Leveraging Chromatin Accessibility and Nucleosome Positioning in Liquid Biopsy for Cancer Classification

Sasha Main
Department of Medical Biophysics
University of Toronto
Toronto, ON. 101 College Street
sasha.main@mail.utoronto.ca

Oliver De Sa
Department of Immunology
University of Toronto
Toronto, ON. 1 King's College Circle
oliver.desa@mail.utoronto.ca

Yuchen Zhou
Department of Immunology
University of Toronto
Toronto, ON. 2075 Bayview Ave.
thomas.zhou@mail.utoronto.ca

Abstract

In cancer patients, both healthy and tumour cells release cell-free DNA (cfDNA) fragments into the bloodstream. The tumour cfDNA can be detected through a simple blood test known as a liquid biopsy, to reveal tumour-specific information throughout a patient's journey with cancer. New developments in the liquid biopsy field have shown that fragmentation patterns of cfDNA can provide highly accurate cancer detection and reveal information about the chromatin structure of the cell of origin, expanding the potential clinical utility of cfDNA. Recently, a new framework for extracting features related to nucleosome positioning was published. The authors showed cancer classification with high accuracy, but their analysis was limited to transcription factor binding sites. Here, we expanded upon this by leveraging cancer and blood-specific genome-wide chromatin accessibility and features of nucleosome positioning to classify cancer. We used a cfDNA dataset, for which these features had not been explored, and generated 376 healthy and 575 cancer *in silico* samples. We applied logistic regression, Random Forest, XGBoost, and multilayer perceptron models to all features and a dimensionality-reduced feature set. We found that the multilayer perceptron trained on the dimensionality-reduced feature set had the best model performance with an accuracy of 0.995 and an F1 score of 0.992. We also extracted feature importance to elucidate biological insights about the most informative chromatin regions. We found that diversity in genomic elements, accessibility sites across cancer types, and both accessible and inaccessible regions were important for classification. This work provides a proof of concept that nucleosome positioning features from cancer and blood-specific chromatin accessibility sites can be used for cancer classification, and diversity in chromatin sites are important in a pan-cancer setting.

1 Introduction

Liquid biopsy technologies focused on detecting tumour-derived cfDNA have revealed biomarkers for cancer management[1]. However, phenotypic cfDNA features, such as fragmentation patterns present within cfDNA whole-genome sequencing (WGS) data, may expand potential diagnostic applications of cfDNA[2,3]. The fragmentation of cell-free DNA is a non-random process where differences in nucleosome occupancy patterns at open versus closed chromatin affect where nucleases can access

and fragment the DNA. Since nucleosome-bound DNA is protected from digestion, the distribution of cfDNA sequencing reads over the genome reflects chromatin accessibility and nucleosome positioning of the cell of origin. Tumour cells are characterized by widespread changes in chromatin accessibility compared to healthy cells, so these fragmentation features can be leveraged for cancer detection, subtype classification, and more[4–8].

2 Related Works

In December 2022, a group published a novel framework for extracting features related to nucleosome positioning called "Griffin"[5]. Specifically, this framework utilizes genome-wide sites of interest where cancer signals may differ from healthy blood and calculates quantitative fragmentation features. This approach circumvents previous challenges faced by the liquid biopsy field, where the generation of coverage profiles to infer nucleosome positioning required large amounts of tumour cfDNA and high sequencing depth[7,9]. Here, Griffin analyzes average signal over many regions of interest, with robust GC correction and normalization measures, to allow for informative features of nucleosome patterns with down to 0.1X WGS coverage. The authors used their framework to detect cancer within a cohort of healthy and cancer patient samples (1-2X cfDNA WGS) with a high AUC (0.94). However, this analysis was limited to transcription factor binding sites and a logistic regression model. Around the same time, a paper by Taklifi et al. published work investigating the integration of chromatin accessibility states into liquid biopsies[10]. This group used ATAC-seq data from TCGA and other sources to perform differential chromatin accessibility analyses between various cancer types and normal healthy blood. They then performed machine learning modeling with ATAC-seq data to select the most important regions of differential chromatin accessibility that could distinguish blood and cancer types. As a result, they published highly annotated lists of these cancer or blood-specific regions of differential chromatin accessibility. However, the potential of using these regions to classify cancer with cfDNA WGS data has yet to be explored and would demonstrate the benefits of this group's analyses.

