Predicting Mitochondrial tRNA Modification

Diego Calderon
Stanford University
dcal@stanford.edu

Mitochondria are integral to proper cell function, and mutations in its small genome (mtDNA) are associated with many diseases, along with the progression of normal aging [9]. While mtDNA has been extensively studied, not much is known about transcriptional variations of mitochondrial genes [3]. Recently, tRNA modifications have been the focus of intense study owing to their putative role in diseases [4]. Identifying genes responsible for mitochondrial tRNA modification is an important step towards a better understanding of transcriptional variation. To the best of my knowledge, a machine learning approach has never been utilized in order to identify such genes. Here, I used penalized linear regression to predict tRNA modification activity using gene expression as features. After controlling for confounding factors, I was able to predict modification activity at several putative modified tRNA positions. These models explained between 19% and 51% of the variance. Most notably, I used preconditioned lasso [7] which yielded four promising gene candidates that affect tRNA modification: ALKBH8 (an *E. coli* homolog DNA repair enzyme), MAD2L1 Binding Protein, C1orf103 (an open reading frame gene with unknown function), and TRMT5 (tRNA methyltransferase 5).

Introduction

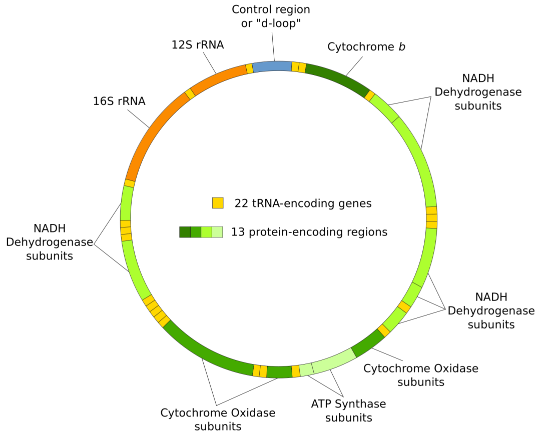
Central Dogma of Biology

In 1956, Francis Crick is credited with first articulating the flow of sequencing information in a cell known as the Central Dogma of Biology [1]. Essentially, this Dogma states that DNA sequence (which serves as data storage) is transcribed to messenger RNA (mRNA transcripts) and later translated into proteins, which are nano-scale machines that do most of the functional work in the cell. This flow of information provides several points in which the cell can regulate production of protein, the most studied of which being gene transcription. However, since then scientists have discovered mechanisms by which mRNA transcripts are postranscriptionally modified, which can have downstream effects on protein expression. These kinds of RNA modifications are typically studied in nuclear genes, and have not been well characterized in mitochondrial RNA transcripts. However, because of the mitochondrion's unique mechanism of gene transcription and translation, I will argue that certain RNA modifications can have significant downstream functional effects.

Mitochondrial Punctuation Model

Independent of nuclear DNA, mitochondria have their own genome (mtDNA), which is circular (similar to bacterial genomes), 16,569 base pairs long, and contains 37 genes which consist of 2 rRNAs, 22 tRNAs, and 13 proteins involved with the oxidative phosphorylation pathway (see figure 1)[1]. Mitochondria are not self-sufficient and rely on many proteins encoded by the nuclear genome. When the mitochondria transcribe the genes located in mtDNA, they produce three polycistronic (transcripts with more than one gene) mRNA sequences where the coding proteins are separated (or punctuated) by tRNA genes. These tRNA (or transfer RNA) genes are different from mRNA because they are not translated, but instead tRNAs aid in the translation process, by helping ribosomes construct proteins from amino acids. However, in the mitochondrion, tRNAs likely have a second function. Scientists hypothesize that the clover-like structure of tRNA regions in the polycistronic reads serve as a binding site for nuclear encoded cleaving proteins that separate the mitochondrial genes from the polycistronic read allowing them to continue through the process of translation [6]. Given the importance of tRNA structure in the punctuation model of mitochondrial gene translation, it is plausible that modifications of tRNA structure could affect protein translation and functional properties of the mitochondrion. In fact, Liu et al., showed that perturbing the editing activity of a tRNA modifier protein resulted in defects of mitochondrial function in cardiomyocytes [4].

