Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Subject Section

Third-Generation Sequencing Allows for Unique Identification of cfDNA Tissue of Origin

Jacob Rast [1]* and associates

[1] Carnegie Mellon University, Pittsburgh, PA 15213, United States

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Discovery of extracellular genomic content in the human bloodstream has potential to allow the creation of a blood-based cancer assay. Analyzing cell free DNA with advanced computation methods has shown promise for longitudinal cancer analysis. However, current methods suffer from poor sensitivity, precluding use for early cancer preventative screening.
Results: In this work a proof-of-concept is developed to demonstrate the use of third-generation sequencing for early cancer detection. The ability of third-generation sequencing to capture the entire length of a strand of cfDNA allows for tumor originating reads to be isolated from background noise computationally. Investigation into the number of unique methylation signatures required to identify tissue of origin for a contiguous strand of cfDNA was undertaken. The results demonstrate the feasibility of using third generation sequencing to increase the sensitivity of blood based cancer assays.
Availability: Data models and full pipeline are available at https://github.com/jacobrast/cf-dna-assay
Contact: jrast@andrew.cmu.edu

1 Introduction

Cancer is the second leading cause of death in the United States, claiming approximately 600,000 lives annually. Preventative care options are limited; only a small subset of cancers have diagnostic testing methods with high enough sensitivity to be recommended by the US Preventative Screening Task Force. Recent discovery of novel molecular cancer biomarkers for in-vitro diagnostics has generated significant research interest. This new class of diagnostic testing–often referred to in literature as "liquid biopsy"–has potential to greatly broaden the scope of cancer screening both in frequency and disease detection. As cancer survival rates have a strong correlation with disease progression at diagnosis this new class of in-vitro diagnostic testing has potential to greatly improve cancer prognosis.

Molecular targets for liquid-biopsy cancer biomarker discovery vary. Frequently analyzed targets include tumor-derived cells, DNA, RNA, or proteins found in patient biofluids. Perhaps the most mature target is Prostate-Specific Antigen (PSA), a protein which over-expression in cancer positive patients caused FDA approval for cancer screening in 1994. Modern examples of FDA-approved liquid biopsy targets include Circulating Tumor Cells (CTCs), enumeration of which allows for the longitudinal monitoring of metastatic breast, prostate, and colorectal cancers in the CellSearch assay. In recent years, extracellular vesicles (EVs) have been subjected to intense investigation as cancer biomarkers due to their involvement in numerous biological processes. Extracellular vesicles are small subcellular particles of range $10\mu m$ to $20nm$. Due to the presence of a lipid bilayer, EV cargos such as DNA, RNA, or protein are protected from degradation in the extracellular environment. EV derived DNA has been therefore shown to be superior to free circulating-tumor DNA or cell-free DNA in diagnostic potential [6].

Excitement over cell-free DNA–both EV derived or free in the bloodstream–has stemmed in part from the well-demonstrated association between tumor genome and the genome found in cfDNA [3]. This properties allows for multi-omic analysis of cfDNA for cancer diagnostics similar to that possible with a traditional tissue biopsy. One issue with this approach is the lack of one-to-one correlation between disease biomarker and underlying cancer. Due to the highly heterogeneous nature of cancer, positive identification of a well-known cancer causing mutation does not allow for a definitive diagnosis of cancer subtype. Rather, longitudinal analysis would be required for this genomic diagnosis be that radiography, traditional tissue biopsy, or monitoring. One approach to overcome this difficulty is the use of epigenomic markers to infer additional information about the underlying disease. Epigenomic markers, such as DNA

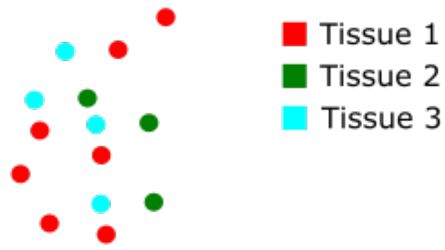


Figure 1. High sensitivity liquid biopsy is a "needle in haystack" problem of detecting tumor-originating cfDNA in a highly heterogeneous mixture of a blood-borne molecules. High molecular weight cfDNA is largely contained in extracellular vesicles, where it is protected from degradation in the extracellular environment. Use computational approaches the isolation of tissue-specific cfDNA from background noise can be demonstrated for further downstream processing. We hypothesize that this can be used to increase sensitivity of blood-based assays for early cancer screening applications.

methylation or histone modification, have been shown to have unique tissue-specific expression. DNA methylation in particular has been shown to have high correlation with gene expression, which in turn is unique for each tissue-specific cellular subtype. Several groups have used this information to build statistical models with power to discriminate tissue type based upon epigenome [2].