We aimed to build off these two key works by investigating the highly informative genome-wide regions of chromatin accessibility from Taklifi et al. and use Griffin to extract features of nucleosome positioning within these sites of interest. We then sought to explore many models for cancer classification and interpret the importance of these features to answer questions about which differentially accessible genomic elements are most informative (e.g. promoters, introns, etc.). Furthermore, we used a different dataset for which these features have not been explored, and applied data augmentation approaches to achieve improved sample sizes.

3 Methods

3.1 Dataset and Pre-processing

Public raw cfDNA sequencing datasets are challenging to find, so we used NucPosDB with the filters set to cfDNA source as blood plasma, experimental method as whole genome sequencing (WGS), and access as open[11]. We identified a cfDNA WGS dataset from Snyder and colleagues (2016) and downloaded raw paired-end sequencing reads in FASTQ format from the European Nucleotide Archive (available at https://www.ebi.ac.uk/ena/browser/view/PRJNA291063)[11, 12]. We trimmed reads using Cutadapt[13]. Next, we aligned trimmed reads to the hg38 version of the human genome using Bowtie2[14]. We then removed duplicate reads using Samtools (Version 1.9)[15], combining samtools sort, samtools fixmate with the -m option, and samtools markdup with the -r option. We then filtered out low-quality mapped reads using samtools view with the -bq option, with a minimum mapping quality of 30. We then split our data into training, validation, and test sets based on the sample's sequence coverage and tried to maintain balance across different cancer types (see Table 4).

3.2 Data Augmentation

Sequencing coverage is essentially the number of reads stacked on top of each other at any given site in the genome. While our dataset had small sample size, (healthy controls n=4, cancer patients n=48), some of the samples had high sequencing coverage (also referred to as high depth). Notably, three healthy controls and five cancer patient samples were deep WGS (up to 104X sequencing

coverage). The remaining samples had an average sequencing coverage of 2.2X. We then split the dataset, as previously described, with the deep WGS samples in the training set. Since Griffin required only 0.1-1X sequencing coverage, we employed a data augmentation technique known as random stratified downsampling with replacement (similar processes have been described[16–20]). Essentially, we randomly bootstrapped and downsampled reads to generate many *in silico* samples with 1X sequencing coverage.

To generate the *in silico* samples, we used downsampleSam from Picard tools (GATK). For each sample, we extracted the total number of reads and then calculated the number of *in silico* samples to generate by dividing the total number of reads by 40 million reads (an approximation for 11X coverage). If the total number of reads was less than 600 million, we multiplied the rounded result by 2, and if it was greater than 600 million, we multiplied it by 3. Additionally, we calculated the proportion of reads to select in the downsampling process by dividing the total number of reads by 40 million. Overall, we generated 951 *in silico* samples from the initial 52 samples.

	Training	Validation	Test	Total
Healthy	168	65	143	376
Cancer	471	36	68	575
Percent Cancer	76%	36%	32%	

Table 1: Overview of the augmented dataset

3.3 Feature Extraction

After generating the *in silico* samples, we downloaded regions with cancer and blood-specific chromatin accessibility determined by Taklfi and colleagues[10]. Specifically, supplementary files 7 and 8 were downloaded. The region files were then formatted into two specific site files (2216 accessible in cancer, and 612 inaccessible in cancer) with the columns Chrom, Start, End and position, where the position was the midpoint of the chromatin sites. Next, we input the site files and *in silico* samples into Griffin, a snakemake pipeline composed of two sections (GC correction and nucleosome positioning). We ran all steps with default settings except for the nucleosome positioning portion, where the individual parameter was set to true. Quantitative features (central coverage, mean coverage, and amplitude) from Griffin were then extracted from each set of regions, for each sample, and used as input features for machine learning algorithms.

