Title
Management, Integration, and Mining of Tumor Data

Author
Cario, Clinton L

Publication Date
2018

Peer reviewed|Thesis/dissertation

Management, Integration, and Mining of Tumor Data

by

Clinton L. Cario

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biological and Medical Informatics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

[redacted]

Dedication

Foremost, I would like to thank my Dissertation Chair, Dr. John Witte, for the outstanding guidance, support, encouragement, and mentorship over the past 5 years. As a first year student, I was told that one's relationship with their Principal Investigator is probably the single most important component of a successful PhD and should be considered accordingly when choosing a lab. I took this advice to heart and can confidently say joining his lab was the single best academic decision of my PhD. He allowed me to explore my own ideas and to blaze a trail into unfamiliar territory, but gave me direction and guidance when I became a bit too lost in the weeds. I respect John deeply on an academic level, but also on a personal one. His calm, thoughtful approach to challenges and life-balance is something I strive to emulate.

Secondly, I would like to thank my Dissertation Committee for the hours of feedback, advice, and suggestions. To Dr. Pam Paris, without your encouragement, warmth, and knowledge of liquid biopsy, my project would have never gotten off the ground. I will always be grateful that I had you as a friendly face in the audience during my first conference talk as I tried to not fumble over my words. To Dr. Mike Keiser, thank you for your inspiration—early discussions with you and your passion for machine learning convinced me that it was something I needed to do, and it became a major theme of my research as a result. I would have been lost without your guidance. And to Dr. Joe. DeRisi, a person who will light up a room in more ways than one—be it through your humor, intellect, or storytelling; thank you for giving me an example of the type of scientist I'd like to be. Your passion for science and boldness in asking important questions is a very big reason why I came to UCSF in the first place.

I would also like to thank my Program, including Directors Patsy Babbitt and Ryan Hernandez, Administrators Julia Molla, Nicole Flowers, and Rebecca Brown, and all the associated faculty and staff. Your dedication to our program has created a rich and diverse environment where people can truly thrive. I've always felt there was someone to turn to when situations seemed out of control, felt that my academic and personal lives were enriched is ways far beyond what I could have expected from a graduate institution. The dedication to academic and personal success of all students here speaks magnitudes about the quality of this program.

Next, I would like to thank my friends, classmates, and labmates. Of everything that I have learned, and in all the ways that I have grown these last few years, it is most of all through you. Our shared challenges, failures, successes, and fun will serve as the most cherished moments of my PhD. Thank you for the beers, the laughs, and for opening my mind to new science and ideas. I would like to thank Cat and Greg Rybka for being my guides to the bay, for their dependability, advice, and warmth. All of you are awesome and I cannot wait to see how wildly successful you become!

I am most grateful and owe the largest thanks to my family. To my father, Sam, who taught me to persevere, and my mother, Jan, who taught me to believe in myself, this dissertation is dedicated to you. To my sisters, Kristen and Jenna, and brother, Jordan, you are my closest friends and best advisors. Your support gives me the courage to continuously challenge and better myself. To Amanda, my best friend, companion, and inspiration. Your kindness, dedication, and love are the reasons I was able to make it through my PhD. You've guided me through one of the most challenging phases of my life, and I will be eternally grateful for that.

Finally, I would like to thank the patients who provided samples for my study. You've placed trust and faith in me as a researcher and I hope that this dissertation reflects the tremendous efforts I've made in trying to advance science in ways that will benefit you and the community at large. I hope you find me deserving of that trust and faith.

Acknowledgments

A version of Chapter 3 of this dissertation has been published in *Bioinformatics*[1]. The published material is substantially the product of Clinton L Cario's period of study of study at UCSF and was primarily conducted and written by him. The work he completed for this published manuscript is comparable to a standard dissertation chapter.

[1] Cario CL, Witte JS. Orchid: a novel management, annotation and machine learning framework for analyzing cancer mutations. *Bioinformatics*, 2018 March; 34(6):936–942.

Management, Integration, and Mining of Tumor Data

by

Clinton L. Cario

Abstract

Genomics is expected to soon overtake astronomy, particle physics, and even YouTube as the biggest creator of digital information (1). Analysis of this information has already led to important and ground breaking discoveries relevant to our health, but ongoing work will require creative solutions to the multitude of challenges arising from this volume of data. Practically speaking, one such challenge comes from determining what data should be collected and how it is to be managed. As cohort sizes in population based studies grow into the hundreds of thousands, practical issues about collection, storage, and filtering have begun to come more into focus. Additionally, frameworks that seamlessly integrate disparate datasets and also allow for flexible analysis will be required. Finally, as technical challenges and limitations arise, new analytical approaches and designs will have to be considered.

This dissertation work was comprised of three projects relating to these questions as approached from the perspective of a bioinformatician. These projects describe the development of new software and methods for sample management, data integration and analysis, and design strategies to improve signal in noisy data.

The first chapter of this dissertation consists of background material relating to the projects, including a description about the state of prostate cancer genomics, the development of

biomarkers for its detection, and an exploration of a promising new biomarker, cell-free DNA (cfDNA). It also includes a discussion about some of the overarching questions of my PhD.

The second chapter describes a web based sample management system, called Samasy. Born out of necessity, this tool addresses a very practical issue of sample subsetting that is often required of resequencing studies. Samasy was used to facilitate the selection of 16,600 samples from a much larger cohort of 54,000 while preserving ethnicity and age balance among cases and controls. This tool integrates with liquid handling systems and provides a visually intuitive interface for plate/sample management and batch sample transfer execution.

The third chapter details Orchid, a framework designed to make machine learning of cancer variant data easy and extendible. It does so by integrating a variety of biological annotations (or features) and simple somatic tumor data available from large repositories like the The Cancer Genome Atlas (TCGA) or the International Cancer Genome Consortium (ICGC). This tool supports an efficient data store, MemSQL, that allows for very fast retrieval and filtering, and extends the popular python pandas and scikit-learn packages to facilitate machine learning of this data.

Finally, the fourth chapter outlines the creation of a custom targeted sequencing panel for prostate cancer that was designed for screening tumor variants in cfDNA. Building upon the power of Orchid, we detail how machine learning on whole genome prostate tumor datasets can be used to rank mutations by likelihood of being found in a patient with few mutations, or in other words, involved in early state disease. This ranking was used to build a targeted sequencing panel for detection of tumor-derived cfDNA variants. This panel was then validated and applied

to a cohort of nine UCSF prostate cancer patients with multiple tumor foci that were collected at time of Radical Prostatectomy (RP).

Taken together, the information described in this dissertation provides tools and methodologies for the analysis of germline and somatic variants in prostate and other cancers. It also attempts to further technological development of cfDNA as biomarker for the detection or monitoring of diseases like cancer.

Table of Contents

List of Tables

List of Figures

Chapter 1

Background

The State of Prostate Tumor Genetics

Approximately 1 in 7 men are diagnosed with prostate cancer in their lifetime, making it the second leading cause of cancer death in the United States. In 2015, there will be an estimated 220,800 new diagnoses and 27,540 deaths resulting from the disease (2). Although widespread screening with prostate-specific antigen (PSA), digital rectal examination (DRE), and early treatment of localized cancer have improved both detection and 5-year survival rates for early-stage disease, overall survival has not changed significantly over the past 10 years (3). In addition to the human costs associated with the disease, economic burden has been estimated at close to 10 billion dollars per year in the US alone (4).

Prostate cancer is highly heterogeneous, with pronounced variability in both pathology and outcome (5). More than 90% of men are diagnosed with localized or regional disease, for which 5-year survival is close to 100%. Among the remaining men diagnosed with aggressive metastatic disease, prognosis is much poorer (28.2% 5-year survival) (3). Pathological heterogeneity further complicates diagnosis; multiple tumor foci with different histological and genomic features are often found within a single gland, and many same-stage histologically identical tumors result in vastly different course of disease (6). Because of this, there is great incentive to better understand the molecular mechanisms underlying prostate cancer's development and progression, and especially the genetic components that contribute to heterogeneity (7,8).

Moving towards these goals, much work has been done to identify significant somatic mutations that arise during tumorigenesis (9). Biochemical and sequence analyses have revealed some 22 common variants in prostate cancer, including fusions of TMPRSS with ETS family members (10), a potential "gatekeeper gene" NKX3.1 (11), hypermethylation of GSTP1 (12), and copy number (CNVs) or single nucleotide variants (SNVs) in PTEN, P53, AR, CDKN1B, SPOP, and IL-6 (5,13,14). The process of somatic variant discovery is driven primarily by large-scale tumor/normal sequencing efforts which have identified thousands of genetic variants in tumors of all types (15). These sequencing projects are now standard practice, shifting the challenge away from data collection to the task of separating biologically meaningful mutations that drive tumorigenesis ("drivers") from those that are simply artifacts of an unstable and mutated tumor genome ("passengers"). To this end, several statistical and computational algorithms have been developed to find genes with higher than expected mutation rates (16,17), to analyze predicted functional effects of mutations (18-21), and to exploit interaction networks of protein pathways to infer relevant disrupting mutations (22,23). Despite these efforts, results are often inconsistent. Additionally, little has been done to address the somatic mutations falling outside of coding regions that still may have functional importance (e.g., enhancers or promoters) (15). Recently, data repositories like The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) have emerged to provide valuable resources for researchers needing large sample sizes to generate robust signal and who wish to implement more sophisticated tools like machine learning (24,25).

The Potential of cfDNA

With the increasing knowledge of genetic alterations that occur within a tumor, and especially of alterations that are diagnostic, predictive, or associated with actionable therapy,

clinical implications are coming into sharper focus. Assessing the genetic structure of a tumor requires that it be sampled, and traditional biopsy has several unfortunate limitations. First, biopsies are invasive and sometimes difficult to obtain. Second, they represent single snapshots in time, making measurement of tumor dynamics difficult. Finally, biopsies are inherently subject to sample bias and potentially miss tumor heterogeneity (26). One fairly recent and potentially transformative technology has risen that could address all of these concerns: cell-free DNA, or cfDNA, a subcategory within "liquid biopsy".

Cell-free DNA was first discovered in 1948, but its clinical utility in cancer was not fully realized until 1989 when tumor specific alterations were found in circulation (27). Since then, cfDNA has been recognized for its use in a diverse set of clinical scenarios, among them prenatal testing (28-30), measurement of transplantation rejection (31,32), observation of cellular injury (eg. myocardial infarction (33)), and in detection and monitoring of tumors (34,35).

Previous studies have analyzed cfDNA in a variety of cancer types and contexts (36-38). It has been used in screening (39), disease subtyping (40,41), detection of resistance mutations/residual disease (42-44), and real-time monitoring of treatment response (45-47). The bulk of research has attempted to quantify cfDNA levels in correlation with clinical measures (39,48,49) or to look for the presence of known variants within the cfDNA pool (38,50,51).

Several groups have specifically explored cfDNA's usefulness in prostate cancer. In a 2010 review, Ellinger noted more than a dozen studies suggesting it's diagnostic and prognostic potential for prostate cancer (52). Other studies have explored its virtue as a biomarker (53-55), its ability to measure treatment response following chemotherapy (56), and have associated mutations within cfDNA and differential response to treatment in metastatic castration-resistant

prostate cancer patients (57). Compared to other cancers, however, the role of cfDNA in prostate cancer is largely underexplored.

Despite the great promise of cfDNA in oncology, there are a few crucial issues that ground expectations. CfDNA is found at levels consistently detected in most patients, however, a very small fraction of molecules typically originate from the tumor (often < 1.0%) of patients with cancer (34). The issue is further complicated by large variation within tumor type and patient (58). Another issue is identifying variants for which to look. Mutations in driver genes and other variants are not universally present in high proportions of patients with some cancer types. One solution is to ignore specific variants altogether and instead use whole genome sequencing to cast a wide net. Unfortunately, owing to the rarity of tumor fragments in the cfDNA pool, extremely deep sequencing is often needed (>>10,000X), inflating cost of whole genome sequencing (WGS) to intractable levels. Balancing the cost and depth-of-coverage extremes is a hybrid approach whereby a modest number of candidate genes is selected from a list of drivers for mutational profiling (a "selector" or "targeted capture library"). One such approach, called CAPP-Seq, has been successfully developed and employed to broadly detect patient-specific mutations in a non-small-cell lung carcinoma (NSCLC) cohort (59). However, sequencing error and a bias toward capturing only coding regions limit this approach.

Overarching Dissertation Questions

My dissertation work leverages methods of somatic driver mutation identification for the purpose of generating a prostate cancer specific targeted sequencing panel to probe cfDNA, and uses this information to discover tumor mutations within a cohort of UCSF patients. This research has potential for creating diagnostic, prognostic, and predictive tools for use in a clinical setting, and revolves around two core questions:

1. Can novel tumor-specific driver variants be identified for prostate cancer using machine learning models trained on public cancer data (e.g. from ICGC)?
2. In a UCSF prostate cancer patient cohort, can a targeted sequencing panel from discovered driver mutations be used to successfully identify mutations in cfDNA isolated from blood, and how do they compare with sequence results from primary tumor tissue?

Exploring the first question in a bit more in depth, algorithms that attempt to discern driver from passenger mutations fall into three broad categories, focused on finding A) genes with mutations rates above baseline, B) mutations that are predicted to impact protein function, or C) sets of mutations affecting similar biological pathways (15,60) . Examples of machine learning algorithms applied specifically to cancer, such as CHASM (a random forest classifier)(61), ParsSNP (Expectation Maximization/Neural Net) (62), and the methods of Chen et al. and Tan et al. (Support Vector Machines; SVMs) also exist

(24,63). In **Chapter 4**, I propose a SVM-based approach with genome-wide coverage, a richer feature set, and more training data than previous models. The goal was to use this model to rank and prioritize mutations for a targeted sequencing panel. The mathematical framework of a SVM is shown in **Figure 1.1**.

CfDNA from prostate cancer patients was also collected and prepared for sequencing using Unique Molecular Identifiers (UMIs) to reduce downstream sequencing noise and improve detection power. The panel was then applied to these samples and processed using standard workflows to assess performance of the capture library by looking at variant coverage across patients (i.e., number of detectable variants per patient).

Figure 1.1: A Support Vector Machine (SVM)

Legend: A SVM finds the optimal hyperplane (defined by normal vector **w** and intercept **b**) that maximizes the margin between the hyperplane and classes of data points (M). Here, Whole Genome Sequence data from prostate cancer patients in the ICGC data repository were used for training. The classification features **f**, encoded by the vector **x**, included genome annotations gathered from various biological databases on the Internet.

$$M = \{[\vec{x_1}, y_1], [\vec{x_2}, y_2], ..., [\vec{x_n}, y_n]\} \quad (1)$$

$$x_i = [f_{i,1}, f_{i,2}, ..., f_{i,m}] \quad (2)$$

$$y_i = \begin{cases} +1, \text{if } \vec{x_i} \in M_{cancer} \\ -1, \text{if } \vec{x_i} \in M_{simulated} \end{cases} \quad (3)$$

$$\vec{w} \cdot \vec{x_+} + b \geq +1 \text{ for } y_i = +1 \quad (4)$$

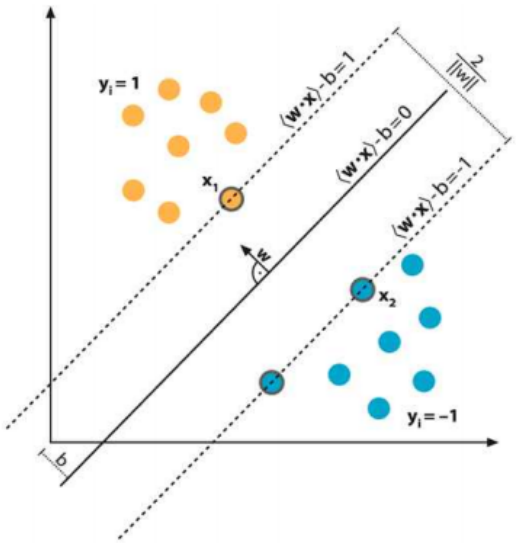
$$\vec{w} \cdot \vec{x_-} + b \leq -1 \text{ for } y_i = -1 \quad (5)$$

$$y_i(\vec{w} \cdot \vec{x} + b) + -1 \geq 0 \quad (6)$$

Goal:

$$max\left(\frac{2}{\|\vec{w}\|}\right) = min\left(2\|\vec{w}\|^2\right) \quad (7)$$

[Lagrangian optimization constrained to eq. 6]



Example : A support vector machine with two features (axes) and a decision boundary $(\vec{w} \cdot \vec{x}) - b = 0$ to classify points.

Chapter 2

Samasy: An Automated System for Sample Selection and Robotic Transfer.

Introduction

Over the past few years the number of samples included in genomic and other 'omic projects has exponentially increased. Such growth allows for teasing apart the complexities of polygenic disease and the effects of less common measures (e.g., rare variants) (64-66). For example, large-scale cohort studies such as the UK Biobank now include genetic and phenotype information on 500,000 individuals (REF), and other enormous cohorts are underway or planned. As another example, the number of samples included in genome wide association studies (GWAS) of schizophrenia has increased more than 10 fold in just 5 years via consortia efforts, which has resulted in the discovery of more than 100 novel genetic associations (67). This trend in increasing numbers of samples is expected to continue into the foreseeable future, raising practical considerations regarding how to effectively and efficiently scale projects.

One such consideration is that of sample management, which includes the storage, transfer, and tracking of samples (68). To address storage, a common, affordable, and convenient solution is the 96-well plate. This format also facilitates sample transfer through a grid-like layout. However, for large cohorts sample transfer is both laborious and error prone unless automated liquid handling robots like the Beckman Biomek is available and utilized (69,70). As far as record-keeping, many research groups depend on spreadsheets (e.g. Microsoft Excel), which work in principal but quickly become unwieldy at large scale and are difficult to use for tracking the transfer of thousands of samples across various batches. Addressing the need for information technology in a laboratory setting, many Laboratory Information Management Systems (LIMS) have been developed, and some have achieved a comprehensive scope in capability (71,72).

However, these systems are growing in complexity, tending to become harder to install, configure, and use. In response, some simple and easy-to-use, but niche, web-based software tools have been developed for high throughput sample processing (73,74). Standing in contrast to both of these system design paradigms, Samasy was developed with the goal of being both easy-to-use and general in applicability.

As a case-in-point, we undertook a large-scale re-genotyping project that required subsetting 16,600 samples from a cohort of 54,000 samples in a manner that preserved ethnicity and age balance among prostate cancer cases and controls (Emami, *et. al.* in preparation). To address the aforementioned sample management issues, and because sample transfer was originally planned to be done by hand, we developed Samasy, a visual web-based sample management system for 96 well plates with support for automated robotic transfer. By Integrating Samasy with a Beckman Biomek, we were able to reduce the number of technician hands-on hours approximately 10-fold (Eunice Wan, University of California Institute for Human Genetics Genomics Core Facility, personal communication) while also decreasing the likelihood of numerous sample transfer errors (e.g. pipetting volume, well/plate mix-up, misspecified robotic well assignments, operator error, etc.). To verify the validity of Samasy's database, batch algorithm, and robotic transfer file generation feature, we performed a sample transfer demonstration using 6 'Source' plates, 1 'Control' plate, and 2 'Destination' plates with colored wells, and observed successful volume transfer with no errors (**Supplemental Figure 2.1**).

Results

We designed Samasy in a user-friendly manner, making installation simple in as little as 3 steps, import of large datasets straight-forward through a drag and drop interface, and tracking of samples visually intuitive. Samasy can visualize plate and sample data several different ways: 1)

as a searchable, sortable, and filterable table; 2) conceptually in a 96-well plate format with samples color coded based on user defined attributes (e.g. age, ethnicity); 3) in batch mode, reflecting plate and sample layout on a transfer platform; or 4) as grouped attribute histograms showing sample distribution per plate or batch. Additionally, sample transfer batches can be imported, viewed, and used to generate sample transfer files for automated systems (a Biomek sample transfer script is included). Samasy also provides a convenient REST API for data access, allowing integration for other uses, transfer systems, and customizations should the user find it necessary.

Samasy is implemented on a modified LAMP (Linux/ Apache/ MySQL/ PHP) software stack, consisting of Ubuntu, a Nginx web server, a SQLite3 database, and the Ruby-based Sinatra web framework. Only Ruby and Sinatra are required; Samasy code can run on other operating systems, with or without Nginx, and use any storage database. It is capable of running on commodity hardware and requires very little memory (less than 512Mb).

Samasy can be installed in as little as three commands in a terminal (see code repository at https://github.com/wittelab/samasy), and data files are easily imported using a drag-and-drop wizard that starts automatically when the interface is first launched. The wizard will also check the integrity of each data file and provide help screens describing file formats. Afterward, the wizard creates an administration account, which allows a privileged user to import batch information and destroy data within the database. The administration user can also create, modify, or destroy additional administration or unprivileged accounts.

To import data into the interface, several types of headered tab-delimited flat files (exportable by Excel) are used. The original dataset file should consist of three required columns

SampleID, PlateID, and Well, and any number of user-defined sample parameters (e.g., age, ethnicity). SampleID and PlateID columns can be any alphanumeric text but the Well should be specified like 'C7'. The second type of flat file is the data dictionary, which defines possible sample values and maps coded data (if exists) to corresponding real labels (e.g., 0 to 'male', 1 to 'female'). This file has three required columns, Attribute, Code, and Value, where Attribute specifies the corresponding column header in the data file, Code the encoded value for a sample (e.g., 0), and Value the actual sample label (e.g., 'male'). Finally, the batch file(s), used to specify well transfer mappings, can be provided. There are five required columns for this file, BatchID, Source Plate, Source Well, Destination Plate, Destination Well, and one optional column Volume. The interface itself will produce a robot file, which is similar to a batch file but readable by sample automation platforms. For Biomek machines, the platform layout and transfer script is provided transfer script in the biomek folder. Example files are located in the example folder in the code repository. Samasy is capable of guessing the datatypes of attributes, first trying numeric types (integers or floats) before defaulting to character strings.