Figure 1: *Mitochondrial genome, which consists of 16,569 bps and contains 37 genes.*


Control region
or "d-loop"
12S rRNA
Cytochrome *b*
16S rRNA
NADH Dehydrogenase subunits
22 tRNA-encoding genes
13 protein-encoding regions
NADH Dehydrogenase subunits
NADH Dehydrogenase subunits
Cytochrome Oxidase subunits
ATP Synthase subunits
Cytochrome Oxidase subunits

Quantifying tRNA Modification

In the last five years there have been significant improvements in the technology used to decode DNA (DNAseq) and RNA (RNAseq) sequencing reads. By sequencing a

[1] upload.wikimedia.org/wikipedia/commons/3/3e/Mitochondrial_DNA_en.svg

tissue or cell sample, a scientist can decode all the DNA and RNA into a human readable format (G,C, T, A, and RNA have a special base pair represented as U instead of T). The output of sequencing are millions of reads that consist of strings of nucleotide base pairs (typically between 75-150) which represent discrete chunks of molecules of genetic information. Researchers then assemble these reads like a jigsaw puzzle (reads typically have some overlap between each other), and the number of reads that align to a specific position in a genome represent the confidence with which they predict a specific base pair to be located there.

In addition to obtaining decoded genetic information, by comparing DNA sequencing to RNA sequencing data, scientists can identify modified RNA positions. For example, if DNA sequencing confidently predicts the presence of an A at a specific position, and RNA sequencing instead (but also confidently) predicts the presence of a G at that same base pair, then the evidence suggests that after transcription the RNA position was modified. Hogdkinson et al. [3], compared mitochondrial DNA and RNA sequencing data from a Canadian population of nearly 1000 individuals to identify RNA modification events. At eleven positions of mitochondrial tRNA (specifically p9 positions) they found systematic significant differences between the DNA and RNA predicted base pair, and from this evidence they hypothesized the presence of a RNA modification event at these positions. To quantify the amount of tRNA modification, they used the RNA sequencing data and calculated the ratio of read counts that predicted the usual base pair at that position and the number of reads that predicted the modified nucleotide. For example, if we look at a read distributions at one of mitochondria's tRNA positions, let's say the DNA sequencing data has 1000 reads that align to A and zero reads that align to the other nucleotides (representing a high likelihood that the position is an adenosine), however the RNA sequencing data at this same position has 400 reads aligned to A, 200 reads to C, 200 reads to G, and 200 to U. Then, a surrogate for the quantity of RNA modification would be the alternative allele frequency, in our example case:

$$\frac{200 + 200 + 200}{200 + 200 + 200 + 400} = 0.6$$

The premise is that during RNA sequencing, modified bases are not correctly identified, but instead are stochastically misread as different nucleotides. With a summary statistic for tRNA modification in place we must find a way to identify genes responsible for this mechanism.

Genes Involved with tRNA Modification

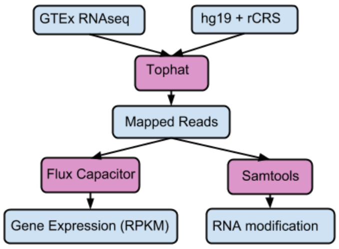
The prototypical method for identifying genes associated with a phenotypic trait is called a genome-wide association study (GWAS). When performing a GWAS, the researcher must associate single nucleotide mutations (referred to as SNPs) to the trait of interest. Let's say we are studying diabetes and are interested in genes that are correlated with blood glucose levels. A GWAS is an iterative search for SNPs that significantly affect the trait of interest, so maybe a interesting target would be a gene where individuals that have a G in one of the positions instead of a C base pair are more likely to have elevated blood glucose levels. Hodgekinson et al., calculated a RNA modification summary statistic for all the putatively modified tRNA positions in each of the nearly 1000 individuals, and then performed a genome-wide association study (GWAS) to identify SNP genotypes associated with variations of RNA modification. They found several significant hits, the most significant hit was located in the MRPP3 gene region that encodes for a protein associated with the mitochondria and methylation activity.

This work ties tRNA modification to a biological mechanism, however GWAS tend to identify genomic regions of interest at a low resolution, and it is quite possible that not all genes responsible for variations of tRNA modification were identified. Instead of performing a GWAS, I utilized a machine learning approach to predict tRNA modification values, and identified genes of interest using feature selection. In this context, my feature matrix will consist of p protein coding gene expression values for n observations, and the continuous response variables are the tRNA modification summary statistics.