3.4 Algorithms

We trained logistic regression, random forest, gradient boosting, neural network, and modeled to predict cancer versus healthy patients using all features and top principle components that can explain 95% of the variance. The hyperparameters were tuned using a validation set. In the end, all models were examined using a test set. Feature importances and distributions of important features were used to gain biological insights from the machine learning models.

Algorithm 1 MLP model to classify cancer versus healthy patients using cfDNA sequence data

Split data into training, validation, and test sets (70%, 10%, 20%)
Use DownsampleSam (Picard) from GATK to generate *in silico* samples
Extract each cancer-specific site's features using Griffin
model $\leftarrow$ MLP (2-layers)
$S \leftarrow$ a set of hyperparameters
for $s \in S$ **do**
    while epochs < 30 **do**
        train MLP using training data with hyperparameters s
    end while
    predValY $\leftarrow$ predict validation data using MLP
    calculate F1 score and accuracy using predValY and true label, valY, of validation data
end for
train MLP using s that gives the highest F1 score and accuracy
predict label of test data using MLP model with the best hyperparameters

4 Results

4.1 Modelling with all features

Metrics	Accuracy	F1 Score
Logistic Regression	0.938	0.894
Random Forest	0.697	0.680
XGBoost	0.578	0.604
MLP	0.986	0.977

Table 2: Performance of machine learning models using all features

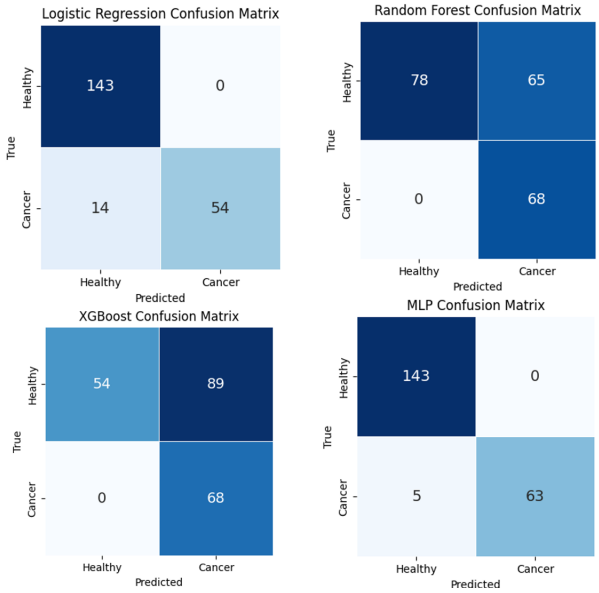

Figure 1: Confusion matrices of logistic regression, random forest, XGBoost, and MLP models using all features.

Once the final feature set was prepared and split between training validation and testing, four models were trained on the full feature sets, tuned on the full validation set and tested on the full test set. The MLP model had the most accurate classification of both healthy and cancer samples, with an accuracy of 0.986 and an F1 score of 0.977 (Table 2). Logistic regression performed second best

at classification with an accuracy of 0.938 and an F1 score of 0.894 (Table 2). This is consistent with the results published by Doebley et al. in the first implementation of the Griffin tool for feature extraction[5]. The random forest model also performed reasonably well, with an accuracy of 0.697 and an F1 score of 0.680 (Table 2). The worst model for the classification of this dataset was XGBoost, with an Accuracy of 0.578 and an F1 score of 0.604, likely due to the imbalanced validation set that the model was tuned with (Table 2).

