Spring 5-10-2023

Integrative analysis of cell-free DNA liquid biopsy data

Irfan Alahi

WASHINGTON UNIVERSITY IN ST. LOUIS

McKelvey School of Engineering
Department of Computer Science & Engineering

Thesis Examination Committee:
Aadel Chaudhuri, Chair
Roger Chamberlain
Cynthia Ma

Integrative Analysis of Cell-Free DNA Liquid Biopsy Data
by
Irfan Alahi

A thesis presented to
the McKelvey School of Engineering
of Washington University in
partial fulfillment of the
requirements for the degree
of Master of Science

May 2023
St. Louis, Missouri

Table of Contents

List of Figures

Acknowledgments

I would like to express my sincere gratitude to my thesis committee and advisor for their support, guidance, and encouragement. Their invaluable feedback, insightful comments, and constructive criticism helped me to shape my ideas and refine my research. I am truly grateful for their time, expertise, and dedication in helping me to achieve my goals. In addition, this study would not be possible without the clinical samples. We are grateful to the patients, their families, and healthy volunteers. The first project was supported by grants from the V Foundation for Cancer Research V Scholar Award (A.A.C.), the Cancer Research Foundation Young Investigator Award (A.A.C.), the Washington University Alvin J. Siteman Cancer Research Fund (A.A.C.), and two grant from the National Institutes of Health (NIGMS R35GM142710 and NCI K08CA238711 to A.A.C.). The second project was supported by the National Cancer Institute (NCI) under award number K08CA238711 (A.A.C.), and the National Center for Advancing Translational Sciences (NCATS) under award number UL1TR002345 (Principal Investigator, Bradley Evanoff; A.A.C.) and was additionally supported by the Washington University School of Medicine Dean's Medical Student Research Fellowship (A.S), the Midwest Stone Institute (Z.L.S.), the Rabushka Bladder Cancer Research Fund (Z.L.S. and A.A.C.), the Alvin J. Siteman Cancer Research Fund (A.A.C.), the Cancer Research Foundation Young Investigator Award (A.A.C.), and the V Foundation V Scholar Award (A.A.C.).

Irfan Alahi

Washington University in St. Louis
May 2023

ABSTRACT OF THE THESIS

Integrative Analysis of Cell-Free DNA Liquid Biopsy Data

by

Irfan Alahi

Master of Science in Computer Science

Washington University in St. Louis, 2023

Assistant Professor Aadel Chaudhuri, Chair

Liquid Biopsy is a revolutionary idea where researchers try to examine tumors from bodily fluids like blood and urine noninvasively. However, bodily fluids carry lots of other things from all over the body making Liquid Biopsy a very challenging task. In this work, we aim to study and develop computational methods to facilitate Liquid Biopsy and thus predict cancer treatment outcomes. We focus on two challenges of Liquid Biopsy: identifying tissue of origin and molecular residual disease (MRD) detection.

Identifying tissue of origin from biofluid is an important component of Liquid Biopsy. In this respect, methylation data is a promising biomarker as different cell states have different methylation patterns. We have developed an algorithm for methylation sequencing data to identify set of informative CpGs to discriminate specific cell types and states in high resolution. We tested the algorithm on publicly available data and then validated its performance on an independent cohort and melanoma patients' cfDNA. The algorithm successfully identified DMRs that were associated with specific cell states, highlighting its potential as a tool for tissue-of-origin detection.

Next, we focus on detecting the molecular residual disease (MRD) on bladder cancer as bladder cancer has a high rate of recurrence even after successful treatment. We integrated ultra-low-pass whole genome sequencing (ULP-WGS) with urine cancer personalized profiling by deep sequencing (uCAPP-Seq) to achieve sensitive MRD detection and predict overall survival. A random forest model incorporating these urine cfDNA-derived factors with leave-one-out cross-validation was 87% sensitive for predicting residual disease in reference to gold-standard surgical pathology.

Overall, the development of the high resolution DMR algorithm and MRD detection model presents promising avenues for Liquid Biopsy. These tools may provide clinicians with a more comprehensive understanding of a patient's disease status and inform treatment decisions.

Chapter 1

Introduction

Tumor biopsy is an invasive method where cells are extracted directly from the tumor for in-depth examination, and is one of the first steps used by clinicians to diagnose cancer. Based on analysis of the extracted cells, doctors determine the pathway of treatment. Though this invasive method is the standard practice for solid tumor malignancies, it can be expensive, risky and sometimes impractical. Particularly, monitoring a treatment response precisely would require serial biopsies which is not feasible. Liquid Biopsy is an alternative idea where researchers try to examine tumors from body fluids like blood [2, 29].

In blood plasma, scientists discovered the presence of cell-free DNA (cfDNA) more than 100 years ago [31]. As the name suggests, these DNA fragments are not contained within cells but rather circulate within blood plasma. When cells die, some of their DNA fragments are released into blood circulation where they can be captured and measured within cfDNA [22, 20]. My mentor and others have shown that a fraction of these circulating DNA fragments arise from malignant cells in patients with cancer [7, 10, 57, 55]. These cancer-specific cell-free DNA fragments are known as circulating tumor DNA (ctDNA) [22, 62].

Mutation as Biomarker

Circulating tumor DNA or ctDNA are DNA fragments coming from cancer cells. Cancer is a disease that begins with genomic mutations. Based on tracking these mutational signatures, ctDNA can be quantified from cfDNA sequencing. While several ctDNA detection technologies exist, they generally work by querying genomic positions likely to be mutated in cancer cells and deeply sequencing these positions (known as targeted sequencing) in plasma cfDNA. After targeted sequencing, the pre-defined genomic positions are interrogated for mutations, and in this way ctDNA molecules are detected and quantified [14, 37, 39].

ctDNA detection is also affected by background noise. Noise can be introduced during sample preparation and sequencing. This can confound results when quantifying rare ctDNA fragments in patients with low burdens of disease (i.e., early-stage cancer or post-curative-intent-treatment minimal residual cancer) [7, 10, 41]. Duplex variant support, where both positive and negative strands are sequenced, and the mutated variant is corroborated in both parent strands of DNA, reduces this noise significantly [24]. However, requiring duplex variant support is an inefficient approach as 80-90% of recovered cell-free DNA sequencing reads are typically single-stranded without duplex support [41]. Another approach to reduce noise is to profile the background error pattern by sequencing healthy donor-derived cell-free DNA and to account for it while querying mutations in patient cfDNA [41]. A recent paper utilizes a different approach to reduce background noise by requiring co-detection of adjacent mutations within the same cfDNA fragment [26].

Methylation as Biomarker

In DNA, cytosine (C) followed by guanine (G) are known as CpG (the ‘p’ stands for the phosphate bond between them). Through an epigenetic mechanism, a methyl (CH3) group can be added to the C of a CpG site. This phenomenon is known as CpG methylation. It turns out that different cell types and states have specific methylation patterns that regulate gene expression [34]. To quantify methylation patterns in DNA, bisulfite treatment followed by next-generation sequencing is commonly used. Briefly, in this bisulfite-based sequencing method, if a C in a CpG site is not methylated, the C converts to Uracil (U) which is subsequently recognized as Thymine (T). On the other hand, if a C in CpG site is methylated, the C remains as it is. Finally, the sequenced reads are aligned to the reference genome and for every CpG position, the ratio of C and T is calculated [30, 25]. Recent studies have shown potential utility of cfDNA methylation to detect cancer early including determining the tumor tissue origin [63, 35, 53, 19, 27]. However, TILs have not been profiled using CpG methylation from cfDNA yet.

Several recent studies demonstrate methylation-based cell-free DNA analysis to detect ctDNA and predict tumor tissue of origin [63, 35, 53, 19, 27]. Methylation-based ctDNA detection mainly has two steps: 1) like mutational signatures, cancer-specific methylation signatures are first identified; 2) deconvolution techniques are used to detect the ctDNA. are used to detect the ctDNA and infer tumor tissue of origin. Moss and colleagues [35] and Shen and colleagues [53] demonstrated that by using differentially methylated CpGs, the tissue origin of ctDNA can be identified. Guo and colleagues [19] showed that instead of single CpGs, co-associated adjacent CpGs (which they termed Methylation Haplotype Blocks or MHBs), can be used to more accurately deconvolve methylation data. They identified deferentially methylated MHBs and detected ctDNA and tissue of origin from cfDNA using a random forest classifier. Another approach can be to classify the aligned reads individually with the help of methylation patterns of that read. CancerDetector [27] is such a method where based on a beta binomial model the authors try to assign every cfDNA sequencing read as either cancer-derived or not cancer-derived.

In this study, we developed two frameworks in the Liquid Biopsy domain. First, we studied the problem of differentiation between closely related cell states and developed a method that considers known biology to select important features. Moreover, we demonstrated clinical application in CRC and melanoma cancer patients. In the second study, we developed a machine-learning model to predict molecular residual disease (MRD) from bladder cancer patients using mutation and copy number alteration as biomarkers. Both tools present promising directions for Liquid Biopsy and can be extended to clinical settings.
Briefly, our contributions are:

- We developed a high-resolution signature matrix generation framework
- Our method can differentiate among closely related cell states with potential application in Liquid Biopsy
- In addition, We developed a multi-modal urine cfDNA method, to sensitively detect MRD and predict pCR in bladder cancer patients.
- Our technology also predicted survival significantly and comparably to gold-standard surgical pathologic analysis of resected tumor tissue.

In Chapter 2, we discuss the first project where we develop an algorithm to identify high-resolution deferentially methylated regions. It can differentiate between different cell states and have potential clinical applications that we demonstrate in CRC and melanoma cancer patients. In Chapter 3, we develop a framework to predict MRD non-invasively in bladder cancer patients based on urine. It correlates with invasive pathological gold standards and associates with survival analysis significantly. Finally, in Chapter 4, we provided a summary of our work and possible future directions.

Chapter 2

High-resolution cell state-specific methylation profile to monitor TME in blood cfDNA

2.1 Introduction

Immune checkpoint inhibitors (ICI) are a promising way to treat advanced-stage cancer patients. However, not all patients benefit from this treatment, and it is challenging to know who will benefit and who will not. Serious side effects with ICI treatment may occur, emphasizing the importance of improving patient selection so that only patients who will benefit get treated. ICI treatment response can be predicted early by tumor biopsy analysis, however there is no noninvasive method to derive this data.

Solid tumors can be divided into two parts: malignant cancer cells and other cells of the body intermixed with the the malignant cancer cells. These non-malignant immune and stromal cells enveloping malignant cancer cells are known as the Tumor Microenvironment (TME) and play a critical role in promoting tumor cell growth versus death [3, 23]. Malignant cells can change the TME in such a way that the immune cells in the TME cannot effectively kill the cancer cells [16, 60]. Immune Checkpoint Inhibitors (ICI) can "take the breaks off" these immune cells and turn them into more potent cancer-killers. The ability of ICI to kill otherwise unresponsive tumors has transformed the treatment of advanced tumors [49, 51]. Unfortunately, not all patients respond to ICI. As ICI transform the immune cell compartment of the TME into cancer-killing cells, the treatment response largely depends on the cellular composition of the tumor [58, 59, 9, 15, 21, 50, 61]. For example, the TME of a tumor may lack immune cells with cancer-killing potential [49, 51, 16, 58]. Therefore, monitoring the TME before and during treatment is critical. However, monitoring the TME requires invasive biopsy. Serial biopsy of a patient is not practical and can suffer from sampling bias due to the heterogeneity of the tumor. cfDNA based deconvolution model can be a promising alternative to serial biopsy.

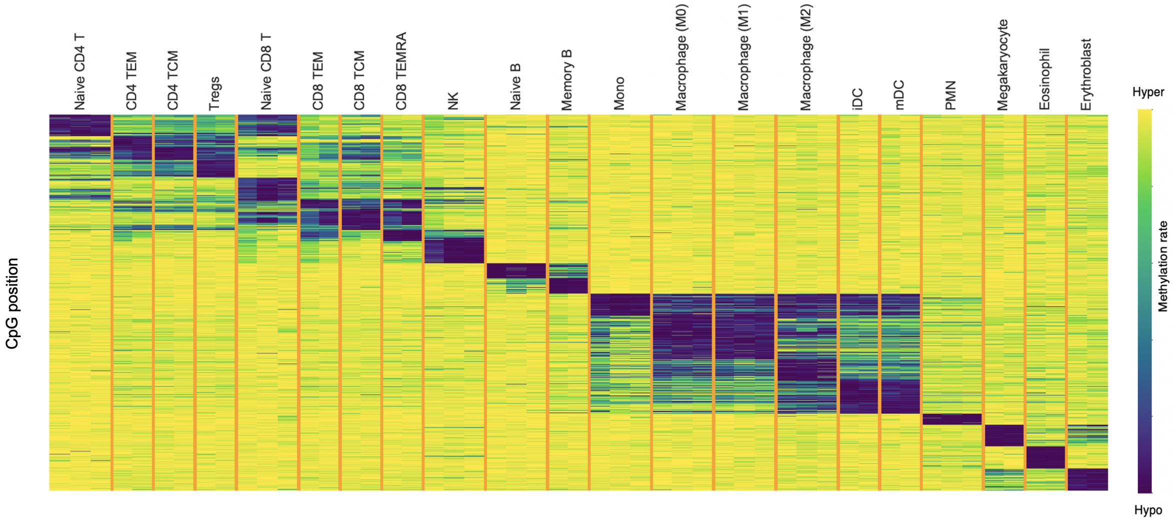


Figure 2.1: A high-resolution SM with 21 cell states using conventional one vs rest method.