For cfDNA epigenomic analysis, third-generation sequencing technologies such as those offered by Oxford Nanopore offer unique advantages. These sequencing platforms offer long-read sequencing directly from non-amplified DNA, with no theoretical upper limit on the length of DNA that can be read. Reads of length up to a megabase pair have been demonstrated [4]. Furthermore, the operational principle allows for the direct recovery of previously undetectable base pair information, for example RNA base uracil or DNA methylation. This includes the common cytosine guanine modification referred to as CpG sites. Third-generation sequencing therefore facilitates the direct, simultaneous analysis of cfDNA genome and epigenome without need for fragmentation or bisulfite conversion. In this work, in order to demonstrate the capabilities of third-generation sequencing and applications to liquid biopsy assay development a proof of concept of tissue of origin localization for a single contiguous cfDNA strand was demonstrated. The model uses existing, low-density epigenomic information widely available in public datasets as a "worst-case" sparsity. Several statistical models including Naive Bayes and Multilayer Perceptron were developed to analyze the epigenetic signatures between two given points and infer the tissue of origin the underlying data was capture from. This problem formulation allows for a clear demonstration that a single strand of cfDNA captured in blood can be traced to its origination tissue using epigenomic information alone.

2 Methods

2.1 Datasets and Preprocessing

Few publicly available thrid-genertion sequencing datasets are available, and of those available none match the critera of sequencing high molecular weight DNA fragments. However, this did not pose a challenge. As a proof of concept, it is only necessary to demonstrate that tissue-of-origin localization can be performed using methylation signatures at or below the information density offered by advanced sequencing techniques. If localization using fewer methylation cites (more sparse data) is possible, addition of more highly granular data can only improve performance. In light of this fact, microarray methlyation data was preprocessed to create numerous synthetic datasets each mimicking a highly-sparse cfDNA read of a given length, centered on a locii of choice. Concretely, 290 datasets were created to mimic reads of high molecular weight cfDNA with length from 1kbp to 580kbp. To overcome sparsity issues, each read was centered a single locii. Namely, the location of well-known proto-oncogene KRAS was selected on chromosome 12 position 25294434.

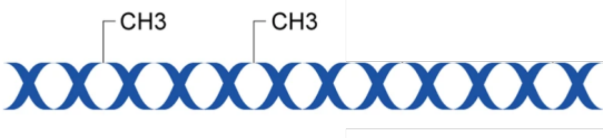


Figure 2. This proof-of-concept explores the length of a contiguous DNA strand necessary to capture sufficient epigenomic data to perform tissue-of-origin localization. Third-generation sequencing allows for the capture of entire strands of high molecular weight cfDNA, allowing for computational techniques to capture sufficient methylation data to infer tissue of origin.

Creation of synthetic cfDNA epigenetic datasets was enabled through the use of publicly available cancer positive genome methylation data. Data was sourced from The Cancer Genome Atlas (TCGA). Illumina Human Methylation 450k microarray data was selected for a balance of epigenome coverage and availability. This dataset reports methylation percentage and 400,000 known methylation sites, all occuring at CpG sites. Each CpG site reports a methylation beta value, a ratio of intensity between methylated and unmethylated sites that correlates with percentage of tissues with methylation at this location. 600 datasets in total across for two cancer subtypes (COAD and LUSC) were selected at random from a pool origination from the same group and same assay. Data from these all of these 600 Illumina 450K assays was processed to create 290 synthetic cfDNA long-read datasets. As all data used in this set originated from the same assay and same group, further data normalization was not performed.

2.2 Model Evaluation and Selection

Several models were evaluated for supervised lung or colon tissue of origin classification. Evaluation was performed over a small test dataset, and the model with best performance was selected for use on all 290 full synthetic datasets. On this test dataset, Naive Bayes achieved a training accuracy of 0.82 compared to 0.91 for MLP. As such, the MLP model was selected.