Materials & Methods

To construct my feature matrix and response variables, I used data from the Genotype-Tissue Expression (GTEx) project. The current data freeze has fastq (sequence read format) and bam (sequence read format after aligned to the genome) files located on the Stanford Genetics department cluster, scg3. I analyzed 3,203 whole transcriptome RNAseq samples from 212 individuals and 30 different tissues.

Figure 2: *Data processing workflow diagram. The blue boxes represent data sources, and the red boxes are data processing programs.*



Data Processing

The reads contained in the fastq files are aligned (jigsaw puzzle solving process relying on read overlap) to the hg19 reference genome concatenated with the rCRS mitochondrial reference genome[2] using TopHat (a popular short read aligner that is aware of gene splicing). Flux capacitor then processes the aligned reads to extract gene expression values (in RPKM units) per gene. Finally, I use the samtools package to calculate the modification summary statistic of the eleven previously identified p9 tRNA positions described in Hodgkinson et al., (585, 1610, 2617, 4271, 5520,

[2] Over millions of years fragments of mtDNA have been inserted into the nuclear genome, these fragments are called nuclear mitochondrial genes (NUMTs). Thus, alignment of the reads to both the nuclear and mitochondrial genome is important to avoid false mapping of NUMT reads to the mitochondria.

7526, 8303, 9999, 10413, 12146, 12274, 13710, and 14734). See figure 2 for the flow diagram of data processing.

Due to quality control concerns (low coverage), I restricted my analysis to only four tRNA positions: 2617, 13710, 14734, 12274. Also, I calculated two meta-response variables, one was the average summary statistic for the four selected tRNA positions, and the other was mean summary statistic over all the tRNA positions (including the low-quality base predictions). My gene expression feature matrix was restricted to protein coding genes (as identified by Gencode annotation) that are differentially expressed in the samples (i.e., the variance of gene expression values is greater than 0). This process narrows down the number of features from over 55,000 to 18,780 genes.

From the total set of observations I randomly selected 20% of the observations (or 641 samples) to use as a test set after generating the prediction model. I normalized both the gene expression values and response variable of the training set to have a mean of 0 and variance of 1 by subtracting each gene's respective mean and dividing by each each gene's respective variance. To normalize the test data, I centered and scaled the values by the mean and variance of the training set's respective gene or response variable's mean and variance. This process allows me to keep the test data independent from the training set and still normalize the data under the assumption that the values were produced from the same distribution.

In addition to regressing on the expression values, I controlled for known confounders by performing regression with GTEx metadata. This included metadata such as: height, weight, gender, age, batch id (samples sequenced separately could have different QC characteristics), and tissue, which were all located on the scg3 server. Controlling for confounding variables is an essential step that removes extra noise from the signal.

Model Generation & Feature Selection

Regularized Linear Regression

Since the number of features (approximately 18,000 differentially expressed protein coding genes) far outnumbers the number of observations (3,203 GTEx RNAseq samples), linear regression is not feasible. Instead, I used the `glmnet` package in R which implements a mixture of L_1 and L_2 penalized linear regression by optimizing the following objective function:

$$\min_{\beta_0,\beta} \frac{1}{N}\sum_{i=1}^{N} w_i l(y_i,\beta_0+\beta^T x_i)+\lambda\left[(1-\alpha)||\beta||_2^2/2+\alpha||\beta||_1\right]$$

The goal is to find the vector of β coefficients that minimizes the objective function. On the left we see the familiar linear regression objective, and on the right we have both the L_1 and L_2 norm of the β coefficients, which represents the penalty. When optimizing the objective function there are two parameters to select, λ, which controls the strength of the penalty, and α that controls the proportion of the L_1 or L_2 norm that is mixed into our regularization penalty. I performed 10-fold cross validation on the training set to select the optimal λ value, and I set α to 1, which is referred to as lasso regression. I chose to use lasso regression, as opposed to ridge regression (α set to 0 resulting in less sparse solutions) or elastic net (α set to between 0 and 1, resulting in coefficient sparsity between lasso and ridge regression), because I prefer a sparser solution (lasso tends to shrink coefficients to 0), and accuracy gains made by varying α were minimal. Models were evaluated by calculating the Mean Square Error (MSE) between the prediction and the response variable, and relying on variance explained as a metric for determining the amount of signal captured, calculated by

$$1-\frac{var(\hat{y})}{var(y)}$$

where y is the original response and $\hat{y}$ is the regression residual.