Heterocellular tissue consists of different cell types and states. Deconvolution methods try to computationally estimate the cellular proportions of these different cell types from bulk sequencing data. Tissue deconvolution was developed primarily for gene expression data where gene expression of the tissue is modeled as a weighted sum of the gene expression of underlying cell types. CIBERSORT is a popular such method that first identified signatures from 22 cell types and then used support vector regression to estimate those 22 cell types from bulk expression data [40]. CIBERSORTx is a recent extension of CIBERSORT which enables the ability to build signature matrices from single-cell RNA-sequencing data and to profile distinct cellular states (e.g., exhausted vs. non-exhausted CD8 T cells) within each deconvolved cell type [42].

This idea of deconvolution can be extended from gene expression data to methylation data considering methylation status of CpG sites as a weighted sum of the methylation status of the underlying cell types. Based on this observation, MethylCIBERSORT [6] uses CIBERSORT applied to methylation sequencing data whereas MethylResolver [4]uses Least Trimmed Squares regression for methylation deconvolution. In addition to modelling the deconvolution problem as a system of linear equations like the discussed methods, some groups have built machine learning classifiers which are trained on some bulk samples and then applied to held-out data [19, 27].

Deconvolution approaches typically use a pre-defined feature matrix, known as Signature Matrix as the reference pattern of the cell types. The resolution of the deconvolution depends on the resolution of the signature matrix. Tumor-infiltrating leukocytes (TILs) are leukocytes (white blood cells) that infiltrate the tumor and make up the tumor microenvironment. If we want to detect TIL content from cfDNA, the molecular profile of TILs must be different from PBLs. Using ATAC-seq, Philip et al. demonstrate distinct epigenetic programs in tumor-specific CD8 T cells [47]. In 2020, Yang et al. [64] showed that the methylation profile of CD8 T cells coming from tumor tissue is different from normal CD8 T cells. They demonstrated that the gene promoter from CD8 T cells isolated from the TME of are hypomethylated for the tumor-reactive marker genes CD39 and CD103. In cfDNA, there are many things from all over the body, so we need a signature matrix that has high resolution consisting of different cell states along with cell types. The conventional signature matrix generation method uses a one vs rest method which may not be able to provide a specific high-resolution signature matrix (Figure 2.1). This is because this one vs rest method does not consider the biological information. For example, we know that CD4 TEM and CD4 TCM both originate from CD4 T cell and thus they will be hard to distinguish. We developed a tiered method to address this issue and generate a better high-resolution signature matrix.

2.2 Methods

In the deconvolution framework, differentially methylated regions or the signature matrix is a very, if not the most, important component as the deconvolution algorithm uses it as a reference to match. Traditionally, the signature for a cell type is generated by a one vs. rest fashion where while preparing a cell type's reference all other cell types are considered together. More specifically, the cell types are split into two groups. One group has the cell type of interest and the other has the rest. Though this technique is efficient and reasonable for smaller numbers of cell types that are clearly distinct from one another, if the signature matrix has lots of closely related cell types and states comprising of public and in-house data, the one vs. rest fashion can be problematic.

We developed a methylation signature matrix where we have several different cell types and states, many of which are closely related. In this scenario, instead of using the one vs.

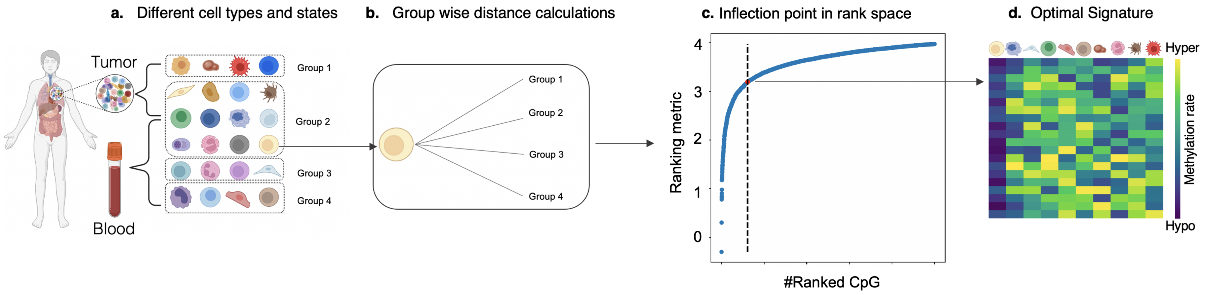


Figure 2.2: **Workflow of signature matrix generation. a,** We start by considering the methylation profile of different cell types and states in the human body. Based on known biology we will group them first. **b,** Then for any cell type/state like the yellow one we will measure the distance from each group. For Group 2, we will measure distance for all cell states separately as the yellow cell is biologically similar to rest of the cells of Group 2. **c,** With different distances and CpG number thresholds we will get different candidate signatures for the yellow cell. **d,** The optimal signature for the yellow cell state from the previous step where columns are cell types/states and rows are CpG positions. The first column (yellow) cell is mostly blue as we are using hypomethylation as the cell state-specific signature.

rest approach, we utilized a tiered approach where the existing biological information will be used. Specifically, we designed the following algorithm to generate the methylation signature matrix:

1. All cell types and states will be grouped into smaller groups based on biological similarity so that closely related cell types/states will be grouped together. For example, we know that CD4 T, CD8 T and Treg cell states are all T cells, thus we will put them together in a single group. This group information will be user-defined so that based on the context we can adjust the granularity (Figure 2.2a).

2. In this groupwise framework, all cell states belong to a unique group. When we generate the profile of a cell state the groups are of two kinds. The group that includes that cell state, we term it as Own Group and other groups as Rest Groups. To generate profile of a cell state, the methylation distance between that cell state and all Rest Groups will be calculated separately. For the Own group, we are planning to compare with all cell states of that group one by one as they are closely related (Figure 2.2b).

3. After all the groupwise distance comparisons, we will rank the CpGs based on some predefined criteria. First, we define some metrics: minimum difference, average difference, own group difference, and other group differences from the previous step. We combine all these metrics using a ranking scheme. There can be a few ranking schemes used. In this work, we used (minimum delta rank + average delta rank)/2. Other scheme can be : (own group delta rank + other group delta rank)/2. For the ranked CpG, we prepare a ROC-like plot for each cell state (Figure 2.2c). Finally, we will consider the signature which is at the inflection point (Figure 2.2d).

2.3 Experiments

We have collected methylation data for 21 purified cell states from the BLUEPRINT public database [5] and generated a signature matrix of these 21 cell states 2.3. This signature matrix shows a more specific compared to the corresponding traditional one vs rest method (Figure 2.1).

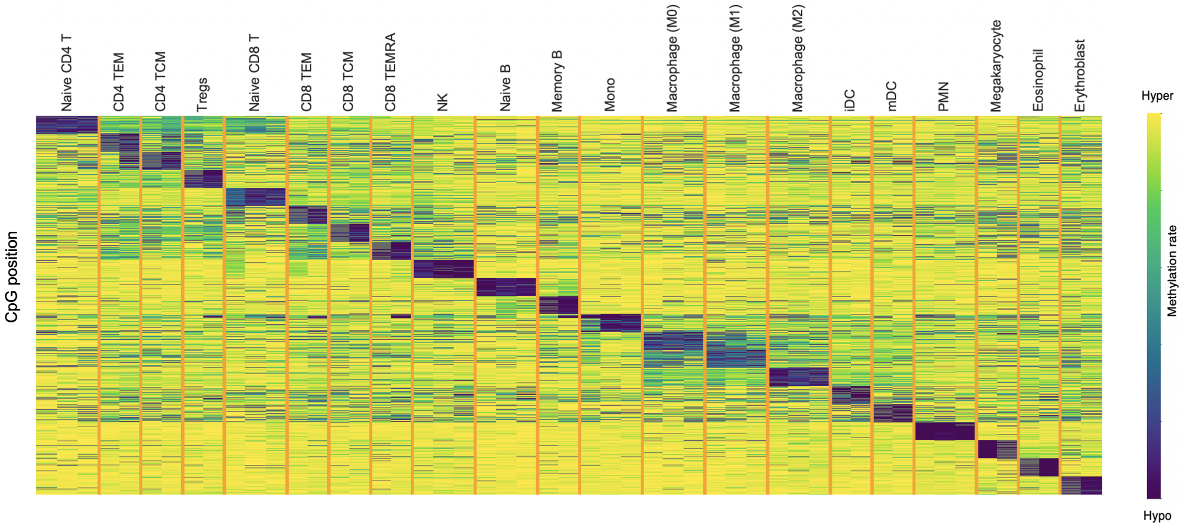


Figure 2.3: Tiered Signature Matrix for 21 cell states

Next, we evaluated the performance of deconvolution using the new signature matrix approach. To conduct this we prepare a simple deconvolution method, particularly targeting methylation data based on aligned bam files. Given a signature matrix of cell states of interest, we will test every read pair or fragment of a mixture one by one. While testing a fragment we will try to match the fragment methylation pattern with the predefined signature matrix. If it matches with a specific cell state's signature, we will classify that fragment to that cell state. Finally, to get the cellular fraction we can take all the fragments classified to that cell state and divide it by the number of available fragments tested for that cell state. We are calling this method Read Counting. Using the generated Signature Matrix, we run our Read Counting approach on seven real BULK PBMC samples where we have wet lab ground truth for some cell states using Flow cytometry or CyTOF (Figure 2.4).

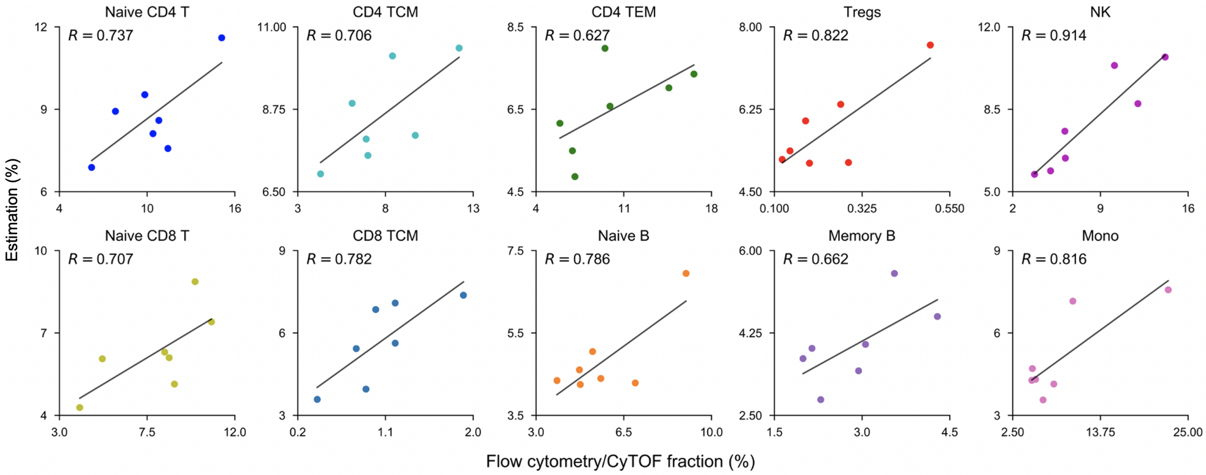


Figure 2.4: Deconvolution performance on healthy PBMC

After testing the signature matrix in healthy donors, we asked if it can be helpful for Liquid Biopsy. To assess whether noninvasive TIL profiling will have utility in vivo, it is important to compare estimated TIL composition in the plasma of colorectal (CRC) and melanoma patients against orthogonal measures of TIL content in paired tumors (e.g., by flow cytometry). We analyzed banked viably preserved tumor, plasma, and PBL samples patients with advanced melanoma. Patients have undergone tumor biopsy and blood draw pre-treatment. To estimate the TIL from patients cfDNA, we prepared signature matrix with TIL 2.5. We have taken the TIL signature and estimated it on 6 CRC patients' cfDNA where we have matched bulk tumor. We hypothesized that the TIL content in bulk tumor tissue should correlate with ctilDNA levels quantified by our LiquidTME approach. Indeed, our method showed a significant positive correlation with wet lab ground truth (Figure 2.6b). Next, we applied our method to 23 melanoma patients' plasma cfDNA to detect TIL content non-invasively and predict response to immunotherapy. All these patients had a diagnosis of metastatic melanoma and received immunotherapy. Our method shows that ctilDNA is higher in patients who responded to the ICI treatment (Figure 2.6c, d).