Once data has been imported and a user has logged in, plate, sample, and batch information can be visualized. A Views drop-down menu switches between Samples (**Figure 2.1**), Plate (**Figure 2.2**), Batch (**Figure 2.3**), and Distribution (**Figure 2.4**) view modes. In the Plate and Batch views, there is an additional Color By drop-down menu that allows sample wells to be color coded by sample parameters. Search and Well Legend panels are shown to the left of these two views. In the Batch view, robot files can be downloaded and batches can be marked as completed, updating sample locations in the database. Robot files can be uploaded to the biomek machine and used to transfer samples when plates are loaded as shown in the web interface and as specified in the biomek transfer file (see biomek folder). Finally, in the Distribution view,

sample attributes across a plate or batch can be viewed as histograms, with the option of grouping sample by another attribute (e.g., to view the distribution of sex after grouping by ethnicity). Distributions can use ordinal or numeric data, and numeric data will be binned if a large range of values exists.

Discussion

While Samasy is designed with broad use in mind, it currently only supports the 96-well plate format and many studies store samples in denser plate formats, like the 384-well plate. Despite this limitation, many studies--including those undertaking DNA genotyping or sequencing--require volumes not amicable to these denser formats. Additionally, archival and retrospective samples are often stored in 96-well plates.

In conclusion, we have developed Samasy, a simple-to-use and intuitive web-based application to improve and optimize sample management, visualization, and transfer encountered by large-scale studies utilizing 96-well plates. Starting with only sample files and (optionally) data dictionaries, Samasy will generate a database to visualize plates and samples across provided parameters. Samasy will also accept sample transfer batches, provide batch views, transfer information and history, and 'robot files' to perform automated sample transfer with robotic platforms. We believe this application will serve as a useful tool for future studies requiring large-scale sample management, especially those involving genotyping or re-sequencing.

Figure 2.1: Samasy sample table view showing sortable and searchable sample attributes.

Legend: To quickly find sample information, a search box is provided that will fuzzy match values in any attribute field. Additionally, attributes can be sorted ascendingly or descendingly.

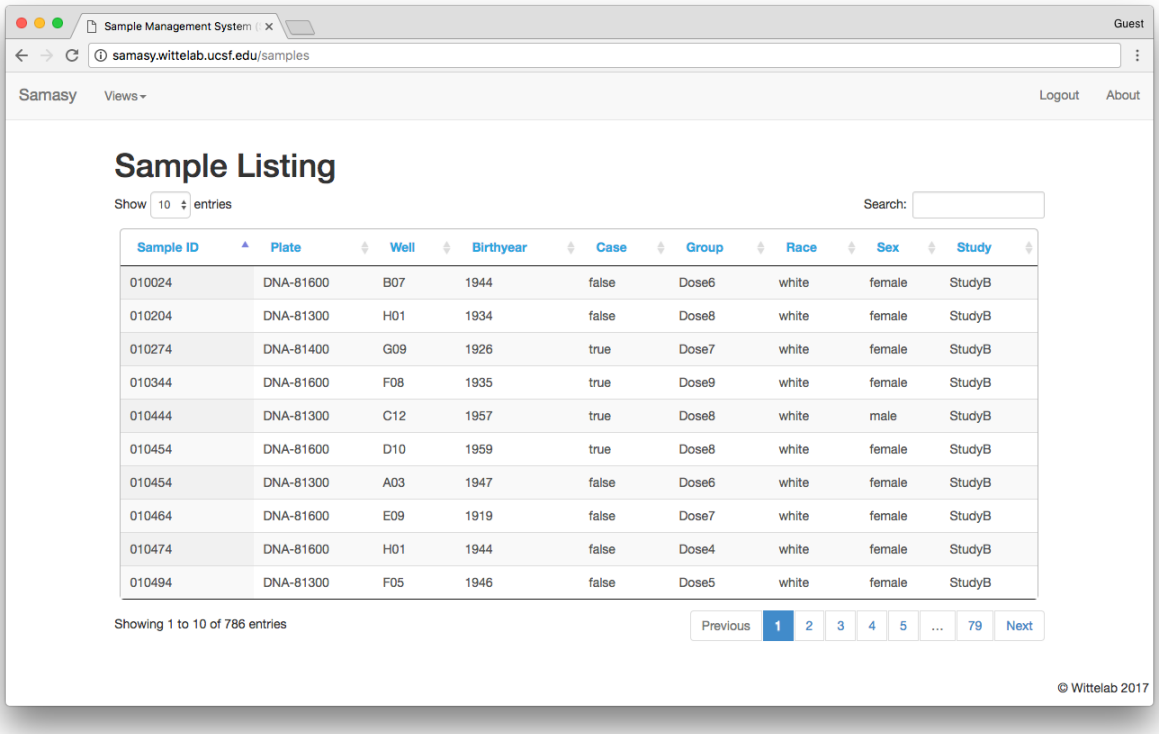


Figure 2.2: Plate view that shows sample locations within a 96 well plate.

Legend: Wells can be clicked to access sample attribute data from the database, including to which batches they belong. Wells are color coded by an attribute of interest using the interface menu.

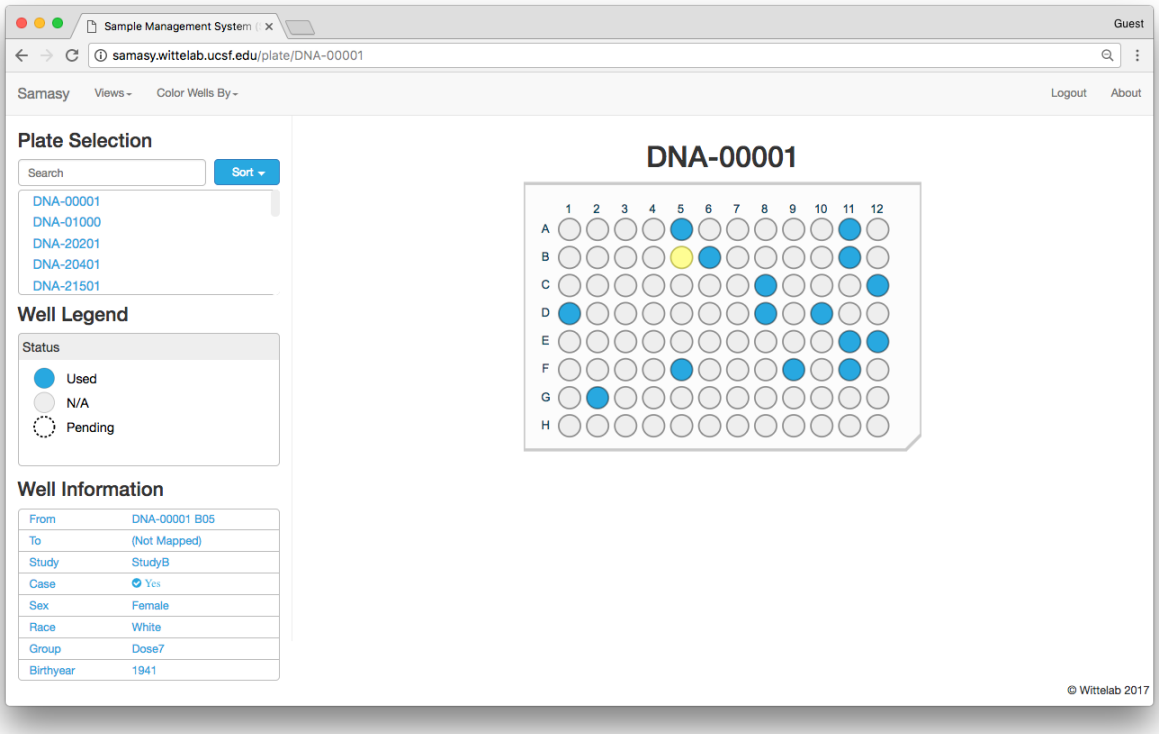


Figure 2.3: Samasy batch view.

Legend: Plates within a batch are shown in a layout reflecting a robotic transform staging area, and plates are color coded by 'source' and 'destination' status. Wells are colored according to the attribute of interest, and well border indicates whether a sample transfer through this batch has occurred.

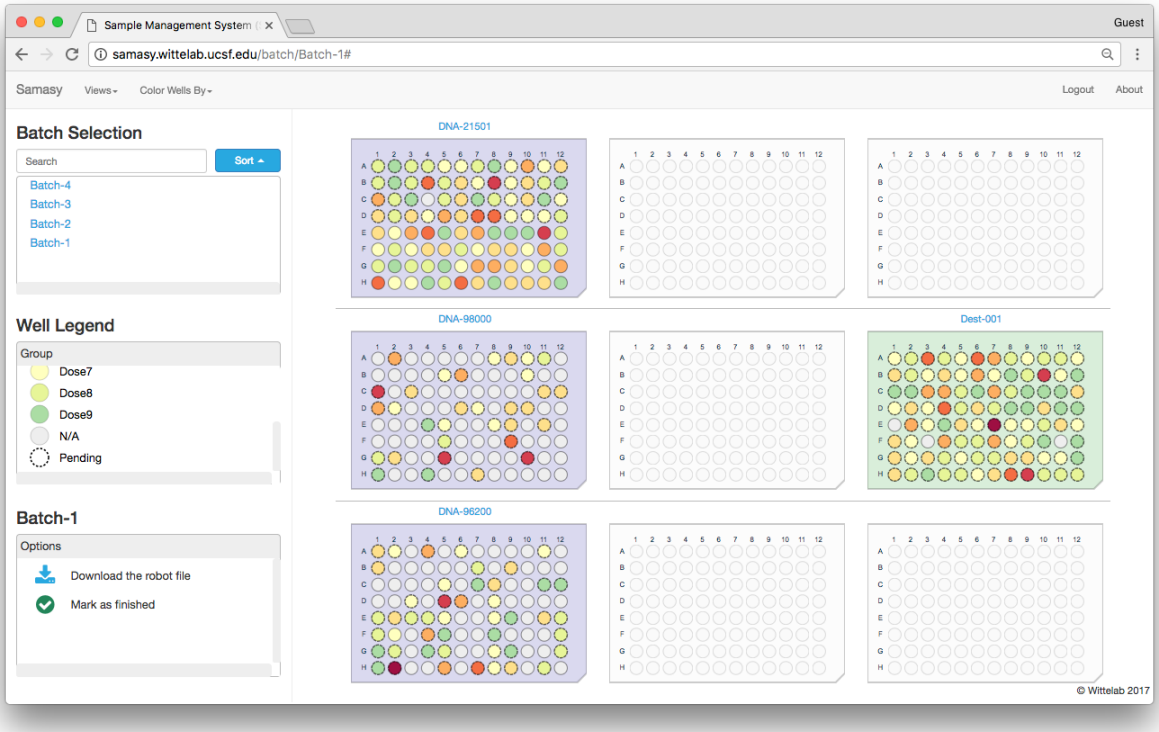
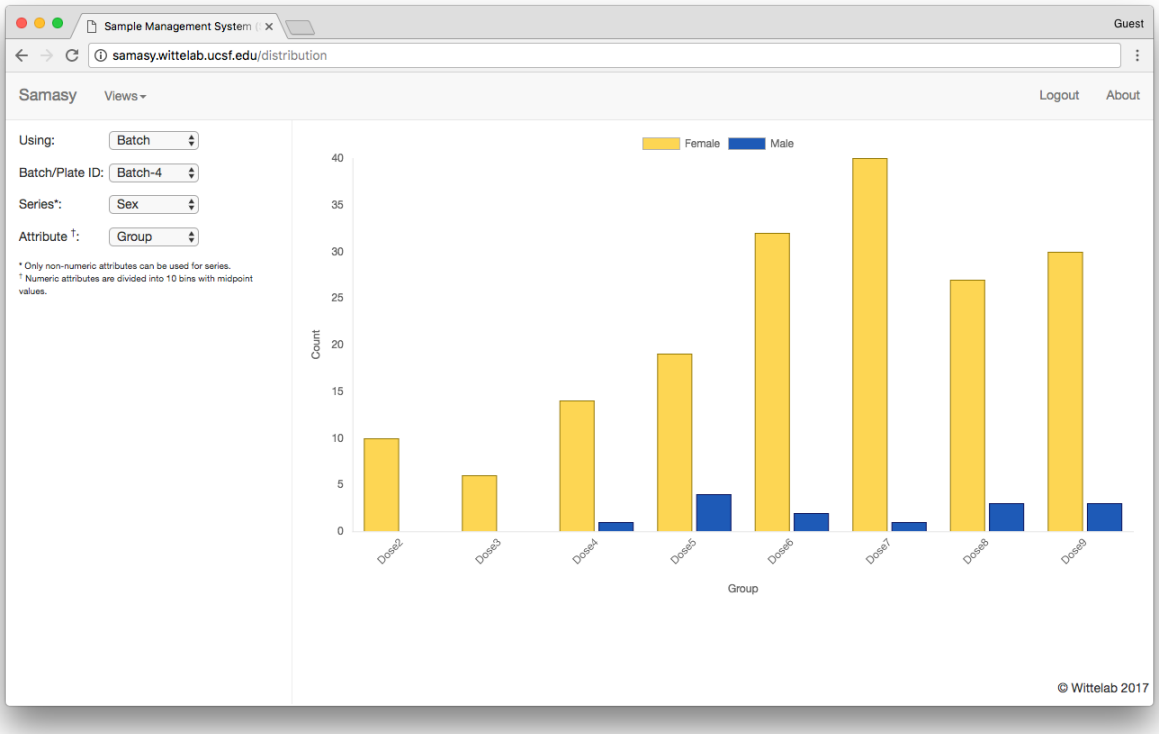


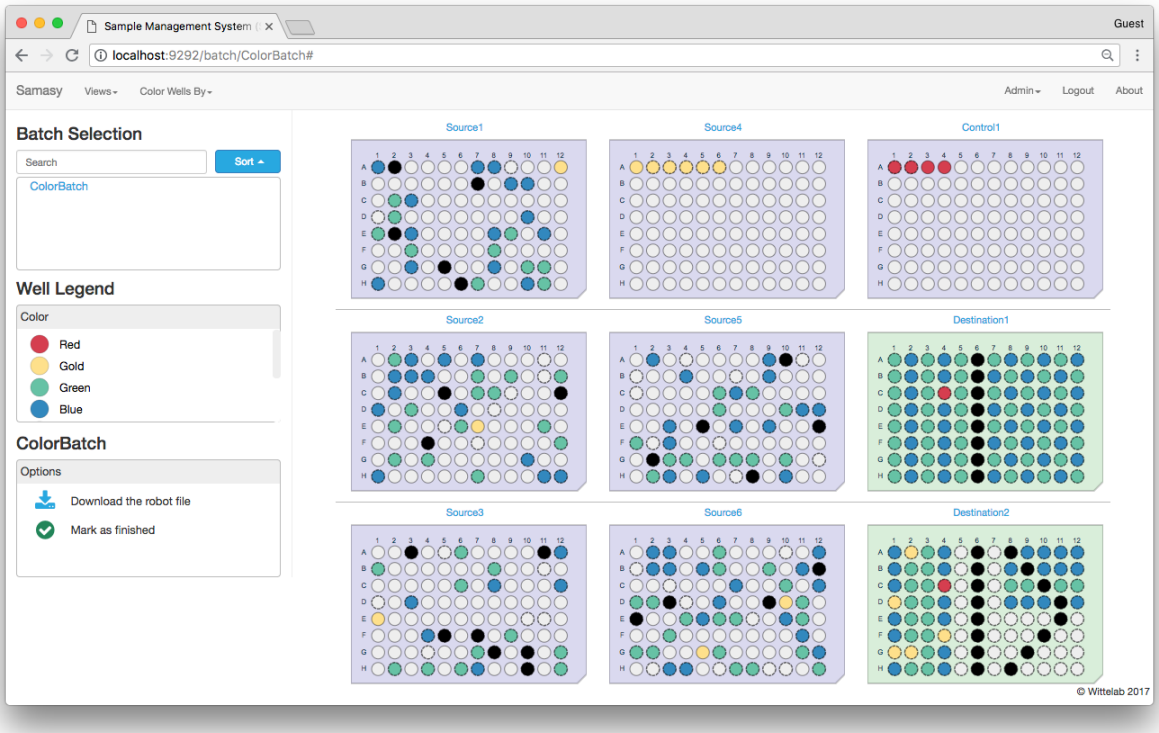
Figure 2.4: Samasy distribution view.

Legend: Sample attribute distributions can be displayed as histograms by plate or batch. If desired samples can also be grouped as series for distribution comparison between series class.



Supplemental Figure 2.1: Sample transfer verification.

Legend: A sample transfer demonstration was performed using 6 sample 'Source' plates, 1 'Control' plate, and 2 sample 'Destination' plates with Samasy's batch import and view features. Source samples consisted of water dyed with blue, black, gold, and green food coloring that were randomly (non-randomly for plate 4) plated on 'Source' plates. An additional 'Control' plate consisting of water samples dyed red was also included. A batch file was imported to map 'Source' and 'Control' samples to 'Destination' plates to produce the patterns indicated in the figure. Using the Samasy interface, the robotic transfer file was then generated and executed by a Beckman Biomek liquid handler, resulting in successful sample transfer with no errors.



Chapter 3

Orchid: a novel management, annotation and machine learning framework for analyzing cancer mutations

Introduction

Cancer is a complicated disease driven largely by genomic alterations. To better understand and characterize the genetic architecture underlying carcinogenesis, thousands of tumor genomes have been sequenced. This work has detected a large number of somatic mutations, gleaning meaningful biological insight such as identification of functional driver mutations in dozens of genes like KRAS, APC, P53, PI3K, SMAD4 (9) that are involved in many cancers.

A key challenge in the analysis of tumor genomes is how to interpret mutations with uncertain function. This is further complicated by the fact that many mutations may have no relevant function, but arise simply as artifacts of an unstable and mutated tumor genome (i.e., as passengers). To address these issues, several statistical and computational algorithms have been developed that attempt to prioritize or annotate mutations by finding genes with higher than expected mutation rates (16,17), by analyzing predicted functional effects of mutations (19) (20), and by exploiting interaction networks of protein pathways to infer relevant disrupting mutations (22,23).

Recently a new class of methods inspired by machine learning paradigms have emerged that determine ‘deleteriousness’ of mutations in both general and cancer-specific contexts. These consist of models trained in evolutionary conservation (75,76), protein sequence, domain and/or structural information(18,21), and parsimonious analysis of a broad range of tumor datasets (62).

Despite successful applications, there remain limitations to such statistical and machine learning approaches. For one, few methods annotate or score mutations that fall outside of coding regions despite the known regulatory importance of many intergenic bodies (e.g., enhancers, promoters, transcription factor binding sites, or microRNAs)(15). There are also issues with the collection, parsing, and integration of tumor and annotation information that is scattered across dozens of databases and in a variety of formats, many of which are not suited for high-throughput analysis. Finally, methods that can score variants at a base-level resolution tend to be general in what they predict (e.g., evolutionary conservation), making more refined predictions difficult (e.g., likelihood of being a prostate cancer driver mutation).

To address some of these issues, we developed *orchid*, an open source tumor mutation management and machine learning analysis framework. Orchid makes the management, annotation, and analysis of tumor mutations more programmatically elegant and computationally efficient by integrating mutation data with popular databases and python-based numeric and machine learning frameworks. Orchid is capable of accepting a wide assortment of feature types and is agnostic to the desired classification task, making it easy to build a variety of models quickly. Furthermore, it accepts and annotates mutations from any region of the genome, allowing for the analysis of non-coding mutations.

To demonstrate orchid, we applied it to the task of inferring cancer tissue-of-origin based upon copy number information and simple somatic mutations found in the genomes of 12 tumor types. This application highlights the value of orchid in generating models that can potentially be used in the diagnosis of metastatic tumors from which primary tumor cannot be located (called "cancers of unknown primary" or CUPS), which represent 2-4% of all cancers(77), or in identifying tissue-of-origin from mutations found within cell-free DNA (cfDNA).

Methods

We created an open-source mutation management and modeling software package called orchid which consists of the *orchid-db* script for loading and annotating mutations into a MySQL-like database system, and *orchid-ml*, a python module that interfaces with the popular python numeric analysis library, pandas (http://pandas.pydata.org/), and with scikit-learn (http://scikit-learn.org/), a python framework for machine learning. Orchid has the ability to parse raw data in various common formats and can be used to generate annotated tumor mutational databases and models in as little as ten lines of code. A diagram of the orchid workflow is shown in (**Figure 3.1**).

Orchid-db

To build a tumor mutational database for subsequent supervised machine learning tasks, we first downloaded and collected raw tumor variants calls, copy number information, and metadata for multiple tumor types as well as variant annotation data from several biological databases. For tumor data, we choose to make use of the International Cancer Genome Consortium (ICGC) given its extensive collection of tumor data across dozens of studies and tissue types. For annotation data, we hand selected biological features to represent a broad range of functional genomic annotations, prioritizing genome-wide annotation datasets when available. Data was populated into a MySQL or MemSQL database (**Supplemental Figure 3.5**) using the orchid software running on either a 2013 MacPro (OSX Sierra) or a PBS Cluster (Red Hat Linux v6.6). A MemSQL version of this database is available for public use; please see https://github.com/wittelab/orchid.

