Predicting cancer subtypes from nucleosome profiling of cell-free DNA

Anna-Lisa Doebley

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:

Gavin Ha, Chair

David MacPherson

Peter Nelson

Program Authorized to Offer Degree:

Molecular and Cellular Biology

University of Washington

Abstract

Predicting cancer subtypes from nucleosome profiling of cell-free DNA

Anna-Lisa R Doebley

Chair of the Supervisory Committee:
Gavin Ha, PhD, Assistant Professor
Public Health Sciences Division, Fred Hutchinson Cancer Center

Modern cancer treatments take advantage of genomic differences between tumors to kill cancer cells using targeted approaches. Typically, this requires a tumor biopsy in order to get tissue for phenotypic and genotypic analysis. However, in late-stage cancer, surgical biopsies of metastases may not be part of the standard of care and repeated biopsies cannot be performed for monitoring. However, late-stage cancer patients can benefit from targeted therapies that address specific tumor phenotypes and resistance mechanisms. Because of this, non-invasive diagnostic approaches are needed. Cell-free DNA (cfDNA) provides a promising approach for non-invasive tumor characterization. In cancer patients, a fraction of cfDNA is derived from tumor cells which have died and released their DNA into the bloodstream. This DNA remains wrapped around nucleosomes which protect it from degradation by apoptotic and plasma nucleases. cfDNA can

be extracted from the blood and sequenced, revealing which regions of the genome were protected by nucleosomes and which were more accessible in the tumor cells. This type of analysis, known as nucleosome profiling, has the potential to be used for phenotypic characterization of tumors, such as determining the activity of key transcription factors. In this thesis, we develop methods to perform nucleosome profiling in cfDNA from cancer patients. First, we quantify the effects of GC bias on nucleosome profiles and implement a GC correction procedure to reduce these impacts. Next, we develop a nucleosome profiling method called Griffin that uses this GC bias correction procedure and generates composite nucleosome profiles around transcription factor binding sites (TFBSs) and other accessible sites. We then apply this method to detect cancer in early-stage cancer patients. Finally, we identify assay for transposase-accessible chromatin using sequencing (ATAC-seq) sites that are differentially accessible in estrogen receptor (ER) positive or ER negative metastatic breast tumors and use Griffin nucleosome profiling around these sites to predict the ER status in cfDNA samples from breast cancer patients. Additionally, in a separate study, we design a targeted sequencing panel to predict the activity of lineage defining transcription factors in small-cell lung cancer (SCLC) and use this approach to predict SCLC subtypes from cfDNA.

A version of Chapter 2 is currently accepted in principle at Nature communications as a journal article entitled:

A framework for clinical cancer subtyping from nucleosome profiling of cell-free DNA

Anna-Lisa Doebley, Minjeong Ko, Hanna Liao, A. Eden Cruikshank, Katheryn Santos, Caroline Kikawa, Joseph B. Hiatt, Robert D. Patton, Navonil De Sarkar, Katharine A. Collier, Anna C.H. Hoge, Katharine Chen, Anat Zimmer, Zachary T. Weber, Mohamed Adil, Jonathan B. Reichel, Paz Polak, Viktor A. Adalsteinsson, Peter S. Nelson, David MacPherson, Heather A. Parsons, Daniel G. Stover, Gavin Ha

A version of Chapter 3 is currently in preparation for submission as a journal article entitled:

Characterizing small-cell lung cancer subtypes from cell-free DNA

Joseph Hiatt*, Anna-Lisa Doebley*, Henry Arnold, Holly Sandborg, Anish Thomas, Charlie Rudin, Keith Eaton, Gavin Ha[#], David MacPherson[#]

*These authors contributed equally to the work

[#]Corresponding authors

Acknowledgements

I would like to thank my doctoral advisor Gavin Ha for his support which has helped me grow as a researcher over the past three and a half years. He has been an invaluable source of ideas, advice, and unwavering enthusiasm for the potential for cfDNA technologies to improve the lives of cancer patients. I would also like to thank my co-advisor David MacPherson for his support and thoughtful mentorship. Additionally, I thank the other members of my committee, Stephen Tapscott, Trevor Bedford, and Pete Nelson for their guidance and expertise. Thanks to the members of the Ha Lab, Patty Galipeau, Adam Kreitzman, Anat Zimmer, Pushpa Itagi, Robert Patton, Sitapriya Morthi, Eden Cruikshank, Abigail Thorpe, David Chen, and Mohamed Adil, who have provided valuable discussion and made the lab a joyful place to work. Additionally thank you to the members of the MacPherson and Berger labs who have provided valuable feedback on my work at lab meetings. I would especially like to thank Joe Hiatt who has been a wonderful collaborator on the SCLC project and a great source of mentorship from an MSTP graduate. Also, thanks to Holly Sandborg and Henry Arnold for their help generating data for the SCLC project. Additionally, I want to thank our collaborators Dan Stover and Heather Parson who provided invaluable data for the MBC project. Thank you to the MSTP and MCB programs who have been supportive of me and my journey through my PhD. And last, thank you to my friends and family who have supported and cheered for me through this whole PhD process. These past few years of the pandemic have made me even more grateful for those around me and the joy they bring into my life.

Table Of Contents

List of Figures

Chapter 1. Introduction

1.1 Discovery of cfDNA

Nucleic acids in the bloodstream, including cell-free DNA (cfDNA), were first described in a 1948 study by Mandel and Métais, but it was many years before the source and potential utility of this DNA, particularly in cancer, was better understood[1]. In 1977, before PCR or Sanger sequencing were available, Leon and colleagues used radioimmune labeling to find that cfDNA concentrations were increased in cancer patients, especially after metastasis[2]. They noted that a decrease in cfDNA levels after therapy was associated with treatment response and better prognosis and proposed the use of cfDNA as a prognostic biomarker indicating successful response to treatment. However, it took many years and a series of technological and scientific advances for cfDNA analysis in cancer patients to grow into a major field of research and begin making its way into the clinic.

1.1.1 *Cellular origins of cfDNA*

Although healthy individuals all have cfDNA in their bloodstreams, it is difficult to determine the exact origin of this cfDNA because healthy cells all share the same genetic sequences regardless of their tissue of origin. A 2002 study by Lui and colleagues got around this limitation by examining patients with sex-mismatched bone marrow transplants[3]. By quantifying the fraction of cfDNA containing Y chromosome sequences, these researchers found that the cfDNA was primarily derived from the bone marrow donor rather than the patient indicating that hematopoietic cells were the primary source of cfDNA. It was many years before the increased availability of whole genome sequencing technologies allowed more specifics to be worked out using genome wide bisulfite sequencing to detect tissue specific methylation patterns. In a 2015 study, Sun and colleagues found that cfDNA was primarily derived from neutrophils, with a smaller contribution from lymphocytes (B cells and T cells) and liver cells based upon cell type-specific methylation marks for 16 tissue types[4]. They validated their approach by showing that it was able to accurately predict the percentage of placental derived cfDNA in pregnant patients (validated

using fetus specific sequences such as the y chromosome) and in patients with liver and bone marrow transplants (validated using donor specific single nucleotide polymorphisms (SNPs)).

In cancer patients, as early as 1989 a study by Stroun and colleagues showed evidence that the high level of cfDNA observed in cancer patients was derived from tumor cells[5]. Later, the development of PCR enabled researchers to detect cancer specific mutations in cfDNA and a study by Sorenson and colleagues demonstrated the presence of KRAS mutations in the cfDNA of patients with pancreatic cancer[6]. Additional studies, such as one by Jahr and colleagues in 2001 further supported this finding by showing cancer specific methylation marks in a fraction of the cfDNA of cancer patients using bisulfite sequencing[7]. These findings raised the possibility of detecting and characterizing cancer from non-invasive collection of cfDNA.

1.1.2 *Mechanisms of cfDNA release*

cfDNA is released through apoptosis, necrosis, and active secretion. The 2001 study by Jahr also provided some of the first experimental evidence that cfDNA was derived from apoptosis and necrosis[7]. This study induced apoptotic or necrotic liver injury in mice and monitored the concentration and fragment length of the resulting cfDNA. They found that both apoptosis and necrosis lead to large increases in cfDNA but with different fragment lengths. Apoptotic cells produced the characteristic 1 or 2 nucleosome (167bp or 334bp) fragments often observed in cfDNA, while necrotic cells produced much longer (>10,000bp) fragments with a much higher concentration than apoptosis. This suggested that the majority of cfDNA is from apoptosis, however, more recent studies such as one by Watanabe and colleagues using mouse knockouts of key nucleases and induced necrosis have suggested that necrotic cfDNA is also broken down into nucleosome sized units and may contribute to short fragments in cfDNA[8,9]. A 2020 study by Rostami and colleagues using mouse xenografts suggested that necrosis was the primary cfDNA release mechanism after treatment with ionizing radiation, although apoptosis also played a role[10]. Proliferating tumors, like other proliferating tissues, undergo apoptosis related to proliferation providing a possible explanation for the release of cfDNA from tumors as discussed in a

recent review by Heitzer and and colleagues[11]. Other studies suggest that it is also possible for cfDNA to be released via active secretion in exosomes from tumor cells[1,12–14].

Recent studies of nucleases, including many from the lab of Dennis Lo, have contributed further evidence to the theory that apoptosis plays a major role in cfDNA fragmentation[9]. A 2018 study by Cheng and colleagues found that knockout of DNASE1 (a key DNAse secreted into the plasma) did not impact the amount of cfDNA or fragment size for either single or double stranded cfDNA in mice[15]. However, the authors of a later study by Serpas and colleagues noted that DNASE1 is known to act primarily on naked DNA, while DNASE1L3 (another plasma DNAse) digests chromatin[16]. When these researchers deleted DNASE1L3 in mice, they found that the fraction of short and long cfDNA fragments was increased, while the number of typical nucleosome-sized (167bp) fragments was decreased in the DNASE1L3 deleted mice. Additionally, the most common fragment endpoint motifs which all ended in 'CC' were less frequently observed in DNASE1L3 deleted mice suggesting that preferential cleaving of cfDNA by DNASE1L3 created these common motifs. These alterations in both the fragment size and end motifs in response to DNASE1L3 loss suggested that this nuclease plays a key role in cfDNA fragmentation and provided further evidence that apoptosis was responsible for cfDNA fragmentation as DNASE1L3 is known to be active during apoptosis[17]. Finally, the preference of DNASE1L3 for chromatin suggested that cfDNA is likely bound to nucleosomes in the plasma, rather than being naked DNA which would be impacted by DNASE1 digestion[18]. A more recent study by Han and colleagues found that DNASE1 may play a minor role in cfDNA fragmentation after all, particularly for shorter DNA fragments that are not believed to be bound to nucleosomes[18]. These fragments show an increased fraction of 'T' nucleotides at fragment ends, suggesting a different nuclease is cutting them. When DNASE1 activity was enhanced using heparin, these short fragments were enriched, especially those with 'T' ends. This did not occur in DNASE1 knockout mice suggesting that DNASE1 may contribute to the creation of subnucleosomal 'T' end fragments in cfDNA. In this same study, intracellular apoptotic DNAse 'DNA fragmentation factor B' (DFFB) was also found to play a key role in intracellular apoptotic DNA cleavage before release into the plasma. This nuclease preferentially creates cfDNA fragments with

'A' ends. When apoptosis occurs after inactivating the plasma DNAses (DNASE1L3 and DNAse1) with EDTA these 'A' fragments are preferentially enriched. This does not occur in DFFB knockout mice suggesting that DFFB is responsible for creating these 'A' fragment ends before cfDNA is released from cells. This study led to a proposed model of (1) intracellular DNA cleavage by intracellular apoptotic DNAses such as DFFB and DNASE1L3 followed by (2) cfDNA release and further degradation by extracellular DNASE1L3 and finally (3) further degradation of short fragments of naked DNA by DNASE1. Meanwhile in necrosis, large fragments >10,000bp are released from cells and subsequently degraded into nucleosome sized fragments by DNASE1L3 followed by further degradation of naked DNA by DNASE1. However, additional studies to confirm this and explore the role of additional nucleases are needed[18].

1.2 Applications of cfDNA for cancer diagnostics

Because a fraction of cfDNA is released from tumor cells in patients with cancer, an expanding field of research has emerged using cfDNA for 'liquid biopsies'. Unlike a traditional tumor biopsy, cfDNA can be obtained non-invasively allowing for screening and monitoring without the risk of a more invasive procedure. Assays to detect genomic alterations have advanced the furthest clinically with some FDA approved assays being available. This has several potential applications including cancer screening, determining somatic alterations for cancer subtyping and treatment selection, prediction of treatment response and prognosis, and monitoring for disease recurrence or treatment resistance[1].

1.2.1 *cfDNA assays for cancer screening*

Many studies have attempted to develop assays for cancer screening from cfDNA in the general population. Such assays have the potential benefit of being able to detect early-stage cancers in a broad variety of tissue types including those such as pancreatic cancer where screening isn't available and disease is often detected at a late stage leading to poor outcomes. In an early proof of concept study, Diehl and colleagues found APC mutations, a driver of colorectal cancer, in the blood of 60% of patients with

early-stage colorectal cancer demonstrating that it may be possible to screen people for cancer using cfDNA[19]. In recent years, clinical trials have expanded this targeted approach to look for multiple driver gene mutations. One example of this type of assay is CancerSEEK, a targeted sequencing panel which allows deep sequencing of 61 regions in 16 genes containing common driver mutations that are seen frequently in various types of cancer[20]. This test was able to detect cancer in a median of 70% of cancer patients with specificity greater than 99%. However, this study was performed in patients with known cancer and healthy controls, potentially overestimating its performance in a true screening population where non-symptomatic cancers may be harder to detect and patients without cancer may have other conditions that would lead to false positives. A study by Lennon and colleagues which utilized an early version of CancerSeek combined with blood protein biomarkers (called the DETECT-A blood test) tested this approach in a true screening population of 10,006 women aged 65-75 without cancer at the time of enrollment[21]. Participants with a positive mutation or protein marker (490 participants, 4.9% of all participants) were tested again to check if the abnormality was still present and exclude clonal hematopoiesis of indeterminant potential (CHIP). Less than a third of patients with an initial positive (134 participants, 1.35% of all participants) were positive on the second test and most of the false positives on the first test were determined to be due to CHIP. If the second test was positive and the patient's medical history did not provide a likely non-cancer explanation, the patient underwent a full body PET-CT scan. Of these individuals, 26 were found to have cancer, 15 of whom had a mutation detected from cfDNA. This study showed that blood screening, including targeted cfDNA mutational screening can detect cancers in the general population without excessive false positives, potentially detecting cancer at an earlier and more treatable stage. However, most of the cancers detected were also detected by standard of care screening or symptoms during the same time frame or were already late stage, likely limiting the impacts of this type of screening[21].

1.2.2 *cfDNA assays for treatment selection*

Another potential use for genomic changes in cfDNA is for identifying tumor susceptibilities to targeted therapies. For instance, in neuroblastoma, MYCN amplifications are associated with worse outcomes and are treated with different therapies, such as immunotherapies, to improve survival rates. In late-stage neuroblastoma patients with MYCN amplifications, the amplification could be detected in cfDNA with high sensitivity[22]. An early application of mutation detection in cfDNA to predict treatment response and prognosis was published by Kimura and colleagues in 2006[23]. In this study, researchers found that EGFR mutations found in cfDNA of non-small-cell lung cancer patients were predictive of EGFR inhibitor response and progression free survival. In the following years, many additional studies looked at EGFR mutation detection from cfDNA and in recent years, the US food and drug administration has approved cfDNA detection of EGFR mutations for therapy detection[1]. In addition to single gene screening, multi-gene cfDNA panels have been developed to look at multiple regions of interests. In 2020, a 61 gene panel known as Guardant360 CDx and a 324 gene panel known as F1 Liquid CDx were approved by the FDA for patients with solid tumors[24]. Although Guardant360 CDx was initially approved specifically as a companion diagnostic for identifying non-small cell lung cancer (NSCLC) patients with EGFR mutations who might benefit from EGFR inhibitors, it can also pick up many other actionable mutations and is approved to be used for clinical decision making regarding these therapies. F1 Liquid CDx is approved as a companion diagnostic for a broader range of cancer types (NSCLC, prostate, ovarian, and breast cancers) and for several actionable mutations (EGFR, ALK, BRCA1/2, ATM, and PIK3CA) but can be used in clinically in any solid tumor to detect many more actionable mutations[24]. Many more assays are in development and the use of cfDNA for detecting actionable mutations in cancer patients is likely to become increasingly prevalent.