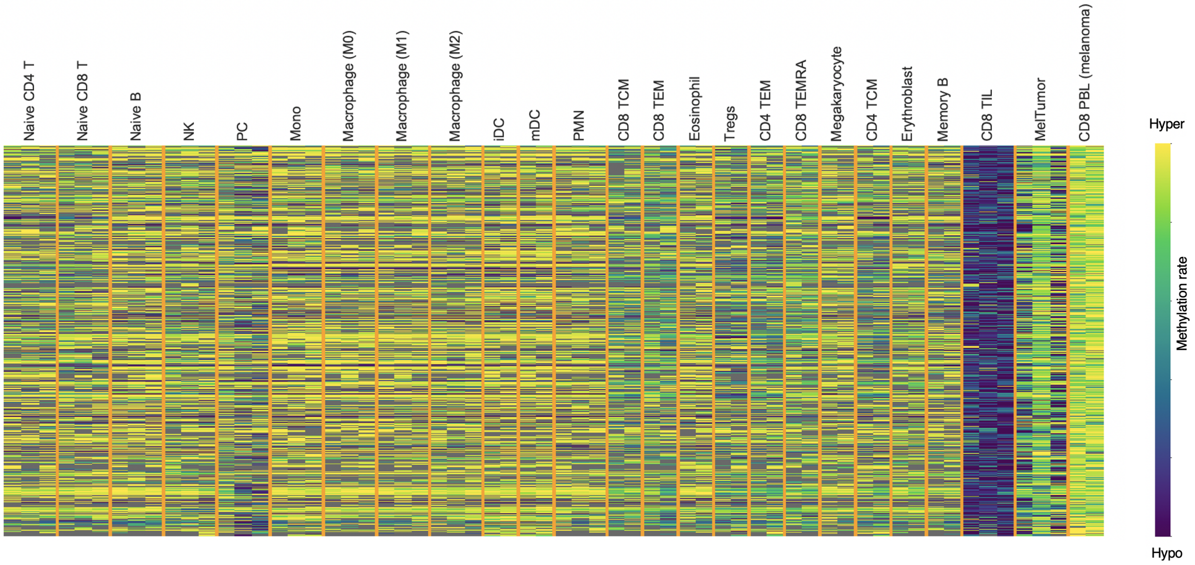


Figure 2.5: **Signature of CD8 TIL**. Columns are different cell types/states and rows are CpG positions. CD8 TILs display a distinct pattern compared to other cell types and states.

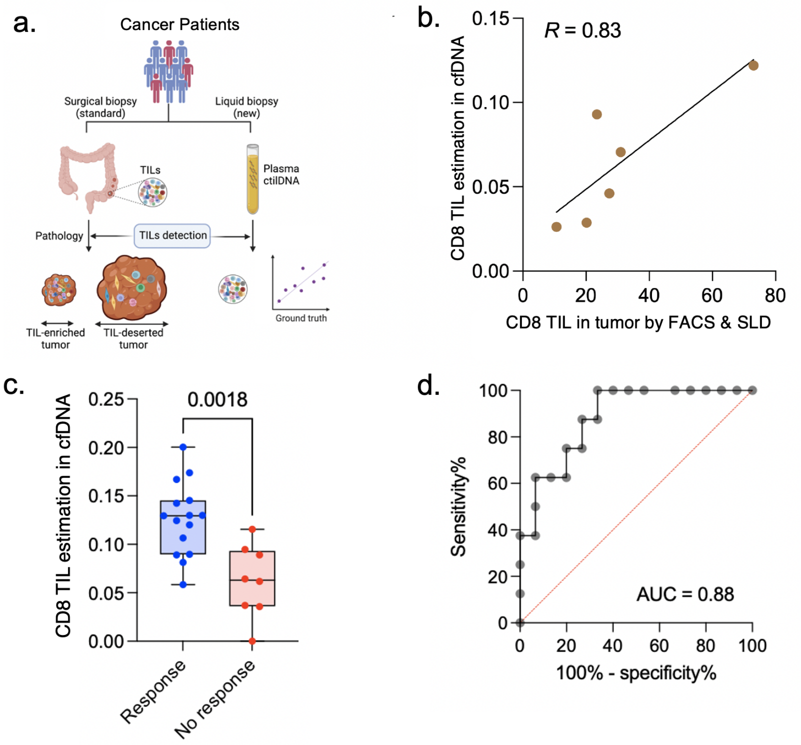


Figure 2.6: **Application in Liquid Biopsy**. **a,** Liquid Biopsy framework. **b,** Correlation of estimated TIL content from cfDNA with the paired tumor biopsy result. The TIL level in tumor is obtained by standard FACS and SLD imaging techniques. **c,** Box plot and **d,** ROC plot showing prediction of melanoma patients' response for ICI.

2.4 Discussion

Here, we developed a new framework for a high-resolution signature matrix generation and demonstrated the performance in classification and deconvolution settings. This new framework provides a way to differentiate between closely related cell states which can be helpful in various clinical applications as we demonstrated using CRC and melanoma patients. This high-resolution signature matrix can be considered a methylation cell atlas consisting of almost all important leukocytes which can be used as a common reference for methylation data. Limitations of the study include a small cohort of clinical samples as collecting this data is challenging though we are planning to extend the cohort in future work.

Chapter 3

Urine cell-free DNA multi-omics to detect MRD and predict survival in bladder cancer patients

3.1 Introduction

Bladder cancer is the 4$^{\text{th}}$ most common malignancy in men, and in 2021 alone, there were an estimated 83,000 new cases of bladder cancer in the United States [54]. Over the course of the diagnostic workflow, pathologic assessment is utilized to stratify patients into categories of muscle-invasive bladder cancer (MIBC) and non-muscle invasive bladder cancer (NMIBC). Approximately 25% of cases are found to be MIBC at the time of initial presentation, while the remaining 75% are NMIBC, with tumors retained within the mucosa and submucosa of the bladder [11].

NMIBC patients are typically managed with TURBT followed by intravesical therapies such as Bacillus Calmette-Guerin (BCG). Standard of care treatment for MIBC and high-risk NMIBC, on the other hand, often involves neoadjuvant chemotherapy (NAC) followed by radical cystectomy (RC). Even today, there remains a great deal of morbidity associated with RC and accompanying urinary diversions. Approximately 50-60% of patients will face at least one perioperative complication, and upwards of 40% will require readmission to the hospital to manage these complications [36]. The 5-year survival for patients with MIBC treated with RC remains only 50-70% [43].

Of all patients that receive recommended NAC, roughly 40% will be found to have no residual cancer, or a pathological complete response (pCR), on their final surgical specimen [18]. In these cases, the cancer was completely ablated with the original TURBT followed by NAC, and the RC could have been foregone. Unfortunately, there is currently a lack of data to help clinicians accurately identify these patients and predict which would be likely to achieve pCR after TURBT/NAC alone. There is also a lack of targetable biomarkers that can be used for empirically grounded disease prognostication.

Previously, our group utilized a single nucleotide variant (SNV) based cell-free DNA liquid biopsy assay (uCAPP-Seq) for the analysis of urine tumor DNA (utDNA) from MIBC patients [8]. Positivity on this assay was highly correlated with both residual disease as well as poorer progression-free survival (PFS) outcomes. Here, we were interested in building upon our previous uCAPP-Seq assay by integrating ultra-low-pass whole genome sequencing (ULP-WGS) data within the context of a machine learning model. Improved residual disease detection and prognostic power allows for better risk stratification. In turn, those deemed low risk may be managed with bladder-sparing approaches, avoiding the significant morbidity associated with cystectomy.

3.2 Results

Cohort characteristics and biofluid samples

Seventy-four localized bladder cancer patients underwent a physician's-choice of neoadjuvant treatment and curative-intent radical cystectomy. Seventy-eight percent (58/74) harbored muscle-invasive bladder cancer, while the rest had treatment-refractory non-muscle-invasive bladder cancer (Supplementary Data 1 [1]). Ninety-two percent (68/74) had urothelial carcinoma, while the remainder had variant histologies. A full description of the cohort is displayed in Supplementary Data 2. Urine cancer personalized profiling by deep sequencing (uCAPP-Seq) libraries prepared from urine cfDNA samples were sequenced to ¿900x median unique depth (Supplementary Data 3) along with comparably sequenced plasma (Supplementary Data 4) and germline DNA (Supplementary Data 5). ULP-WGS libraries prepared from urine cfDNA were sequenced to a median unique coverage of 2x (Supplementary Data 6).

[1] Supplementary Data is provided in excel in the published version of this work. Link: https://www.nature.com/articles/s41698-022-00345-w

Cell-free DNA biomarker differences in relation to pCR status

Copy number-derived tumor fraction (TFx) levels, estimated from ULP-WGS of urine cfDNA, ranged from 0 to 62% with a median value of 4.3% in this cohort (Supplementary Data 2). Genome-wide analysis of urine cfDNA revealed focal copy number alteration of genes previously reported by The Cancer Genome Atlas (TCGA) to be recurrently altered in MIBC (Supplementary Fig. 3.2) [38, 52], with PPARG, ZNF703, and E2F3 being the most frequently amplified. Further, uCAPP-Seq analysis of single nucleotide variant (SNV) data from our full 74 patient cohort revealed that the TERT promotor and TP53 were the most commonly mutated genes (Supplementary Fig. 3.3), again consistent with prior tissue sequencing data [38, 52, 48]. Indicative of specificity, neither copy number alterations nor SNVs were detected with significance in healthy adult urine cfDNA (Supplementary Figs. 3.2, 3.3). Additionally, results of our copy number (Supplementary Fig. 3.2) and uCAPP-Seq (Supplementary Fig. 3.3) analyses demonstrated clear differences in urine cfDNA based on pathologic complete response (pCR) status, which was determined by examination of surgical specimens by board certified genitourinary pathologists.

Bladder cancer patients who achieved pCR had significantly lower variant allele frequency (VAF) levels measured by uCAPPSeq compared to those who did not (Fig. 3.1b) despite having similar baseline characteristics (Supplementary Data 7). Strikingly, urine cfDNA significantly outperformed plasma circulating tumor DNA (Supplementary Fig. 3.4). We also measured the tumor mutational burden inferred from the number of non-silent mutations detected in urine cfDNA (iTMB). The median iTMB was 170 (range 0–476) across the cohort, consistent with previous reports in bladder cancer [17]. Comparing between subgroups, patients with no pCR had significantly higher iTMB levels than patients with pCR (median 204 vs. 117, $p = 0.001$) (Fig. 3.1c). This result is consistent with findings in breast cancer, suggesting that increased TMB is a negative predictor of pCR to neoadjuvant chemotherapy [28]. TFx, which was inferred from genome-wide copy number alterations in urine cfDNA, also differed significantly based on pCR status (median 2.4% for pCR vs. 9.9% for no pCR, $p < 0.0001$) (Fig. 3.1d), suggesting that genome-wide copy number alterations, like SNVs, could be utilized for urine-based MRD detection in bladder cancer.

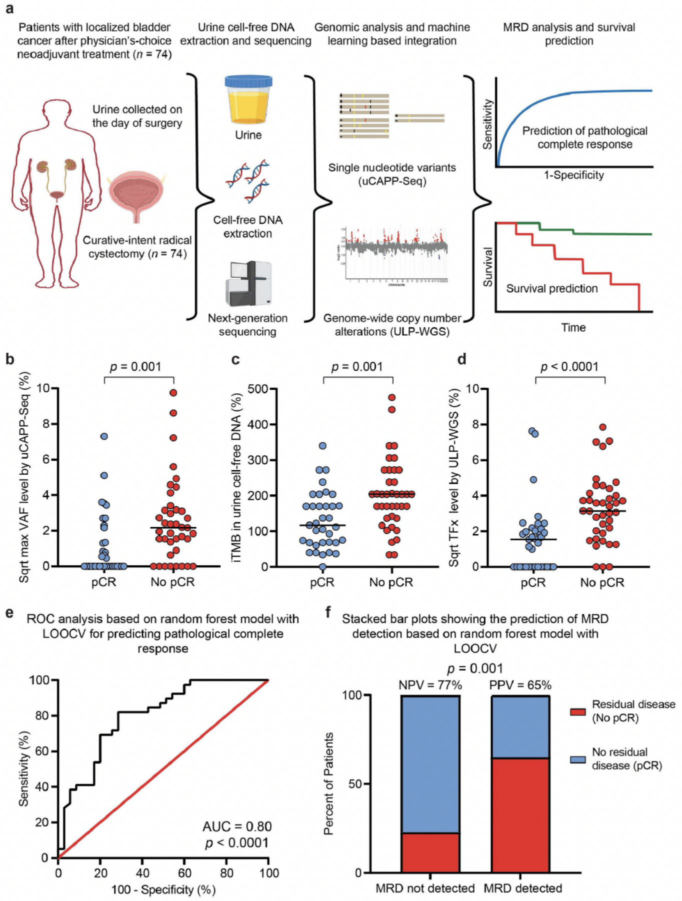


Figure 3.1: **Pathologic complete response prediction using a random forest model based on urine tumor DNA. a** Urine was collected prospectively from 74 localized bladder cancer patients pre-operatively on the day of curative-intent radical cystectomy after physician'schoice neoadjuvant treatment. Urine cell-free DNA was sequenced by uCAPP-Seq (for single nucleotide variants) and ULP-WGS (for genomewide copy number alterations) and then correlated with residual tumor in the surgical resection specimen and with patient survival. This figure panel was created with BioRender.com. **b** SNV-derived maximum VAFs, **c** inferred tumor mutational burden, and **d** CNA-derived tumor fraction levels in urine cell-free DNA from patients with localized bladder cancer. Scatter plots display these three different urine cell-free DNA metrics, stratified by pathologic complete response status, with significance determined by the Mann–Whitney U-test. VAF and CNA-derived tumor fraction data are shown after square root transformation. **e** ROC analysis of random forest model integrating urine tumor DNA metrics and other pretreatment clinical variables (Supplementary Fig. 3.5). ROC curve demonstrating the model's performance for predicting pCR after LOOCV (AUC = 0.80, p ¡ 0.0001). **f** Stacked bar plot depicting NPV and PPV of the random forest model with LOOCV, with significance determined by the Fisher's exact test. AUC area under the curve, cfDNA cell-free DNA, CNA copy number alteration, iTMB inferred tumor mutational burden, LOOCV leave-one-out cross-validation, max maximum, MRD molecular residual disease, NPV negative predictive value, pCR pathologic complete response, PPV positive predictive value, ROC receiver operating characteristic, SNV single nucleotide variant, Sqrt square root, TFx tumor fraction, uCAPP-Seq urine cancer personalized profiling by deep sequencing, ULP-WGS ultra-low-pass whole genome sequencing, VAF variant allele frequency.

