`

A METHOD FOR IDENTIFICATION OF PANCREATIC CANCER THROUGH METHYLATION SIGNATURES IN CIRCULATING CELL-FREE DNA

A DISSERTATION IN
Biomedical and Health Informatics
and
Computer Science

Presented to the Faculty of the University
of Missouri-Kansas City in partial fulfillment of
the requirements for the degree

DOCTOR OF PHILOSOPHY

by

NEIL ANDREW MILLER

B.A. Tufts University, 1992
M.S. University of Missouri-Kansas City, 2021

Kansas City, Missouri

2022

`

A METHOD FOR IDENTIFICATION OF PANCREATIC CANCER THROUGH METHYLATION SIGNATURES IN CIRCULATING CELL-FREE DNA

Neil Andrew Miller, Candidate for the Doctor of Philosophy Degree

University of Missouri-Kansas City, 2022

ABSTRACT

Pancreatic cancer has high mortality rates in comparison to other cancers due to limited treatment options and challenges in detecting the disease. Early diagnosis is difficult; successful outcomes are directly tied to detection of the cancer before it can spread throughout the body. Evaluation of circulating cell-free DNA (cfDNA), specifically, detection of circulating tumor DNA (ctDNA), is being explored as an approach for non-invasive 'liquid biopsy' that can be deployed widely and cost-effectively to screen for early signs of disease. DNA methylation signatures found in cfDNA can serve as a biomarker for detection of cancer. Previous efforts to detect pancreatic cancer using cfDNA showed limited sensitivity for detection. This manuscript describes work to develop a method for early detection of pancreatic cancer in circulating cell-free DNA using publicly available data from The Cancer Genome Atlas (TCGA) and the Gene Expression Omnibus (GEO). The work includes the development of methods for simulation of samples and cell-free DNA data at multiple ctDNA concentrations as well as a pilot implementation and evaluation of a machine learning model for automated classification of cfDNA samples.

Chapter 1 introduces background of pancreatic cancer, cfDNA, liquid biopsy, DNA methylation and DNA methylation in cancer as well as a review of previous efforts to develop early detection applications.

Chapter 2 describes the identification of DNA methylation markers that distinguish pancreatic tumor from normal pancreas and blood using publicly available data from the TCGA and GEO. In addition, the chapter outlines the development of a machine learning model to classify samples as tumor or normal based on these markers.

Chapter 3 gives an overview of the challenges of genetic data simulation and the development of a novel tool, Heisenberg, for simulating DNA methylation data and cfDNA methylation data. This chapter also illustrate the use of Heisenberg to simulate normal blood samples and pancreatic cancer cfDNA samples.

Chapter 4 describes the development of a neural network classification model for detection of pancreatic tumor in different concentrations of cfDNA using simulated samples. The chapter also reports the detection performance of the model using different model training strategies.

APPROVAL PAGE

The faculty listed below, appointed by the Dean of the School of Graduate Studies have examined a thesis titled "A Method for Identification of Pancreatic Cancer through Methylation Signatures in Cell-Free DNA," presented by Neil A. Miller, candidate for the Doctor of Philosophy degree, and certify that in their opinion it is worthy of acceptance.

Supervisory Committee

Gerald J. Wyckoff, Ph.D., Committee Chair
Department of Biomedical and Health Informatics
Division of Pharmacology and Pharmaceutical Sciences

Reza Derakhshani, Ph.D.
Department of Computer Science and Electrical Engineering

Monica Gaddis, Ph.D.
Department of Biomedical and Health Informatics

J. Steven Leeder, PharmD, Ph.D.
Division of Pharmacology and Pharmaceutical Sciences
Department of Pediatrics

Mark Hoffman, Ph.D.
Department of Biomedical and Health Informatics
Department of Pediatrics

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Figure Page

LIST OF TABLES

Table Page

ACKNOWLEDGMENTS

I would like to offer my heartfelt thanks to the numerous and varied people who helped me in my studies:

To my advisor and committee chair Dr. Gerald Wyckoff for his persistent patience and for showing me the way forward especially when I thought there was none. To the other members of my committee for their expertise and guidance: Dr. Monica Gaddis, Dr. Reza Derakhshani, Dr. Steve Leeder and Dr. Mark Hoffman. Additionally, to Dr. Hoffman, Dr. Leeder, Dr. Laura Fitzmaurice, Dr. Dinakar Dinakarpandian and Dr. Callum Bell who gave support at the earliest stages when the first steps were still in contemplation. To Ada Solidar for a first and extensive review of the draft manuscript.

To my friends and colleagues, too many to list here, including but not limited to those at Children's Mercy, Kansas City and Bionano Genomics.

To the Millers: my parents Roger and Ellen, my brother Roger, my brother Justin, my sister-in-law Lara and my nephews Oliver and Felix. Thank you for listening and not asking about completion dates too often. To my children Ruby and Jasper Miller who remain an inspiration and whose own academic awakenings have helped push me forward. To Natalie Walker, whose independent scholastic achievements served as an example. And, finally, to my wife, Dr. Sarah Soden, who counseled and consoled through it all.

Some of the computing for this project was performed on the Beocat Research Cluster at Kansas State University, which is funded in part by NSF grants CNS-1006860, EPS-1006860, EPS-0919443, ACI-1440548, CHE-1726332, and NIH P20GM113109.

CHAPTER 1

INTRODUCTION

Pancreatic Cancer

Overview

Pancreatic cancer is the third most common cause of cancer-related deaths in the US [1] and is projected to be the second most common by 2030 [2]. For the years 2013-17, overall incidence of pancreatic cancer was approximately 12.9 per 100,000. An estimated 60 thousand new cases will be diagnosed in the US in 2021 with more than 48 thousand deaths. Pancreatic cancer is rare in younger patients, with less than .8% of new cases occurring in patients under age 35 [3]. Median age of diagnosis is 71 [3, 4].

Pancreatic cancer is the deadliest common cancer [4], with an overall 5-year survival rate of 10% compared to greater than 90% for breast, skin melanoma, testicular, thyroid and prostate cancers [1]. Approximately 90% of those diagnosed with the disease eventually die from it, with 70% succumbing due to pervasive metastases. The remaining 30% with limited metastatic disease die with substantial primary tumors [4]. Prognosis for late stage pancreatic cancer patients is particularly poor, with 5-year survival rates of 3% compared to 34% for early stage tumors localized to the pancreas without metastasis [1].

Pancreatic cancer is broadly divided into two groups: exocrine cancers, which account for 95% of cases, and endocrine cancers, such as pancreatic neuroendocrine tumors, which make up the remaining 5%. Pancreatic adenocarcinoma (PAAD) is the primary histological sub-type of pancreatic cancer, accounting for nearly 95% of the exocrine cancers and about 90% of pancreatic cancers overall [5].

Risk factors for the disease include smoking, obesity, chronic pancreatitis, and diabetes. Inherited predisposition is estimated to account for 5-10% of cases, but a genetic basis has not been established [4].

Tumor genomes have been studied to identify and characterize somatic mutations, including attempts to determine which mutations are "driver mutations" conferring selective growth advantage to cancer cells [6]. Genomic studies have shown that more than 90% of pancreatic tumors show gain of function mutations in the *KRAS* gene, a gene involved in cell signaling pathways that control cell growth and differentiation [4, 7]. Accompanying loss of function variants are frequently observed in tumor suppressor genes such as *TP53, CDKN2A* and *SMAD4* [4, 8-11]. Mouse models of pancreatic cancer are consistent with a hypothesis of progression in which *KRAS* lesions contribute to the initiation of neoplasms, while the rate of growth increases or decreases according to the severity of mutations in the suppressor genes [4].

As with other cancers, exome sequencing of pancreatic tumors has shown significant genomic heterogeneity within each tumor, with multiple distinct sub-clonal cell populations developing from a common originating cell. Metastases are thought to arise from specific subclones. Molecular sub-types with an associated metastatic phenotype have been identified based on *SMAD4* and *TP53* mutations[4].

Diagnosis and Treatment

The high mortality rate of pancreatic cancer is due to a combination of diagnostic and treatment challenges. First, pancreatic cancer is usually not diagnosed until it has reached an advanced stage. Early detection is made difficult by the lack of reliable molecular diagnostic markers[5]. Two common performance measures of a diagnostic test are sensitivity, the

ability of a test to correctly identify an individual with disease as positive for the disease, and specificity, the ability of a test to designate an individual without disease as negative for the disease. Serum carbohydrate antigen 19 (CA-19-9) is an FDA approved biomarker for prognostic monitoring of patients with pancreatic cancer but generally has inconsistent sensitivity (41-86%) and specificity (33-100%) when used as a marker for detection [12-14]. False positive results are common in patients with other pancreatic conditions such as obstructive jaundice and chronic pancreatitis. False negative results occur in patients with the Lewis-negative blood group who do not have detectable CA19-9 levels and which occurs in 5-10% of the population [5, 15]. Its use as a diagnostic is generally discouraged[14, 15].

Growth and spread are both rapid and silent so that patients with early-stage tumors are frequently asymptomatic [16]. Late-stage symptoms are often due to metastasis rather than the primary tumor itself, with patients presenting with fatigue, loss of appetite, weight loss, itchy skin, jaundice and abdominal or back pain as well as other signs and symptoms that are not specifically indicative of pancreatic cancer on their own. Primary pancreatic tumors are not able to be palpated during medical exams, though concomitant swelling of the liver or gall bladder may sometimes be detected by a physician [17].

Diagnosis is currently made through imaging procedures such as positron emission tomography (PET) scans, computed tomography (CT) scans, magnetic resonance imaging (MRI), endoscopic or laparoscopic ultrasound and endoscopic retrograde cholangiopancreatography (ECRP) [4]. Differentiation between benign and malignant masses can only be determined via biopsy and histological evaluation [18]. Diagnostic procedures are complex, invasive and impractical for broad application as screening measures [14, 18].

A second factor contributing to the lethality of pancreatic cancer is the lack of treatment options [4, 19]. Pancreatic cancer is aggressive and shows poor response to standard chemotherapeutic agents, making complete tumor resection the singular curative treatment [14, 18, 20] . Tumor treatment is based on tumor stage, which is usually determined by abdominal CT scan [4].

Stage I or II cancers, which are localized and have not spread, are categorized as resectable or unresectable, depending on the involvement of local blood vessels including the celiac artery, superior mesenteric artery and vein, portal vein and hepatic artery. Only 15-20% of patients are candidates for surgical resection [4]. Tumors in the head and neck of the pancreas are removed with a pancreaticoduodenectomy (the Whipple procedure) while tumors in the body or tail of the pancreas are removed with a distal pancreatectomy, often accompanied by a splenectomy [4].

Chemotherapy and radiation therapy may be used as neoadjuvant treatments to try to shrink tumors prior to surgery, especially in tumors classified as borderline resectable. Chemotherapy as an adjuvant therapy post-surgery is often used to reduce the risk of relapse and metastasis and has been shown to improve overall survival rates. Radiation therapy as an adjuvant therapy has been evaluated, however results are conflicting and inconclusive [4].

Stage III cancers, which are locally advanced but not resectable, are usually treated with chemotherapy in an effort to slow cancer growth and prolong life. These cancers are generally treated without expectation of cure. Stage IV cancer, where the cancer has metastasized to distant organs, has extremely poor outcomes, with median survival times measured in months; however, survival up to two years has been observed in approximately

10% of patients receiving a multiagent chemotherapy regimen including fluorouracil, irinotecan, oxaliplatin and leucovorin (FOLFIRINOX) [4].

The difficulty in early detection of pancreatic cancer combined with the dismal outcomes of late-stage patients highlight a need for improved diagnostic methods. Early diagnosis has significant therapeutic benefits. A non-invasive, comprehensive test for pancreatic cancer with high sensitivity could greatly aid the treatment of the disease by enabling early detection through routine monitoring in a manner that is not currently feasible.

Non-invasive Screening with Liquid Biopsy

Precision medicine is an approach to medical care in which a patient's individual makeup, as assessed through genomic, genetic and epigenetic profiles, as well as lifestyle, environment and other biomarkers, is used to tailor treatment of an individual for maximum therapeutic benefit [21]. Pharmacogenomics, for example, uses an individual's genome to inform drug selection and dosing, while precision oncology uses the genetic profile of a patient's cancer to guide treatment and define therapies targeted at specific mutations observed in the patient [22].

Stratification is the practice of dividing patients into clinically and biologically meaningful groups based on shared characteristics. In molecular stratification, biomarkers such as genetic mutations, transcriptomic profiles or quantification of metabolites are used to define subtypes that share patterns of disease progression, prognosis or response to treatment [23, 24]. In precision oncology, molecular stratification is used to classify tumors into specific subgroups to enable targeted treatment and management [25]. Tissue-based biopsy remains the standard of care for diagnosis and molecular stratification in precision medicine applications; however, it has several limitations. Tissue biopsies require samples taken

directly from matter such as tumor or bone marrow that may not be easily accessed and which may require time-intensive or painful, invasive procedures to extract. These procedures carry the risk of complications such as infection, internal bleeding, and significant patient discomfort [26].

An additional challenge with solid tumor specimens is that common preservation methods such as formalin fixation, aimed at slowing tissue necrosis, have been shown to degrade the sample, adding noise and false positives to many molecular analyses based on samples [26, 27]. Another drawback is that tissue biopsies may not fully represent the heterogeneity of the tumor [28]. As sampling is localized, the risk of missing the genetic diversity present in sub-clonal cell populations that are spatially distributed across the mass of the tumor and which have distinct mutational profiles is increased [26].

Cell-free DNA

Circulating cell-free DNA (cfDNA) are fragments of extracellular DNA that are present in the blood stream or other bodily fluids including urine, saliva and cerebrospinal fluid [26, 28]. While cfDNA is released from both normal and cancer cells, fragments of DNA that originate in primary tumors or metastases are called circulating tumor DNA (ctDNA) [29]. cfDNA fragments are relatively small, averaging around 167 base pairs in size, and have a short half-life between 15 minutes and 4 hours [29]. cfDNA concentrations range from 1 to 100,00 fragments per milliliter of plasma [30]. Cell-free DNA was first found in human blood samples in 1948 [31], however the exact origin and mechanism of release of cfDNA is not fully understood. It is theorized that cfDNA is released from source tissue into the body during apoptosis and necrosis of normal and malignant cells or through

the secretion of extracellular vesicles such as exosomes. cfDNA is found in both healthy and diseased individuals and at elevated levels in cancer patients [32].

Liquid biopsy is the practice of using cfDNA in biological fluids, particularly blood, for diagnostic and prognostic purposes [26]. Liquid biopsy presents a quick, comprehensive, easily obtained and minimally invasive alternative to tissue-based biopsy. Liquid biopsy via assessment of cfDNA levels cannot fully replace tissue biopsy. Liquid biopsy may be used in conjunction with tissue biopsy, especially in applications where it may be superior, such as in screening and monitoring tests where repeated samples are required and the impact on patient health can be reduced through its use. In addition, liquid biopsy may be applied using a variety of biological materials including urine, stool, saliva, pleural fluid and cerebrospinal fluid. Cerebrospinal fluid-based samples are of particular utility in the assessment of primary or metastatic brain tumors, which have limited cellular presence in peripheral blood due to the blood-brain barrier [26].

ctDNA contains tumor-specific mutations and epigenetic profiles which are detectable and can act as diagnostic and predictive biomarkers. Analysis of ctDNA consists of identification, quantification and qualitative evaluation of the tumor-specific genetic material. The primary task of detection requires the pinpointing of tumor-specific signatures in the overall totality of cell-free DNA [5, 26].

cfDNA fragments are primarily found in multiples of 167 base pairs (bp), which corresponds to the unit size of a nucleosome. Factors affecting the rate of cfDNA release include physical trauma, strenuous exercise, pregnancy, inflammation, autoimmune disorders and other physiological processes and disorders such as stroke, myocardial infarction and cancerous cell growth. The short half-life of cfDNA in blood makes it theoretically suitable

for real-time assessment of patient condition, including response to treatment [29]. ctDNA shows higher fragmentation than cfDNA, with shorter fragments between 132 and 145 bp being observed[26]. DNA size selection, a laboratory technique for the targeted capture of DNA fragments of a specific size range, can improve sample quality by increasing ctDNA quantities up to 11-fold [26].

Multiple proof-of-concept applications using liquid biopsy have been developed for finding early-stage cancer and long-term monitoring of tumors both during and after treatment [33]. As an example, the Guardant360 CDx test is an FDA-approved liquid biopsy test used to guide treatment of advanced solid tumors [34]. A prospective study of 323 patients with advanced non-small cell lung cancer showed an increase in detection of actionable mutations over tissue biopsy alone [35]. This increase in detection enabled targeted treatment for 35 patients for whom no actionable mutation was found in tissue biopsy alone. Further, for 31 patients, actionable mutations were found in plasma testing making tissue biopsy unnecessary. Concordance between the Guardant360 test and standard-of-care tissue testing was greater than 90% for four tested biomarkers with an average turnaround time being a week faster than the standard methods [36].

Many applications work by detecting and evaluating of low levels of ctDNA in circulating blood as well as urine, stool, saliva and cerebrospinal fluid [26]. Detection and analysis of ctDNA is complicated by its low concentration in cfDNA [28, 37]. This low fraction of ctDNA makes the differentiation between low-signal variant calls and random noise challenging [38].

The amount and fraction of ctDNA observed in blood is variable according to tumor size and stage, as well as other factors such as the amount of tumor vascularization, lymph-

node involvement, lymphovascular invasion, tumor histology, treatments and the total amount of cfDNA found in the blood [28, 38]. In cancer patients, the amount of ctDNA may vary from < .01% to as high as 60% of cfDNA; ctDNA levels are dynamic and have been shown to change rapidly in response to successful therapy [26].

Extraction and analysis of cfDNA has a number of complications. Specialized kits and protocols are required to isolate cfDNA, as standard DNA extraction methods will result in loss and further fragmentation of the short cfDNA fragments [26]. cfDNA is also difficult to store and preserve, being highly susceptible to contamination from high molecular weight genomic DNA [5]. Transport of samples, with accompanying shaking and moving of tubes, increases cell lysis and contamination and necessitates the use of cell-stabilizing tubes and reagents[39]. The highest quality extraction is therefore from samples processed within 4-6 hours after collection [26]. However, suitable quality cfDNA can be isolated for a period of up to seven days [38]. While samples can be frozen, significant degradation in cfDNA has been observed after a year of storage at temperatures of -20°C to -80°C. It has not been definitively determined whether blood plasma has higher quantity and quality of cfDNA than serum, however there is evidence that serum is more vulnerable to contamination due to the routine delay of processing in comparison to plasma[26].

Evaluation of cfDNA

Diagnostic methods to evaluate cfDNA include Sanger sequencing [40], quantitative PCR, digital and digital droplet PCR and multiple variations based on Next Gen Sequencing (NGS) [5, 26]. Sanger sequencing remains the gold standard for mutation detection in tissue-based biopsies but presents challenges for use in liquid biopsy. First, the small quantity of cfDNA that can be isolated from a sample limits the number of loci that can be interrogated,

as each Sanger sequencing of each target requires separate preparation and handling and therefore consumption of the sample [26]. This means that Sanger sequencing cannot be used for a whole genome approach but must be targeted to a small number of genes or other genomic regions of interest. Second, Sanger sequencing is not sensitive enough to detect allele fractions that make up less than 20% of the total sample [27]. Given that ctDNA is expected to be from between 1% and .1% or lower, Sanger sequencing tests will suffer from unacceptably high false negative rates.

Quantitative PCR (qPCR) has similar issues with sample consumption and sensitivity to Sanger sequencing. While many qPCR kits have been validated for use with Formalin-fixed paraffin-embedded (FFPE) samples, which are common in pathology practice and which introduce technical artifacts with other methods, the method consumes DNA serially as multiple genes are tested [26]. Targeted commercial qPCR kits enrich for common oncogenic mutational hot spots and variants, but the approach cannot be used genome-wide and has no ability to detect rare or novel genomic events. Further, evaluations have shown that qPCR's sensitivity does not allow for detection of variants at low allelic frequencies [26, 30].

Digital PCR (dPCR) and Digital Droplet PCR (ddPCR) have been shown to have superior detection performance as compared to qPCR, enabling more precise quantification of ctDNA [41]. However, the targeting of specific alleles has the recurrent problem of sample consumption, as well as the inability to detect novel variants [26].

Next Gen Sequencing (NGS) has had a profound effect on molecular diagnostics and is becoming routine for evaluation of tumors in precision oncology [42, 43]. A key advance is the ability to assess the entire genome with a single test [26, 44]. NGS enables the

detection of a range of genomic variants, including single-nucleotide substitutions and small insertions and deletions as well as larger structural variation including copy number variants and translocations [43]. NGS decodes DNA using short 'reads' of nucleotide sequence ranging from 75-400 bp in size, making the technology suitable for reading short cfDNA fragments in their entirety. NGS experiments read each nucleotide in the genome multiple times with the average number of times being referred to as 'sequencing depth' or 'coverage depth'. Sequence variant calls are made by evaluating the consensus call at each position. A coverage depth of approximately 30x is standard for germline whole genome sequencing while detection of lower allele fraction somatic mutations in tumors may require up to ten times as much sequence [43].

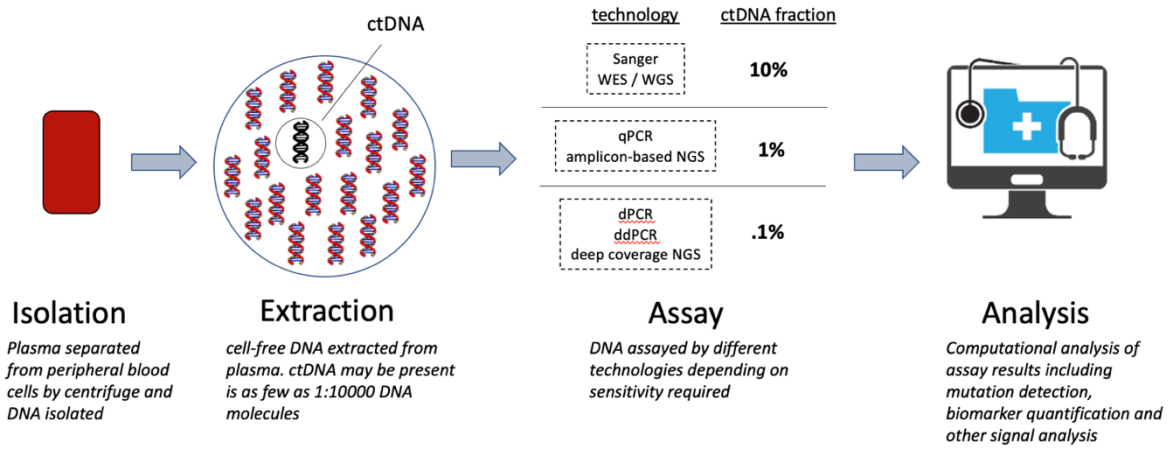

Figure 1.-- Overview of liquid biopsy process

The small allelic fraction of cfDNA represents a challenge for NGS (Figure 1) [5, 26, 45]. For NGS to be viable for detection and analysis of ctDNA, sequencing depth must be greatly increased to capture low frequency events. Collecting additional sequencing data means elevating the instrument time and consumption of sequencing reagents for each sample. Both factors raise costs. Additionally, the increase in the amount of sequencing data

introduces the need for improved error correction methods to distinguish signal from noise. Targeted approaches, in which a subset of the genome is enriched to increase efficiency, have proven effective, though they may introduce technical artifacts of their own.

Whole exome sequencing is a targeted sequencing method that focuses on the 2% of the genome that codes for proteins. By contrast, whole genome sequencing is a comprehensive analysis of the genome without limitation [43]. As an alternative to oncogene panels, whole exome sequencing and whole genome sequencing are utilized for liquid biopsy applications [26]. Whole genome sequencing in particular, has the benefit of minimizing technical biases and enabling superior detection of complex genomic events such as structural variation [43]. Whole genome bisulfite sequencing (WGBS) allows for the detection of DNA methylation, which offers an alternative method of distinguishing tumor-derived DNA from germline [26] [29]. The development of 5-base sequencing from sequencing vendor PacBio promises simultaneous sequencing of nucleotides as well as methylation events. While this technology is in early stages, it holds the potential for evaluation of sequence variation and epigenetic profiling in a single assay [46].

While technical challenges exist, liquid biopsy holds significant promise as a tool for early detection of cancers. Liquid biopsy also allows for the longitudinal monitoring of patients after treatment [28]. The early detection prospects are particularly compelling, given the advantage that early detection offers in management of all cancers. This is of particular importance to pancreatic cancer, where early detection represents the only meaningful chance for successful treatment [4].

Many approaches for ctDNA detection rely on the identification of somatic mutations at significant quantities in cfDNA; however, the interpretation of these variants is difficult

[26]. High throughput whole exome or whole genome sequencing through NGS routinely detects high numbers of variants of uncertain significance (VUS), which present reporting challenges [47]. These variants increase the amount of time needed to analyze a genomic test result since an unclear or incomplete evaluation of these data can lead to errors in patient management[47].

