TGEN WOMEN'S PHILANTHROPY COUNCIL

Fueling the Future of Medicine

tgen part of City of Hope

Project Update

Early Detection of Cancer in Blood by Quantifying Chromosome Changes

Reporting Period: 12/01/22 — 05/30/23

Kamel Lahouel, PhD

Project Summary

Detection of aneuploidy—the presence of an abnormal number of chromosomes in a cell—has been explored as a promising method for the evaluation of various cancer types. DNA from cancer cells is shed into the bloodstream, enabling the analysis of cell-free DNA to identify chromosomes with pronounced alterations.

Our project focuses on building a robust test for the early detection of cancer by detecting aneuploidy from an amplicon-based approach called BestSeq that targets repetitive elements of the genome. Amplicons can be thought of as magnifying glasses that help us zoom in on specific genome regions. These amplicons are designed to target specific genomic regions and amplify the signals originating from those regions by creating multiple copies of the DNA template in those areas, provided that the template region is intact and not fragmented.

In our specific case, the regions that are targeted are repetitive elements of the DNA: these are sequences in the DNA that are redundant and share patterns. We hypothesize that the difference between normal and cancer DNA patterns is more pronounced in those regions. The advantage of an amplicon-based approach resides in the low amount of DNA needed as a template to amplify only the signal coming from repetitive elements, compared to a whole genome sequencing approach that is not magnifying any specific region, where more sequencing depth and hence higher cost are required to pick up the same signal at the targeted regions.

The idea is to quantify the amount of DNA coming from repetitive regions that are amplified by our approach. Each region can be mapped to a specific chromosome. Therefore, we can estimate the amount of amplification of every chromosome and determine whether it has a normal signal, under-expressed or over-expressed. This allows us to pinpoint aneuploidy coming from a particular chromosome for a given patient and hence detect a cancerous signal.

Budget Utilization

We were fortunate to receive funding from the Women's Philanthropy Council which enabled us to hire and provide partial salary support to Ms. Manasa Upadhyaya as a bioinformatician in the Tomasetti lab, working on implementing the quantification algorithm and testing its performance.

In addition, the funding allowed us to acquire a pilot dataset for proving concordance between the aneuploidy signal detected from cfDNA and chromosomal gains or losses observed in the primary tumor. This dataset consists of 11 patient-matched primary (gastric) tumors with their respective plasmas. This specific dataset will be used in the last phase to confirm that the classifier is calling aneuploidy in plasma on the chromosomal arms where aneuploidy was observed in the primary tumor.

Current State of the Methodology

We are currently developing and testing our methodology on an already existing amplicon-based technology called RealSeq. Amplicon-based technology is a highly targeted approach that enables researchers to analyze genetic variation in specific genome regions. This technology generates data similar to BestSeq when it comes to aneuploidy analysis. We currently have large, RealSeq-generated datasets on which our methodology can be tested in a robust fashion. This dataset consists of a total of 5,481 cancer patients, of which approximately 40% were diagnosed with cancer and 60% were healthy.The data generated through an amplicon-based approach comprises counts representing the magnification frequency of each target DNA region. In our context, we deal with hundreds of thousands of regions. To distill and extract relevant insights from this extensive dataset, we employ two computational techniques: GSVA (Gene Set Variation Analysis) and NMF (Non-negative Matrix Factorization).

Gene Set Variation Analysis is typically utilized to determine whether a set of genes within specific gene pathways collectively exhibit upregulation or downregulation. In simpler terms, it helps assess whether a particular set of genes is more or less active compared to a standard set of genes. This technique assigns a score to each gene set by considering the rank of individual genes within all genes based on their activity. Consistently low ranks in a gene set result in a low score, indicating underexpression, while consistently high ranks yield a high score, signifying overexpression. In both cases, we refer to the set as enriched for underexpressed or overexpressed genes. In our scenario, instead of genes, we apply the GSVA method to combine scores associated with individual amplicons, as described below.

