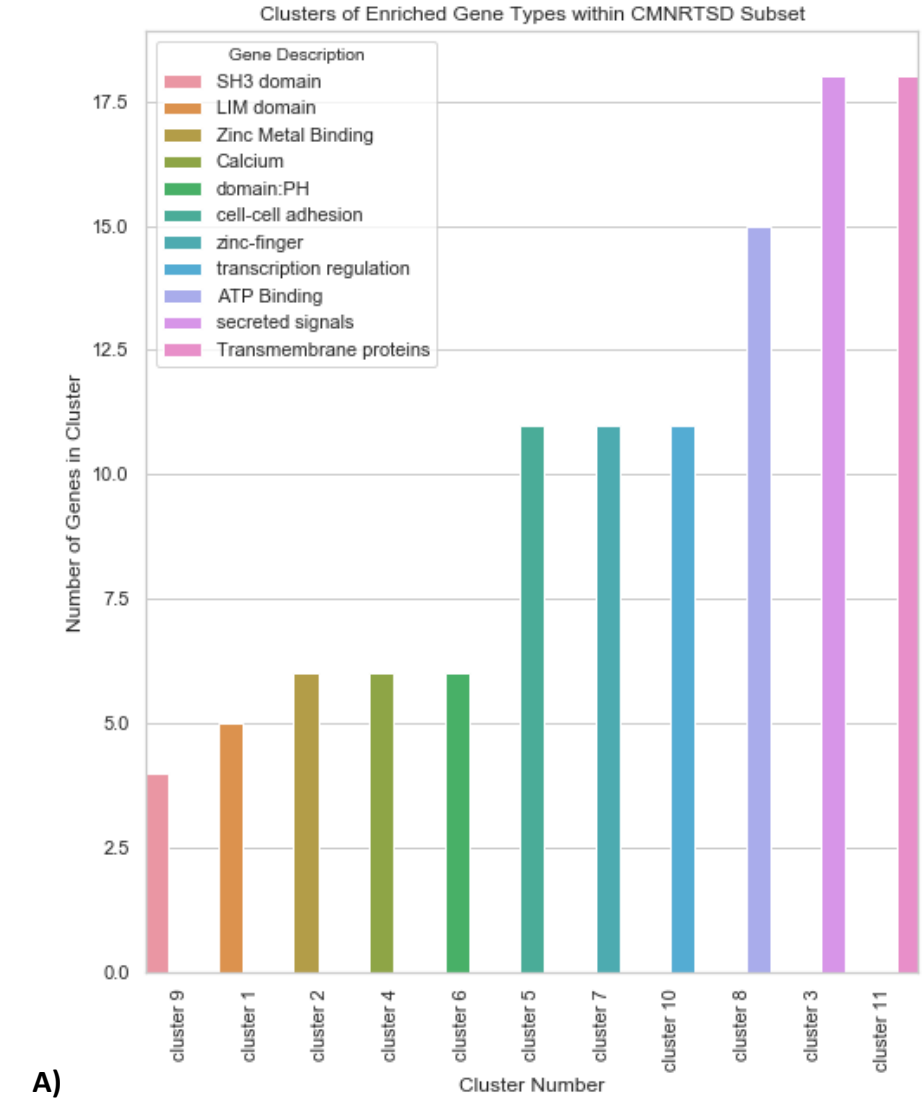
This investigation provided considerable evidence to support our hypothesis that as with the *mdx* mice, endogenous mediative alternative splicing is occurring and is present in human genomes and can be complicit in preventing disease phenotypes. This is novel research and brings with it novel methodologies to two of the most critical components of biomedical development, namely therapeutic development and diagnostics.

4.2 Therapeutic Implications

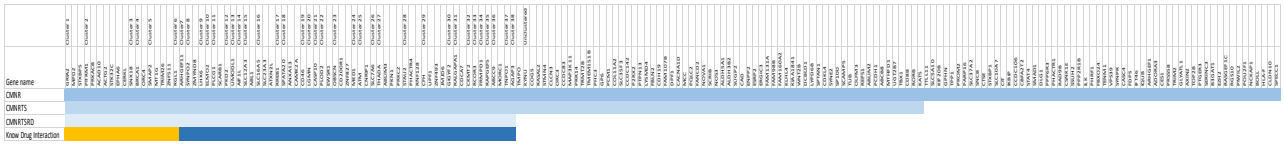
Generally, investigations into diseases involving alternative splicing focus on mediating or repairing aberrant splicing mechanisms. Similarly investigations into diseases that involve SNP or point mutations are generally focused on repairing the mutated gene or substituting/compensating for a nonfunctional protein. The results of this investigation suggest there may be another option, instead of affecting the gene as a whole or replacing the splicing mechanism we can work within the endogenous alternative splicing machinery. As evidenced by this investigation there is already naturally occurring alternative splicing processes to explicitly splice out would be harmful mutations and still yield functional proteins. This suggests that a critical determinant for disease phenotype may not be simply the presence of these mutations but the degree to which they are transcribed. Therefore, there is a clear opportunity to intervene therapeutically by simply promoting for or upregulating whatever natural process results in the healthy spliced variant compared to the diseased canonical. Previous investigations into exploiting alternative splicing mechanisms have sought to accomplish similar results but relied

on manually instituting the desired splice by introducing agents such as oligonucleotides (McClorey et al, *Gene Therapy* 2006) and generally been focused on diseases for which there is a clear point mutation cause such as muscular dystrophy. Contrarily the approach we are recommending from the results of this study would negate the need for this introduction and rather simply utilize the existing splicing mechanisms we have shown are already occurring in the tissues.