4.2 Modelling with dimensionality reduced features

Metrics	Accuracy	F1 Score
Logistic Regression	0.933	0.885
Random Forest	0.512	0.569
XGBoost	0.858	0.819
MLP	0.995	0.992

Table 3: Performance of machine learning models using principle components

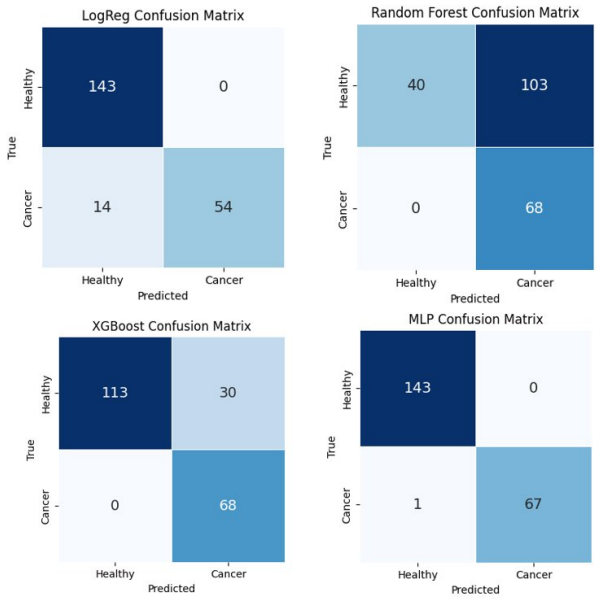

Figure 2: Confusion matrices of logistic regression, random forest, XGBoost, and MLP models using principle components explaining 95% of the variance in the data.

To see if the dimensionality of our feature set could be reduced while retaining the ability of our models to make accurate classification, we clustered our data by PCA and re-trained the same four models on principle components that explained 95% of the variance in the dataset. The models trained on the principle components performed similarly to the models trained on the full feature set, with MLP being the most successful classifier (Table 3). The Logistic Regression classifier also retained performance when trained on principal components (Tables 2 & 3). Classification with XGBoost was improved greatly when trained on principle components, increasing from accuracy and F1 of 0.578 and 0.604 to 0.858 and 0.819 (Tables 2 & 3). In contrast, random forest performance was decreased when trained on principle components decreasing from accuracy and F1 of 0.697 and 0.680 to 0.512 and 0.569 (Tables 2 & 3).

4.3 Biological Interpretations

Having fitted various models, we were interested in which chromatin sites were most important for our classification. We used feature importance and magnitude of coefficients to extract feature importance across our models. We then selected the top 10 important features from each model, which provided us with 18 important chromatin sites.

First, we sought to look at the types of genomic elements of the top sites. We found that most sites were gene-centric (i.e., introns and promoters), but there was a diversity across other sites as well (e.g., distal intergenic regions) (Fig 3A). Next, we explored which cancers the chromatin accessibility sites were found to originate from within the work by Taklifi and colleagues[10]. We saw a large diversity in chromatin sites from different cancers, with many regions from colon and kidney cancers (Fig 3B). Lastly, we investigated whether the top chromatin accessibility sites were accessible or inaccessible in cancer, and found both to be important (Fig 3C).

Moreover, we then explored the coverage profiles across the top individual chromatin sites to visualize the metrics extracted and used in modelling. For example, we explored an accessible promoter found in prostate cancer, VTRNA1-1: human non-coding vault RNA1-1, which has been previously associated with cancer (Fig 4 left)[21]. The coverage patterns in this region differed between cancer and healthy patients, implying different nucleosome organization. In the cancer patients, there was a clear decrease in coverage at the center of the chromatin site, corresponding to a nucleosome-depleted region (Fig 4 left). Whereas in the healthy patients, this region was inaccessible, implying the DNA was protected by a nucleosome, resulting in increased coverage (Fig 4 left). In contrast, we observed the opposite association for an inaccessible intron from prostate cancer, TBC1D7-LOC100130357 (Fig 4 right). This gene is documented as over-expressed in whole blood, and accordingly, we saw a decrease in coverage at the center of the chromatin site for healthy patients, suggesting this region is open and unprotected (Fig 4 right)[22]. However, there was a clear increase in coverage at the chromatin site for cancer patients, implying this region is inaccessible in cancer (Fig 4 right).