From the ICGC data portal, we selected patients from release 25 with genome wide simple somatic mutation and/or copy number tumor data that were publically available (non

PCAWG), for a total of 3,604 individuals. We then excluded outliers by removing those individuals whose tumors had less than 10 or more than 30,000 mutations. Finally, we excluded tissues with fewer than 80 tumors and randomly sampled 80 from the remaining, resulting in 960 tumors from twelve cancer types (Bladder, Blood, Bone, Brain, Breast, Esophagus, Head and neck, Pancreas, Prostate, Skin, Stomach, and Uterus). In total, 3,489,978 mutations were populated into the database with tissue means ranging from 281 for Bladder to 15,202 for Esophagus (**Supplemental Figure 3.1**). Conceptually, we grouped this data into two levels of specificity for analysis: 1) the mutational level -- where real mutations found within patients of a single tumor type are compared to mutations that might occur by chance through careful simulation; and 2) the tumoral level -- where real mutations from patients (either of the same tumor type of different tumors) are compared to each other. Possible classification outcomes (i.e. labels) are 'observed' (or 'real') and 'simulated' for mutational level classifications, and any patient-level stratifier (e.g. tumor tissue-of-origin, stage, aggressiveness) for the tumoral level. The tissue-of-origin application presented in this paper represents classification of tumoral-level data.

We collected, downloaded, and curated mutational annotation from 15 biological databases and annotation tools (**Supplemental Table 3.1**; http://wittelab.ucsf.edu/orchid). These features include functional annotation (*SnpEff*, (78)); cancer gene network presence (*KEGG*, (79)); phylogenetic conservation (*phylop*, (80)); location within snoRNA and microRNA regions (*wgrna*, (81)); locations within predicted enhancers, promoters, and transcription start sites (*segmentation*, (82, 83); *rfecs*, (84); *dbsuper*, (85); *encode*, (86)); locations within DNAse I hypersensitivity sites (*dnase*, (87)); trinucleotide contexts (88); assorted composite scores

(*funseq2*, (89); *cadd*, (75); *dann*, (76)); and various other measures (*targetscans*, (90); *remap*, (91); *gwas*, (66)).

There are a wide variety of suitable biological annotations and file types that can serve as mutation features. We therefore designed our annotation software tool to be flexible in the file formats it accepts. Features in tabix, bed, or wig file formats can be added to the modeling process with no modification while other formats can be integrated with minimal effort—namely by converting to bed format or by providing orchid with UNIX awk commands to pre- and post-parse feature lookup data.

All annotation and mutational data is based on genome coordinates from human reference sequence version GRCh37 (hg19). Data should be in the same coordinate system for database population. For convenience, GRCh38 (hg38) coordinates are also provided in our publically accessible database.

Orchid-ml

Orchid-ml exists as a standalone python module that can be imported into any python script. To load, transform, model, and visualize mutation data, we designed the *MutationMatrix* object, an extension of the pandas *DataFrame* object. Transformations of the data include loading, encoding, imputing, feature scaling, and feature selection. Modeling consists of selecting a prediction label and running orchid's built-in support vector machine (SVM) or random forest (RF) wrapper functions or any of the scikit-learn classifiers. Finally, visualization produces ROC curves, confusion matrices, and other performance metric tables.

We implemented several functions to load and encode data as a *MutationMatrix*. The first, *load_mutations()*, will take a MySQL connection string for a database populated by orchid-

db and load all (or a desired subset of) mutations and their basic associated metadata (chromosome, position, donor_id, sequence, etc.). The second, *load_features()*, will import all (or a desired subset of) annotation features. Finally, *encode()* will transform categorical features into numeric values so they can be properly modeled. This is accomplished through the specification of encoding strategies given as a dictionary to the function (strategies={feature: strategy}). Choices for strategy are 'one-hot', 'binary', 'label', or 'rarity' (i.e. a feature value's frequency). Alternatively, or if not specified, orchid will use a one-hot encoder.

In some situations, it may be desirable to aggregate mutational level data to the tumoral level to compare tumor mutational profiles with each other. For this purpose, we created the *collapse()* function to aggregate feature values within each patient (i.e. tumor) using feature median or mean values. In practice this is can be done with any grouping column by passing the column name as the 'by' parameter. Collapsing should be performed after encoding has occurred but before normalization.

Most machine learning algorithms require numeric, non-missing, feature-scaled data for effective learning. With orchid-ml, one can specify strategies for imputation and scaling using the *set_normalize_options()* function which takes parameters 'nan_strat' and 'scaler_strat' to respective missing and scaling strategies. Imputation strategies include setting all unknown feature values to 0 ('zero'), or to the feature mean ('mean'), median ('median'), or most frequent ('most_frequent') values. Feature scaling is performed using a min-max scaler ('mms'), where feature values are transformed to a [0, 1] range based on the minimum and maximum values or a z-score based method ('standard'), where feature values are subtracted by their mean and divided by their unit variance. We used orchid-ml's default values for normalization, 'median' and 'standard', unless otherwise stated.

Classifiers with a large numbers of features can potentially begin to model noise specific to the training dataset (a.k.a. overfitting), which decreases overall performance and classification generalizability. To avoid this pitfall, we employ a feature selection method that reduces feature number to a desired subset size—generally one-tenth the number of training examples. This is accomplished through orchid-ml's *select_features()* function. This function normalizes, shuffles, and divides data into training and testing sets in a 75:25 ratio. Next, it trains a user-specified model (or by default a random forest) with training data and accuracy is assessed in test data. Then, for each feature, it shuffles the feature values, remodels the data, and then compares the resulting accuracy to the original model to generate an error percentage for that feature. It repeats this process 50 times and reports mean error percentages for each feature. The specified top number of features whose permutation caused the largest decrease in model accuracy are retained for subsequent modeling.

To model tumor data, orchid-ml first requires a label column to be set with the *set_label_column(column_name)* function. This flags one of the data columns in the *MutationMatrix* for use as class labels during supervised learning and test prediction. Orchid-ml can then perform modeling with the *svm()* or *random_forest()* function, which interface with scikit-learn's sklearn.ensemble.RandomForestClassifier and sklearn.svm.SVC modules, respectively. The Mutation Matrix() is also compatible with other sk-learn classifiers. For random forest models, we set default values of max_features='auto', max_depth=None, min_samples_split=2, and min_samples_leaf=1. For support vector machine models, we set the kernel default to 'linear', C=1.0 and probability=True. Orchid-ml uses default scikit-learn values for all other parameters, but a user can pass custom sklearn parameter value pairs through orchid. To estimate model stability, orchid performs k-fold cross-validation (k=10 by default) and

reports mean accuracy and standard deviation. Optionally, it will also permute class labels, remodel data, and report accuracy for comparison with a null model, which has an expected accuracy equivalent to randomly guessing a class (that is 1/C where C is the number of classes). This 'sanity check' helps ensure no systematic bias—such as large class imbalance—is falsely contributing to classification accuracy. Modeling can also be performed with custom train/test sizes by specifying the proportion of samples to withhold for testing.

Orchid-ml includes several functions for visualizing data and reporting model performance. These functions depend on the python modules seaborn (https://seaborn.pydata.org/), matplotlib (https://matplotlib.org/), and sci-kit learn. Orchid has the ability to generate dendrograms of mutations clustered by both feature and sample in the form of the *show_dendrogram()* function and can also easily generate performance metric reports (*print_report()*); display confusion matrices (*show_confusion_matrix()*); draw violin plots to compare classification probability distributions (*show_confidence_plot()*; **Supplemental Figure 3.3**); show Receiver Operating Characteristic (ROC) or Precision-Recall (PR) curves (*show_curves();* **Supplemental Figure 3.4**); and indicate feature importance for classification (*show_feature_importances()*). The orchid code repository provides further documentation on each of these.

Application of Orchid: Tissue of Origin Dataset

We first downloaded whole-genome sequencing data from ICGC and biological annotation features as described, populating data into the multi25_20170710 database (see repository). Next, we used orchid-ml to load mutations and features, encode ordinal features, and collapse mutations by patient tumor using mean feature values (this is accomplished with orchid_ml's *load_mutations()*, *load_features()*, *encode()*, and *collapse()* functions respectively).

This resulted in a total 960 tumor tissue profiles. From these profiles, we imputed missing data with a 'median' strategy, normalized the entire matrix with the 'mms' min/max transformation, and selected the 20 most-performant features as described. Model performance was then assessed with 10-fold cross validation and label permutation. Finally, a predictive model was generated with 65% of the data, and tissue predictions were made in the remaining 35%.

Results

Orchid

To facilitate the task of machine learning on tumor mutational profiles, we created orchid, an open-source software framework to efficiently annotate, manage, and model tumor mutations on a genome-wide scale. A user can begin with mutational data from ICGC or in VCF format and annotation feature data in various formats, and then use orchid to import, manage, annotate, and model data. Orchid is divided into two components, *orchid-db*, which loads and annotates mutations into a database system (e.g. MySQL or MemSQL), and *orchid-ml*, a python module that facilitates machine learning using the popular scikit-learn framework.

Orchid-db

We designed orchid-db to efficiently process, parse, and transform raw data into a structured MySQL-like database to maximize subsequent access speed and analysis. Mutation and feature data can be imported individually using two orchid-db subcommands (*populate* and *annotate*), or simultaneously, in parallel, with the workflow management tool, nextflow (v. 0.17; https://www.nextflow.io) (92). For the latter option, we provide the *make_database* shell script to control nextflow execution through a single configuration file that specifies data locations and processing options. Nextflow is capable of executing seamlessly on a desktop machine, on a

cluster, or in cloud environments (Amazon, DNANexus, Docker, Singularity, Apache Ignite, PBS, SGE, SLURM), and can interface with a local or remote SQL-like database system. Orchid specifically supports software compatibility with a performant, distributed, in-memory database system, MemSQL (http://www.memsql.com/) making the import of tens of millions of mutations and hundreds of feature annotations possible on the order of a few hours (see supplemental materials for a details).'

Orchid-ml

We also developed orchid-ml, a python module that interfaces with orchid-db data and provides convenience functions for machine learning of tumor variant data. Our module extends the pandas *DataFrame* class object into a *MutationMatrix* that adds support for importing, encoding, and subsetting tumor mutation data from the database produced by orchid-db. Our modeling functions use the scikit-learn framework for machine learning due to its flexibility, excellent documentation, and large variety of algorithms. Additionally, orchid-ml is capable of visualizing data and model performance that generate plots with seaborn and matplotlib.

Once populated in the database by orchid-db, data is easily accessed and modeled with orchid-ml. A typical workflow is summarized as follows:

1. Specify access to the database generated by orchid-db with a SQL connection string.
2. Load mutations and features either in their entirety or by a desired subset (e.g., by tumor).
3. Encode categorical features using default or user-defined strategies (e.g., one-hot).
4. Optionally collapse mutations by tumor (e.g., by averaging).
5. Set a prediction label and select features.
6. Model data with any of the scikit-learn machine learning algorithms.

For convenience, random forest and support vector machine functionality is built directly into orchid-ml, automatically performing data normalization, train/test splitting, cross-validation, and label permutation for null model generation. Orchid-ml visualization functions can be used to assess performance and explore relationships within the data; these include mutation dendrograms, feature weight boxplots, class prediction and confusion matrix heatmaps, receiver operating characteristic (ROC) curves, and Precision Recall (PR) curves.

Application of Orchid: Tissue of Origin

To demonstrate orchid's ability to facilitate machine learning with biologically relevant classification tasks, we applied a classification model used to determine tissue-of-origin from 12 tissues. The code for this task is provided as a jupyter notebook (http://jupyter.org/) in the orchid software repository.

For this application, we prepared data as described in Methods and randomly sampled 100 tumor profiles for visualization with orchid-ml's *show_dendrogram()* function, using complete linkage hierarchical clustering on both features and tissues. This was done to assess segregation of tumor profiles by feature groups, and to see if patterns emerged that correspond to biology of underlying tissue type (**Figure 3.2**).

From this we observed a small amount of tissue level grouping with particular feature combinations differentially driving segregation. For example, cancers of the stomach, uterus, head and neck, and bladder appeared to show increased mutation burden in transcribed regions (**Figure 3.2A**), and conversely lower mutation burden in repressed regions (**Figure 3.2E**) of encode cell lines. For head and neck cancer, mutations were of higher frequency and enriched in both the G [T>A] G trinucleotide context and 3'UTR regions (**Figure 3.2B**). Upon cross

referencing the ICGC data portal, we noted the very common chr7:140453136 C [A>T] C mutation in the 3'UTR region of BRAF, representing 44% of patients, as potentially driving this signal. Finally, when considering the V, D, J and C immunoglobulin biotype features, separation of blood cancers was observed (**Figure 3.2C and 3.2D**). We also clustered the full 960 tissues on just tissue-of-origin an observed similar patterns (**Supplemental Figure 3.2**), as well as additional trinucleotide signatures that corresponded to those previously reported by Alexandrov *et.al* (88).

Next, we modeled these profiles using a random forest classifier. First, we first employed feature selection to guard against overfitting by reducing the number of features from 339 to 20 using the permutation method described in Methods. Of the retained features, their permutation caused an increase of between 2.5% and 5.2% in classification error. Ten of the twenty most important features were trinucleotide context features, four were transcript biotypes, and two were related to cancer pathways. The remaining retained features were the modifier impact category, Nhlf enhancer, HeLa-S3 transcription, and CADD. From this reduced dataset, we performed 10-fold cross validation with a random forest classifier using orchid_ml's *random_forest()* function. The resulting models had a mean accuracy for tissue classification of 0.94 +/- 0.02. To help ensure systematic artifacts such as class label imbalance were not driving signal accuracy, orchid was used to re-train the models after permuting training labels, and the expected null performance was observed (accuracy = 0.08 ± 0.09; expected = 0.08). Finally, the *random_forest()* function was called to build a predictive model using a randomly subset population of patients (n=624; 65%), while the remaining were withheld for testing (n=336; 35%). For this final model, we plotted the feature weights on a per-tissue basis (**Figure 3.3A**), showing that several features were particularly useful for classifying just one of the tissue types

(e.g. IG variable segment for blood and many of the C>T trinucleotide context features). We also used orchid-ml's *show_curves()* function to produce ROC curves in a one-vs-rest fashion for each tissue (**Figure 3.3B**). Tissues have an AUC range between 0.80 (brain) and 0.98 (bone).

To observe whether consistent tissue misclassifications were present, we generated a confusion matrix using orchid-ml's *show_confusion_matrix()* function (**Figure 3.3C**). For this analysis, we assigned each tumor profile the tissue with the highest predictive probability and compared the predicted tissues with their actual types. Tissues most often confused as others (False Negative Rate) include pancreas, prostate, and uterus, while bone, head and neck, and stomach were rarely confused. Likewise, tissues were often confused as prostate, brain, and breast (False Discovery Rate), but not as often as blood, skin, uterus, esophagus. Interestingly, we also found that while some tissue types were confused in bi-directional manner (e.g. breast ⇔ prostate) others were not (e.g. pancreas ⇒ breast).

Discussion

To better aid the analysis of tumor genomes, we present orchid, a powerful mutation management and machine learning framework. We also demonstrate orchid's ability to determine with high accuracy tissue-of-origin from tumor mutation data, which may have potential use in diagnosing tumors of unknown origin and for screening cfDNA. To our knowledge, orchid represents the first cancer mutation analysis framework with an in-memory database data storage, parallelization/cluster support, and integration with python numeric analysis and machine learning modules.

While orchid does not represent the first software to annotate mutations or produce mutation profile models within machine-learning, it does offer some advantages over other

methods. For one, it has the ability to quickly integrate new biological features by employing a flexible parsing system with parallel processing support for both a desktop machine or cluster. Secondly it allows seamless integration with existing python analysis workflows and provides functionality for basic machine learning tasks within the scikit-learn ecosystem. Finally, it provides many convenience functions to aid the visualization and analysis models.

Despite these advantages, orchid has a few limitations when compared to other software tools designed to model tumor mutations. For one, it's design centers around the analysis of simple somatic mutations and copy number variation data, and some cancers are largely driven by biological mechanisms of higher order genetic architecture, such as gene fusions (e.g. prostate cancer TMPRSS:ERG), large-scale structural rearrangements, epigenetic and gene expression changes. Nevertheless, the analysis of such mechanisms could be incorporated in future versions of orchid. Secondly, due to dependence on the scikit-learn ecosystem, some popular machine learning algorithms (e.g., neural networks) are not available for analysis or are not as fully featured as in other frameworks. And finally, orchid makes use of copy number variation data on a very granular mutational level, potentially missing important associations that could be seen when such data is analyzed over larger genomic regions.

With regard to our application of orchid to classify tumor tissue-of-origin, it is important to note that related methods have been previously developed. In particular, Snyder et al. used a novel nucleosome footprint window protection score to demonstrate correlation with patterns of protection and pathological states such as cancer (93). Likewise, Marquard et al. developed *TumorTracer* to classify tissue-of-origin with 85% accuracy across 6 primary sites using both somatic point mutation as well as copy number information (94). Orchid was able to achieve slightly a better accuracy of 94% among 12 tumor types, improving upon these initial methods.

While the tissue-of-origin task demonstrates one potential use of orchid, it is possible to model other types of data. For example, one can use orchid to generate a set of null, simulated mutations in conjunction with observed mutations to see if a particular feature set can be used to distinguish between the two classes, or even to assign a probability of class membership. This follows a similar strategy used by several driver/passenger and other base-level scoring tools (62,75,76,89) and has application in developing models for mutation prioritization for the design of custom sequencing panels for cancer detection.

Figure 3.1: Diagram of the orchid workflow.

Legend: The *make_database* shell script builds a database of annotated cancer mutations from raw source data using the *orchid_db* populate and annotate subscripts and can be run on a single computer or in a cluster environment. Afterwards, data can be quickly imported and analyzed with machine learning algorithms using *orchid_ml*.

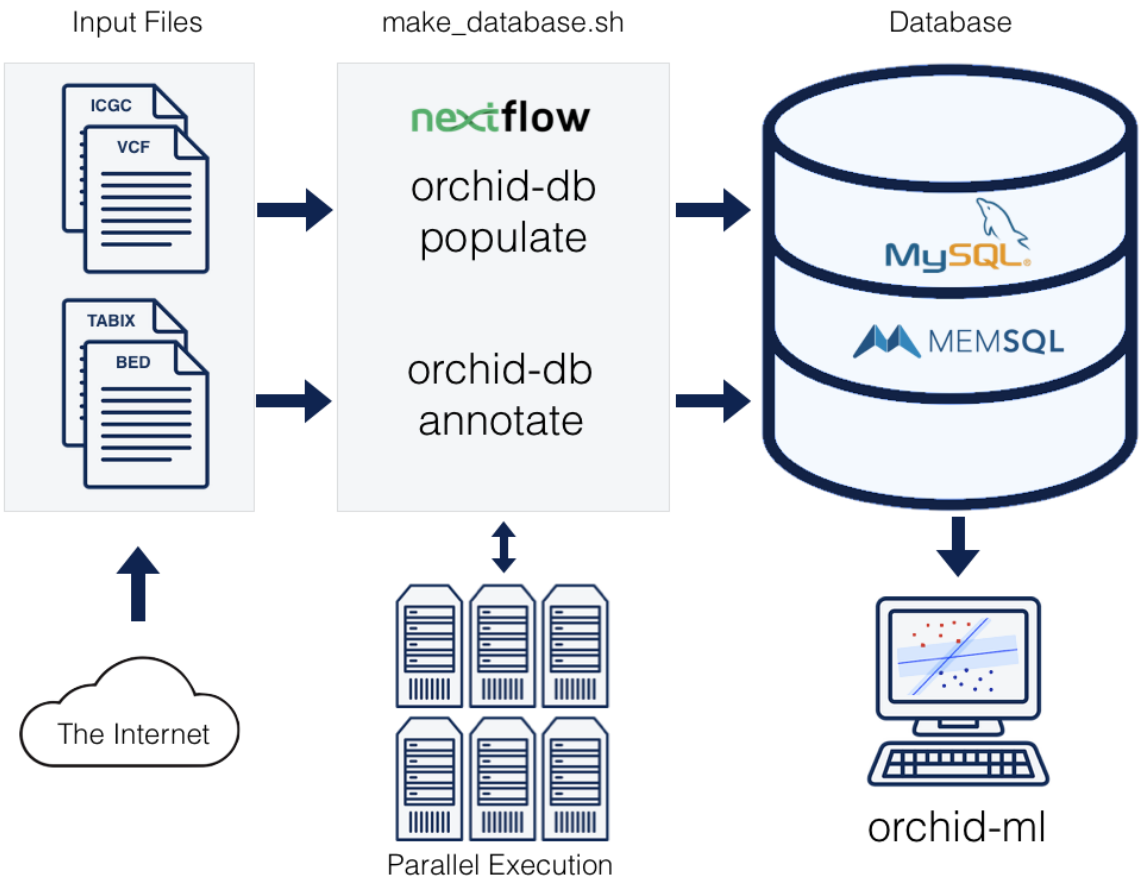


Figure 3.2: A tumor mutational profile dendrogram.