The lack of definitive information of the anatomic source of detected ctDNA is potentially a greater issue than the challenges of variant interpretation [24]. Comprehensive characterization of variants as pathogenic mutations of specific tissue or tumor origin is confounded by the shared mutational profile of many cancers; detection of a deleterious variant cannot on its own be diagnostic of a specific cancer [24]. In the absence of definitive information regarding the tissue of origin, making a confirmatory diagnosis is difficult. Theoretically, a full-body CT scan could be performed in response to a liquid biopsy result for a known tumor driver mutation; however, the financial costs would make this method infeasible [26].

While selected studies have demonstrated the ability of liquid biopsies to detect cancers as early as two years before clinical symptoms have manifested, those same studies reported that less than half of the patients who tested positive for known oncogenic mutations went on to develop cancer. [48, 49]. The prospect of a test with a 50% false positive rate has significant financial and psychosocial implications, particularly given the inability to treat a patient who has tested positive for an oncogenic mutation but shows no other signs of disease. In such cases, the psychological burden of a positive test result when nothing can be done may outweigh the potential benefit of early detection [26].

While the sensitivity and specificity of mutation-based detection methods may be improved, the predictive value of a variant-based test may be reduced by clonal hematopoiesis, benign tumors with mutations, and deaminating mutations that mimic somatic mutation [26]. Taken together, these challenges highlight a need for methods that can more accurately identify source tissue and cell type, as well as distinguish disease state from normal.

DNA Methylation and Human Disease

DNA methylation is a known mechanism for regulation of gene expression that operates through the addition of a methyl (CH_3) group to DNA [14, 29]. Like the transcriptome, which refers to the full range of messenger RNA present in a cell, the complete set of methylation modifications in an individual's genome is referred to as the methylome. The most common and best-characterized DNA methylation pattern in humans occurs at CpG sites, where a cytosine nucleotide is adjacent to a guanine nucleotide [50]. In most cell types, up to 70-80% of autosomal CpG sites are methylated, with the remaining percentage dynamically regulated [29, 51].

DNA methylation acts to limit gene expression by inhibiting gene transcription, with the majority of the variable methylation sites proximal to gene transcription start sites and regulatory elements [51]. Regulatory elements that may be methylated include DNA sequences immediately preceding genes where RNA polymerase binds (promoters), more distant sequences that increase promoter activity (enhancers), and regions where proteins that regulate gene expression bind to a DNA molecule (transcription factor binding sites) [51]. This epigenetic mechanism regulates embryonic development, X chromosome inactivation, chromatin structure, and genomic imprinting, an inheritance process through which the

paternal or maternal allele is preferentially expressed [50]. While DNA methylation does not alter the DNA sequence itself, it is maintained through mitosis and stably inherited by offspring [50, 52].

The regulation of gene expression through the epigenetic code allows for the development of phenotypically different cell types from a single DNA blueprint [50]. While methylation of these regions varies according to aging processes, developmental stage, and external influences that include environmental and lifestyle factors [53], methylation signatures in healthy somatic tissues are generally conserved and stable over a lifetime [51]. Cell identity is defined by the DNA methylation profile, as well as its accompanying transcriptomic state, and distinct differentially methylated regions (DMR) have been identified for tissue types [53] [51].

Aberrant methylation patterns have been associated with a range of diseases in humans. Loss of normal imprinting is associated with Prader-Willi, Angelman and Beckwith-Wiedemann syndromes, as well as Albright hereditary osteodystrophy and pseudohypoparathyroidism Ia [54-56]. Expansion of a triplet CGG repeat accompanied by hypermethylation leads to silencing of the *FMR1* gene, causing Fragile X syndrome [57]. Conversely, contraction of a 3.3kb, CpG rich, D4Z4 repeat region on chromosome 4 leads to hypomethylation, which affects the expression of *DUX4* and causes facioscapulohumeral muscular dystrophy (FSHD) [57]. Finally, deleterious nucleotide variants that disrupt DNA methylation patterns in *DNMT3B* and *ATRX* lead to centromeric instability, facial anomalies (ICF) syndrome, immunodeficiency and Alpha-thalassemia/mental retardation syndrome [58].

DNA Methylation in Cancer

Abnormal DNA methylation patterns are particularly pronounced in cancer, with the most extreme changes associated with the disease state [53]. Global disruptions of methylation are present, with both genome-wide hypomethylation and gene-specific hypermethylation occurring together, each playing a part in all stages of tumorigenesis [53, 59]. Hypermethylation, causes inactivation or reduced expression of the tumor-suppressor genes responsible for DNA damage repair [53]. These states lead to a loss of those functions and contributes to the development of cancer. Similarly, hypomethylation may lead to an upregulation of oncogenes along with reactivation of transposable elements and increased levels of recombination and mutation [59]. Given the role of methylation in cancer, mutations in epigenetic modifier genes such as *DNMT1*, *DNMT3A, MBD1, MBD4, TET1, TET2* and *TET3* have been studied to understand their role in the development and progression of the disease [60].

In addition to global patterns of methylation observed for cancer in general, there has also been significant effort to describe methylation signatures that are unique to particular cancers [59]. The inherent differences in methylation profiles for cell and tissue types, as well as the varying characteristics of somatically acquired DNA methylation changes in different cancers, have led to the identification of type-specific methylation patterns such as the methylation of *BRCA1* and *BRCA2* which is seen in breast and ovarian cancers but not colon or liver cancers [59].

Increasingly, it is recognized that individual cancers are not homogeneous diseases, but a range of sub-types within each cancer type. Molecular sub-types based on somatic mutation, transcriptome and methylome profiles have been identified for a range of cancers

including breast cancer [59], hepatocellular carcinoma [61], leukemia [62], lymphoma [63], lung cancer [23] and pancreatic cancer [24, 64]. These can be used not only to enhance understanding of the mechanisms of the disease, but also to drive individualized treatment decisions and predict variation in progression and clinical outcomes [23, 65].

As an example, combined transcriptomic and epigenetic analysis of 230 lung adenocarcinoma samples defined three major sub-types associated with differing patterns of expression and co-mutation [66]. Each sub-type has a unique mutation rate, transition to transversion rate, and association to smoking history, with membership in the terminal respiratory unit (TRU) sub -type significantly correlated with favorable outcomes and response to treatment. Methylation-based clustering divided samples into three additional groups: an altered CpG island methylator phenotype-high (CIMP-H(igh) group, a group with intermediate levels of methylation at CIMP sites and more normal CIMP-L(ow) group. CIMP-H tumors were observed to have WNT pathway genes significantly enriched as well as *MYC* overexpression [66]. In another study, unsupervised clustering of methylation profiles from central nervous system tumors defined 82 distinct to treatment [67]. As many as 29 of these were associated to a single category in the four-class cancer staging system used by the World Health Organization, meaning that significant differences between groups might not be accounted for in the clinical guidelines [67, 68]. In addition to adding a new level of granularity for stratifying tumors, the use of methylation as a molecular marker has been shown to significantly reduce the number of conflicting interpretations made by multiple analysts for the histopathologic diagnosis of CNS tumors [67].

Molecular stratification and machine learning

Classification is a computational task that uses machine learning algorithms to assign class labels to input data sets based on examples [69]. Classification models are trained with known data to learn to distinguish between classes and then applied prospectively to new data [69]. Machine learning classification models enable the development of automated tools for analyzing large datasets and have great utility in identification of molecule sub-types using biomarkers [69]. As an example, multi-omics analysis of hepatocellular carcinoma integrating genomic, transcriptomic and methylomic data has led to the definition of multiple molecular sub-types with differing survival profiles as well as the development of machine learning models for classifying tumor vs non-tumor liver samples and predicting outcomes [61, 65, 70]. Multiple additional examples exist of both identification of DNA methylation biomarkers and machine learning classification models for acute lymphoblastic leukemia [71], lymphoma, nasopharyngeal carcinoma, breast and colon cancers [23].

While classification algorithms provide the ability to analyze data at a genome-wide scale, they require large numbers of samples to work effectively. Models developed with a reduced sample set may not be generalizable to a broader sample set without additional data, which can be difficult to obtain for more rare sample types such as early-stage pancreatic cancer.

Molecular sub-types of pancreatic cancer

Significant attention has also been paid to defining molecular sub-types for pancreatic cancer using machine learning and genome-wide profiling data [23, 24]. This research aims to gain biological insight to the heterogeneity of the disease as well as to aid risk and treatment stratification[24] . However, molecular stratification of pancreatic cancer lags

behind work done for other cancers, with conflicting evidence on the clinical relevance of detected sub-types. None are currently used in standard practice [24]. Efforts to identify and refine these sub-types into clinically actionable groups would be significantly aided by additional early-stage tumor samples which are relatively rare due to the difficulties in timely detection.

Many studies have focused on the mutational landscape defined by small nucleotide variants and structural variation. Multiple studies, using both array-based hybridization and NGS RNAseq, have reported transcriptomic sub-types with varying degrees of overlap, but generally defining two major lineages that appear to be driven by epigenetic events: Classical-Pancreatic and Squamous [16, 24, 64, 72].

Mishra et al. reported on analysis of DNA methylation data in pancreatic cancer samples that identified differentially methylated regions between normal pancreas and tumor tissue. Profiles included those of pancreas development-related genes and homeobox genes such as *HOXA1, HOXA2, HOXB1, HOXB3, HOXB7, HOXC4, HOXC9, HOXD4, HOXD8, HOXD10, HOXD11, HOXD12* and *HOXD3*. Differential methylation analysis also identified pancreatic cancer-specific hyper- and hypo-methylated distal enhancer sites, while unsupervised clustering of differentially methylated regions was iteratively performed to optimally define three distinct clusters in pancreatic cancer patients [16].

DNA methylation has been shown to play a distinct role in developmental biology as well in disease. Methylation profiles have been identified for multiple diseases and disease states and have been used effectively to distinguish tissue and cell types. Specific patterns have been identified for cancer, broadly, as well as for specific cancer types. Further, methylation profiles offer the ability to define molecular sub-types that stratify samples with

a new level of detail. These factors taken together demonstrate the utility of incorporating methylation into precision medicine applications.

DNA Methylation as a Biomarker in cfDNA

Several factors make DNA methylation signals in cfDNA an attractive biomarker for identification of ctDNA in plasma [14, 29, 73]. Hypermethylation of tumor suppressor genes is an early event in tumorigenesis, meaning that methylation changes may be some of first biological signals of neoplasm, making it suitable for early detection applications [14] .

However, identification of ctDNA through detection of somatic mutations that distinguish tumor DNA from normal DNA is hampered both by the small fragment size of cfDNA and by the low volume of ctDNA in the background of cfDNA. In addition, only a limited number of known recurrent mutations can be reliably interpreted as indicative of cancer.

As discussed, aberrant DNA methylation is present genome-wide in cancer, providing a signal over multiple target regions that is present in both cancer tissue and normal cfDNA. These patterns are repeated across multiple altered individual CpG sites in each region. This redundancy provides a signal that is both robust and resistant to technical artifacts and dropouts of individual portions of the genome that mutation-based analyses cannot tolerate [29, 37]. A key shortcoming in the interpretation of cfDNA through nucleotide variants alone is that the source of the DNA cannot be determined [74]. This can be addressed through the analysis of methylation profiles that are different across tissues and cell types [29]. Because methylated DNA is present in plasma, the identification of tissue of origin of cfDNA is made possible through analysis of these profiles [73, 75].

Methylation-based liquid biopsy has been explored for multiple types of cancer [73]. For example, Xu et al. developed a comprehensive, genome-wide approach for the detection of hepatocellular carcinoma using cfDNA. Through the analysis of 377 tumor samples from The Cancer Genome Atlas as well as blood leukocyte samples from healthy controls, the study identified ten informative candidate methylation sites that could be used to distinguish tumor from blood [65]. Nassiri et al. applied a similar approach to intracranial tumors to define markers for detection, as well as for distinguishing between extracranial and subtypes of tumors that could be used to provide preoperative diagnostic detail. Differentiating between these subtypes through liquid biopsy allows physicians to avoid the invasive surgery needed to collect tissue specimens [76]. In another study, Nuzzo et al. identified differentially methylated regions in plasma and urine cell-free methylomes of 120 patients with renal cell carcinoma and 28 healthy controls [77]. These studies can be considered proof-of-concept, with the next ideal step being broad validation and testing through clinical trials.

Diagnosis of Pancreatic Cancer through Liquid Biopsy

Several studies have been conducted to investigate the feasibility of liquid biopsy for early detection and management of pancreatic cancer [5, 18, 78, 79]. The *KRAS* oncogene has the highest known mutation rate across all cancers and mutations in the gene are associated with lung, colorectal and pancreatic cancers [7]. Given the prevalence of *KRAS* mutations, some studies have focused on detecting these mutations in cfDNA. However, circulating *KRAS* mutations are generally detectable in only 26-73% of patients with pancreatic cancer, while the incidence of the mutations in pancreatic tumors has been assessed at greater than 90%, meaning that many *KRAS* mutations are not being detected in circulating samples [14]. The low rate of detection of *KRAS* mutations may be due to

technical limitations or biological limitations such as the possibility that for some variations of mutant *KRAS* pancreatic cancer cfDNA is not released to the bloodstream [80]. Regardless, while driver mutations in *KRAS* have prognostic value for pancreatic cancer, they are not suitable as a marker for early detection in cfDNA [14].

Given the limitations of variant-based analysis, the use of methylation signals in cfDNA has been explored as an alternative biomarker for the diagnosis and stratification of pancreatic cancer. Levenson et al. compared methylation profiles of 30 patients suffering from pancreatic cancer to those from 30 age-matched control subjects without signs of pancreatic disease. The study was based on the hypothesis that absence of methylation could be indicative of disease and focused primarily on the evaluation of hypomethylation in the promoters of 56 frequently methylated genes. Promoters from five genes (*CCND2, SOCS1, THBS1, PLAU* and *VHL)* were combined into a single composite biomarker and used in a machine learning model for classifying cancer versus normal samples. Five-fold cross-validation results yielded an estimated 76% sensitivity and 59% specificity [81].

In contrast to Levenson, other studies of cfDNA in pancreatic cancer have focused on hypermethylation. Park et al. examined DNA methylation of six genes (*CDKN2A, NPTX2, PENK, SFRP1, RASSF1A* and *UCHL1*) previously reported to be hypermethylated in pancreatic cancer. The study investigated methylation levels in a cohort of 16 pancreatic patients, 29 healthy controls and 13 patients with chronic pancreatitis. DNA hypermethylation of at least one of the six genes was detected in 13 of the pancreatic cancer patients (85%) and only 1 of the controls. However, it was also detected in 8 of the chronic pancreatitis patients, highlighting the potential difficulties in discriminating between the two conditions by methylation analysis [82].

In other studies, Henriksen et al. investigated the possibilities of early detection using the promoter sites of 28 genes nominated through literature review. The group employed a logistic regression model using the methylation status of each gene as a binary variable (i.e., methylated vs. unmethylated) and demographic and behavioral factors such as age and smoking status. An optimal combination of features with the highest predictive power was devised by stepwise backward elimination of variables, resulting in a composite marker that included eight genes (*APC, BMP3, BNC1, MESTv2, RASSF1A, SFRP1, SFRP2* and *TFP12*), plus a demographic factor for age > 65 years. Detection sensitivity, as measured in a set of known cases and controls, was reported as 76% with a specificity of 83% [18, 79, 83].

Finally, Shinjo et al. used data from The Cancer Genome Atlas (TCGA) plus methylation levels from 37 pancreatic cancer samples containing *KRAS* mutations, and three normal samples to identify differentially methylated markers with a high predictive value. The group performed an iterative analysis to first identify differentially methylated genes between tumor and normal pancreas followed by further curation to remove sites consistently methylated in normal whole blood samples. The remaining sites were then compared to methylation data from publicly available samples for lung, breast, colorectal, bladder, kidney, AML, stomach, and skin cancers to find genes that appear to have a unique methylation status in pancreatic cancer. The final list of differentially methylated genes was composed of five genes, *ADAMTS2, HOXA1, PCDH10, SEMA5A, SPSB4*. These markers were then tested using cell-free DNA from 47 patients with pancreatic cancer and 14 normal volunteers. In addition to the methylation predictors, the research group also included *KRAS* mutation status as a diagnostic marker. While direct tumor samples from all cancer patients showed *KRAS* mutations, these mutations were detected in only 23 (49%) of the cancer cfDNA samples.

This highlights the difficulties of detecting the cancer using mutation analysis alone by leaving 24 patients undiagnosed. A combined analysis using mutation status plus the five differentially methylated genes yielded a diagnostic sensitivity of 68% and specificity of 86%, making it potentially useful to rule out pancreatic cancer, but being relatively insensitive at early detection [13].

Discussion

In summary, early detection of pancreatic cancer through non-invasive methods represents a significant opportunity to improve patient outcomes. Currently, pancreatic cancer is often detected after it can be successfully treated leading to a low survival rate for patients who develop the cancer. Liquid biopsy represents a promising method for early detection due its ability to serve as non-invasive, comprehensive screening test that could be applied as part of routine healthcare. However, there are multiple challenges both in the lab and in the identification of cancer signals in data produced by existing assays. Detection of mutations in cell-free DNA is difficult and non-specific to pancreatic cancer. The use of DNA methylation as a biomarker may address some of these issues by enabling tissue origin and cancer-specific detection, however while there have been several investigations into using DNA methylation in cell-free DNA as a platform for early detection, there remains significant room for improvement in test sensitivity and specificity. Further, there is little to no overlap in the candidate gene sets evaluated by previous studies suggesting that the search for predictive methylation features may benefit from an expanded, genome-wide approach. These factors lead to the observation that the field may benefit from exploration of additional methods for early detection of pancreatic cancer.

CHAPTER 2

DEVELOPMENT OF A MODEL TO DISCRIMINATE BETWEEN PANCREATIC TUMOR, NORMAL PANCREAS, AND NORMAL BLOOD

Rationale

An early detection model for pancreatic cancer must be capable of identifying pancreatic cancer ctDNA while also distinguishing circulating tumor cells, not only from normal blood cells, but also from any normal pancreatic cells present in the bloodstream. This may be achieved by identifying a subset of differentially methylated sites that define each of the three tissues. Once these differential sites are defined, a classification model trained using them can be adjusted to account for the reduced tumor DNA signal that is present as a fraction of the cfDNA found in a patient.

Materials and Methods

Samples

The Cancer Genome Atlas (TCGA) was a joint effort between the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) [84]. TCGA generated genomic, epigenomic, transcriptomic and proteomic data for over 20,000 primary cancer samples and matched normal samples for more than 33 cancer types [85]. The TCGA-PAAD project produced data for patients with pancreatic adenocarcinoma including methylation data for 185 primary tumor samples and 10 normal pancreatic tissue samples[85] (Table 1).

Table 1.--TCGA-PAAD tumor sample overview

Tumor stage	Male	Female	Total
i	0	1	1
ia	3	2	5
ib	7	8	15
iia	20	10	30
iib	67	55	122
iii	1	3	4
iv	2	3	5
not reported	2	1	3
Total	102	83	185

Methylation data was generated using the Illumina Infinium HumanMethylation450 BeadChip microarray platform [86], which measures methylation at more than 450,000 distinct sites at single-nucleotide resolution. Methylation status is measured through the use of two types of probes at each of the survey sites; one for the methylated allele and one for the unmethylated allele [87]. Methylation status at each position is calculated as a ratio of signal intensities from the allele-specific probes and reported as a continuous beta value (β) ranging from 0 (unmethylated) to 1 (completely methylated). Methylation data from TCGA are provided as plain text files that define the genomic position of the array probe and a beta value that specifies the detected methylation signal for each site included in the assay. These data are available from the National Cancer Institute Genomic Data Portal [88, 89].

The Gene Expression Omnibus (GEO) is a public functional genomics data repository maintained by the National Center for Biotechnology Information (NCBI) [90]. Normal, cancer free samples were downloaded from the GEO study GSE67393 (*Sex differences of leukocytes DNA Methylation*) [91] and used with pancreatic tumor and paired normal samples from TCGA-PAAD. A total of 117 healthy samples, 63 male and 54 female, were

used. Methylation data was generated using the Illumina Infinium HumanMethylation450 BeadChip platform.

In order to define genomic sites that were differentially methylated between normal pancreas and tumor, novel machine learning models were created using a collection of publicly available packages implemented in the Python programming language including scikit-learn [92], pandas [93], numpy [94] and TensorFlow [95]. Additional models were created using DeepLearning4J, a deep learning toolkit implemented in the Java programming language [96]. Computation was performed in multiple environments including a consumer-grade laptop computer, a High-Performance Computing cluster and Amazon Web Services. Computation was done using the Linux and Mac OS X operating systems. These computations required a modest amount of physical memory (8 GB of RAM) with the primary benefit of the HPC environment being the ability to run multiple simulations in parallel.

A decision tree is a machine learning method that can be used for classification of datasets by creating a tree-like model where each node represents a test of an attribute of the data [69, 97]. Classification of an input dataset is made by following the pathway through the tree to arrive at a specific label assigned to training elements that shared these characteristics [69]. Random forest is a classifier that makes predictions using an ensemble of decision trees. Each individual decision tree makes a prediction according to its own specific rules and the random forest aggregates the results to make a final prediction based on the majority decision produced by the collection of decision trees [97]. Random forest models quantify the importance of individual features in making a decision using the Gini index or Gini importance score [97]. The Gini importance, or mean decrease in impurity, calculates the

feature importance as the total decrease in node impurity averaged over all trees of the ensemble [97]. The ability of random forest models to identify the most relevant features for discriminating between labeled classes makes it a common tool for feature selection in machine learning applications [98].

Previous analyses with TCGA-PAAD methylation samples identified 23,688 sites that are differentially methylated between normal pancreas and tumor tissue [16]. A random forest model was employed to statistically identify a subset of methylation features that could be used for automated discrimination between tumor and normal pancreas, as well as between tumor and normal plasma drawn from individuals without cancer. This approach was modeled on previous efforts to define methylation-based molecular subtypes in hepatocellular carcinoma [61] and tumors of the central nervous system [67].[67]. Tumor and normal pancreatic tissue data from TCGA were combined with methylation data from peripheral blood available from GEO to create an expanded dataset for use in training and testing the model.

Identification of feature sets

A random forest classifier to distinguish between normal pancreas and tumor was created in python using scikit-learn by splitting 195 TCGA-PAAD samples (185 tumor, 10 normal) into training and test sets, with 60% of the data for training and 40% of the data for testing. Relevant features were ranked by Gini importance score, and multiple candidate feature subsets were defined. Subsets included all features with a Gini score > .003 as well as four additional sets containing features with the top ranked 50, 100, 200 and 500 Gini scores. Classification accuracy was improved by implementing feature subsets versus a model that used all differentially methylated sites (Table 2).

Table 2.--Confusion matrices from representative cross-validation run of two random forest models. Correct predictions are shown in the cells pred:Normal/true:Normal and pred:tumor/true:Tumor. Incorrect predictions are shown in the cells pred:Normal/true:Tumor and pred:Tumor/true:Normal.

All differentially methylated sites		
	pred:Normal	pred:Tumor
true:Normal	1	4
true:Tumor	1	72

Relevant features (Gini > .003)		
	pred:Normal	pred:Tumor
true:Normal	2	3
true:Tumor	1	72

Cross-validation is a machine learning technique for estimating a model's performance by repeating the process of model training and testing using different subsets of the dataset [69]. Model accuracy is measured by aggregating the results from all repetitions to minimize bias in any one iteration [69]. Candidate feature sets were evaluated using 10-fold cross-validation, where the process of splitting the dataset into training and test subsets, training the model and testing it was repeated 10 times. Initial results showed the top 50 and top 500 probe sets performing the best, with a prediction accuracy of 97.9 % +/- .0051.

Following the hypothesis that stage III and IV cancers represent extremes and that inclusion of those samples might skew the detection of stage I and II cancers vital to an early detection application, random forest training and feature set identification was repeated excluding 4 stage III and 5 stage IV samples (*N*=176). Results from 10-fold cross-validation showed an improvement in accuracy using the Gini > .003 (99% +/- .068) and top 50 (98.4% +/- .047) probe sets, with a decreased performance observed in the top 500 set (Table 3).

Table 3.--Cross-validation results of model trained with tumor stages all vs stage I + II samples only

	All Tumor Stages		**Stage I + II only**	
probe set	*# probes*	*accuracy / 95% CI*	*# probes*	*accuracy / 95% CI*
Gini > .003	23	.974 (+/- 0.068)	31	.990 (+/- 0.041)
top 50	50	.979 (+/- 0.051)	50	.984 (+/- 0.047)
top 100	100	.974 (+/- 0.068)	100	.974 (+/- 0.068)
top 200	276	.974 (+/- 0.068)	356	.974 (+/- 0.068)
top 500	500	.979 (+/- 0.051)	508	.969 (+/- 0.067)

Evaluation of methylation values in target probe sets

To further evaluate the feasibility of using methylation values to discriminate between normal blood, normal pancreas and pancreatic tumor, density plots were created in R to visualize the difference in methylation signals using the candidate probe sets (all differential, Gini > .003, top 50, 100, 200 and 500 probes). In all sets, pancreatic tumor signals were observed to be distinct, with most probes showing a unique clustering between beta values of .25 and .75. Normal blood and pancreas showed a marked bimodal distribution at 0 and 1 (Figure 2). The difference in patterns between tissue types highlights the potential for using methylation values for these specific probes to distinguish pancreatic tumor from normal blood or pancreas. The identification of tissue- and cancer-type specific methylation patterns suggests that pancreatic tumor and normal pancreas patterns can be expected to be distinct from other cancer types, however a direct comparison remains to be done.