Within this study alone, 71 genes were identified as possible therapeutic targets for breast cancer having significantly increased incidence of pathogenic mutations, expression levels in tumor samples, and reduced survival (Figure 7.B). Of these genes only 18 currently have any sort of drug interaction, either as a direct target or accessory according to the Drug Gene Interaction Database. Perhaps even more surprising is that of the remaining 53 genes none of them are commonly associated with breast cancer or any cancer in general. A DAVID functional clustering of the genes revealed significant enrichment in 11 functional clusters; SH3 domain, LIM domain, Zinc Metal Binding, Calcium, domain:PH, cell-cell adhesion, zinc-finger, transcription regulation, ATP Binding, secreted signals, and Transmembrane proteins (Figure 7.A). While the relationship to some of these functional clusters and cancer is understandable, such as cell to cell adhesion and secreted signals, some of the others are not so obvious such as the zinc binding and SH3 and LIM domains.



A)



B)

Figure 7. Gene clusters within CMNRTSRD subset which are not currently associated with any drug interaction: A) The 53 CMNRTSRD variants which do not have a drug interaction can be effectively grouped into 11 clusters by functional gene annotation. Of these clusters the largest are membrane and metal binding, two functions not generally associated with cancer development. Known functions affected by cancer are contained by smaller clusters. **B)** The evolution of the specification of variants for by each characteristic as well as the clustering of the original 191 CMNR variants. Of these 191 original variants 119 remained unclustered however all CMNRTSRD variants were part of a CMNR clustering.

The implications of the findings of this study for therapeutic development are considerable. If we can apply these findings to human diseases it would provide an effective and simpler alternative to large scale gene therapy and provide hope to thousands of individuals suffering from all types of illness. However, to maximize the impact of these findings similar statistical investigations should be conducted on other cancer cohorts. It would be worthwhile to examine the composition of gene types of this investigation on other cancer types as the genes identified here have for the most part not been implicated in the disease yet it would be interesting to see if the same sleuth genes are identified across cancer types or if an entirely different previously unknown subset emerges. Finally, while cancer was chosen as the disease of interest for this initial investigation due to the exhaustive publicly available datasets it is important to remember these findings could as efficiently be applied to any disease with a genetic component.

4.3 Diagnostic Implications

In addition to the considerable implications of this study for therapeutic development there are also equally impactful implications for diagnostics. From this study it is clear that a machine learning based approach to sorting samples could contribute to faster and more accurate diagnoses. Samples could be accurately classified as tumor or normal by expression of less than 0.01% the number of variants of a full mRNA sequencing. By testing a tissue sample for expression levels of just the handful of variants identified here the sample can be automatically sorted and classified. As this approach is expanded to include different types of cancers it is also foreseeable this algorithm can be expanded to classify the tissue not only as tumor or normal but the that the expression of just a few variants may reveal the stage and type of cancer, currently

largely still determined by hand through standard histological procedures. Furthermore, there is the possibility this technique could be expanded beyond tissue specificity to cellular specificity given the increasing feasibility of single cell RNAseq methods. By combining machine learning with the findings of this investigation diagnostic methods can improve efficiency and accuracy on less information meaning faster answers or patients and potentially earlier detection.

References:

Aartsma-Rus, Annemieke, et al. "Therapeutic antisense-induced exon skipping in cultured muscle cells from six different DMD patients." *Human Molecular Genetics* 12.8 (2003): 907-914.

Broad Institute TCGA Genome Data Analysis Center (2016): Analysis Overview for Breast Invasive Carcinoma (Primary solid tumor cohort) - 28 January 2016. Broad Institute of MIT and Harvard. doi:10.7908/C1V40TJ9

Daniel R Zerbino, Premanand Achuthan, Wasiu Akanni, M Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, Carla Cummins, Astrid Gall, Carlos García Girón, Laurent Gil, Leo Gordon, Leanne Haggerty, Erin Haskell, Thibaut Hourlier, Osagie G Izuogu, Sophie H Janacek, Thomas Juettemann, Jimmy Kiang To, Matthew R Laird, Ilias Lavidas, Zhicheng Liu, Jane E Loveland, Thomas Maurel, William McLaren, Benjamin Moore, Jonathan Mudge, Daniel N Murphy, Victoria Newman, Michael Nuhn, Denye Ogeh, Chuang Kee Ong, Anne Parker, Mateus Patricio, Harpreet Singh Riat, Helen Schuilenburg, Dan Sheppard, Helen Sparrow, Kieron Taylor, Anja Thormann, Alessandro Vullo, Brandon Walts, Amonida Zadissa, Adam Frankish, Sarah E Hunt, Myrto Kostadima, Nicholas Langridge, Fergal J Martin, Matthieu Muffato, Emily Perry, Magali Ruffier, Dan M Staines, Stephen J Trevanion, Bronwen L Aken, Fiona Cunningham, Andrew Yates, Paul Flicek, Ensembl 2018, *Nucleic Acids Research*, Volume 46, Issue D1, 4 January 2018, Pages D754–D761, https://doi.org/10.1093/nar/gkx1098

Johnson, Jason M., et al. "Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays." *Science* 302.5653 (2003): 2141-2144.