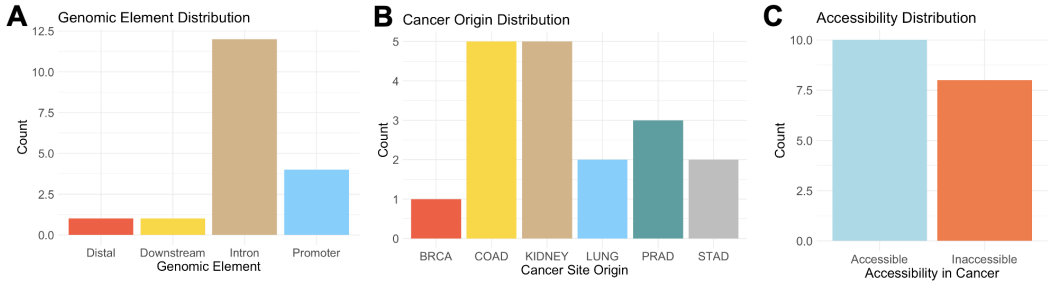

Figure 3: Distributions of the most important chromatin sites. A) The distribution of genomic elements of the top chromatin sites. B) The distribution of cancer origins across the top chromatin sites. BRCA = Breast invasive carcinoma, COAD = Colon adenocarcinoma, KIDNEY = kidney cancer, LUNG = lung cancer, PRAD = Prostate adenocarcinoma, STAD = Stomach adenocarcinoma. C) Distribution of accessibility in cancer across the top chromatin sites.

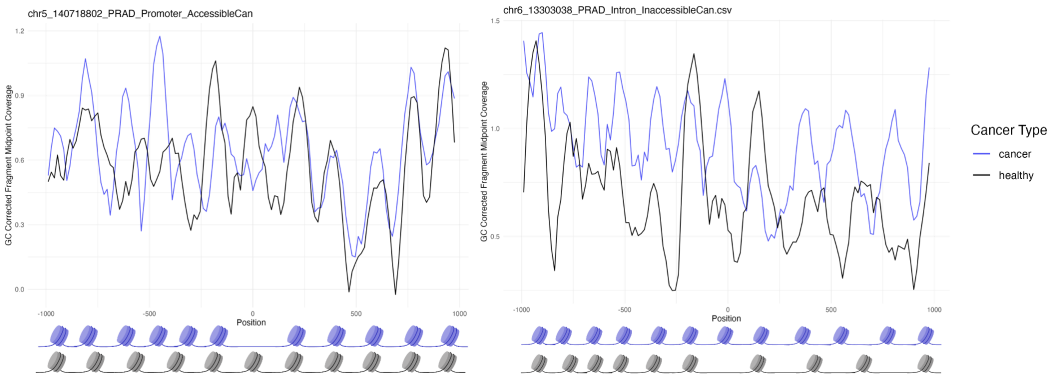

Figure 4: Coverage plots at specific chromatin accessibility sites across all cancer and healthy patient samples. A) Coverage profile (mean) at an accessible promoter from prostate cancer, VTRNA1-1: human non-coding vault RNA1-1. B) Coverage profile (mean) at an inaccessible intron from prostate cancer, TBC1D7-LOC100130357.

5 Discussion and Conclusion

In this work, we demonstrate that nucleosome positioning features of cancer and blood-specific chromatin accessibility sites can be used for cancer classification. We combined ideas and resources from two recent publications to approach cancer classification from a novel angle. We used Griffin to extract our nucleosome positioning features, a framework which has only been implemented twice in the literature. We also explored cancer and blood-specific chromatin accessibility sites originating from ATAC-seq data that had not ever been profiled in cfDNA for nucleosome positioning features.