2.2.1 Gaussian Naive Bayes

Many problems in bioinformatics can be posed as conditional likelihood. Classification frequently takes the form of asking the probability of class Y given known variables $X_1, X_2, ..., X_n$. That is to say,

$$P(Y|X_1, X_2, ..., X_n)$$

If n is a small number, these probabilities can be easily calculated from the laws of conditional probability.

$$P(Y = y_i|X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = \frac{P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n|Y = y_i)P(Y = y_i)}{\sum_j P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n|Y = y_j)P(Y = y_j)}$$

However, as our feature space n grows the number of data points required to estimate full conditional probabilities grows as 2^n. Naive Bayes

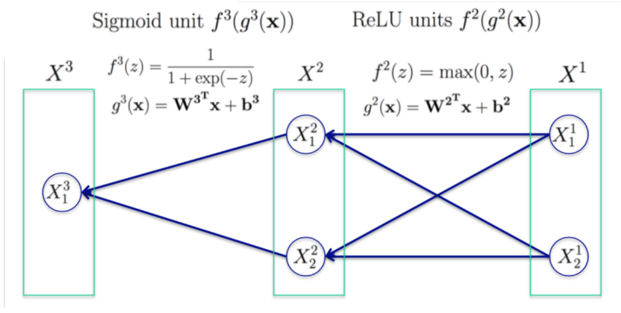

Sigmoid unit $f^3(g^3(\mathbf{x}))$
ReLU units $f^2(g^2(\mathbf{x}))$
X^3
$f^3(z) = \frac{1}{1+\exp(-z)}$
$g^3(\mathbf{x}) = \mathbf{W^3}^\mathbf{T}\mathbf{x} + \mathbf{b^3}$
X^2
$f^2(z) = \max(0, z)$
$g^2(\mathbf{x}) = \mathbf{W^2}^\mathbf{T}\mathbf{x} + \mathbf{b^2}$
X^1
X_1^3
X_1^2
X_2^2
X_1^1
X_2^1

Figure 3. Multi-layer perceptron statistical model architecture. Note that data flows from left to right. Figure drawn from slides by Tom Mitchell

makes the assumption that each X in $X_1, X_2, ..., X_n$ is conditionally independent from each other. That is to say,

$$P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n | Y = y_k)$$
$$= \prod_i^n P(X_i = x_i | Y = y_k)$$

For continuous (rather than Boolean) values of X, we must assume an underlying distribution of the data. A common assumption is that X follows a Gaussian distribution. This implies that

$$P(X_i = x_i | Y = y_k) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2})$$

2.2.2 Multi-layer Perceptron

The MLP or multi-layer perceptron is a powerful statistical model that has been applied to a large class of problems. The model architecture can be found in figure . Each layer in the MLP can be conceptualized as a series of parallel matrix multiplications followed by a nonlinear element-wise operation. Full discussion of the MLP model can be found in [1].

2.3 Learning and Inference for model

2.3.1 Gaussian Naive Bayes

Training Gaussian Naive Bayes consists only of calculation of conditional probabilities. For this reason, Naive Bayes is known to be a fast algorithm.

2.3.2 Multi-layer Perceptron

MLP models are trained using the well known backpropagation algorithm. In our case, the stochastic gradient descent algorithm was with the negative log liklihood loos function was used for training. The SGD loss function, algorithm and full backpropagation algorithm can be found in [1].

3 Results

3.1 Classification accuracy

Results for classification accuracy can be found in figure 4. Each dot in this model represents a psuedo-dataset. In other words, the feature vector for each point in this graph was defined by the set of methylation CpG loci within a given distance from a fixed location in the genome. As can be seen, highly accurate read classification can be achieved for this proof-of-concept toy example with read lengths in the 100 kbp range. Using convectional technologies this would be impossible to achieve, yet with long-read sequencing the state of the art has achieved sequencing of high-molecular weight DNA fragments an order of magnitude longer. [4] Additional, groups have reported in literature discovery of EV associated cfDNA with length of up to a megabase pair. [5] The results support the conclusion that tissue of origin localization can be performed on a per-strand basis for DNA fragments such as those found in blood.