1.2.3 *cfDNA assays for detection of minimal residual disease*

A third possible use for cfDNA in cancer screening is for detection of minimal residual disease or recurrence. This can be achieved by identifying tumor specific mutations in the primary tumor and

designing a patient specific panel to look for the emergence of these mutations at low levels in cfDNA. For instance, a 2019 prospective study by Garcia-Murillas and colleagues examined tumors from 170 early-stage breast cancer patients found breast cancer driver mutations in 101 primary tumors and then designed custom mutation panel for each of the 101 patients to track these mutations in cfDNA[25]. 16 patients had detectable mutations in their cfDNA during follow up and these patients were more likely to have disease relapse during this time. Other studies have also shown similar ability to detect relapse using patient specific mutation panels in cfDNA. However, an open question remains of how to treat patients with evidence of ctDNA relapse prior to clinically apparent relapse and more studies are needed to determine the best use for these assays[24].

1.2.4 *cfDNA assays for quantifying tumor content*

One key concern which impacts any application of cfDNA to cancer genomics is the tumor fraction or amount of the cfDNA that is derived from the tumor. As discussed previously, healthy tissues, including nucleated blood cells and liver cells, shed DNA into the bloodstream during normal cell turnover. This process does not stop in cancer patients. Consequently, in the absence of large tumors shedding large amounts of DNA, the vast majority of cell-free DNA will be derived from healthy cells. This presents a challenge for both detecting and characterizing tumors from cfDNA because it is difficult to tell whether the cancer or mutation of interest is absent or simply below the threshold of detection for the given assay. Because of this, it is often necessary to quantify the tumor fraction in order to correctly interpret cfDNA results. This can be accomplished by examining mutations or copy number alterations. Mutation analysis requires deep sequencing either from whole exomes or whole genomes with matched germline sequencing[26,27]. A less costly alternative involves looking for copy number alterations in ultra-low pass (0.1x) whole genome sequencing as demonstrated by Adalsteinsson and colleagues[28]. Although this depth is too shallow to detect mutations, large (chromosome arm level) copy number alterations are common in metastatic cancer and can be detected using copy ratios. The ichorCNA software uses a probabilistic model to identify copy number alterations and predict the tumor fraction of cell-free DNA

samples. This allows detection of cancer down to 3% tumor fraction from ultra-low pass sequencing, providing a cost-effective option for determining whether a sample has enough tumor content to make deeper sequencing worthwhile.

1.3 Fragmentation patterns in cfDNA

While much work on cfDNA has focused on genomic alterations, the unique biological processes that generate cfDNA have led to the development of a new field known as fragmentomics, or the study of cfDNA fragmentation patterns. Unlike traditional sequencing where long strands of genomic DNA need to be sheared into small pieces before sequencing, cfDNA fragmentation occurs as the result of biological processes in-vivo. This fragmentation is non-random and instead is dictated by mechanisms of cell death, intracellular nucleases, and plasma nucleases, and the chromatin organization in the cells of origin[17,29]. Segments of DNA that are wrapped around nucleosomes are protected from degradation while regions of open chromatin lack this protection[29]. This raises the possibility of inferring the epigenetic state of cells of origin from the fragmentation patterns seen in cfDNA.

1.3.1 *cfDNA fragment size and endpoints*

cfDNA has a typical fragment length of approximately 167 bp and a characteristic fragment size distribution with additional peaks at approximately 10bp intervals below 167bp. This pattern is evidence for the apoptotic origin of cfDNA and the persisting nucleosome protection in the blood, which causes DNA to be cleaved at 10bp intervals[17]. However, different cfDNA cells of origin and epigenetic states can contribute cfDNA with different fragment length distributions and there has been much study of these fragment length profiles. Early studies of fragment length had mixed results with some studies reporting that cancer derived cfDNA had better integrity (i.e. few short reads), while others reported a longer length, and still others reported a shorter length[30]. Analysis of tumor specific mutations in cfDNA began to shed light on the true fragment length distribution of tumor derived cfDNA. In a 2005 proof of concept study of colorectal cancer detection, Diehl and colleagues found that short fragments (100bp) were

enriched for APC driver mutations, suggesting that tumor derived fragments were shorter than healthy blood fragments[19]. In another 2011 study by Mouliere and colleagues, researchers used mouse xenografts to examine the fragment lengths of tumor and non-tumor DNA and also found that short (<100bp) fragments were highly enriched for tumor DNA and that this was true in both xenografts and human colorectal cancers[31]. As whole genome sequencing (WGS) became more widely available, this technology enabled more thorough investigation of size profiles in cfDNA. A 2015 study by Jiang and colleagues used whole genome sequencing of cfDNA in hepatocellular carcinoma patients and showed a shift towards shorter fragments in cancer patients, relative to controls and showed that these shorter fragments were relatively more enriched in regions amplified in the tumor relative to regions deleted in the tumor[30]. This demonstrated that the shorter fragments were derived specifically from tumor cells rather than due to some global process that impacted all cfDNA during cancer. In another study, Underhill and colleagues used whole genome sequencing of cfDNA from cell-line derived xenografts and found a similar shortening of tumor derived cfDNA with a typical length of 132-145bp and confirmed this finding in human melanoma and human lung cancer patients and showed that enriching for short fragments also enriched for mutant alleles[32]. Another study in 2018 by Mouliere and colleagues further confirmed the finding that tumor derived cfDNA was shorter than healthy cfDNA in a variety of tumor types[33]. In addition, this study showed that selecting for shorter fragments either before or after sequencing could increase the ability to detect cancer associated mutations or copy number alterations. Additionally, a proof of concept study by Chabon and colleagues leveraged the shorter fragments observed in cancer derived cfDNA to distinguish tumor mutations from mutations observed in clonal hematopoiesis, a common condition in older individuals in which a hematopoietic stem cell acquires a mutation and expands over time without becoming cancerous[34]. Mutations found in clonal hematopoiesis are often similar to those found in cancer and can cause false positives on cfDNA cancer screening tests, so the ability to distinguish them by fragment size may be useful.

The relative enrichment of short fragments in cancer derived cfDNA lead to attempts to use fragment size profiles, rather than mutations, for cancer detection. One method, called DELFI was

developed by Cristiano and colleagues and analyzes fragment sizes in shallow (1-2x) whole genome sequencing[35]. This method uses the ratio of short (<150bp) to long (>150bp) fragments in 5mb bins across the genome. These researchers examined a dataset of several hundred cancer and healthy cfDNA samples and found that the short to long fragment ratios in cancer were often outside of the typical range for healthy donors. They then built a machine learning model based on these values and were able to detect cancer with high sensitivity and specificity. A follow up study by Mathios and colleagues used a similar methodology in a prospective study of patients at high risk of lung cancer[36]. cfDNA was collected from patients with a high risk of lung cancer, often due to smoking and either symptoms of lung cancer or incidental radiologic findings suggestive of lung cancer. Of the 365 patients enrolled, 129 were diagnosed with lung cancer within 44 days of the cfDNA collection. Shallow WGS was performed on the cfDNA and the fragment size profiles and copy number alterations were examined. The researchers then trained a machine learning model and were able to detect cancer with 0.90 AUC, although this varied significantly between stage I (0.76) and stage IV (0.92) likely due to the lower tumor cfDNA contribution in patients with early-stage tumors.

Despite the extensive evidence that cancer cfDNA is shorter, and the multitude of studies that have attempted to use this knowledge for cancer detection, the cause of these shorter fragments remains unclear. Changes in size profiles have been better studied in fetal cfDNA, which, like cancer cfDNA, is shorter (around 142bp) than the background of haematopoietically derived maternal cfDNA (around 167bp). In the case of fetal cfDNA, Sun and colleagues demonstrated in a 2018 study that the shorter fragments correspond to the stretches of DNA protected by nucleosome cores, while longer fragments are protected by both cores and linkers[37]. The loss of the linker coverage has been proposed to be due to hypomethylation of placental DNA compared to hematopoietic DNA[38]. Hypomethylation results in less tightly packed chromatin which renders it accessible to DNAses. Although the biological mechanism that leads to short fragments in cancer has been less well studied, cancer cells often have hypomethylated DNA, potentially leading to the same increased DNAse accessibility[38]. Another hypothesis for the cause of shorter cfDNA fragments is a decrease in DNASE1L3 expression seen in tumor cells relative to similar

non-tumor tissues. DNASE1L3 loss in mice results in changes in fragment end points, with a loss of common 'CC' motifs and an increased proportion of short fragments, similar to that observed in tumor derived cfDNA so potentially, the lower expression in cancer tissues could have a similar effect[16,39]. The potential to use this change in fragment endpoints for cancer detection from an extremely small number of reads has been shown by Jiang and colleagues[39], highlighting the potential value of better understanding of the biology of cfDNA fragmentation.

1.4 Prediction of nucleosome positions from cfDNA

Nucleosomes protect cfDNA from degradation and consequently can provide information about the location of nucleosomes in the cell of origin. This has the potential to provide important information because the location of nucleosomes is not random and instead is determined by the epigenetic state of the cell, and the activity of elements such as transcription factor binding sites (TFBSs) and transcriptional start sites (TSS). Recent work has focused on predicting and interpreting nucleosome positions from cfDNA with the aim of being able to gain information about the cells of origin including the presence of cancer, type of cancer, cancer subtypes, resistance mechanisms, and gene expression. This emerging field is known as nucleosome profiling.

The first demonstrations of nucleosome profiling were in 2015 when Snyder and colleagues used deep whole genome sequencing of healthy and cancer cfDNA to discover the position of nucleosomes and TFs in the cells of origin[29]. They initially sequenced standard whole genome cfDNA and developed a metric known as the windowed protection score (WPS) which quantified the amount of nucleosome protection at a given location by looking at the nucleosome sized fragments which fully overlapped it and subtracting the fragment endpoints. They demonstrated that cfDNA was characterized by clear nucleosome footprints spanning nearly the entire genome. They found that these nucleosome footprints were organized as expected around key genomic elements such as transcriptional start sites and CTCF binding sites, a constitutively active transcription factor that is involved in chromatin organization. They also used a single strand library preparation method to better capture the damaged ends of cfDNA and

found that this preparation method enriched for short fragments. There was increased coverage of these short fragments at active TFBSs such as CTCF, ETS, and MAFK sites, suggesting that they represented DNA that was protected by binding to the TF itself. They also found that active TSS, in aggregate, had stronger nucleosome positioning than inactive TSS. In addition, nucleosome spacing across gene bodies was correlated with the gene expression and this correlation was strongest for hematopoietic gene expression. Finally, they correlated the nucleosome spacing in cancer cfDNA samples to gene expression in various tissues and found that the correlation was often best for the tissue corresponding to the tumor types[29]. Although this analysis required deep sequencing and high tumor fraction, this landmark study demonstrated that cfDNA fragmentation patterns contained information about the chromatin structure and nucleosome positions in tumor cells and that these signals could be interpreted to gain information about the tumor type and possibly, gene expression.

Another study from that same year by Ulz and colleagues also explored gene expression prediction from cfDNA, with a focus on changes in coverage around TSS[40]. They also found a relationship between gene expression and nucleosome protection around TSS. They found decreased coverage and increased nucleosome amplitude around housekeeping genes and other highly expressed genes and increased coverage with weaker nucleosome amplitudes for lower expression genes. They found they could use this decrease in coverage to distinguish the most highly expressed genes from less expressed ones. They also showed that this was possible in two breast cancer cfDNA samples but predicted it would only be possible for genes in regions with above 0.75 tumor fraction. Although this technique was also reliant on deep whole genome sequencing and high tumor fraction, it was a promising start to gene expression prediction from cfDNA.

1.4.1 *Nucleosome profiling to predict gene expression*

The first studies demonstrating differences in nucleosome profiles between active and inactive genes sparked further research into using nucleosome profiling to predict gene expression. Because of the necessity of using deep sequencing to gain information about TSS, many of these studies have focused on

targeted panel sequencing of TSS to enable deep coverage of these sites without the cost of deep whole genome sequencing.

In a 2021 study, Zhu and colleagues validated that they saw a similar inverse correlation between gene expression and cfDNA coverage at the nucleosome depleted region (NDR) surrounding the TSSs in deep WGS[41]. They then examined genes that were differentially expressed in blood compared to colorectal cancer and used machine learning to find a set of only 6 genes (3 upregulated in blood and 3 upregulated in cancer) that could be used to predict tumor fraction in cfDNA samples. They used a targeted panel which captured a window of 4,000 bp around each of these 6 TSS to deeply sequence samples and showed that this relatively small panel was able to accurately predict tumor fraction and detect cancer. However, it required a large amount of training data to select a handful of genes that showed good predictive value and did not attempt to estimate gene expression from these TSS.

A later study by Esfahani and colleagues also examining targeted panel sequencing of TSS but took a somewhat different approach[42]. They observed that the fragment size distributions at TSS were more predictive of gene expression than the NDR coverage, although both metrics provided complementary data. They then identified groups of genes that were differentially expressed between lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) and between subtypes of diffuse large B cell lymphoma while also having relatively low expression in healthy blood. They designed a targeted panel to capture a 2kb window around these genes (an approach that they called EPIC-seq) and showed that the features extracted from this targeted panel could be used to predict LUAD or LUSC and to predict clinical response to immune checkpoint inhibitors. They showed a similar ability to predict prognostic subtypes of diffuse large B cell lymphoma. EPIC-seq appears to be a promising approach for cancer subtyping and tumor fraction estimation that could potentially be adapted to different types of cancer.

1.4.2 *Nucleosome profiling to predict TFBS activity*

The early studies of nucleosome profiling also sparked research into using cfDNA nucleosome profiling to predict transcription factor (TF) activity. In many types of cancer, key TFs are responsible for tumor cell characteristics, response to therapy, and prognosis. In many cases, TFs are also the targets of therapy including androgen deprivation therapy in prostate cancer[43] and estrogen receptor antagonists in breast cancer[44]. Because of this, predicting TF expression from cfDNA could be hugely beneficial for a multitude of purposes including detecting cancer, subtyping cancer, and monitoring response or resistance to treatment.

A 2019 study by Ulz and colleagues was the first to take an in-depth look at TFBS accessibility in cancer cfDNA[45]. This study used a database of chromatin immunoprecipitation sequencing (ChIP-seq) experiments to find a thousand high confidence TFBSs each for hundreds of transcription factors. They then used WGS of cfDNA to generate composite profiles around these factors. In healthy donors, they found evidence of nucleosome organization around TFs involved in hematopoiesis. They also found that cancer samples had nucleosome organization around some cancer specific TFs and reduced organization at some blood TFs, corresponding to the reduced fraction of blood derived cfDNA. They used these signals to build a cancer detection model. But more interestingly, they also demonstrated that a set of subtype specific factors including AR, HOXB13, NKX3-1 and REST were differentially accessible between androgen receptor active prostate cancer (ARPC) and neuroendocrine prostate cancer (NEPC), demonstrating that this method might be able to distinguish TF based subtypes.

A few other studies have explored this concept of nucleosome profiling around transcription factor binding sites in order to predict transcription factor activity in cfDNA studies. A 2021 report by Mathios and colleagues that was mostly focused on fragment length and lung cancer detection, looked briefly at ASCL1 binding sites to distinguish small cell lung cancer from non-small cell lung cancer[36]. ASCL1 is a key transcription factor expressed in a large fraction of SCLC tumors. Their study showed that accessibility at these sites was much higher in SCLC cfDNA compared to NSCLC cfDNA and that

this accessibility could be used to distinguish these two tumor types. Additionally, a study by Herberts and colleagues primarily looking at genomic alterations and subclonal evolution in prostate cancer also examined AR binding sites and showed that AR accessibility was seen in ARPC but not NEPC[46]. They also showed an individual in which AR accessibility was lost post treatment, suggesting an evolution to NEPC in response to androgen deprivation.

1.4.3 *Differential accessibility for tumor subtyping*

While many studies of nucleosome profiling have focused on sites with known biological importance such as TSS and TFBSs, these sites are not necessarily the most informative for applications such as subtyping or cancer detection. Several studies have instead used chromatin accessibility data from published assays on tumors to identify the sites with the most differential accessibility regardless of biological function. This improves the chance that sites will be informative for distinguishing the cell types of interest, although less information is gained about the biological significance of the differences. Sun and colleagues published one of the first studies to examine nucleosome profiles around differential open chromatin sites using tissue specific DNAse hypersensitivity sites (DHS)[47]. They then examined the number of fragment starts and ends flanking the open chromatin sites and developed a metric called orientation-aware cfDNA fragmentation (OCF) to quantify the amount of protection at the site, relative to the protection at the flanking nucleosomes. They found that OCF for sites specific to a given tissue type correlated with the contribution of that tissue type such as healthy blood, placental fraction in pregnant patients, and cancer fraction in hepatocellular carcinoma patients.