Random forest model for pCR and survival prediction

We next integrated the three urine cfDNA-derived metrics— maximum VAF, iTMB, and TFx—with pretreatment clinical variables using a machine learning random forest model that we validated by leave-one-out cross-validation (LOOCV) (Supplementary Fig. 3.5a). Area under the receiver operating characteristic curve (AUROC) for the random forest model was 0.80 ($p < 0.0001$) (Fig. 3.1e), with a sensitivity of 87%, a negative predictive value (NPV) of 77%, and a positive predictive value (PPV) of 65% for determining pCR (Fig. 3.1f). The combinatorial urine cfDNA metric was by far the most important predictive feature in the model (Supplementary Fig. 3.5b). Indeed, when we developed a LOOCV model including only urine cfDNA features (maximum VAF, iTMB, and TFx), its performance remained high with AUROC of 0.76 for determining pCR (Supplementary Fig. 3.6).

Using our LOOCV model, we also aimed to predict survival outcomes within our 74-patient localized bladder cancer cohort. Therefore, we performed Kaplan–Meier and Cox regression landmark analyses starting from the time of surgery (Fig. 3.2 and Supplementary Data 8, 9). Strikingly, patients predicted by our model to harbor MRD also had significantly worse progression free survival (PFS) (HR = 3.00, $p = 0.01$; Fig. 3.2a) and overall survival (OS) (HR = 4.81, $p = 0.009$; Fig. 3.2b), comparable to the presence of residual disease in the radical cystectomy specimen itself (PFS HR = 3.13, $p = 0.005$; OS HR = 3.57, $p = 0.03$; Fig. 3.2c, d). Univariate and multivariate Cox proportional hazards models confirmed the significance of our MRD predictions (Supplementary Data 8, 9). The model remained predictive for both PFS and OS when restricted to only MIBC patients (Supplementary Fig. 3.7) and patients treated with NAC (Supplementary Fig. 3.8). Furthermore, the model remained significant for predicting PFS when applied to an independent held-out validation cohort (Supplementary Fig. 3.9a) with a trend toward predicting OS significantly as well (Supplementary Fig. 3.9b).

3.3 Discussion

Here, we developed a multi-modal urine cfDNA method to sensitively detect MRD and predict pCR in bladder cancer patients. Our technology also predicted survival significantly and comparably to gold-standard surgical pathologic analysis of resected tumor tissue [46]. Limitations of our study include patients having only a single timepoint assessment of urine cfDNA. Other investigations utilizing plasma have shown that multiple samples obtained in surveillance settings can achieve greater sensitivity for detecting circulating tumor DNA MRD [33, 45]. We nevertheless achieved high MRD sensitivity by multimodally analyzing urine, the biofluid most proximal to localized bladder cancer. While our study was prospective, all samples were obtained from a single medical center. It will be important to corroborate our findings in a multiinstitutional setting. Finally, given the prospective nature of our study with all patients enrolled between 2019 and 2021, the median follow-up time was modest at 23 months. It will be important to perform a study with a longer follow-up to confirm the dramatic survival differences we observed.

In conclusion, our multi-omic urine-based cell-free DNA analysis allowed for the detection of MRD with high sensitivity and risk stratified patients by survival. In the future, this type of integrative analysis could potentially be used to facilitate more personalized clinical decision-making for bladder cancer.

3.4 Methods

Patient recruitment and sample collection

We enrolled 74 patients with localized bladder cancer who proceeded with curative-intent radical cystectomy at the Washington University Siteman Cancer Center. Eligible patients were required to be at least 18 years old and to have a diagnosis of bladder cancer confirmed by histologic or cytologic assessment. Urine and blood collection was performed at the time of enrollment. We also utilized urine and blood samples from 15 healthy adult volunteers for comparison. The methods were performed in accordance with relevant guidelines and regulations and approved by the institutional review board at the Washington University in St. Louis School of Medicine. Patients and healthy donors were enrolled in NCT04354064 (ClinicalTrials.gov). Written informed consent was obtained from all trial participants in accordance with the Declaration of Helsinki. This study followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines for observational studies.

Pathologic response assessment

Surgical resection specimens from radical cystectomy procedures were processed consistently using a standardized institutional approach, including specimen collection, handling, and submission to the Pathology Department at the Washington University School of Medicine. Resected surgical specimens were microscopically reviewed by blinded board-certified genitourinary surgical pathologists. AJCC 8th edition pathologic stage T0, Tis, and Ta were defined as pathologic complete response (pCR) in our study. Non-pathologic complete response (no pCR) was defined as stages T1, T2, T3, or T4, with or without evidence of nodal disease (N1–N2) and/or evidence of metastatic disease.

Urine cell-free DNA extraction

Urine samples were collected in cups pre-filled with 1–2mL of 0.5M EDTA. Shortly following collection, cfDNA was extracted from 22 to 90 ml of urine with Q-sepharose resin slurry (GE Healthcare, Chicago, Illinois)3. Briefly, Q-sepharose resin was added to urine at a ratio of 10 ul slurry per ml of urine and mixed for 30 min. After centrifuging the mixture at 1800 × g for 10 min, the supernatant was discarded. The resin was washed twice with 0.3M LiCl/10mM sodium acetate (pH 5.5), transferred to a Micro Bio-Spin column (Bio-Rad, Hercules, California, USA), and the bound DNA was eluted with 70% ethanol and passed over a QIAquick column (Qiagen, Hilden, Germany). Columns were then washed with 2M LiCl in 70% ethanol, followed by 75mM potassium acetate (pH 5.5) in 80% ethanol. Finally, DNA was eluted in nuclease-free water or 10mM Tris-Cl (pH 8.5). Urine cfDNA was quantified using the Qubit dsDNA High Sensitivity Assay kit (Thermo Fisher Scientific, Waltham, Massachusetts). cfDNA quality was assessed on an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, California).

Germline DNA extraction

A peripheral blood sample was collected from each subject using EDTA tubes (Becton Dickinson, Franklin Lakes, New Jersey). Plasma-depleted whole blood (PDWB) was collected by centrifugation and then frozen at −80°C prior to the isolation of germline DNA. Germline

DNA was extracted from 50 to 100 ul of PDWB using the QIAmp DNA Micro Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. DNA was then quantified by the Qubit dsDNA High Sensitivity Assay to determine yield (Thermo Fischer, Waltham, Massachusetts).

Cancer personalized profiling by deep sequencing (CAPP-Seq)

Urine CAPP-Seq was performed on urine cfDNA along with matched germline DNA [13, 8]. Briefly, urine cfDNA and germline DNA were fragmented to 180 bp size fragments prior to library preparation using a LE220-focused ultrasonicator (Covaris, Woburn, Massachusetts). Approximately 32 ng of sheared urine cfDNA or germline DNA was used for library preparation using the KAPA HyperPrep kit with barcoded adapters containing demultiplexing, deduplicating, and duplexed unique molecular identifiers. Targeted hybrid capture was performed per the standard uCAPPSeq method [13, 8]. We used a focused MRD gene panel spanning 145 kb in size and consisting of 49 consensus driver genes frequently mutated in bladder cancer for the VAF estimation in each sample [8]. For TMB estimation, we utilized an expanded panel of 387 kb in size which covers 536 genes [8]. Following hybridization capture, libraries were sequenced deeply on a HiSeq 4000 (Illumina, San Diego, California) with 2 × 150 bp paired-end reads. Sequencing results were analyzed for single nucleotide variants using the CAPP-Seq bioinformatic pipeline [39, 41]. CAPP-Seq was similarly performed on plasma with matched germline DNA12–14 [39, 41, 7].

Single nucleotide variant analysis from cfDNA

Only non-silent mutations with duplex support and with no germline support were considered when querying MRD from cfDNA [8]. Specifically, we defined maximum VAF as the maximum variant allele fraction among all non-silent mutations with duplex support detected by CAPP-Seq using our 145 kb driver genefocused MRD gene panel [8], regardless of the number of other mutations detected and their frequencies. Maximum VAF was selected as the metric representing tumor DNA by CAPP-Seq, and was correlated with MRD status in the surgical specimen. Nonsilent SNVs in urine cfDNA with ¿2.3% VAF [8] are represented in the Supplementary Fig. 3.3 heatmap. We additionally inferred tumor mutational burden using our urine CAPP-Seq results. Briefly, we utilized our TMB gene panel, which is 387 kb in size and covers 536 genes, and applied the equation determined previously by linear regression while accounting for potential dropout in order to infer exome-wide TMB [8].

Ultra-low-pass whole genome sequencing (ULP-WGS)

ULP-WGS libraries were prepared from 32 to 50 ng of sheared urine cfDNA using the Kapa HyperPrep kit (Roche, Basel, Switzerland). Libraries were balanced, pooled, and sequenced on a HiSeq 4000 (Illumina, San Diego, California) to a median deduplicated depth of 2x (Supplementary Data 6). FASTQ files were demultiplexed and raw reads were quality-filtered using fastp v.0.20.0. Quality-filtered reads were then aligned to the hg19 human genome assembly using BWA v.0.7.17. Aligned reads were deduplicated with Samtools v.1.13. ichorCNA v0.2.015 [1] was then used to infer tumor fractions in each urine cfDNA sample. Briefly, reads were summed in nonoverlapping bins of 106 bases; local read depth was corrected for GC bias and known regions of low mappability, and artifacts were removed by comparison to ichorCNA's built-in healthy control reference. Copy number alterations (CNAs) were then predicted across the whole genome using low tumor fraction parameters for cfDNA samples; X and Y chromosomes were excluded from copy number calculations. ichorCNA then used these binned, bias-corrected copy number values to model a two component mixture of tumor-derived and non-tumor-derived fragments, from which it inferred the fraction of reads in each sample originating from the tumor (tumor fraction) [1].

The visualization of aggregate genome-wide CNAs (Supplementary Fig. 3.2) was generated from compiled log_2 ratios of copy number, broken down into three categories: No pathologic complete response (n = 39), pathologic complete response (n = 35), and healthy adults (n = 15). Following the removal of artifacts, regions were classified as exhibiting copy number gain if log2 of the copy number ratio was > 0.58 ($\log_2$ (3/2)) or loss if log_2 of the copy number ratio was < -1.0 (log2 (1/2)) [56]. Midpoints of genes previously shown to be commonly altered in whole exome sequencing data of muscle-invasive bladder cancer, based on their annotation in Fig. 1 of the respective TCGA publications [38, 52] are specifically highlighted (Supplementary Figs. 3.2, 3.3).

Machine learning model to predict pathologic complete response and survival

We implemented a random forest model for the prediction of pCR, which we validated using LOOCV. We used the maximum VAF, iTMB, and ULP-WGS-inferred tumor fraction (TFx) in urine cfDNA, which were combined together into one urine tumor DNA feature for the random forest model via multiplication followed by the square root of the product. Other features in the model included age, gender, ethnicity, smoking status, receipt of neoadjuvant chemotherapy, and tumor invasion status (Supplementary Fig. 3.5). We additionally developed another LOOCV random forest model using only urine cfDNA features (VAF, iTMB, and TFx) without the clinical variables (Supplementary Fig. 3.6). We used the Python scikit-learn package (v0.24.2) [44] to implement the random forest algorithm, with the following parameters: n_estimators = 2000; criterion = gini; bootstrap = True. The performance of the model after LOOCV for predicting pCR was assessed by receiver operating characteristic (ROC) area under the curve (AUC) analysis.

Patients predicted by the LOOCV model to not achieve pCR were defined as MRD-positive, while those predicted to have pCR were defined as MRD-negative. LOOCV model MRD predictions were compared to gold-standard surgical pathology results (Fig. 3.1f) and were also stratified by Kaplan–Meier analysis from the time of surgical resection for progression-free survival (PFS) and overall survival (OS) (Fig. 3.2). The model was additionally generated using independent training and held-out validation cohorts (Supplementary Fig. 3.9). Furthermore, we calculated feature importance levels by assessing mean decrease in impurity18, to determine how classifications of pCR (MRD-negative) versus no pCR (MRD-positive) were affected if a particular feature was left out of the random forest model (Supplementary Fig. 3.5b).