Legend: Patient mutation values were averaged over all features and labeled with the tissue-of-origin. The orchid-ml *show_dendrogram()* function was then used to generate a clustered heatmap. **A)** and **E)** Fairly strong separation of stomach, head and neck, bladder, and uterus tissues based on encode cell line transcribed regions was observed. **B)** In head and neck cancers, a frequent, G [T>A] G context, 3'UTR mutational signal was present. **C)** and **D)** Blood cancers showed separation from other tissues based on the V,D,J, and C immunoglobulin biotype features.

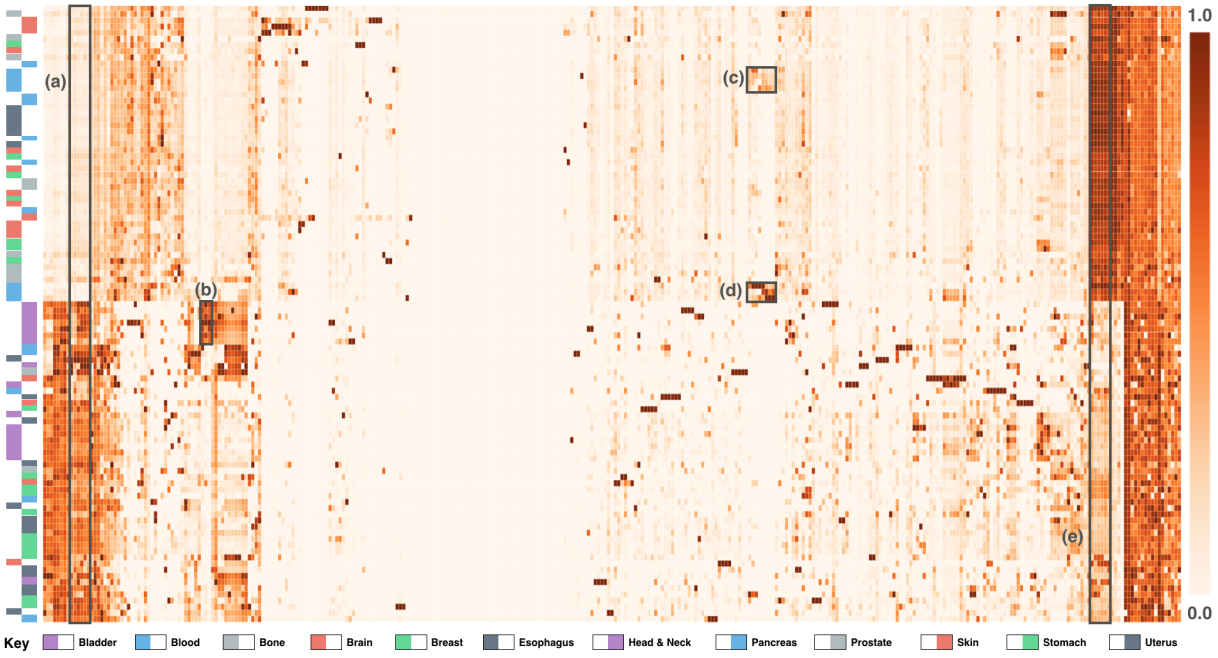
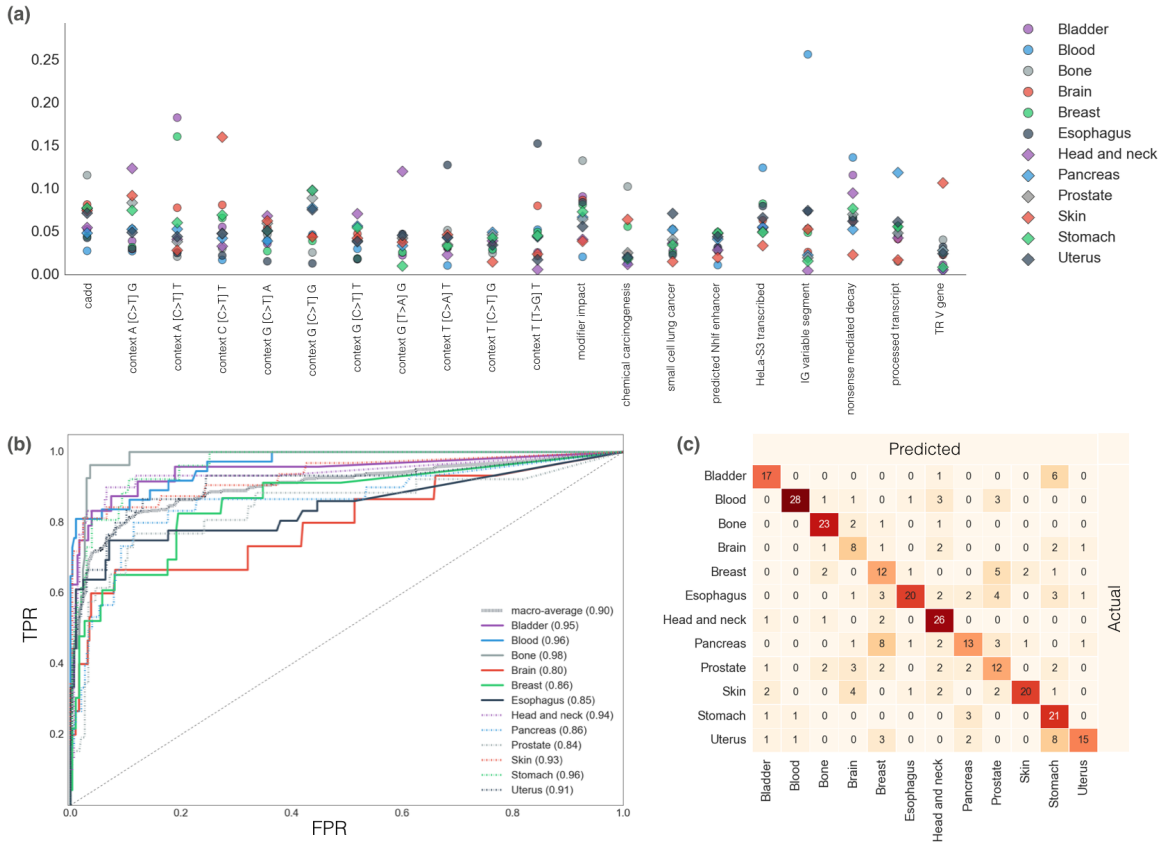


Figure 3.3: The models performance.

Legend: A) The twenty most important features for classification were selected using orchid_ml's *select_features()* function and plotted on a per-tissue basis after modeling. **B)** The true positive vs false positive classification rates for each tissue are plotted in a one-vs-rest fashion. The dashed diagonal line indicates random classification. The macro average over all models is shown and a heavy dashed line and Area Under the Curves (AUCs) are given in parenthesis for each tissue. **C)** A matrix indicating classification predictions from the tissue model. Rows labels are actual tissues and columns labels are tissues predicted by our model. True positive counts can be found along the diagonal, and of the remaining, false positives are along columns, and false negatives are along rows.



Supplemental Table 3.1: A table of features used to annotate tumor mutations.

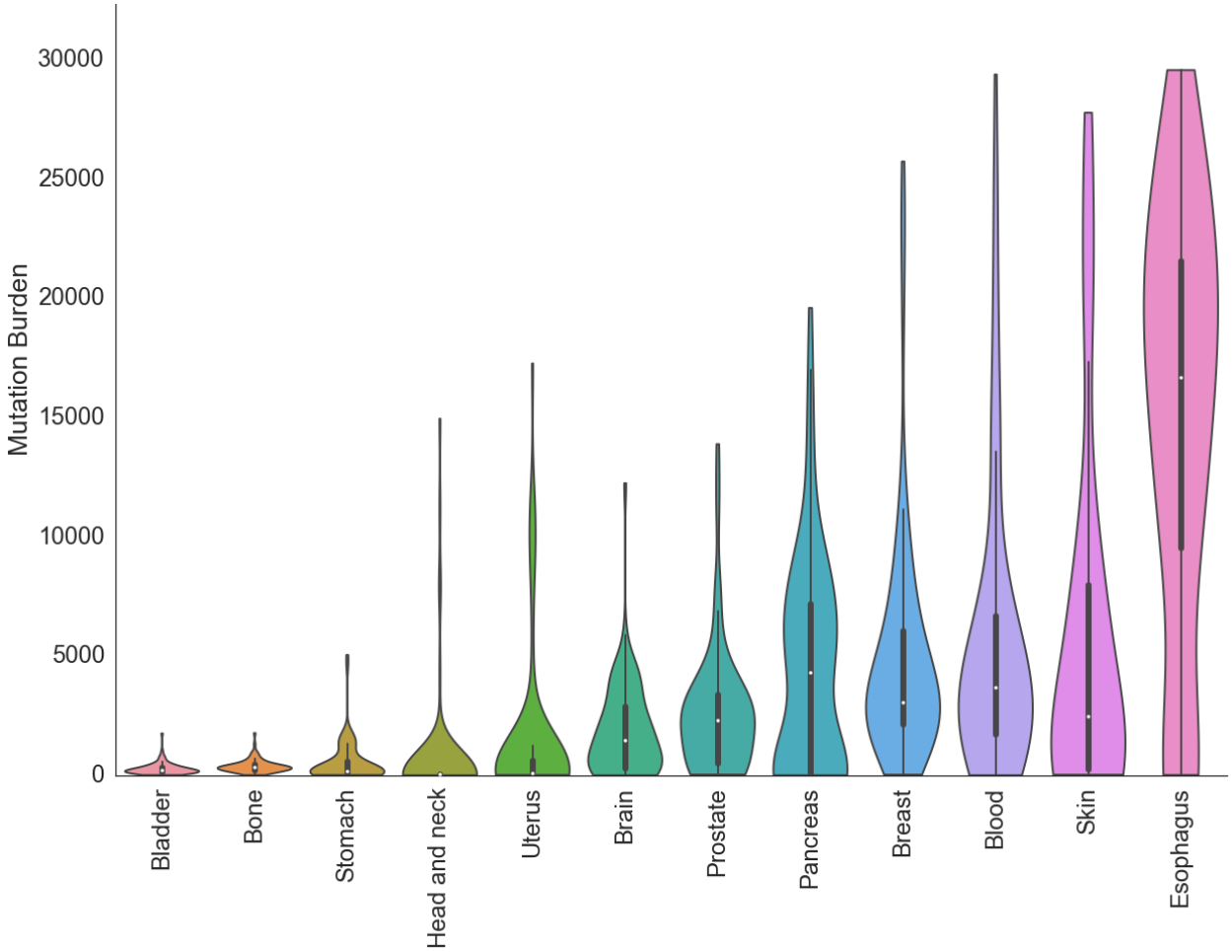
Legend: This table gives list of feature classes used in modeling with orchid, including links to feature websites and file downloads are located at http://wittelab.ucsf.edu/orchid.

Feature	Description
aa_class_change	Twenty six amino class change categories (e.g. Positive => Aromatic)
cadd; cadd_raw	Combined Annotation Dependent Depletion scores (c-scores and raw values)
consequence_type	snpEff consequences (e.g. nsSNP, 5'UTR, etc…)
context	The trinucleotide context as per Alexandrov *et. al.*
copy_number	segment mean $\log_2$ copy number values, averaged over ICGC donors
dann	Deep Artificial Neural Network (deleteriousness)
dbsuper	Super enhancers (any cell type)
dnase	Encode DNASE I Hypersensitivity distal and proximal (any cell type)
encode	Encode enhancer, promoter histone marks (distal and proximal; any cell type)
frequency	Number of donors sharing this mutation
funseq2	Somatic mutation variant prioritization (deleteriousness)
gwas	Boolean presence of GWAS Annotated SNPs
impact	Four SNPEff predicted functional impact categories (Modifier, Low, Moderate, High)
kegg_cancer_gene	Boolean presence/absence within a cancer pathway

Feature	Description
kegg_pathways	Biological cancer pathway membership (22 pathways)
phyloP	Primate evolutionary conservation score (deleteriousness)
remap	ChIP-seq analysis of regulatory elements
rfecs	Random Forest Enhancer / Chromatin States (11 cell types)
segmentation	ChromHMM and Segway genomic state predictors
targetscans	miRNAs sites
transcript_biotype	Transcript biotype as reported by SNPEff (35 categories)
wgrna	Whole Genome RNA predicted binding site (any)

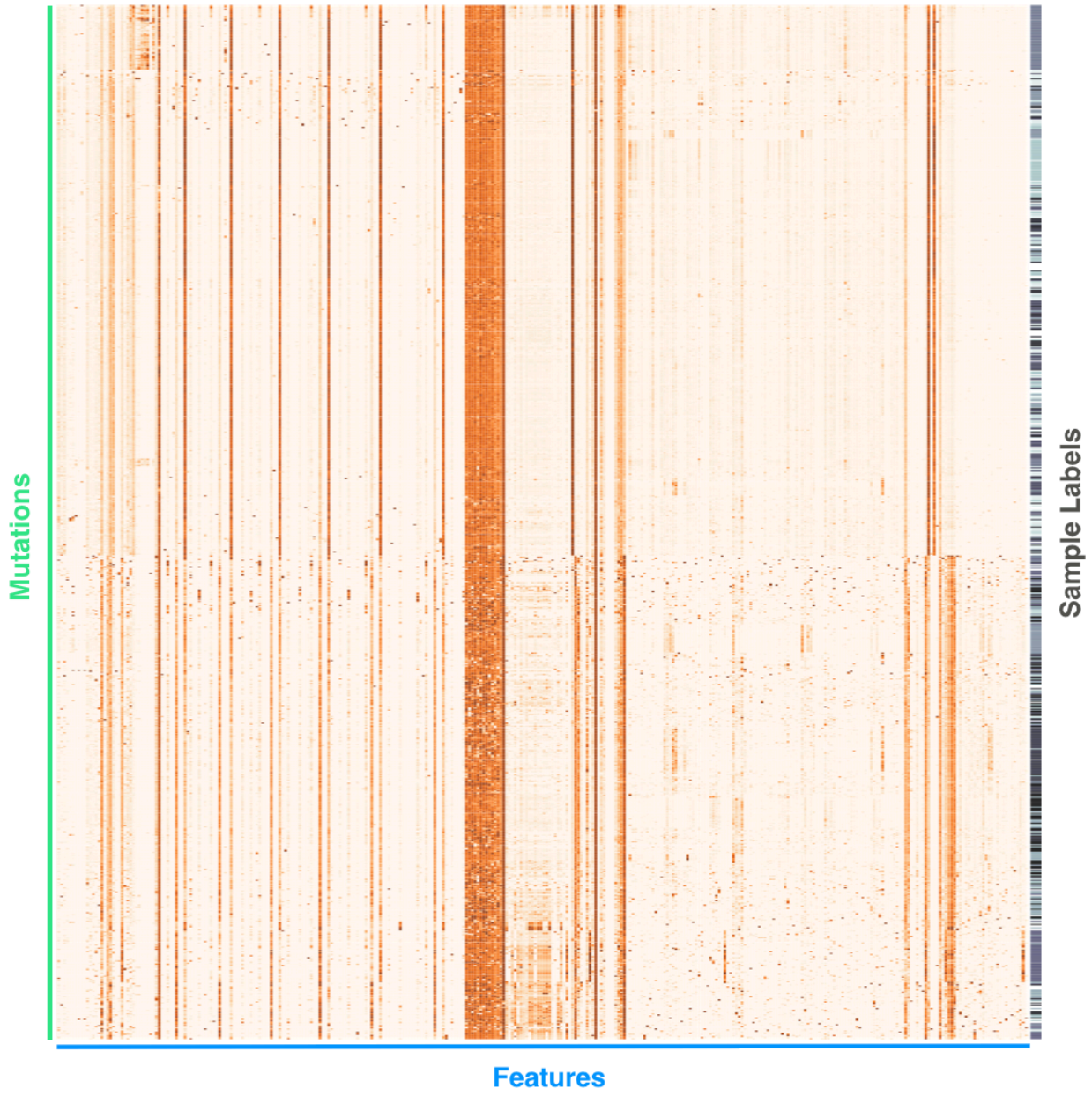
Supplemental Figure 3.1: A violin plot of mutational burden.

Legend: Mutational burden across the 12 tumor types used to generate the tissue-of-origin model is shown. Each tissue class consists of 80 tumors.



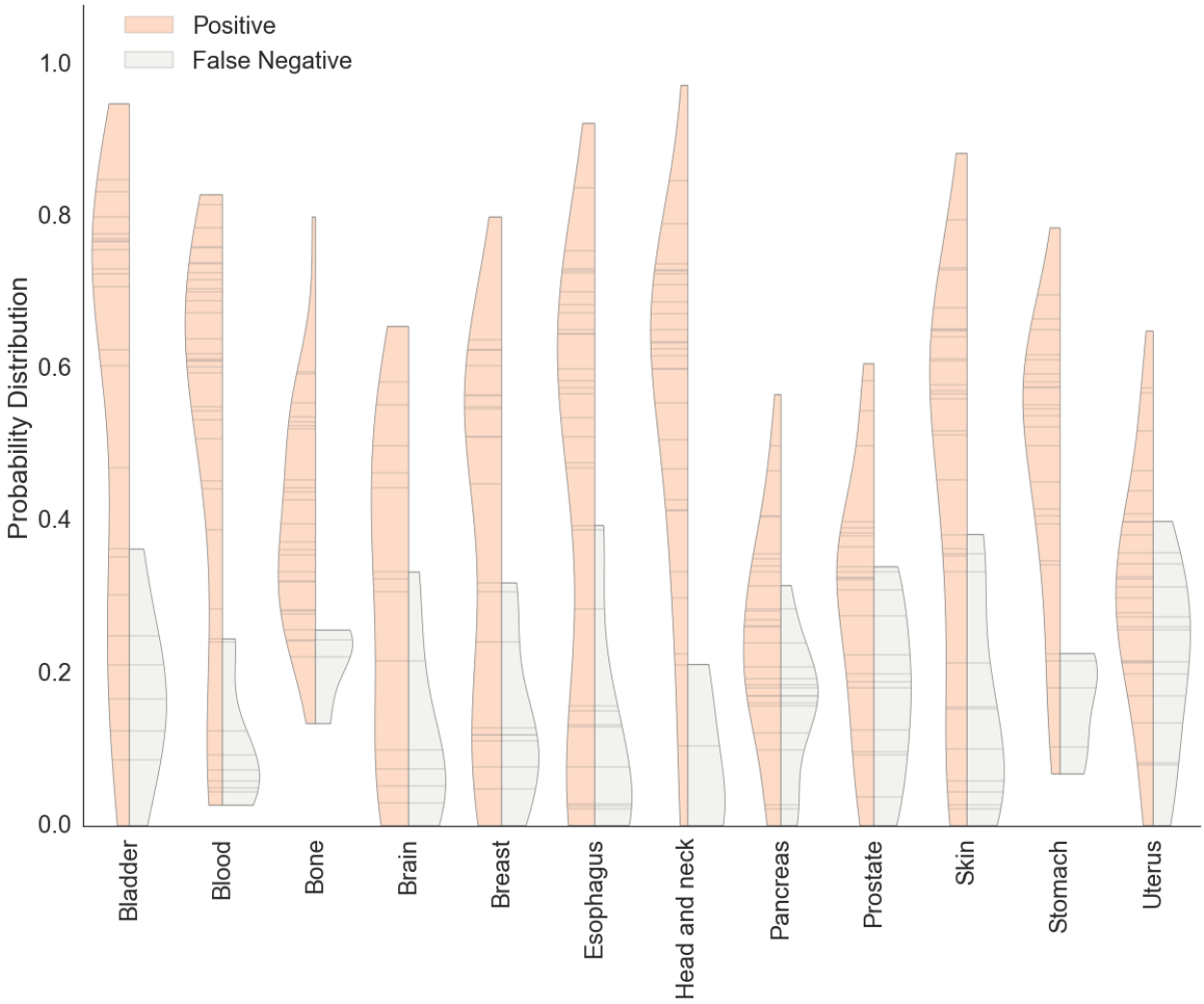
Supplemental Figure 3.2: Full cluster plot of all 960 tumor mutational profiles used for tissue-of-origin classification.

Legend: The 339 features are shown across columns and individual tumors across rows. Tumor labels are indicated as greyscale labels. Due to the density of the plot, feature and tumor labels are omitted. The data is normalized between using a min-max scaling transformation.