Another study by Peneder and colleagues, looking at detection of pediatric cancers examined DHS that were specific to Ewing Sarcoma or alveolar rhabdomyosarcoma[48]. They found that these sites were only accessible in the specified tumor type and that these differences in signal could be used to distinguish these tumor types.

1.5 Open question and areas of research

Previous studies have clearly demonstrated the value of cfDNA for non-invasive cancer detection, characterization, and monitoring. Tumor genotyping from cfDNA is used for treatment selection in the clinic. However, many potential applications of cfDNA are still in the early stages of development, especially those characterizing tumor phenotypes. In particular, nucleosome profiling remains an emerging field and there are many areas with unexplored potential. These include studying the impact of GC sequencing biases on nucleosome profiles, characterization of different site selection assays, and investigation of nucleosome profiling in additional cancer types. The research in this thesis explores some of these open areas and addresses unanswered questions in order to bring this promising technique closer to the clinic where it can improve lives.

1.5.1 *Objective 1: Study the impact of GC bias on cfDNA coverage profiles*

The first objective of this research is to study the impact of GC bias on cfDNA coverage profiles and determine whether GC bias impacts the prediction of chromatin accessibility from low coverage WGS. GC bias has long been known to have an impact on sequencing depth in WGS[49] and GC bias correction is essential for copy number analysis[28]. This bias is known to be based on the GC content of individual DNA fragments, but most early studies of cfDNA fragmentomics did not consider correcting for it[29,40]. More recent studies have begun to incorporate GC bias correction into their pipelines including the LUCAS study which performed GC bias correction on fragments before computing short to long fragment ratios[36]. However, they used a custom GC bias correction pipeline that matched the GC bias to a target bias distribution, rather than using established methods, and did not perform GC correction prior to nucleosome profiling around ASCL1 sites[36]. The paper developing LIQUORICE also built a GC bias correction tool and applied it for nucleosome profiling[48], but it only performed a correction in bins, which has been shown to be less accurate than individual fragment length correction[49]. Additionally, neither study systematically evaluated the effects of GC bias on nucleosome profiling. It would be valuable to

implement established fragment-wise GC bias correction[49] and evaluate the impacts of this correction on nucleosome profiles. This is especially important for TFBSs where differences in GC content between the binding site and flanking regions may create GC bias related changes in coverage that obscure accessibility signals[50]. Thus, we hypothesize that GC sequencing biases are obscuring true differences in nucleosome protection and that removing these GC biases will enable us to better distinguish differences in chromatin accessibility from cfDNA. This research is discussed in Chapter 2.

1.5.2 *Objective 2: Characterize sites from various site selection assays*

The second objective of this research is to characterize sites from different selection assays and determine which selection assays are best for specific cfDNA nucleosome profiling applications. Existing work has shown that TFBSs identified with ChIP-seq and DHS can both be used to find tissue specific accessible sites for nucleosome profiling[45,47,48]. However, other chromatin accessibility assays such assay for transposase-accessible chromatin using sequencing (ATAC-seq) could also produce high quality differential sites. More studies exploring and comparing these different site types in additional cancer types will contribute information on the benefits and drawbacks of each assay type and help identify the best site selection assays for future nucleosome profiling studies. We hypothesize that differential sites identified using ATAC-seq will have differential accessibility in cfDNA and that these sites can be used for subtyping. This is research is discussed in Chapter 2.

1.5.3 *Objective 3: Apply nucleosome profiling to new types of cancer*

The third objective of this thesis is to apply nucleosome profiling to new types of cancer, in order to expand the range of patients who might benefit from this technology. While nucleosome profiling has been shown in many types of cancer, in-depth studies have been conducted in only a few of these types. However, there are many additional cancer types with known transcription factor driven subtypes which could be characterized from analysis of transcription factor binding sites in cfDNA. Such cancer types include breast cancer, where estrogen receptor (ER) driven tumors have a different treatment and

prognosis than tumors that don't express this transcription factor[44], and small cell lung cancer (SCLC), where recent studies have begun to characterize four or more transcriptional subtypes[51]. We hypothesize that differences in chromatin accessibility between ER positive and ER negative breast cancer can be observed in cfDNA and that these differences will allow us to develop methods to determine ER status from cfDNA. The development of this method is discussed in Chapter 2. Additionally, we hypothesize that differences in transcription factor activity and gene expression between SCLC subtypes will enable us to find differences in chromatin accessibility that can be used to characterize SCLC subtypes from cfDNA. This research is addressed in Chapter 3.

Chapter 2. A framework for clinical cancer subtyping from nucleosome profiling of cell-free DNA

2.1 Abstract

Cell-free DNA (cfDNA) has the potential to inform tumor subtype classification and help guide clinical precision oncology. Here we develop Griffin, a framework for profiling nucleosome protection and accessibility from cfDNA to study the phenotype of tumors using as low as 0.1x coverage whole genome sequencing data. Griffin employs a GC correction procedure tailored to variable cfDNA fragment sizes, which generates a better representation of chromatin accessibility and improves the accuracy of cancer detection and tumor subtype classification. We demonstrate estrogen receptor subtyping from cfDNA in metastatic breast cancer. We predict estrogen receptor subtype in 139 patients with at least 5% detectable circulating tumor DNA with an area under the receive operator characteristic curve (AUC) of 0.89, and validate performance in independent cohorts (AUC=0.96). In summary, Griffin is a framework for accurate tumor subtyping and can be generalizable to other cancer types for precision oncology applications.

2.2 Introduction

Accurate cancer diagnosis and subtype classification are critical for guiding clinical care and precision oncology. Current approaches to determine tumor subtype require a tissue biopsy, which is often difficult to obtain from patients with metastatic cancer. Therefore, at the time of recurrence or metastatic cancer diagnosis, treatment options may often be informed by clinical diagnostics from the primary tumor. However, molecular changes in the tumor can emerge during metastatic progression and in the context of therapeutic resistance. Moreover, surveying molecular changes is challenging because repeated biopsies are problematic and not routine in clinical practice for solid tumors.

Cell-free DNA (cfDNA) is DNA released into circulation by cells during apoptosis and necrosis.[52] In patients with cancer, a portion of this cfDNA is released from tumor cells, called circulating tumor DNA (ctDNA). The analysis of ctDNA can address the challenges in tissue accessibility and has demonstrated great potential for clinical utility.[53–60] Much of the current research and clinical efforts have focused on the detection of genetic alterations in ctDNA. Shallow coverage sequencing of cfDNA, including ultra-low pass whole genome sequencing (ULP-WGS, 0.1x), provides a cost-effective and scalable solution for estimating the tumor fraction (fraction of the cfDNA that is tumor derived) from the analysis of genomic copy number alterations.[61–64] Sequencing analysis of genomic alterations from ctDNA have helped to distinguish molecular subsets of tumors.[65,66] However, these genomic alterations, including somatic mutations, may not always fully explain treatment failure or identify therapeutic targets, exemplifying a major limitation of cancer precision medicine.

Tumor subtypes are often characterized by distinct transcriptional regulation, which can change during treatment resistance, leading to different clinical tumor phenotypes. For example, prostate and lung cancers may undergo trans-differentiation from adenocarcinoma to small-cell neuroendocrine phenotypes.[67–71] For metastatic breast cancer (MBC), treatment is guided based on clinical subtypes determined by the expression of the estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2), often in the primary tumor[72]; endocrine therapies are prescribed to patients with ER-positive (ER+) or PR-positive (PR+) carcinomas while patients with HER2 positive tumors are prescribed anti-HER2 drugs. Patients with tumors absent for expression of all three receptors have triple negative breast cancer (TNBC) and receive chemotherapy.[73] However, receptor conversions during primary and metastatic disease progression have been frequently observed, including ~20% of patient tumors switching from ER+ to ER-negative (ER-) subtypes.[74–79] Furthermore, similar to the presence of intra-tumor genomic heterogeneity in breast cancer, mixtures of clinical subtypes may also co-exist across or within metastatic lesions in the same patient, presenting major clinical challenges.[80,81] Therefore, accurate subtype classification and identification of transcriptional patterns underlying emergent clinical

phenotype during therapy has critical implications for studying mechanisms of resistance and informing treatment decisions.

Recent studies have shown that the computational analysis of cfDNA fragmentation patterns from genome sequencing data can reveal the occupancy of nucleosomes in cells-of-origin.[82–87] When DNA is released into the peripheral blood following cell death, they are protected from degradation by nucleosomes.[52] At accessible genomic locations, such as at actively bound transcription factor binding sites (TFBSs) and open chromatin regions, nucleosomes are positioned in an organized manner that allows access for DNA binding proteins[88] (Fig. 2.1a). This nucleosome organization results in a loss of sequencing coverage, reflecting DNA degradation at the unprotected binding site with peaks of coverage at the surrounding protected locations.

Analysis of the protected and unprotected regions, termed nucleosome profiling, has been demonstrated for cancer detection and tumor tissue-of-origin prediction, including the analysis of shorter cfDNA fragments which tend to be enriched from tumor cells.[89–94] Tumor subtyping from cfDNA has been explored in castration-resistant prostate cancer (CRPC) and lung cancer by analyzing fragmentation patterns.[95,96] However, subtype classification from cfDNA has been less studied in other cancer types. Specifically, to our knowledge, predicting receptor-based subtypes from cfDNA in breast cancer has not been shown. Furthermore, current cfDNA nucleosome profiling approaches have not been optimized for ULP-WGS data. Studying the clinical phenotype of tumors from ctDNA remains challenging due to lack of robust computational methods but has obvious potential clinical benefits for guiding treatment decisions in patients with metastatic cancer.

In this present study, we develop a computational framework called Griffin to classify tumor subtypes from nucleosome profiling of cfDNA. Griffin overcomes current analytical challenges to profile the nucleosome accessibility and transcriptional regulation from the analysis of standard cfDNA genome sequencing, including ULP-WGS (0.1x) coverage. Griffin employs a GC correction procedure that is specific for DNA fragment sizes and therefore uniquely suited for cfDNA sequencing data. We apply Griffin to perform cancer detection with high performance. Then, we demonstrate breast cancer ER

subtyping from cfDNA, showing high classification accuracy and insights into tumor monitoring and heterogeneity, all achieved from analysis of ULP-WGS data. Overall, Griffin is a generalizable framework that can accurately profile chromatin accessibility from cfDNA for cancer subtype prediction and has the potential to direct personalized treatment to improve patient outcomes.

2.3 Results

2.3.1 *Griffin framework for nucleosome profiling to predict tumor phenotype*

We developed Griffin as an analysis framework with a GC correction procedure to accurately profile nucleosome occupancy from cfDNA. Griffin processes fragment coverage to distinguish accessible and inaccessible features of nucleosome protection (Fig. 2.1a). Griffin is designed to be applied to whole genome sequencing (WGS) data of cfDNA from patients with cancer to quantify nucleosome protection around sites of interest and is optimized to work for ULP-WGS data (Fig. 2.1b). Sites of interest can be selected from various chromatin-based assays, such as from assay for transposase-accessible chromatin using sequencing (ATAC-seq) and are tailored to address specific problems including cancer detection and tumor subtyping.

The analysis workflow begins with computing the genome-wide fragment-based GC bias for each sample. Then, for the region at each individual site of interest, the fragment midpoint coverage is computed and reweighted to remove GC biases (Methods). Midpoint coverage rather than full fragment coverage is used because it produces higher amplitude nucleosome protection signals (Appendix: Supplementary Fig. 1a). Next, a composite coverage profile is computed as the mean of the GC-corrected coverage across the set of sites differential for a tissue type, tumor type, transcription factor (TF), or any phenotypic comparison of interest. By examining these coverage profiles around known cancer-specific and blood-specific TFs, we identified three quantitative features that distinguish a site as accessible and inaccessible: (a) the coverage in the window between ± 30 bp (central coverage), where lower values represent increased accessibility, (b) the coverage in a window between ± 1000 bp (mean coverage), and (c) the overall

nucleosome peak amplitude calculated using Fast Fourier transform (amplitude). These features can be used to quantify transcription factor activity or chromatin accessibility and be used as features for detection of cancer, tumor subtyping, or studying other phenotypes of interest.

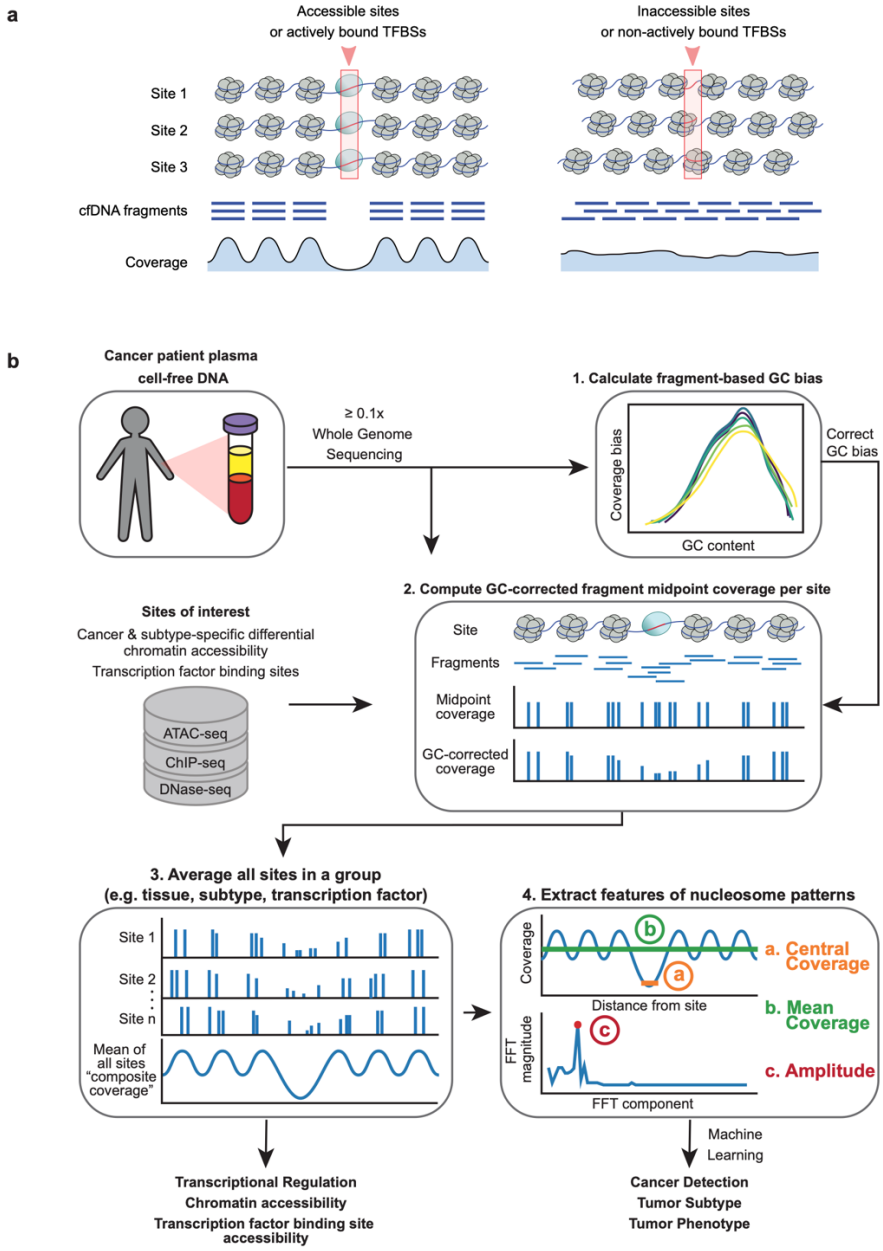


Figure 2.1 Griffin framework for cfDNA nucleosome profiling to predict cancer subtypes and tumor phenotypes

Fig. 2.1 Griffin framework for cfDNA nucleosome profiling to predict cancer subtypes and tumor phenotypes. (a) Illustration of a group of accessible sites (left panel) and inaccessible sites (right panel), such as a TFBS. The nucleosomes (in grey) are positioned in an organized manner around the accessible sites (red box; left panel), but not around the inaccessible ones (right panel). These nucleosomes protect the DNA from degradation when it is released into peripheral blood. The protected fragments from the plasma are sequenced and aligned, leading to a coverage profile which reflects the nucleosome protection in the cells of origin. **(b)** Griffin workflow for cfDNA nucleosome profiling analysis. cfDNA whole genome sequencing (WGS) data with ≥ 0.1x coverage is aligned to hg38 genome build. (1) For each sample, fragment-based GC bias is computed for each fragment size. (2) Sites of interest are selected from any assay. Paired-end reads aligned to each site are collected, fragment midpoint coverage is counted and corrected for GC bias to produce a coverage profile. (3) Coverage profiles from all sites in a group (e.g., open chromatin for tumor subtype) are averaged to produce a composite coverage profile. Composite profiles are normalized using the surrounding region (-5 kb to +5 kb). (4) Three features are extracted from the composite coverage profile: central coverage (coverage from -30 bp to +30 bp from the site; orange 'a'), mean coverage (between -1 kb to +1 kb; green 'b'), and amplitude calculated using a Fast-Fourier Transform (FFT) (red 'c').

