PROBING MULTIVARIATE BREAST CANCER DATA TO PREDICT SILENT CODON MUTATION EFFECTS ON PROTEIN EXPRESSION

Brentsen Wolf, Alec Blair, Dr. Ronald Worthington

Southern Illinois University Edwardsville

Introduction

Synonymous mutations, those which represent a change in the codon nucleotides without affecting the resultant amino acid, have largely been considered inconsequential because the primary sequence of the peptide is not altered. Recently, evidence has surfaced that these mutations do have consequences that affect protein folding kinetics which can lead to structural and functional changes that alter substrate affinity.[1] The mechanism of consequence begins with aminoacyl-transfer RNA (tRNA), which carries a single amino acid and is used to elongate a peptide chain. Ribosomes read codons on the messenger RNA (mRNA), which have corresponding anti-codons on the tRNA, and use both to build proteins. There is redundancy in both mRNA and tRNA in that multiple codons and anti-codons have the same amino acid specification. Codons, and thus their corresponding anti-codons, are used at different frequencies when coding for specific amino acids. For example, isoleucine has three codons that code for it, ATC, ATT, and ATA which are used in human cells 48%, 36%, and 16% respectively.

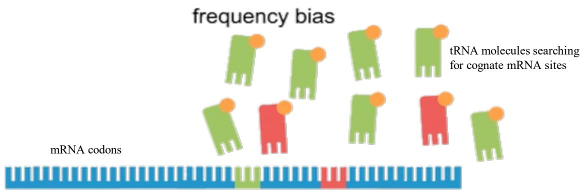

Figure 1: Green and red codons in an mRNA molecule are shown with corresponding floating tRNA molecules whose colors match their cognate mRNA targets. This shows the frequency bias for the green codon as there are more readily available green tRNA molecules.

The abundance of each tRNA is proportional to the use frequencies of each codon. Synonymous mutations that lead to a large change in codon frequency, and thus tRNA abundancy, can alter protein structure and potentially lead to altered function and disease progression.[2] Mucins (MUC) have been implicated in a variety of cancers. Functionally speaking, mucins help to maintain homeostasis through epithelial protection from environmental damage or insult. The secreted mucins form physical barriers in the respiratory and intestinal tracts as well as in organs like the kidney, breast, and pancreas.[3] However, mucins can contribute to oncogenesis and metastasis through aberrant glycosylation and expression which frequently leads to immunosuppression via receptor masking and cytolytic inhibition.[4] Mucin genes span from MUC1 through MUC22 and are comprised of both transmembrane mucins and secreted mucins. MUC6 is a secreted mucin that has been marginally linked to a variety of cancers including pancreatic, breast, and intestinal cancers. MUC6 is located on chromosome 11 and consists of 2439 amino acids. Most of the mucins associated with cancer are transmembrane mucins, most notably MUC1. In fact, there are a multitude of clinical trials targeting MUC1 using a variety of techniques.[5] However, secreted mucins such as MUC2, MUC5b, and MUC7 have also been linked to cancer through chronic inflammation and accumulation of inflammatory cells. It is possible that the function of MUC6 is most like that of MUC2, which plays an anti-inflammatory role, as both are secreted mucins that are clustered on chromosome 11 (along with MUC5AC and MUC5B). Loss of MUC2 function has been linked to colon cancer, activation of inflammatory responses, ulcerative colitis, and superficial erosions. The purpose of this research is to explore the effects of synonymous mutations in breast cancer cells. MUC6, a secreted mucin demonstrating high susceptibility to synonymous somatic mutation, was critically analyzed in an attempt to explore how it might propagate human breast cancer specifically.