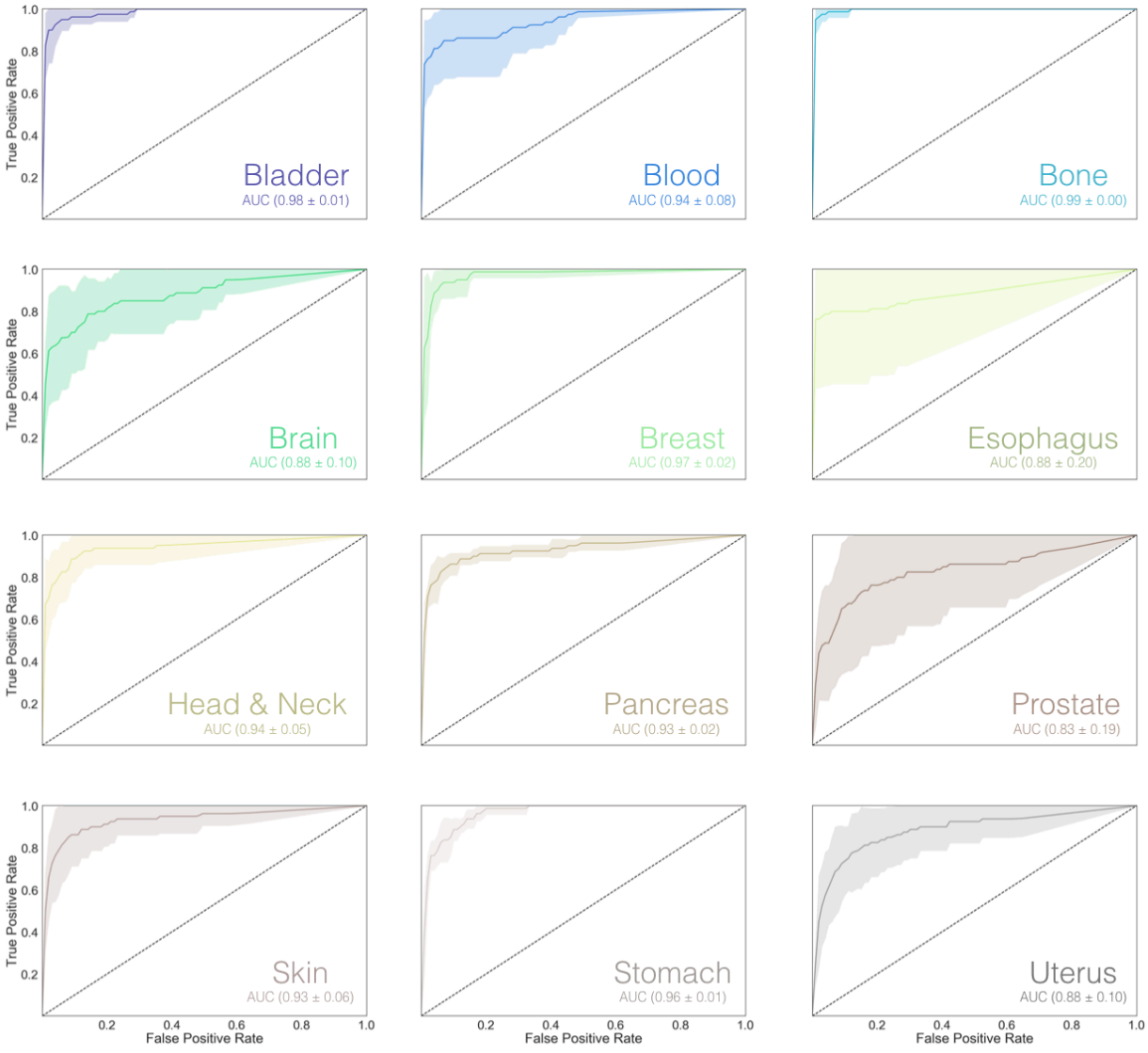
Supplemental Figure 3.3: An example of an orchid-ml confidence plot.

Legend: This plot shows tissue-of-origin positive vs false negative probability distributions for each tissue as a split violin plot. These violin splits can be generated with data from any combination of False Positive, False Negative, True Negative, True Positive, Negative, or Positive values.



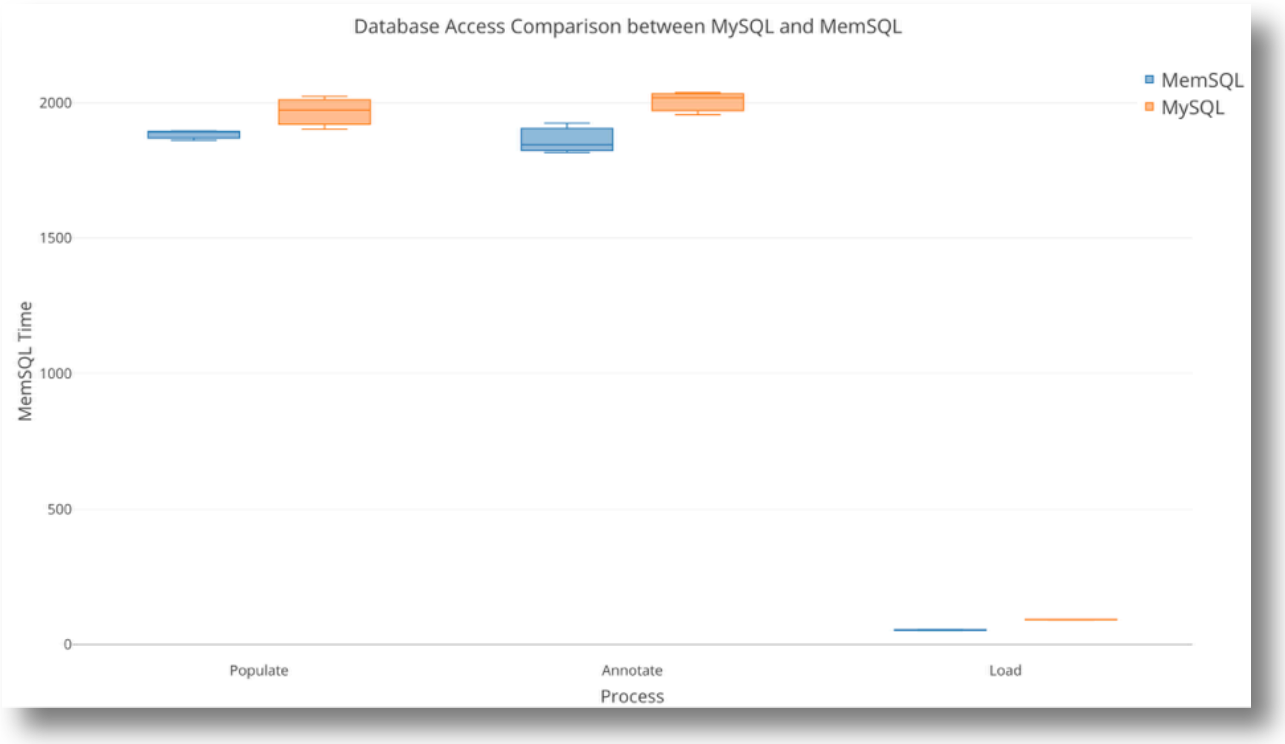
Supplemental Figure 3.4: Plots of tissue specific ROC curves.

Legend: Plots showing mean individual tissue ROC curves over 5-fold cross validation ± 1 standard deviation. The AUC of the mean curve is also shown. These curves can be produced using the show_roc_curves(separate=True, show_folds=False) function call.



Supplemental Figure 3.5: Comparison between MySQL and MemSQL.

Legend: Times for database population, annotation, and loading into the orchid-ml framework were compared between MySQL and MemSQL using the same set of data. This was repeated three times, and overall time in seconds is reported as boxplots.



Notes:

MemSQL is a modern in-memory distributed relational database system which is built to scale easily with very large datasets. Its syntax is almost identical to MySQL and connection is ODBC-compatible. Internally data is stored in-memory as rows and on-disk as colums. For transactional workloads, such as referencing individual variants from a large set, the in-memory row store provides fast and efficient lookup. MemSQL can be installed on commodity hardware (a quad core machine with 8GB RAM). More information about MemSQL, including hardware, software, and network requirements can be found in the MemSQL documentation:

https://docs.memsql.com/introduction/latest/how-memsql-works/

https://docs.memsql.com/installation/v5.8/system-requirements/

To compare the difference in performance between MySQL and MemSQL, we populated a test database from 1 million annotated prostate mutations from the International Cancer Genome Consortium (ICGC). Performance was compared for orchid's three most database intensive processes: population, annotation, and loading (into orchid-ml) and by measuring mean access time over three trials. MemSQL is significance faster for all three of these tasks, which is more pronounced for even larger datasets (not shown).

Chapter 4

A machine learning approach to optimizing cell-free DNA sequencing panels in prostate cancer

Introduction

Stroun et. al. first reported the presence of cell-free DNA (cfDNA) in the plasma of patients with cancer in 1989, and by 1994 tumor derived KRAS mutations were confirmed in the cfDNA of patients with pancreatic cancer using allele specific PCR primers (95). Since then a flurry of research has been conducted to explore potential oncological applications of cfDNA, including early detection, monitoring residual disease, and recurrence following treatment (47,96-98). Many of these efforts have been met with success, as demonstrated by the FDA's approval of the first liquid biopsy test: a qPCR based test called cobas that can detect 42 mutations in the EGFR gene using the cfDNA from patients with Non-Small Cell Lung Cancer (NSCLC).

Despite some promising initial results, there are still many challenges for cfDNA as a biomarker (37). One of the most important limitations, especially in the context of variant detection, is the weak signal-to-noise ratio of circulating tumor of ctDNA to cfDNA derived from healthy tissue. Often, ctDNA represents much less than 1% of the total cfDNA fraction (34). To circumvent this issue, two strategies have been developed. The first focuses on probing a single mutation or a small number of genomic locations and utilizes qPCR or Unique Molecular Identifiers (UMIs) coupled with sequencing to reach sensitivities required for detection (51,99,100). The second approach uses targeted sequencing (or possibly even whole exome sequencing) to expand the breadth of genomic coverage thereby increasing the likelihood of detection through multiple independent sampling. Both of these strategies have trade-offs. For

example, qPCR based methods have high sensitivity but require a priori or patient-specific knowledge of mutation locations. In some cancers, like prostate, this can be especially problematic as even the most common driver mutations exist at frequencies too low to be of clinical utility (101). Targeted sequencing can be used for broader patient coverage and de novo discovery, but at the cost of reduced sensitivity or greater risk of false positives.

We propose a solution that leverages strengths of both approaches while minimizing their weaknesses. This is done through several strategies: 1) generating a medium-sized (2.5Mb) targeted sequencing panel, but instead of including coding regions of entire genes, focusing on small (~350bp) regions of the genome that are most likely to harbor mutations; 2) incorporating mutations from non-coding regions that may have important regulatory effects; 3) selecting candidates for inclusion on the panel by a machine learning model trained on actual tumor data, and optimized to detect driver-like mutations; and 4) using UMIs to suppress technical errors induced by sequencing. Here we present this novel approach to building a targeted sequencing panel and its validation, and then demonstrate its performance in comparison with other conventional panels for screening cfDNA mutations that were somatically validated using matched tumor/normal data from multiple foci of prostate cancer patients.

Methods

Model Training Data

Prostate tumor variant data from the International Cancer Genome Consortium (ICGC; https://icgc.org), release 23, was obtained to build a classification model and rank mutations for inclusion on a custom hybrid capture panel. Using the ICGC Data Portal Advanced Search tool, 'Prostate' tissue was subsetted from Primary Site; 'WGS' from Donor Analysis Type; and then

'Simple Somatic Mutation', 'Copy Number Somatic Mutation', and 'Clinical Data' data was downloaded and extracted. Data was then imported into a MemSQL database using the make_database.sh script from our previously described orchid software (102). All feature data, available at the orchid website (http://wittelab.ucsf.edu/orchid), was also included in database population. Copy number information was preprocessed using the parse_cnv.sh script in the orchid repository before import.

Union of Existing and Frequency Comparison Panels

To benchmark performance of the orchid generated mutation panel in detecting tumor variants, two other panel designs were explored. The first, "Union of Existing", consisted of coding regions from the aggregated set of 530 genes found from four clinically available cancer-specific targeted sequencing gene panels (Supplemental Table 1). This was generated by intersecting gene hg19 coordinates (as queried through Ensembl Biomart at http://feb2014.archive.ensembl.org/) with the SeqCap EZ Exome v3 capture panel to include only exons (n=9,470). These regions were then randomly downsampled to match the size of the orchid generated panel (n=7,034). The second comparison panel, "Frequency", was queried from the ICGC prostate release 23 database. In this case, all mutations present in more than one donor were considered ± a 175 bp window to match the orchid panel region sizes (n=5,824 regions).

In Silico *Analysis*

We assessed the orchid panel's potential performance using an in silico analysis of tumor foci DNA and matched cfDNA. We whole exome sequenced multiple tumor foci and a normal tissue control sample from 5 prostate cancer patients who underwent radical prostatectomy, and called somatic variants (see Initial Cohort Sample Extraction). We also extracted patient matched cfDNA and exome sequenced it at 200X followed by variant calling (see cfDNA Extraction and

Sequencing). Next, we generated in silico capture probes for the orchid panel by expanding the genomic coordinates of panel mutations by ± 175 bp (the mode size of cfDNA molecules). Tumor and cfDNA variants were intersected with this generated panel and the two comparison panels described above.

Hybrid Capture Probes

We designed and ordered Custom MyBaits™ hybrid capture probes targeting the final corrected set of mutations, ± 175 bp and tiled 3X through Arbor Biosciences (https://arborbiosci.com; 16 reaction; catalog 300116). This targeted capture panel was then applied by Arbor Biosciences to library prepped cfDNA samples from Study Sample Patients.

cfDNA Extraction and Sequencing

Between 10 and 20 mL of whole blood was collected from a cohort of prostate cancer patients at time of radical prostatectomy or during treatment for metastatic cancer. Blood was first spun at 1,900 g for 10 minutes and collected plasma was respun at 16,000 g for 10 minutes to remove any residual cell debris. Samples were then processed using the Qiagen Circulating Nucleic Acid Kit (Catalog # 55114), double eluted with 40 uL of Qiagen Elution Buffer (EB) for 80 uL total and run on the Agilent Bioanalyzer with High Sensitivity DNA chips (Catalog # 5067-4627) to assess concentration and fragment size distribution. Next, for samples meeting a 10 ng threshold, the Zymo Clean and Concentrator Kit (Catalog # 4013) was used to concentrate DNA into 10 uL of distilled water and resulting samples were library prepped using the UMI tagging Rubicon ThruPLEX Tag-Seq 48S kit (Catalog # R400585) following kit recommendations for final PCR amplification (7-11 cycles). After AMPure XP bead cleanup (Catalog # A63881) samples were bioanalyzed for quality control and either sent for sequencing directly (for Initial Cohort samples) or sent to Arbor Biosciences for hybrid capture (for Study

samples) using the orchid generated panel. In cases where multiple sequencing strategies were used, or when DNA yield was less than required for sequencing, reamplification for 3 cycles was performed followed by AMPure cleanup and bioanalyzer. Samples were sent for sequencing with a target depth of either 200X for whole exome sequencing (Initial Cohort) or 2,500X (or ~90 Million reads) for panel captured samples (Study).

Sequencing data analysis was performed using either a hisat2/connor/freebayes workflow (Initial Cohort) or the Curio Genomics (www.curiogenomics.com) web platform (Study). Default parameters were used with the exception of connor (consensus frequency threshold of 0.6, min family size threshold of 3, umi distance threshold of 1) and freebayes (min alternate fraction of 0.01, min alternate count of 3). For samples run on the Curio platform the following parameters were used: 1) Alignment UMI demultiplexing with a 6 UMI and 100 max stem length, 2) Variant calling with the orchid panel genome feature, 75% family threshold, 2 base hamming distance, and minimum family size of 4 reads, and 3) Filtering with a minimum quality of phred 30, 2 family minimal coverage, and rare allele frequency between 0% and 20%. In some cases, when mentioned, stricter filtering was applied during cfDNA variant analysis. Parameters for this analysis were the same as above except 1) 10 family minimal coverage and 2) 100 total family minimal coverage.

Tumor/Normal Sample Extraction and Sequencing

For Initial Cohort samples, multiple tumor tissue foci and a normal tissue control samples (seminal vesicle or whole blood if not available) from 5 stage 1 or 2 prostate cancer patients who underwent radical prostatectomy were collected and processed with the Qiagen DNeasy Blood and Tissue Kit (Catalog # 69504). Samples were sent for whole exome sequencing at 200X.

Tumor normal variant calling was performed using the bcbio GATK workflow (https://github.com/bcbio/bcbio-nextgen).

For Study samples, multiple tumor foci and a normal tissue control samples (seminal vesicle or whole blood if not available) from 9 stage 1 or 2 prostate cancer patients who underwent radical prostatectomy were collected and processed with the Qiagen DNeasy Blood and Tissue Kit. Samples were sent for whole exome sequencing at 40X. Tumor normal variant calling with then performed with the SpeedSeq workflow (103) using a minimum of 5% allele fraction somatic variant calling. Additional filtering was applied to variants where mentioned.

Results

Training Data

To develop a machine learning model for generating a prostate cancer specific targeted sequencing panel that detects tumor mutations in cfDNA, we first downloaded whole genome mutation and copy number data from prostate cancer patients the International Cancer Genome Consortium (ICGC) data portal (see Methods). In addition, we collected more than 300 genome annotation features as previously described (102). Using the orchid software developed by our lab, we populated this information into a MemSQL database (102). In total, data from 550 cancer patients (274 with copy number information), 1,717,507 mutations (1,588,558 single base substitutions, 66,202 insertions ≤ 200 bp, and 90,255 deletions ≤ 200 bp), and 339 features was included in the database.

For training labels, we defined two classes of mutations by dividing the prostate cancer patients into two groups (n=275 each) based on their number of mutations: 1) Driver Enriched (DE), consisting of mutations from men with a lower mutational burden, and 2) Passenger

Enriched (PE), consisting of mutations from men with a greater burden (**Figure 1A**). The cutoff used to classify patients into DE and PE groups (675 mutations) was chosen to create two equal sized groups. We chose this classification strategy because tumors with few mutations are generally earlier in the carcinogenic process, and the corresponding mutations are more likely to drive disease (or be biologically relevant). In contrast, tumors with many mutations are usually at a later stage of cancer, or have compromised DNA maintenance machinery, and the corresponding mutations have greater likelihood of representing an artifact of unstable genome maintenance (62). This classification scheme was also reflective of our desire to create a panel for early, localized prostate cancer. With these two classes defined, to reduce the computation complexity we randomly downsampled the data to a total of 16,334 unique mutations while preserving the original DE:PE mutation ratio (0.026).

Initial Modeling and Performance

To guard against overfitting, we used orchid's feature selection method to reduce the number of features from >300 to 20 and performed 10-fold cross validation with a linear support vector classifier (SVC) to generate a DE predictive model. A ROC curve of model performance in the test sets is given in **Figure 1B**, indicating a 0.76 (± 0.12) classification accuracy. Additionally, we observed classification probabilities across all test cases and discovered strong model 'confidence' in its predictions, tending to call mutations as PE or DE with high probability and at rates reflective of the unbalanced DE:PE ratio (**Figure 1C**).

Feature Selection and Significance

We then visualized feature weights used for classification of mutations as being drivers versus passengers (**Figure 1D**). In this figure the sign and magnitude of features define a vector orthogonal to the hyperplane that maximizes the margin between DE and PE classes. Since the

dot product of this vector with mutation vectors form the basis for class assignment in a linear SVC, these weights can give insights into the predictive model. For example, we observed segregation by ChromHMM functional genomic segmentations of ENCODE cell-lines (82). In some cases, reciprocal functional states within a cell-line appeared to help classify mutations in opposite directions (e.g. K562: Repressed ➡ DE, Transcribed ➡ PE; HUVEC: Repressed ➡ DE, Transcribed ➡ PE). As another example, for mutations classified as DE, a higher scaled CADD score was predictive, indicating mutations tended to be highly deleterious. In contrast, for mutations classified as PE, a positive raw CADD feature indicated these mutations appeared simulated-like deleterious, which could correspond to genomic instability of late-stage disease (75). Curiously these features were in opposite directions, perhaps explained in part by the compressed PHRED-like scoring used for the scaled CADD score.

We also noticed other patterns with potential biological explanations, including that coding mutations classified as DE tended towards preserving amino acid chemical properties (Unchanged AA feature). This suggests that PE classified coding mutations alter amino acid chemical properties through a variety of the other possible mechanisms (e.g., aliphatic ➡ aromatic, polar ➡ neutral, polar ➡ aliphatic), which could occur later in disease progression. Finally, we observed that mutations in bases with high 'deleteriousness' scores (i.e., related to strong evolutionary conservation), including cancer-specific scores like FunSeq2 and copy number, were particularly useful for classification PE mutations; this further suggests genomic instability of late-stage disease.

Mutation Ranking

We then used the selected features to build a new linear SVC with all downsampled data and applied it to the full ICGC dataset to score DE probability for all prostate cancer mutations. Mutation distances from the fit hyperplane were then used to rank them. Those with the greatest magnitudes in the DE direction ranked highest (i.e., the most "driver like"), and were further considered for inclusion on a targeted sequencing panel.

Standardizing Mutation Scores

We annotated the candidate DE mutations with associated gene information, if available, using SNPEff (78) functional impact information and transcript length from the UCSC genome database. We then binned genes according to their length, and the number of mutations per gene was visualized (**Figure 2A**). As one might expect, longer genes had more candidate mutations (Pearson's correlation = 0.20, p=6.03e-39). Such candidates, however, could be over represented on a panel if they had marginal individual scores but strong gene-level annotations. To address this issue and to increase diversity on the panel, we implemented a corrective standardization (**Supplemental Figure 1**) and applied it to the distance scores of mutations (**Figure 2B**). This standardization reduced to the Pearson's correlation between gene length and number of candidate mutations to 0.05 (p=0.0015). Mutations that were non-coding or without gene annotation were unaffected by this method. After applying the standardization, the top 7,034 mutations were then selected for the panel. This panel represented 0.41% of the total number of potential mutations.

Panel Composition

Once our standardized panel was established we plotted hyperplane distances for the 7,034 mutations against their logged rank (**Figure 2C**). The top 5 coding mutations are labeled

with their corresponding genes, all of which are involved in cancer (104-106). While the most highly ranked genes were coding, many non-coding genes were also included on the panel (~18%).