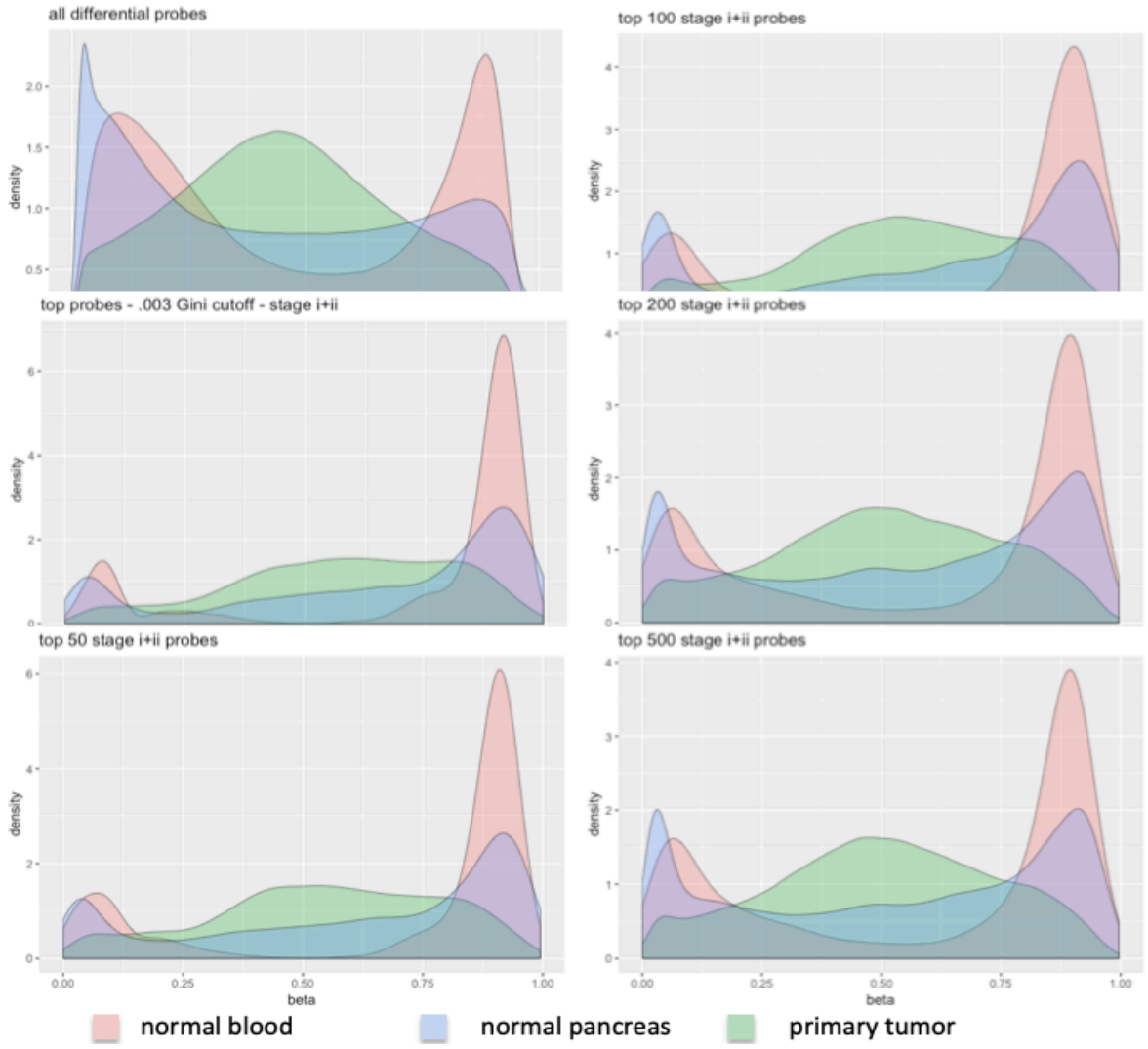

Figure 2.--Density plots of methylation values in probe sets

Evaluation of genes associated to target probe sets

Target probe sets were evaluated to assess the genes associated with the differentially methylated probes. Genomic coordinates of promoter regions and identifiers of their associated genes were downloaded from the UCSC Genome Browser for the hg19/GRCh37 human reference genome [99, 100]. These data were merged and cross referenced with gene positions from NCBI Annotation Release 105.20211022 to create a lookup map of genomic

positions and features. Probe positions for the Infinium HumanMethylation450 BeadChip were compared to this reference set to identify features that overlapped with each probe position. Finally, these data were used to annotate the probes of each candidate target set with their associated genes. A superset of 426 distinct genes was identified across all probe sets (Table 4, detailed list in Appendix A).

Table 4.-- Counts of distinct genes associated with methylation probes in each probe set

probe set	**genes**
Gini > .003	25
top 50	42
top 100	87
top 200	301
top 500	426

A biological pathway describes a set of interactions among molecules in a cell that result in a change in the cell such as the activation or inactivation of genes, synthesis of a cell product or metabolism of a substance [101]. The Reactome Knowledgebase is a curated database of pathways that describe the details of a range of biological processes including signal transduction, transport, DNA replication and metabolism as networks of molecular actions [102]. To further the characterize the genes found associated to each probe set, a pathway enrichment analysis was performed using Reactome. The analysis compared input gene lists with the pathway database to identify pathways that appear to be represented more often in the input list than would be expected by chance. Each pathway is assigned a probability score and ranked.

Pathway enrichment analysis of the probe sets produced a list of significantly overrepresented pathways in each set. Ketosis related pathways including Synthesis of

Ketone Bodies (Reactome identifier R-HSA-77111), Ketone body metabolism (R-HSA-74182) and Utilization of Ketone Bodies (R-HSA-77108) were found in the 25 most significant pathways in all analyses and were found in three of the top 6 in the Gini > .003 and top 50 probe sets. These pathways all include the *BDH1* and *AACS* genes whose methylation statuses are measured by probes cg18163092 and cg00417147 respectively on the Infinium HumanMethylation450 BeadChip microarray. Methylation values for these probes are consistently high in normal blood while showing lower average methylation and ten-fold higher variability in pancreatic tumor samples (Table 5). Alterations in these pathways are potentially of interest given the involvement of these pathways in high-energy consuming states such as rapid cell division as well as the relationship between insulin and ketosis.

Table 5.--Differential methylation status of genes in top Reactome pathways

gene	**pathways**	**probe**	**blood**			**tumor**		
			range	*avg*	*stdev*	*range*	*avg*	*stdev*
BDH1	Synthesis of Ketone Bodies Ketone body metabolism Utilization of Ketone Bodies	cg18163092	.931-1.0	.977	.015	.115-.986	.759	.177
AACS	Synthesis of Ketone Bodies Ketone body metabolism	cg00417147	.630-.941	.851	.033	.127-.942	.541	.151
TSNAX	Small interfering RNA (siRNA) biogenesis	cg23050705	.321-.882	.674	.067	.284-.826	.537	.128
AGO2	Small interfering RNA (siRNA) biogenesis Post-transcriptional silencing by small RNAs	cg23731089	.343-.960	.792	.077	.038-.758	.271	.145

The top 3 pathways in the top-200 probe list also included two small RNA pathways, Small interfering RNA (siRNA) biogenesis (R-HSA-426486) and post-transcriptional silencing by small RNAs (R-HSA-426496), both of which include the genes *TSNAX* and *AGO2.* As with the ketosis genes, methylation status for these genes is lower on average than

that of blood and generally much more variable. *AGO2* in particular is of interest given its average state of hypomethylation and assumed activation since it is known to play a role in tumorigenesis, and its expression has been associated with liver cancer [103] .

An examination of these pathways may be useful for elucidating both the genesis of pancreatic tumors, as well as giving a new view into the early progression of pancreatic cancer. As such, these pathways and genes represent novel biological results that may be useful for further exploration by cancer researchers. These specific pathways have not been previously identified in early pancreatic cancer.

Neural network classifier

Neural networks are a class of machine learning algorithms that are inspired by the human brain and use connected nodes that mimic the connections of biological neurons [104, 105]. Neural networks organize nodes, or artificial neurons, into interconnected layers where each node has an associated weight and activation threshold [105]. Like other machine learning algorithms, neural networks use training data to learn discriminating features of a dataset and can then be applied prospectively on previously unseen data [105]. Neural networks have been successfully employed for a variety of generalized tasks including image and speech recognition and classification as well as specific applications in cancer genomics [69, 105, 106]

A proof-of-concept neural network classifier was created to evaluate the feasibility of a deep learning model. The model was initially implemented using TensorFlow, with the final implementation done in Java using DeepLearning4J. The model was created as a multi-layer perceptron, with an input layer, one feed forward hidden layer and an output layer with two classes (tumor/normal). The rectified linear unit (ReLU) activation function was used for

the input and hidden layers using a L1 normalization of 1e-4 and L2 normalization of 1e-6 with a dropout rate of .8. The Adam optimization algorithm was used along with a negative log likelihood loss function. Finally, the output classification layer was created using the softmax activation function, which calculates probabilities for each class defined.

A training program was written to simplify iterative training and testing using different probe sets and model hyperparameters such as batch size and learning rate. To improve efficiency, early stopping was implemented, meaning model training was halted when performance ceased to improve. Implementation of early stopping led to a significant decrease in model training time, reducing a full training run of >1000 epochs from days to minutes.

Training and testing were performed 50 times using different combinations of batch size, learning rate and probe set. Model sensitivity ranged from 41% to 100%, with five variations repeatedly producing sensitivities of 98.48%. The best-performing models were created using the top 200 and 500 probe sets (with and without stage iii and iv samples) with a batch size of 200 and learning rate of .01. One iteration of the top 200 probe set model had 100% sensitivity when early stopping was configured to allow 100 iterations without improvement. The second model created with the top 200 set and configured to allow a maximum of 20 iterations without improvement and the top 500 probe set both had an F1 score of .985, 99.5% accuracy, 99.72% specificity with a false negative rate of .015. All models took less than 25 minutes to train using a consumer grade laptop computer (Table 6).

Table 6.--Classification performance of top six neural network models

Probe set	TP	FP	TN	FN	Sens.	Spec.	PPV	NPV	Accuracy	F1	Run time (mm:ss)
Top 200 Stage-i-ii	65	1	353	1	0.985	0.997	0.985	0.997	0.995	0.985	06:52
Top 500 Stage-i-ii	65	4	350	1	0.985	0.989	0.942	0.997	0.988	0.963	05:08
Top 500 Stage-i-iv	65	1	353	1	0.985	0.997	0.985	0.997	0.995	0.985	12:09
Top 200 Stage-i-iv	66	2	352	1	0.985	0.994	0.970	0.997	0.993	0.977	11:16
Top 200 Stage-i-ii	66	2	352	0	1.000	0.994	0.971	1.000	0.995	0.985	23:49
Top 500 Stage-i-ii	65	4	350	1	0.985	0.989	0.942	0.997	0.988	0.963	09:00

Discussion

The investigation of methylation data from pancreatic tumor, normal pancreas and blood yielded a number of insights. Implementing a random forest classifier and training it to classify samples as tumor vs. normal yielded five candidate probe sets in addition to the full set of differentially methylated site previously identified by Mishra et al. Retraining the random forest with these probe sets led to an increase in classification accuracy over a model created with all differential sites. Visualization of the distribution of methylation values for these candidate sets across tissue types shows a pattern of methylation that distinguishes pancreatic tumor from both normal pancreas and blood.

The identification of these probe sets also enabled an investigation into genes that are differentially methylated between the tissue types. Pathway enrichment analysis nominated ketosis and small interfering RNA pathways that have not received previous study in pancreatic cancer. Additionally, there is relatively little crossover observed between the genes associated with the predictive features nominated here and the genes targeted by previous work on early detection methods for pancreatic cancer. Specifically, of the 426

genes represented, only two were incorporated into the best performing methods from Henriksen et al. (*BNC1)* and Shinjo et al. (*SPSB4*). These gene findings underscore the utility of a comprehensive data driven approach to identifying potentially unconsidered predictors for pancreatic cancer.

The proof-of-concept development of a neural network classifier gave the opportunity to evaluate the candidate probe sites as well as some of the possible variations in parameterization that could be used for a final classification model. Overall feasibility of a model to distinguish between tissue types has been demonstrated. These results serve as a basis for expansion of the automated classification of pancreatic tumor from normal blood with the best-performing models and probe sets used to guide further development and extension to detection in cell-free DNA.

CHAPTER 3

HEISENBERG: A TOOL FOR AUGMENTING DNA METHYLOME DATASETS AND SIMULATING CELL-FREE TUMOR DNA SIGNALS

Overview

An obvious obstacle to studying any complex biological phenomenon of interest is lack of available representative data. While some initial research can be performed with small amounts of data, computational and data science applications require vast quantities of data upon which models can be built and analyses run.

In the life sciences, a great deal of biological data is available in public repositories or through controlled access mechanisms. However, relying exclusively on prior studies to have collected and published the exact datum needed for future research cannot possibly succeed for all areas of interest. More commonly, researchers unable to collect the specific data they need directly--either through lack of resources, opportunity, or access--must instead look for pragmatic ways to use available data to approximate reality.

Study of the methylation signals of pancreatic cancer in cell-free DNA suffers from three distinct gaps in publicly available data. First, The Cancer Genome Atlas project (TCGA), while an invaluable resource that represents significant effort and expenditure for data collection and dissemination, has collected little data on pancreatic cancer. Of the 33 cancer types and 11,315 cases available from the TCGA, only 185 pancreatic cancer cases are included, making the TCGA-PAAD the 21st most populous study, well short of the numbers in a cohort like the TCGA-BRCA breast cancer project, which includes 1095 cases [88]. Second, as a data type, methylation lags other -omics sources, such as DNA and RNA sequencing. Of 21,819 data collection platforms at the Gene Expression Omnibus (GEO), only 47 were methylation-related on 1/31/2021 [90]. Finally, while multiple studies have

made methylation data available for peripheral blood samples, cfDNA samples known to contain pancreatic cancer are essentially non-existent in the public domain.

Biological simulation

Applications

Many tools have been created to computationally simulate genetic data , using a variety of approaches and end goals. These efforts can be broadly categorized into two groups: tools for generating data to develop and test the performance of statistical methods, and tools for simulating complex biological interactions, with multiple dimensions and scenarios modeled [107].

In the first case, data simulation is used for testing the ability of computational approaches to detect known patterns or higher-level data structures and associations in high dimensional genomic data. For example, simulators have been developed for testing the ability of statistical methods to detect genetic features, such as haplotype blocks, through phasing of independent single-nucleotide-polymorphism (SNPs) calls or to detect disease-associated SNPs [107]. In these instances, simulation serves a valuable purpose in creating datasets that can be used for methods development by providing data where the features of interest are deliberately embedded into the dataset and used as a target for detection. As the embedded true state of the dataset is known, a computational method can be implemented and refined to correctly identify features of interest while ignoring features that are not relevant.

The controlled nature of simulated datasets also enables the measurement of a method's accuracy through the defining of statistical measures used to assess test performance (i.e., true positives, true negatives, false positives, and false negatives) [108]. In

this way, data simulation enables methods development specifically by limiting the data to those cases where ground truth of the sample is known definitively [109]. After a method has been developed to perform satisfactorily on the known dataset, it can then be applied as a tool for discovery by being run on real datasets where the underlying truth is not known [110]. Confidence in the method's ability to correctly interpret data is established in the simulation set, and then refined through the real-world application [111]. Ideally, both a simulation method and a method for interpreting those simulated data can be refined through the application of the method to non-simulated datasets paired with validation of results via an independent method [109, 110].

In the second case, where tools are used to simulate complex biological interactions, the simulation is not a means for methods development, but rather serves as an end by enabling the study of complex scenarios with multifactorial variables [107]. This can be done through the creation of models whose parameters can be adjusted to observe the effects of a variable of interest. Population and evolutionary simulation have received a great deal of focus, with models created to study effects such as environmental factors, population bottleneck and expansion, natural selection, random mutation and recombination [107]. Additional models generate theoretical case/control populations with realistic linkage disequilibrium and analyze allele frequency patterns to help evaluate hypotheses on phenotype-to-genotype associations [107, 112].

Simulation methods

Population simulation algorithms primarily fall into three main categories: backward-time or coalescent simulation, forward-time simulation, and resampling [112]. Briefly, coalescent simulations work by starting with a current population of individuals, and then

tracing alleles back to a most recent common ancestor. Then, individuals are regenerated back to the current generation introducing random mutations into the newly created genealogy [112]. The forward-time method is similar, but starts simulation at a current population, incorporating mutations and other effects without first tracing to a most recent common ancestor. Common parameters of these simulation methods include demographic factors, such as migration and population bottlenecks, along with selection pressures and genomic factors, including recombination rates and variation hot spots [112].

In contrast, resampling approaches generally do not attempt to manage the evolutionary complexities incorporated by backward- and forward-time approaches, but rather generate samples through random selection from an existing data set. While resampling approaches lack the fine-grained control provided by backward- and forward-time methods, they have significant advantages in computational efficiency, as well as providing the ability to generate a potentially unlimited number of samples [112].

A significant issue with data simulation is the fact that data are generated according to inherent assumptions of how the data would appear *in vivo*. In other words, the validity of the generated data is directly related to how well the rules of simulation reflect the reality of the factors involved in producing a biological condition of interest [112]. In this area, resampling approaches provide an advantage by avoiding many of the pitfalls of incorrect assumptions potentially present in other methods using actual samples as a source [112].

These factors make re-sampling approaches attractive for the study of real genomic data while limiting their suitability as a method for generating diverse population data for use in observing the evolutionary process or effect of recombination and mutation parameters. Coalescent and forward-time approaches are considered to be superior for these applications

[112]. Further, resampling approaches are inherently limited by the makeup of the original dataset. For example, a lack of ethnic diversity in the original source data may limit the applicability of the simulation to populations that are not well represented [113] .

Existing tools

Researchers have produced an array of genetic simulation programs. Recognizing the value of these simulation programs for their range of usability, reliability, performance and application areas, the National Cancer Institute initiated the Genetic Simulation Resources (GSR) website both to catalog genetic simulation tools and provide researchers a means for comparing tools and selecting the most appropriate ones for their study [107]. The GSR website currently catalogs more than 100 genetic simulation tools and offers easy-to-use facilities for searching tools of interest. Further, the GSR has established a certification program to set standards for simulation programs that include documentation, support, and ease of installation and use. The GSR provides an extensive list of simulators, including tools for simulating a range of Next Gen Sequencing (NGS) data types, DNA sequences, and population/evolution simulators. However, there are no existing tools for the creation of epigenetic data sets based on DNA methylation as there are for diploid and haploid DNA, protein, or RNA sequences. There are no tools of any kind for simulating cfDNA sets [114].

A common analysis method for studying epigenetic variation is through the identification of differentially methylated regions (DMRs) between two sets of samples. Recognizing the lack of formal evaluation of the performance of DMR identification methods, Lacey *et al.* developed an algorithm to simulate realistic datasets based on reduced representation bisulfite sequencing (RRBS). Like whole exome sequencing, RRBS employs laboratory techniques for target selection to sequence only the areas of the genome with a

high CpG content to reduce costs. The algorithm produced by this work effectively models differentially methylated regions using a two-state Hidden Markov Model combined with a random noise function. The algorithm simulates sources of technical variation in sequencing-based approaches that reduce accuracy. These factors include input source and quality of DNA, errors in lab procedures to isolate and process DNA and the overall sequencing error rate inherent in NGS [115].

Using this algorithm, researchers generated multiple datasets that mirror the characteristics of RBBS datasets produced by the Encyclopedia of DNA Elements (ENCODE) project, a public research consortium aimed at identifying functional elements in the human genome [116, 117]. Using these data, researchers were able to evaluate existing DMR detection tools through the iterative permutation of simulation parameters. However, the algorithm developed is limited to 10,000 CpG sites that are assessed by RRBS, a subset of the potentially methylated regions in a whole genome, and further limits its simulation to modeling technical differences rather than actual biological condition.

Several additional tools have a similar focus on generating data for assessing DMR detection performance but vary in their approaches. Whole-Genome Bisulfite Sequencing (WGBS) is a protocol for detecting methylated cytosines in genomic DNA using NGS, and many of these tools focus on simulating data that approximate this type of sequencing output. For example, WGBSSuite, a methylation data simulator created by Rackham et al., uses a pair of dependent Hidden Markov Models to create genome-wide, single-base resolution DNA methylation data for unbiased benchmarking [118]. WGBSSuite improves on the methods developed by Lacey *et al.* [115] by expanding to the whole genome and includes simulation parameters, such as sequencing read coverage and inter-CpG distance

distributions, which may affect the accuracy of methylation detection. WGBSSuite infers these parameters from user-supplied datasets.

Frith and team created DNemulator, a tool for simulating WGBS sequencing reads, to test the specific problem of aligning WGBS reads to a reference genome, a common first step in analyzing NGS data [119]. DNemulator creates methylation values to individual cytosine positions by randomly assigning one of five possible methylation rates to each cytosine on both strands of all chromosomes. DNemulator attempts to incorporate biological and technical variance by introducing simulated common genetic variants (polymorphisms) based on allele frequency into the dataset. DNemulator also introduces random sequencing errors based on the error rates of Illumina sequencing to create simulations that reflect the real-world characteristics of NGS.

Finally, Chung et al. developed the WGBS data simulator pWGBSSimla to simulate methylation quantitative trait loci (meQTLs), genetic variants that influence methylation levels and are associated with gene expression changes [120]. These variants, called methylated quantitative trait loci (meQTLs), as well as allele specific methylation and DMRs are simulated by pWGBSSimla to approximate methylation data that are tissue-type specific. The tool does this by using WGBS profiles derived from 41 datasets of 29 human cell types to simulate both SNP and WGBS data. This approach produces data that represents healthy tissue but is not aimed at modeling disease states or cancer tumor profiles.

There are a limited number of publicly available cancer repositories. In an attempt to meet the specific needs of generating synthetic data that represent cancer-specific methylation profiles, a neural-network based tool, methCancer-Gen, was developed to generate cancer-type specific data using TCGA samples as a source [121]. methCancer-Gen

uses a machine learning technique called a conditional variational autoencoder to generate datasets that are similar to source data. Conditional variational autoencoders are generative models that attempt to recreate data by first compressing, or encoding, an input and then decompressing, or decoding it to an approximation of the input. methCancer-Gen employs this technique to simulate samples that are similar to the input, but which include random variability that ensures the produced samples are not identical.

The authors of methCancer-Gen produced 100 methylation datasets for each cancer type in TCGA and subsequently evaluated the datasets using multiple machine learning classification algorithms. The overall results showed that methCancer-Gen was able to produce data with an average accuracy of .823, which exceeded that of other methods of data generation; However, the results for pancreatic cancer, based on the TCGA-PAAD study, were significantly lower, yielding a classification accuracy of .434. This rate was the lowest for all cancer types tested, with the next highest being .733 for kidney renal clear cell carcinoma (TCGA-KIRC) [121].

In summary, while several methods for simulating genetic data have been developed, there remains a gap in creating data specifically targeted at methylation data for pancreatic cancer in cfDNA. In addition to the lack of any publicly available tools for generating cfDNA, the methods for simulating methylation data are focused on creating data for purposes other than realistic simulation of a particular condition (e.g., pancreatic cancer). The exception is methCancer-Gen, which while showing promising performance in creating simulations based on TCGA data, showed a decreased ability to accurately recreate pancreatic cancer samples.

Heisenberg

To address these issues, I present Heisenberg – a collection of software utilities for the augmentation of methylation sample numbers and simulation of cfDNA datasets that contain a mixture of two samples at expected concentrations. The resulting sample mixtures are intended to represent the case where ctDNA is present as a small fraction of the total cfDNA, Heisenberg includes utility scripts for managing and transforming methylation datasets downloaded from the Gene Expression Omnibus and Sequence Read Archive [122], as well as programs for the simulation of datasets.

"Jitter" is a term used in electronics and telecommunications to describe deviation in a high frequency signal [123]. For example, in computer networking, jitter refers to small delays during data transfers that result in irregular intervals between packets [124]. These variations may be due to network congestion or other signal interference and may be problematic for applications such as online gaming or video-conferencing. Jitter may be generally split into two general classes: random jitter which is unpredictable and deterministic jitter which is reproducible. The combination of the two is the total jitter of a system [123, 124].