Methods

Using the NCBI Conserved Domains resource, we searched for regions of homology with a variety of superfamilies including von Willebrand factor type D domain (VWD) and Herpes BLLF1.[6] Gene expression and mutation data was collected and generated from The Cancer Genome Atlas (TCGA) Research Network, and an analysis was performed on 69 breast cancer patients that had passed a quality assessment in a previous publication by Mertins, Philipp et al.[7] The gene expression data consisted of RNA-seq data in the form of fragments per kilobase of transcript per million mapped reads (FPKM) with a corresponding Ensembl Stable ID. All of the Ensembl Stable IDs were converted to HGNC symbols to match that of the HUGO Gene Nomenclature Committee database.[8] This conversion helped to analyze and visualize the data. Protein production data in the form of iTRAQ ratios which were collected and analyzed for each patient was also downloaded from the TCGA project. The mutation and gene expression data were analyzed using a variety of linear and non-linear methods including PCA, MFA, t-SNE, and random forest. PCA and MFA analysis were conducted in R-Studio using the FactoMiner package, while both t-SNE and random forest were conducted using Python in the Anaconda integrated development environment. Looking at MUC6 specifically, an analysis was conducted to explore which positions were most affected by both synonymous and missense mutations, which amino acids were most frequently mutated, and if there was any conservation of specific mutations across a large percentage of the cohort.

Results

Using the NCBI Conserved Domains resource, we found that the region from position 1402 to 2102, a span of 700 amino acids, shares homology with Epstein Barr Virus (EBV) as part of the Herpes BLLF1 superfamily (Figure 2). Interestingly, every synonymous and missense mutations in MUC6 across all of our patients fall within the EBV range of homology according to the NCBI Conserved Domain prediction. EBV, a herpes virus spread through saliva, increases the risk of a variety of cancers including Burkitt lymphoma, some types of Hodgkin's and non-Hodgkin's lymphoma, and perhaps most notably, gastric cancers.[9]

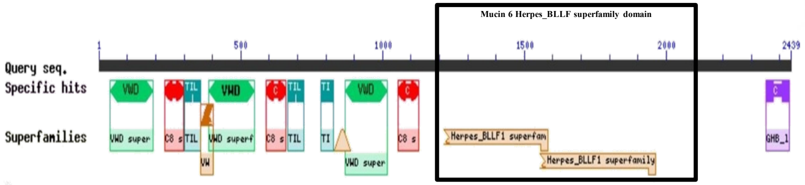

Figure 2: MUC6 is a member of the large family of glycoproteins that is typically secreted by epithelial cells. MUC6 is located in chromosome 11 containing 2439 amino acids. The region from 1402-2102 is of special interest as this correlates with *all* of the synonymous and missense mutations in MUC6 across all of our patients.

Additionally, MUC6 proved to be the most hypermutated gene across the entire patient population. When comparing MUC6 to the top 20 most synonymously mutated genes, patients had an average of 20.6 mutations in MUC6 followed by 1.5 mutations in MUC16, and 0.4 mutations in SNF587 (Figure 3).

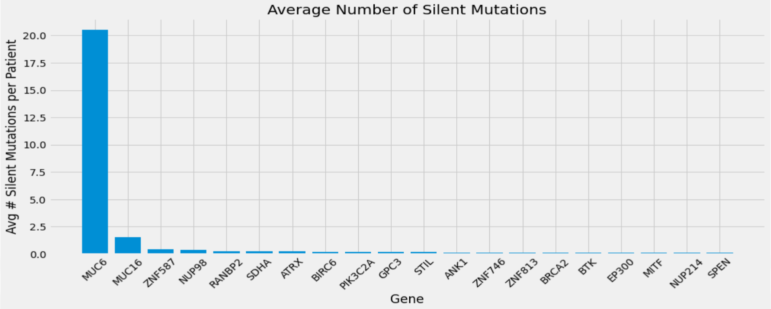

Figure 3: Top 20 most synonymously mutated genes in dataset ranked 1-20

Interestingly, when exploring which amino acids were being affected by the synonymous mutations in the MUC6 gene, there appeared to be a large bias towards threonine (Thr/T) with close to 300 mutations in the ACC codon followed by the CAC codon which codes for histidine (His/H) as the second most affected amino acid (Figure 4).