Kelemen, Olga et al. “Function of alternative splicing.” *Gene* vol. 514,1 (2012): 1-30. doi:10.1016/j.gene.2012.07.083

Kelsy C Cotto, Alex H Wagner, Yang-Yang Feng, Susanna Kiwala, Adam C Coffman, Gregory Spies, Alex Wollam, Nicholas C Spies, Obi L Griffith, Malachi Griffith, DGIdb 3.0: a redesign and expansion of the drug–gene interaction database, *Nucleic Acids Research*, Volume 46, Issue D1, 4 January 2018, Pages D1068–D1073, https://doi.org/10.1093/nar/gkx1143

Kriventseva, Evgenia V., et al. "Increase of functional diversity by alternative splicing." *Trends in Genetics* 19.3 (2003): 124-128.

Mann, Christopher J., et al. "Antisense-induced exon skipping and synthesis of dystrophin in the mdx mouse." *Proceedings of the National Academy of Sciences* 98.1 (2001): 42-47.

McClorey, G., Moulton, H. M., Iversen, P. L., Fletcher, S., & Wilton, S. D. (2006). Antisense oligonucleotide-induced exon skipping restores dystrophin expression in vitro in a canine model of DMD. *Gene therapy*, *13*(19), 1373.

McClorey, G., et al. "Induced dystrophin exon skipping in human muscle explants." *Neuromuscular Disorders* 16.9-10 (2006): 583-590.

Modrek, Barmak, and Christopher Lee. "A genomic view of alternative splicing." *Nature genetics* 30.1 (2002): 13.

Nilsen, Timothy W., and Brenton R. Graveley. "Expansion of the eukaryotic proteome by alternative splicing." *Nature*463.7280 (2010): 457.

Simon A. Forbes, David Beare, Harry Boutselakis, Sally Bamford, Nidhi Bindal, John Tate, Charlotte G. Cole, Sari Ward, Elisabeth Dawson, Laura Ponting, Raymund Stefancsik, Bhavana Harsha, Chai Yin Kok, Mingming Jia, Harry Jubb, Zbyslaw Sondka, Sam Thompson, Tisham De, Peter J. Campbell, COSMIC: somatic cancer genetics at high-resolution, *Nucleic Acids Research*, Volume 45, Issue D1, January 2017, Pages D777–D783, https://doi.org/10.1093/nar/gkw1121

UCSC Genome Browser: Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. *Genome Res.* 2002 Jun;12(6):996-1006.

Wilton, Stephen D., et al. "Specific removal of the nonsense mutation from the mdx dystrophin mRNA using antisense oligonucleotides." *Neuromuscular disorders* 9.5 (1999): 330-338.

Supplementary Materials:

Supplementary Table S1. Gene information for all genes with canonical variants in the CMNR subset: Ensembl ID's, functions and names for each gene is supplied in addition to a Boolean array for each derived subset in the investigation. A value of 1 indicates the gene was present in the given subset, a value of 0 indicates it was not.

Gene stable ID	Gene description	Gene name	CMNR	CMNRTS	CMNRTSRD	Drug Interaction
ENSG00000275931	glycoprotein VI platelet [Source:NCBI gene;Acc:51206]	GP6	1	0	0	0
ENSG00000117335	CD46 molecule [Source:HGNC Symbol;Acc:HGNC:6953]	CD46	1	1	0	0
ENSG00000143344	ral guanine nucleotide dissociation stimulator like 1 [Source:HGNC Symbol;Acc:HGNC:30281]	RGL1	1	0	0	0
ENSG00000282608	adenosine A3 receptor [Source:HGNC Symbol;Acc:HGNC:268]	ADORA3	1	1	1	1
ENSG00000170190	solute carrier family 16 member 5 [Source:HGNC Symbol;Acc:HGNC:10926]	SLC16A5	1	0	0	0
ENSG00000144354	cell division cycle associated 7 [Source:HGNC Symbol;Acc:HGNC:14628]	CDCA7	1	1	0	0
ENSG00000119865	cannabinoid receptor interacting protein 1 [Source:HGNC Symbol;Acc:HGNC:24546]	CNRIP1	1	1	1	0
ENSG00000117625	REST corepressor 3 [Source:HGNC Symbol;Acc:HGNC:25594]	RCOR3	1	1	0	0
ENSG00000132694	Rho guanine nucleotide exchange factor 11 [Source:HGNC Symbol;Acc:HGNC:14580]	ARHGEF11	1	1	0	0
ENSG00000148908	regulator of G protein signaling 10 [Source:HGNC Symbol;Acc:HGNC:9992]	RGS10	1	1	0	0
ENSG00000100600	legumain [Source:HGNC Symbol;Acc:HGNC:9472]	LGMN	1	1	1	0
ENSG00000109572	chloride voltage-gated channel 3 [Source:HGNC Symbol;Acc:HGNC:2021]	CLCN3	1	1	0	0
ENSG00000232541	collagen type XI alpha 2 chain [Source:HGNC Symbol;Acc:HGNC:2187]	COL11A2	1	0	0	0
ENSG00000118096	intraflagellar transport 46 [Source:HGNC Symbol;Acc:HGNC:26146]	IFT46	1	1	1	0
ENSG00000154358	obscurin, cytoskeletal calmodulin and titin-interacting RhoGEF [Source:HGNC Symbol;Acc:HGNC:15719]	OBSCN	1	1	1	0
ENSG00000213901	solute carrier family 23 member 3 [Source:HGNC Symbol;Acc:HGNC:20601]	SLC23A3	1	0	0	0
ENSG00000143190	POU class 2 homeobox 1 [Source:HGNC Symbol;Acc:HGNC:9212]	POU2F1	1	1	0	0
ENSG00000125686	mediator complex subunit 1 [Source:HGNC Symbol;Acc:HGNC:9234]	MED1	1	0	0	0
ENSG00000070495	jumonji domain containing 6, arginine demethylase and lysine hydroxylase [Source:HGNC Symbol;Acc:HGNC:19355]	JMJD6	1	1	1	1
ENSG00000213923	casein kinase 1 epsilon [Source:HGNC Symbol;Acc:HGNC:2453]	CSNK1E	1	1	0	0
ENSG00000062598	engulfment and cell motility 2 [Source:HGNC Symbol;Acc:HGNC:17233]	ELMO2	1	1	1	0
ENSG00000136237	Rap guanine nucleotide exchange factor 5 [Source:HGNC Symbol;Acc:HGNC:16862]	RAPGEF5	1	1	1	0
ENSG00000230230	tripartite motif containing 26 [Source:HGNC Symbol;Acc:HGNC:12962]	TRIM26	1	1	0	0