Here, I adopt the term jitter to refer to variance that is present in quantitative data like methylation beta values. Real life data often contain random irregularities, referred to as statistical noise. Methylation signals may contain noise due a variety of factors [125, 126]. First, there may be technical sources of variance, such as the differences between lab instruments, procedures or personnel used to prepare samples or reagent manufacturing, all of which may end up affecting data collected from a microarray chip [125]. Another source of variance is due to biological differences between individuals or samples [127]. Actual

deviation between samples can be expected to result in some variability in methylation signals, however genetic differences might also result in measurement error due to the interplay between an individual's DNA and the assay platform itself, leading to variation that is not biologically meaningful [127]. All of these together might be called the jitter in biological data that is assessed through laboratory tests. Approximating this jitter is a key component of realistically simulating sample methylation signals.

Like other tools previously described that use resampling approaches, Heisenberg uses actual samples as input and combines them to create simulated outputs. The software employs novel methods to add noise to generated datasets, with the goal of modeling statistical, technical, and biological sources of variance to create output samples that reflect the inherent characteristics of the input samples while adding non-deterministic error components to avoid the hazards of simple copying and bootstrapping. Heisenberg software is implemented in Python and made available for research use. Heisenberg is available for download at https://github.com/milleneil/heisenberg.

Below, I describe the development and innovation of Heisenberg using a specific use case: the creation of simulated cell-free pancreatic cancer samples through the combination of pancreatic tumor and peripheral blood samples. However, the techniques employed by the software make it applicable to a range of methylation data simulation tasks, such as creating data for different types of cancers or biological states and increasing raw numbers of datasets for use in the development of machine learning models.

Materials and Methods

GEO/SRA samples

Dataset augmentation was performed with pancreatic tumor and normal blood samples to increase the numbers of input samples available for training a classification model. TCGA-PAAD samples served as the exclusive source for pancreatic tumor samples; normal blood samples were identified through a search of publicly available studies at the NCBI Gene Expression Omnibus (GEO).

A comprehensive search for studies conducted using the Illumina Infinium Human Methylation 450K BeadChip yielded 1,444 studies. These were manually reviewed to identify studies using cancer-free, peripheral blood where data were available for download. Manual review produced a list of 16 projects containing 4148 subjects (Table 7).

Table 7.-- Initial List of GEO studies with normal blood methylation data

Accession	Title	N
GSE105018	*Whole blood DNA methylation profiles in participants of the Environmental Risk (E-Risk) Longitudinal Twin Study at age 18*	1658
GSE109905	*Genome-wide DNA methylation analysis in autism spectrum disorders (ASD patients vs Controls)*	69
GSE32148	*DNA methylation in peripheral blood from individuals with Crohn's disease or ulcerative colitis and normal controls*	48
GSE110043	*Epigenome analysis of alcohol consumption in whole blood samples*	94
GSE36369	*DNA methylation contributes to natural human variation*	309
GSE42861	*Differential DNA methylation in Rheumatoid arthritis*	689
GSE59489	*DNA methylation modifications associated with Chronic Fatigue Syndrome*	24
GSE62003	*Blood methylomic signatures of pre-symptomatic dementia in elderly subjects with Type 2 Diabetes Mellitus*	70
GSE63499	*DNA Methylation Changes in Whole Blood and CD16+ Neutrophils in Response to Chronic Folic Acid Supplementation in Women of Childbearing Age*	60
GSE64495	*DNA methylation profiles of human blood samples from a severe developmental disorder and controls*	113
GSE72774	*DNA methylation profiles of human blood samples from Caucasian subjects with Parkinson's Disease*	508
GSE72776	*DNA methylation profiles of human blood samples from Hispanic subjects with Parkinson's disease*	84
GSE89218	*Characterization of Whole Genome DNA Methylation Profile Associated with Post-Traumatic Stress Disorder in OIF/OEF Veterans*	163
GSE77056	*Genome-wide DNA methylation profile in the peripheral blood of cocaine and crack dependents*	47
GSE67393	*Sex differences of leukocytes DNA methylation*	117
GSE41169	*Blood DNA methylation profiles in a Dutch population*	95

Demographic pairing and identification of age groups

Sample augmentation was performed by combining existing samples to produce simulated ones. To maximize the validity of the combinations, samples were segregated into groups defined by sex and age to account for variance in methylation levels that has been specifically attributed to those factors [128]. To do this, an additional review of the list of peripheral blood methylation studies was carried out to inventory the superset of samples and determine the extent of demographic information.

Using the clinical data available for TCGA-PAAD subjects as a baseline, studies were evaluated to determine the presence or absence of sample metadata including subject age, sex, and ethnicity. Further, an audit for completeness of methylation data from each subject was conducted to account for variation observed between studies, as some projects deposited methylation values for *all* probes on the Illumina Infinium 450k array, while others filtered data according to QC cutoffs, leaving only a partial set available for download.

Sex was recorded for all selected studies with age reported for 8 of the 16 initial studies. Ethnicity was recorded for only 4 studies. Other factors such as control vs. affected status and disease state were highly variable and not consistently reported across studies.

Sex and age were therefore selected as the demographic factors most useful for minimizing variance and maximizing the number of samples for downstream analyses. After review, 7 studies (836 subjects) were excluded due to incomplete demographics where both sex and age were not available. An additional 14 samples from 2 included studies (GSE32148 and GSE62003) were excluded due to incomplete metadata. A large cohort of 1658 samples from subjects at 18 years of age was excluded due to being unrepresented in the TCGA-PAAD dataset and therefore not relevant for use in constructing simulated cfDNA datasets for pancreatic cancer.

The distribution of ages in the remaining samples was evaluated in comparison to TCGA-PAAD samples to define age ranges for grouping samples. Initial age groups listed by the National Cancer Institute of 0-19, 20-34, 35-44, 45-54, 55-64, 65-74, 75-84 and > 84 were refined to 0-34, 35-54, 55-64, 65-74 and 75+. An additional 259 samples under the age of 35 were removed to match the minimum age of TCGA-PAAD samples leaving a final set of 1381 normal blood samples from which simulated samples were generated (Figure 3).

	GEO			TCGA		
age	male	female	total	male	female	total
0-35	137	122	259	0	0	0
35-54	151	307	458	20	15	35
55-64	134	214	348	20	29	49
65-74	178	167	345	40	22	62
75+	133	97	230	22	16	38

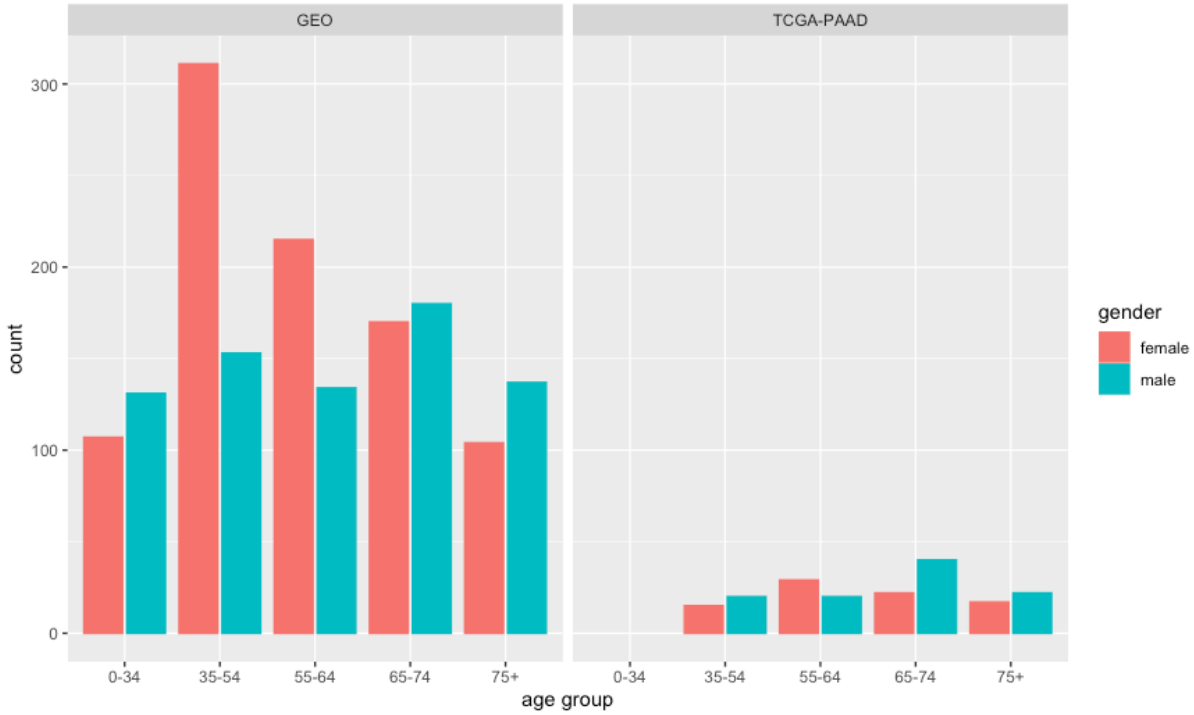

Figure 3.-- Sample distribution by sex and age group in the GEO and TCGA datasets

Results

Using curated source samples from GEO and TCGA, Heisenberg was developed to incorporate methods for creating new samples from combinations of source samples while adding sources of noise. Various methods for applying noise components were evaluated and data sets were generated for use in the development of classification models for detection of pancreatic cancer in cell-free DNA.

Normal and Tumor Sample augmentation – basic combination

Normal blood and pancreatic tumor samples were augmented by combining methylation values from individuals matched by age and sex to create simulated, unique individuals that are presumed to share the underlying biological state of the samples as represented through the methylation signals of each component sample while differing enough to be more than simple copies. Synthetic methylation beta values for each individual probe (β_S) were created using the mean beta values for the probe site across multiple actual samples (β_μ). The mean was then adjusted randomly in three distinct ways to approximate individual-to-individual variation (Figure 4).

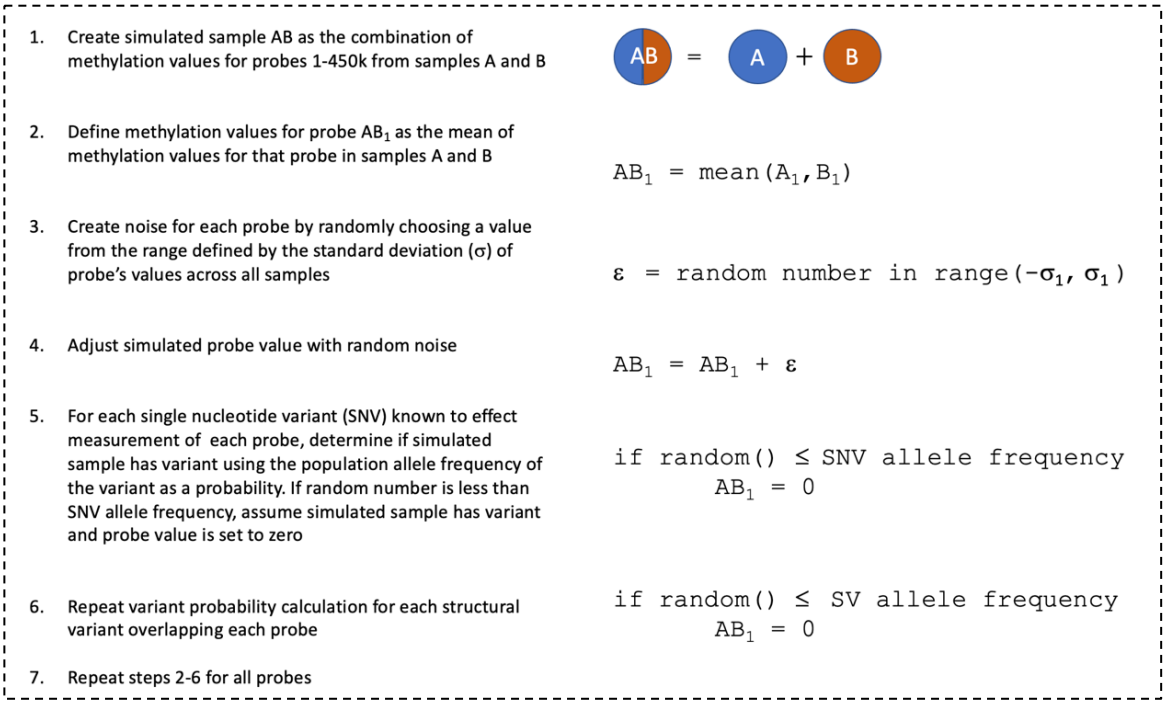

Figure 4.-- Steps to create methylation values from source samples

Error components were applied at random, meaning that theoretically an infinite number of unique datasets could be created with this method; however, the simulation was limited to creating a single synthetic individual for each unique combination of source samples. Simulations were performed using unique combinations of two samples from each

group of age- and sex- matched source samples. Normal blood samples were increased from the 1,381 source samples to 134,427 (male = 45,363, female = 89,064) individuals while tumor samples were increased from 173 to 2,478 (male = 1,534, female = 944) (Table 8).

Table 8.--Counts of simulated samples representing every possible combination of two samples

	Blood		**Tumor**	
age	*male*	*female*	*male*	*female*
35-54	11476	47278	210	120
55-64	9045	23005	210	435
65-74	15931	14028	861	253
75+	8911	4753	253	136
	total: 134,427		total: 2,478	

This technique of creating a new sample by merging traits from other individuals can be easily extended by increasing the number of samples in each combination. This expansion raises the number of possible combination and therefore the final number of samples that can be created. For example, using the demographic groups defined above while increasing the number of samples to 3 would create a simulated sample set of 9,825,258 individuals (male 2,352,171, female = 7,473,087) (Figure 5).

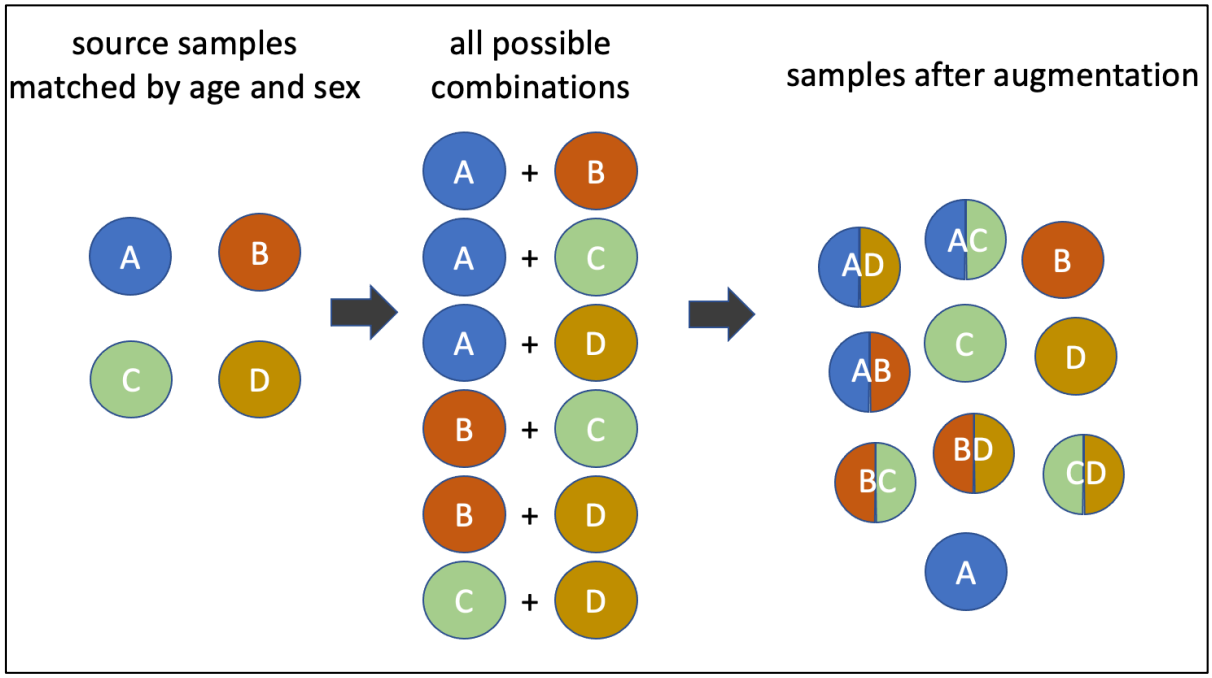

Figure 5.-- Overview of sample augmentation using combinations of two samples

Statistical jitter

An error component reflecting the observed variation across samples was created and incorporated into the simulation. Descriptive statistics were calculated for each probe separately for the normal blood and tumor source datasets. The standard deviation for each probe was then used as input to a random error function to add a small amount of noise to each simulated probe value.

The standard deviation of all individual probe values ranged from 0.001 to 0.423 for tumor samples and 0.006 to 0.346 for normal blood (Figure 6). Noise was calculated by randomly choosing a value from a uniform distribution between the negative and positive

values of the standard deviation (σ), or ε = rand(-σ, σ). The mean probe value was then adjusted by adding the randomly selected noise component.

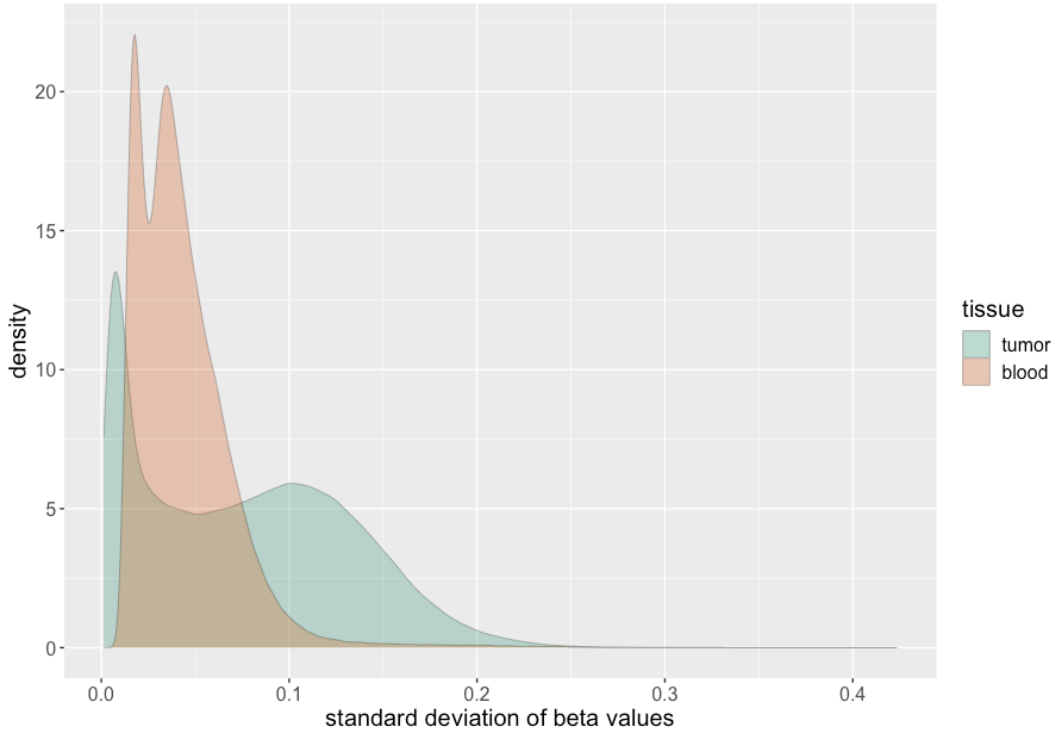

Figure 6.--Density plot of standard deviation of beta values for all probes in normal blood and tumor samples

Technical jitter – Confounding SNPs

A second error component was incorporated into the simulation formula to capture the interaction of subject genetic variability with technical performance of the Illumina Infinium HumanMethylation450 BeadChip array. Previous studies have shown that the occurrence of single nucleotide polymorphisms (SNPs) and short insertions and deletions (in/dels) in the probe target regions of an individual may confound the array's ability to accurate accurately measure methylation at that site due to the variants' effect on the probe hybridization process [127, 129].

The result is that beta values in individuals who carry select genetic variants may be noisy or non-existent for probes that overlap these variants. An additional complication of this is the difficulty in distinguishing between a true methylation value of 0 and a missing value due to the presence of a cofounding variant.

Using a set of SNPs and in/dels which affect methylation results as published by Illumina [86], the theoretical presence or absence of confounding small variants was included as a step in the simulation. The incorporation of these confounding variants was done according to population allele frequencies of the variants in order to add additional random variation based on biological factors to the simulated samples. The Illumina SNP and in/del dataset consists of array probe identifiers along with one or more nucleotide variants found to affect results in the Infinium HumanMethylation450 BeadChip array. Genetic variants are identified by their identifiers (rsIDs) in the NCBI dbSNP database [130]. The dataset also contains the observed minor allele frequency for each variant in the general population (MAF) as well as the distance from the variant position to the methylation probe site where beta values are measured (Figure 7). A total of 459, 774 variants in 273,661 probes were recorded for an average of 1.68 variants / probe.

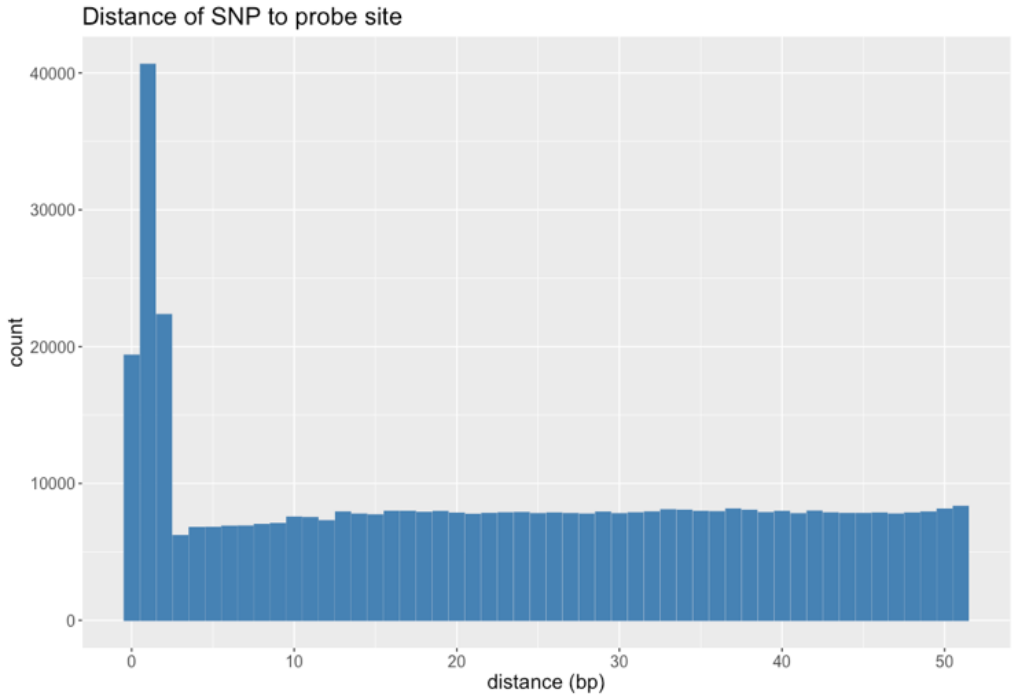

Figure 7.-- Histogram of distances between small variants and methylation probe sites

Presence or absence of confounding small variants in simulated individuals was modeled by treating each variant's MAF as the probability that the given variant would be present in the subject. Variant probabilities are assumed to be independent since the rate of variant co-occurrence is not known. As each probe value is simulated, a random number is chosen for each variant found to affect that probe; a random number less than or equal to the variant's MAF is used to determine that the individual has the variant in question. The effect of having the theoretical variant is to set the corresponding probe beta value to zero, reflecting the absence of signal that would be recorded by the assay.

Biological jitter – Structural variants

Structural variants are thought to affect a larger fraction of the genome than small nucleotide variants [131]. Structural variants include deletions and insertions of more than 50 bp as well as more complex events such as inversions, where a segment of a chromosome is

reversed, and balanced translocations, where segments of different chromosomes have exchanged places and segmental duplications [132]. These types of variants might affect actual methylation by disrupting chromosomal structure but also might interfere with the measurement of methylation values through their interaction with the array probes used for the quantification. Because of this, Heisenberg includes a third error component to account for the presence of large structural variants in an individual.

The Genome Aggregation Database (gnomAD) was constructed from 14,891 genomes across a diverse population and contains 387,478 structural variants [133], along with their observed global allele frequency and observed frequency as homozygous or heterozygous variants. To mimic the potential occurrence of structural variants, version 2.1 of the gnomAD structural variant reference was downloaded and included in the simulation process. Overlap between gnomAD variants and methylation probe sites was calculated using each feature's coordinates on the GRCh37 human reference genome sequence to determine which probes would be affected by the occurrence of each structural variant. Genomic coordinates for Illumina methylation probes were determined using the R Bioconductor IlluminaHumanMethylation450kprobe library, which translates positions originally published in human genome reference build 36 coordinates to GRCh37.