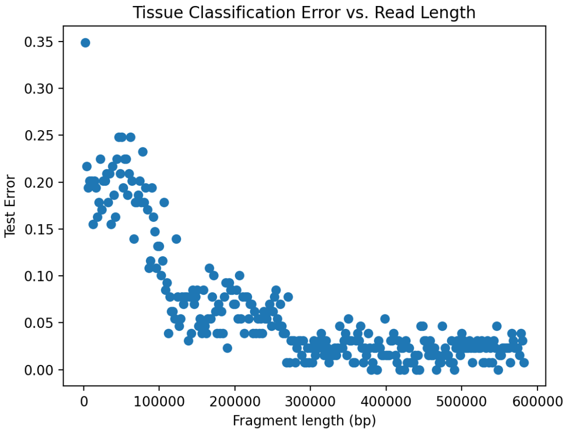


Figure 4. This proof-of-concept explores the length of a contiguous DNA strand necessary to capture sufficient epigenomic data to perform tissue-of-origin localization. Third-generation sequencing allows for the capture of entire strands of high molecular weight cfDNA, allowing for computational techniques to capture sufficient methylation data to infer tissue of origin.

4 Implementation Details

4.1 Data

All data used in this experiment is publicly available on the Cancer Genome Atlas. A small subset of all available COAD and LUSC Illumina 450K Methylation assay data was used. In order to enable reproduction of the results reported in this publication, the full GDC Manifest is available on GitHub.

4.2 Data Preprocessing

Creation of synthetic datasets to mimic high molecular weight cfDNA was a central challenge of this work. Datasets were chosen to center around the location of KRAS proto-oncogene, chromosome 12 position 25294434. All methylation sites within distance n of this location were selected for inclusion in the synthetic dataset, with n acting as a hyperparameter.

4.3 Models

For both Gaussian Naive Bayes and Multilayer Perceptron, off-the-shelf models were used for train, test and validation. Models can be found in popular Python package scikit-learn v0.24. Model hyperparameters such as number of layers, hidden layer size, and learning rate were evaluated via cross-validation dataset. Default parameters were found to be sufficient for most parameters. Full source code can be found online at the author's GitHub.

5 Discussion and future direction

Research excitement, recent discoveries, and the emergence of numerous commercial ventures in the liquid biopsy space signal the emergence of extracellular cell free DNA as a strong potential biomarker for cancer diagnostics. However, existing work such as that of [7] has to date focused on traditional "NGS" short-read sequencing. These models are similar in nature to "bulk" sequencing, and as a result suffer from poor sensitivity on patients in early cancer stage. In this work, we seek to demonstrate an approach to overcome these issues and obtain tumor or tissue specific genomic data from the highly heterogenous mixture of blood-based cell free DNA.

Broading this technology to allow application in clinical setting would require several further developments. Firstly, it is necessary to test the underlying hypothesis that tissue of origin pre-processing will increase sensitivity in cfDNA assays. Use of a toy cancer/not cancer classification scheme in conjunction with the tissue of origin localization model described in this publication would allow for demonstration of this. Additionally, this proof-of-concept examined only two tissue classifications: colon and lung cancers. Development of a single classification scheme–for example lung vs all classification–would allow for increase confidence in the robustness of the result. This one vs. all classification as this is more similar to challenge posed by real clinical data in the creation of a single cancer diagnostic step. Finally, this proof-of-concept presupposed that all cfDNA discovered was filtered such that it originated from a given location on chromosome 12. In reality, this would only hold true for a very small percentage of cfDNA found. Broadening the classifier to cover reads originating from any location in the genome would raise issues of sparsity requiring further computational analysis to overcome.