After extensive data pre-processing and feature preparation, we applied four different modeling approaches to all features and a dimensionality-reduced feature set. We found a high similarity and crossover between healthy and cancer samples in the validation set when visualized by the first and second components in a PCA (Figure 5). We also observed that the healthy samples in the training set and the healthy samples in the validation set are quite different from each other. This difference might be due to different library preparation methods. We still tuned our hyperparameters using the validation set and found out that MLP had the best performance for classifying test set data. Our model performances were comparable to those from the original Griffin publication, where their AUCs for logistic regression on a dimensionality-reduced features set ranged from 0.93 (early-stage cancers) to 0.99 (late-stage). However, our approach differed significantly since we used 2,828 cancer and blood-specific chromatin accessibility sites. Whereas, they used 30,000 transcription factor binding sites for each of the 270 transcription factors of interest, resulting in 8.1 million sites total. Therefore, our feature set is less than 1 percent of their size, yet we achieve high performance.

Furthermore, we were able to make biological insights about which chromatin regions were most important in our models. In general, we found that diversity across chromatin accessibility sites was important, in terms of genomic elements, cancer origins, and states of accessibility in cancer. We also found that gene-centric elements were common across the top features. These findings were consistent with previous literature highlighting altered nucleosome organization and sequencing coverage at promoters and first intron-exon junctions of active genes compared to inactive ones7,8,16. In addition, we saw a large diversity in chromatin sites from different cancers. Some of these important sites were found within stomach adenocarcinoma patient ATAC-seq data, but we did not have any patients with stomach adenocarcinoma in the cfDNA dataset (Fig 3B). This suggests that some chromatin sites may be consistent across cancer types. Lastly, we found that both inaccessible and accessible sites in cancer are pertinent to classification. We highlighted this through coverage plots for two distinct cancer and blood-associated genes, which permit visualization of differing coverage peaks at the center of the site between cancer and healthy patients (Figure 4).

While we were able to make some intriguing insights, our work has many limitations. First, our cfDNA dataset had different library preparation techniques across samples, which may be one of many covariates contributing to batch effects or noise in our feature set. This is likely a main contributor

to the differences in our healthy controls in the validation set (these were single-stranded library prep, whereas the other healthy samples were all double-stranded). In addition, our cfDNA dataset was limited in sample size, so we resorted to *in silico* sample generation. Similar techniques have been performed in the liquid biopsy field previously, but we did not fully characterize the similarities between our *in silico* samples across the varying coverages. Intuitively, generating two samples from one low-coverage sample, versus 100 samples from one high-depth sample may have implications for our models. Furthermore, our cfDNA cancer samples were exclusively from metastatic patients, which is potentially problematic for two reasons. First, the amount of tumour-derived cfDNA is related to tumour burden, so metastatic patients generally have a higher proportion of tumour-derived cfDNA than early-stage patients. This means that the cancer signal is less diluted by cfDNA from white blood cells, likely resulting in more prominent features and simpler cancer classification. However, we did not quantify the amount of tumour-derived cfDNA in our *in silico* samples, so we cannot confirm this is the case across all cancer types and patients in our dataset. In the future, similar works should use a tool like ichorCNA20, which permits quantification of the tumour fraction of cfDNA and normalization to help mitigate this bias. Secondly, a metastatic patient cohort has limited clinical relevance for cancer versus healthy classification, since typically, this classification problem would be used for screening, and very few patients would present with metastatic disease in this setting. The approach and results from our work are likely applicable to early-stage cancer as well or could be expanded upon to detect recurrence in metastatic patients. However, further exploration into appropriate cohorts would be needed to test the sensitivity and applicability in these clinical settings. Lastly, we wanted to validate our results with an independent external cfDNA dataset, but due to limited pan-cancer public datasets available and time constraints involved in preprocessing and feature extraction, external validation remains outside the scope of this report.

Despite the limitations of our project, we highlight a novel approach of using nucleosome positioning features extracted from cancer and blood-specific chromatin accessibility sites. We used this workflow for cancer classification, but the findings and concepts from this work could be applied to future liquid biopsy cfDNA projects spanning many clinical applications, such as treatment response, monitoring, and subtype classification.