While gnomAD catalogs a variety of types of structural variants, the effect of many variant types on methylation array results is not clear. For example, while a balanced translocation may contain a methylation probe site, the concrete effect of it on the observed beta signal is uncertain. Similarly, the effect of each variant's zygosity on methylation probe values cannot be definitively predicted.

For this reason, only large, homozygous deletions were incorporated into the simulation algorithm, as the effect of such an event can be safely assumed to negate the methylation value that would be observed at the deleted location (Figure 8). After filtering, the final supplemental data set used in simulation consisted of 22,107 deletion variants ranging in size from 52 to 6,454,328 bp (mean=34,592, median=7,039) and overlapping 94,292 probes (average 5.6 probes / deletion).

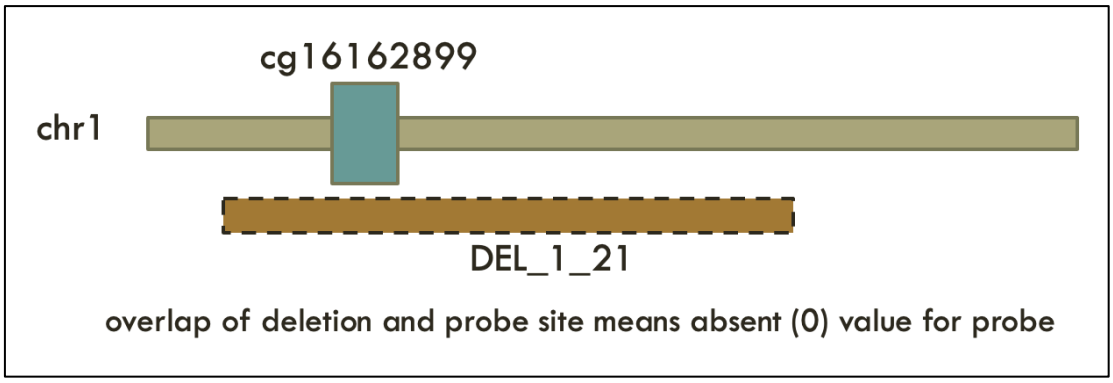

Figure 8.—Visualization of array probe loss due to homozygous deletion

The occurrence of deletion variants was included in the simulation in a manner like the incorporation of confounding small nucleotide variants. The allele frequency for each deletion as a homozygous event was used as the independent probability that the simulated individual carried the deletion. Determination was made through the generation of a pseudo-random number. If a deletion was randomly selected, the beta values recording methylation values for all probes overlapped by the variant were set to zero.

Simulation and Validation

Demographically matched sample sets were used for all simulations. Multiple simulations were carried out for both normal and pancreatic tumor samples using different configurations for the statistical, confounding SNP and structural variant-based error

components. Iterations were performed with no error adjustment (mean of input samples only), different cutoffs for small variant and structural variant allele frequency, and by including only probes within a variable maximum distance from probe site. Allele frequency cutoffs were implemented to differentiate between random inclusion of all variation and rare variants only. Rare variants were defined as small variants with an allele frequency cutoff of 0.2 and structural variants with a maximum homozygous allele frequency of 0.3. Maximum distance from probe sites to known confounding variants was adjusted to include only variants within 2bp, 4bp, 10bp and all distances (max 51bp).

Simulations were initially performed for all 485,512 probes for but were eventually limited to include only the 23,688 probes found to be differentially methylated between normal pancreas and pancreatic tumor. This reduction greatly minimized the computational requirements for generating simulated data. Generation of these data was done in a Linux high performance computing environment, with each iteration being done in a single thread on a compute node with 32 GB of RAM.

The effect of the error components, including the different settings for SNP and structural variant MAF cutoff and SNP proximity, was quantified by calculating the number of adjustments made to the simulated set. Metrics were gathered for the number of SNPs and structural variants that were randomly included in the simulated set, as well the number of probe values that were adjusted per simulated sample (Table 9).

Table 9.--Simulated datasets and parameterization

set	filter max SNP MAF	filter max SV MAF	max SNP distance	stdev	SNPs / sample	SNP probes / sample	SVs / sample	SV probes /sample
Blood								
no-err	n/a	n/a	n/a	N	0	0	0	0
common snp - all	1	1	51	Y	753.2	767.0	14.3	16.3
common snp – 2bp	1	1	2	Y	50.1	50.1	14.3	16.3
common snp – 4bp	1	1	4	Y	62.1	62.1	14.3	16.3
common snp – 10bp	1	1	10	Y	93.9	94.4	14.1	16.1
rare snp – all	.2	.3	51	Y	186.5	187.8	11.1	12.8
rare snp – 2bp	.2	.3	2	Y	7.0	7.0	11.1	12.8
rare snp - 4bp	.2	.3	4	Y	8.1	8.1	11.1	12.8
rare snp – 10bp	.2	.3	10	Y	13.0	13.0	11.1	12.8
Tumor								
no-err	n/a	n/a	n/a	N	0	0	0	0
common snp - all	1	1	51	Y	857.4	874.3	22.1	27.9
common snp – 2bp	1	1	2	Y	58.6	58.6	22.2	28
common snp – 4bp	1	1	4	Y	73.9	73.9	22.1	27.8
common snp – 10bp	1	1	10	Y	122.6	123.1	22.2	27.9
rare snp – all	.2	.3	51	Y	203.2	204.6	18.8	24.3
rare snp – 2bp	.2	.3	2	Y	7.6	7.6	18.9	24.3
rare snp - 4bp	.2	.3	4	Y	9.0	9.0	18.8	24.3
rare snp – 10bp	.2	.3	10	Y	15.7	15.7	18.9	24.3

Validation of the simulated datasets was performed using the neural network classifier previously developed to distinguish between pancreatic tumor and normal pancreas and blood. Validation metrics for all simulated sets were calculated as the percentage of samples identified correctly by the neural network classifier. Classification was performed on source blood and tumor sets as a control. Classification accuracy of source blood samples was 99.70%. Mean prediction accuracy for simulated sets was 99.93% (min=99.89%, max=99.96%, median=99.94%, SD=0.0002), with the highest accuracy occurring in the dataset that used all variants and included no allele frequency or distance cutoff (common

SNP – all). Classification accuracy of source tumor samples was 96.22%. Mean prediction accuracy in simulated sets was 99.40% (min=99.19%, max=99.64%, median=99.44%, SD=0.0015), with the highest accuracy in the dataset that used all variants and included a maximum distance from methylation probe site of 2bp (Table 10).

Table 10.--Classification accuracy of simulated samples

tissue	set	predicted correctly	predicted incorrectly	accuracy
blood	source	1649	5	0.9970
blood	no-err	134344	83	0.9994
blood	common SNP - all	134376	51	0.9996
blood	common SNP – 2bp	115189	123	0.9989
blood	common SNP – 4bp	134306	121	0.9991
blood	common SNP – 10bp	134327	100	0.9993
blood	rare SNP – all	134324	103	0.9992
blood	rare SNP – 2bp	134341	86	0.9994
blood	rare SNP - 4bp	134347	80	0.9994
blood	rare SNP – 10bp	134347	80	0.9994
tumor	source	178	7	0.9622
tumor	no-err	2458	20	0.9919
tumor	common SNP - all	2462	16	0.9935
tumor	common SNP – 2bp	2469	9	0.9964
tumor	common SNP – 4bp	2464	14	0.9944
tumor	common SNP – 10bp	2462	16	0.9935
tumor	rare SNP – all	2467	11	0.9956
tumor	rare SNP – 2bp	2464	14	0.9944
tumor	rare SNP - 4bp	2458	20	0.9919
tumor	rare SNP – 10bp	2464	14	0.9944

Cell-free DNA sample simulation

Heisenberg implements a model for simulation of cfDNA methylation datasets by combining source blood samples with secondary tissue samples. Simulations can optionally include demographic matching of blood and secondary tissue samples by age group and sex as well as the multi-factor error component consisting of random statistical noise, variation due to confounding SNPs and variation due to structural variation.

Heisenberg simulates cell-free samples by combining blood and tumor methylation signals. after adjusting the methylation values for each probe according to the expected concentration in the final cell-free sample. The expected fraction of secondary tissue in the primary blood is provided as a parameter to Heisenberg. The simulated methylation value at each probe site (β_A) is calculated by first adjusting the methylation signal from normal tissue (β_N) by the expected fraction of normal DNA (F_N), altering the methylation signal from secondary tissue DNA (β) according to its expected fraction (F_T) and then combining the two values along with the multifactorial error component (ε):

$$\beta_A = (F_N * \beta_N) + (F_T * \beta_T) + \varepsilon$$

Simulation of cell-free pancreatic cancer samples

Synthetic cell-free DNA sample sets were created using the 185 TCGA-PAAD pancreatic tumor samples and the 1381 normal peripheral blood samples obtained from the SRA. Simulation was performed using demographic matching by age and sex and all small structural variants without filtering by allele frequency or proximity to the probe site. Simulated samples were created to simulate ctDNA fractions of 10%, 5%, 1% and .1% using every unique combination of matched blood and tumor sample in each demographic group. A total of 31,783 sample sets (male=15,746, female=16,037) were created for each tumor fraction, yielding a total of 127,132 samples (male=62,984, female=64,148) (Table 11).

Table 11.-- Overview of simulated pancreatic cancer cfDNA sets created with Heisenberg

age group	males	females	total
35-54	3020	4605	7625
55-64	2680	6206	8886
65-74	7120	3674	10794
75+	2926	1552	4478
total per fraction	15746	16037	31783

Heisenberg modules

The Heisenberg software program provides different functionality grouped into modules to simplify execution. Each module has its own specific command line arguments and outputs (Table 12).

Table 12.-- Heisenberg modules and description

module	description
simulate	augment samples through combination
mix	simulate cell-free DNA mixtures from two sample types
stats	gather descriptive statistics for each probe in preparation for simulation
extract_sra_probe	extract probe values from SRA series_matrix.txt files
invert	turn file with probes as rows into wide file with probes as columns as needed by Heisenberg
extract_sra_meta	print metadata for sample characteristics in SRA series_matrix.txt.gz
subset	extract subset of sample rows and/or probe columns given list
combine	safely combine multiple wide files into one, ensure that probe vals are uniform and that minimum set of required probes is included
cell	print value of master file cell[x,y]

Discussion

In summary, a novel method for simulating methylation data and cell-free DNA mixtures was developed to serve the data generation needs of the project. These methods have been implemented in the Heisenberg software and made available for community research use. Heisenberg provides a scalable, resampling-based method for increasing sample numbers while incorporating a multi-factor error model to add variation to the newly created samples. The error component integrates statistical noise and population-level genetic variation to incorporate both technical and biological variance into the samples created.

While the current methods of Heisenberg are primarily applicable to the Illumina microarray platform, the principles could be extended to sequencing based methods if the data were available. The technical variance introduced by the simulation of confounding SNPs would not have a direct sequencing equivalent, however the modeling of structural variation would still have value. Further, techniques such as surveying methylation sites across a phenotypically similar sample set in order to identify the standard deviation of methylation signals could be readily performed and the results supplied to the software as a parameter to the simulation.

An unexplored, potential limitation of the simulation model is the assumption that cfDNA fragments would be evenly distributed both across probe sites and between cfDNA and ctDNA. The simulation of ctDNA data does not take into account the possibility of dropout or overrepresentation at specific probe sites. Rather, it is assumed that all methylation sites have an equal chance of being present in cfDNA and that coverage of these sites by the tumor and blood fractions will be consistent and balanced. Investigation into these assumptions using real cfDNA samples, ideally from pancreatic cancer, may yield some insight as to their validity. If systematic biases or uneven representation can be identified, the simulation model might be improved by incorporating additional random dropout of probe signals or signal imbalance into the error model.

CHAPTER 4

EVALUATION OF CLASSIFICATION MODEL USING SIMULATED DATASETS

Overview

The analytical performance of computational models must be assessed at multiple points during development. In the initial phases, models are regularly tested using training and test subsets [134]. These train/test cycles are rapid and provide immediate feedback that can be used to guide development including, choice of algorithm and parameterization of the model [109, 110]. After an initial classification model has been established and refined, more extensive testing must be carried out in multiple stages with increasing rigor before the model can be deployed to a production setting [69, 110]. For a translational research application that may be used in a medical context, testing and validation should be carried out using datasets of increasing value and size [111]. Development of a computational model intended for use with data that are rare or costly to acquire will often start with simulated data created *in silico* to facilitate the rapid train/test development cycle [134, 135]. Models that show promise may then be applied to actual data and finally validated prospectively at scale in a clinical trial [69].

To assess the feasibility of using a neural network model for early detection of pancreatic cancer in cell-free DNA using methylation signals, a testing exercise was performed using methods and datasets previously described. This exercise provided an evaluation of classification performance at multiple different cell-free fractions. Taken together, these steps of identifying relevant markers from publicly available data at from TCGA, simulating cell-free datasets, creating a classification model and assessing its

performance describe a development template that can be used for a range of potential applications in the future.

Methods

Creation of simulated datasets

Simulated cell-free DNA methylation datasets were created using Heisenberg as previously described. Sample sets representing simulated ctDNA fractions of 10%, 5%, 1% and .1% were created with 31,783 samples each for a total of 127,132 synthetic samples (male=62,984, female=64,148). Normal blood samples (N=1,381) were augmented using Heisenberg to create 134,427 simulated samples (male = 45,363, female = 89,064). Both ctDNA and normal samples were created using the compound error component including statistical noise, occurrence of common structural variants and small nucleotide variants at any distance from the probe site (SNP common all). These data were used to train and test classification models through ten-fold cross-validation.

Novel neural network classification model

A novel neural network classification model was created using Keras and TensorFlow 2.6.0. The model was created as a multi-layer perceptron with an input layer, two feed forward hidden layers and an output layer with two classes (tumor/normal). The rectified linear unit (ReLU) activation function was used for the input and hidden layers. A dropout layer was added with a dropout rate of .2. The Adam optimization algorithm was used with a learning rate of 0.001 along with a sparse categorical cross-entropy loss function. The output classification layer was created using the softmax activation function for calculating probabilities for each class. The model was trained using a validation split of 80/20 with early stopping when model loss did not improve by at least 1e-2 over 20 epochs. The

classification model was trained using the probe set previously defined and containing the top 200 probes as ranked by their Gini importance score and using only stage i and stage ii samples (top-200.stage-i-ii.probes).

Evaluation modes

Model accuracy was initially measured using two training modes to assess classification performance at different ctDNA fractions. Ten iterations of training and testing were performed for each mode. In the 'per concentration' mode, the classification model was trained and tested on a single ctDNA fraction. For example, performance at 1% ctDNA was assessed by training a model using only simulated 1% ctDNA samples and then tested on a set of samples with the same concentration. In the 'all concentrations' mode, the classification model was trained using a set of samples containing all ctDNA fractions and then tested on a single fraction to obtain metrics on performance at that level. In this mode, performance at 1% ctDNA was measured by training a model with samples at 10%, 5%, 1% and .1% and then testing on a 1% test set alone.

In addition to gathering overall performance metrics, the two training modes were used to explore the benefits of the two differing strategies. Theoretically, a 'per concentration' model might offer superior performance for detection at that specific level which might enable an estimate of tumor concentration, and therefore disease progression, in addition to the binary tumor/normal classification. The final design of a model to detect cancer at all ctDNA fractions could then be achieved by combining multiple individual models, each trained for a specific fraction, into an ensemble detection program. Conversely, the 'all concentrations' mode might offer superior performance for detection at any level by minimizing potential model overfitting through training on all concentrations together. In this

mode, a more flexible model that would be sensitive to all fractions might be enabled and strengthened by allowing model training to use a greater number of samples, since all available samples could be used.

Results

Ten iterations were run for 'per concentration' and 'all concentration' modes for each simulated ctDNA fraction for a total of 80 runs. True positives, false positives, true negatives, and false negatives were recorded for each iteration and used to calculate performance metrics for each model at each ctDNA fraction (Figure 9).

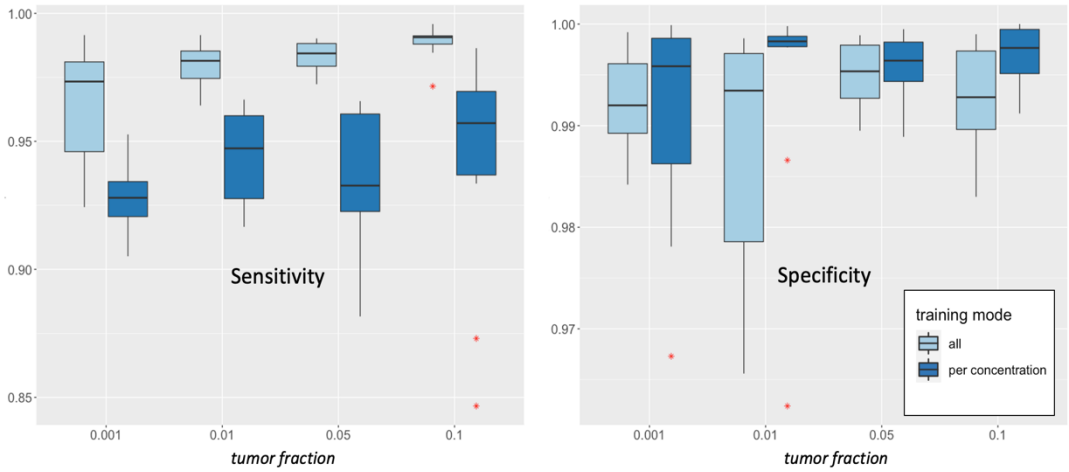

Figure 9.--Sensitivity and specificity of two different model training modes

The 'per concentration' models showed greater specificity while the 'all concentrations' model consistently showed greater sensitivity at all ctDNA fractions. Median sensitivity of the 'per concentration' model ranged from .928 to .957 for ctDNA fractions of .1% to 10% respectively while specificity ranged from .996 to .998. The 'all concentrations' model showed a median sensitivity ranging from .973 to .991 with a specificity range of .992 to .993 (Table 13).

Table 13.--Median performance metrics of classification models at four ctDNA fractions using two model training modes

ctDNA fraction	sensitivity	specificity	PPV	NPV	accuracy	F1
per concentration model						
0.001	0.928	0.996	0.982	0.983	0.982	0.953
0.01	0.947	0.998	0.992	0.988	0.986	0.962
0.05	0.933	0.996	0.984	0.984	0.984	0.956
0.1	0.957	0.998	0.990	0.990	0.989	0.970
all concentrations model						
0.001	0.973	0.992	0.965	0.994	0.987	0.967
0.01	0.981	0.993	0.973	0.996	0.990	0.975
0.05	0.984	0.995	0.980	0.996	0.993	0.983
0.1	0.991	0.993	0.970	0.998	0.992	0.978

Single test evaluation

While the 'per concentration' results show the performance of a model trained at a specific ctDNA fraction with a test set of that same fraction, the evaluation does not show how that same model might perform against other fractions. This is relevant since an actual test would have no way of determining *a priori* the actual tumor content in the sample. For a single model to be used as the test for all samples, it must show adequate performance across all fractions. Similarly, if a 'per-concentration' model performs well at its own native concentration while performing badly at others, it lends strength to the idea of an ensemble classifier that could give a binary response of tumor vs normal, as well as an estimate of the ctDNA fraction present in the sample. To examine these questions, a final evaluation was performed to assess the ability of a model trained at a single concentration to detect cancer at all other fractions. As an example, a model trained only on .1% ctDNA was then tested on .1%, 1%, 5% and 10% to determine which single model might offer the best performance regardless of the true concentration. These results were compared to the 'all concentrations

model' where a single model trained with all ctDNA fractions was used to classify all test samples.

As before, 10 iterations of the training/test cycle were performed for each model type for a total of 200 runs. True positives, false positives, true negatives, and false negatives were again recorded for each iteration and used to calculate performance metrics for each model at each ctDNA fraction (Figure 10).

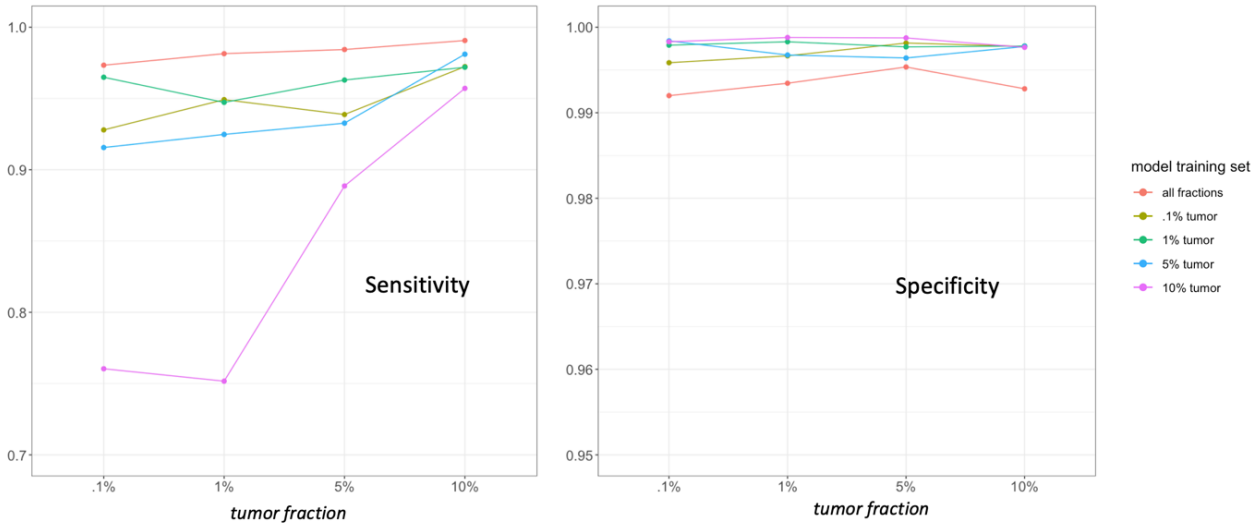

Figure 10.-- Median classification sensitivity and specificity of model training sets against all ctDNA

The model trained with all concentrations again showed the highest sensitivity of detection for all ctDNA fractions while the per-concentration models showed greater specificity. Median sensitivity of the 'all concentrations' model for all tests was .964 and median specificity of was .997. The best performing 'per concentration' model was the classifier trained with 1% ctDNA with a median sensitivity of .962 and median specificity of .998. The 10% ctDNA model was both the least sensitive, with a median sensitivity of .837, and the most specific, with a median specificity of .999 (Table 14).

Table 14.--Median test performance for models across all ctDNA fractions

training set	sensitivity	specificity	PPV	NPV	accuracy	F1
.1%	.945	.997	.987	.987	.987	.967
1%	.962	.998	.992	.991	.989	.971
5%	.933	.998	.989	.984	.984	.956
10%	.837	.999	.992	.963	.960	.891
all	.964	.997	.983	.992	.983	.958

Discussion

Contrary to expectation, models trained specifically for the ctDNA fraction they were tested on were less sensitive than models trained on all fractions, both for their target fractions as well as for others. This performance could be due to the greater number of training samples available for the 'all concentrations' model, but also could be that the greater variability in the samples leads to a more flexible model that was more easily able to classify a range of input signals. Concurrently, specificity was overall superior for the 'per concentration' models indicating that fewer false positive results might be expected with these models. Again, this may be due to the restricted nature of the training set since a model trained with a smaller number of representative samples will have a smaller range of inputs that it will call true, leading it to be more conservative in its calls and therefore yielding fewer false positive results.

While the 'all concentrations' model provides the best sensitivity at all levels, the decreased specificity and therefore increased number of false positives might present problems in a clinical setting. However, while specificity is lower for this model, it is not a drastic difference (median .997 vs. .999 for the best performing 'per concentration' model) meaning that the practical results would be very similar. Further, it could be argued that for

an early detection model, sensitivity might be the more important factor since a positive result from a screening test would rarely serve as the primary diagnostic method of pancreatic cancer and would not on its own inform treatment. Rather, a positive test result would be an indication for confirmatory testing with an orthogonal, more invasive method. While false positive results could be managed in this way, there is no similar avenue for addressing false negatives. In this case, a negative result would receive no follow up with the end result being a failure to diagnose and a lost opportunity for early treatment. Given the relatively small difference in specificity compared to the large gap in sensitivity, it seems clear that the broader model would represent the best single approach for further study.

The utility of an ensemble approach, in which a sample would be tested against multiple individual models trained at specific ctDNA fractions and then receive a classification based on the outputs of all models, is uncertain. While the performance of the individual ctDNA fraction models lagged in sensitivity, it has not been tested whether a consensus approach might yield superior results to the single, 'all concentrations' model. It is also not clear how reliably these individual models might predict the overall ctDNA fraction. However, the ability to estimate the tumor fraction is a distinctly secondary application behind the ability of the model to detect the presence of pancreatic cancer. There would be no change during action taken by a clinician in response to a 1% or a 5% ctDNA fraction estimate from a screening test – in either case, the next step would be confirmatory testing with the subsequent determination of tumor stage and level of advancement or metastasis determine through traditional means.