In comparison with random prostate tumor mutations, our panel was enriched for general cancer (p=4.18e-228) and prostate cancer genes (p=1.19e-61). In addition, our panel was significantly enriched for regions associated with regulation of cellular response to growth factors (p=2.68 e-4), MAP kinase activity (p=8.66e-8), and Integrin signaling (p=1.74e-13) among others (107-109) (Enrichr; http://amp.pharm.mssm.edu/Enrichr/). A table of consequence mutations is shown in **Figure 2D**. A majority of coding mutations were classified as high or moderate impact, including 50 induced stop gains and 3,386 missense mutations.

For non-coding mutations, we discovered significant enrichment for several general/prostate cancer transcription factor binding sites (**Supplemental Figure 2**), including BRD4 (e=329), CTCF (e=254), FOXA1 (e=188), MYC (e=181), and AR (e=159), and a microRNA involved in angiogenesis (mir-126) (91,107) (ReMap; http://tagc.univ-mrs.fr/remap/).

In silico *Analysis*

We then compared how well our panel detected somatic variants in relation to two other panels: 1) the union of four existing sequencing panels (Fluxion Biosciences, Foundation Medicine, Guardant Health, and UCSF 500; **Supplemental Table 1**); 2) and a frequency-based panel (most common mutations; see **Methods**). We first assessed the panels' ability to identify somatic variants in 5 prostate cancer patient's primary tumors. Orchid's panel detected more variants than the two other panels across all patients (**Figure 3A**), and these differences were statistically significant for patients P0014 and P0023 (p=0.043 and p=0.017 in comparison with

the union-existing panel; p=0.011 and 0.014 in comparison with the frequency panel, respectively; from T-test). The second scenario assessed how well the panels detected variants found in both the patient's cfDNA and their matched primary tumor foci. Again the orchid panel performed better than the other two other panel approaches, detecting more variants for 4 out of the 5 patients (all $p<5x10^{-7}$), where the fifth patient only had a single tumor focus (**Figure 3B**).

Panel Performance: cfDNA Variant Detection

After confirming that our machine learning panel improved upon the union-existing and frequency based panels, we ordered hybrid capture probes for our regions of interest, which totaled ~2.5Mb. We then sequenced 9 patients with multiple prostate tumor foci and normal tissues @ 40X WES. Matched cfDNA was also collected from these patients at time of radical prostatectomy, extracted, and targeted sequenced with our panel at 2,500X. We then assessed how well our panel performed in detecting variants that were also somatically found within patient tumors (**Figure 4A**). The orchid panel detected tumor variants in all patients, ranging between 3 (S083) and 119 mutations (S041). The allele frequency of detected variants ranged between 0.2% to 20% (the default germline threshold used in the curio analysis), though the majority of variants were at frequencies >1%, the theoretical threshold at 2,500X sequencing.

To see how our panel would perform at greater sequencing depths and in healthy patients, we compared strictly filtered cfDNA variants from 4 patients that were sequenced at 5,000X to those observed when their sequence data was downsampled to 2,500X. On average, 2.5 times as many variants were detected with deeper sequencing. These results suggest that the orchid panel performance will directly improve with increasing sequencing depth. Finally, we extracted cfDNA from 20 healthy volunteers, applied the orchid panel, and sequenced at 2,500X. When compared to the cfDNA of 28 prostate cancer patients sequenced at the same level, we observed

significantly fewer variants than in the prostate cancer patients ($p < 6.0e-5$; Supplemental **Figure 3**). This suggests that our panel could be used to specifically detect variants in prostate cancer from those found at low frequency in clones of blood cells (i.e., the Clonal Hematopoiesis of Indeterminate Potential or CHIP effect (110)).

Discussion

Our machine learning approach allows for designing an optimal targeted sequencing panel for tumor-derived cfDNA. We used prostate cancer data from ICGC and twenty biologically relevant genome annotation features to train a classification model, which then ranked candidate mutations for a panel. Furthermore, our panel outperformed two alternatives—one generated from a combination of several existing panels, and one based on tumor mutation frequencies—as assayed with an *in silico* screen to detect tumor-derived cfDNA mutations. Applying our panel to prostate cancer patients showed that it was able to detect tumor-derived cfDNA mutations in every patient, and indicated a significantly increased detection rate when compared to healthy controls. Based on our results, the sensitivity of the panel should scale at least linearly with increased sampling depth.

When training our machine learning model, we made several design choices that could be further optimized in the future. Foremost, we used a linear SVC classifier and trained a model with 76% classification accuracy. While SVCs allow for interpretability of features used in classification and are generally a robust model type, other (perhaps more accurate) classifiers could have been used instead (e.g. random forest, neural net). However, accuracy is also impacted by noise inherent to our training labels (i.e. driver/passenger mutation hypothesis), which reflects the uncertainty of assigning accurate biological function to any given mutation in cancer datasets. Other labeling schemes or the employment of continuous class labels may have

been used as an alternative. Finally, to reduce overfitting, we used a standard feature selection method that does not necessarily reduce correlation among features. Depending on computational time, future models could use more sophisticated algorithms that take feature correlation into account.

Another consideration when using methods described here relates to cfDNA as a biomarker itself. The most important of these is the physical limit to the number of molecules obtained using conventionally available blood volumes (5-20 mL). From our experience in prostate cancer, this equates to ~20 ng of DNA, ~100 billion molecules, or a maximum genome coverage of ~5,400X. Further complicating this issue is library preparation protocols with multiple cleanups (losing molecules) and amplification steps (introducing false positives), which makes mutation detection by sequencing difficult, especially in tumor tissues like prostate where mutation rate is lower relative to other tissues. While the use of UMIs ameliorates the issue, clearly more starting material (blood) is needed. Improvements in hybrid capture probe design, or multiplexing of targeted amplicons is also required. And finally, specificities of cfDNA based assays must improve to accommodate high background mutation rates induced by the CHIP effect.

In conclusion, we have developed a novel model to rank coding and non-coding mutational hotspots for inclusion on a targeted sequencing panel. We show how this panel compares to two other panel design paradigms and demonstrate its performance in detecting actual tumor-derived cfDNA variants. This provides a useful strategy for broad—yet sensitive—future panel design for mutation detection in cfDNA isolated from cancer patients.

Figure 4.1: Modeling simple somatic mutations.

Legend: A) We divided ICGC prostate cancer donors into two classes, Driver Enriched (DE) or Passenger Enriched (PE), based on the number of somatic mutations in their tumors and labeled their mutations accordingly. **B)** After modeling with a linear Support Vector Classifier (SVC), we generated a ROC curve of DE classification. Accuracy was 76% +/- 12%. **C)** We visualized classification probabilities for test mutations. The model predicts fewer DE mutations and classifies both classes with high confidence. **D)** We show model feature weights for both classes. Several features, including those in ENCODE gene segments annotation were useful for classifying both DE and PE variants.

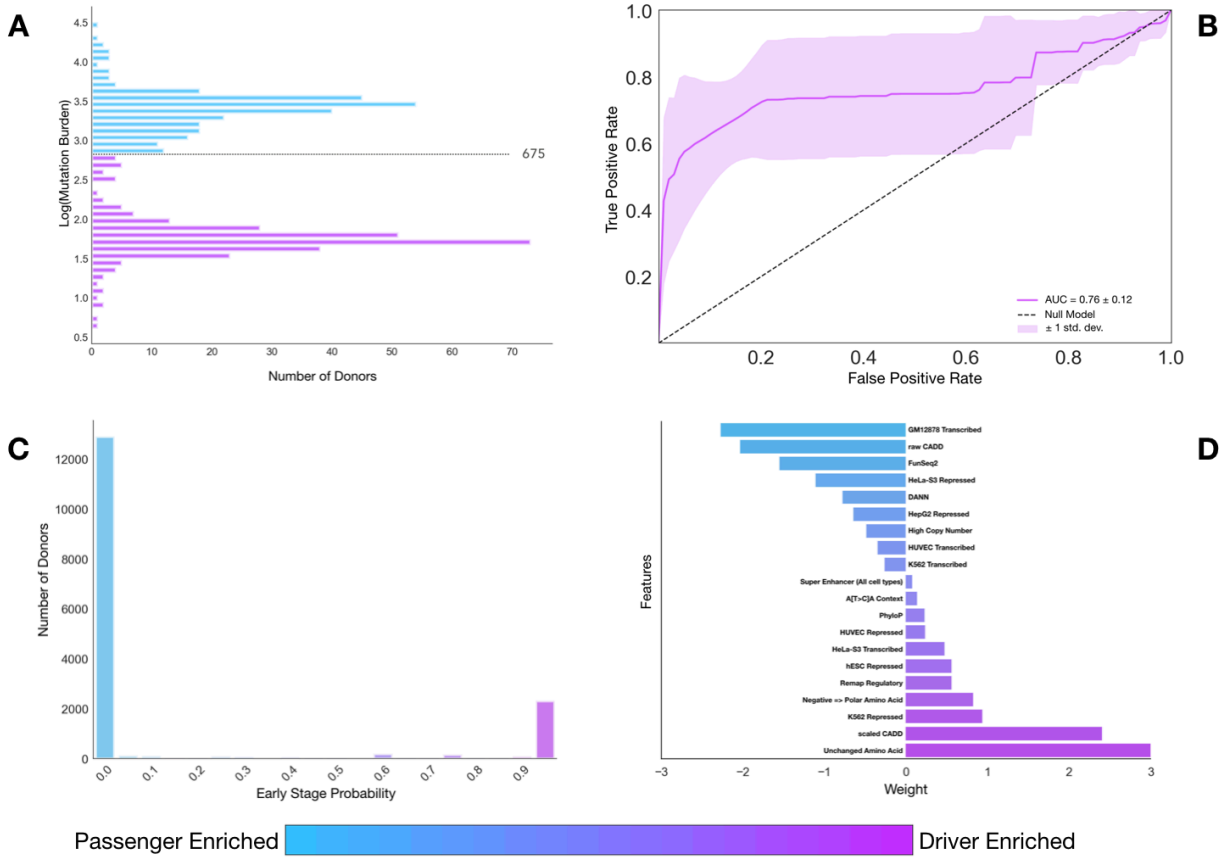
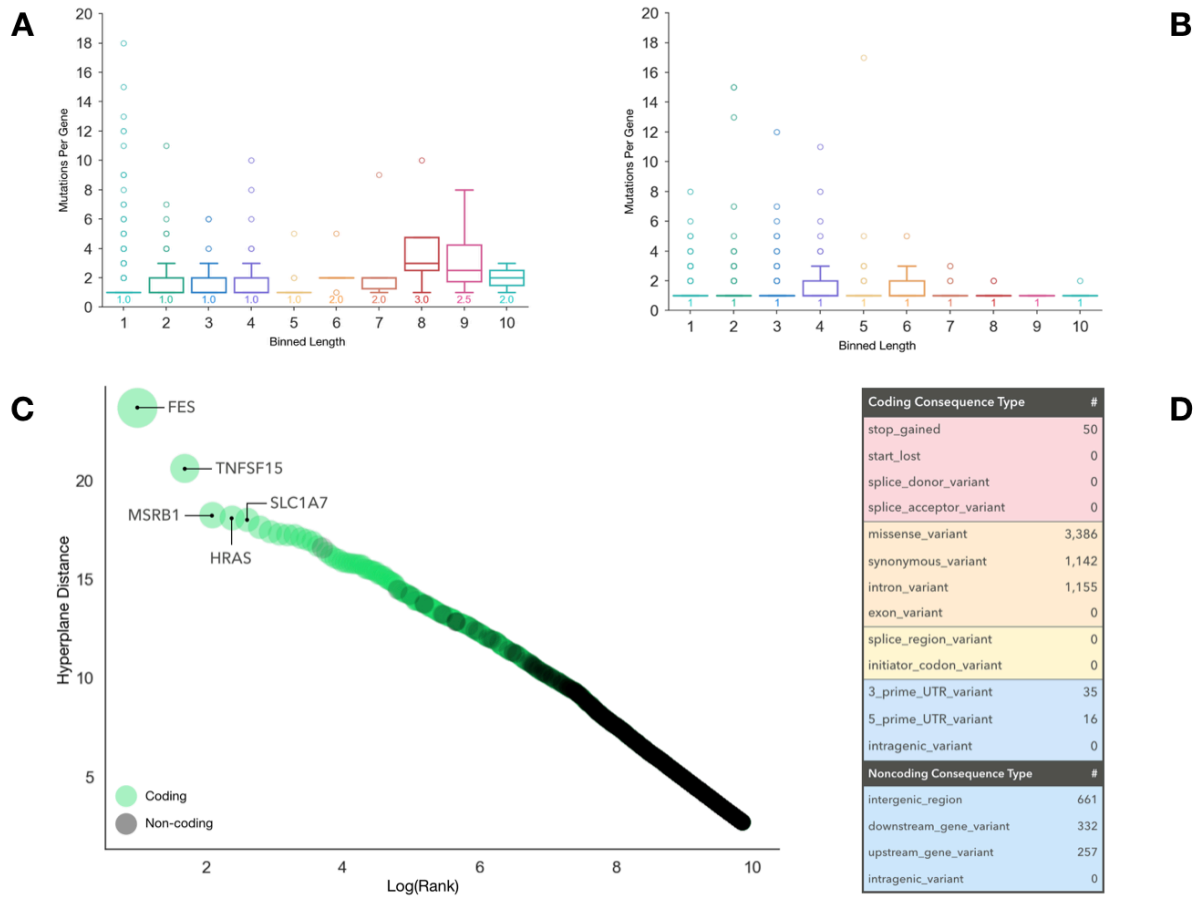


Figure 4.2: Generating a targeted sequencing library for hybrid capture of DE mutations.

Legend: We generated a candidate panel consisting of probes targeting the ~7,000 highest ranked DE mutation loci. **A)** We binned genes represented by candidate mutations into 10 groups based on length and show the distribution in number of mutations. Gene length correlated with the number of mutations on the panel (Pearson's correlation = 0.20, p=6.03e-39). B) We employed a distance standardization to mutation hyperplane distances to increase gene diversity on the panel. After standardization the correlation between gene length and number of mutations decreased significantly (Pearson's correlation = 0.05, p=0.0015). C) We plotted the hyperplane distances of retained mutations after standardization against logged mutation rank. Mutations are labeled as coding (green) or non-coding (grey). We labeled the top 5 coding mutations with their corresponding genes. D) We show a table of panel mutation consequence types and counts, colored by impact severity (red=high, orange=moderate, yellow=low, blue=modifier).



Coding Consequence Type	#
stop_gained	50
start_lost	0
splice_donor_variant	0
splice_acceptor_variant	0
missense_variant	3,386
synonymous_variant	1,142
intron_variant	1,155
exon_variant	0
splice_region_variant	0
initiator_codon_variant	0
3_prime_UTR_variant	35
5_prime_UTR_variant	16
intragenic_variant	0
Noncoding Consequence Type	**#**
intergenic_region	661
downstream_gene_variant	332
upstream_gene_variant	257
intragenic_variant	0

Figure 4.3: Panel performance using *in silico* capture cfDNA.

Legend: Five patients with multiple prostate cancer tumor foci, normal tissue DNA, and matched cfDNA were whole exome sequenced at 200X-fold coverage. Discovered variants were compared to three in silico capture panels: 1) Our orchid generated panel, 2) A panel consisting of genes on any of 4 clinically used panels (union-existing), and 3) A panel consisting of all mutations in the ICGC prostate cancer dataset with a frequency > 1. A) The log number of somatic mutations called from patient foci are shown (purple) in comparison with those also present on the three panels. Orchid detected significantly more mutations in patients P0014 and P0023 (p=0.043 and p=0.017 in comparison with the union-existing panel; p=0.011 and 0.014 in comparison with the frequency panel, respectively, using a T-test) B) The log number of tumor mutations also found in cfDNA are shown (green) in comparison with the three panels. Orchid detected significantly more mutations in all cases.

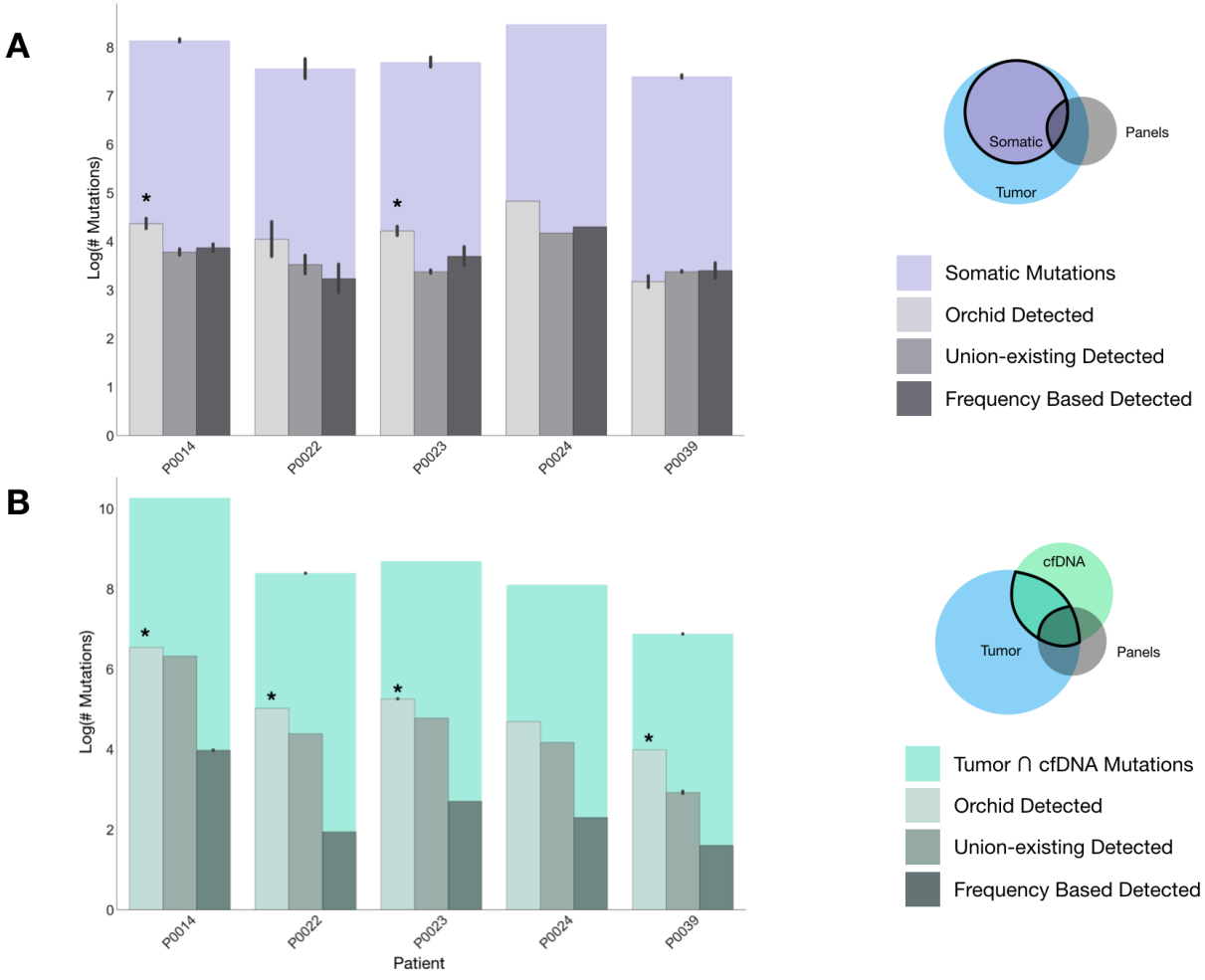
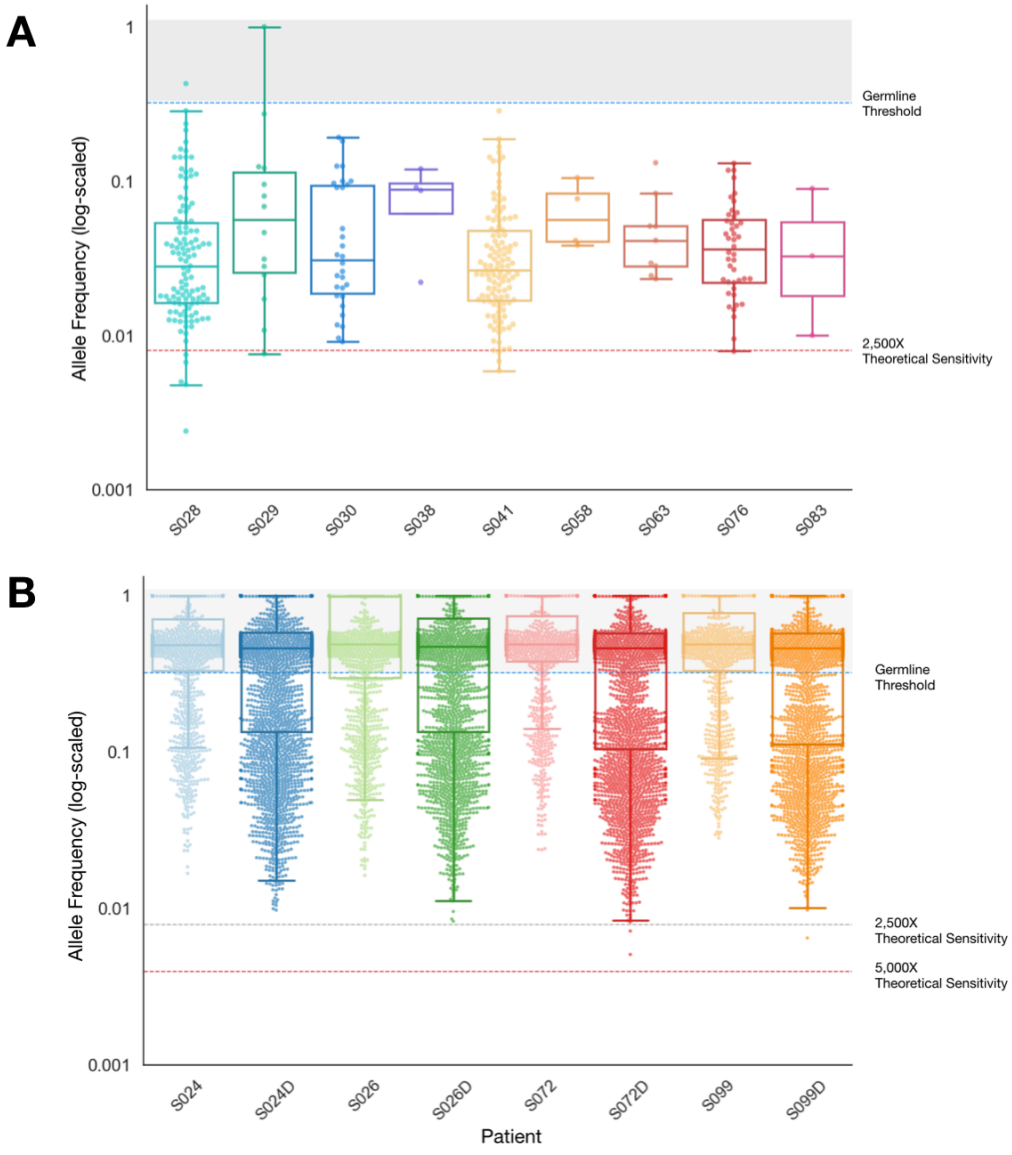