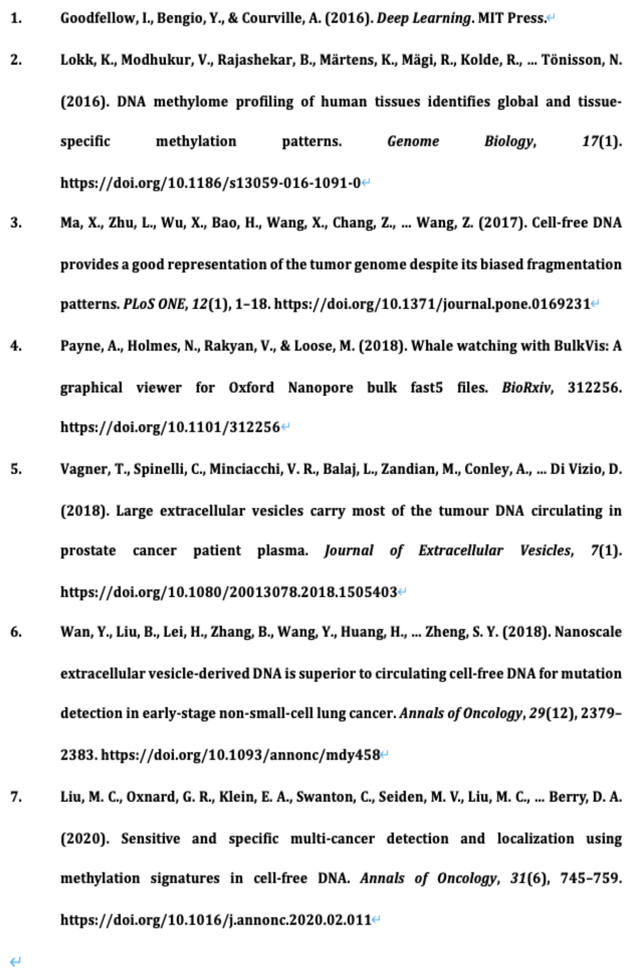
Beyond broadening the scope of the proposed computation technique, further optimization could be performed in order to reduce length of the minimum fragment required for tissue of origin localization. Current modeling suggest that 300 kbp length fragments are required for perfect data localization. While this is within the reported capabilities of long read sequencing [4] and the known length of cfDNA fragments found in EVs, this would require disuse of a large number of short reads. Such practice may reduce the sensitivtiy of the cancer assay and potentially introducing bias into the final read selection. While preliminary reports indicate that high molecular weight cfDNA is predominately cancer-positive in origin [5], this bias may not be favorable. More sophisticated techniques such as graph neural networks or a deeper neural network might be applied to the problem in order to improve classification accuracy of short reads.

Acknowledgements

We thank the course staff of Carnegie Mellon University's Spring 2021 Computation Genomics course.

CompGene Report.bib

References



1. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.↵
2. Lokk, K., Modhukur, V., Rajashekar, B., Märtens, K., Mägi, R., Kolde, R., … Tõnisson, N. (2016). DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. *Genome Biology*, *17*(1). https://doi.org/10.1186/s13059-016-1091-0↵
3. Ma, X., Zhu, L., Wu, X., Bao, H., Wang, X., Chang, Z., … Wang, Z. (2017). Cell-free DNA provides a good representation of the tumor genome despite its biased fragmentation patterns. *PLoS ONE*, *12*(1), 1–18. https://doi.org/10.1371/journal.pone.0169231↵
4. Payne, A., Holmes, N., Rakyan, V., & Loose, M. (2018). Whale watching with BulkVis: A graphical viewer for Oxford Nanopore bulk fast5 files. *BioRxiv*, 312256. https://doi.org/10.1101/312256↵
5. Vagner, T., Spinelli, C., Minciacchi, V. R., Balaj, L., Zandian, M., Conley, A., … Di Vizio, D. (2018). Large extracellular vesicles carry most of the tumour DNA circulating in prostate cancer patient plasma. *Journal of Extracellular Vesicles*, *7*(1). https://doi.org/10.1080/20013078.2018.1505403↵
6. Wan, Y., Liu, B., Lei, H., Zhang, B., Wang, Y., Huang, H., … Zheng, S. Y. (2018). Nanoscale extracellular vesicle-derived DNA is superior to circulating cell-free DNA for mutation detection in early-stage non-small-cell lung cancer. *Annals of Oncology*, *29*(12), 2379–2383. https://doi.org/10.1093/annonc/mdy458↵
7. Liu, M. C., Oxnard, G. R., Klein, E. A., Swanton, C., Seiden, M. V., Liu, M. C., … Berry, D. A. (2020). Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Annals of Oncology*, *31*(6), 745–759. https://doi.org/10.1016/j.annonc.2020.02.011↵

↵