6 Appendix

References

[1] Ignatiadis, M., Sledge, G.W. & Jeffrey, S.S. Nat. Rev. Clin. Oncol. 18, 297–312 (2021).

[2] Lo, Y.M.D., Han, D.S.C., Jiang, P. & Chiu, R.W.K. Science 372, (2021).

[3] Liu, Y. British Journal of Cancer (2021).doi:10.1038/s41416-021-01635-z

[4] Corces, M.R. et al. Science 362, eaav1898 (2018).

[5] Doebley, A.-L. et al. Nat. Commun. 13, 7475 (2022).

[6] De Sarkar, N. et al. bioRxiv 2022.06.21.496879 (2022).doi:10.1101/2022.06.21.496879

[7] Ulz, P. et al. Nat. Genet. 48, 1273–1278 (2016).

[8] Shahrokh, E.M. et al. Nat. Biotechnol. xxx, xxx (2022).

[9] Murtaza, M. Caldas, C. Nat. Genet. 48, 1105–1106 (2016).

[10] Taklifi, P., Palizban, F. & Mehrmohamadi, M. Sci. Rep. 12, 10447 (2022).

[11] Shtumpf, M., Piroeva, K.V., Agrawal, S.P., Jacob, D.R. & Teif, V.B. Chromosoma 1, 2021.11.24.469884 (2022).

[12] Snyder, M.W., Kircher, M., Hill, A.J., Daza, R.M. Shendure, J. Cell 164, 57–68 (2016).

[13] Martin, M. EMBnet.journal 17, 10–12 (2011).

[14] Langmead, B. & Salzberg, S.L. Nature Methods 2012 9:4 9, 357–359 (2012).

[15] Li, H. et al. Bioinformatics 25, 2078 (2009).

[16] Zhu, G. et al. Nat. Commun. 12, 2229 (2021).

[17] Markus, H. et al. Sci. Rep. 12, 1–11 (2022).

[18] Widman, A.J. et al. (2022).

[19] Zviran, A. et al. Nat. Med. doi:10.1038/s41591-020-0915-3

[20] Adalsteinsson, V.A. et al. Nat. Commun. 8, 1324 (2017).

[21] Bracher, L. et al. Biomolecules 10, (2020).

[22] GeneCards Human Gene Databaseat <https://www.genecards.org/cgi-bin/carddisp.pl?gene=LOC100130357>

6.1 Supplemental

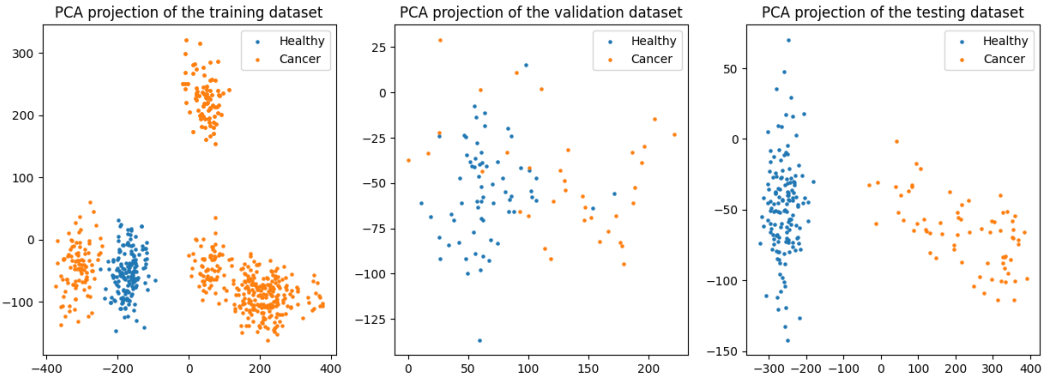

Figure 5: PCA clustering plots for training, testing and validation sets with principle componenets explaining 95% of the variance in the data. The first 2 principle components of each set are plotted against eachother for visualization.