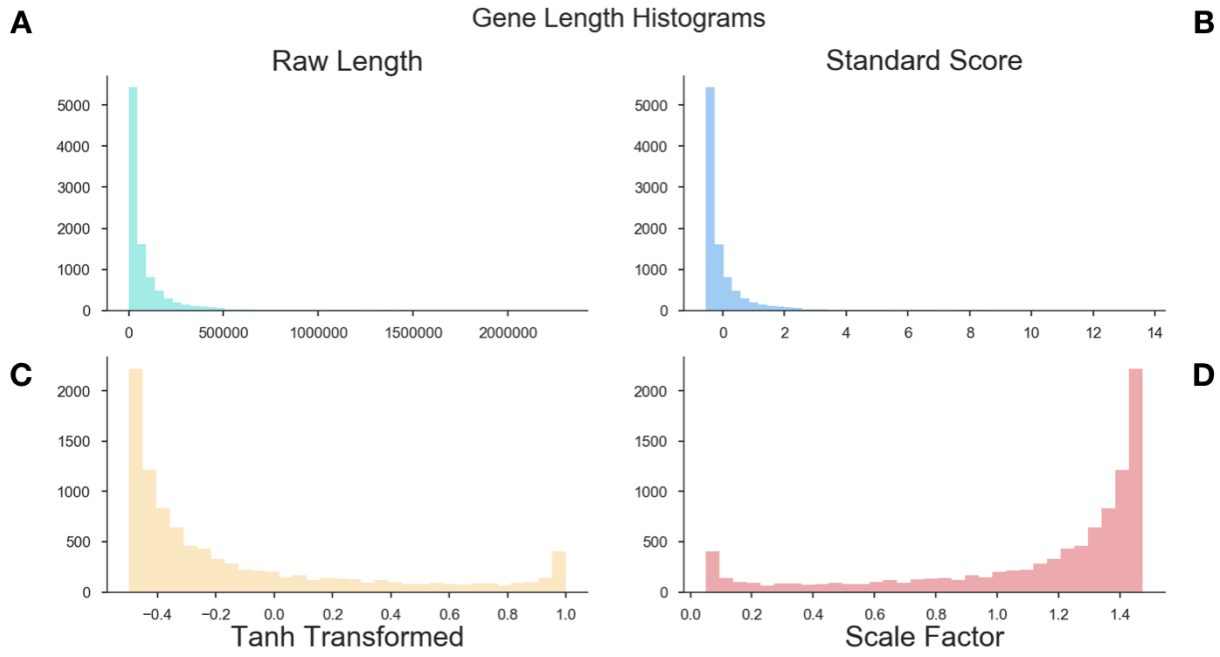
Figure 4.4: Variant detection using the Orchid generated targeted sequencing panel.

Legend: A) Nine patients with multiple tumor foci and normal tissue DNA were whole exome sequenced at 40X-fold coverage. Matched cfDNA was sequenced at 2,500X after targeted capture using the orchid generated panel. The number of detected variants (i.e., both somatic and present in cfDNA) and allele frequency distribution are shown. **B)** For four patients, cfDNA was sequenced at 5,000X with the targeted capture panel. Changes in the number and distribution frequency of detected variants are shown, demonstrating an increase in the number of detected raw variants at higher sequencing depth.



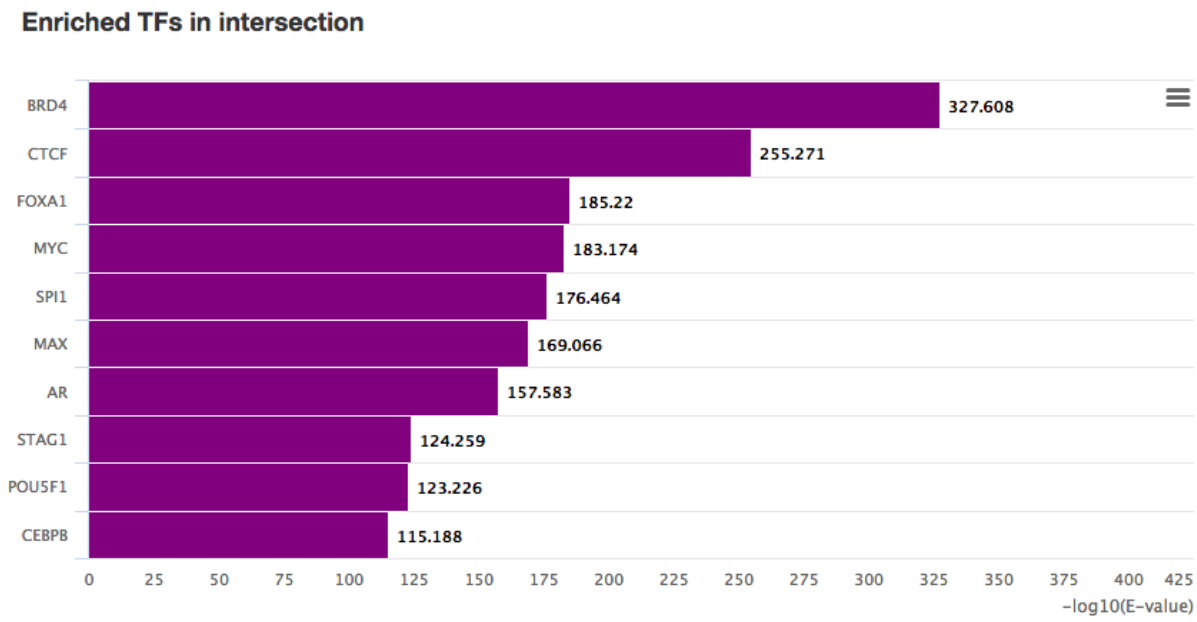
Supplemental Figure 4.1: Distribution of gene transcript lengths and scale factors for normalization.

Legend: A) A histogram of transcript lengths of genes associated with ranked mutations is shown. **B)** Gene lengths after standardization **C)** Due to the long tail of the distribution of standard scores, a tanh transform was used to compress scores between a -1 to 1 range. **D)** The tanh transformed scores were reversed and multiplied to mutation distance values, effectively down-weighting values of mutations in long genes and up-weighted values of short genes.



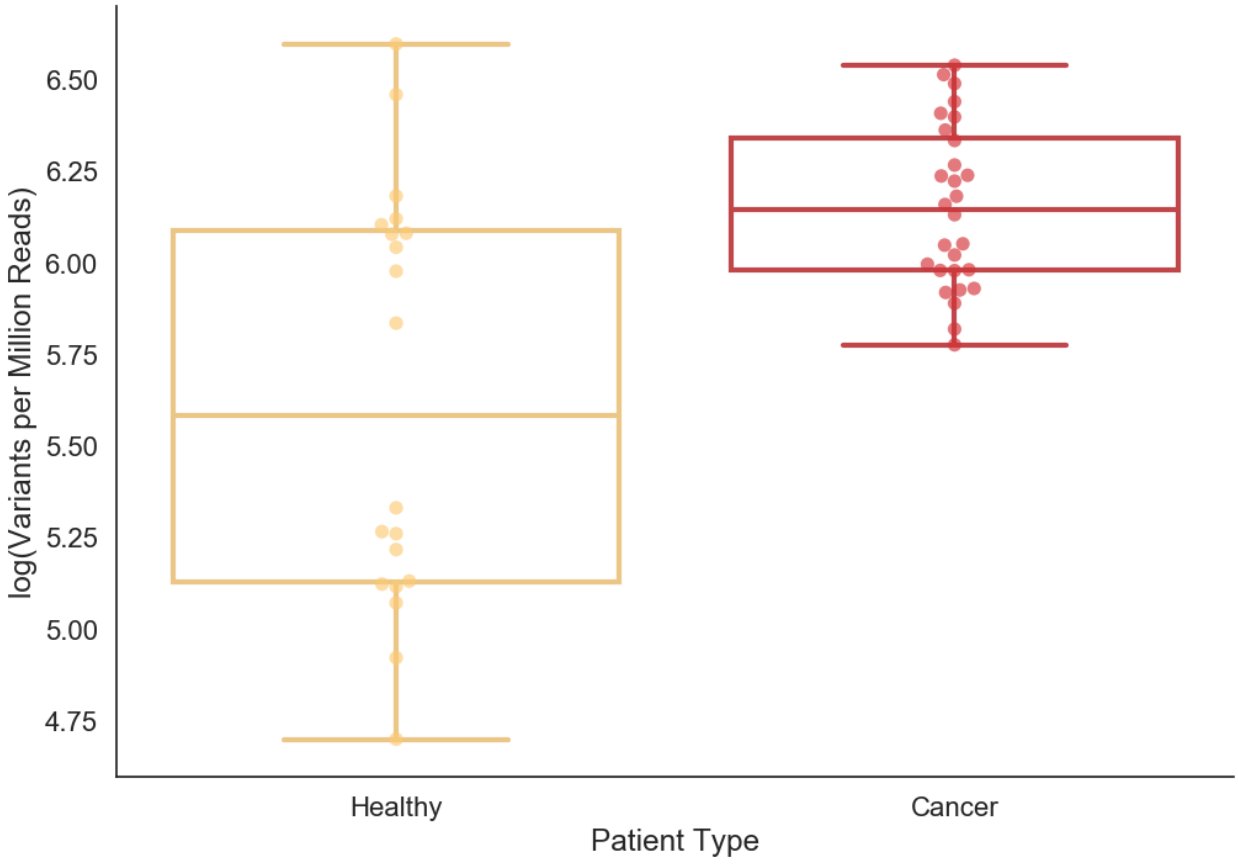
Supplemental Figure 4.2: Transcription Factors Enriched on the Orchid Panel.

Legend: All non-coding regions on the orchid panel were submitted to the Remap server to predict transcription factor enrichment using 10% overlap with our regions. Shown here are top 10 most enriched transcription factors and their enrichment factors. Most of these are associated with the regulation of cancer, and in some cases, prostate cancer.



Supplemental Figure 4.3: CHIP effects and cfDNA in healthy patients.

Legend: CfDNA from twenty-eight prostate cancer patients (19 with localized prostate cancer 9 with metastatic disease) and 20 healthy volunteers was sequencing at 2,500X after targeted capture using the orchid panel. The distribution in the number of variants per million reads is shown. The number of variants in cancer patients is significantly higher than in healthy volunteers (p-values < 6.0e-5) using a T-test.



Supplemental Table 4.1: List of existing panel genes.

Legend: Below is the list of genes from 4 combined cfDNA gene panels that were used for the 'union existing' panel design in this publication. Genes listed were found from online published material and come from the following panels:

1. UCSF 500: http://labmed.ucsf.edu/labmanual/db/resource/UCSF_CCGL500_REQUISITION_FORM_2016_withGeneList.pdf
2. Fluxion Biosciences: https://support.fluxionbio.com/hc/en-us/article_attachments/214690287/_634-0042_-_Spotlight_59_Data_Sheet_RevA.pdf
3. Foundation Medicine: https://www.foundationmedicineasia.com/dam/assets/pdf/FOne_Current_Gene_List.pdf
4. Guardant Health: http://www.guardant360.com/img/G360MicroSite73GenePanel.jpg

Note: These panels assess other types of tumor variants beyond simple somatic mutations (including amplifications and gene fusions) and cannot be compared directly to the orchid generated panel for this reason. This list was compiled as a consensus representation of the genes involved in cancer and suggestions a reasonable starting point for defining a targeted sequencing panel.

ABL1	ERRFI1	MPL	SMC1A
ABL2	ESPL1	MRE11A	SMC3
ACVR1	ESR1	MSH2	SMO
ACVR1B	ESR2	MSH3	SNCAIP
AJUBA	ETS1	MSH6	SOCS1
AKT1	ETV6	MTOR	SOS1
AKT2	EWSR1	MUTYH	SOS2
AKT3	EZH1	MYB	SOX10
ALK	EZH2	MYBL1	SOX2
AMER1	FAM46C	MYC	SOX9
APC	FANCA	MYCL	SPEN
APOBEC3G	FANCC	MYCN	SPOP
AR	FANCE	MYD88	SPRED1
ARAF	FANCF	MYH9	SPRY1
ARFRP1	FANCG	NAV3	SPRY2
ARHGAP35	FANCL	NBN	SPRY4
ARID1A	FAT1	NCKAP5	SPTA1

ARID1B	FAT3	NCOA2	SRC
ARID2	FBXW7	NCOA3	SRSF2
ARID5B	FGF10	NCOR1	SS18
ASH2L	FGF14	NF1	STAG2
ASXL1	FGF19	NF2	STAT3
ASXL2	FGF23	NFE2L2	STAT4
ATF1	FGF3	NFKBIA	STAT6
ATM	FGF4	NFKBIE	STK11
ATR	FGF6	NIPBL	SUFU
ATRX	FGFR1	NKX2-1	SYK
AURKA	FGFR2	NOTCH1	SYNE1
AURKB	FGFR3	NOTCH3	TADA1
AXIN1	FGFR4	NPM1	TBX3
AXIN2	FH	NRAS	TCEB1
AXL	FLCN	NSD1	TCF7L2
BAP1	FLT1	NT5C2	TERT
BARD1	FLT3	NTRK1	TET2
BCL2	FLT4	NTRK2	TFE3
BCL2A1	FOXA1	NTRK3	TFEB
BCL2L1	FOXL2	NUP93	TGFBR2
BCL2L12	FOXO1	NUTM1	TLR4
BCL2L2	FOXP1	OR5L1	TMPRSS2
BCL6	FRS2	PAK1	TNFAIP3
BCOR	FUBP1	PAK3	TNFRSF14
BCORL1	FUS	PALB2	TOP1
BLM	FYN	PARK2	TOP2A
BRAF	GAB2	PAX3	TP53
BRCA1	GATA1	PAX5	TRAF3
BRCA2	GATA2	PAX7	TRAF7
BRD4	GATA3	PAX8	TRIM28
BRIP1	GLI1	PBRM1	TSC1
BTG1	GLI2	PDCD1LG2	TSC2
BTK	GNA11	PDGFB	TSHR
C11orf30	GNA13	PDGFRA	TSHZ2
CALR	GNAQ	PDGFRB	TSHZ3
CARD11	GNAS	PDK1	TSLP
CBFB	GPC3	PHF6	TTYH1
CBL	GPR124	PHOX2B	TYK2
CBLB	GRIN2A	PIK3CA	U2AF1
CCND1	GRM3	PIK3CG	USP7

CCND2	GSK3B	PIK3R1	VEGFA
CCND3	H3F3A	PIK3R2	VHL
CCNE1	H3F3B	PLAG1	WHSC1
CD274	HDAC4	PLCB4	WISP3
CD79A	HDAC9	PMS1	WRN
CD79B	HEY1	POLD1	WT1
CDC42	HGF	POLE	XBP1
CDC73	HIF1A	POLQ	XPO1
CDH1	HIST1H3B	POT1	YAP1
CDK12	HMGA2	POU3F2	YWHAE
CDK4	HNF1A	PPM1D	ZBTB20
CDK6	HOXB13	PPP2R1A	ZFHX3
CDK8	HRAS	PPP6C	ZFHX4
CDKN1A	HSP90AB1	PRDM1	ZMYM3
CDKN1B	HSPA2	PREX2	ZNF217
CDKN2A	HSPA5	PRKACA	ZNF703
CDKN2B	ID3	PRKAG2	ZRSR2
CDKN2C	IDH1	PRKAR1A	HER2
CEBPA	IDH2	PRKCA	MEK1
CHD1	IGF1R	PRKCH	MEK2
CHD2	IGF2	PRKDC	ERK2
CHD4	IGF2R	PTCH1	MAPK3
CHD5	IKBKE	PTCH2	ERK1
CHEK1	IKZF1	PTEN	NOCH1
CHEK2	IKZF2	PTK2B	FAM123B
CIC	IKZF3	PTPN1	EMSY
CLDN18	IL2RB	PTPN11	PD-L1
CNOT3	IL7R	PTPRB	CRLF2
COL1A1	INHBA	PTPRD	DAXX
COL2A1	INPP4B	PTPRK	ERRFl1
CRCT1	IPMK	PTPRT	FANCD2
CREB1	IRF4	RAC1	FAS
CREBBP	IRS2	RAD21	GABRA6
CRKL	JAK1	RAD50	GATA4
CSF1R	JAK2	RAD51	GATA6
CSF3R	JAK3	RAD51C	GID4
CTCF	JAZF1	RAD51D	C17orf39
CTNNA1	KAT6A	RAF1	GLl1
CTNNB1	KDM5A	RARA	HSD3B1
CUL3	KDM5C	RASA1	HSP90AA1

CUX1	KDM6A	RASA2	IRF2
CXCR4	KDR	RB1	JUN
CYLD	KEAP1	RBM10	MYST3
DCC	KIT	REL	KEL
DDIT3	KLF4	RELA	MLL
DDR2	KLHL6	RET	KMT2C
DDX3X	KMT2A	RHEB	MLL3
DDX41	KMT2B	RHOA	MLL2
DGKH	KMT2D	RICTOR	LMO1
DICER1	KNSTRN	RIT1	LYN
DIS3	KRAS	RNF43	MAGI2
DNAJB1	LEF1	ROBO1	(MEK2)
DNMT3A	LIFR	ROS1	MYCL1
DOT1L	LRP1B	RPL10	NOTCH2
DUSP2	LZTR1	RPTOR	PD-L2
DUSP4	MALAT1	RRAGC	PIK3C2B
DUSP6	MAML2	RRAS	PIK3CB
DYNC1I1	MAP2K1	RRAS2	PLCG2
EBF1	MAP2K2	RSPO2	PMS2
EDNRB	MAP2K4	RSPO3	PRKCI
EGFR	MAP3K1	RUNX1	PRSS8
EGR1	MAP3K2	RUNX1T1	QKI
EIF1AX	MAP3K5	SDHB	RANBP2
ELF3	MAP3K7	SDHD	SDHA
EP300	MAP3K9	SETBP1	SDHC
EPCAM	MAPK1	SETD2	TAF1
EPHA2	MCL1	SF3B1	TERC
EPHA3	MDM2	SH2B3	ZBTB2
EPHA5	MDM4	SHH	
EPHA7	MED12	SIN3A	
EPHB1	MEF2B	SLIT2	
EPOR	MEN1	SLITRK6	
ERBB2	MET	SMAD2	
ERBB3	MGA	SMAD3	
ERBB4	MGMT	SMAD4	
ERCC1	MITF	SMARCA2	
ERCC2	MLH1	SMARCA4	
ERG	MLH3	SMARCB1	

References

1. Big Data: Astronomical or Genomical? Public Library of Science; 2015 Jul 7;13(7):e1002195. Available from: http://dx.plos.org/10.1371/journal.pbio.1002195

2. American Cancer Society. 8 ed. ACS Prostate Cancer.

3. Seer Cancer Statistics. 8 ed. Seer Cancer Statistics.

4. Roehrborn CG, Black LK. The economic burden of prostate cancer. BJU International. Wiley/Blackwell (10.1111); 2011 Aug 25;108(6):806–13.

5. Barbieri CE, Bangma CH, Bjartell A, Catto JWF, Culig Z, Grönberg H, et al. The Mutational Landscape of Prostate Cancer. European Urology. Elsevier; 2013 Oct 1;64(4):567–76.

6. Humphrey PA. Gleason grading and prognostic factors in carcinoma of the prostate. Modern Pathology 2004 17:3. Nature Publishing Group; 2004 Mar 1;17(3):292–306.

7. Boyd LK, Mao X, Lu Y-J. The complexity of prostate cancer: genomic alterations and heterogeneity. Nature Reviews Urology 2012 9:11. Nature Publishing Group; 2012 Nov 1;9(11):652–64.