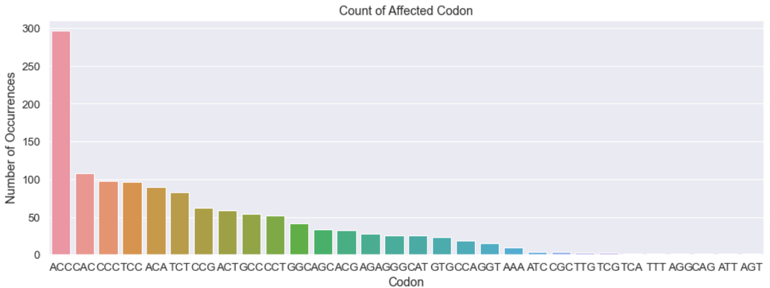

Figure 4: Ranking of the most synonymously mutated codons

When looking at the 20 most frequently mutated positions, threonine occupies nearly half of the locations (40%) at positions 1949, 1881, 1954, 2004, 1997, 1993, 1973, and 1967. Many of the synonymous mutations appear to be conserved across the patient population with over 80% of patients having the exact same mutations at positions 2039 and 1949. A similar phenomenon occurs when looking at missense mutations. The number of missense mutations in MUC6 far exceeds that of all other genes with an average of 29.8 missense mutations in MUC6 compared to the second most mutated gene TAS2R31 with an average of 6.6 mutations per patient (Figure 5).

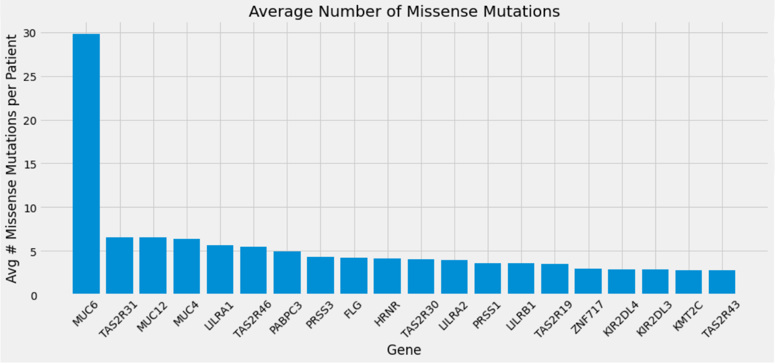

Figure 5: Top 20 most missense mutated genes in dataset ranked 1-20

PCA was utilized in R-studio to help visualize a possible relationship between codon ratio change and iTRAQ value or cancer stage for synonymous mutations in MUC6 specifically. The codon ratio is the frequency of use of the old codon divided by that of the new codon. For example, in the case of an isoleucine mutation from ATC to ATA there would be a frequency change from 48% to 16% and therefore a codon ratio of 3. We then took the log2 of these ratios for further analysis using PCA. The correlation graph (Figure 7) shows the relationship between the codon ratios, iTRAQ ratios, and cancer stage as a result of PCA.

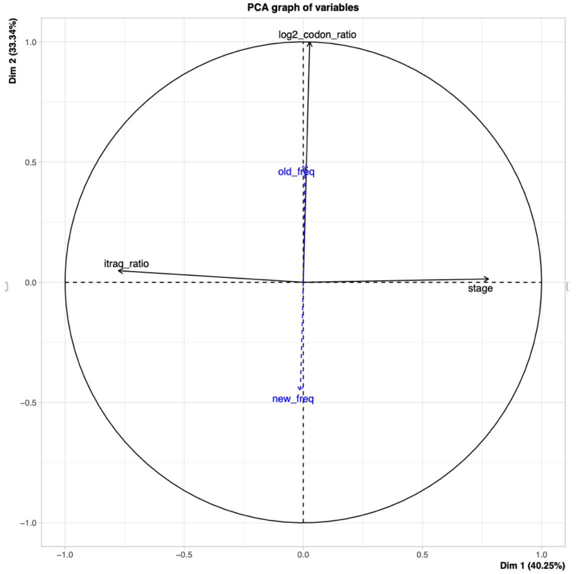

Figure 6: Correlation circle graph showing relationship of log2 codon ratio, iTRAQ, and stage across all PAM50 subtypes with synonymous mutations in MUC6