Disease	Sample	Coverage	*in silico* samples	Dataset
Healthy	BH01	96	168	Train
Ovarian cancer (undefined)	IC02	2	4	Test
Skin cancer (Melanoma)	IC03	2	4	Test
Breast cancer (Invasive/infiltrating ductal)	IC04	2	4	Test
Lung cancer (Adenocarcinoma)	IC05	1	2	Test
Lung cancer (Mesothelioma)	IC06	2	4	Train
Uterine cancer (undefined)	IC08	1	2	Test
Ovarian cancer (serous tumors)	IC09	2	4	Train
Lung cancer (adenocarcinoma)	IC10	1	4	Val
Colorectal cancer (undefined)	IC11	2	4	Val
Breast cancer (Invasive/infiltrating lobular)	IC12	2	4	Val
Prostate cancer (undefined)	IC13	2	4	Val
Head and neck cancer (undefined)	IC14	2	4	Train
Lung cancer (Small cell)	IC15	29	81	Train
Bladder cancer (undefined)	IC16	2	4	Train
Liver cancer (Hepatocellular carcinoma)	IC17	42	99	Train
Kidney cancer (Clear cell)	IC18	2	4	Test
Testicular cancer (Seminomatous)	IC19	2	2	Test
Lung cancer (Squamous cell carcinoma)	IC20	23	57	Train
Pancreatic cancer (Ductal adenocarcinoma)	IC21	1	1	Test
Lung cancer (Adenocarcinoma)	IC22	2	4	Train
Liver cancer (Hepatocellular carcinoma)	IC23	2	4	Val
Pancreatic cancer (Ductal adenocarcinoma)	IC24	2	4	Val
Pancreatic cancer (Ductal adenocarcinoma)	IC25	2	4	Train
Prostate cancer (Adenocarcinoma)	IC26	2	4	Train
Uterine cancer (undefined)	IC27	2	4	Train
Lung cancer (Squamous cell carcinoma)	IC28	2	4	Val
Head and neck cancer (undefined)	IC29	1	4	Test
Esophageal cancer (undefined)	IC30	1	2	Test
Lung cancer (Small cell)	IC32	2	4	Val
Colorectal cancer (Adenocarcinoma)	IC33	1	4	Test
Breast cancer (Invasive/infiltrating lobular)	IC34	2	4	Train
Breast cancer (Ductal carcinoma in situ)	IC35	18	48	Train
Liver cancer (undefined)	IC36	2	4	Test
Colorectal cancer (Adenocarcinoma)	IC37	38	99	Train
Bladder cancer (undefined)	IC38	1	2	Test
Kidney cancer (undefined)	IC39	2	4	Train
Prostate cancer (Adenocarcinoma)	IC40	1	2	Test
Testicular cancer (Seminomatous)	IC41	2	4	Train
Lung cancer (Adenocarcinoma)	IC42	2	2	Train
Skin cancer (Melanoma)	IC43	2	4	Train
Esophageal cancer (undefined)	IC44	2	4	Train
Breast cancer (Ductal carcinoma in situ)	IC46	2	4	Val
Pancreatic cancer (Ductal adenocarcinoma)	IC47	2	4	Train
Breast cancer (Invasive/infiltrating lobular)	IC48	2	2	Train
Pancreatic cancer (Ductal carcinoma)	IC49	11	12	Train
Pancreatic cancer (Ductal carcinoma)	IC50	10	14	Test
Pancreatic cancer (undefined)	IC51	11	14	Train
Pancreatic cancer (undefined)	IC52	11	14	Test
Healthy	IH01	104	143	Test
Healthy	IH02	30	63	Val
Healthy	IH03	2	2	Val
Total		462	951	

Table 4: Table showing the distribution of each patient's data throughout the dataset, including sample sequencing coverage and number of *in silico* samples generated.