After applying GSVA to our data, we obtained scores for each set of amplicons. In essence, every genomic region containing these amplicons receives a score. Typically, we have thousands of such genomic regions for each chromosome, resulting in thousands of scores representing each chromosome in a tested sample. To condense this information into a more manageable set of features per chromosome, we turn to NMF. NMF is a computational tool that enables the explanation of a large number of scores using a typically smaller set of features. To draw an analogy, consider characterizing a student's performance using test scores. If we have records of hundreds of tests across various subjects, extracting meaningful information about the student can be daunting. However, if we identify that all these test scores are highly correlated with two factors—Verbal and Analytical skills—we can effectively summarize the student's performance using just two scores instead of hundreds. NMF addresses a similar challenge in our case, where we have thousands of scores related to genomic intervals and aim to characterize them using a reduced number of factors, scoring these factors rather than the numerous genomic intervals. For a more detailed technical description of GSVA and NMF, please see below.

Following the application of NMF, we obtain scores corresponding to these factors for each chromosome. These factors represent aneuploidy features per chromosome, which we use for classifying cancer and healthy samples. Our chosen classifier is the Support Vector Machine Classifier (SVM). SVMs are trained on labeled datasets comprising healthy and cancer samples to learn decision boundaries based on input features. These boundaries define the separation between cancer and normal cases for future testing of patients.

GSVA or Gene Set Variation Analysis is an unsupervised method for detecting variation of gene set enrichment using a gene expression dataset. What is meant by “unsupervised” is that there is no prior need to label a training sample set as healthy or cancerous or categorize them in any other way in order to apply the method. GSVA assesses the relative enrichment of defined gene sets using a non-parametric approach. GSVA transforms the data from a gene expression by sample matrix (in this case) to a gene set by sample matrix with enrichment scores for each predefined gene set. It is important to note that we are not analyzing gene expressions in our case. Our input data consists of read counts corresponding to fixed DNA regions called amplicons. These regions can be mapped to specific genomic intervals contained in a specific chromosome. For the purposes of this analysis strategy, the counts corresponding to amplicons play the role of gene expression and the genomic intervals play the role of gene sets. Therefore, using the GSVA technique to detect Gene Set Enrichment entries with high expression will correspond to a gain of copies in the corresponding genomic interval, and thus also in the chromosome containing the interval. The same reasoning is true for detecting Gene Set Enrichment with low expression, i.e., low counts, as this will indicate a loss of copies in the containing chromosome.

NMF or non-negative matrix factorization is an unsupervised reduction technique (in which machine learning is based on unlabeled or raw data) used to decompose a matrix into two non-negative matrices. NMF is used to break down genomic interval scores into a small number of factors that we call “signatures of aneuploidy”. This technique decomposes the input matrix representing the genomic interval enrichment score for every patient (V, m samples by p genomic intervals) into two matrices, where the first is sample-independent and represents the signatures and the second is patient-dependent and represents the intensity of every aneuploidy signature for the given patient. The second matrix contains the features that are used to generate an aneuploidy score for a tested sample.

Steps in the pipeline

In order to build a classifier using the data from our amplicon-based approach, we tested our hypothesis on one subset of patients with normal chromosome counts, where aneuploidy was simulated (created bioinformatically) on a subset of these patients in order to validate our approach. To check the validity of the features extracted by our technique, we introduced aneuploidy in half of the “normal” samples, i.e., half of 1500 patients. For 750 patients, the read counts of the amplicons from one randomly chosen chromosome arm were multiplied with a randomly chosen multiplier between 0.5 and 2. In other words, 750 patients were artificially assigned a gain or loss of reads in the amplicons of a random chromosome arm. These were pooled back into the rest of the unmodified normal samples to be used for further testing. For the purposes of this report, this data will be referred to as the “simulated data”.