Controlling for Covariates

Originally, my model explained tRNA modification as a function of gene expression and random standard gaussian distributed noise or

$$y \sim \text{expression} + \epsilon$$

however, this does not take into account known covariates that could confound the signal. To identify metadata correlated with the response variable, I modified my model so that tRNA modification is a result of a linear function of batch id, tissue, gene expression and random standard gaussian distributed noise or

$$y \sim \text{batch} + \text{tissue} + \text{expression} + \epsilon$$

The other metadata values did not have much predictive power so they were excluded. To observe the quantity of signal loss, I first ran regression on the metadata then regressed the gene expression values on the residuals.

Tissue-Specific Models

In an attempt to minimize the amount of signal lossage when I regress out tissue-specific effects, I generated several models, separately, using only the more numerous single tissue types. Specifically, I generated models using observations from brain, skin, blood, and muscle, of which there were 424, 347, 271, and 176 respective samples. Once again, I held out 20% for testing. In the results section, for brevity, I only include analysis on the model generated with muscle tissue samples.

Training Set Size Analysis

To evaluate instability of training a model on a smaller data set, I performed a training set size analysis. This consisted of iteratively training lasso models on a training sets of sizes 100 to 2600 by increments of 100 observations, and testing the generated model on a test set of size 600.

Preconditioning Lasso

While successfully predicting tRNA modification was one of this project's main goals, it was equally important to identify a subset of genes that putatively affect mitochondrial tRNA modification. Preconditioning lasso is a method used in high-dimensional problems and can reduce negative

Table 1: *Lasso model error rates for all the response variables, before and after controlling for confounding factors.*

Response	Train1	Test1	Train2	Test2
Top 4	0.40	0.70	0.20	0.58
All	0.62	0.79	0.45	0.69
2617	0.48	0.68	0.39	0.60
13710	0.35	0.61	0.27	0.56
14734	0.46	0.62	0.22	0.50
12274	0.92	0.63	0.60	0.59

effects of a noisy response variable under the assumption of a latent variable model, as described in [7]. This method effectively balances the tradeoff between low test-error rate from supervised PCA and sparse lasso solutions [2].

The first step of this method is to run supervised Principal Components (which is implemented in the `superpc` package in R). Briefly, this consists of calculating univariate regression coefficients for each of the normalized features, and then for several monotonically increasing threshold values, we calculate the first m principal components (PCs) on a subset feature matrix consisting of all variables with larger absolute value univariate coefficients than the specified threshold (going through the list of thresholds iteratively), and using these PCs in a regression model to predict the outcome. Threshold values and m are selected using cross validation.

After we select the m most significant PCs (constructed from features with coefficients greater than the selected threshold value), we generate a predicted response variable from all the training set observations, and finally run lasso (with all the features) to try and predict the response variables generated with supervised PCA.

Results & Discussion

Regularized Regression & Controlling for Covariates

The first training and test set error rates were calculated using a model generated with batch id and tissue metadata information, while the second set was calculated using the residuals after controlling for tissue and batch (see table 1). Metadata was best able to explain tRNA modification at position 13710, and least capable at position 12274. Strangely, the latter position had a lower test error rate than training error rate. For most of the positions there was a disparity between the metadata predicted training and testing error rate, signifying that there was still some overfitting occurring. This could be the case because batch id is a categorical variable and it is possible that there were not enough samples in a single batch to form a useful model.

After removing batch and tissue effect confounding, we see more modest decreases in error rate (columns labeled with 2 in table 1). The biggest decrease in error rate was at position 14734 for both the training set and the test set. However, once again there was a disparity between training and test set error.