Power and statistical analyses

We powered the current study assuming a substantial difference in urine tumor DNA levels between patients who achieved pCR or healthy donors, compared to patients with no pCR. Assuming a large effect size estimated by Cohen's f=0.5, we accrued subjects to this study until there were at least 14 subjects per group (groups= healthy donors, bladder cancer with pCR, bladder cancer with no pCR) in order to detect a difference between healthy or pCR, and no pCR with an estimated power of 80% and significance level of 0.05 as determined by one-way ANOVA. Patient characteristics such as age, gender, ethnicity, smoking history, tumor stage, neoadjuvant chemotherapy, and histology were statistically compared between groups of pCR and no pCR patients using Fisher's exact test for categorical variables and Student's t-test for normally distributed continuous variables (Supplementary Data 7). SNV-derived maximum VAFs, inferred tumor mutational burden, and CNA-derived tumor fraction levels in urine cell-free DNA from patients with localized bladder cancer was statistically compared between groups of pCR and no pCR using the Mann–Whitney U-test (Fig. 3.1b–d and Supplementary Figs. 3.4a, 3.7a–c, 3.8a–c). The Python scikit learn package (v0.24.2) was used for random forest modeling with LOOCV (Supplementary Figs. 3.5, 3.6) or with separate training and validation datasets (Supplementary Fig. 3.9). ROC analysis was carried out to assess the performance of the LOOCV random forest model and the corresponding AUC was calculated for the full cohort of 74 localized bladder cancer patients with and without pretreatment clinical variables (Fig. 3.1e and Supplementary Fig. 3.6b) and for MIBC patients (Supplementary Fig. 3.7d). MRD predictions based on the LOOCV random forest model was compared to surgical ground truth by Fisher's exact test (Fig. 3.1f and Supplementary Fig. 3.7e). Survival curves for PFS and OS were analyzed by the Kaplan–Meier method and statistical significance was determined by the log-rank test (Fig. 3.2 and Supplementary Figs. 3.7f-g, 3.8d-e, 3.9). The Mantel–Haenszel method was used to estimate hazard ratios. Cox proportional hazards model (PHM) univariate and multivariate analyses were developed to assess both PFS and OS (Supplementary Data 8, 9). In addition to random forest model prediction, hematocrit, body mass index, and urine cfDNA concentration were included in the multivariate models. For OS, there were no deaths during the follow-up period among patients predicted by the random forest model to achieve pCR. Given this, the assumption of proportional hazards was not met. We performed all Kaplan–Meier and Cox regression analyses starting from the time of surgery. The reverse Kaplan–Meier method was used to calculate the median follow-up time (Supplementary Data 1). All statistical analyses were performed using Prism 9 (GraphPad Software, San Diego, California) or SAS version 9.4 (SAS, Cary, North Carolina).

3.5 Acknowledgements

We are grateful to the patients, their families, and healthy volunteers involved in this study. All wet-lab works were performed and supervised by our wet-lab colleagues. This work was supported by the National Cancer Institute (NCI) under award number K08CA238711 (A.A.C.), and the National Center for Advancing Translational Sciences (NCATS) under award number UL1TR002345 (Principal Investigator, Bradley Evanoff; A.A.C.). This work was additionally supported by the Washington University School of Medicine Dean's Medical Student Research Fellowship (A.S), the Midwest Stone Institute (Z.L.S.), the Rabushka Bladder Cancer Research Fund (Z.L.S. and A.A.C.), the Alvin J. Siteman Cancer Research Fund (A.A.C.), the Cancer Research Foundation Young Investigator Award (A.A.C.), and the V Foundation V Scholar Award (A.A.C.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

3.6 Supplementary

Data Information: All supplementary data is provided in excel format of our published version [2] of this work.
Figures:

[2] https://www.nature.com/articles/s41698-022-00345-w

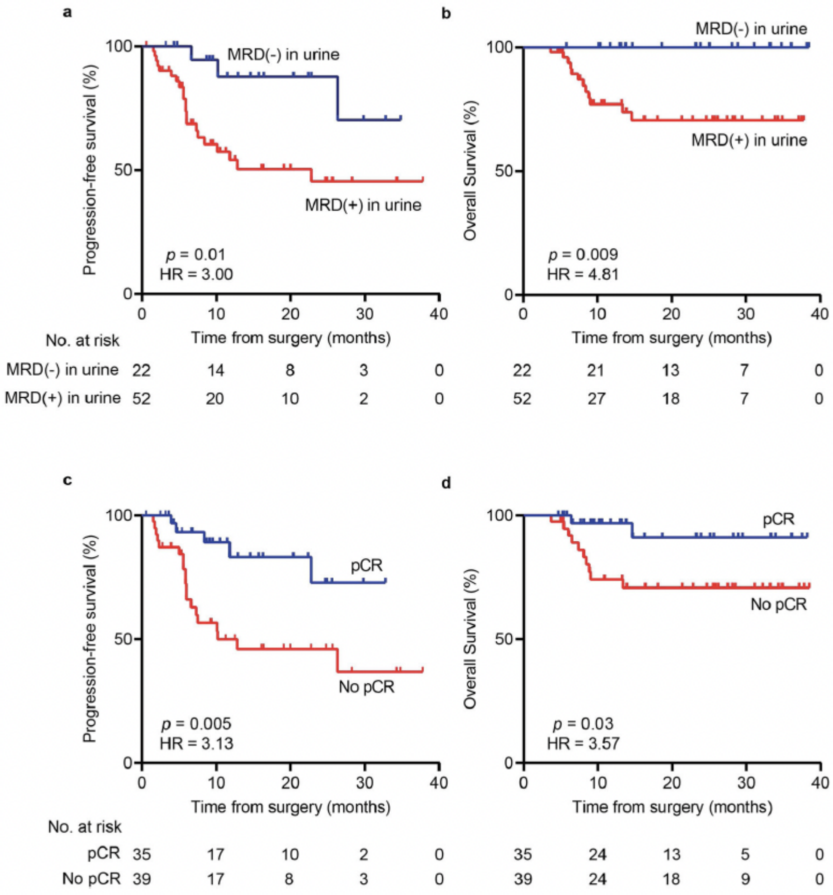
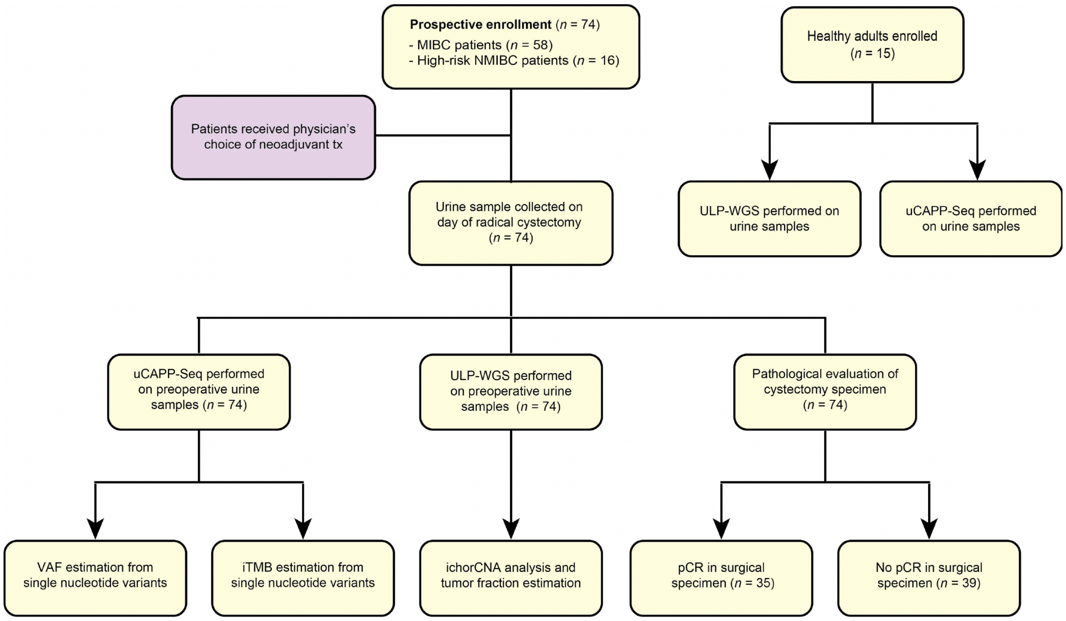
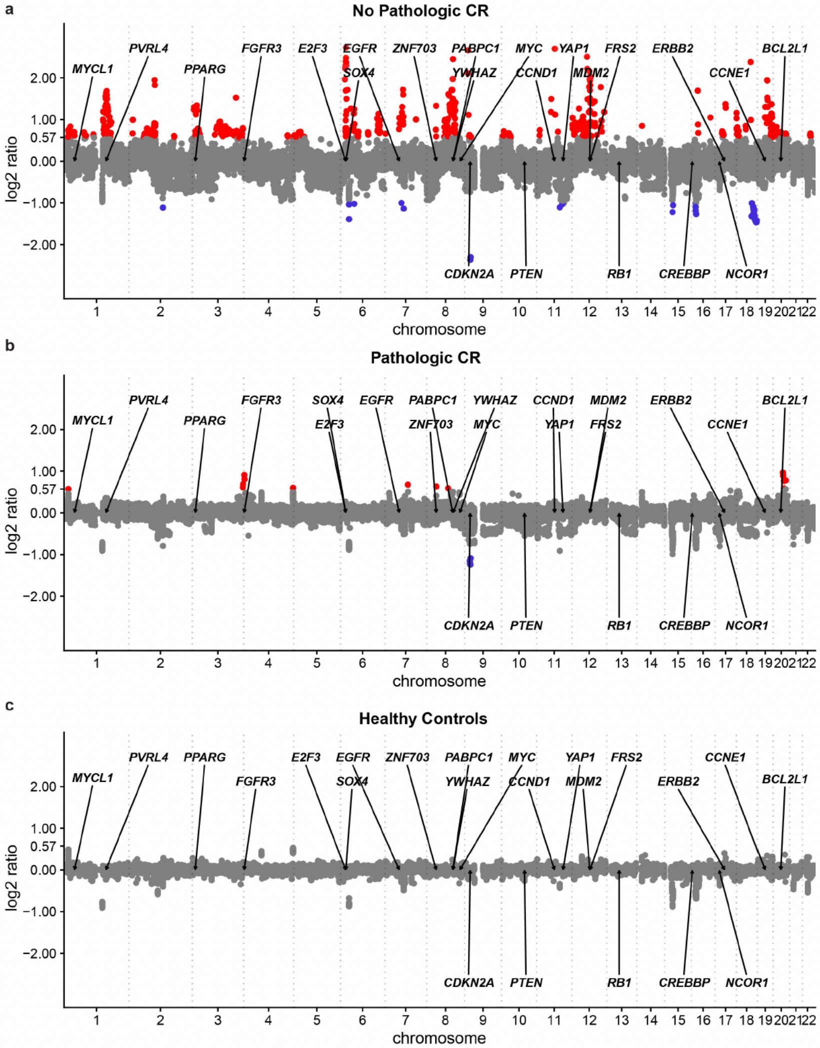


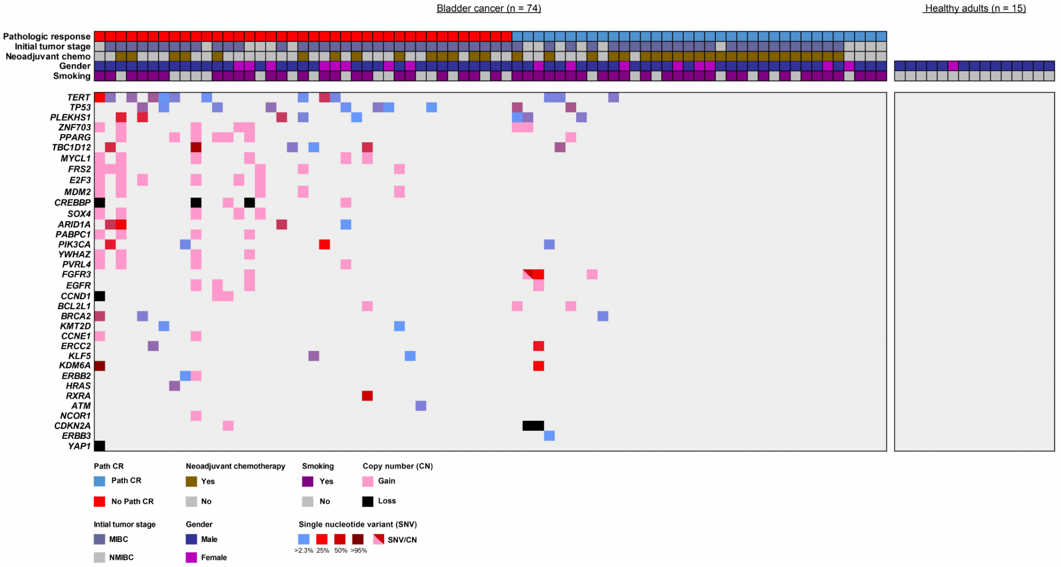
Figure 3.2: **Survival analysis comparing urine MRD detection to pathologic analysis of the resection specimen.** **a** progression-free survival and **b** overall survival stratified by MRD detection in urine, determined by the LOOCV random forest model (Supplementary Fig. 3.5). **c** Progression-free survival and **d** overall survival stratified by pCR determined by microscopic analysis of the radical cystectomy specimen. Survival times shown are relative to the time of radical cystectomy. p values were calculated by the log-rank test and HRs by the Mantel–Haenszel method. HR hazard ratio, LOOCV leave-one-out cross-validation, MRD molecular residual disease, pCR pathologic complete response.