ENSG000	pitrilysin metallopeptidase 1 [Source:HGNC	PITR				
00107959	Symbol;Acc:HGNC:17663]	M1	1	1	0	0
ENSG000	FA complementation group D2 [Source:HGNC	FANC				
00144554	Symbol;Acc:HGNC:3585]	D2	1	1	0	0
ENSG000	ras responsive element binding protein 1 [Source:HGNC	RREB				
00124782	Symbol;Acc:HGNC:10449]	1	1	1	1	0
ENSG000	tubulin tyrosine ligase like 11 [Source:HGNC	TTLL1				
00175764	Symbol;Acc:HGNC:18113]	1	1	1	1	0
ENSG000						
00106852	LIM homeobox 6 [Source:HGNC Symbol;Acc:HGNC:21735]	LHX6	1	1	1	0
ENSG000	major histocompatibility complex, class I, F [Source:HGNC					
00234487	Symbol;Acc:HGNC:4963]	HLA-F	1	0	0	0
ENSG000	polyhomeotic homolog 3 [Source:HGNC					
00173889	Symbol;Acc:HGNC:15682]	PHC3	1	1	1	0
ENSG000		PCDH				
00156453	protocadherin 1 [Source:HGNC Symbol;Acc:HGNC:8655]	1	1	1	0	0
ENSG000	Rho guanine nucleotide exchange factor 1 [Source:HGNC	ARHG				
00076928	Symbol;Acc:HGNC:681]	EF1	1	1	0	0
ENSG000	ATM serine/threonine kinase [Source:HGNC					
00149311	Symbol;Acc:HGNC:795]	ATM	1	1	1	1
ENSG000						
00184058	T-box 1 [Source:HGNC Symbol;Acc:HGNC:11592]	TBX1	1	1	0	0
ENSG000	phosphatase and actin regulator 4 [Source:HGNC	PHAC				
00204138	Symbol;Acc:HGNC:25793]	TR4	1	1	1	0
ENSG000	tubby bipartite transcription factor [Source:HGNC					
00166402	Symbol;Acc:HGNC:12406]	TUB	1	1	0	0
ENSG000	staufen double-stranded RNA binding protein 2 [Source:HGNC	STAU				
00040341	Symbol;Acc:HGNC:11371]	2	1	0	0	0
ENSG000						
00123143	protein kinase N1 [Source:HGNC Symbol;Acc:HGNC:9405]	PKN1	1	1	0	0
ENSG000	family with sequence similarity 98 member B [Source:HGNC	FAM9				
00171262	Symbol;Acc:HGNC:26773]	8B	1	1	1	0
ENSG000		SYNP				
00163630	synaptoporin [Source:HGNC Symbol;Acc:HGNC:16507]	R	1	0	0	0
ENSG000	SURP and G-patch domain containing 2 [Source:HGNC	SUGP				
00064607	Symbol;Acc:HGNC:18641]	2	1	1	1	0
ENSG000	SMAD family member 1 [Source:HGNC	SMAD				
00170365	Symbol;Acc:HGNC:6767]	1	1	1	0	0
ENSG000	CBFA2/RUNX1 translocation partner 2 [Source:HGNC	CBFA				
00078699	Symbol;Acc:HGNC:1536]	2T2	1	1	1	0
ENSG000		SMAP				
00084070	small ArfGAP2 [Source:HGNC Symbol;Acc:HGNC:25082]	2	1	1	1	0
ENSG000	phosphatase and actin regulator 1 [Source:HGNC	PHAC				
00112137	Symbol;Acc:HGNC:20990]	TR1	1	0	0	0
ENSG000	calcium/calmodulin dependent protein kinase II alpha	CAMK				
00070808	[Source:HGNC Symbol;Acc:HGNC:1460]	2A	1	1	0	0
ENSG000		STMN				
00015592	stathmin 4 [Source:HGNC Symbol;Acc:HGNC:16078]	4	1	1	1	1
ENSG000	zinc finger protein 827 [Source:HGNC	ZNF82				
00151612	Symbol;Acc:HGNC:27193]	7	1	1	0	0
ENSG000	tripartite motif containing 2 [Source:HGNC	TRIM				
00109654	Symbol;Acc:HGNC:15974]	2	1	1	1	0
ENSG000	polycystin 2, transient receptor potential cation channel					
00118762	[Source:HGNC Symbol;Acc:HGNC:9009]	PKD2	1	1	1	0
ENSG000	aldehyde dehydrogenase 5 family member A1 [Source:HGNC	ALDH				
00112294	Symbol;Acc:HGNC:408]	5A1	1	1	0	0