All the amplicons in our simulated data were pooled into 2600 genomic intervals, each composed of amplicons in 500kb genomic regions. Each interval was representative of the amplicons in that 500kb region, and the enrichment score for each of these gene sets was computed using GSVA.

These enrichment scores were further transformed to positive values for NMF (feature selection step). NMF was used to reduce the scores into a set of 15 features/arms that were representative of the scores of that arm. The resulting matrix contains the weights of the 1500 samples for each of the selected 15 features per arm. To ensure the accuracy of the technique, the correlation between the NMF features and the corresponding multipliers of the modified samples was assessed.

These features were further used to build the classifier for the simulated data. Using a 50:50 train-test split, half of the modified patients were dubbed “cancerous”, and half of the unmodified patients were dubbed “normal”. Using these features, we trained an SVM model to make predictions for the test data. The model, assessed based on those predictions, performed at 99% specificity.

The above-mentioned techniques were further tested on a subset of the real data. 1500 samples consisting of 750 normal patients and 750 cancer patients (randomly chosen) were used, and the predictions of the resulting SVM model were tested at 99% specificity (1% false positive rate).

Preliminary Data Results

Simulation results

To validate the meaning of the signatures extracted by our approach, the correlation between the features' (signatures') intensities, obtained from the signature's matrix for each patient, and their respective multipliers was calculated per arm. The results of the correlation were assessed in terms of the Pearson estimate and the p-values. Plots illustrating the correlation between the intensity of the signatures and the multiplier simulating aneuploidy are depicted in Figures 1 and 2. In the two examples illustrated below, every point represents a sample in which aneuploidy was simulated on one particular arm (3p or 22q in our example). We can observe the trend of the intensity of the signature (feature) as a function of the multiplier simulating aneuploidy. In the example corresponding to 3p, a higher signature intensity corresponds to a loss of copies, while it corresponds to a gain of copies in the 22q arm example.

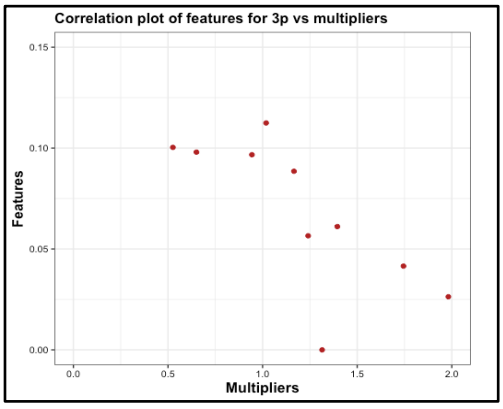

Fig 1: Correlation plot for 3p
Increasing signature intensity correlates to gene increasing gene copies.

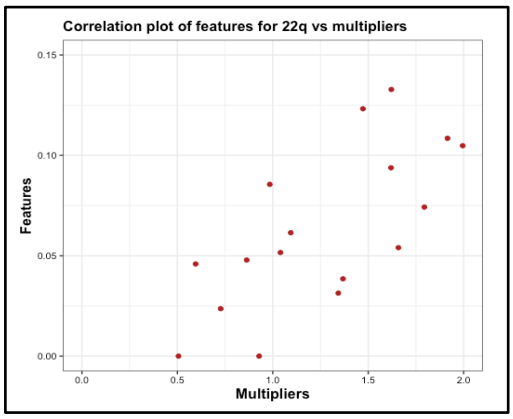

Fig 2: Correlation plot for 22q
Increasing signature intensity correlates to decreasing copies.