Prospective enrollment (n = 74)
- MIBC patients (n = 58)
- High-risk NMIBC patients (n = 16)
Healthy adults enrolled
(n = 15)
Patients received physician's choice of neoadjuvant tx
Urine sample collected on day of radical cystectomy
(n = 74)
ULP-WGS performed on urine samples
uCAPP-Seq performed on urine samples
uCAPP-Seq performed on preoperative urine samples (n = 74)
ULP-WGS performed on preoperative urine samples (n = 74)
Pathological evaluation of cystectomy specimen
(n = 74)
VAF estimation from single nucleotide variants
iTMB estimation from single nucleotide variants
ichorCNA analysis and tumor fraction estimation
pCR in surgical specimen (n = 35)
No pCR in surgical specimen (n = 39)

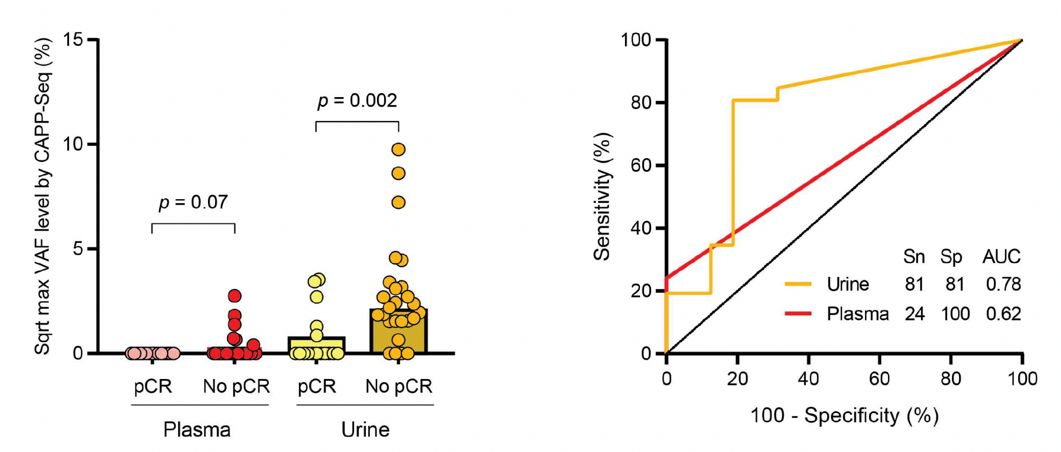
Supplementary Fig 3.1: **Study schema.** Patients with localized bladder cancer who were candidates for radical cystectomy were prospectively enrolled onto this study. Urine samples were then collected for uCAPP-Seq and ULP-WGS analysis as shown in the schema. Urine samples from 15 healthy adults were also used for ULP-WGS and uCAPP-Seq analysis. iTMB, inferred tumor mutational burden; MIBC, muscle-invasive bladder cancer; NMIBC, non-muscle-invasive bladder cancer; pCR, pathologic complete response; tx, treatment; uCAPP-Seq, urine Cancer Personalized Profiling by deep Sequencing; ULP-WGS, ultra-low-pass whole genome sequencing; VAF, variant allele frequency.



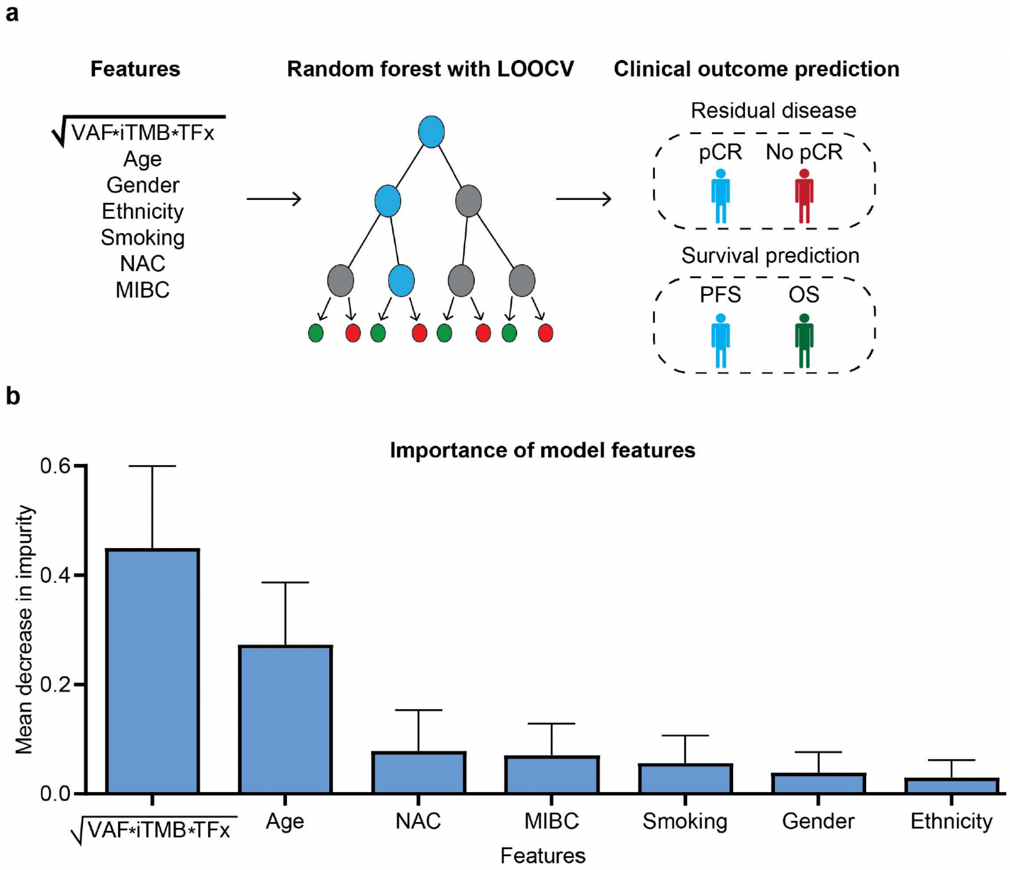
Supplementary Fig 3.2: **Genome-wide copy number plots with annotation of genes important in bladder cancer.** Plots represent the aggregate copy number alterations compiled from urine cell-free DNA data in **(a)** Patients with no pCR (n = 39), **(b)** Patients with pCR (n = 35) or **(c)** Healthy adults (n = 15). Each panel depicts log2 copy number ratios across the genome. Red represents copy number gain while blue represents copy number loss (Methods). Annotated genes are those previously reported in TCGA to be copy-number altered in bladder cancer (Methods). pCR, pathologic complete response; TCGA, the cancer genome atlas.



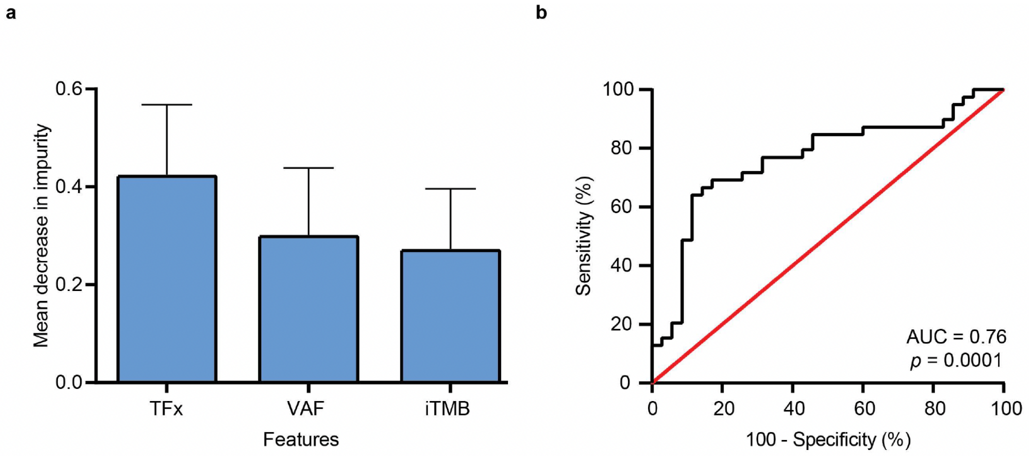
Supplementary Fig 3.3: **Subject characteristics and detected genomic alterations.** Co-mutation plot showing genomic alterations (mutations and copy number alterations) detected in pre-operative urine cell-free DNA from each patient with no pCR versus pCR and healthy adults. Mutational data represent non-silent SNVs detected within the MRD uCAPP-Seq gene panel, while copy number alterations represent ultra-low-pass whole genome sequencing data, focusing on genes reported by TCGA to be altered in muscle-invasive bladder cancer (Methods). Patient and healthy donor characteristics are represented by the upper heatmaps. NMIBC, non-muscle-invasive bladder cancer; MIBC, muscle-invasive bladder cancer; MRD, molecular residual disease; pCR, pathological complete response; SNV, single nucleotide variant; TCGA, the cancer genome atlas.



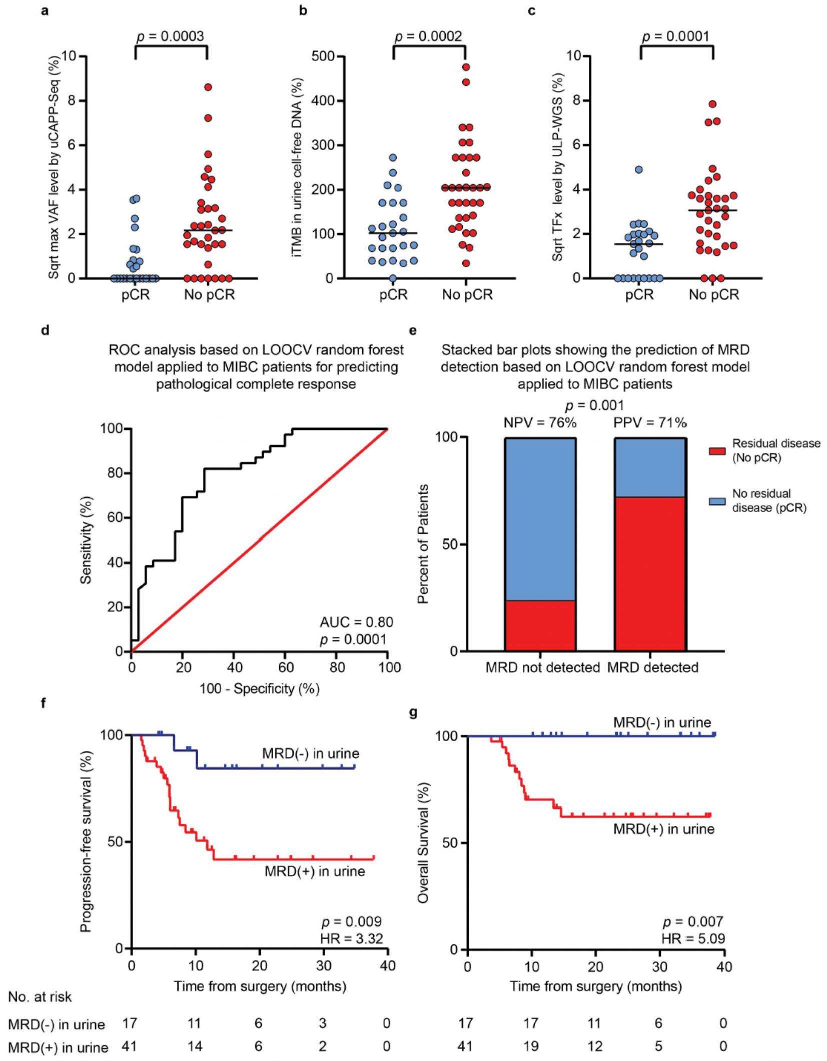
Supplementary Fig 3.4: **Performance of CAPP-Seq in matched urine and plasma samples for detecting MRD and predicting pathologic response.** **(a)** Scatter plot of maximum VAF levels after square-root transformation in urine versus plasma from 40 localized bladder cancer patients, compared to gold-standard surgical pathology. **(b)** ROC analysis for classifying pCR from no pCR patients by CAPP-Seq. CAPP-Seq in urine cell-free DNA classified pathologic response more accurately than in paired plasma (AUC 0.78 versus 0.62). AUC, area under the curve; CAPP-Seq, Cancer Personalized Profiling by deep Sequencing; MRD, molecular residual disease; pCR, pathologic complete response; ROC, receiver operating characteristic; Sqrt, square root; VAF, variant allele frequency.



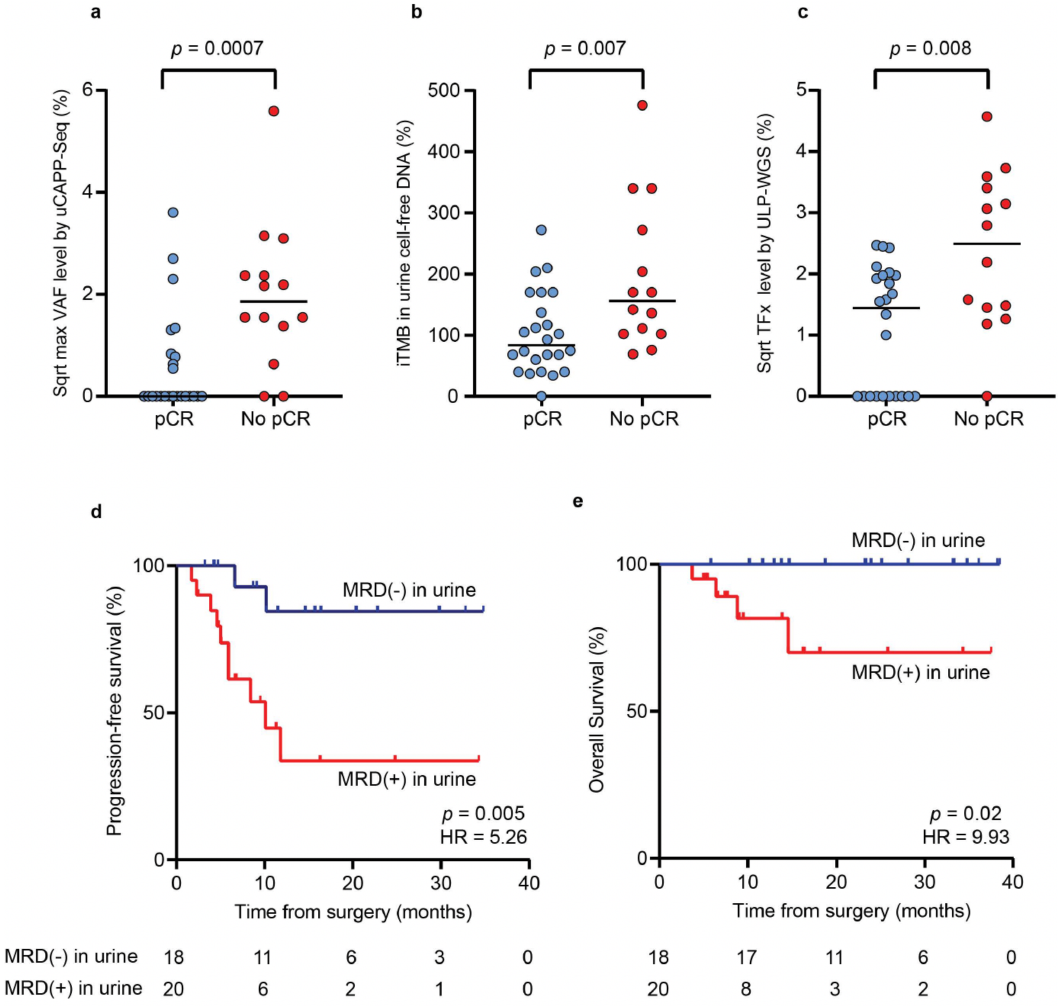
Supplementary Fig 3.5: **Random forest model with LOOCV to predict pathologic complete response status.** **(a)** Schema depicting the model's development, validation, and application. **(b)** Importance of features in the random forest model used for predicting pCR status. Error bars represent the standard deviation. iTMB, inferred tumor mutational burden; LOOCV, leave-one-out cross-validation; MIBC, muscle-invasive bladder cancer; NAC, neoadjuvant chemotherapy; OS, overall survival; pCR, pathologic complete response; PFS, progression-free survival; TFx, tumor fraction; VAF, variant allele frequency.