2.3.2 *Griffin reduces GC biases enabling detection of differential tissue accessibility*

A unique aspect of Griffin is the implementation of a fragment-based GC bias correction developed by Benjamini and Speed and previously demonstrated on genomic DNA[49]. At open chromatin regions, especially at TFBSs, GC-content is non-uniform between the binding site and flanking regions, which leads to GC-related coverage biases (Fig. 2.2a-c, Appendix: Supplementary Fig. 1b,c, Supplementary Data 1).[97] GC bias varies between samples and between different fragment lengths within a sample[49] (Fig. 2.2b), which can have a major impact on nucleosome accessibility prediction (Fig. 2.2c). To correct for this GC bias, for each sample and each fragment length, Griffin computes the global estimated mean fragment coverage ("expected") using a fragment length position model[49] (Methods, Fig. 2.2b). Then, when

calculating coverage around sites of interest, each fragment is assigned a weight based on the expected coverage for its GC content. This correction eliminates unexpected increases (or decreases) in coverage at binding sites, removing technical biases to enhance the tissue-associated accessibility when analyzing WGS (9-25x, Fig. 2c) cancer patient cfDNA and ULP-WGS (0.1-0.3x, Fig. 2d).

To test the performance of nucleosome profiling following Griffin GC-bias correction, we compared the estimated TFBSs accessibility with the amount of tumor-derived DNA (i.e. tumor fraction) predicted by ichorCNA. From analysis of WGS data for 14 CRPC, two MBC, and two healthy donor samples[61,66], we observed stronger correlations between nucleosome profiles derived from shorter (35-100 bp) fragments and tumor fraction when using GC correction for multiple fragment lengths, which lead us to choose this correction strategy (Appendix: Supplementary Fig. 2, Supplementary Data 2). However, in ULP-WGS data from 191 MBC cfDNA samples[61] with ≥ 0.1 tumor fraction, we focused on the nucleosome sized fragments (100-200bp) due to the low number of short fragments (<100 bp). For nucleosome sized fragments, we expected the tumor fraction to be negatively corrected with the central coverage around tumor-specific sites, and positively correlated for blood-specific sites. For a blood-specific TF, LYL1, we observed that the central coverage at TFBSs was positively correlated with tumor fraction before GC correction (Pearson's r=0.41) as expected, but this correlation was much stronger after GC correction (Pearson's r=0.63, Fig. 2.2e). For a tumor-specific TF, GRHL2, we observed a negative correlation between the central coverage and tumor fraction, as expected (Pearson's r=-0.62, Appendix: Supplementary Fig. 3a). The mean coverage and amplitude features are also correlated to tumor fraction but appeared to be less influenced by GC bias (Appendix: Supplementary Fig. 3a,b, Supplementary Data 3). Similar correlations between nucleosome profile features and tumor fraction following GC correction were also observed for blood and cancer specific DNase I hypersensitivity sites (DHSs) (Appendix: Supplementary Fig. 3a).

To quantify whether GC correction reduces signal variability between samples, we examined the central coverage in the 191 MBC cfDNA ULP-WGS samples for 377 TFs in the Gene Transcription Regulation Database (GTRD).[95,98] For each factor, we compared the variability between the central coverage and tumor fraction using the root mean squared error (RMSE) from a linear regression fit before

and after GC correction. For LYL1, the RMSE decreased (0.062 to 0.046), indicating less inter-sample variation in the data after GC correction (Fig. 2e). Similarly, for 351 (93.1%) TFs, the RMSE was decreased after GC correction, indicating reduced inter-sample variability after accounting for the correlation between tumor fraction and central coverage (two-sided Wilcoxon signed rank test $p = 1.0x10^{-58}$, test statistic = 1421, Fig. 2f, Appendix: Supplementary Fig. 1d, Supplementary Data 3). Next, in the cfDNA samples, we systematically analyzed differentially expressed TFs between blood cells and breast cancer (Methods, Supplementary Data 4). We found that central coverage and tumor fraction were correlated for a subset of these TFs (11 of 35 cancer and 15 of 22 blood TFs, adjusted p-value < 0.05), most correlations were in the expected direction, and that these correlations increased for blood TFs after GC correction (Appendix: Supplementary Fig. 4a).

Additionally, we examined the central coverage for the 377 TFs in a cohort of 215 healthy donors[89] before and after GC correction. Because healthy donor samples have no tumor content, we evaluated the mean absolute deviation (MAD) for each TF to compare inter-sample variability. We found that the MAD decreased after GC correction for 365 (96.8%) TFs (two-sided Wilcoxon signed rank test $p = 6.28x10^{-62}$, test-statistic = 466, Fig. 2g, Appendix: Supplementary Fig. 3c, Supplementary Data 5), indicating lower inter-sample variability for nearly all TFs. Finally, we tested the impact of mappability biases and copy number alterations (CNA) and found that explicit correction accounting for these factors did not improve RMSE values in the MBC cfDNA samples (Methods, Appendix: Supplementary Fig. 4b-f, Supplementary Data 3). Altogether, these results suggest that the GC correction strategy in the Griffin framework reduces the variability in chromatin accessibility signals due to GC biases between samples and allows for improved detection of differential tissue accessibility in ULP-WGS data.

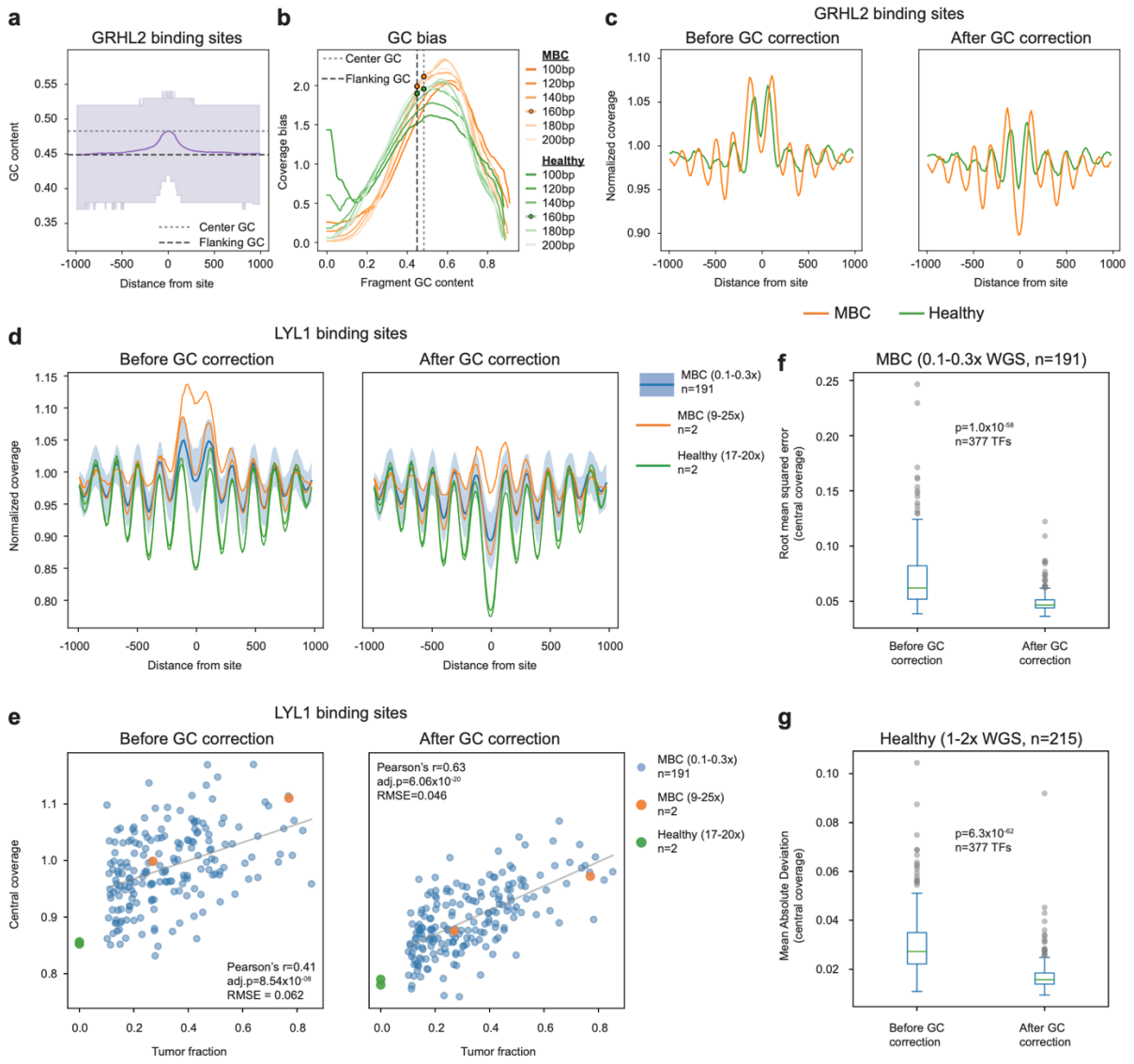


Figure 2.2 Griffin GC bias correction improves detection of tissue specific accessibility from cfDNA.

Fig. 2.2 Griffin GC bias correction improves detection of tissue specific accessibility from cfDNA. (a) Mean ± IQR of GC content around 10,000 GRHL2 sites. **(b)** GC bias of various fragment sizes for cfDNA from a healthy donor (HD_46; green) and a metastatic breast cancer (MBC_315; orange) sample. GRHL2 center and flanking GC content are noted with dashed lines (same as [a]). The MBC sample (orange dots) has a larger difference between center (2.11) and flanking (1.99) for 165bp fragments than the healthy sample (1.90 center, 1.96 flanking; green dots). This means that, for GRHL2, GC bias will cause increased

central coverage relative to the flanking coverage and this effect will be more pronounced in the MBC sample. **(c)** Composite coverage profile of 10,000 GRHL2 sites before and after GC correction, shown for HD_46 and MBC_315. Before GC correction, the center has increased coverage due to GC bias. After GC correction, the MBC sample has lower central coverage, which is consistent with increased GRHL2 activity in tumor cells. **(d)** Composite coverage profiles of 10,000 LYL1 sites before and after GC correction, shown for two MBC samples with deep WGS (9-25x, orange), two healthy samples (17-20x, green), and 191 MBC samples with ULP-WGS (0.1-0.3x, median ± IQR, blue). Lower central coverage in the healthy samples is consistent with LYL1 activity in hematopoiesis. **(e)** cfDNA tumor fraction and central coverage correlation for LYL1. GC correction increases the strength of the Pearson correlation (n=191 MBC ULP-WGS samples; 2 sided with Benjamini-Hochberg FDR correction). Root mean squared error (RMSE) of the linear fit is shown. **(f)** Distribution of the RMSE (linear fit between central coverage and tumor fraction (n=191 MBC ULP-WGS samples) across 377 TFs, before and after GC correction. Boxed range: median ± IQR, whiskers: non-outlier data (maximum extent is 1.5x IQR), grey dots: outliers. p-value from the Wilcoxon signed-rank test (two-sided). **(g)** Distribution of the mean absolute deviation (of the central coverage across 215 healthy donors [1-2x WGS]) for 377 TFs, before and after GC correction. Box elements are the same as (f). p-value from the Wilcoxon signed-rank test (two-sided).

2.3.3 *Griffin analysis at TFBSs enables cancer detection*

To determine if Griffin can perform cancer detection, we analyzed a published WGS (1-2X) dataset of cfDNA samples from healthy donors (n = 215) and early-stage cancer patients (n = 208) (DELFI cohort).[89] We generated nucleosome profiles around the top TFBSs for each TF and extracted three features from each (central coverage, mean coverage, and amplitude). Due to the large number of features, we used principal components analysis (PCA) to select the top components that explained 80% of the variance (Methods). Using logistic regression on these components, we determined that the best performance was achieved when using the top 30,000 TFBSs for each of 270 TFs that contained at least this many sites (Methods, Appendix: Supplementary Fig. 5a). We achieved a high performance for predicting the presence

of cancer with an area under the receiver operating curve (AUC) of 0.94 (Fig. 3a, Supplementary Data 6) We observed the highest performance for stage IV cancers (AUC=0.99) and moderately lower performance in stage I cancers (AUC=0.93, Fig. 3a, Appendix: Supplementary Fig. 5b). The performance was likely reflective of the higher tumor fractions observed in late-stage cancer relative to early-stage cancer. As anticipated, we observed higher performance for samples with tumor fraction ³ 0.05 (AUC=0.99) than samples with < 0.03 tumor fraction (AUC=0.92, Appendix: Supplementary Fig. 6a). By cancer type, we achieved the highest performance for lung and ovarian cancers (AUC ≥ 0.99) and the lowest for pancreatic cancer (AUC=0.85, Appendix: Supplementary Fig. 6b). To test the ability to detect cancer at ULP-WGS coverage (0.1x), we applied Griffin to the same cfDNA data downsampled to 0.1x coverage and achieved an AUC of 0.89 (Fig. 3a, Appendix: Supplementary Fig. 6a,b).

Next, we systematically evaluated various configurations and comparisons of Griffin for cancer detection (Appendix: Supplementary Fig. 7a). First, because fragments <150bp are enriched for tumor derived DNA[89], we tested whether different fragment size ranges, such as short (35-150bp) or all (35-500 bp) fragments may improve our ability to detect cancer in this framework but observed a decreased performance (0.91 and 0.92 AUC, respectively, Appendix: Supplementary Fig. 7a). Next, when omitting GC correction, we also observed decreased overall performance for 1-2X WGS (AUC=0.83, Fig. 3b, Appendix: Supplementary Fig. 7a) and ULP-WGS (AUC=0.85) for all disease stages (Fig. 3b). Then, we tested the use of mappability and copy number correction, exclusion of Griffin features, and analysis at DHSs in place of TFBSs and observed similar or lower performance (Appendix: Supplementary Fig. 7a). Finally, we compared our results with the method by Ulz et al.[95], which analyzed cfDNA fragments of all lengths at TFBSs, and found it had lower performance for 1-2X WGS (AUC=0.82) and ULP-WGS (AUC=0.55) coverages. (Appendix: Supplementary Fig. 7a,b).

To validate Griffin for the application of cancer detection, we analyzed a published cfDNA WGS (1-2X) dataset consisting of 129 lung cancer patients and 158 healthy individuals (LUCAS cohort).[96] A validation cohort of 46 cancer patients and 385 healthy individuals was also available in this same study. There was a notable batch effect between the DELFI and LUCAS cohorts in the initial fragment size

distributions and Griffin coverage profiles before and after GC correction, which prevented use of the same model on both cohorts (Methods, Appendix: Supplementary Fig. 8, Supplementary Data 7). Using the 270 TFs in the Griffin analysis, we built a new model and observed an AUC of 0.76 in 1-2X WGS and 0.65 for ULP-WGS (downsampled to 0.1X) coverages in the LUCAS cohort (Fig. 3c, Appendix: Supplementary Fig. 5c, Supplementary Data 8). We observed an AUC of 0.91 for samples with ≥ 0.05 tumor fraction, which was higher than samples with 0.03-0.05 (AUC=0.78) and < 0.03 (AUC=0.65) tumor fractions (Appendix: Supplementary Fig. 6c). Applying the trained model from the LUCAS cohort to the LUCAS validation cohort, we achieved an AUC of 0.86 across all stages, including an AUC of 0.83 for stage I cancers (Fig. 3d, Appendix: Supplementary Fig. 5d, Supplementary Data 9). The performance was 0.87 and 0.81 AUC for tumor fractions of < 0.03 and ≥ 0.03, respectively (Appendix: Supplementary Fig. 6d). For ULP-WGS coverage, the performance was 0.69 AUC for stage I and 0.69 AUC across all stages (Fig 3d, Appendix: Supplementary Fig. 5d). Overall, while cancer detection has been demonstrated from nucleosome profiling analysis in ctDNA[89,94–96], we show that Griffin may also be applied in this setting.