Classification Results Simulated Study

To test the sensitivity of our classifier, the 750 modified samples from the pilot data were modified at 19 arms, as opposed to a single random arm. The features from the resulting NMF were used to train an SVM classifier for the pilot data. The unmodified "normal" samples in the pilot data were labeled "healthy," while the modified "normal" samples were labeled "cancerous". The pilot data was split into train and test data, with 50% of each class represented in both train and test datasets. At 99% specificity, the classifier had a sensitivity of 91.7%. The performance of the classifier is depicted using a ROC curve (Fig. 3). The area under the ROC curve (AUC) was 0.971(359/393), stating that the classifier has a 97% probability of ranking a random modified normal more highly than a random unmodified normal. In conclusion, our modified normals stand in for simulated aberrant samples. If we randomly pick a simulated aberrant sample and an unmodified normal sample, the method has a chance of 97% to predict a higher aneuploidy risk for the aberrant sample compared to the unmodified one.

Real (RealSeq) data

The same steps were carried out for the original patient cohort. This data represented 1500 patients, with 750 healthy patients and 750 cancer patients. The features generated from the signature's intensity matrix were used to build an SVM-based classifier. The dataset was again split into train and test data, with equal representation from various cohorts of the dataset. At 99% specificity (1% chance of calling a false positive), the classifier had a sensitivity of 58.5% (This is the chance of calling a true positive, i.e., identifying true cancer). The ROC curve for the real data performed well, with an AUC of 0.95. In other words, if one randomly picks a true cancer sample and a normal sample, the method has a chance of 95% to predict a higher aneuploidy risk for the cancer sample compared to the normal one.

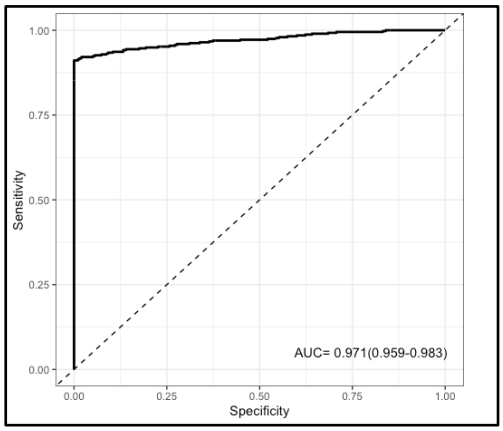

Fig 3: ROC curve for the pilot data
Classifier sensitivity vs. specificity based on simulated pilot data.

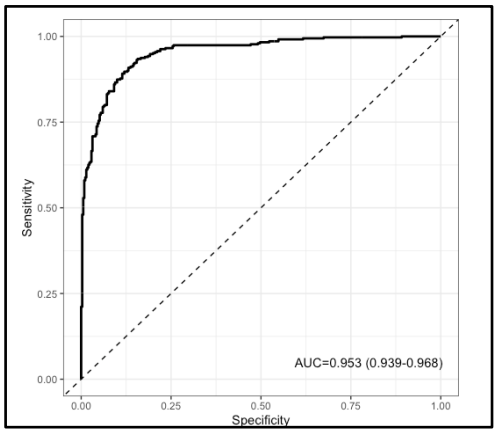

Fig 4: ROC curve for the real patient data
Classifier sensitivity v. specificity based on real patient data.

Conclusion and Future Steps

We developed a procedure designed for an amplicon-based approach to detect cancer in cell-free DNA by extracting an aneuploidy signal. As illustrated in Fig. 3, this procedure detected cancer at a 58% true positive rate, controlling the false positive rate at a level of less than 1%. However, it is important to reiterate that the method used here was based on a large dataset generated by the RealSeq technology. Our next goal is to train and test our model on the novel technology that we are developing, BestSeq, once we generate a large enough dataset based on this new technology. We anticipate analysis of this new dataset will begin in the last trimester of 2023. Finally, we will test for concordance between primary tumor aneuploidy and the score generated by our procedure on the 11 acquired samples wherein the primary tumor tissue and its corresponding aneuploidy have been analyzed.

Developing such a method will constitute an important building block for a BestSeq-based test combining features of aneuploidy and DNA fragmentation. Such a test has the advantage of being cost-effective and significantly impacts and shifts the stages of cancer detection.