8. Prensner JR, Rubin MA, Wei JT, Chinnaiyan AM. Beyond PSA: The Next Generation of Prostate Cancer Biomarkers. Sci Transl Med. American Association for the Advancement of Science; 2012 Mar 28;4(127):127rv3–127rv3.

9. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer Genome Landscapes. Science. American Association for the Advancement of Science; 2013 Mar 29;339(6127):1546–58.

10. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun X-W, et al. Recurrent Fusion of TMPRSS2 and ETS Transcription Factor Genes in Prostate Cancer. Science. American Association for the Advancement of Science; 2005 Oct 28;310(5748):644–8.

11. Bieberich CJ, Fujita K, He WW, Jay G. Prostate-specific and androgen-dependent expression of a novel homeobox gene. J Biol Chem. American Society for Biochemistry and Molecular Biology; 1996 Dec 13;271(50):31779–82.

12. Millar DS, Ow KK, Paul CL, Russell PJ, Molloy PL, Clark SJ. Detailed methylation analysis of the glutathione S-transferase π (GSTP1) gene in prostate cancer. Oncogene. Nature Publishing Group; 1999 Feb 1;18(6):1313–24.

13. Nelson WG, de Marzo AM, Lippman SM. Prostate Cancer Prevention. In: Cancer Chemoprevention. Totowa, NJ: Humana Press; 2005. pp. 185–203.

14. Abate-Shen C, Shen MM. Molecular genetics of prostate cancer. Genes Dev. Cold

Spring Harbor Lab; 2000 Oct 1;14(19):2410–34.

15. Raphael BJ, Dobson JR, Oesper L, Vandin F. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. Genome Medicine 2014 6:1. BioMed Central; 2014 Dec 1;6(1):5.

16. Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, et al. MuSiC: identifying mutational significance in cancer genomes. Genome Res. 2012 Aug;22(8):1589–98.

17. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. Nature Publishing Group; 2013 Jul 11;499(7457):214–8.

18. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. Nature Publishing Group; 2010 Apr;7(4):248–9.

19. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. de Brevern AG, editor. PLoS ONE. Public Library of Science; 2012;7(10):e46688.

20. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. Nature Publishing Group; 2009;4(7):1073–81.

21. Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, et al. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer Res. 2009 Aug 15;69(16):6660–7.

22. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS. National Academy of Sciences; 2005 Oct 25;102(43):15545–50.

23. Vandin F, Upfal E, Raphael BJ. De novo discovery of mutated driver pathways in cancer. Genome Res. Cold Spring Harbor Lab; 2012 Feb;22(2):375–85.

24. Chen Y, Sun J, Huang L-C, Xu H, Zhao Z. Classification of Cancer Primary Sites Using Machine Learning and Somatic Mutations. BioMed Research International. Hindawi; 2015 Oct 11;2015(6):1–9.

25. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nature Reviews Genetics 2015 16:6. Nature Publishing Group; 2015 Jun 1;16(6):321–32.

26. Holdhoff M, Schmidt K, Cancer RDOTN, 2009. Analysis of Circulating Tumor DNA to Confirm Somatic KRAS Mutations. academicoupcom.

27. Stroun M, Anker P, Maurice P, Lyautey J, Lederrey C, Beljanski M. Neoplastic characteristics of the DNA found in the plasma of cancer patients. OCL. Karger Publishers; 1989;46(5):318–22.

28. Lo YMD. Non-invasive prenatal diagnosis by massively parallel sequencing of maternal plasma DNA. Open Biology. Royal Society Journals; 2012 Jun 1;2(6):120086–6.

29. Brar H, Wang E, Struble C, Musci TJ, Norton ME. The fetal fraction of cell-free DNA in maternal plasma is not affected by a priori risk of fetal trisomy. The Journal of Maternal-Fetal & Neonatal Medicine. Taylor & Francis; 2012 Aug 23;26(2):143–5.

30. Bianchi DW, Parker RL, Wentworth J, Madankumar R, Saffer C, Das AF, et al. DNA Sequencing versus Standard Prenatal Aneuploidy Screening. http://dxdoiorg/101056/NEJMoa1311037. Massachusetts Medical Society; 2014 Feb 26;370(9):799–808.

31. De Vlaminck I, Valantine HA, Snyder TM, Strehl C, Cohen G, Luikart H, et al. Circulating Cell-Free DNA Enables Noninvasive Diagnosis of Heart Transplant Rejection. Sci Transl Med. American Association for the Advancement of Science; 2014 Jun 18;6(241):241ra77–7.

32. Beck J, Bierau S, Balzer S, Andag R, Kanzow P, Schmitz J, et al. Digital Droplet PCR for Rapid Quantification of Donor DNA in the Circulation of Transplant Recipients as a Potential Universal Biomarker of Graft Injury. Clin Chem. Clinical Chemistry; 2013 Jan 1;59(12):clinchem.2013.210328–1741.

33. Lou X, Hou Y, Liang D, Peng L, Chen H, Ma S, et al. A novel Alu-based real-time PCR method for the quantitative detection of plasma circulating cell-free DNA: Sensitivity and specificity for the diagnosis of myocardial infarction. International Journal of Molecular Medicine. Spandidos Publications; 2015 Jan 1;35(1):72–80.

34. Diaz LA, Bardelli A. Liquid biopsies: genotyping circulating tumor DNA. J Clin Oncol. 2014 Feb;32(6):579–86.

35. LO YMD. Circulating Nucleic Acids in Plasma and Serum: An Overview. Annals of the New York Academy of Sciences. Wiley/Blackwell (10.1111); 2001 Sep 1;945(1):1–7.

36. Diaz LA, Bardelli A. Liquid biopsies: genotyping circulating tumor DNA. J Clin Oncol. 2014 Feb 20;32(6):579–86.

37. Heitzer E, Ulz P, Geigl JB. Circulating tumor DNA as a liquid biopsy for cancer. Clin Chem. Clinical Chemistry; 2015 Jan;61(1):112–23.

38. Schwarzenbach H, Hoon DSB, Pantel K. Cell-free nucleic acids as biomarkers in cancer patients. Nat Rev Cancer. Nature Publishing Group; 2011 Jun 1;11(6):426–37.

39. Sozzi G, Conte D, Leon M, Ciricione R, Roz L, Ratcliffe C, et al. Quantification of free circulating DNA as a diagnostic marker in lung cancer. J Clin Oncol. 2003 Nov;21(21):3902–8.

40. Chan KCA, Jiang P, Zheng YWL, Liao GJW, Sun H, Wong J, et al. Cancer Genome Scanning in Plasma: Detection of Tumor-Associated Copy Number Aberrations, Single-Nucleotide Variants, and Tumoral Heterogeneity by Massively Parallel Sequencing. Clin Chem. Clinical Chemistry; 2012 Jan 1;59(1):clinchem.2012.196014–224.

41. Leary RJ, Sausen M, Kinde I, Papadopoulos N, Carpten JD, Craig D, et al. Detection of Chromosomal Alterations in the Circulation of Cancer Patients with Whole-Genome Sequencing. Sci Transl Med. American Association for the Advancement of Science; 2012 Nov 28;4(162):162ra154–4.

42. Shaw JA, Page K, Blighe K, Hava N, Guttery D, Ward B, et al. Genomic analysis of circulating cell-free DNA infers breast cancer dormancy. Genome Res. Cold Spring Harbor Lab; 2012 Feb;22(2):220–31.

43. Siravegna G, Bardelli A. Genotyping cell-free tumor DNA in the blood to detect residual disease and drug resistance. Genome Biol. BioMed Central; 2014 Aug 1;15(8):449.

44. Murtaza M, Dawson S-J, Tsui DWY, Gale D, Forshew T, Piskorz AM, et al. Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. Nature. Nature Publishing Group; 2013 May;497(7447):108–12.

45. Thierry AR, Messaoudi SE, Lopez-crapez E. Circulating Nucleic Acids in Early Diagnosis, Prognosis and Treatment Monitoring. Gahan PB, editor. Dordrecht: Springer Netherlands; 2015;5.

46. Valtorta E, Misale S, Bianchi AS, Nagtegaal ID, Paraf F, Lauricella C, et al. KRAS gene amplification in colorectal cancer and impact on response to EGFR-targeted therapy. Int J Cancer. Wiley-Blackwell; 2013 Sep 1;133(5):1259–65.

47. Diehl F, Schmidt K, Choti MA, Romans K, Goodman S, Li M, et al. Circulating mutant DNA to assess tumor dynamics. Nat Med. Nature Publishing Group; 2008 Sep;14(9):985–90.

48. NO JH, KIM K, PARK KH, KIM Y-B. Cell-free DNA level as a prognostic biomarker for epithelial ovarian cancer. Anticancer Res. International Institute of Anticancer Research; 2012 Aug;32(8):3467–71.

49. Kim K, Shin DG, Park MK, Baik SH, Kim TH, Kim S, et al. Circulating cell-free DNA as a promising biomarker in patients with gastric cancer: diagnostic validity and significant reduction of cfDNA after surgical resection. Annals of Surgical Treatment and Research. 2014 Mar 1;86(3):136–42.

50. Frattini M, Gallino G, Signoroni S, Balestra D, Lusa L, Battaglia L, et al. Quantitative and qualitative characterization of plasma DNA identifies primary and recurrent colorectal cancer. Cancer Letters. Elsevier; 2008 May 18;263(2):170–81.

51. Heitzer E, Auer M, Hoffmann EM, Pichler M, Gasch C, Ulz P, et al. Establishment of tumor-specific copy number alterations from plasma DNA of patients with cancer. Int J Cancer. Wiley-Blackwell; 2013 Jul 15;133(2):346–56.

52. Ellinger J, Bastian PJ. Cell-Free DNA: A Novel Biomarker for Patients with Prostate Cancer? The Open Prostate Cancer Journal. 2010 May;3(1):57–62.

53. Feng J, Gang F, Li X, Jin T, Houbao H, Yu C, et al. Plasma cell-free DNA and its DNA integrity as biomarker to distinguish prostate cancer from benign prostatic hyperplasia in patients with increased serum prostate-specific antigen. International urology and nephrology. 2013 Aug;45(4):1023–8.

54. Delgado PO, Alves BCA, De Sousa Gehrke F, Kuniyoshi RK, Wroclavski ML, Del Giglio A, et al. Characterization of cell-free circulating DNA in plasma in patients with prostate cancer. Tumor Biology. 2013;34:983–6.

55. Elshimali Y, Khaddour H, Sarkissyan M, Wu Y, Vadgama J. The Clinical Utilization of Circulating Cell Free DNA (CCFDNA) in Blood of Cancer Patients. International Journal of Molecular Sciences 2013, Vol 14, Pages 18925-18958. Multidisciplinary Digital Publishing Institute; 2013 Sep 13;14(9):18925–58.

56. Kwee S, Song MA, Cheng I, Loo L, Tiirikainen M. Measurement of Circulating Cell-Free DNA in Relation to 18F-Fluorocholine PET/CT Imaging in Chemotherapy-Treated Advanced Prostate Cancer. Clinical and Translational Science. Wiley/Blackwell (10.1111); 2012 Feb 1;5(1):65–70.

57. Azad AA, Volik SV, Wyatt AW, Haegert A, Le Bihan S, Bell RH, et al. Androgen receptor gene aberrations in circulating cell-free DNA: biomarkers of therapeutic resistance in castration-resistant prostate cancer. Clin Cancer Res. American Association for Cancer Research; 2015 Feb 23;21(10):clincanres.2666.2014–324.

58. Heitzer E, Ulz P, Geigl JB. Circulating Tumor DNA as a Liquid Biopsy for Cancer. Clin Chem. 2014 Nov;123.

59. Newman AM, Bratman SV, To J, Wynne JF, Eclov NCW, Modlin LA, et al. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. Nat Med. Nature Publishing Group; 2014 May 1;20(5):548–54.

60. Ding L, Wendl MC, McMichael JF, Raphael BJ. Expanding the computational toolbox for mining cancer genomes. Nature Reviews Genetics 2015 16:6. Nature Publishing Group; 2014 Aug 1;15(8):556–70.

61. Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, et al. Cancer-Specific High-Throughput Annotation of Somatic Mutations: Computational

Prediction of Driver Missense Mutations. Cancer Res. American Association for Cancer Research; 2009 Aug 15;69(16):6660–7.

62. Kumar RD, Swamidass SJ, Bose R. Unsupervised detection of cancer driver mutations with parsimony-guided learning. Nat Genet. Nature Publishing Group; 2016 Oct;48(10):1288–94.

63. Tan H, Bao J, Bioinformatics XZ, 2012. A novel missense-mutation-related feature extraction scheme for “driver”mutation identification. academicoupcom.

64. Lindquist KJ, Jorgenson E, Hoffmann TJ, Witte JS. The impact of improved microarray coverage and larger sample sizes on future genome-wide association studies. Genet Epidemiol. 2013 May;37(4):383–92.

65. Marouli E, Graff M, Medina-Gomez C, Lo KS, Wood AR, Kjaer TR, et al. Rare and low-frequency coding variants alter human adult height. Nature. Nature Publishing Group; 2017 Feb 9;542(7640):186–90.

66. Figueiredo JC, Stram DO, Haiman CA. The Impact of GWAS Findings on Cancer Etiology and Prevention. Current Epidemiology Reports. Springer International Publishing; 2014 Jul 3;1(3):130–7.

67. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet. 2017 Jul 6;101(1):5–22.

68. Van Rossum T, Tripp B, Daley D. SLIMS--a user-friendly sample operations and inventory management system for genotyping labs. Bioinformatics. 2010 Jun 30;26(14):1808–10.

69. Andersen D, Rasmussen B, Linnet K. Validation of a fully automated robotic setup for preparation of whole blood samples for LC-MS toxicology analysis. J Anal Toxicol. 2012 May;36(4):280–7.

70. Kong F, Yuan L, Zheng YF, Chen W. Automatic liquid handling for life science: a critical review of the current state of the art. J Lab Autom. SAGE PublicationsSage CA: Los Angeles, CA; 2012 Jun;17(3):169–85.

71. Voegele C, Tavtigian SV, de Silva D, Cuber S, Thomas A, Le Calvez-Kelm F. A Laboratory Information Management System (LIMS) for a high throughput genetic platform aimed at candidate gene mutation screening. Bioinformatics. 2007 Sep 15;23(18):2504–6.

72. Thurow K, Göde B, Dingerdissen U, Stoll N. Laboratory Information Management Systems for Life Science Applications. Organic Process Research & Development. American Chemical Society; 2004 Nov;8(6):970–82.

73. Ortiz L, Pavan M, McCarthy L, Timmons J, Densmore DM. Automated Robotic

Liquid Handling Assembly of Modular DNA Devices. J Vis Exp. 2017 Dec 1;(130):e54703–3.

74. Arvidsson S, Kwasniewski M, Riaño-Pachón DM, Mueller-Roeber B. QuantPrime--a flexible tool for reliable high-throughput primer design for quantitative PCR. BMC Bioinformatics. BioMed Central; 2008 Nov 1;9(1):465.

75. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. Nature Publishing Group; 2014 Mar;46(3):310–5.

76. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics. 2015 Mar 1;31(5):761–3.

77. Pavlidis N, Pentheroudakis G. Cancer of unknown primary site. Lancet. 2012 Apr 14;379(9824):1428–35.

78. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). Taylor & Francis; 2012 Apr;6(2):80–92.

79. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. Oxford University Press; 2000 Jan 1;28(1):27–30.

80. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. Cold Spring Harbor Lab; 2010 Jan;20(1):110–21.

81. Griffiths-Jones S. The microRNA Registry. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D109–11.

82. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods. Nature Publishing Group; 2012 Feb 28;9(3):215–6.

83. Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat Methods. Nature Publishing Group; 2012 Mar 18;9(5):473–6.

84. Rajagopal N, Xie W, Li Y, Wagner U, Wang W, Stamatoyannopoulos J, et al. RFECS: A Random-Forest Based Algorithm for Enhancer Identification from Chromatin State. Singh M, editor. PLoS Comput Biol. Public Library of Science; 2013 Mar 14;9(3):e1002968.

85. Khan A, Zhang X. dbSUPER: a database of super-enhancers in mouse and human genome. Nucleic Acids Res. 2016 Jan 4;44(D1):D164–71.

86. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M,

Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. Nature Publishing Group; 2015 Feb 19;518(7539):317–30.

87. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. Nature. Nature Publishing Group; 2012 Sep 1;489(7414):75–82.

88. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. Nature Publishing Group; 2013 Aug 22;500(7463):415–21.

89. Fu Y, Liu Z, Lou S, Bedford J, Mu XJ, Yip KY, et al. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. Genome Biol. BioMed Central; 2014;15(10):480.

90. Agarwal V, Bell GW, Nam J-W, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. Elife. eLife Sciences Publications Limited; 2015 Aug 12;4:101.

91. Griffon A, Barbier Q, Dalino J, van Helden J, Spicuglia S, Ballester B. Integrative analysis of public ChIP-seq experiments reveals a complex multi-cell regulatory landscape. Nucleic Acids Res. 2015 Feb 27;43(4):e27–7.

92. Di Tommaso P, Chatzou M, Floden EW, Nature PB, 2017. Nextflow enables reproducible computational workflows. naturecom

93. Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. Cell. Cell Press; 2016 Jan 14;164(1-2):57–68.

94. Marquard AM, Birkbak NJ, Thomas CE, Favero F, Krzystanek M, Lefebvre C, et al. TumorTracer: a method to identify the tissue of origin from the somatic mutations of a tumor specimen. BMC Med Genomics. BioMed Central; 2015 Oct 1;8(1):58.

95. Agah S, Akbari A, Talebi A, Masoudi M, Sarveazad A, Mirzaei A, et al. Quantification of Plasma Cell-Free Circulating DNA at Different Stages of Colorectal Cancer. Cancer Investigation. Taylor & Francis; 2017 Dec 15;35(10):625–32.

96. Sozzi G, Musso K, Ratcliffe C, Goldstraw P, Pierotti MA, Pastorino U. Detection of microsatellite alterations in plasma DNA of non-small cell lung cancer patients: a prospect for early diagnosis. Clin Cancer Res. 1999 Oct;5(10):2689–92.

97. Tie J, Semira C, Gibbs P. Circulating tumor DNA as a biomarker to guide therapy in post-operative locally advanced rectal cancer: the best option? Expert Review of Molecular Diagnostics. Taylor & Francis; 2017 Oct 6;18(1):1–3.

98. Dawson S-J, Tsui DWY, Murtaza M, Biggs H, Rueda OM, Chin S-F, et al. Analysis

of circulating tumor DNA to monitor metastatic breast cancer. N Engl J Med. Massachusetts Medical Society; 2013 Mar 28;368(13):1199–209.

99. Forshew T, Murtaza M, Parkinson C, Gale D, Tsui DWY, Kaper F, et al. Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. Sci Transl Med. 2012 May;4(136):136ra68.

100. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. Detection and quantification of rare mutations with massively parallel sequencing. Proceedings of the National Academy of Sciences of the United States of America. 2011 Jul;108(23):9530–5.

101. Barbieri CE, Baca SC, Lawrence MS, Demichelis F, Blattner M, Theurillat J-P, et al. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nat Genet. Nature Publishing Group; 2012 May 20;44(6):685–9.

102. Cario CL, Witte JS. Orchid: a novel management, annotation and machine learning framework for analyzing cancer mutations. Hancock J, editor. Bioinformatics. 2018 Mar 15;34(6):936–42.

103. Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, et al. SpeedSeq: ultra-fast personal genome analysis and interpretation. Nat Methods. Nature Publishing Group; 2015 Oct;12(10):966–8.

104. Miyata Y, Watanabe S-I, Matsuo T, Hayashi T, Sakai H, Xuan JW, et al. Pathological significance and predictive value for biochemical recurrence of c-Fes expression in prostate cancer. Prostate. 2012 Feb 1;72(2):201–8.

105. Zhou J, Yang Z, Tsuji T, Gong J, Xie J, Chen C, et al. LITAF and TNFSF15, two downstream targets of AMPK, exert inhibitory effects on tumor growth. Oncogene. Nature Publishing Group; 2011 Apr 21;30(16):1892–900.

106. De Luca A, Sacchetta P, Nieddu M, Di Ilio C, Favaloro B. Important roles of multiple Sp1 binding sites and epigenetic modifications in the regulation of the methionine sulfoxide reductase B1 (MsrB1) promoter. BMC Mol Biol. BioMed Central; 2007 May 22;8(1):39.

107. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. BioMed Central; 2013 Apr 15;14(1):128.

108. Dhillon AS, Hagan S, Rath O, Kolch W. MAP kinase signalling pathways in cancer. Oncogene. 2007 May 14;26(22):3279–90.

109. Desgrosellier JS, Cheresh DA. Integrins in cancer: biological implications and therapeutic opportunities. Nat Rev Cancer. Nature Publishing Group; 2010 Jan;10(1):9–22.

110. Hu Y, Ulrich BC, Supplee J, Kuang Y, Lizotte PH, Feeney NB, et al. False-Positive Plasma Genotyping Due to Clonal Hematopoiesis. Clin Cancer Res. 2018 Mar 22.

Funding

This work conducted in this dissertation was supported by the following funding sources:

National Institute of Health (NIH) grants: CA088164 and CA201358

The UCSF Goldberg-Benioff Program in Cancer Translational Biology

Amazon Web Services

Microsoft Azure Web Services.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Author Signature Christine Cassi Date 9/10/2018