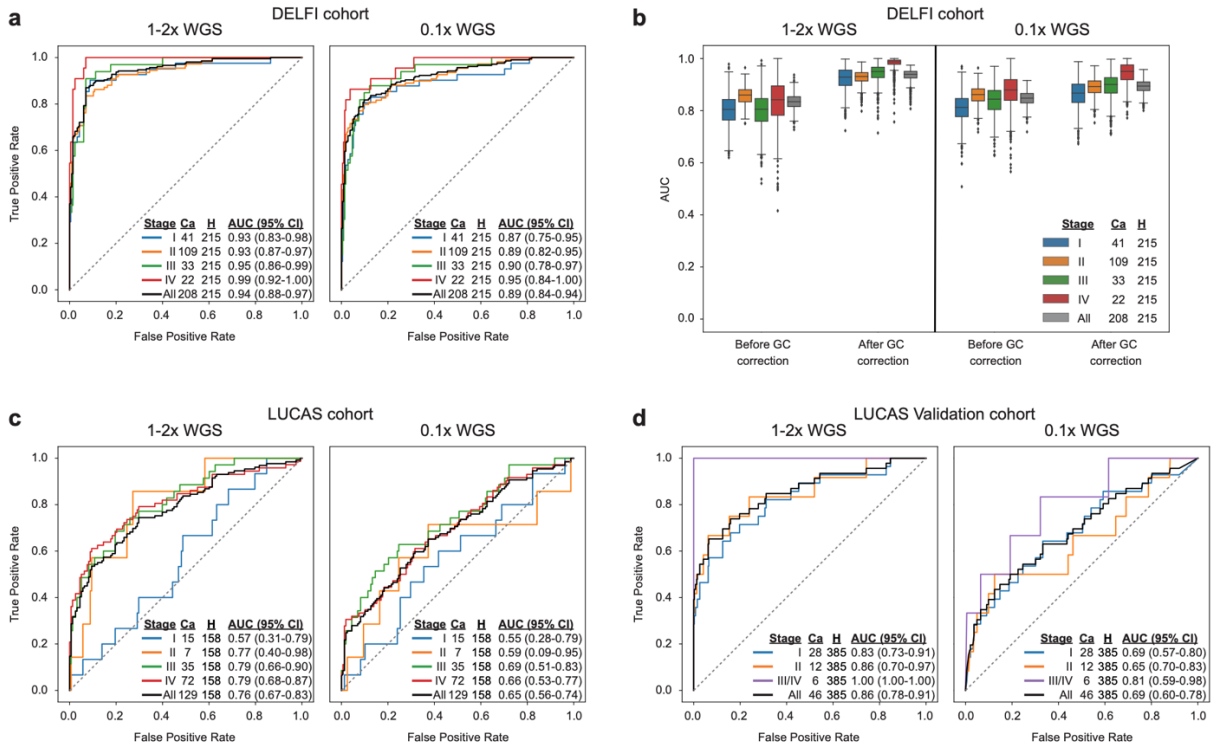


Figure 2.3 Griffin enables accurate cancer detection

Fig. 2.3 Griffin enables accurate cancer detection. Receiver operator characteristic (ROC) curves for logistic regression classification of cancer vs. healthy controls in three cohorts. Logistic regression was performed on the top PCA components which explained 80% of the variance in the features (central coverage, mean coverage, and amplitude) extracted from nucleosome profiles around 30,000 TFBS for each of 270 TFs. ROC and area under the ROC curve (AUC) performance is shown for each disease stage. The number of cancer samples (Ca) is indicated for each stage. Each ROC curve also includes all healthy controls (H) from that cohort. 95% confidence intervals (CI) were obtained from 1,000 bootstrap iterations. **(a)** Performance for DELFI cohort[89] consisting of plasma samples for 208 early-stage cancers and 215 healthy controls. **(b)** Comparison of the performance in the DELFI cohort before and after GC correction using Griffin. Samples are the same as in (a). Boxplots indicate median, interquartile range (IQR), whiskers for 1.5 x IQR, and outliers. **(c)** Performance of the LUCAS cohort[96] consisting of plasma from 129 lung cancer patients and 158 healthy patients. **(d)** Performance of the LUCAS validation cohort[96] consisting of

plasma for 46 lung cancers and 385 healthy controls. For each dataset, performance is shown for both the original low pass (1-2x) WGS and ultra-low pass (0.1x) WGS generated by in-silico downsampling.

2.3.4 *Griffin enables accurate prediction of breast cancer subtypes from ultra-low pass WGS*

Breast cancer tumor classification relies on accurate clinical determination of hormone receptor status primarily by immunohistochemistry (IHC) to quantify the expression of ER, but no ctDNA approach exists for this application. We set out to determine whether Griffin can be used to predict ER subtype status from ULP-WGS (0.1x) of cfDNA from MBC patients. We analyzed 254 samples with tumor fraction greater than 0.05 from 139 patients.[61,62] First, we inspected the Griffin profiles at TFBSs for key factors, including ESR1, FOXA1, and GATA3, which are known to be associated with ER positive tumors.[99] We observed that these TFBSs were more accessible in cfDNA samples from patients with ER+ metastases compared to ER-; central coverage was significantly lower in ER+ samples after accounting for tumor fraction (ANCOVA FDR adjusted p-value < 3.8×10^{-2}, Appendix: Supplementary Fig. 9, Supplementary Data 10). To predict ER status, we initially built a logistic regression classifier using features from the Griffin profiles for all 270 TFs and achieved an accuracy of 0.71 (AUC of 0.79, Appendix: Supplementary Fig. 10). We also used TFBSs features computed by the Ulz method for ER subtyping and observed an accuracy of 0.53 (AUC=0.55, Appendix: Supplementary Fig. 10), likely because it was not designed for ULP-WGS data.

Next, we used a more tailored site selection approach by analyzing regions of differential chromatin accessibility. Using ATAC-seq data generated from 44 ER+ and 15 ER- primary breast tumors by The Cancer Genome Atlas (TCGA)[100], we identified open chromatin sites that were differentially accessible between ER subtype (Methods, Fig. 4a, Appendix: Supplementary Fig. 11, Supplementary Data 11-12). ER+ sites (n=28,170) were enriched for the TFBSs of ESR1, PGR, FOXA1 and GATA3, and ER- sites (n=41,712) were enriched for the TFBSs of STAT3 and NFKB1 (Supplementary Data 13). We observed differences in coverage profiles between differential sites that were shared (9,930 ER+, 22,365 ER-) and not shared (18,240 ER+, 19,347 ER-) with accessible chromatin in hematopoietic cells[101] and analyzed

them separately (Fig. 4b, Appendix: Supplementary Fig. 12). We applied Griffin to profile nucleosome accessibility at these four sets of ER differential accessible chromatin sites, extracting a total of 12 features. We built a logistic regression classifier to predict ER subtype from these chromatin accessibility features (Fig. 4c, Supplementary Data 14, Methods). We achieved an overall accuracy of 0.81 (AUC=0.89, n=139) with a higher performance for samples having high tumor fraction (accuracy 0.86, AUC=0.92, n=101, tumor fraction ³ 0.1) compared to those with lower tumor fraction (accuracy 0.69, AUC=0.75, n=38, tumor fraction 0.05 to 0.1) (Fig. 4d). Systematic evaluation of different configurations and comparisons with Griffin, including fragment size ranges and data correction strategies, resulted in similar or lower performance (Appendix: Supplementary Fig. 10, Methods).

We validated the trained model from the MBC dataset by evaluating its performance on independent cohorts consisting of additional ULP-WGS data or data obtained from published studies[102,103] (Methods). Using PCA, we did not observe batch effects between the cohorts, but rather signals could be attributed to the known ER status (by metastatic tumor IHC) and estimated tumor fraction (Appendix: Supplementary Fig. 13a). In 36 patients (25 ER+, 11 ER-) with tumor fraction ≥ 0.05, we observed an overall accuracy of 0.92 (AUC=0.96), including 0.96 accuracy (AUC=0.98) for samples with higher tumor fraction (≥ 0.1, n=24) and 0.85 accuracy (AUC=0.90) for lower tumor fraction (0.05 – 0.1, n=12) (Fig. 4e, Appendix: Supplementary Fig. 13b,c, Supplementary Data 15). For samples with tumor fraction < 0.05 (n=35), the accuracy was 0.54 (AUC=0.39), indicating the lower limit of accurate ER classification is likely 0.05 tumor fraction (Appendix: Supplementary Fig. 13). These results illustrate the utility of using chromatin accessibility for cancer subtyping from ULP-WGS data and showcase ER status prediction in breast cancer from cfDNA.

2.3.5 *Analysis of ER status from longitudinal cfDNA suggests potential subtype heterogeneity*

To further investigate the ER predictions, we inspected the classification results for 91 patients with known primary ER status and cfDNA tumor fraction of ≥ 0.1 (Fig. 4c,f, Supplementary Data 14). In 40 patients who had ER- status for both primary and metastatic tumors determined by IHC, we predicted 39

(95.1%) to have ER- subtype from plasma (Fig. 4f). In 41 patients who had ER+ primary and metastatic tumors, we classified 36 (85.4%) to have ER+ subtype. Intriguingly, in the nine patients who had clinical primary ER+ and metastatic ER- status (i.e., ER loss), five (55.6%) were predicted to be ER+, and this higher prevalence of ER+ prediction was statistically significant when compared to patients with no subtype switches (ER- group, $p = 3.7 \times 10^{-4}$ and ER+ group, $p = 0.043$, two-sided Fisher's exact test, Fig. 4f). We observed a performance of 0.74 AUC for classifying ER status among patients who had ER+ primary tumor status, suggesting Griffin may have some potential to classify patients with ER loss (Fig. 4g). These results demonstrate that Griffin has relatively high performance for ER classification in MBC patients with no subtype switches but ER status prediction is more challenging for patients with subtype switches perhaps due to ER subtype heterogeneity.

To further investigate the ER status predictions and subtype heterogeneity, we examined eight patients who had ULP-WGS of cfDNA from plasma collected at different timepoints and ER expression by IHC available for one or more metastatic biopsies (Fig. 4h, Appendix: Supplementary Fig. 14, Supplementary Data 16).[62,104] As an interesting example, MBC_1413 was initially diagnosed with an ER- metastatic pleural effusion but a second biopsy of the liver metastasis revealed ER expression in 5% of cells. The initial cfDNA sample was collected 178 days after and was predicted to have ER+ status (0.74 probability), in agreement with the metastatic liver biopsy. A third biopsy from the pleural fluid was ER-, which was consistent with the ER- prediction (0.23 probability) from a cfDNA sample taken 26 days prior. In another example, MBC_1009 had two ER- biopsies of the bone and liver, but a third biopsy had 5% ER expression, which was consistent with ER+ predictions (> 0.68 probability) for cfDNA samples taken 251 days before and 52 days after. These results suggest that Griffin may be detecting ER status changes or heterogeneity of tumor biopsies and that that subtype monitoring during therapy may be a potential application.

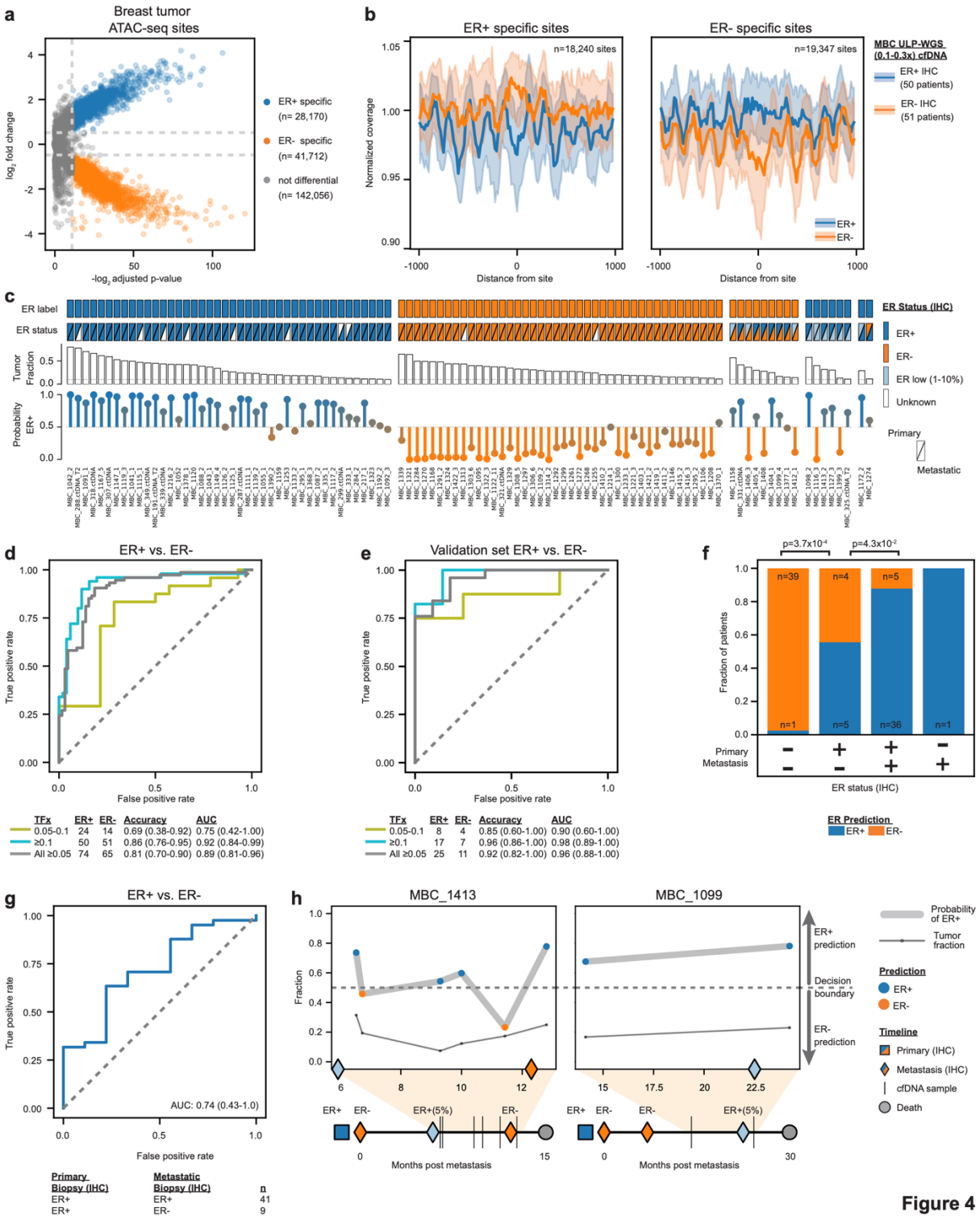


Figure 4

Figure 2.4 Griffin enables accurate prediction of breast cancer estrogen receptor subtypes from ultra-low pass WGS

Fig. 2.4 Griffin enables accurate prediction of breast cancer estrogen receptor subtypes from ultra-low pass WGS. (a) ER+ and ER- open chromatin sites from assay for transposase-accessible chromatin using sequencing (ATAC-seq) in ER+ (n=44) and ER- (n=15) breast tumors from The Cancer Genome Atlas (TCGA).[100] Differential sites were identified using DESeq2[105] which employs a Wald test with Benjamini-Hochberg FDR correction. Sites with an adjusted p-value $<5x10^{-4}$ and a $\log_2$ fold-change >0.5 or <-0.5 (dashed lines) were considered differential and are shown in blue (ER+) or orange (ER-). **(b)** Composite coverage profiles (median ± IQR) for ER+ (n=18,240) and ER- (n=19,347) sites in MBC patients (≥0.1 tumor fraction; ER+, n=50; ER-, n=51). Differential sites shared with hematopoietic cells have been excluded and are shown in Appendix: Supplementary Fig. 12a.[101] **(c)** Tumor and cfDNA characteristics for 101 MBC patients with ≥0.10 tumor fraction. Statuses are from immunohistochemistry on tumor tissue. Top row: Binary ER status used for training and testing the model. Second row: primary (upper left triangle) and metastatic (lower right triangle) ER status. Third row: tumor fraction from ichorCNA[61]. Fourth row: median probability ER+ predicted across 1,000 bootstrap iterations. **(d)** Receiver operator characteristic (ROC) curve for predicting ER status. 95% CIs from 1000 bootstrap iterations. **(e)** Performance of the trained model on samples from three validation cohorts. **(f)** Predictions in patients grouped by primary and metastatic ER status. P-values from Fisher's exact test (two-sided). **(g)** ROC curve for predicting ER loss among patients with a primary ER positive tumor. **(h)** Timelines for two patients with multiple biopsies and cfDNA samples. Top: predicted probability of ER+ and tumor fraction for cfDNA samples with ≥0.05 tumor fraction and ≥0.1x coverage. Bottom: timeline in months from metastatic diagnosis. The square indicates primary ER status (timeline from primary to metastatic diagnosis is not to scale). Diamonds indicate each metastatic ER status. Patient MBC_1413 had 3 metastatic biopsies, ER- at zero months (pleural fluid), weak ER+ (5%) at 5.9 months (liver), and ER- at 12.3 months (pleural fluid). Patient MBC_1099 had 3 metastatic biopsies, ER- at 0 months (bone), ER- at 7 months (liver), and ER low (5%) at 22.5 months (liver).