The results, however, demonstrate the feasibility of an early detection method for pancreatic cancer based on methylation in cell-free DNA. Detection sensitivity and

specificity at even .1% tumor fraction is sufficiently high to merit further development and investigation. The study is limited by the use of simulated data, however the fact that differential methylation probes have been determined from real pancreatic tumor, pancreas and peripheral blood samples lends weight to their validity as predictors. A logical next step for this work would to be to evaluate the performance of the conceptual model presented here using actual samples from patients known to have pancreatic cancer.

The performance evaluation of the various models has some limitations beyond the use of simulated data. A more rigorous comparison would be to account for a classification confidence threshold that could be applied to each model. The models evaluated here report the highest scoring class prediction produced by the classifier regardless of the confidence. Utilization of confidence cutoff, below which no class would be assigned, would introduce a concept of a 'no-call' or indeterminate result where currently calls are strictly tumor or normal. A more fine-grained approach might allow for evaluating and tuning of each model to find the specific threshold at each ctDNA fraction that provides the best balance between true positives and false positives. An evaluation of this kind, which would enable evaluation of each model by calculating the Area Under the ROC Curve (AUC) could provide a more nuanced view of the strengths and weaknesses of the different model creation strategies. Additionally, the evaluation here would benefit by an increase in statistical power by adding samples.

A third limitation is the use of microarrays to identify methylation sites that have predictive value. Microarray data are inherently limited to the specific sites targeted by an array design. The discovery of predictive features could benefit from a genome-wide survey with whole genome bisulfite sequencing or the emerging 5-base sequencing from PacBio

which would include all regions of the genome including those not expected to be relevant at the time a particular microarray was designed. It is therefore possible that more informative features could be identified if whole genome methylation data were available for a number of pancreatic cancer samples. Future work based on the findings presented here could include a wider search for better predictive features as well as an investigation into determining sequence-based genomic equivalents for the probes identified from microarrays to assess whether a sequencing approach offers improvement in detection.

While the reduced view of the methylome that a microarray platform provides may be a limitation in the initial definition of predictive features, the significant efficiency and cost advantages of the platform make it attractive as a method for implementation as a high-volume screening method. Once features have been defined, such as those presented in the candidate probe sets described here, an optimized assay can be designed and implemented in a manner that could offer significant throughput at a low-cost. Further, the computational requirements for data acquisition and analysis are greatly reduced for the microarray platform in comparison to whole genome sequencing, making the platform an attractive technology choice for a range of lab settings where specialized computing hardware and facilities may not be available.

Finally, this study focused on pancreatic cancer as the primary use case. The methods for feature identification, data simulation and automated classification that are described here present a template for development that could apply to additional cancer types. Additional research to extend these methods using other data freely available at resources like the TCGA may offer an opportunity to address the significant health care burden that cancers together represent.

APPENDIX

Table A1.-- Genes associated to probes differentially methylated between normal pancreas, blood and

Symbol	GeneID	full_name	top 50	top 100	top 200	top 500	Gini > .003
AACS	65985	acetoacetyl-CoA synthetase	X	X	X	X	.
AATK	9625	apoptosis associated tyrosine kinase	X	X	X	X	X
ACAN	176	aggrecan	.	.	.	X	.
ACOT7	11332	acyl-CoA thioesterase 7	X	X	X	X	X
ACSM6	142827	acyl-CoA synthetase medium chain family member 6	.	.	X	X	.
ACTA1	58	actin alpha 1, skeletal muscle	.	.	X	X	.
ADAM12	8038	ADAM metallopeptidase domain 12	.	.	X	X	.
ADAP1	11033	ArfGAP with dual PH domains 1	.	X	X	X	.
ADGRD1	283383	adhesion G protein-coupled receptor D1	.	X	X	X	.
ADTRP	84830	androgen dependent TFPI regulating protein	.	.	X	X	.
AFG3L1P	172	AFG3 like matrix AAA peptidase subunit 1, pseudogene	.	X	X	X	.
AGO2	27161	argonaute RISC catalytic component 2	.	.	X	X	.
AHDC1	27245	AT-hook DNA binding motif containing 1	.	.	.	X	.
AIFM2	84883	apoptosis inducing factor mitochondria associated 2	.	.	.	X	.
AJAP1	55966	adherens junctions associated protein 1	.	.	X	X	.
AKR1E2	83592	aldo-keto reductase family 1 member E2	.	.	X	X	.
ALDH3A1	218	aldehyde dehydrogenase 3 family member A1	X	X	X	X	.
ANKRD36BP2	645784	ankyrin repeat domain 36B pseudogene 2	.	.	.	X	.
AP3B2	8120	adaptor related protein complex 3 subunit beta 2	.	.	.	X	.
APOD	347	apolipoprotein D	.	.	X	X	.
AREG	374	amphiregulin	X	X	X	X	.
ARHGEF28	64283	Rho guanine nucleotide exchange factor 28	.	.	X	X	.
ARRDC2	27106	arrestin domain containing 2	.	.	X	X	.
ASCL1	429	achaete-scute family bHLH transcription factor 1	.	.	X	X	.
ATF6B	1388	activating transcription factor 6 beta	.	.	.	X	.
ATXN1	6310	ataxin 1	.	.	X	X	.
B3GNT3	10331	UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase 3	.	.	X	X	.

Table A1.--Continued.

symbol	GeneID	full_name	top 50	top 100	top 200	top 500	Gini > .003
B4GALNT1	2583	beta-1,4-N-acetyl-galactosaminyltransferase 1	.	.	X	X	.
BAHCC1	57597	BAH domain and coiled-coil containing 1	.	X	X	X	.
BAIAP2	10458	BAR/IMD domain containing adaptor protein 2	.	.	.	X	.
BAIAP2L2	80115	BAR/IMD domain containing adaptor protein 2 like 2	X	X	X	X	X
BAK1	578	BCL2 antagonist/killer 1	.	.	X	X	.
BDH1	622	3-hydroxybutyrate dehydrogenase 1	X	X	X	X	X
BLCAP	10904	BLCAP apoptosis inducing factor	.	.	X	X	.
BMERB1	89927	bMERB domain containing 1	.	.	X	X	.
BNC1	646	basonuclin 1	X	X	X	X	X
BTBD16	118663	BTB domain containing 16	X	X	X	X	X
C11orf53	341032	chromosome 11 open reading frame 53	.	.	X	X	.
C12orf42	374470	chromosome 12 open reading frame 42	X	X	X	X	.
C14orf39	317761	chromosome 14 open reading frame 39	.	.	X	X	.
C1QTNF5	114902	C1q and TNF related 5	.	.	X	X	.
C1orf94	84970	chromosome 1 open reading frame 94	.	.	.	X	.
C5orf38	153571	chromosome 5 open reading frame 38	X	X	X	X	.
C7orf50	84310	chromosome 7 open reading frame 50	.	.	X	X	.
CACNA1C	775	calcium voltage-gated channel subunit alpha1 C	.	.	.	X	.
CACNA1H	8912	calcium voltage-gated channel subunit alpha1 H	X	X	X	X	.
CACNB2	783	calcium voltage-gated channel auxiliary subunit beta 2	.	.	.	X	.
CALHM3	119395	calcium homeostasis modulator 3	.	.	.	X	.
CAPN2	824	calpain 2	.	.	X	X	.
CAPN9	10753	calpain 9	X	X	X	X	.
CARD11	84433	caspase recruitment domain family member 11	.	.	X	X	.
CASC15	401237	cancer susceptibility 15	.	.	X	X	.
CASC16	643714	cancer susceptibility 16	.	.	X	X	.
CBS	875	cystathionine beta-synthase	.	X	X	X	.
CCL15	6359	C-C motif chemokine ligand 15	X	X	X	X	X
CCL15-CCL14	348249	CCL15-CCL14 readthrough (NMD candidate)	X	X	X	X	X
CD81-AS1	101927682	CD81 antisense RNA 1	.	.	.	X	.
CD96	10225	CD96 molecule	.	.	X	X	.
CEACAM1	634	CEA cell adhesion molecule 1	.	.	X	X	.

Table A1.--Continued.

symbol	GeneID	full_name	top 50	top 100	top 200	top 500	Gini > .003
CHST8	64377	carbohydrate sulfotransferase 8	.	.	X	X	.
CIB3	117286	calcium and integrin binding family member 3	.	.	X	X	.
CIDEC	63924	cell death inducing DFFA like effector c	.	.	.	X	.
CMIP	80790	c-Maf inducing protein	.	.	X	X	.
CNIH3	149111	cornichon family AMPA receptor auxiliary protein 3	.	X	X	X	.
CNPY4	245812	canopy FGF signaling regulator 4	.	.	X	X	.
CNTFR	1271	ciliary neurotrophic factor receptor	.	.	.	X	.
CNTFR-AS1	415056	CNTFR antisense RNA 1	.	.	.	X	.
COL16A1	1307	collagen type XVI alpha 1 chain	.	.	.	X	.
CPEB1-AS1	283692	CPEB1 antisense RNA 1	.	.	.	X	.
CPNE6	9362	copine 6	.	.	.	X	.
CREB3L2	64764	cAMP responsive element binding protein 3 like 2	.	.	X	X	.
CREB3L2-AS1	100130880	CREB3L2 antisense RNA 1	.	.	X	X	.
CSNK1D	1453	casein kinase 1 delta	.	X	X	X	.
CTSB	1508	cathepsin B	X	X	X	X	.
CUX1	1523	cut like homeobox 1	.	.	X	X	.
CYFIP1	23191	cytoplasmic FMR1 interacting protein 1	.	.	.	X	.
CYP2B6	1555	cytochrome P450 family 2 subfamily B member 6	.	.	.	X	.
DAB1	1600	DAB adaptor protein 1	.	X	X	X	.
DCLK2	166614	doublecortin like kinase 2	X	X	X	X	X
DENND3	22898	DENN domain containing 3	.	.	X	X	.
DEPP1	11067	DEPP1 autophagy regulator	.	X	X	X	.
DHX30	22907	DExH-box helicase 30	.	X	X	X	.
DLX5	1749	distal-less homeobox 5	.	.	X	X	.
DNAJA4	55466	DnaJ heat shock protein family (Hsp40) member A4	.	.	X	X	.
DOCK2	1794	dedicator of cytokinesis 2	.	.	X	X	.
DOK7	285489	docking protein 7	X	X	X	X	X
EEF1A2	1917	eukaryotic translation elongation factor 1 alpha 2	.	.	X	X	.
EFL1	79631	elongation factor like GTPase 1	.	X	X	X	.
EGFL8	80864	EGF like domain multiple 8	.	.	.	X	.
EGR4	1961	early growth response 4	.	.	X	X	.
EHBP1L1	254102	EH domain binding protein 1 like 1	.	.	.	X	.
ELDR	102725541	EGFR long non-coding downstream RNA	X	X	X	X	X

Table A1.--Continued.

symbol	GeneID	full_name	top 50	top 100	top 200	top 500	Gini > .003
ELK3	2004	ETS transcription factor ELK3	.	.	.	X	.
ELMO1	9844	engulfment and cell motility 1	.	.	.	X	.
EYA4	2070	EYA transcriptional coactivator and phosphatase 4	.	.	.	X	.
F11R	50848	F11 receptor	.	.	X	X	.
FAM110A	83541	family with sequence similarity 110 member A	.	X	X	X	.
FAM110B	90362	family with sequence similarity 110 member B	.	.	X	X	.
FAM25A	643161	family with sequence similarity 25 member A	.	.	.	X	.
FBN1	2200	fibrillin 1	.	.	.	X	.
FHDC1	85462	FH2 domain containing 1	.	.	X	X	.
FIGN	55137	fidgetin, microtubule severing factor	.	.	.	X	.
FMN2	56776	formin 2	.	.	X	X	.
FOXI2	399823	forkhead box I2	.	X	X	X	.
FOXP1	27086	forkhead box P1	.	.	X	X	.
FOXP4	116113	forkhead box P4	.	.	X	X	.
FTCD	10841	formimidoyltransferase cyclodeaminase	.	.	X	X	.
GAA	2548	alpha glucosidase	.	.	X	X	.
GABRG3	2567	gamma-aminobutyric acid type A receptor subunit gamma3	.	X	X	X	.
GALNT2	2590	polypeptide N-acetylgalactosaminyltransferase 2	.	.	.	X	.
GAS7	8522	growth arrest specific 7	.	.	X	X	.
GDNF	2668	glial cell derived neurotrophic factor	.	X	X	X	.
GDPD3	79153	glycerophosphodiester phosphodiesterase domain containing 3	.	.	.	X	.
GJC2	57165	gap junction protein gamma 2	.	.	X	X	.
GMDS	2762	GDP-mannose 4,6-dehydratase	.	.	X	X	.
GMPPA	29926	GDP-mannose pyrophosphorylase A	.	.	X	X	.
GNG7	2788	G protein subunit gamma 7	.	.	X	X	.
GRM5	2915	glutamate metabotropic receptor 5	X	X	X	X	X
GRP	2922	gastrin releasing peptide	.	.	X	X	.
GUK1	2987	guanylate kinase 1	.	.	X	X	.
GYPC	2995	glycophorin C (Gerbich blood group)	X	X	X	X	.
H19	283120	H19 imprinted maternally expressed transcript	.	.	X	X	.
HCG17	414778	HLA complex group 17	.	.	X	X	.
HDAC5	10014	histone deacetylase 5	.	X	X	X	.

Table A1.--Continued.

symbol	GeneID	full_name	top 50	top 100	top 200	top 500	Gini > .003
HECW1	23072	HECT, C2 and WW domain containing E3 ubiquitin protein ligase 1	.	.	X	X	.
HLA-L	3139	major histocompatibility complex, class I, L (pseudogene)	.	.	X	X	.
HMX3	340784	H6 family homeobox 3	.	.	.	X	.
HNF4A	3172	hepatocyte nuclear factor 4 alpha	.	.	X	X	.
HOXA7	3204	homeobox A7	.	.	X	X	.
HOXD3	3232	homeobox D3	X	X	X	X	.
IGLON5	402665	IgLON family member 5	.	.	.	X	.
IL17REL	400935	interleukin 17 receptor E like	.	.	X	X	.
IMMP2L	83943	inner mitochondrial membrane peptidase subunit 2	.	.	X	X	.
INAVA	55765	innate immunity activator	.	X	X	X	.
IQSEC3	440073	IQ motif and Sec7 domain ArfGEF 3	X	X	X	X	X
IRF7	3665	interferon regulatory factor 7	.	.	X	X	.
IRX1	79192	iroquois homeobox 1	.	X	X	X	.
ITPR2	3709	inositol 1,4,5-trisphosphate receptor type 2	.	.	.	X	.
KCNC2	3747	potassium voltage-gated channel subfamily C member 2	.	.	.	X	.
KCNMA1	3778	potassium calcium-activated channel subfamily M alpha 1	.	.	.	X	.
KCNN2	3781	potassium calcium-activated channel subfamily N member 2	.	.	X	X	.
KCNN4	3783	potassium calcium-activated channel subfamily N member 4	.	.	.	X	.
KCNS1	3787	potassium voltage-gated channel modifier subfamily S member 1	.	.	.	X	.
KCTD19	146212	potassium channel tetramerization domain containing 19	.	.	.	X	.
KHDRBS2	202559	KH RNA binding domain containing, signal transduction associated 2	.	X	X	X	.
KLHL25	64410	kelch like family member 25	.	.	X	X	.
LFNG	3955	LFNG O-fucosylpeptide 3-beta-N-acetylglucosaminyltransferase	.	X	X	X	.
LHFPL4	375323	LHFPL tetraspan subfamily member 4	.	.	X	X	.
LINC00391	101927248	long intergenic non-protein coding RNA 391	.	.	.	X	.
LINC00461	645323	long intergenic non-protein coding RNA 461	.	.	X	X	.
LINC01248	102723818	long intergenic non-protein coding RNA 1248	.	.	.	X	.
LINC01932	105377682	long intergenic non-protein coding RNA 1932	.	.	X	X	.
LINC02743	646522	long intergenic non-protein coding RNA 2743	.	.	X	X	.
LIPE-AS1	100996307	LIPE antisense RNA 1	.	.	X	X	.

Table A1.--Continued.

symbol	GeneID	full_name	top 50	top 100	top 200	top 500	Gini > .003
LMNB2	84823	lamin B2	.	.	X	X	.
LMO3	55885	LIM domain only 3	.	.	X	X	.
LMX1A	4009	LIM homeobox transcription factor 1 alpha	.	.	.	X	.
LOC100130121	100130121	-	X	X	X	X	X
LOC101928682	101928682	-	.	.	X	X	.
LOC102724050	102724050	-	.	.	.	X	.
LOC105370362	105370362	-	.	.	X	X	.
LOC105372480	105372480	-	.	.	X	X	.
LOC388282	388282	-	.	.	X	X	.
LRP5	4041	LDL receptor related protein 5	.	.	X	X	.
LRRC36	55282	leucine rich repeat containing 36	.	.	.	X	.
LRRN3	54674	leucine rich repeat neuronal 3	.	.	X	X	.
LSP1	4046	lymphocyte specific protein 1	.	.	X	X	.
MAB21L4	79919	mab-21 like 4	.	.	X	X	.
MACF1	23499	microtubule actin crosslinking factor 1	.	.	.	X	.
MAEA	10296	macrophage erythroblast attacher, E3 ubiquitin ligase	.	X	X	X	.
MAP2K6	5608	mitogen-activated protein kinase kinase 6	.	X	X	X	.
MAPK8	5599	mitogen-activated protein kinase 8	.	.	.	X	.
MBLAC1	255374	metallo-beta-lactamase domain containing 1	.	.	X	X	.
MCF2L	23263	MCF.2 cell line derived transforming sequence like	X	X	X	X	X
MDFI	4188	MyoD family inhibitor	.	.	.	X	.
MEGF11	84465	multiple EGF like domains 11	.	.	X	X	.
MEGF6	1953	multiple EGF like domains 6	.	.	X	X	.
MELTF	4241	melanotransferrin	.	.	X	X	.
METRNL	284207	meteorin like, glial cell differentiation regulator	.	X	X	X	.
MFRP	83552	membrane frizzled-related protein	.	.	X	X	.
MIR1200	100302113	microRNA 1200	.	.	.	X	.
MIR1226	100302232	microRNA 1226	.	X	X	X	.
MIR1250	100302229	microRNA 1250	X	X	X	X	X
MIR1276	100302121	microRNA 1276	.	.	X	X	.
MIR129-2	406918	microRNA 129-2	.	.	X	X	.
MIR137	406928	microRNA 137	.	X	X	X	.
MIR137HG	400765	MIR137 host gene	.	X	X	X	.
MIR200A	406983	microRNA 200a	.	.	X	X	.

Table A1.--Continued.

symbol	GeneID	full_name	top 50	top 100	top 200	top 500	Gini > .003
MIR200B	406984	microRNA 200b	.	.	X	X	.
MIR2682	100616452	microRNA 2682	.	X	X	X	.
MIR3936HG	553103	MIR3936 host gene	.	X	X	X	.
MIR429	554210	microRNA 429	.	.	X	X	.
MIR663A	724033	microRNA 663a	.	.	.	X	.
MIR663AHG	284801	MIR663A host gene	.	.	.	X	.
MIR675	100033819	microRNA 675	.	.	X	X	.
MOGAT2	80168	monoacylglycerol O-acyltransferase 2	.	.	.	X	.
MST1R	4486	macrophage stimulating 1 receptor	.	.	.	X	.
MTSS1	9788	MTSS I-BAR domain containing 1	.	.	X	X	.
MUC17	140453	mucin 17, cell surface associated	.	.	X	X	.
MVP	9961	major vault protein	.	.	.	X	.
MYADML2	255275	myeloid associated differentiation marker like 2	.	.	X	X	.
MYEOV	26579	myeloma overexpressed	.	.	.	X	.
MYO10	4651	myosin X	.	.	.	X	.
MYO1E	4643	myosin IE	.	.	X	X	.
MYRF	745	myelin regulatory factor	.	X	X	X	.
NCOA7	135112	nuclear receptor coactivator 7	X	X	X	X	.
NCOA7-AS1	104355145	NCOA7 antisense RNA 1	X	X	X	X	.
NDUFAF6	137682	NADH:ubiquinone oxidoreductase complex assembly factor 6	.	.	X	X	.
NEFM	4741	neurofilament medium chain	.	.	X	X	.
NFATC1	4772	nuclear factor of activated T cells 1	.	.	X	X	.
NGEF	25791	neuronal guanine nucleotide exchange factor	.	.	.	X	.
NKX3-2	579	NK3 homeobox 2	.	.	X	X	.
NOS2	4843	nitric oxide synthase 2	.	.	X	X	.
NPR3	4883	natriuretic peptide receptor 3	.	X	X	X	.
NPY	4852	neuropeptide Y	.	.	X	X	.
NR3C1	2908	nuclear receptor subfamily 3 group C member 1	.	.	X	X	.
NRCAM	4897	neuronal cell adhesion molecule	.	.	.	X	.
NRXN1	9378	neurexin 1	.	.	.	X	.
NTHL1	4913	nth like DNA glycosylase 1	.	.	.	X	.
NTM	50863	neurotrimin	.	.	.	X	.
NYAP1	222950	neuronal tyrosine phosphorylated phosphoinositide-3-kinase adaptor 1	.	.	.	X	.

Table A1.--Continued.

symbol	GeneID	full_name	top 50	top 100	top 200	top 500	Gini > .003
OBI1-AS1	100874222	OBI1 antisense RNA 1	.	.	X	X	.
OBSCN	84033	obscurin, cytoskeletal calmodulin and titin-interacting RhoGEF	.	.	X	X	.
OCA2	4948	OCA2 melanosomal transmembrane protein	.	.	X	X	.
OPCML	4978	opioid binding protein/cell adhesion molecule like	.	.	X	X	.
OTX2	5015	orthodenticle homeobox 2	.	.	X	X	.
PAGR1	79447	PAXIP1 associated glutamate rich protein 1	.	.	.	X	.
PANTR1	100506421	POU3F3 adjacent non-coding transcript 1	.	.	X	X	.
PARM1	25849	prostate androgen-regulated mucin-like protein 1	.	.	X	X	.
PARP3	10039	poly(ADP-ribose) polymerase family member 3	.	.	X	X	.
PAX6-AS1	440034	PAX6 antisense RNA 1	.	.	.	X	.
PC	5091	pyruvate carboxylase	.	.	X	X	.
PCDHG@	56115	protocadherin gamma cluster	.	.	.	X	.
PCDHGA1	56114	protocadherin gamma subfamily A, 1	.	.	.	X	.
PCDHGA10	56106	protocadherin gamma subfamily A, 10	.	.	.	X	.
PCDHGA11	56105	protocadherin gamma subfamily A, 11	.	.	.	X	.
PCDHGA2	56113	protocadherin gamma subfamily A, 2	.	.	.	X	.
PCDHGA3	56112	protocadherin gamma subfamily A, 3	.	.	.	X	.
PCDHGA4	56111	protocadherin gamma subfamily A, 4	.	.	.	X	.
PCDHGA5	56110	protocadherin gamma subfamily A, 5	.	.	.	X	.
PCDHGA6	56109	protocadherin gamma subfamily A, 6	.	.	.	X	.
PCDHGA7	56108	protocadherin gamma subfamily A, 7	.	.	.	X	.
PCDHGA8	9708	protocadherin gamma subfamily A, 8	.	.	.	X	.
PCDHGA9	56107	protocadherin gamma subfamily A, 9	.	.	.	X	.
PCDHGB1	56104	protocadherin gamma subfamily B, 1	.	.	.	X	.
PCDHGB2	56103	protocadherin gamma subfamily B, 2	.	.	.	X	.
PCDHGB3	56102	protocadherin gamma subfamily B, 3	.	.	.	X	.
PCDHGB4	8641	protocadherin gamma subfamily B, 4	.	.	.	X	.
PCDHGB5	56101	protocadherin gamma subfamily B, 5	.	.	.	X	.
PCDHGB6	56100	protocadherin gamma subfamily B, 6	.	.	.	X	.
PCDHGB7	56099	protocadherin gamma subfamily B, 7	.	.	.	X	.
PDCD4	27250	programmed cell death 4	.	.	X	X	.
PDZD2	23037	PDZ domain containing 2	.	.	X	X	.
PFKP	5214	phosphofructokinase, platelet	.	.	.	X	.