ENSG00000134873	claudin 10 [Source:HGNC Symbol;Acc:HGNC:2033]	CLDN10	1	0	0	0
ENSG00000120519	solute carrier family 10 member 7 [Source:HGNC Symbol;Acc:HGNC:23088]	SLC10A7	1	0	0	0
ENSG00000173327	mitogen-activated protein kinase kinase kinase 11 [Source:HGNC Symbol;Acc:HGNC:6850]	MAP3K11	1	1	0	0
ENSG00000135336	origin recognition complex subunit 3 [Source:HGNC Symbol;Acc:HGNC:8489]	ORC3	1	0	0	0
ENSG00000171444	MCC, WNT signaling pathway regulator [Source:HGNC Symbol;Acc:HGNC:6935]	MCC	1	1	1	0
ENSG00000162929	KIAA1841 [Source:HGNC Symbol;Acc:HGNC:29387]	KIAA1841	1	1	0	0
ENSG00000012048	BRCA1, DNA repair associated [Source:HGNC Symbol;Acc:HGNC:1100]	BRCA1	1	0	0	0
ENSG00000166333	integrin linked kinase [Source:HGNC Symbol;Acc:HGNC:6040]	ILK	1	0	0	0
ENSG00000166340	tripeptidyl peptidase 1 [Source:HGNC Symbol;Acc:HGNC:2073]	TPP1	1	0	0	0
ENSG00000102271	kelch like family member 4 [Source:HGNC Symbol;Acc:HGNC:6355]	KLHL4	1	1	1	0
ENSG00000166887	VPS39, HOPS complex subunit [Source:HGNC Symbol;Acc:HGNC:20593]	VPS39	1	1	1	0
ENSG00000067606	protein kinase C zeta [Source:HGNC Symbol;Acc:HGNC:9412]	PRKCZ	1	1	0	0
ENSG00000168646	axin 2 [Source:HGNC Symbol;Acc:HGNC:904]	AXIN2	1	0	0	0
ENSG00000276785	serine protease 58 [Source:HGNC Symbol;Acc:HGNC:39125]	PRSS58	1	0	0	0
ENSG00000171234	UDP glucuronosyltransferase family 2 member B7 [Source:HGNC Symbol;Acc:HGNC:12554]	UGT2B7	1	0	0	0
ENSG00000198794	secretory carrier membrane protein 5 [Source:HGNC Symbol;Acc:HGNC:30386]	SCAMP5	1	1	0	0
ENSG00000172977	lysine acetyltransferase 5 [Source:HGNC Symbol;Acc:HGNC:5275]	KAT5	1	0	0	0
ENSG00000106336	F-box protein 24 [Source:HGNC Symbol;Acc:HGNC:13595]	FBXO24	1	0	0	0
ENSG00000131730	creatine kinase, mitochondrial 2 [Source:HGNC Symbol;Acc:HGNC:1996]	CKMT2	1	1	1	1
ENSG00000198848	carboxylesterase 1 [Source:HGNC Symbol;Acc:HGNC:1863]	CES1	1	1	1	1
ENSG00000173581	coiled-coil domain containing 106 [Source:HGNC Symbol;Acc:HGNC:30181]	CCDC106	1	1	0	0
ENSG00000040487	PQ loop repeat containing 2 [Source:HGNC Symbol;Acc:HGNC:26001]	PQLC2	1	1	0	0
ENSG00000115970	THADA, armadillo repeat containing [Source:HGNC Symbol;Acc:HGNC:19217]	THADA	1	1	1	0
ENSG00000163995	actin binding LIM protein family member 2 [Source:HGNC Symbol;Acc:HGNC:19195]	ABLIM2	1	1	1	0
ENSG00000197694	spectrin alpha, non-erythrocytic 1 [Source:HGNC Symbol;Acc:HGNC:11273]	SPTAN1	1	1	1	0
ENSG00000051009	family with sequence similarity 160 member A2 [Source:HGNC Symbol;Acc:HGNC:25378]	FAM160A2	1	1	0	0
ENSG00000009694	teneurin transmembrane protein 1 [Source:HGNC Symbol;Acc:HGNC:8117]	TENM1	1	0	0	0
ENSG00000103064	solute carrier family 7 member 6 [Source:HGNC Symbol;Acc:HGNC:11064]	SLC7A6	1	0	0	0