2.4 Discussion

In this study, we described the development of Griffin, a framework and analysis tool for studying transcriptional regulation and tumor phenotypes. Griffin applies a fragment length specific GC-correction procedure to remove the GC biases that obscure chromatin accessibility signals in cfDNA. We demonstrated that Griffin can be used to detect cancer from low pass WGS with high accuracy. We also developed an approach to perform ER subtyping in breast cancer from ULP-WGS of ctDNA.

Griffin is versatile and can be used for various applications in cancer. We highlighted cancer detection and tumor subtype use-cases. However, Griffin can also be used for any biological comparison where transcriptional regulation and chromatin accessibility differences can be delineated. The applications described here use TFBSs from chromatin immunoprecipitation sequencing (ChIP-seq) and accessible chromatin sites from ATAC-seq. However, Griffin differs from existing frameworks due to its ability to analyze custom sites of interest that are specific to any biological context. These sites may be obtained from external sources and different assays, such as ChIP-seq, DHS, ATAC-seq or cleavage under targets and release using nuclease (CUT&RUN). As additional epigenetic data are collected by the cancer research community, including from single-cell experiments[106,107], Griffin will be integral for advancing tumor phenotype studies from liquid biopsies.

Griffin is designed for the analysis of ULP-WGS (0.1x) of cfDNA, while other nucleosome profiling methods have focused on deeper coverage sequencing. Griffin takes advantage of analyzing the breadth of sites as opposed to individual loci, which was inspired by a similar strategy used by Ulz et al[95]. We showed that Griffin had better performance for both detecting cancer and predicting ER status from ULP-WGS data when compared to the Ulz method, likely because of its GC bias correction strategy and versatility to analyze any set of genomic regions. We observed improved performance after GC-correction consistently for all analyses, suggesting the benefit of the approach, although this improvement was minor for ER status prediction in ULP-WGS data. While the GC correction strategy was able to reduce inter-sample variability, we found that it was not able to eliminate batch effects between datasets potentially

caused by different cfDNA processing and sequencing workflows, thus preventing cancer detection models from being compatible across all datasets. However, Griffin provides a framework to extract cfDNA features, enabling users to train models on new datasets, as we showed with the LUCAS and validation cohorts. Griffin can be applied to future large prospective studies using standardized plasma collection and workflows to carefully assess the performance of cancer detection in real clinical scenarios.

Although this study focused on the analysis of ULP-WGS (0.1x) of cfDNA, Griffin is not limited to low coverage data. Increased cfDNA sequencing coverage can allow for analysis of specific gene promoters and cis-regulatory elements and may enable gene expression prediction.[82] While recent studies show the promise of cfDNA methylation and cfRNA analysis for tumor phenotype analysis and cancer detection,[108–114] these analytes may be challenging to isolate from clinical specimens or require specialized assays. Overall, Griffin provides a cost-effective and scalable framework requiring only standard low coverage WGS of cfDNA, which can be more rapidly incorporated into existing platforms to predict clinical cancer phenotypes.

A limitation of the binary ER classification (ER+ or ER-) is the decreased accuracy for samples with lower tumor fraction (< 10%) and a 5% limit for accurate prediction, suggesting that it may be challenging to use Griffin for early-stage and minimal residual disease settings. However, in MBC, previous reports have suggested that up to 34% of MBC patients may have at least 10% tumor fraction in plasma[61], which highlights potential utility for this disease stage. TNBC patients with cfDNA tumor fraction ≥ 10% have poorer prognosis[115] and would benefit more from tumor monitoring. It may be possible to improve performance of ER subtyping for lower tumor fraction samples with additional sequencing depth, using TFBSs identified directly from ER+/-tumors, or joint analysis of multiple cfDNA timepoints from the same patient.

The application of Griffin to predict ER status from cfDNA of MBC patients led to interesting results for patients with ER loss, suggesting potential tumor heterogeneity. Intriguingly, we noticed that for the patients with ER- tumors by IHC, ER+ predictions were significantly enriched when the primary tumor was ER+. Moreover, in some patients with multiple cfDNA biopsies we observed changes in predicted ER

status that might be explained by the presence of metastatic tumors with both subtypes. This subtype heterogeneity and switching would typically not be captured from a single metastatic biopsy, but our results demonstrate the possibility of using the predicted ER probability to monitor subtype status over time during therapy using ctDNA. Future studies using synchronous tumor biopsy and plasma sequencing data for more patients will be needed to establish clinical utility.

We focus our breast cancer subtyping on ER prediction because its status has important utility in predicting likely benefit to endocrine therapy.[116] While PR expression is also determined in the clinic and ER-/PR+ tumors are considered hormone receptor positive, these are rare, not reproducible or less useful for prognosis.[117] In our cohort, only 2 out of 139 (1.4%) patients were ER-/PR+. HER2 overexpression is important for prognosis and determining treatment such as with trastuzumab.[118] However, we were unable to identify sufficient number of open chromatin sites that were differentially accessible between HER2 positive and HER2 negative tumors. Since ERBB2 (encodes the HER2 protein) is amplified in ~20% breast cancers, one can instead assess ERBB2 copy number amplification from ctDNA genomic analysis.[103,119] Alternatively, a model to predict PAM50 status could be useful as this may be a better indicator of prognosis than ER/PR/HER2 IHC alone.[120]

In summary, the Griffin framework enables prediction of tumor phenotypes from ULP-WGS. In this study, we demonstrate the use of this framework to detect cancer in early-stage cancer patients and to predict ER status in metastatic breast cancer patients. Combined with methods for predicting tumor fraction and copy number alterations[61] Griffin joins a suite of tools for in depth analysis of ULP-WGS of cfDNA enabling cost effective, non-invasive monitoring of tumors. Griffin has the potential to reveal clinically relevant tumor phenotypes, which will support the study of therapeutic resistance, inform treatment decisions, and accelerate applications in cancer precision medicine.

2.5 Methods

The research described in this study complies with all relevant ethical regulations. New patient data (Independent MBC Cohort) was obtained under protocols which were approved by the institutional review

board of the Dana Farber Cancer Institute (DFCI-09204) or Ohio State University (2007C0066, 2018C0211). Use of additional clinical data for the previously published MBC ULP-WGS cohort was approved by an institutional review board (Dana-Farber Cancer Institute IRB protocol identifiers 05-246, 09-204, 12-431 [NCT01738438; Closure effective date 6/30/2014]). Patients in all studies provided written informed consent for the study in which they were enrolled. See descriptions of human subjects and datasets below.

Griffin: GC-content bias correction procedure

GC content influences the efficiency of amplification and sequencing leading to different expected coverages (coverage bias) for fragments with different GC contents and fragment lengths. This is called GC bias and is unique to each sample. We calculated the GC bias of each bam file using an implementation of the method developed by Benjamini and Speed 2012[49] which was previously implemented in deepTools[121]. However, unlike the deepTools implementation, which assumes that all fragments have the same length, we used the 'fragment length model' which calculates a separate GC bias curve for each fragment length. This is helpful for cfDNA where different samples may have different fragment size distributions and different fragment lengths have biological significance.[83]

Mappability filtering

Prior to performing GC bias calculation, we identified all mappable regions of the genome (as described by Benjamini and Speed and implemented in deepTools) using the Umap multi-read mappability track for 100bp reads downloaded from UCSC genome browser[122] (https://hgdownload.soe.ucsc.edu/gbdb/hg38/hoffmanMappability/k100.Umap.MultiTrackMappability.bw). We used pybedtools (0.8.0)[123] to find the mappable regions (defined as mappability score = 1) and further excluded regions with known mapping problems including the encode unified GRCh38 exclusion list (https://www.encodeproject.org/files/ENCFF356LFX/), centromeres, fix patches, and

alternative haplotypes for hg38 downloaded from UCSC table browser (https://genome.ucsc.edu/cgi-bin/hgTables).

Multi-fragment length GC bias model and correction

We then examined all remaining regions of the genome and, for each fragment length, counted the observed GC content of every possible fragment overlapping those positions. The observed frequencies of each GC content for each fragment length are the 'genome GC frequencies' and are specific to the genome build. We then developed the 'griffin GC bias' pipeline to compute the GC bias in a given bam file. The pipeline takes a bam file, bedGraph file of valid (mappable, non-excluded) regions, and genome GC frequencies for those regions. For each given sample, we fetched all reads aligning to the valid regions on autosomes using pysam v0.15.4 (https://github.com/pysam-developers/pysam)[124]. We counted the number of observed reads for each length and GC content, excluding reads with low mapping quality (<20), duplicates, unpaired reads, and reads that failed quality control. These read counts are the 'GC counts' for that sample. We then divided the GC counts for a sample by the GC frequencies for the genome to obtain the GC bias for that bam file and normalized the mean GC bias for each fragment length to 1, resulting in a GC bias value for every combination of fragment size and GC content (except those combinations that are never observed in the genome). We then smoothed the GC bias curves. For each fragment size we took all GC bias values for fragments of a similar length (+/- 10 bp). We sorted these values by the GC content of the fragment to create a vector of GC bias values for similar sized fragments. We then smoothed this vector by taking the median of k nearest neighbors (where k = 5% of the vector length or 50, whichever is greater) and repeated for each possible fragment length. We then normalized to a mean GC bias of 1 for each possible fragment length (excluding GC contents that are never observed) to generate a smoothed GC bias value for every possible fragment length and GC content observed in the genome.

Griffin: Nucleosome profiling

We designed the griffin nucleosome profiling pipeline to perform nucleosome profiling around sites of interest. This pipeline takes a bam file, GC bias for that bam file, and site list, and assorted other parameters described below. For a given bam file and site list, we fetched all reads in a window (-5000 to +5000bp) around each site using pysam (excluding those that failed quality control measures). We then filtered read pairs by fragment length and selected those in a range of fragment lengths (100-200bp for all analyses in this study unless otherwise specified). For each read pair, we determined the GC bias for the fragment and assigned a weight of $\frac{1}{GC\ bias}$ to that fragment and identified the location of the fragment midpoint. We split the site into 15bp bins and summed the weighted fragment midpoints in each bin to get a GC corrected midpoint coverage profile (see Fig. 1b for a schematic). Next, we excluded bins that overlapped regions with known mapping problems (described in Griffin: GC-content bias correction procedure) and bins with at least one unmappable position using pyBigWig for fetching data (0.3.17). We also identified bins with extremely high coverage (10 standard deviations above the mean) and removed these bins. We repeated this for every site on the site list and took the mean of all sites (ignoring excluded bins within those sites) to generate the mean coverage profile for that site list. We then smoothed the coverage profiles using a Savitzky-Golay filter with window length 165bp and polynomial order of 3. Finally, to make samples with different depths comparable, we normalized the coverage profile to a mean coverage of 1 across the ±5000bp window and retained the central region (+/- 1000 bp) for further analysis.

Griffin: Nucleosome profile feature quantification

To quantify coverage profiles, we extracted 3 features from each coverage profile. First, we calculated the coverage value at the site (± 30bp). Second, we calculated the 'mean coverage' value ±1000 bp from the site. And third, we calculated the amplitude of the nucleosome peaks surrounding the site by using a Fast Fourier Transform (as implemented in Numpy v1.21.2[125]) on the window ±960 bp from the site and taking the amplitude of the 10th frequency term. This window and frequency were chosen due to the observed

nucleosome peak spacing at an active site (190bp) which results in approximately 10 peaks in the window ±960bp.

Mappability correction

To test the impact of mappability bias on Griffin profiles, we implemented a per fragment mappability bias correction. First, for each sample, we obtained an approximate coverage distribution by sampling 1,000 random positions within the genome (excluding positions which overlapped regions with known mapping problems see 'Griffin: GC-content bias correction procedure') and determined the cutoff for extreme outliers >5 standard deviations above the mean. Next, we split the genome into 5Mbp segments resulting in 587 segments spanning the genome (autosomes only). For each segment, we sampled every 100th position, skipping positions with known mapping problems. At sampled positions, we obtained the mappability value from Umap multi-read mappability track for 100bp reads (described in Griffin: GC-content bias correction procedure) and the number of reads overlapping that position (excluding unpaired reads, reads with mapping quality <20, duplicates and reads that failed quality control). Sampled positions with read counts >5 SD above the mean were excluded. After obtaining the mappability values and read counts for all sampled positions, we calculated the mappability bias for each mappability value within that 5Mbp bin by dividing the total number of reads observed at positions with a given mappability by the total number of positions with that mappability value. We repeated this procedure for all bins. Finally, we took the mappability biases for all mappability values in all bins and smoothed them using loess regression as implemented in python statsmodels (version 0.13.2)[126]. When calculating coverage profiles, we calculated the mappability value for each fragment as the mean mappability of all positions covered by the forward and reverse read. We then assigned a weight of $\frac{1}{Mappability\ bias}$ to that fragment and multiplied this by the weight from GC bias $\frac{1}{GC\ bias}$ to get the total fragment weight used when calculating the mappability and GC bias corrected coverage profiles. This correction did not improve performance of any correlations or models and was not used in the final Griffin models.

Copy number alteration (CNA) correction

To assess whether CNA correction improved Griffin performance, we performed CNA correction at each site prior to merging sites into composite coverage profiles. This correction was performed by dividing the coverage at each position in the profile by the mean coverage in the surrounding ±50Kbp window. We found that the addition of CNA correction had a minimal impact on coverage profiles and did not improve the correlations to tumor fraction or performance of the cancer detection model and resulted in only a small improvement in the ER status prediction model. We did not use CNA correction in our final Griffin models, however we did leave an option to turn it on for future users who might find it useful.

Single fragment length GC correction

In order to assess whether to use a single fragment length model or a multiple fragment length model was better able to correct GC biases around accessible sites in cfDNA, we implemented a GC correction model that assumes a single fragment length (165bp) for all fragments similar to the method implemented by deepTools. This model used the same procedure as described in Griffin: GC-content bias correction procedure, with a few modifications. When calculating the GC counts, it assumed that every read had a fragment length of 165bp (starting from the read start position). The resulting GC counts were then divided by the GC frequencies for 165bp to generate the GC biases for each GC possible GC content for 165bp fragments. Next, when generating coverage profiles, we found the GC content of each fragment and then found the GC bias for the 165bp fragment with the most similar GC content and used this value to reweight the fragment. This single fragment length procedure was found to not perform as well for short (35-100bp) fragments (Appendix: Supplementary Fig. 2a-c) and perform similarly for nucleosome sized (100-200bp) fragments (Appendix: Supplementary Fig. 2d-f, Appendix: Supplementary Fig. 7a, Appendix: Supplementary Fig. 10). Consequently, the multi-fragment length model was used for all subsequent analysis.

Early-stage cancer and healthy donor cfDNA samples

DELFI cohort

Whole genome sequencing (WGS) cfDNA from patients with various types of early stage cancer and healthy donors were obtained from an existing dataset published in Cristiano et al[89]. Bam files were downloaded from EGA (dataset ID: EGAD00001005339). This data consisted of 1-2x low pass whole genome sequencing from 100bp paired end Illumina sequencing reads. For our analyses, we used a subset of samples with 1-2X WGS of cfDNA from 208 cancer patients with no previous treatment and 215 healthy donors. These were the same samples used for the cancer detection analysis in the original Cristiano et al. study. cfDNA tumor fraction was estimated using ichorCNA (github commit 15B1D336) [61]. An hg38 panel of normals (PoN) with a 1mb bin size was created using all 215 healthy donors in the dataset. ichorCNA was then run on all cancer and healthy samples to estimate tumor fraction. ichorCNA_fracReadsInChrYForMale was set to 0.001. Defaults were used for all other settings.

LUCAS cohort and LUCAS validation cohort

Whole genome sequencing (WGS) cfDNA from a prospective study of patients with lung cancer and without cancer were obtained from an existing dataset published by Mathios and colleagues[96]. Bam files were downloaded from EGA (dataset ID: EGAD00001007796). This data consisted of 1-2x low pass whole genome sequencing from 100bp paired end Illumina sequencing reads. For our analyses, we used the subset of samples described in the paper as the 'LUCAS' cohort and a second subset of samples described as the LUCAS validation cohort. The LUCAS cohort included 158 patients who had no history of cancer and no future cancer diagnosis and 129 patients who were diagnosed with lung cancer within days of blood draw (0-44 days). The LUCAS validation cohort included 46 patients with cancer and 385 patients without cancer. All samples were realigned to hg38 as described below in 'sequence data processing'. Tumor fraction was determined using ichorCNA (as described for the DELFI cohort) with a new panel of normals constructed using 54 separate healthy donor samples (not included in either the LUCAS or LUCAS validation cohorts) from the LUCAS study.