Table A1.--Continued.

symbol	GeneID	full_name	top 50	top 100	top 200	top 500	Gini > .003
PHF12	57649	PHD finger protein 12	X	X	X	X	.
PHF21B	112885	PHD finger protein 21B	.	.	.	X	.
PIK3C2B	5287	phosphatidylinositol-4-phosphate 3-kinase catalytic subunit type 2 beta	.	.	X	X	.
PITX1	5307	paired like homeodomain 1	.	.	X	X	.
PITX2	5308	paired like homeodomain 2	.	.	.	X	.
PLEKHA6	22874	pleckstrin homology domain containing A6	.	.	.	X	.
PLXNA4	91584	plexin A4	.	.	X	X	.
PNKY	105447646	-	.	.	.	X	.
POLM	27434	DNA polymerase mu	.	.	X	X	.
POLR1A	25885	RNA polymerase I subunit A	.	.	.	X	.
POU3F3	5455	POU class 3 homeobox 3	.	X	X	X	.
POU4F1	5457	POU class 4 homeobox 1	.	.	X	X	.
PP12613	100192379	-	.	.	X	X	.
PPFIBP2	8495	PPFIA binding protein 2	.	.	.	X	.
PPT2-EGFL8	100532746	PPT2-EGFL8 readthrough (NMD candidate)	.	.	.	X	.
PRDM16	63976	PR/SET domain 16	.	.	X	X	.
PRKCA	5578	protein kinase C alpha	.	.	X	X	.
PRKCB	5579	protein kinase C beta	.	.	.	X	.
PRKCE	5581	protein kinase C epsilon	.	.	X	X	.
PROKR2	128674	prokineticin receptor 2	.	.	X	X	.
PRR35	146325	proline rich 35	X	X	X	X	.
PRRT1	80863	proline rich transmembrane protein 1	X	X	X	X	X
PSMG3	84262	proteasome assembly chaperone 3	.	.	X	X	.
PTGDR	5729	prostaglandin D2 receptor	X	X	X	X	X
PTPN14	5784	protein tyrosine phosphatase non-receptor type 14	.	.	.	X	.
PTPRN2	5799	protein tyrosine phosphatase receptor type N2	.	X	X	X	.
PUS1	80324	pseudouridine synthase 1	.	.	.	X	.
PXN	5829	paxillin	.	.	X	X	.
PYCARD	29108	PYD and CARD domain containing	.	.	X	X	.
RAI1	10743	retinoic acid induced 1	.	.	X	X	.
RAP1GAP2	23108	RAP1 GTPase activating protein 2	X	X	X	X	.
RAPGEFL1	51195	Rap guanine nucleotide exchange factor like 1	.	.	X	X	.
RASSF4	83937	Ras association domain family member 4	.	X	X	X	.

Table A1.--Continued.

symbol	GeneID	full_name	top 50	top 100	top 200	top 500	Gini > .003
RBFOX3	146713	RNA binding fox-1 homolog 3	.	.	X	X	.
RDH16	8608	retinol dehydrogenase 16	.	.	.	X	.
RESP18	389075	regulated endocrine specific protein 18	.	.	.	X	.
RHOBTB2	23221	Rho related BTB domain containing 2	.	.	.	X	.
RNF212	285498	ring finger protein 212	.	.	X	X	.
RRM2	6241	ribonucleotide reductase regulatory subunit M2	.	.	X	X	.
RRP9	9136	ribosomal RNA processing 9, U3 small nucleolar RNA binding protein	.	.	X	X	.
RTEL1	51750	regulator of telomere elongation helicase 1	.	X	X	X	.
RTEL1-TNFRSF6B	100533107	RTEL1-TNFRSF6B readthrough (NMD candidate)	.	X	X	X	.
RTN4R	65078	reticulon 4 receptor	.	.	X	X	.
RUFY1	80230	RUN and FYVE domain containing 1	.	.	X	X	.
RUNDC3A	10900	RUN domain containing 3A	.	.	X	X	.
RUVBL1	8607	RuvB like AAA ATPase 1	.	.	X	X	.
RYR2	6262	ryanodine receptor 2	.	.	X	X	.
S100A16	140576	S100 calcium binding protein A16	.	.	X	X	.
SALL1	6299	spalt like transcription factor 1	.	.	.	X	.
SALL3	27164	spalt like transcription factor 3	.	.	X	X	.
SARM1	23098	sterile alpha and TIR motif containing 1	.	.	.	X	.
SCRT2	85508	scratch family transcriptional repressor 2	.	.	X	X	.
SEL1L3	23231	SEL1L family member 3	.	.	X	X	.
SELENOI	85465	selenoprotein I	.	.	X	X	.
SFN	2810	stratifin	.	.	.	X	.
SHH	6469	sonic hedgehog signaling molecule	.	.	.	X	.
SIX6	4990	SIX homeobox 6	.	.	X	X	.
SLC12A7	10723	solute carrier family 12 member 7	.	X	X	X	.
SLC16A3	9123	solute carrier family 16 member 3	.	.	X	X	.
SLC22A10	387775	solute carrier family 22 member 10	.	.	X	X	.
SLC22A18	5002	solute carrier family 22 member 18	.	.	X	X	.
SLC22A18AS	5003	SLC22A18 antisense RNA	.	.	X	X	.
SLC22A23	63027	solute carrier family 22 member 23	.	X	X	X	.
SLC22A4	6583	solute carrier family 22 member 4	.	X	X	X	.
SLC34A3	142680	solute carrier family 34 member 3	.	.	X	X	.
SLC38A10	124565	solute carrier family 38 member 10	.	.	X	X	.

Table A1.--Continued.

symbol	GeneID	full_name	top 50	top 100	top 200	top 500	Gini > .003
SLC38A3	10991	solute carrier family 38 member 3	.	.	X	X	.
SLC41A3	54946	solute carrier family 41 member 3	.	.	X	X	.
SLC6A19	340024	solute carrier family 6 member 19	X	X	X	X	.
SMG6	23293	SMG6 nonsense mediated mRNA decay factor	.	.	X	X	.
SMIM43	132332	small integral membrane protein 43	.	.	X	X	.
SMU1	55234	SMU1 DNA replication regulator and spliceosomal factor	.	.	X	X	.
SNTG2	54221	syntrophin gamma 2	.	.	.	X	.
SNTG2-AS1	101060391	SNTG2 antisense RNA 1	.	.	.	X	.
SNX25	83891	sorting nexin 25	.	.	.	X	.
SOX11	6664	SRY-box transcription factor 11	.	X	X	X	.
SPEG	10290	striated muscle enriched protein kinase	.	.	X	X	.
SPEGNB	100996693	SPEG neighbor	.	.	X	X	.
SPON1	10418	spondin 1	.	.	.	X	.
SPSB4	92369	splA/ryanodine receptor domain and SOCS box containing 4	.	.	.	X	.
SPX	80763	spexin hormone	.	.	X	X	.
SSBP2	23635	single stranded DNA binding protein 2	.	.	.	X	.
SSH1	54434	slingshot protein phosphatase 1	X	X	X	X	X
SSPOP	23145	SCO-spondin, pseudogene	.	.	.	X	.
STOML1	9399	stomatin like 1	.	.	X	X	.
STRA6	64220	signaling receptor and transporter of retinol STRA6	.	.	X	X	.
STRADB	55437	STE20 related adaptor beta	.	.	X	X	.
STX18	53407	syntaxin 18	.	.	X	X	.
SUN1	23353	Sad1 and UNC84 domain containing 1	.	.	.	X	.
TACC2	10579	transforming acidic coiled-coil containing protein 2	.	X	X	X	.
TBC1D16	125058	TBC1 domain family member 16	X	X	X	X	X
TBR1	10716	T-box brain transcription factor 1	.	.	.	X	.
TBX15	6913	T-box transcription factor 15	.	.	X	X	.
TBX2-AS1	103689912	TBX2 antisense RNA 1	.	.	X	X	.
TBX5	6910	T-box transcription factor 5	.	.	X	X	.
TCEA3	6920	transcription elongation factor A3	X	X	X	X	X
TEAD3	7005	TEA domain transcription factor 3	.	.	X	X	.
TENM4	26011	teneurin transmembrane protein 4	.	.	X	X	.
TENT5A	55603	terminal nucleotidyltransferase 5A	.	.	X	X	.

Table A1.--Continued.

symbol	GeneID	full_name	top 50	top 100	top 200	top 500	Gini > .003
TENT5C	54855	terminal nucleotidyltransferase 5C	.	.	X	X	.
TFAP2D	83741	transcription factor AP-2 delta	.	.	.	X	.
TFF1	7031	trefoil factor 1	.	.	X	X	.
TGFB3	7043	transforming growth factor beta 3	.	.	X	X	.
TIMM13	26517	translocase of inner mitochondrial membrane 13	.	.	X	X	.
TLDC2	140711	TBC/LysM-associated domain containing 2	X	X	X	X	X
TLK1	9874	tousled like kinase 1	.	.	.	X	.
TLX3	30012	T cell leukemia homeobox 3	.	.	X	X	.
TM6SF1	53346	transmembrane 6 superfamily member 1	.	.	.	X	.
TMEM105	284186	TMEM105 long non-coding RNA	.	X	X	X	.
TMEM151A	256472	transmembrane protein 151A	.	.	.	X	.
TMEM196	256130	transmembrane protein 196	.	.	X	X	.
TMEM220-AS1	101101775	TMEM220 antisense RNA 1	.	.	.	X	.
TMEM234	56063	transmembrane protein 234	.	.	X	X	.
TMEM238L	100289255	transmembrane protein 238 like	.	.	.	X	.
TMEM52	339456	transmembrane protein 52	.	.	X	X	.
TNFRSF10D	8793	TNF receptor superfamily member 10d	.	.	.	X	.
TNFRSF6B	8771	TNF receptor superfamily member 6b	.	X	X	X	.
TNK2	10188	tyrosine kinase non receptor 2	X	X	X	X	X
TOX3	27324	TOX high mobility group box family member 3	.	.	X	X	.
TPSD1	23430	tryptase delta 1	.	.	X	X	.
TRAPPC9	83696	trafficking protein particle complex subunit 9	X	X	X	X	.
TRERF1	55809	transcriptional regulating factor 1	.	.	X	X	.
TRIM15	89870	tripartite motif containing 15	X	X	X	X	X
TRIM58	25893	tripartite motif containing 58	.	.	.	X	.
TRIO	7204	trio Rho guanine nucleotide exchange factor	.	.	X	X	.
TSNAX	7257	translin associated factor X	.	.	X	X	.
TSNAX-DISC1	100303453	TSNAX-DISC1 readthrough (NMD candidate)	.	.	X	X	.
TTC39A	22996	tetratricopeptide repeat domain 39A	X	X	X	X	X
USH1C	10083	USH1 protein network component harmonin	.	.	X	X	.
VAMP5	10791	vesicle associated membrane protein 5	.	.	.	X	.
VANGL1	81839	VANGL planar cell polarity protein 1	.	X	X	X	.
VAT1L	57687	vesicle amine transport 1 like	.	.	.	X	.

Table A1.--Continued.

symbol	GeneID	full_name	top 50	top 100	top 200	top 500	Gini > .003
VILL	50853	villin like	.	.	.	X	.
VMO1	284013	vitelline membrane outer layer 1 homolog	.	.	X	X	.
VOPP1	81552	VOPP1 WW domain binding protein	.	.	X	X	.
YBX3P1	440359	Y-box binding protein 3 pseudogene 1	.	.	X	X	.
ZBED2	79413	zinc finger BED-type containing 2	.	.	X	X	.
ZDHHC14	79683	zinc finger DHHC-type palmitoyltransferase 14	.	.	X	X	.
ZDHHC7	55625	zinc finger DHHC-type palmitoyltransferase 7	.	.	.	X	.
ZFHX3	463	zinc finger homeobox 3	.	.	.	X	.
ZFP41	286128	ZFP41 zinc finger protein	.	X	X	X	.
ZNF184	7738	zinc finger protein 184	.	.	X	X	.
ZNF382	84911	zinc finger protein 382	.	.	X	X	.
ZNF385A	25946	zinc finger protein 385A	.	.	.	X	.
ZNF529	57711	zinc finger protein 529	.	.	X	X	.
ZNF578	147660	zinc finger protein 578	.	.	X	X	.
ZNF732	654254	zinc finger protein 732	.	.	X	X	.
ZSCAN22	342945	zinc finger and SCAN domain containing 22	.	.	X	X	.
uc001ulg.1	.	.	.	.	X	X	.
uc002duh.2	.	.	.	.	.	X	.
uc002fmo.1	.	.	.	.	.	X	.
uc002hdl.3	.	.	.	.	X	X	.
uc002igh.3	.	.	.	.	X	X	.
uc002wly.1	.	.	.	.	X	X	.
uc002xlw.1	.	.	.	.	X	X	.
uc003hii.3	.	.	.	.	X	X	.
uc003wpj.2	.	.	.	.	X	X	.
uc009yar.1	.	.	.	X	X	X	.
uc010luc.2	.	.	.	.	X	X	.
uc010ofi.1	.	.	.	.	.	X	.
uc021osq.1	.	.	.	.	X	X	.
uc021qbx.2	.	.	.	.	X	X	.
uc021qoa.1	.	.	.	.	X	X	.
uc021rcu.1	.	.	.	.	X	X	.
uc021stw.1	.	.	.	.	X	X	.

Table A1.--Continued.

symbol	GeneID	full_name	top 50	top 100	top 200	top 500	Gini > .003
uc021ufi.1	.	.	.	.	X	X	.
uc021vwu.1	.	.	.	.	X	X	.
uc021xci.1	.	.	.	.	X	X	.
uc022aig.1	.	.	.	.	X	X	.
uc031pkr.1	.	.	.	.	X	X	.
uc031pks.1	.	.	.	.	X	X	.
uc031pzv.1	.	.	.	.	X	X	.

Table A2.—Median test performance for models trained at a single concentration but tested against all others. Here, ctDNA fraction refers to the tumor fraction the model was trained with while performance metrics show the model's performance in classifying others.

ctDNA fraction	sensitivity	specificity	PPV	NPV	accuracy	F1
0.001	0.961	0.996	0.994	0.973	0.982	0.978
0.01	0.957	0.999	0.998	0.971	0.981	0.977
0.05	0.932	0.998	0.998	0.954	0.967	0.960
0.1	0.780	0.999	0.997	0.865	0.908	0.876

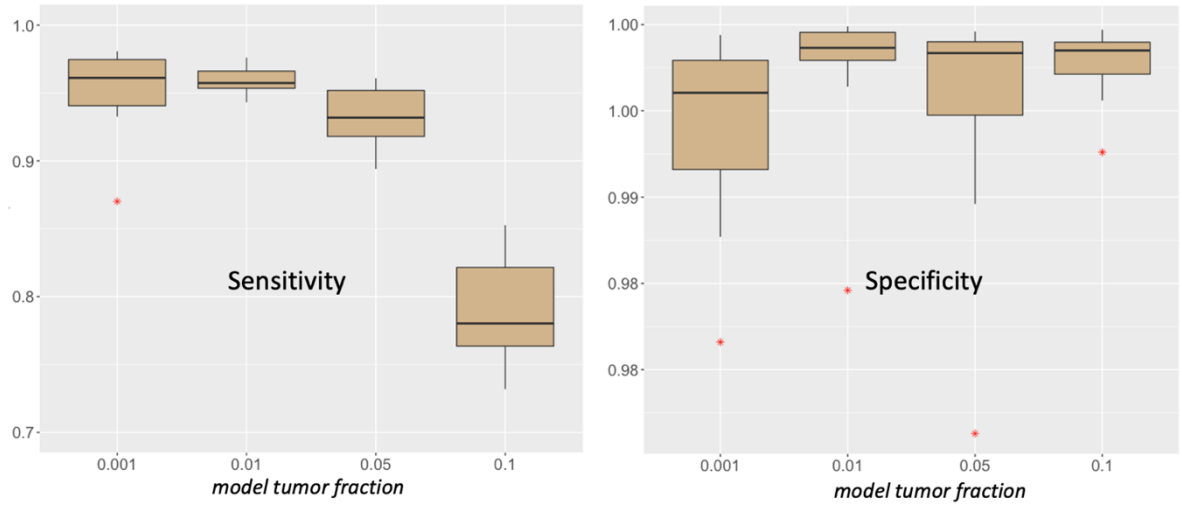

Figure A1.--Sensitivity and specificity of models trained at a single concentration but tested against all others.

Table A3.-- Median classification sensitivity and specificity of models trained at a single concentration

training set	ctDNA fraction	sensitivity	specificity	PPV	NPV	accuracy	F1
all fractions	0.001	0.973	0.992	0.965	0.994	0.987	0.967
all fractions	0.01	0.981	0.993	0.973	0.996	0.990	0.975
all fractions	0.05	0.984	0.995	0.980	0.996	0.993	0.983
all fractions	0.1	0.991	0.993	0.970	0.998	0.991	0.978
.1% tumor	0.001	0.928	0.996	0.982	0.983	0.982	0.953
.1% tumor	0.01	0.949	0.997	0.985	0.988	0.987	0.965
.1% tumor	0.05	0.939	0.998	0.992	0.986	0.986	0.963
.1% tumor	0.1	0.972	0.998	0.990	0.994	0.993	0.982
1% tumor	0.001	0.965	0.998	0.991	0.992	0.990	0.974
1% tumor	0.01	0.947	0.998	0.992	0.988	0.986	0.962
1% tumor	0.05	0.963	0.998	0.990	0.991	0.988	0.968
1% tumor	0.1	0.972	0.998	0.991	0.993	0.991	0.975
5% tumor	0.001	0.916	0.998	0.992	0.980	0.981	0.949
5% tumor	0.01	0.925	0.997	0.985	0.983	0.983	0.955
5% tumor	0.05	0.933	0.996	0.984	0.984	0.983	0.956
5% tumor	0.1	0.981	0.998	0.990	0.996	0.995	0.987
10% tumor	0.001	0.760	0.998	0.990	0.946	0.951	0.856
10% tumor	0.01	0.752	0.999	0.993	0.944	0.951	0.855
10% tumor	0.05	0.889	0.999	0.994	0.974	0.975	0.934
10% tumor	0.1	0.957	0.998	0.990	0.990	0.988	0.970

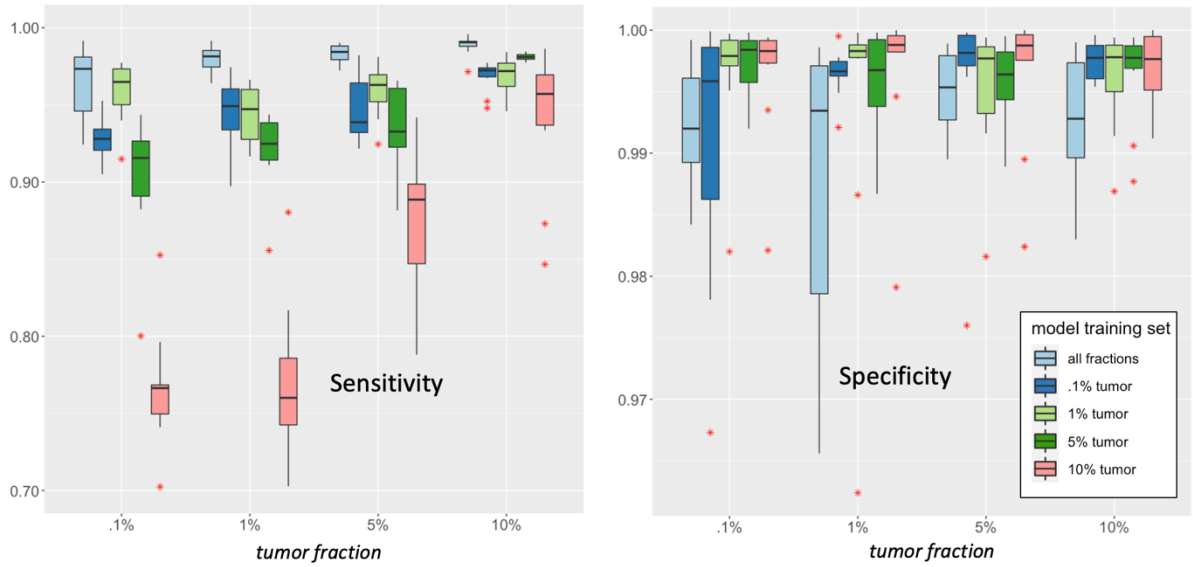

Figure A2.--Sensitivity and specificity of models trained at a single concentration across all test sets

REFERENCES

1. *American Cancer Society - Cancer Statistics Center*. Available from: https://cancerstatisticscenter.cancer.org/.
2. Rahib, L., et al., *Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States.* Cancer Res, 2014. **74**(11): p. 2913-21.
3. *National Cancer Institute - Surveillance, Epidemiology and End Results Program*. Available from: https://seer.cancer.gov/statfacts/html/pancreas.html.
4. Ryan, D.P., T.S. Hong, and N. Bardeesy, *Pancreatic adenocarcinoma.* N Engl J Med, 2014. **371**(11): p. 1039-49.
5. Lee, J.S., et al., *Liquid biopsy in pancreatic ductal adenocarcinoma: current status of circulating tumor cells and circulating tumor DNA.* Mol Oncol, 2019. **13**(8): p. 1623-1650.
6. Bozic, I., et al., *Accumulation of driver and passenger mutations during tumor progression.* Proc Natl Acad Sci U S A, 2010. **107**(43): p. 18545-50.
7. Huang, L., et al., *KRAS mutation: from undruggable to druggable in cancer.* Signal Transduct Target Ther, 2021. **6**(1): p. 386.
8. Ogura, T., et al., *Prognostic value of K-ras mutation status and subtypes in endoscopic ultrasound-guided fine-needle aspiration specimens from patients with unresectable pancreatic cancer.* J Gastroenterol, 2013. **48**(5): p. 640-6.
9. Aguirre, A.J., et al., *Activated Kras and Ink4a/Arf deficiency cooperate to produce metastatic pancreatic ductal adenocarcinoma.* Genes Dev, 2003. **17**(24): p. 3112-26.
10. Kanda, M., et al., *Presence of somatic mutations in most early-stage pancreatic intraepithelial neoplasia.* Gastroenterology, 2012. **142**(4): p. 730-733 e9.
11. Hustinx, S.R., et al., *Concordant loss of MTAP and p16/CDKN2A expression in pancreatic intraepithelial neoplasia: evidence of homozygous deletion in a noninvasive precursor lesion.* Mod Pathol, 2005. **18**(7): p. 959-63.
12. Herreros-Villanueva, M., et al., *Molecular markers in pancreatic cancer diagnosis.* Clin Chim Acta, 2013. **418**: p. 22-9.
13. Shinjo, K., et al., *A novel sensitive detection method for DNA methylation in circulating free DNA of pancreatic cancer.* PLoS One, 2020. **15**(6): p. e0233782.
14. Brancaccio, M., et al., *Cell-Free DNA methylation: the new frontiers of pancreatic cancer biomarkers' ciscovery.* Genes (Basel), 2019. **11**(1).
15. Goggins, M., *Identifying molecular markers for the early detection of pancreatic neoplasia.* Semin Oncol, 2007. **34**(4): p. 303-10.
16. Mishra, N.K. and C. Guda, *Genome-wide DNA methylation analysis reveals molecular subtypes of pancreatic cancer.* Oncotarget, 2017. **8**(17): p. 28990-29012.
17. *American Cancer Society - Signs and Symptoms of Pancreatic Cancer*. Available from: https://www.cancer.org/cancer/pancreatic-cancer/detection-diagnosis-staging/signs-and-symptoms.htm.
18. Henriksen, S.D., et al., *Cell-free DNA promoter hypermethylation in plasma as a diagnostic marker for pancreatic adenocarcinoma.* Clin Epigenetics, 2016. **8**: p. 117.
19. Feig, C., et al., *The pancreas cancer microenvironment.* Clin Cancer Res, 2012. **18**(16): p. 4266-76.

20. Puri, A. and M.W. Saif, *Pharmacogenomics update in pancreatic cancer.* JOP, 2014. **15**(2): p. 114-7.
21. Collins, H., et al., *Information needs in the precision medicine era: how genetics home reference can help.* Interact J Med Res, 2016. **5**(2): p. e13.
22. Green, E.D., M.S. Guyer, and I. National Human Genome Research, *Charting a course for genomic medicine from base pairs to bedside.* Nature, 2011. **470**(7333): p. 204-13.
23. Zhao, L., et al., *Molecular subtyping of cancer: current status and moving toward clinical applications.* Brief Bioinform, 2018.
24. Collisson, E.A., et al., *Molecular subtypes of pancreatic cancer.* Nat Rev Gastroenterol Hepatol, 2019.
25. Liu, Z. and S. Zhang, *Tumor characterization and stratification by integrated molecular profiles reveals essential pan-cancer features.* BMC Genomics, 2015. **16**: p. 503.
26. Volckmar, A.L., et al., *A field guide for cancer diagnostics using cell-free DNA: From principles to practice and clinical applications.* Genes Chromosomes Cancer, 2018. **57**(3): p. 123-139.
27. Endris, V., et al., *Molecular diagnostic profiling of lung cancer specimens with a semiconductor-based massive parallel sequencing approach: feasibility, costs, and performance compared with conventional sequencing.* J Mol Diagn, 2013. **15**(6): p. 765-75.
28. Diaz, L.A., Jr. and A. Bardelli, *Liquid biopsies: genotyping circulating tumor DNA.* J Clin Oncol, 2014. **32**(6): p. 579-86.
29. Huang, J. and L. Wang, *Cell-Free DNA methylation profiling analysis-technologies and bioinformatics.* Cancers (Basel), 2019. **11**(11).
30. Wan, J.C.M., et al., *Liquid biopsies come of age: towards implementation of circulating tumour DNA.* Nat Rev Cancer, 2017. **17**(4): p. 223-238.
31. Mandel, P. and P. Metais, *Nuclear acids in human blood plasma.* C R Seances Soc Biol Fil, 1948. **142**(3-4): p. 241-3.
32. Leon, S.A., et al., *Free DNA in the serum of cancer patients and the effect of therapy.* Cancer Res, 1977. **37**(3): p. 646-50.
33. Martins, I., et al., *Liquid biopsies: applications for cancer diagnosis and monitoring.* Genes (Basel), 2021. **12**(3).
34. *US Food and Drug Administration - Guardant360 CDx*. Available from: https://www.fda.gov/medical-devices/recently-approved-devices/guardant360-cdx-p200010s001.
35. Aggarwal, C., et al., *Clinical implications of plasma-based genotyping with the delivery of personalized therapy in metastatic non-small cell lung cancer.* JAMA Oncol, 2019. **5**(2): p. 173-180.
36. Leighl, N.B., et al., *Clinical utility of comprehensive cell-free DNA analysis to identify genomic biomarkers in patients with newly diagnosed metastatic non-small cell lung cancer.* Clin Cancer Res, 2019. **25**(15): p. 4691-4700.
37. Bai, Y. and H. Zhao, *Liquid biopsy in tumors: opportunities and challenges.* Ann Transl Med, 2018. **6**(Suppl 1): p. S89.
38. Aravanis, A.M., M. Lee, and R.D. Klausner, *Next-generation sequencing of circulating tumor DNA for early cancer detection.* Cell, 2017. **168**(4): p. 571-574.