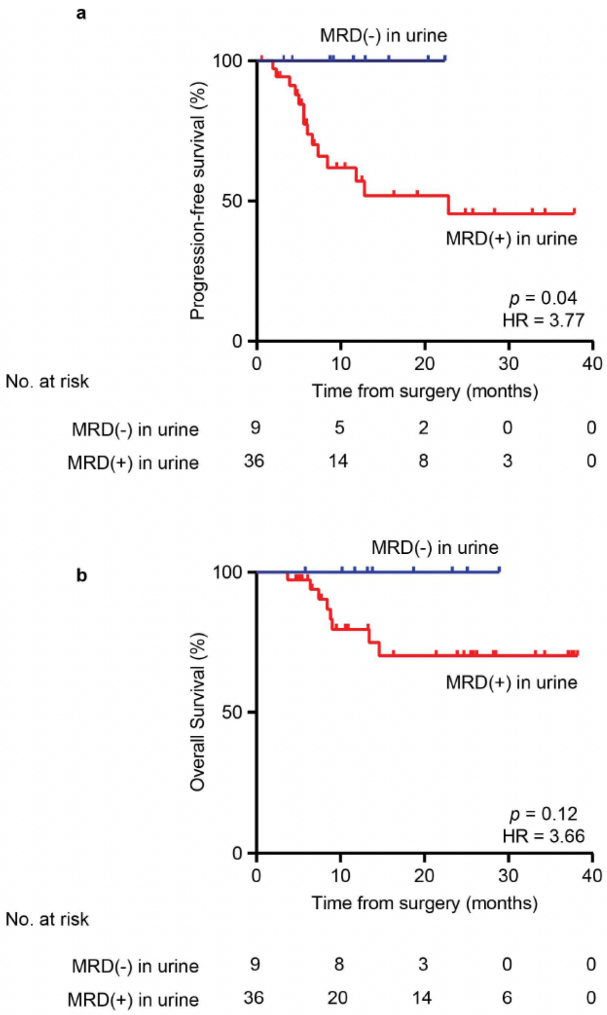
Supplementary Fig 3.6: **Random forest model based on urine cell-free DNA features with LOOCV to predict pathologic complete response status.** **(a)** Importance of features in the random forest model used for predicting pCR status based on urine cell-free DNA features only (TFx, maximum VAF and iTMB). Error bars represent the standard deviation. **(b)** ROC analysis of random forest model for predicting pCR after LOOCV (AUC = 0.76, p = 0.0001). AUC, area under the curve; iTMB, inferred tumor mutational burden; LOOCV, leave-one-out cross-validation; pCR, pathologic complete response; ROC, receiver operating characteristic; TFx, tumor fraction; VAF, variant allele frequency.



Supplementary Fig 3.7: **LOOCV random forest model applied to MIBC patients to predict pathologic response and survival outcomes.** Scatter plots displaying **(a)** maximum VAF (square-root transformed), **(b)** iTMB, and **(c)** TFx (square-root transformed), stratified by pathologic response status among MIBC patients (n = 58), with significance determined by the Mann-Whitney U test. **(d)** ROC analysis demonstrating the LOOCV random forest model's performance in classifying MIBC patients by pCR status; AUC of 0.80 (p = 0.0001). **(e)** Stacked bar plot depicting NPV and PPV of the LOOCV random forest model with significance determined by the Fischer's exact test. Kaplan-Meier analysis of (f) progression-free survival and **(g)** overall survival based on the LOOCV random forest model applied to patients with MIBC (n = 58). p values were calculated by the log-rank test and HRs by the Mantel-Haenszel method. AUC, area under the curve; cfDNA, cell-free DNA; iTMB, inferred tumor mutational burden; LOOCV, leave-one-out cross-validation; MIBC, muscle-invasive bladder cancer; MRD, molecular residual disease; NPV, negative predictive value; pCR, pathologic complete response; PPV, positive predictive value; ROC, receiver operating characteristic; Sqrt, square root; TFx, tumor fraction; VAF, variant allele frequency.



Supplementary Fig 3.8: **LOOCV random forest model applied to MIBC patients who received neoadjuvant chemotherapy to predict pathologic response and survival outcomes.** Scatter plots displaying **(a)** maximum VAF (square-root transformed), **(b)** iTMB, and **(c)** TFx (square-root transformed), stratified by pathologic response status among MIBC patients who received NAC (n = 38). Significance was determined by the Mann-Whitney U test. Kaplan-Meier analysis of **(d)** progression-free survival and **(e)** overall survival based on the LOOCV random forest model applied to MIBC patient who received NAC. p values were calculated by the log-rank test and HRs by the Mantel-Haenszel method. cfDNA, cell-free DNA; iTMB, inferred tumor mutational burden; MIBC, muscle-invasive bladder cancer; MRD, molecular residual disease; NAC, neoadjuvant chemotherapy; pCR, pathologic complete response; Sqrt, square root; TFx, tumor fraction; VAF, variant allele frequency.



Supplementary Fig 3.9: **Random forest model evaluated for survival outcomes in a held-out validation cohort.** Kaplan-Meier analysis of **(a)** progression-free survival and **(b)** overall survival in a held-out validation cohort of 45 localized bladder cancer patients, after random forest model training using data from 29 localized bladder cancer patients (Methods). p values were calculated by the log-rank test and HRs by the Mantel-Haenszel method. HR, hazard ratio; MRD, molecular residual disease.

Chapter 4

Conclusion

In this study, we presented two works on Liquid Biopsy. Based on methylation data, the first work focused on how to detect tissue of origin from cfDNA samples. We modeled this as a deconvolution problem and developed a new signature matrix generation framework enabling high-resolution deconvolution for closely related cell types. In addition, we demonstrated potential Liquid Biopsy applications in CRC and melanoma patients where CRC patients predicted TIL correlates positively with ground truth and in the melanoma patients, more TIL was detected for the response case.

In the second work, we developed a framework to detect molecular residual disease (MRD) on bladder cancer using based on urine. Our analysis shows that mutation, tumor burden, and copy number alterations are key features to detect MRD which is in line with biology. We developed a random forest model to predict MRD which is 87% sensitive for predicting residual disease in reference to gold-standard surgical pathology. Along with this, the survival analysis showed that our model is can be effective in clinical settings. Limitations of our study include the majority of patients having only single timepoint assessment of urine cell-free DNA. Other investigations utilizing plasma have shown that multiple samples obtained in surveillance settings can achieve greater sensitivity for detecting ctDNA MRD. We nevertheless achieved high ctDNA MRD sensitivity by multimodally analyzing urine. While our study was prospective, all samples were obtained from a single medical center. It will be important to corroborate our findings here in a multi-institutional prospective setting. Finally, given the prospective nature of our study with all patients enrolled between 2019 and 2021, the median follow-up time was modest at 23 months. It will be important to perform a study with longer follow-up to confirm the dramatic survival differences we observed.

In the future, we are hoping to extend our analysis based on these two works and combine all discussed modalities. In fact, there are many possibilities in this field such as fragment length-based analysis and nucleosome profiling.

Fragment length-based analysis uses the fact that the length of cfDNA fragment coming from the tumor is different from the normal cfDNA fragment. Usually, the genes in tumors are more expressed than in normal cells. As a result, the fragments coming from tumor cfDNA are less protected and get more fragmented, resulting in the shorter fragment. Based on this observation, there are several studies that differentiate cancer from healthy using the length of the fragments [32]. We are hoping to explore this area in the future.

In addition, transcription factor binding sites (TFBS) of a gene need to be accessible for that gene to be expressed. It means the TFBSs which are corresponding to expressed genes will lose coverage. Profiling the coverage of such sites can be an additional feature [12]. We are also thinking to include this feature in our future analysis. Hopefully, by extending our work using these features we will be able to develop an end-to-end Liquid Biopsy tool that will be able to help clinical decision-making.

References

[1] V. A. Adalsteinsson, G. Ha, S. S. Freeman, A. D. Choudhury, D. G. Stover, H. A. Parsons, G. Gydush, S. C. Reed, D. Rotem, J. Rhoades, et al. Scalable whole-exome sequencing of cell-free dna reveals high concordance with metastatic tumors. *Nature communications*, 8(1):1324, 2017.

[2] C. Alix-Panabieres. The future of liquid biopsy. *Nature*, 579(7800):S9–S9, 2020.

[3] N. M. Anderson and M. C. Simon. The tumor microenvironment. *Current Biology*, 30(16):R921–R925, 2020.

[4] D. Arneson, X. Yang, and K. Wang. Methylresolver—a method for deconvoluting bulk dna methylation profiles into known and unknown cell contents. *Communications biology*, 3(1):1–13, 2020.

[5] Blueprint. http://dcc.blueprint-epigenome.eu/#/home. Accessed: 2022-03-05.

[6] A. Chakravarthy, A. Furness, K. Joshi, E. Ghorani, K. Ford, M. J. Ward, E. V. King, M. Lechner, T. Marafioti, S. A. Quezada, et al. Pan-cancer deconvolution of tumour composition using dna methylation. *Nature communications*, 9(1):1–13, 2018.

[7] A. A. Chaudhuri, J. J. Chabon, A. F. Lovejoy, A. M. Newman, H. Stehr, T. D. Azad, M. S. Khodadoust, M. S. Esfahani, C. L. Liu, L. Zhou, et al. Early detection of molecular residual disease in localized lung cancer by circulating tumor dna profiling. *Cancer discovery*, 7(12):1394–1403, 2017.

[8] P. S. Chauhan, K. Chen, R. K. Babbra, W. Feng, N. Pejovic, A. Nallicheri, P. K. Harris, K. Dienstbach, A. Atkocius, L. Maguire, et al. Urine tumor dna detection of minimal residual disease in muscle-invasive bladder cancer treated with curative-intent radical cystectomy: A cohort study. *PLoS medicine*, 18(8):e1003732, 2021.

[9] P.-L. Chen, W. Roh, A. Reuben, Z. A. Cooper, C. N. Spencer, P. A. Prieto, J. P. Miller, R. L. Bassett, V. Gopalakrishnan, K. Wani, et al. Analysis of immune signatures in longitudinal tumor samples yields insight into biomarkers of response and mechanisms of resistance to immune checkpoint blockade. *Cancer discovery*, 6(8):827–837, 2016.

[10] R.-I. Chin, K. Chen, A. Usmani, C. Chua, P. K. Harris, M. S. Binkley, T. D. Azad, J. C. Dudley, and A. A. Chaudhuri. Detection of solid tumor molecular residual disease (mrd) using circulating tumor dna (ctdna). *Molecular diagnosis & therapy*, 23(3):311–331, 2019.

[11] M. G. K. Cumberbatch, I. Jubber, P. C. Black, F. Esperto, J. D. Figueroa, A. M. Kamat, L. Kiemeney, Y. Lotan, K. Pang, D. T. Silverman, et al. Epidemiology of bladder cancer: a systematic review and contemporary update of risk factors in 2018. *European urology*, 74(6):784–795, 2018.

[12] A.-L. Doebley, M. Ko, H. Liao, A. E. Cruikshank, K. Santos, C. Kikawa, J. B. Hiatt, R. D. Patton, N. De Sarkar, K. A. Collier, et al. A framework for clinical cancer subtyping from nucleosome profiling of cell-free dna. *Nature Communications*, 13(1):7475, 2022.

[13] J. C. Dudley, J. Schroers-Martin, D. V. Lazzareschi, W. Y. Shi, S. B. Chen, M. S. Esfahani, D. Trivedi, J. J. Chabon, A. A. Chaudhuri, H. Stehr, et al. Detection and surveillance of bladder cancer using urine tumor dna. *Cancer discovery*, 9(4):500–509, 2019.