ENSG00000274287	scribbled planar cell polarity protein [Source:HGNC Symbol;Acc:HGNC:30377]	SCRIB	1	1	0	0
ENSG00000125046	ssu-2 homolog (C. elegans) [Source:HGNC Symbol;Acc:HGNC:24809]	SSUH2	1	0	0	0
ENSG00000142875	protein kinase cAMP-activated catalytic subunit beta [Source:HGNC Symbol;Acc:HGNC:9381]	PRKACB	1	0	0	0
ENSG00000006282	spermatogenesis associated 20 [Source:HGNC Symbol;Acc:HGNC:26125]	SPATA20	1	1	0	0
ENSG00000188917	tRNA methyltransferase 2 homolog B [Source:HGNC Symbol;Acc:HGNC:25748]	TRMT2B	1	0	0	0
ENSG00000134245	Wnt family member 2B [Source:HGNC Symbol;Acc:HGNC:12781]	WNT2B	1	0	0	0
ENSG00000151164	RAD9 checkpoint clamp component B [Source:HGNC Symbol;Acc:HGNC:21700]	RAD9B	1	0	0	0
ENSG00000000971	complement factor H [Source:HGNC Symbol;Acc:HGNC:4883]	CFH	1	1	1	0
ENSG00000174373	Ral GTPase activating protein catalytic alpha subunit 1 [Source:HGNC Symbol;Acc:HGNC:17770]	RALGAPA1	1	1	0	0
ENSG00000177398	uromodulin like 1 [Source:HGNC Symbol;Acc:HGNC:12560]	UMODL1	1	1	0	0
ENSG00000101040	zinc finger MYND-type containing 8 [Source:HGNC Symbol;Acc:HGNC:9397]	ZMYND8	1	1	1	0
ENSG00000168488	ataxin 2 like [Source:HGNC Symbol;Acc:HGNC:31326]	ATXN2L	1	1	0	0
ENSG00000054654	spectrin repeat containing nuclear envelope protein 2 [Source:HGNC Symbol;Acc:HGNC:17084]	SYNE2	1	1	1	0
ENSG00000069431	ATP binding cassette subfamily C member 9 [Source:HGNC Symbol;Acc:HGNC:60]	ABCC9	1	1	1	1
ENSG00000154025	solute carrier family 5 member 10 [Source:HGNC Symbol;Acc:HGNC:23155]	SLC5A10	1	1	0	0
ENSG00000125144	metallothionein 1G [Source:HGNC Symbol;Acc:HGNC:7399]	MT1G	1	1	0	0
ENSG00000164465	discoidin, CUB and LCCL domain containing 1 [Source:HGNC Symbol;Acc:HGNC:21479]	DCBLD1	1	1	1	0
ENSG00000011332	double PHD fingers 1 [Source:HGNC Symbol;Acc:HGNC:20225]	DPF1	1	1	1	0
ENSG00000114487	MORC family CW-type zinc finger 1 [Source:HGNC Symbol;Acc:HGNC:7198]	MORC1	1	0	0	0
ENSG00000157388	calcium voltage-gated channel subunit alpha1 D [Source:HGNC Symbol;Acc:HGNC:1391]	CACNA1D	1	1	0	0
ENSG00000182944	EWS RNA binding protein 1 [Source:HGNC Symbol;Acc:HGNC:3508]	EWSR1	1	0	0	0
ENSG00000163520	fibulin 2 [Source:HGNC Symbol;Acc:HGNC:3601]	FBLN2	1	1	1	0
ENSG00000115947	origin recognition complex subunit 4 [Source:HGNC Symbol;Acc:HGNC:8490]	ORC4	1	1	0	0
ENSG00000077097	DNA topoisomerase II beta [Source:HGNC Symbol;Acc:HGNC:11990]	TOP2B	1	1	1	1
ENSG00000070047	PHD and ring finger domains 1 [Source:HGNC Symbol;Acc:HGNC:24351]	PHRF1	1	0	0	0
ENSG00000111271	acyl-CoA dehydrogenase family member 10 [Source:HGNC Symbol;Acc:HGNC:21597]	ACAD10	1	0	0	0
ENSG00000074317	synuclein beta [Source:HGNC Symbol;Acc:HGNC:11140]	SNCB	1	0	0	0
ENSG00000132824	serine incorporator 3 [Source:HGNC Symbol;Acc:HGNC:11699]	SERINC3	1	1	0	0

ENSG00000070915	solute carrier family 12 member 3 [Source:HGNC Symbol;Acc:HGNC:10912]	SLC12A3	1	0	0	0
ENSG00000109956	beta-1,3-glucuronyltransferase 1 [Source:HGNC Symbol;Acc:HGNC:921]	B3GAT1	1	1	1	1
ENSG00000150556	LY6/PLAUR domain containing 6B [Source:HGNC Symbol;Acc:HGNC:27018]	LYPD6B	1	1	0	0
ENSG00000104537	annexin A13 [Source:HGNC Symbol;Acc:HGNC:536]	ANXA13	1	1	0	0
ENSG00000171723	gephyrin [Source:HGNC Symbol;Acc:HGNC:15465]	GPHN	1	1	1	1
ENSG00000106462	enhancer of zeste 2 polycomb repressive complex 2 subunit [Source:HGNC Symbol;Acc:HGNC:3527]	EZH2	1	1	1	1
ENSG00000150676	coiled-coil domain containing 83 [Source:HGNC Symbol;Acc:HGNC:28535]	CCDC83	1	0	0	0
ENSG00000164597	component of oligomeric golgi complex 5 [Source:HGNC Symbol;Acc:HGNC:14857]	COG5	1	0	0	0
ENSG00000065809	family with sequence similarity 107 member B [Source:HGNC Symbol;Acc:HGNC:23726]	FAM107B	1	0	0	0
ENSG00000120251	glutamate ionotropic receptor AMPA type subunit 2 [Source:HGNC Symbol;Acc:HGNC:4572]	GRIA2	1	0	0	0
ENSG00000115919	kynureninase [Source:HGNC Symbol;Acc:HGNC:6469]	KYNU	1	0	0	0
ENSG00000125991	ERGIC and golgi 3 [Source:HGNC Symbol;Acc:HGNC:15927]	ERGIC3	1	1	0	0
ENSG00000128487	sperm antigen with calponin homology and coiled-coil domains 1 [Source:HGNC Symbol;Acc:HGNC:30615]	SPECC1	1	0	0	0
ENSG00000172116	CD8b molecule [Source:HGNC Symbol;Acc:HGNC:1707]	CD8B	1	1	0	0
ENSG00000146090	RasGEF domain family member 1C [Source:HGNC Symbol;Acc:HGNC:27400]	RASGEF1C	1	1	1	0
ENSG00000152952	procollagen-lysine,2-oxoglutarate 5-dioxygenase 2 [Source:HGNC Symbol;Acc:HGNC:9082]	PLOD2	1	1	1	1
ENSG00000163017	actin, gamma 2, smooth muscle, enteric [Source:HGNC Symbol;Acc:HGNC:145]	ACTG2	1	1	1	0
ENSG00000100714	methylenetetrahydrofolate dehydrogenase, cyclohydrolase and formyltetrahydrofolate synthetase 1 [Source:HGNC Symbol;Acc:HGNC:7432]	MTHFD1	1	1	1	1
ENSG00000135597	RALBP1 associated Eps domain containing 1 [Source:HGNC Symbol;Acc:HGNC:15578]	REPS1	1	0	0	0
ENSG00000083312	transportin 1 [Source:HGNC Symbol;Acc:HGNC:6401]	TNPO1	1	1	0	0
ENSG00000132746	aldehyde dehydrogenase 3 family member B2 [Source:HGNC Symbol;Acc:HGNC:411]	ALDH3B2	1	0	0	0
ENSG00000198963	RAR related orphan receptor B [Source:HGNC Symbol;Acc:HGNC:10259]	RORB	1	1	1	1
ENSG00000050628	prostaglandin E receptor 3 [Source:HGNC Symbol;Acc:HGNC:9595]	PTGER3	1	1	1	1
ENSG00000165752	serine/threonine kinase 32C [Source:HGNC Symbol;Acc:HGNC:21332]	STK32C	1	1	0	0
ENSG00000134955	solute carrier family 37 member 2 [Source:HGNC Symbol;Acc:HGNC:20644]	SLC37A2	1	1	0	0
ENSG00000107186	multiple PDZ domain crumbs cell polarity complex component [Source:HGNC Symbol;Acc:HGNC:7208]	MPDZ	1	1	0	0
ENSG00000138617	poly(ADP-ribose) polymerase family member 16 [Source:HGNC Symbol;Acc:HGNC:26040]	PARP16	1	0	0	0
ENSG00000097007	ABL proto-oncogene 1, non-receptor tyrosine kinase [Source:HGNC Symbol;Acc:HGNC:76]	ABL1	1	1	1	1