39. El Messaoudi, S., et al., *Circulating cell free DNA: preanalytical considerations.* Clin Chim Acta, 2013. **424**: p. 222-30.
40. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors.* Proc Natl Acad Sci U S A, 1977. **74**(12): p. 5463-7.
41. Busser, B., et al., *Plasma circulating tumor DNA levels for the monitoring of melanoma patients: landscape of available technologies and clinical applications.* Biomed Res Int, 2017. **2017**: p. 5986129.
42. Morash, M., et al., *The role of next-generation sequencing in precision medicine: a review of outcomes in oncology.* J Pers Med, 2018. **8**(3).
43. Cummings, C.A., et al., *The role of next-generation sequencing in enabling personalized oncology therapy.* Clin Transl Sci, 2016. **9**(6): p. 283-292.
44. Saunders, C.J., et al., *Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units.* Sci Transl Med, 2012. **4**(154): p. 154ra135.
45. Newman, A.M., et al., *An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage.* Nat Med, 2014. **20**(5): p. 548-54.
46. *PacBio - How 5-Base HiFi sequencing calls methylation status*. Available from: https://www.pacb.com/products-and-services/applications/epigenetics/.
47. Li, M.M., et al., *Standards and guidelines for the interpretation and reporting of sequence variants in cancer: A joint consensus recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists.* J Mol Diagn, 2017. **19**(1): p. 4-23.
48. Mao, L., et al., *Detection of oncogene mutations in sputum precedes diagnosis of lung cancer.* Cancer Res, 1994. **54**(7): p. 1634-7.
49. Gormally, E., et al., *TP53 and KRAS2 mutations in plasma DNA of healthy subjects and subsequent cancer occurrence: a prospective study.* Cancer Res, 2006. **66**(13): p. 6871-6.
50. Tost, J., *DNA methylation: an introduction to the biology and the disease-associated changes of a promising biomarker.* Mol Biotechnol, 2010. **44**(1): p. 71-81.
51. Ziller, M.J., et al., *Charting a dynamic DNA methylation landscape of the human genome.* Nature, 2013. **500**(7463): p. 477-81.
52. Ferguson-Smith, A.C., *Genomic imprinting: the emergence of an epigenetic paradigm.* Nat Rev Genet, 2011. **12**(8): p. 565-75.
53. Vidal, E., et al., *A DNA methylation map of human cancer at single base-pair resolution.* Oncogene, 2017. **36**(40): p. 5648-5657.
54. Kandi, V. and S. Vadakedath, *Effect of DNA methylation in various diseases and the probable protective role of nutrition: a mini-review.* Cureus, 2015. **7**(8): p. e309.
55. Jin, Z. and Y. Liu, *DNA methylation in human diseases.* Genes Dis, 2018. **5**(1): p. 1-8.
56. Elli, F.M., et al., *Mosaicism for GNAS methylation defects associated with pseudohypoparathyroidism type 1B arose in early post-zygotic phases.* Clin Epigenetics, 2018. **10**: p. 16.
57. Avitzour, M., et al., *FMR1 epigenetic silencing commonly occurs in undifferentiated fragile X-affected embryonic stem cells.* Stem Cell Reports, 2014. **3**(5): p. 699-706.
58. Robertson, K.D., *DNA methylation and human disease.* Nat Rev Genet, 2005. **6**(8): p. 597-610.

59. Agrawal, A., R.F. Murphy, and D.K. Agrawal, *DNA methylation in breast and colorectal cancers.* Mod Pathol, 2007. **20**(7): p. 711-21.
60. Lee, C.J., et al., *Impact of mutations in DNA methylation modification genes on genome-wide methylation landscapes and downstream gene activations in pan-cancer.* BMC Med Genomics, 2020. **13**(Suppl 3): p. 27.
61. Zheng, Y., et al., *Genome-wide DNA methylation analysis identifies candidate epigenetic markers and drivers of hepatocellular carcinoma.* Brief Bioinform, 2018. **19**(1): p. 101-108.
62. Golub, T.R., et al., *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* Science, 1999. **286**(5439): p. 531-7.
63. Alizadeh, A.A., et al., *Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.* Nature, 2000. **403**(6769): p. 503-11.
64. Bailey, P., et al., *Genomic analyses identify molecular subtypes of pancreatic cancer.* Nature, 2016. **531**(7592): p. 47-52.
65. Xu, R.H., et al., *Circulating tumour DNA methylation markers for diagnosis and prognosis of hepatocellular carcinoma.* Nat Mater, 2017. **16**(11): p. 1155-1161.
66. Cancer Genome Atlas Research, N., *Comprehensive molecular profiling of lung adenocarcinoma.* Nature, 2014. **511**(7511): p. 543-50.
67. Capper, D., et al., *DNA methylation-based classification of central nervous system tumours.* Nature, 2018. **555**(7697): p. 469-474.
68. Louis, D.N., et al., *The 2016 World Health Organization classification of tumors of the central nervous system: a summary.* Acta Neuropathol, 2016. **131**(6): p. 803-20.
69. Kourou, K., et al., *Machine learning applications in cancer prognosis and prediction.* Comput Struct Biotechnol J, 2015. **13**: p. 8-17.
70. Chaudhary, K., et al., *Deep learning-based multi-omics integration robustly predicts survival in liver cancer.* Clin Cancer Res, 2018. **24**(6): p. 1248-1259.
71. Chatterton, Z., et al., *Validation of DNA methylation biomarkers for diagnosis of acute lymphoblastic leukemia.* Clin Chem, 2014. **60**(7): p. 995-1003.
72. Aguirre, A.J., et al., *Real-time genomic characterization of advanced pancreatic cancer to enable precision medicine.* Cancer Discov, 2018. **8**(9): p. 1096-1111.
73. Liu, L., et al., *Targeted methylation sequencing of plasma cell-free DNA for cancer detection and classification.* Ann Oncol, 2018. **29**(6): p. 1445-1453.
74. Guo, S., et al., *Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA.* Nat Genet, 2017. **49**(4): p. 635-642.
75. Kang, S., et al., *CancerLocator: non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA.* Genome Biol, 2017. **18**(1): p. 53.
76. Nassiri, F., et al., *Detection and discrimination of intracranial tumors using plasma cell-free DNA methylomes.* Nat Med, 2020.
77. Nuzzo, P.V., et al., *Detection of renal cell carcinoma using plasma and urine cell-free DNA methylomes.* Nat Med, 2020.
78. Cohen, J.D., et al., *Combined circulating tumor DNA and protein biomarker-based liquid biopsy for the earlier detection of pancreatic cancers.* Proc Natl Acad Sci U S A, 2017. **114**(38): p. 10202-10207.

79. Henriksen, S.D., et al., *Cell-free DNA promoter hypermethylation in plasma as a predictive marker for survival of patients with pancreatic adenocarcinoma.* Oncotarget, 2017. **8**(55): p. 93942-93956.
80. Le Calvez-Kelm, F., et al., *KRAS mutations in blood circulating cell-free DNA: a pancreatic cancer case-control.* Oncotarget, 2016. **7**(48): p. 78827-78840.
81. Melnikov, A.A., et al., *Methylation profile of circulating plasma DNA in patients with pancreatic cancer.* J Surg Oncol, 2009. **99**(2): p. 119-22.
82. Park, J.W., I.H. Baek, and Y.T. Kim, *Preliminary study analyzing the methylated genes in the plasma of patients with pancreatic cancer.* Scand J Surg, 2012. **101**(1): p. 38-44.
83. Henriksen, S.D., et al., *Promoter hypermethylation in plasma-derived cell-free DNA as a prognostic marker for pancreatic adenocarcinoma staging.* Int J Cancer, 2017. **141**(12): p. 2489-2497.
84. Cancer Genome Atlas Research Network. Electronic address, a.a.d.h.e. and N. Cancer Genome Atlas Research, *Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma.* Cancer Cell, 2017. **32**(2): p. 185-203 e13.
85. *The Cancer Genome Atlas - Pancreatic Adenocarcinoma Project*. Available from: https://portal.gdc.cancer.gov/projects/TCGA-PAAD.
86. *Infinium HumanMethylation450K BeadChip*. Available from: https://support.illumina.com/array/array_kits/infinium_humanmethylation450_beadchip_kit.html.
87. Dedeurwaerder, S., et al., *A comprehensive overview of Infinium HumanMethylation450 data processing.* Brief Bioinform, 2014. **15**(6): p. 929-41.
88. *National Cancer Institute - Genomic Data Portal*. Available from: https://portal.gdc.cancer.gov/.
89. Grossman, R.L., et al., *Toward a shared vision for cancer genomic data.* N Engl J Med, 2016. **375**(12): p. 1109-12.
90. Edgar, R., M. Domrachev, and A.E. Lash, *Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.* Nucleic Acids Res, 2002. **30**(1): p. 207-10.
91. Inoshita, M., et al., *Sex differences of leukocytes DNA methylation adjusted for estimated cellular proportions.* Biol Sex Differ, 2015. **6**: p. 11.
92. Pedregosa, F.V., G; Gramfort, A; Michel, V; Thirion, B; Grisel, O; Blondel, M; Prettenhofer, P; Weiss, R; Dubourg, V; Vanderplas, J; Passos, A; Cournapeau, D; Brucher, M; Perrot, M; Duchesnay, E, *Scikit-learn: machine learning in python.* Journal of Machine Learning Research, 2011. **12**: p. 2825-2830.
93. McKinney, W., *Data structures for statistical computing in python.* Proceedings of the 9th Python in Science Conference, 2010: p. 56-61.
94. Harris, C.R., et al., *Array programming with NumPy.* Nature, 2020. **585**(7825): p. 357-362.
95. Abadi, M.A., Ashish; Barhan, Paul; Brevdo, Eugene; Chen, Zhifeng; Citro, Craig; Corrado, Greg S.; Corrado, Andy; Dean, Jeffrey; Zheng, Xioqiang, *TensorFlow: Large-scale machine learning on heterogeneous distributed systems.* arXiv, 2015.
96. Team, E.D.j.D., *Deeplearning4j: Open-source distributed deep learning for the JVM, Apache Software Foundation License 2.0*.
97. Breiman, L., *Random Forests.* Machine Learning, 2001. **45**(1): p. 5-32.

98. Celli, F., F. Cumbo, and E. Weitschek, *Classification of large DNA methylation datasets for identifying cancer drivers.* Big Data Research, 2018. **13**: p. 21-28.
99. Lee, B.T., et al., *The UCSC Genome Browser database: 2022 update.* Nucleic Acids Res, 2022. **50**(D1): p. D1115-D1122.
100. Kent, W.J., et al., *The human genome browser at UCSC.* Genome Res, 2002. **12**(6): p. 996-1006.
101. *Biological Pathways Fact Sheet*. Available from: https://www.genome.gov/about-genomics/fact-sheets/Biological-Pathways-Fact-Sheet.
102. Fabregat, A., et al., *The Reactome pathway Knowledgebase.* Nucleic Acids Res, 2016. **44**(D1): p. D481-7.
103. Yang, Y. and Q. Mei, *Accumulation of AGO2 facilitates tumorigenesis of human hepatocellular carcinoma.* Biomed Res Int, 2020. **2020**: p. 1631843.
104. Hopfield, J.J., *Neural networks and physical systems with emergent collective computational abilities.* Proc Natl Acad Sci U S A, 1982. **79**(8): p. 2554-8.
105. Oustimov, A.V., Vincent, *Artificial neural networks in the cancer genomics frontier.* Translational Cancer Research, 2014. **Vol 3, No 3**.
106. Sundaram, L., et al., *Predicting the clinical impact of human mutation with deep neural networks.* Nat Genet, 2018. **50**(8): p. 1161-1170.
107. Peng, B., et al., *Genetic data simulators and their applications: an overview.* Genet Epidemiol, 2015. **39**(1): p. 2-10.
108. Fawcett, T., *An introduction to ROC analysis.* Pattern Recognition Letters, 2006. **27**(8): p. 861-874.
109. Tharwat, A., *Classification assessment methods.* Applied Computing and Informatics, 2020. **17**(1): p. 168-192.
110. Doughtery, E.R., H. Jianping, and M.L. Bittner, *Validation of computational methods in genomics.* Curr Genomics, 2007. **8**(1): p. 1-19.
111. Mattocks, C.J., et al., *A standardized framework for the validation and verification of clinical molecular genetic tests.* Eur J Hum Genet, 2010. **18**(12): p. 1276-88.
112. Yuan, X., et al., *An overview of population genetic data simulation.* J Comput Biol, 2012. **19**(1): p. 42-54.
113. Landry, L.G., et al., *Lack of diversity in genomic databases is a barrier to translating precision medicine research into practice.* Health Aff (Millwood), 2018. **37**(5): p. 780-785.
114. *Genetic Simulation Resources (GSR)*. Available from: https://surveillance.cancer.gov/genetic-simulation-resources/.
115. Lacey, M.R., C. Baribault, and M. Ehrlich, *Modeling, simulation and analysis of methylation profiles from reduced representation bisulfite sequencing experiments.* Stat Appl Genet Mol Biol, 2013. **12**(6): p. 723-42.
116. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
117. Davis, C.A., et al., *The Encyclopedia of DNA elements (ENCODE): data portal update.* Nucleic Acids Res, 2018. **46**(D1): p. D794-D801.
118. Rackham, O.J., et al., *WGBSSuite: simulating whole-genome bisulphite sequencing data and benchmarking differential DNA methylation analysis tools.* Bioinformatics, 2015. **31**(14): p. 2371-3.

119. Frith, M.C., R. Mori, and K. Asai, *A mostly traditional approach improves alignment of bisulfite-converted DNA.* Nucleic Acids Res, 2012. **40**(13): p. e100.
120. Chung, R.H. and C.Y. Kang, *pWGBSSimla: a profile-based whole-genome bisulfite sequencing data simulator incorporating methylation QTLs, allele-specific methylations and differentially methylated regions.* Bioinformatics, 2020. **36**(3): p. 660-665.
121. Choi, J. and H. Chae, *methCancer-gen: a DNA methylome dataset generator for user-specified cancer type based on conditional variational autoencoder.* BMC Bioinformatics, 2020. **21**(1): p. 181.
122. *Sequence Read Archive (SRA)*. Available from: https://www.ncbi.nlm.nih.gov/sra.
123. Blagov, A.V., *Modeling a jitter in telecommunication data networks for studying adequacy of traffic patterns.* Modern Applied Science, 2015. **9**(4).
124. Li, M., *Jitter, noise and signal integrity at high speed*. 2007: Prentice Hall.
125. Perrier, F., et al., *Identifying and correcting epigenetics measurements for systematic sources of variation.* Clin Epigenetics, 2018. **10**: p. 38.
126. Siegmund, K.D., *Statistical approaches for the analysis of DNA methylation microarray data.* Hum Genet, 2011. **129**(6): p. 585-95.
127. Daca-Roszak, P., et al., *Impact of SNPs on methylation readouts by Illumina Infinium HumanMethylation450 BeadChip Array: implications for comparative population studies.* BMC Genomics, 2015. **16**: p. 1003.
128. Horvath, S., *DNA methylation age of human tissues and cell types.* Genome Biol, 2013. **14**(10): p. R115.
129. Naeem, H., et al., *Reducing the risk of false discovery enabling identification of biologically significant genome-wide methylation status using the HumanMethylation450 array.* BMC Genomics, 2014. **15**: p. 51.
130. *Database of Single Nucleotide Polymorphisms (dbSNP)*. Available from: https://www.ncbi.nlm.nih.gov/snp/.
131. Mills, R.E., et al., *Mapping copy number variation by population-scale genome sequencing.* Nature, 2011. **470**(7332): p. 59-65.
132. *Overview of Structural Variation*. Available from: https://www.ncbi.nlm.nih.gov/dbvar/content/overview/.
133. Collins, R.L., et al., *A structural variation reference for medical and population genetics.* Nature, 2020. **581**(7809): p. 444-451.
134. Simon, R., *Roadmap for developing and validating therapeutically relevant genomic classifiers.* J Clin Oncol, 2005. **23**(29): p. 7332-41.
135. Li, C. and M. Li, *GWAsimulator: a rapid whole-genome simulation program.* Bioinformatics, 2008. **24**(1): p. 140-2.

VITA

Neil Andrew Miller was born on August 25, 1970 in Rochester, New York. In 1988 he graduated from Santa Fe High School in Santa Fe, New Mexico. He attained a Bachelor of Arts degree in English from Tufts University in Medford, Massachusetts in 1992.

Mr. Miller held software engineering positions at Genome Therapeutics Corporation and iXL, Inc. before joining the National Center for Genome Resources (NCGR) in 2001 where he held multiple positions, eventually becoming Deputy Director, Software Engineering. In 2006, Mr. Miller led development of a novel computational system for nucleotide variant detection using Next Generation Sequencing for which NCGR was awarded the Bio-IT World Best Practices Award (2009) and was named Laureate, 21st Century Achievement Award of the Computerworld Honors Program (2009).

In 2011, Mr. Miller joined Children's Mercy, Kansas City as a founding member of the Center for Pediatric Genomic Medicine, where he held the position of Director of Bioinformatics until 2021. In this role, he developed bioinformatics methods and novel software for genomic medicine with a focus on the diagnosis of rare disease, cancer genomics and pharmacogenomics. Using these methods, the CPGM collectively developed the STAT-seq program for ultra-rapid whole genome sequencing for patients in the neonatal intensive care unit (NICU) which was voted one of Time Magazine's Top Ten Medical Breakthroughs of 2012. Under Mr. Miller's direction, the bioinformatics team of the CPGM received the HPCWire Readers' and Editors' Choice for Best Use of High-Performance Computing in Life Sciences award (2014), the IDC HPC Innovation and ROI award (2014) and the Datanami Readers' Choice Awards – Top Big Data Achievement (2016).

In May 2021, Mr. Miller joined Bionano Genomics, Inc. as Senior Director, Bioinformatics, where he currently leads the development of the commercial analysis software package and algorithms for analysis of optical genome mapping data and genomic structural variant detection.

In 2016, Mr. Miller began his studies at the University of Missouri – Kansas City as part of the Master of Science in Bioinformatics program. In 2017 he entered the Interdisciplinary Ph.D program at the School of Medicine. He was awarded a Master of Science degree in Bioinformatics in December 2021.

PUBLICATIONS

1. **Miller NA**, Farrow EG, Gibson M, Willig LK, Twist G, Yoo B, Marrs T, Corder S. Krivohlavek L, Walter A, Petrikin JE, Saunders CJ, Thiffault I, Soden SE, Smith LD, Dinwiddie DL, Herd S, Cakici JA, Catreux S, Ruehle M, Kingsmore SF *A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases*. Genome Medicine. 2015;7:100
2. Saunders CJ, **Miller NA**, Soden SE, Saunders CJ, Dinwiddie DL, Noll A, Abu Alnadi N, Andraws N, Patterson M, Krivohlavek L, Fellis J, Humphray S, Saffrey P, Kingsbury Z, Weir JC, Betley J, Grocock RJ, Artman M, Petrikin JE, Heese B, Hall KP, Kingsmore SF. STATseq: *Rapid whole genome sequencing for monogenic disease diagnosis in neonatal intensive care units.* Sci. Transl. Med. 2012 4, 154ra135
3. **Miller NA**, Kingsmore SF, Farmer AD, Langley RJ, Joann Mudge J et al. *Management of high-throughput DNA sequencing projects: Alpheus.* J Comp Sci Syst Biol 2008 1: 132-148
4. Noll AC, **Miller NA**, Smith LD, Yoo B, Fiedler S, Cooley LD, Willig LK, Petrikin JE, Caciki JA, Lesko J, Newton A, Detherage K, Thiffault I, Saunders CJ, Farrow EG, Kingsmore SF *Clinical detection of deletion structural variants in whole-genome sequences*. Nature Genomic Medicine, 2016
5. Mudge J, **Miller NA**, Khrebtukova I, Lindquist IE, May GD, Huntley JJ, Luo S, Zhang L, van Velkinburgh JC, Farmer AD, Lewis S, Beavis WD, Schilkey FD, Virk SM, Black CF, Myers MK, Mader LC, Langley RJ, Utsey JP, Kim RW, Roberts RC, Harlan R, Garcia M, Khalsa SK, Ambriz-Griffith V, Czika W, Martin S, Wolfinger RD, Perrone-Bizzozero N, Schroth GP, Kingsmore SF. *Genomic convergence analysis of schizophrenia: mRNA sequencing reveals altered synaptic vesicular transport in post-mortem cerebellum.* PLoS ONE 2008 3:e3625

6. Weems D, **Miller N**, Garcia-Hernandez M., Huala E., Rhee SY (2004) *Design, implementation and maintenance of a model organism database for Arabidopsis thaliana*. Comparative and Functional Genomics 2004 5(4):362-369
7. Gaedigk A, Whirl-Carrillo M, Pratt VM, **Miller NA**, Klein TE. *PharmVar and the Landscape of Pharmacogenetic Resources.* Clin Pharmacol Ther. 2020;107(1):43-46.
8. Gaedigk A, Twist GP, Farrow EG, Lowry JA, Soden SE, **Miller NA** *In vivo characterization of CYP2D6*12, *29 and *84 using dextromethorphan as a probe drug: a case report.* Pharmacogenomics 2017
9. Twist GP, Gaedigk A, **Miller NA**, Farrow EG, Willig LK, Dinwiddie DL, Petrikin JE, Soden SE, Herd S, Gibson M, Cakici J, Riffel AR, Leeder JS, Dinakarpandian D, Kingsmore SF *Constellation: a tool for rapid automated phenotype assignment of a highly polymorphic pharmacogene, CYP2D6, from whole-genome sequences*. Nature Genomic Medicine, 2016
10. Bell CJ, Dinwiddie DL, **Miller NA**, Hateley SL, Ganusova EE, Mudge J, Langley RJ, Zhang L, Lee CC, Schilkey FD, Sheth V, Woodward JE, Peckham HE, Schroth GP, Kim RW, Kingsmore *Carrier testing for severe childhood recessive diseases by next-generation sequencing* Sci Transl Med 12 January 2011 3:65ra4
11. Baranzini SE, Mudge J, van Velkinburgh JC, Khankhanian P, Khrebtukova I, **Miller NA**, Zhang L, Farmer AD, Bell CJ, Kim RW, May GD, Woodward JE, Caillier SJ, McElroy JP, Gomez R, Pando MJ, Clendenen LE, Ganusova EE, Schilkey FD, Ramaraj T, Khan OA, Huntley JJ, Luo S, Kwok PY, Wu TD, Schroth GP, Oksenberg JR, Hauser SL, Kingsmore SF. *Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis.* Nature 2010 464:1351-6
12. Kim JI, Ju YS, Park H, Kim S, Lee S, Yi JH, Mudge J, **Miller NA**, Hong D, Bell CJ, Kim HS, Chung IS, Lee WC, Lee JS, Seo SH, Yun JY, Woo HN, Lee H, Suh D, Lee S, Kim HJ, Yavartanoo M, Kwak M, Zheng Y, Lee MK, Park H, Kim JY, Gokcumen O, Mills RE, Zaranek AW, Thakuria J, Wu X, Kim RW, Huntley JJ, Luo S, Schroth GP, Wu TD, Kim H, Yang KS, Park WY, Kim H, Church GM, Lee C, Kingsmore SF, Seo JS. *A highly annotated whole genome sequence of a Korean individual*. Nature 2009. 460, 1011–1015. PMID: 19587683

Full bibliography:

https://www.ncbi.nlm.nih.gov/myncbi/12efsp7g_eo5x/bibliography/public/