[14] T. Forshew, M. Murtaza, C. Parkinson, D. Gale, D. W. Tsui, F. Kaper, S.-J. Dawson, A. M. Piskorz, M. Jimenez-Linan, D. Bentley, et al. Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma dna. *Science translational medicine*, 4(136):136ra68–136ra68, 2012.

[15] W. H. Fridman, F. Pages, C. Sautes-Fridman, and J. Galon. The immune contexture in human tumours: impact on clinical outcome. *Nature Reviews Cancer*, 12(4):298–306, 2012.

[16] T. F. Gajewski, H. Schreiber, and Y.-X. Fu. Innate and adaptive immune cells in the tumor microenvironment. *Nature immunology*, 14(10):1014–1022, 2013.

[17] M. D. Galsky, A. Saci, P. M. Szabo, G. C. Han, G. Grossfeld, S. Collette, A. Siefker-Radtke, A. Necchi, and P. Sharma. Nivolumab in patients with advanced platinum-resistant urothelial carcinoma: Efficacy, safety, and biomarker analyses with extended follow-up from checkmate 275checkmate 275 extended follow-up and biomarker analyses. *Clinical Cancer Research*, 26(19):5120–5128, 2020.

[18] H. B. Grossman, R. B. Natale, C. M. Tangen, V. Speights, N. J. Vogelzang, D. L. Trump, R. W. d. White, M. F. Sarosdy, D. P. Wood Jr, D. Raghavan, et al. Neoadjuvant chemotherapy plus cystectomy compared with cystectomy alone for locally advanced bladder cancer. *New England Journal of Medicine*, 349(9):859–866, 2003.

[19] S. Guo, D. Diep, N. Plongthongkum, H.-L. Fung, K. Zhang, and K. Zhang. Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma dna. *Nature genetics*, 49(4):635–642, 2017.

[20] D. S. Han and Y. D. Lo. The nexus of cfdna and nuclease biology. *Trends in Genetics*, 2021.

[21] J. Harper and R. C. Sainson. Regulation of the anti-tumour immune response by cancer-associated fibroblasts. In *Seminars in cancer biology*, volume 25, pages 69–77. Elsevier, 2014.

[22] E. Heitzer, L. Auinger, and M. R. Speicher. Cell-free dna and apoptosis: how dead cells inform about the living. *Trends in molecular medicine*, 26(5):519–528, 2020.

[23] J. A. Joyce and J. W. Pollard. Microenvironmental regulation of metastasis. *Nature reviews cancer*, 9(4):239–252, 2009.

[24] S. R. Kennedy, M. W. Schmitt, E. J. Fox, B. F. Kohrn, J. J. Salk, E. H. Ahn, M. J. Prindle, K. J. Kuong, J.-C. Shen, R.-A. Risques, et al. Detecting ultralow-frequency mutations by duplex sequencing. *Nature protocols*, 9(11):2586–2606, 2014.

[25] F. Krueger and S. R. Andrews. Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *bioinformatics*, 27(11):1571–1572, 2011.

[26] D. M. Kurtz, J. Soo, L. Co Ting Keh, S. Alig, J. J. Chabon, B. J. Sworder, A. Schultz, M. C. Jin, F. Scherer, A. Garofalo, et al. Enhanced detection of minimal residual disease by targeted sequencing of phased variants in circulating tumor dna. *Nature biotechnology*, pages 1–11, 2021.

[27] W. Li, Q. Li, S. Kang, M. Same, Y. Zhou, C. Sun, C.-C. Liu, L. Matsuoka, L. Sher, W. H. Wong, et al. Cancerdetector: ultrasensitive and non-invasive cancer detection at the resolution of individual reads using cell-free dna methylation sequencing data. *Nucleic acids research*, 46(15):e89–e89, 2018.

[28] H. Liang, J. Huang, X. Ao, W. Guo, Y. Chen, D. Lu, Z. Lv, X. Tan, W. He, M. Jiang, et al. Tmb and tcr are correlated indicators predictive of the efficacy of neoadjuvant chemotherapy in breast cancer. *Frontiers in Oncology*, 11:740427, 2021.

[29] Liquid biopsy: Using dna in blood to detect, track, and treat cancer. https://www.cancer.gov/news-events/cancer-currents-blog/2017/liquid-biopsy-detects-treats-cancer. Accessed: 2021-11-13.

[30] Y. Liu, K. D. Siegmund, P. W. Laird, and B. P. Berman. Bis-snp: combined dna methylation and snp calling for bisulfite-seq data. *Genome biology*, 13(7):1–14, 2012.

[31] P. Mandel. Nucleic acids in blood plasma in 1 man. *CR Seances Soc Biol Fil*, 142:241–243, 1948.

[32] D. Mathios, J. S. Johansen, S. Cristiano, J. E. Medina, J. Phallen, K. R. Larsen, D. C. Bruhm, N. Niknafs, L. Ferreira, V. Adleff, et al. Detection and characterization of lung cancer using cell-free dna fragmentomes. *Nature communications*, 12(1):1–14, 2021.

[33] E. J. Moding, B. Y. Nabet, A. A. Alizadeh, and M. Diehn. Detecting liquid remnants of solid tumors: Circulating tumor dna minimal residual diseasectdna minimal residual disease in solid tumors. *Cancer discovery*, 11(12):2968–2986, 2021.

[34] L. D. Moore, T. Le, and G. Fan. Dna methylation and its basic function. *Neuropsychopharmacology*, 38(1):23–38, 2013.

[35] J. Moss, J. Magenheim, D. Neiman, H. Zemmour, N. Loyfer, A. Korach, Y. Samet, M. Maoz, H. Druid, P. Arner, et al. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free dna in health and disease. *Nature communications*, 9(1):1–12, 2018.

[36] M. Mossanen, R. E. Krasnow, D. V. Zlatev, W. S. Tan, M. A. Preston, Q.-D. Trinh, A. S. Kibel, G. Sonpavde, D. Schrag, B. I. Chung, et al. Examining the relationship between complications and perioperative mortality following radical cystectomy: a population-based analysis. *BJU international*, 124(1):40–46, 2019.

[37] M. Murtaza, S.-J. Dawson, D. W. Tsui, D. Gale, T. Forshew, A. M. Piskorz, C. Parkinson, S.-F. Chin, Z. Kingsbury, A. S. Wong, et al. Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma dna. *Nature*, 497(7447):108–112, 2013.

[38] C. G. A. R. Network et al. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature*, 507(7492):315, 2014.

[39] A. M. Newman, S. V. Bratman, J. To, J. F. Wynne, N. C. Eclov, L. A. Modlin, C. L. Liu, J. W. Neal, H. A. Wakelee, R. E. Merritt, et al. An ultrasensitive method for quantitating circulating tumor dna with broad patient coverage. *Nature medicine*, 20(5):548–554, 2014.

[40] A. M. Newman, C. L. Liu, M. R. Green, A. J. Gentles, W. Feng, Y. Xu, C. D. Hoang, M. Diehn, and A. A. Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nature methods*, 12(5):453–457, 2015.

[41] A. M. Newman, A. F. Lovejoy, D. M. Klass, D. M. Kurtz, J. J. Chabon, F. Scherer, H. Stehr, C. L. Liu, S. V. Bratman, C. Say, et al. Integrated digital error suppression for improved detection of circulating tumor dna. *Nature biotechnology*, 34(5):547–555, 2016.

[42] A. M. Newman, C. B. Steen, C. L. Liu, A. J. Gentles, A. A. Chaudhuri, F. Scherer, M. S. Khodadoust, M. S. Esfahani, B. A. Luca, D. Steiner, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature biotechnology*, 37(7):773–782, 2019.

[43] J. C. Park, D. E. Citrin, P. K. Agarwal, and A. B. Apolo. Multimodal management of muscle-invasive bladder cancer. *Current problems in cancer*, 38(3):80–108, 2014.

[44] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[45] B. Pellini and A. A. Chaudhuri. Circulating tumor dna minimal residual disease detection of non–small-cell lung cancer treated with curative intent. *Journal of Clinical Oncology*, 40(6):567–575, 2022.

[46] F. Petrelli, A. Coinu, M. Cabiddu, M. Ghilardi, I. Vavassori, and S. Barni. Correlation of pathologic complete response with survival after neoadjuvant chemotherapy in bladder cancer treated with cystectomy: a meta-analysis. *European urology*, 65(2):350–357, 2014.

[47] M. Philip, L. Fairchild, L. Sun, E. L. Horste, S. Camara, M. Shakiba, A. C. Scott, A. Viale, P. Lauer, T. Merghoub, et al. Chromatin states define tumour-specific t cell dysfunction and reprogramming. *Nature*, 545(7655):452–456, 2017.

[48] E. J. Pietzak, A. Bagrodia, E. K. Cha, E. N. Drill, G. Iyer, S. Isharwal, I. Ostrovnaya, P. Baez, Q. Li, M. F. Berger, et al. Next-generation sequencing of nonmuscle invasive bladder cancer reveals potential biomarkers and rational therapeutic targets. *European urology*, 72(6):952–959, 2017.

[49] M. A. Postow, M. K. Callahan, and J. D. Wolchok. Immune checkpoint blockade in cancer therapy. *Journal of clinical oncology*, 33(17):1974, 2015.

[50] N. Riaz, J. J. Havel, V. Makarov, A. Desrichard, W. J. Urba, J. S. Sims, F. S. Hodi, S. Martín-Algarra, R. Mandal, W. H. Sharfman, et al. Tumor and microenvironment evolution during immunotherapy with nivolumab. *Cell*, 171(4):934–949, 2017.

[51] A. Ribas and J. D. Wolchok. Cancer immunotherapy using checkpoint blockade. *Science*, 359(6382):1350–1355, 2018.

[52] A. G. Robertson, J. Kim, H. Al-Ahmadie, J. Bellmunt, G. Guo, A. D. Cherniack, T. Hinoue, P. W. Laird, K. A. Hoadley, R. Akbani, et al. Comprehensive molecular characterization of muscle-invasive bladder cancer. *Cell*, 171(3):540–556, 2017.

[53] S. Y. Shen, R. Singhania, G. Fehringer, A. Chakravarthy, M. H. Roehrl, D. Chadwick, P. C. Zuzarte, A. Borgida, T. T. Wang, T. Li, et al. Sensitive tumour detection and classification using plasma cell-free dna methylomes. *Nature*, 563(7732):579–583, 2018.

[54] R. Siegel, K. Miller, H. Fuchs, and A. Jemal. Cancer statistics, 2021. *CA: a Cancer Journal for Clinicians*, 71(1):7–33, 2021.

[55] M. Stroun, P. Anker, P. Maurice, J. Lyautey, C. Lederrey, and M. Beljanski. Neoplastic characteristics of the dna found in the plasma of cancer patients. *Oncology*, 46(5):318–322, 1989.

[56] J. J. Szymanski, R. T. Sundby, P. A. Jones, D. Srihari, N. Earland, P. K. Harris, W. Feng, F. Qaium, H. Lei, D. Roberts, et al. Cell-free dna ultra-low-pass whole genome sequencing to distinguish malignant peripheral nerve sheath tumor (mpnst) from its benign precursor lesion: A cross-sectional study. *PLoS Medicine*, 18(8):e1003734, 2021.

[57] A. Thierry, S. El Messaoudi, P. Gahan, P. Anker, and M. Stroun. Origins, structures, and functions of circulating dna in oncology. *Cancer and metastasis reviews*, 35(3):347–376, 2016.

[58] D. S. Thommen and T. N. Schumacher. T cell dysfunction in cancer. *Cancer cell*, 33(4):547–562, 2018.

[59] V. Thorsson, D. L. Gibbs, S. D. Brown, D. Wolf, D. S. Bortone, T.-H. O. Yang, E. Porta-Pardo, G. F. Gao, C. L. Plaisier, J. A. Eddy, et al. The immune landscape of cancer. *Immunity*, 48(4):812–830, 2018.

[60] P. C. Tumeh, C. L. Harview, J. H. Yearley, I. P. Shintaku, E. J. Taylor, L. Robert, B. Chmielowski, M. Spasic, G. Henry, V. Ciobanu, et al. Pd-1 blockade induces responses by inhibiting adaptive immune resistance. *Nature*, 515(7528):568–571, 2014.

[61] E. M. Van Allen, D. Miao, B. Schilling, S. A. Shukla, C. Blank, L. Zimmer, A. Sucker, U. Hillen, M. H. G. Foppen, S. M. Goldinger, et al. Genomic correlates of response to ctla-4 blockade in metastatic melanoma. *Science*, 350(6257):207–211, 2015.

[62] J. C. Wan, C. Massie, J. Garcia-Corbacho, F. Mouliere, J. D. Brenton, C. Caldas, S. Pacey, R. Baird, and N. Rosenfeld. Liquid biopsies come of age: towards implementation of circulating tumour dna. *Nature Reviews Cancer*, 17(4):223–238, 2017.

[63] R.-h. Xu, W. Wei, M. Krawczyk, W. Wang, H. Luo, K. Flagg, S. Yi, W. Shi, Q. Quan, K. Li, et al. Circulating tumour dna methylation markers for diagnosis and prognosis of hepatocellular carcinoma. *Nature materials*, 16(11):1155–1161, 2017.

[64] R. Yang, S. Cheng, N. Luo, R. Gao, K. Yu, B. Kang, L. Wang, Q. Zhang, Q. Fang, L. Zhang, et al. Distinct epigenetic features of tumor-reactive cd8+ t cells in colorectal cancer patients revealed by genome-wide dna methylation analysis. *Genome biology*, 21(1):1–13, 2020.