Metastatic breast cancer (MBC) cfDNA samples

Sequencing data

WGS of cfDNA from patients with metastatic breast cancer (MBC) and healthy donors were obtained from an existing dataset published by Adalsteinsson and colleagues[61]. Bam files were downloaded from dbGaP (accession: phs001417.v1.p1). This data consisted of ~0.1x ultra-low pass whole genome sequencing (ULP-WGS) from 100bp paired end Illumina sequencing reads. For our analyses, we used a subset of 254 ULP samples with >0.1X coverage WGS, >0.05X tumor fraction and known estrogen receptor (ER) status. Of these 254 samples 133 were ER positive (from 74 unique patients) and 121 were ER negative (from 65 unique patients). Coverage and tumor fraction metrics were obtained from the supplementary data in the publication[61]. Additionally, we used two deep (9-25x) WGS from two MBC patients (MBC_315 and MBC_288) from the same source and two deep (17-20x) WGS from two healthy donors (HD45 and HD46) from the same source for designing and demonstrating the pipeline.

Human subjects and clinical data

Primary and metastatic ER status was determined by immunohistochemistry and obtained from pathological review. Metastatic survival time was also abstracted from the medical records. Use of this data was approved by an institutional review board (Dana-Farber Cancer Institute IRB protocol identifiers 05-246, 09-204, 12-431 [NCT01738438; Closure effective date 6/30/2014]).

For training and assessing the ER status classifier we labeled each sample as ER+ or ER- using information about the ER status from medical records. If metastatic ER status was not known, the sample was labeled according to the primary tumor ER status (20 samples from 11 patients). ER low (1-10% ER+ staining) samples (15 samples from 6 patients) were labeled ER+ for the purpose of the binary classifier. For eight patients (MBC_1413, MBC_1405, MBC_1399, MBC_1099, MBC_1408, MBC_331, MBC_1312, and MBC_1404) we had information about multiple metastatic biopsies, some with multiple ER statuses among

the biopsies. In these cases, we used the last biopsy taken prior to the initial cfDNA collection for the purpose of training and testing the binary ER status classifier. In a partially overlapping set of 8 patients, we also had information about multiple primary biopsies, two with multiple ER statues among the primary biopsies. In these cases, we used the first ER status to determine if there had been a subtype switch (see Supplementary Data 16 for details about biopsy statues, locations, and timelines).

Metastatic breast cancer (MBC) validation cohorts

Three independent validation cohorts were used to assess the performance of the ER status prediction model, two of these were from previously published studies. The first cohort was from the study by Ahuno et al.[102], which included WGS of cfDNA from 14 breast cancer (BRCA) patients in Ghana with known ER status and ULP WGS (0.1x) sequencing (dbGaP accession: phs002387.v1.p1). ER status and tumor fraction were obtained from the publication. Samples were then realigned to hg38 as described in 'Sequence data processing'. The second cohort was from the study by Bujak et al.[103], which included WGS of cfDNA from 27 patients with ER+ MBC (NCBI BioProject accession: PRJNA578569). ER status was obtained from the publication. The third cohort was the 'Independent MBC cohort' which consisted of ULP-WGS data generated as described below (Methods: Independent MBC cohort).

Tumor fraction for the Bujak et al cohort was estimated using ichorCNA. Samples were aligned to hg19 in order to use the default panel of normal provided with ichorCNA. 'ichorCNA_fracReadsInChrYForMale' was set to 0.001 and all other parameters were defaults. For Griffin analyses, samples from this cohort were aligned to hg38 as described in 'Sequence data processing' and downsampled to 0.1x WGS as described in 'Downsampling cfDNA sequencing data to 0.1x coverage' prior to Griffin analysis.

Independent MBC cohort

Human subjects

Patients were enrolled on clinical data collection and biospecimen banking protocols. Eligible patients had biopsy-proven metastatic breast cancer. Hormone receptor status was performed using Clinical Laboratory Improvement Amendments (CLIA) approved assays. Estrogen receptor (ER) positivity was defined as ≥5% of cells positive by immunohistochemistry (IHC). Human epidermal growth factor receptor 2 (HER2) negativity was defined as IHC score 0 or 1+ and/or HER2:CEP17 fluorescent in situ hybridization (FISH) ratio <2.0. HER2 positivity was defined as IHC score 3+, or IHC score 2+ with HER2:CEP17 FISH ratio ≥2.0. Triple negative breast cancer (TNBC) was defined as < 5% staining for ER and progesterone receptor (PR), as well as HER2 negativity as previously defined. The protocols were approved by the institutional review board of the Dana Farber Cancer Institute (DFCI-09204) or Ohio State University (2007C0066, 2018C0211). All patients provided written informed consent for blood sample collection, genomic analyses, and collection of clinicopathologic data. A total of 103 samples from 30 patients were used for this study. This included 15 hormone receptor positive patients and 15 TNBC patients.

Blood sample processing and plasma extraction

Venous blood samples (10 mL) were collected in EDTA (BD, Franklin Lakes, NJ), CellSave Preservative (Cell Search, Raritan, NJ) or Cell-Free DNA BCT (Streck, Omaha, NE) tubes. EDTA tubes were processed within 4 hours of collection and Streck tubes within 48 hours. Whole blood was centrifuged at 1900g for 10 minutes at room temperature with the brake off. Plasma was removed and transferred to Eppendorf DNA LoBind tubes, then centrifuged at 1900g for 10 minutes at room temperature. Plasma was transferred to cryovials and frozen at -80C for storage.

Frozen aliquots of plasma were thawed at room temperature. cfDNA was extracted using the QIAsymphony DSP Circulating DNA Kit according to the manufacturer's instructions, with ~4 mL of plasma as input and with a 60 uL DNA elution.

Library Construction

Initial DNA input is normalized to be within the range of 25-52.5 ng in 50 uL of TE buffer (10mM Tris HCl 1mM EDTA, pH 8.0) according to picogreen quantification. Library preparation is performed using a commercially available kit provided by KAPA Biosystems (KAPA HyperPrep Kit with Library Amplification product KK8504) and IDT's duplex UMI adapters. Unique 8-base dual index sequences embedded within the p5 and p7 primers (purchased from IDT) are added during PCR. Enzymatic clean-ups are performed using Beckman Coultier AMPure XP beads with elution volumes reduced to 30μL to maximize library concentration.

Post Library Construction Quantification and Normalization

Library quantification was performed using the Invitrogen Quant-It broad range dsDNA quantification assay kit (Thermo Scientific Catalog: Q33130) with a 1:200 PicoGreen dilution. Following quantification, each library is normalized to a concentration of 35 ng/μL, using Tris-HCl, 10mM, pH 8.0.

Library Pool Creation and Ultra-low Pass Sequencing

In preparation for the sequencing of the ultra-low pass libraries (ULP), approximately, 4 μL of the normalized library is transferred into a new receptacle and further normalized to a concentration of 2ng/μL using Tris-HCl, 10mM, pH 8.0. Following normalization, up to 95 ultra-low pass WGS samples are pooled together using equivolume pooling. The pool is quantified via qPCR and normalized to the appropriate concentration to proceed to sequencing. Cluster amplification of library pools was performed according to the manufacturer's protocol (Illumina) using Exclusion Amplification cluster chemistry and HiSeq X flowcells. Flowcells were sequenced on v2 Sequencing-by-Synthesis chemistry for HiSeq X flowcells. The flowcells are then analyzed using RTA v.2.7.3 or later. Each pool of ultra-low pass whole genome libraries is run on one lane using paired 151bp runs.

Castration resistant prostate cancer (CRPC) samples

Deep WGS (16-61x) of cfDNA from patients with castration resistant prostate cancer (CRPC) and healthy donors were obtained from an existing dataset published by Viswanathan and colleagues and Adalsteinsson and colleagues.[61,66] Bam files were downloaded from dbGaP (accession: phs001417.v1.p1). Coverage and tumor fraction metrics were obtained from the supplementary data in the publications. These samples were used for designing and demonstrating the pipeline.

Sequence data processing

All sequencing data used in this study was realigned to the hg38 version of the human genome (downloaded from http://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.gz). Bam files were unmapped from their previous alignment using Picard (v2.18.29) SamToFastq. They were then realigned to the human reference genome according to GATK best practices[127] using the following procedure. Fastq files were realigned using BWA-MEM (v0.7.17).[128] Files were then sorted with samtools (v1.9)[129], duplicates were marked with Picard, and base recalibration was performed with GATK (v4.1.0.0), using known polymorphisms downloaded from the following locations: https://console.cloud.google.com/storage/browser/genomics-public-data/resources/broad/hg38/v0/Mills_and_1000G_gold_standard.indels.hg38.vcf.gz and https://ftp.ncbi.nih.gov/snp/organisms/human_9606_b151_GRCh38p7/VCF/GATK/All_20180418.vcf.gz .

Transcription factor binding site (TFBS) selection

Transcription factor binding sites (TFBSs) were downloaded from the GTRD database[98]. This database contains a compilation of ChIP-seq data from various sources. For our analyses, we used the meta clusters data (version 19.10, downloaded from https://gtrd.biouml.org/downloads/19.10/chip-seq/Homo%20sapiens_meta_clusters.interval.gz). This contains meta peaks observed in one or more ChIP-seq experiments. The GTRD database contains some ChIP-seq experiments for targets that are not transcription factors (TFs). These were excluded by comparing against a list of TFs with known binding

sites in the CIS-BP database[130] (v2.00 downloaded from http://cisbp.ccbr.utoronto.ca/bulk.php). The site position was identified as the mean of 'Start' and 'End'. For GC, mappability, and CNA correction analyses as well as TFBSs nucleosome profiling in MBC, TFs with less than 10,000 sites on autosomes were excluded resulting in 377 TFs. For each remaining TF, the top 10,000 sites were selected by choosing those with the highest 'peak.count' (number of times that peak has been observed across all experiments). For cancer detection we tried several cutoffs (1,000 to 50,000 TFBS) and selected an optimal cutoff of 30,000 sites, resulting in 270 TFs (see number of sites analysis below). For the MBC ER status classifier shown in Appendix: Supplementary Fig. 10, we also used the top 30,000 sites.

Identification of differential TFs in blood and cancer

To identify transcription factors that were differentially expressed between blood cells and breast cancer, we used the University of California Santa cruise (UCSC) Xena online differential gene expression analysis tool (http://xena.ucsc.edu/)[131] which uses the Appyter bulk RNA-seq analysis pipeline to run Limma-Voom differential gene expression analysis.[132] After launching the tool via a web browser, we selected the publicly available 'TCGA TARGET GTEX' study which includes RNA seq from TCGA tumors as well as RNA seq from GTEX healthy tissues. The version of the data was 2016-04-12. We selected the phenotypic variables 'main category' which groups samples by tissue or tumor type and 'study' which groups samples by study. We then ran a differential gene expression analysis on the 'main category' variable and selected GTEX Blood (337 samples) and TCGA_Breast_Invasive_Carcinoma (1099 samples) as the two subgroups to compare in the analysis. All other parameters were left as defaults. We used the outputs to determine which of the 377 TFs (see 'Transcription factor binding site (TFBS) selection' above) were differentially expressed between blood cells and breast cancer cells (using default cutoffs: adjusted p-value ≤ 0.05 and absolute value of log2 fold-change ≥ 1.5). This yielded 107 TFs that were upregulated in BRCA and 82 TFs that were upregulated in blood. We noted that some TFs shared a large number of binding sites with TFs that were upregulated in the opposite tissue type. For instance, MECOM (also called EVI1) has previously shown to be less accessible in BRCA than blood[95] and we saw the same trend of increased accessibility in

blood in our data. However, according to the differential RNA seq analysis, MECOM is actually upregulated in BRCA relative to blood. We suspected that this is due to the fact that it shares almost half of its top 10,000 sites (4,465 sites) with blood specific LYL1 and >15% of the top 10,000 sites each with of a number of other factors that are upregulated in blood including ZBTB16 (2884 sites), TBX21 (1837 sites), STAT5A (1936 sites), and SPI1 (2075 sites). Because of this type of extensive site overlap seen in some differential TFs, we implemented a filter to exclude differential TFs which shared too many sites with the opposite tissue type. For the top 10,000 TFBSs for each TF, we examined how many of them overlapped (binding site was within ± 250bp) with the top 10,000 TFBSs for each TF of the opposite tissue type (i.e. for each TF that was upregulated in blood, we looked at how many sites it shared with each of the 107 TFs that were upregulated in MBC and took the mean of these 107 values). If a TF overlapped with an average of 400 or more sites for the factors expressed in the opposite tissue type, it was excluded from the list of differentially expressed TFs because it was considered to share too many sites with the opposite class, potentially confounding tissue specific accessibility. This left us with 22 TFs that were upregulated in blood and 35 factors that were upregulated in cancer.

DNase I hypersensitivity site selection

DNase I hypersensitivity sites for a variety of tissue types were downloaded from https://zenodo.org/record/3838751/files/DHS_Index_and_Vocabulary_hg38_WM20190703.txt.gz[133]. These sites were split by tissue type for a total of 16 site lists. The 'summit' column was used as the site position. The sites were sorted by the number of samples where that site had been observed ('numsamples') and the top 10,000 most frequently observed sites were selected for each tissue type.

ATAC-seq site selection for ER subtyping

Assay for transposase-accessible chromatin using sequencing (ATAC-seq) site accessibility for primary breast cancer samples from The Cancer Genome Atlas (TCGA) were downloaded from the TCGA ATAC-seq hub (https://gdc.cancer.gov/about-data/publications/ATACseq-AWG)[100]. A file containing raw counts

for all cancer type specific sites were downloaded ('All cancer type-specific count matrices in raw counts') and the file containing breast cancer specific sites was used ('BRCA_raw_counts.txt'). The locations of these sites and patient metadata were obtained from the supplementary tables in the paper[100]. Sites on autosomes were kept for further analysis for a total of 211,938 sites. Differentially accessible sites between ER+ (n=44) and ER- (n=15) tumors were identified using the DESeq2 software[105]. The software was run using default settings described in the 'quick start' guide with no co-variates. A differential accessibility experiment was run using the 'DESeq' and 'results' functions followed by log fold change shrinkage using the 'lfcShrink' function. Sites with an adjusted p-value < 5×10^{-4} were selected. Additionally, selected sites were further filtered based on the log2 fold-change between ER+ and ER- tumors (see 'Selection of cutoffs for ER status differential ATAC-seq sites' below). Sites with a log2 fold change >0.5 were classified as ER+, while sites with a log2 fold change <-0.5 were classified as ER-. These site lists were further split into sites shared with hematopoietic cells and those not shared with hematopoietic cells. Hematopoietic sites were obtained from a database of single cell ATAC-seq data[101] (GEO accession number: GSE129785, peak file available here: https://ftp.ncbi.nlm.nih.gov/geo/series/GSE129nnn/GSE129785/suppl/GSE129785%5FscATAC%2DHematopoiesis%2DAll%2Epeaks%2Etxt%2Egz). Hematopoietic sites were lifted over to hg38 using the UCSC liftover command line tool and sites that changed size during liftover (0.2% of sites) were discarded. ER differential ATAC-seq sites were overlapped with hematopoietic sites (Overlapping sites were defined as site centers being within 500bp of one another) using pybedtools intersect[123,134]. This resulted in a total of 4 differential site lists: ER positive sites that were not shared with hematopoietic cells (18,240 sites), ER positive sites that were shared with hematopoietic cells (9,930 sites), ER negative sites that were not shared with hematopoietic cells (19,347 sites), and ER negative sites that were shared with hematopoietic cells (22,365 sites).

To further characterize these sites, overlapped the four site lists with the top 10,000 sites for each of 377 transcription factors (TFs) using pybedtools intersect. An overlapping pair of sites was defined as having <500bp between site centers. Each differential ATAC-seq site list was compared against each list of TFBSs

and the total number of ATAC sites overlapping one or more TFBSs on the given list was recorded (Supplementary Data 13).

Selection of cutoffs for ER status differential ATAC-seq sites

In order to select the optimal p-value and fold change cutoffs for identifying ER+ and ER- differential ATAC-seq sites, we tried several different cutoffs for the values output by DESeq2. First we tried 4 different log2 fold-change cutoffs (no cutoff, 0.5, 1, 2) while holding the adjusted p-value cutoff constant at 0.05. Second, we tried 3 additional p-value cutoffs while holding the log2 fold-change constant at 0.5. For each cutoff, we ran the griffin nucleosome profiling analysis on the selected ATAC-seq sites, using 100-200bp fragments. We then extracted features and used these in a logistic regression model to predict ER status as described below ('Machine learning, bootstrapping, and performance evaluation procedure' and 'ER status classification in the MBC cohort') and calculated the mean accuracy across all bootstraps. We found that there was a relatively small difference between cutoffs (~2% accuracy) but chose the cutoff with the highest accuracy (adjusted p-value $\leq 5x10^{-4}$ and absolute value of log2 fold-change >0.5) for our final model.