ENSG00000070601	FERM and PDZ domain containing 1 [Source:HGNC Symbol;Acc:HGNC:29159]	FRMPD1	1	1	1	0
ENSG00000155066	prominin 2 [Source:HGNC Symbol;Acc:HGNC:20685]	PROM2	1	1	0	0
ENSG00000084774	carbamoyl-phosphate synthetase 2, aspartate transcarbamylase, and dihydroorotase [Source:HGNC Symbol;Acc:HGNC:1424]	CAD	1	1	0	0
ENSG00000196268	zinc finger protein 493 [Source:HGNC Symbol;Acc:HGNC:23708]	ZNF493	1	1	1	0
ENSG00000156858	proline rich 14 [Source:HGNC Symbol;Acc:HGNC:28458]	PRR14	1	1	0	0
ENSG00000178233	transmembrane protein 151B [Source:HGNC Symbol;Acc:HGNC:21315]	TMEM151B	1	0	0	0
ENSG00000080224	EPH receptor A6 [Source:HGNC Symbol;Acc:HGNC:19296]	EPHA6	1	0	0	0
ENSG00000103067	epithelial splicing regulatory protein 2 [Source:HGNC Symbol;Acc:HGNC:26152]	ESRP2	1	1	0	0
ENSG00000178996	sorting nexin 18 [Source:HGNC Symbol;Acc:HGNC:19245]	SNX18	1	1	1	0
ENSG00000166840	glycine-N-acyltransferase like 1 [Source:HGNC Symbol;Acc:HGNC:30519]	GLYATL1	1	1	0	0
ENSG00000136854	syntaxin binding protein 1 [Source:HGNC Symbol;Acc:HGNC:11444]	STXBP1	1	1	1	0
ENSG00000198546	zinc finger protein 511 [Source:HGNC Symbol;Acc:HGNC:28445]	ZNF511	1	1	0	0
ENSG00000163606	CD200 receptor 1 [Source:HGNC Symbol;Acc:HGNC:24235]	CD200R1	1	1	1	0
ENSG00000003400	caspase 10 [Source:HGNC Symbol;Acc:HGNC:1500]	CASP10	1	1	1	1
ENSG00000164867	nitric oxide synthase 3 [Source:HGNC Symbol;Acc:HGNC:7876]	NOS3	1	1	1	1
ENSG00000197181	piwi like RNA-mediated gene silencing 2 [Source:HGNC Symbol;Acc:HGNC:17644]	PIWIL2	1	1	1	0
ENSG00000138378	signal transducer and activator of transcription 4 [Source:HGNC Symbol;Acc:HGNC:11365]	STAT4	1	0	0	0
ENSG00000126878	allograft inflammatory factor 1 like [Source:HGNC Symbol;Acc:HGNC:28904]	AIF1L	1	1	1	0
ENSG00000197150	ATP binding cassette subfamily B member 8 [Source:HGNC Symbol;Acc:HGNC:49]	ABCB8	1	1	0	0
ENSG00000075391	RAS protein activator like 2 [Source:HGNC Symbol;Acc:HGNC:9874]	RASAL2	1	1	1	0
ENSG00000133135	ring finger protein 128, E3 ubiquitin protein ligase [Source:HGNC Symbol;Acc:HGNC:21153]	RNF128	1	1	1	0
ENSG00000067066	SP100 nuclear antigen [Source:HGNC Symbol;Acc:HGNC:11206]	SP100	1	0	0	0
ENSG00000136960	ectonucleotide pyrophosphatase/phosphodiesterase 2 [Source:HGNC Symbol;Acc:HGNC:3357]	ENPP2	1	1	1	0
ENSG00000172403	synaptopodin 2 [Source:HGNC Symbol;Acc:HGNC:17732]	SYNPO2	1	1	1	0
ENSG00000135637	coiled-coil domain containing 142 [Source:HGNC Symbol;Acc:HGNC:25889]	CCDC142	1	1	0	0
ENSG00000073060	scavenger receptor class B member 1 [Source:HGNC Symbol;Acc:HGNC:1664]	SCARB1	1	1	0	0
ENSG00000183060	LysM domain containing 4 [Source:HGNC Symbol;Acc:HGNC:26571]	LYSMD4	1	1	0	0