Discussion

The mucin class, especially MUC1, has been subject to extensive analysis regarding its role in cancer. However, MUC6, and the secreted mucins in general, have far less data to support both their biological role and ability to contribute to the propagation of certain cancers. Here we provide evidence that breast cancer may be partially propagated in our patient population by driving synonymous and missense mutations in MUC6 leading to decreased expression, and therefore creating an environment of chronic inflammation and an accumulation of inflammatory cells. Furthermore, the results of the PCA, MFA, and Random Forest analysis provide evidence that synonymous mutations are related to protein expression in breast cancer. However, these results are limited due to a small patient population which may lead to an overfitting of the models. Further research can build on the MUC6 story, especially in the gastrointestinal cancer spaces such as colon and stomach cancer, which is where most of the data supporting the role of mucins in cancer exists. Exploring whether the extreme hypermutability of MUC6 for both synonymous and missense mutations exists in cancers outside of breast cancer could provide further insight into its oncogenic potential and mechanism of propagation. Additionally, it should be observed whether the region of EBV homology remains the location of a significant portion of the synonymous and missense mutations, and if this homology may lead to further explanation of MUC6's role in cancer. In summary, we have discovered a new biomarker in breast cancer of all four PAM50 subtypes. Silent mutations in MUC6 may have predictive value in understanding the status of the tumor regarding protein expression and the selective pressures being imposed on the tumor by the patient's immunological responses at the time of biopsy. Using expression and mutation data in addition to machine learning techniques such as Random Forest may develop a path towards the creation of cancer risk stratification tools, protein expression prediction algorithms, and several other potentially useful assessments that can provide clinical benefit to patients and reduce costs to researchers.

References

1. Komar A (2007) NPs, Silent but not invisible. Science 315:456–457
2. Sharp PM, Tuohy TMF, Mosurski KR (1986) Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res 14(13):5125–5143
3. Purvis IJ, Bettany AJ, Santiago TC et al (1987) The efficiency of folding of some proteins is increased by controlled rates of translation in vivo. A hypothesis. J Mol Biol 193:413–417
4. DW. Mucins in cancer: function, prognosis and therapy. Nat Rev Cancer. 2009;9(12):874-885. doi:10.1038/nrc2761
5. Taylor-Papadimitriou J, Burchell JM, Graham R, Beatson R. Latest developments in MUC1 immunotherapy. Biochem Soc Trans. 2018;46(3):659-668. doi:10.1042/BST20170400
6. Shennan Lu et al. (2020), "CDD/SPARCLE: the conserved domain database in 2020.", Nucleic Acids Res.48(D1)265-8.
7. Mertins, Philipp et al. "Proteogenomics connects somatic mutations to signalling in breast cancer." Nature vol. 534,7605 (2016): 55-62. doi:10.1038/nature18003
8. Tweedie S, Braschi B, Gray KA, Jones TEM, Seal RL, Yates B, Bruford EA. Genenames.org: the HGNC and VGNC resources in 2021. Nucleic Acids Res. PMID: 33152070 PMCID: PMC7779007 DOI: 10.1093/nar/gkaa980
9. Nishikawa, Jun et al. "Clinical Importance of Epstein⁻Barr Virus-Associated Gastric Cancer." Cancers vol. 10,6 167. 29 May. 2018, doi:10.3390/cancers10060167

We gratefully acknowledge research funding support from the SIUE URCA program (B. Wolf) the SIUE School of Pharmacy Research Program (B. Wolf, A. Blair, R. Worthington), and the AFPE Gateway to Research Fellowship (B. Wolf).