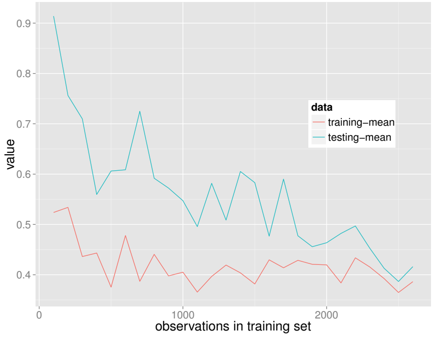
Tissue-Specific Models

Visualizing predictions compared to the top 4 base pair averaged response, we see a dramatic decrease in linearity after controlling for confounding factors (figure 3A to B). While controlling for batch is obviously beneficial, I hypothesized that some of the noise removed when I controlled by tissue also removed some of the signal. Thus, I created several models using single tissues, which, are more genetically homogenous. Figure 3 shows the results of generating a model using 118 muscle samples. Notice that from 3C to 3D, we see that less signal is lost, however the R^2 values are quite low. This is likely the case because there are simply not enough training observations.

Training Set Size Analysis

To explore this idea further, I generated lasso models predicting the top 4 base pair averaged response using increasing numbers of observations as a training set and testing on a fixed test set of 600 observations. Figure 4 shows the results of 10 replicate runs over 26 different training set values. As the number of training set observations increases, we see that the training set and test set mean error converge, however they both plateau at approximately 0.4. For the previous muscle tissue analysis, the model would be towards the left side of figure 4, thus explaining the lack of an appropriate fit.

Figure 4: *Plot of training data size analysis. MSE and training set size, plotted in y and x-axis, respectively.*



Preconditioned Lasso

Each of the models I described previously had, on average, 200 features, even though I chose lasso for its tendency to produce sparse solutions. For predicting the amount of tRNA modification this is fine, however for gene validation I need a method of reducing or ranking these genes of interest to something more manageable. For this reason, I ran preconditioned lasso.

After removing batch and tissue effects, running preconditioned lasso using the top 4 base pair average response resulted in a training error rate of 0.4 and a test error rate of 0.6, from a model with only four gene features. While this method did not produce as low of a training error rate, the disparity with the test error decreased significantly, a sign of less overfitting. In addition, four gene features is a manageable list of interesting genes. The four genes were (in decreasing order of model coefficient size): ALKBH8, MAD2L1BP, C1orf103, and TRMT5. ALKBH8 is an alkylated DNA repair protein homolog, with known tRNA

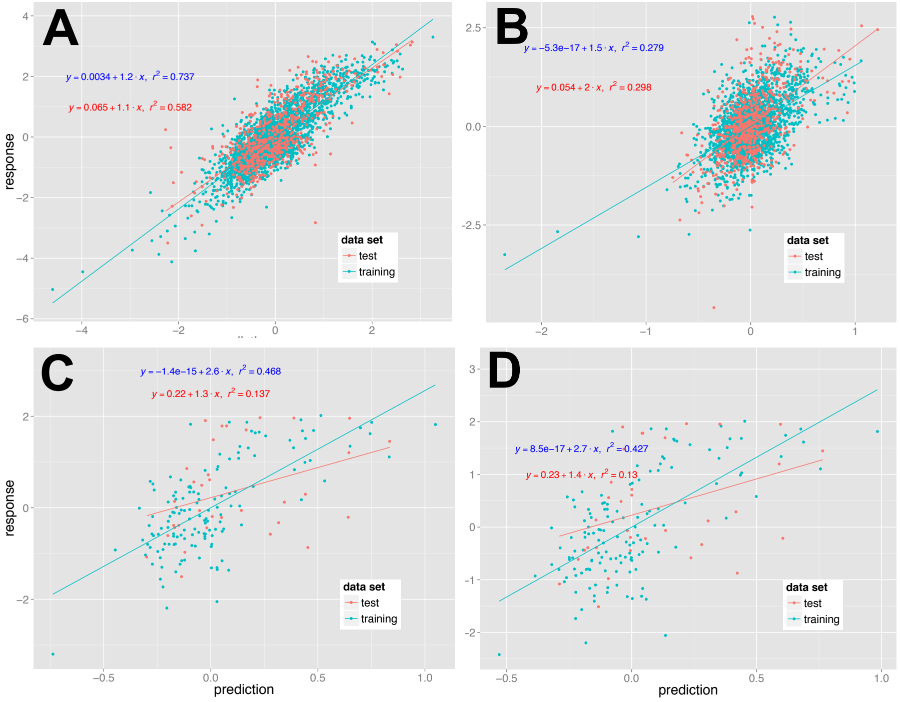


Figure 3: *Response and model prediction values, in y and x-axis. A) Model predictions including metadata trained on the training set. B) Model predictions after removing batch and tissue confounding effects. C) Model predictions including metadata trained on only 118 muscle samples. D) Model predictions after removing batch and tissue confounding effects trained on only muscle samples.*