ENSG000	protein phosphatase 2 scaffold subunit Abeta [Source:HGNC	PPP2R				
00137713	Symbol;Acc:HGNC:9303]	1B	1	1	1	0
ENSG000	NOVA alternative splicing regulator 1 [Source:HGNC	NOVA				
00139910	Symbol;Acc:HGNC:7886]	1	1	1	1	0
ENSG000	farnesyl diphosphate synthase [Source:HGNC					
00160752	Symbol;Acc:HGNC:3631]	FDPS	1	1	0	0
ENSG000		CDH1				
00071991	cadherin 19 [Source:HGNC Symbol;Acc:HGNC:1758]	9	1	0	0	0
ENSG000	actin binding LIM protein family member 3 [Source:HGNC	ABLI				
00173210	Symbol;Acc:HGNC:29132]	M3	1	1	1	0
ENSG000	pericentriolar material 1 [Source:HGNC					
00078674	Symbol;Acc:HGNC:8727]	PCM1	1	1	1	0
ENSG000	cyclin dependent kinase like 4 [Source:HGNC					
00205111	Symbol;Acc:HGNC:19287]	CDKL4	1	0	0	0
ENSG000	UPF1, RNA helicase and ATPase [Source:HGNC					
00005007	Symbol;Acc:HGNC:9962]	UPF1	1	1	0	0
ENSG000		CASC				
00166734	cancer susceptibility 4 [Source:HGNC Symbol;Acc:HGNC:24892]	4	1	1	1	0
ENSG000	syntaxin binding protein 5 [Source:HGNC	STXBP				
00164506	Symbol;Acc:HGNC:19665]	5	1	0	0	0
ENSG000	protein tyrosine phosphatase, non-receptor type 13	PTPN				
00163629	[Source:HGNC Symbol;Acc:HGNC:9646]	13	1	1	1	0
ENSG000	IKBKB interacting protein [Source:HGNC					
00166130	Symbol;Acc:HGNC:26430]	IKBIP	1	0	0	0
ENSG000	boule homolog, RNA binding protein [Source:HGNC					
00152430	Symbol;Acc:HGNC:14273]	BOLL	1	1	0	0
ENSG000	NCK associated protein 1 [Source:HGNC	NCKA				
00061676	Symbol;Acc:HGNC:7666]	P1	1	1	0	0
ENSG000	solute carrier family 35 member F3 [Source:HGNC	SLC35				
00183780	Symbol;Acc:HGNC:23616]	F3	1	1	0	0
ENSG000	purinergic receptor P2X 2 [Source:HGNC					
00187848	Symbol;Acc:HGNC:15459]	P2RX2	1	0	0	0
ENSG000		OLFM				
00118733	olfactomedin 3 [Source:HGNC Symbol;Acc:HGNC:17990]	3	1	0	0	0
ENSG000	kinesin family member 20B [Source:HGNC	KIF20				
00138182	Symbol;Acc:HGNC:7212]	B	1	1	1	0
ENSG000	IQ motif and ubiquitin domain containing [Source:HGNC					
00164675	Symbol;Acc:HGNC:21995]	IQUB	1	0	0	0
ENSG000	phospholipase C gamma 1 [Source:HGNC					
00124181	Symbol;Acc:HGNC:9065]	PLCG1	1	1	1	0
ENSG000	GRB10 interacting GYF protein 2 [Source:HGNC	GIGYF				
00204120	Symbol;Acc:HGNC:11960]	2	1	1	1	0
ENSG000						
00188452	ceramide kinase like [Source:HGNC Symbol;Acc:HGNC:21699]	CERKL	1	1	0	0
ENSG000						
00120802	thymopoietin [Source:HGNC Symbol;Acc:HGNC:11875]	TMPO	1	1	1	0
ENSG000	discs large MAGUK scaffold protein 1 [Source:HGNC					
00075711	Symbol;Acc:HGNC:2900]	DLG1	1	1	1	0
ENSG000	protein phosphatase 6 regulatory subunit 3 [Source:HGNC	PPP6R				
00110075	Symbol;Acc:HGNC:1173]	3	1	0	0	0
ENSG000						
00196188	cathepsin E [Source:HGNC Symbol;Acc:HGNC:2530]	CTSE	1	0	0	0
ENSG000	citron rho-interacting serine/threonine kinase [Source:HGNC					
00122966	Symbol;Acc:HGNC:1985]	CIT	1	1	0	0
ENSG000	family with sequence similarity 131 member A [Source:HGNC	FAM1				
00175182	Symbol;Acc:HGNC:28308]	31A	1	1	0	0

ENSG000 00133612	ArfGAP with GTPase domain, ankyrin repeat and PH domain 3 [Source:HGNC Symbol;Acc:HGNC:16923]	AGAP 3	1	1	0	0

Supplementary Figure S1. Entity Relationship Diagram (ERD): Description of how data sources were merged in this investigation and the identical keys used to integrate them. There were 4 initial data sources each with its own form of unique ID we were able to use some accessory tables to create essentially one centrally used ID for each variant, the Ensembl Transcript ID and one centrally used ID for each sample, the TCGA barcode unique identifier.