modification functionality. In fact, there is a paper titled, “ALKBH8-mediated formation of a novel diastereomeric pair of wobble nucleosides in mammalian tRNA”. Also, it appears as though other homologs of ALKBH8 have mitochondria specific activity. MAD2L1BP is a protein that binds to MAD2L1, a cell cycle regulator, otherwise not much known is about it, however it does localize to the mitochondria[3]. Again, not much is known about C1orf103, the name itself simply stands for Chromosome 1 Open Reading Frame. Finally, there is TRMT5, which stands for tRNA methyltransferase 5 homolog. According to Suzuki et al., TRMT5 has a yeast homolog that modifies tRNAs in the cytoplasm and mitochondria in yeast, and thus is suspected to play a role in mitochondrial function.

Conclusions & Future work

Models generated with lasso achieved varying degrees of success in predicting the surrogate for tRNA modification. After controlling for batch and tissue-specific effects, my best model was able to explain 41% of the training set variance and 20% of the test set variance (position 14734). Tissue-specific gene expression was correlated with the response, and the two signals were difficult to untangle (compare linearity of figure 3A and 3B). Therefore, I tested models trained on a single tissue type (figure 3C & 3D). However, after running a training set size analysis (figure 4), I noticed that there were not enough samples to make accurate predictions.

Besides prediction, the various models resulted in over 200 genes of interest. To reduce this list I ran preconditioned lasso, which is a two part method involving supervised principal components analysis followed by lasso. Preconditioning lasso resulted in a model with only four genes of interest: ALKBH8, MAD2L1BP, C1orf103, and TRMT5. Not much is known about MAD2L1BP or C1orf103. However, ALKBH8 is a known tRNA modifier, with homologs that are associated with mitochondria, and TRMT5 is tRNA methyltransferase with mitochondrial activity. It would be interesting to follow up with functional studies conducted on both of these genes.

I have some doubts as to the effectiveness of some of the tools I used in my pipeline for generating the feature matrix and response variable (flux capacitor has received significant criticism as of late). My first step would be to re-run it. Secondly, I would prefer to look at a RNA sequencing data from the same tissues so that there is not so much tissue-specific noise. Finally, validating genes of interest (ALKBH8 and TRMT5) would be an extremely important future step. One possibility would be to perform a knockout (eliminate gene expression of target gene) of these genes in human cell lines, then sequence the mitochondria to observe any changes in tRNA modification.

Acknowledgements

I would like to thank Dr. Jonathan Pritchard for advising me, as well as the rest of the Pritchard Lab for their feedback and suggestions. Funding for the opportunity to do this research was generously provided by the NLM training grant.

References

[1] Crick, Francis H. (1958) On protein synthesis. *Symposia of the Society for Experimental Biology*, Vol. 12.

[2] Hastie, Trevor, et al. (2009) The elements of statistical learning *New York: Springer*, Vol. 2. No. 1.

[3] Hodgkinson, Alan, et al. (2014). High-Resolution Genomic Analysis of Human Mitochondrial RNA Sequence Variation. *Science*, 344(6182), 413-415.

[4] Liu, Ye, et al. (2014) Deficiencies in tRNA synthetase editing activity cause cardioproteinopathy. *Proceedings of the National Academy of Sciences*, 201420196.

[6] Ojala, Deanna, et al. (1981). tRNA punctuation model of RNA processing in human mitochondria. *Nature*, 470-474.

[7] Paul, Debashis, et al. (2008). “Preconditioning” for feature selection and regression in high-dimensional problems. *The Annals of Statistics*, 1595-1618.

[8] Suzuki, Tsutomu, et al. (2011) Human mitochondrial tRNAs: biogenesis, function, structural aspects, and diseases. *Annual review of genetics*, 45: 299-329.

[9] United Mitochondrial Disease Foundation. Accessed November 14th, 2014. `www.umdf.org`

[3] `www.genecards.org/cgi-bin/carddisp.pl?gene=MAD2L1BP`