Quantification of GC content at TFBS

For 377 TFs (see Transcription factor binding site (TFBS) selection above), the GC content around the top 10,000 sites was quantified (Shown in Figure 2a and Appendix: Supplementary Fig. 1). The sequence at each site (±1000bp) was fetched from the genome and the GC content was calculated. Positions within sites that overlapped the exclusion lists or had zero mappability were excluded. GC content at individual sites was then smoothed using a Savitsky-Golay filter with length 165bp and polynomial order zero. The mean GC content at the site center was calculated as the mean of the smoothed GC content across all sites in the window ±30bp from the site. The mean flanking GC content was calculated as the mean of the GC content in the window ±1,000 bp from the site, excluding the central region (±30bp).

Assessment of Griffin before and after GC correction, mappability correction, and CNA correction

Tumor fraction correlations at TFBS

For 191 MBC ULP samples with >0.1 tumor fraction, nucleosome profiling with and without GC correction was performed on the top 10,000 sites for each of 377 transcription factors (TFs) using nucleosome sized fragments (100-200bp). For each TF, the relationship between central coverage and tumor fraction was modeled using scipy.stats.linregress[135] producing a Pearson correlation (r) and line of best fit (scipy version 1.7.1). Pearson p-values for each feature type were adjusted using a Benjamini-Hochberg FDR correction. Root mean squared error (RMSE) was calculated from the line of best fit. This was performed both before and after GC correction as illustrated for LYL1 in Fig. 2e. For all 377 TFs, the RMSE values before and after GC correction were compared using a Wilcoxon signed-rank test (two-sided). This same procedure was applied to test the benefit of an additional mappability correction step and an additional copy number correction step.

Mean absolute deviation (MAD) at TFBS

For 215 healthy donors, nucleosome profiling with and without GC correction was performed on the top 10,000 sites for each of 377 TFs. For each TF, the MAD of the central coverage values was calculated both before and after GC correction. For all 377 TFs, the MAD values before and after GC correction were compared using a Wilcoxon signed-rank test (two-sided).

Quantification of differential accessibility of TFBS and ATAC sites in MBC

To determine whether nucleosome profiles around TFBSs were differentially accessible between ER+ and ER- samples we performed an analysis of covariance (ANCOVA) as implemented in Pingouin (v0.5.1)[136]. For this analysis, we used the 191 MBC samples with ≥0.1 TFx and ≥0.1 coverage. Nucleosome profile feature (central coverage, mean coverage, or amplitude) was the dependent variable, primary tumor status was the independent variable ('between'), and tumor fraction was a covariate. We performed this analysis on all 3 features for all 377 TFs with 10,000 or more sites. We then used Benjamini-Hochberg FDR

correction to perform multi-test correction for on the p-values for ER status and tumor fraction for each feature.

We performed the same ANCOVA analysis on the features for the 4 types of ER differential ATAC sites but without FDR correction as there were only a total of 12 features (central coverage, mean coverage, and amplitude for each of the 4 site types).

Machine learning, bootstrapping, and performance evaluation procedure

To detect cancer or predict ER subtype, we used logistic regression with Ridge regularization (i.e. L2 norm) as implemented in scikit-learn (v0.23.2)[137]. All feature values were scaled to a mean of 0 and a standard deviation of 1 prior to performing bootstrapping and fitting the models. We used the following bootstrapping procedure to train and assess the performance of our models. First, we selected n samples with replacement from the full set of n samples and used this as a training set. Samples that weren't selected were used as the test set. We then used 10-fold cross-validation on the training set to select the hyperparameter 'C' (inverse of the regularization strength). To account for class imbalances in the data we used set the 'class weight' parameter to 'balanced' to adjust the sample weighs inversely proportional to the class frequencies. We trained a final model on all the training data using the selected regularization strength. Finally, we tested this model on the test set and recorded the performance (accuracy and AUC values) and sample probabilities. Then, a new training set was selected, and the procedure was repeated for 1,000 iterations. After completing the bootstrap iterations, we calculated the AUC and accuracy from each bootstrap iteration and used these to generate the mean and 95% confidence interval around each of these values and to create boxplots. To visualize the ROC curve, we used the median probability from all bootstraps where each sample was included in the test set. For further downstream analyses, including the comut plot, barplots, and timelines we used this same median probability.

Features used for the final cancer detection classification

To detect cancer, we applied the logistic regression approach described above and built a logistic regression classifier on features extracted from the DELFI cohort cancer and healthy samples. First, we performed nucleosome profiling in these samples (selecting fragments 100-200bp in length). For our finalized model we used 30,000 TFBS each for 270 TFs with at least 30,000 sites in the GTRD database (see 'Selection of number of TFBS for cancer detection' below). We extracted three features (as described above 'Griffin: nucleosome profile feature quantification') from each coverage profile for a total of 810 features. We then scaled these features to a mean of 0 and a SD of 1. Within each bootstrap iteration, we reduced the dimensionality of the feature using PCA as implemented in scikit learn[137] on the training set and selected the features that explained 80% of the variance. We then applied this same PCA transformation to the test set for that bootstrap. These PCA components were then used as the inputs for the logistic regression model which was trained on the training set, and tested on the test set.

For the LUCAS cohort, we found that there were batch effects that prevented using the same model trained on the DELFI cohort. Because of this we trained and tested a new model on the LUCAS cohort using the same bootstrapping approach and performed a final validation of this model in the LUCAS validation cohort (described below, 'Validation of the cancer detection model').

Finally, we downsampled both the DELFI and LUCAS cohorts to ~0.1x coverage (procedure described below) and performed the same cancer detection analysis in this lower coverage data.

Validation of the cancer detection model

After training and testing a logistic regression model using the bootstrapping approach on the LUCAS cohort, we applied the final model to the previously unseen LUCAS validation cohort (385 healthy samples and 46 cancer samples). To get this final model, first, we performed PCA on the full LUCAS cohort (not including the LUCAS validation cohort) and extracted 35 features that explained 80% of the variance in that cohort. Then, we used these 35 features to build a logistic regression model using the regularization strength most frequently chosen by the 1,000 bootstraps on the LUCAS cohort ('C'=0.01). Finally, we applied the PCA transformation and logistic regression model to the LUCAS validation cohort, extracted

the same 35 features, and got a probability of cancer for each sample. We obtained confidence intervals for the AUC using a bootstrap procedure in which we selected 431 samples with replacement from the original 431 samples and calculated the AUC for the selected samples. We then repeated this 1,000 times to get 1,000 AUC values which we used to obtain confidence intervals and boxplots.

We repeated this same procedure for the downsampled LUCAS validation cohort.

Selection of number of TFBS for cancer detection

In order to select the optimal number of TFBS for cancer detection, we tried several different cutoffs for the number of sites (1000, 5000, 10000, 20000, 30000, and 50000 sites). For each cutoff we identified all TFs with at least that many sites in GTRD resulting in 566, 446, 377, 316, 270 and 202 TFs respectively for the cutoffs above. We then picked the top sites by choosing those with the highest 'peak.count'. We next used the logistic regression with bootstrapping and PCA dimensionality reduction described above ('Features used for the final cancer detection classification') to train and test models on both the original 1-2x WGS DELFI cohort samples and the downsampled 0.1x DELFI cohort samples. We found that the number of sites had a greater impact on the downsampled data (Appendix: Supplementary Fig. 5a) so we selected the cutoff with the highest AUC in downsampled data which was 30,000 sites.

Cancer detection from DNAse hypersensitivity sites

In addition to examining TFBS, we also performed nucleosome profiling at the 16 tissue-specific DHS site lists described above. We extracted the same 3 features from each site profile for a total of 48 features and used the same bootstrapping plus PCA dimensionality reduction described above to test the performance of this model.

Downsampling cfDNA sequencing data to 0.1x coverage

WGS data for the DELFI cohort, LUCAS cohorts (training and validation), and Bujak et al. dataset was aligned to aligned to hg38 and subsequently downsampled using Picard DownSampleSam. The probability

used by DownSampleSam was calculated based on a target of 2,463,109 read pairs which resulted in approximately 0.11x coverage as calculated by Picard CollectWgsMetrics. Downsampled bam files from the DELFI dataset were realigned to hg19 for use in the Ulz pipeline for comparison. The realignment procedure was the same as above but using the hg19 genome (downloaded from https://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/hg19.fa.gz) and hg19 known polymorphic sites for base recalibration (downloaded from ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/hg37/Mills_and_1000G_gold_standard.indels.hg37.vcf.gz and ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606_b151_GRCh37p13/VCF/GATK/All_20180423.vcf.gz)
.

ER status classification in the MBC cohort

To predict ER status, we applied the logistic regression approach described above to features extracted from the MBC patient samples. Because some patients had multiple samples, we modified the bootstrapping procedure to select 139 patients (rather than samples) with replacement from a full set of 139 patients. For each selected patient, all samples from that patient were added to the training set (If a patient was selected multiple times, all their samples were included multiple times). This ensured that separate samples from the same patient (biological replicates) could not appear in both the training and test set. Samples from patients that weren't selected were used as the test set.

For our model, we applied nucleosome profiling using 100-200bp fragments to the 4 ER differential ATAC seq lists and extracted 3 features per profile for a total of 12 features. For evaluating the model, we only included the first timepoint for each patient in the test set when calculating the accuracy and AUC for each bootstrap iteration. This prevented a small number of patients with many samples from having a large impact on the scores.

ER status prediction from TFBS

In order to assess whether ER status could be predicted from the nucleosome profiles around TFBSs, we performed nucleosome profiling for the top 30,000 sites for 270 TFs and extracted 3 features each for a total of 810 features. We then used the bootstrapping approach described above ('ER status classification in the MBC cohort') to train and test the model. Because of the high dimensionality of the data, within each bootstrap, we performed PCA on the training set and selected the top PCA components that described 80% of the variance. We then used these components as the features in our logistic regression model. This model did not perform as well as the differential ATAC site model and was not used for further analysis.

Validation of the ER status prediction model

After training and testing a logistic regression model using the bootstrapping approach on the MBC cohort and features from griffin profiles around differential ATAC sites, we applied the final model to the three previously unseen validation cohorts. To get this final model, we trained a logistic regression model on the full MBC dataset (254 samples) using the regularization strength most frequently chosen by the 1,000 bootstraps on the MBC cohort ('C'=0.1). We then applied this model to the three validation cohorts and got the probability of ER+ for each sample. For patients with multiple samples (in the independent MBC cohort) we used the first timepoint when evaluating performance, resulting in a total of 71 samples from unique patients across all three cohorts. We obtained confidence intervals for the accuracy and AUC using a bootstrap procedure in which we selected 71 samples with replacement from the original 71 samples and calculated the AUC and accuracy for the selected samples. We repeated this procedure 1,000 times to get 1,000 AUC and accuracy values which we used to obtain confidence intervals and boxplots.

Transcription factor profiling using pipeline from Ulz et al.

We downloaded the Transcription Factor Profiling pipeline published by Ulz and colleagues from Github (https://github.com/PeterUlz/TranscriptionFactorProfiling)[95] and ran it using the following procedure as described in the paper. hg19 aligned bam files were used because the pipeline was written to for this version of the genome. Scripts were modified so that they worked in python3. We trimmed the reads in each bam

to 60bp using ‘trim from bam single end’ with modifications to skip unaligned reads. We ran ichorCNA on the original (untrimmed) bam using the default ichorCNA settings for hg19 except the bin size, which was modified to 50,000bp and no panel of normals. We then ran the transcription factor profiling analysis on the trimmed bam using the script run_tf_analyses_from_bam.py with options ‘-calccov’ and ‘-a tf_gtrd_1000sites’ and the ichorCNA corrected depth file as the ‘-norm-file’. This ran transcription factor profiling on 1,000 sites for each of 504 TFs. Finally, we ran the scoring pipeline. We used the high frequency amplitude ('HighFreqRange') for each of the 504 TFs in the accessibility output file (Accessibility1KSitesAdjusted.txt) as the features for a logistic regression model using the same bootstrapping with PCA dimensionality reduction as described for cancer detection and ER status prediction from TFBS above.

Data availability

The Independent MBC Cohort data generated for this study (0.1x WGS of cfDNA) have been deposited in dbGaP under accession code (we are still in the process of obtaining this code). All other datasets used in this study were published datasets from previous studies and can be obtained by authorization through the database (WGS of cfDNA) or downloaded freely (ATAC-seq, RNA-seq, ChIP-seq, DNAse-seq). WGS of cfDNA from castration resistant prostate cancer, metastatic breast cancer, and healthy donors was obtained from dbGaP (accession phs001417.v1.p1 [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001417.v1.p1]). WGS of cfDNA from breast cancer patients in the Ahuno et al. study was obtained from dbGaP (accession phs002387.v1.p1 [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs002387.v1.p1]). WGS of cfDNA from ER+ breast cancer patients in the Bujak et al. study was obtained from NCBI (BioProject accession number PRJNA578569 [https://www.ncbi.nlm.nih.gov/bioproject/PRJNA578569]). The DELFI cohort (WGS of cfDNA from early stage cancer patients and healthy donors) was obtained from EGA (dataset ID EGAD00001005339 [https://ega-archive.org/datasets/EGAD00001005339]). The LUCAS and LUCAS validation cohorts (WGS of cfDNA from lung cancer patients and healthy donors) were also

obtained from EGA (EGAD00001007796 [https://ega-archive.org/datasets/EGAD00001007796]). ATAC seq peak counts and sample metadata were downloaded from TCGA (https://gdc.cancer.gov/about-data/publications/ATACseq-AWG) and is freely available without authorization. DNAse-seq was downloaded from zenodo (https://zenodo.org/record/3838751/files/DHS_Index_and_Vocabulary_hg38_WM20190703.txt.gz). ChIP-seq was downloaded from GTRD version 19.10 (https://gtrd.biouml.org/downloads/19.10/chip-seq/Homo%20sapiens_meta_clusters.interval.gz). RNA seq to obtain differential gene lists was accessed from https://toil-xena-hub.s3.us-east-1.amazonaws.com/download/TcgaTargetGtex_RSEM_Hugo_norm_count.gz (version 2016-04-12) using the UCSC Xena online tool.

Code availability

Griffin software and the subtype classifier tool can be obtained from https://github.com/adoebley/Griffin. Code for analysis and machine learning models can be accessed at https://github.com/adoebley/Griffin_analyses.

Chapter 3. Characterizing small-cell lung cancer subtypes from cell-free DNA

3.1 Introduction

Small-cell lung cancer (SCLC) is an aggressive, neuroendocrine tumor which causes approximately 15% of lung cancers worldwide[138]. Each year, 40,000 new cases are diagnosed in the US, primarily among elderly individuals with a history of heavy smoking[139]. The prognosis is extremely poor with a 7% 5-year survival rate[138] and a median survival of 8-20 months depending upon the stage[140]. A number of factors play into this poor prognosis. SCLC is often diagnosed at an advanced stage when metastasis has already occurred. Most tumors initially respond to chemotherapy; however, treatment resistance usually arises and there are few options for treating chemotherapy resistant disease[138]. Recent work has suggested that there are four or more subtypes defined by expression of key transcription factors or other transcriptional programs[141–145]. Initially, these were proposed to consist of ASCL1, NEUROD1, POU2F3, and YAP1 although the YAP1 subtype has been suggested to be non-exclusive[142] and additional subtypes have been proposed including ATOH1[145], double negative[142] and inflamed[144]. The clinical significance of these subtypes is not well characterized, however there is evidence that targeted therapy could have different effects on different subtypes. For instance, drugs that target cells expressing DLL3 are more likely to work in ASCL1 subtype tumors because ASCL1 drives DLL3 expression[51]. Aurora kinase inhibitors maybe be more effective in POU2F3 tumors, because tumors with high MYC expression are more sensitive to these inhibitors[51]. Therapies targeting these subtypes might improve outcomes and reduce side effects from ineffective treatment[141], but in many patients, it is not possible to determine subtype because it is not the standard-of-care to obtain a surgical biopsy.

Cell-free DNA (cfDNA) offers a non-invasive alternative to traditional biopsy. In cancer patients, a fraction of cfDNA is derived from tumor cells. This DNA can be sequenced in order to non-invasively identify tumor characteristics[1]. Currently, targeted panels are used in the clinic to identify common

actionable mutations[24] but recent work has shown that it may also be feasible to study chromatin structure from nucleosome footprints in cfDNA[29,45], raising the possibility of identifying tumor phenotypes, such as the activity of transcription factors, from non-invasive biopsies. This could be beneficial in SCLC patients because tumor tissue from biopsies is not always available and cfDNA could allow study of tumor subtypes in patients even when tumor tissue is unavailable for immunohistochemistry or RNA sequencing. Here we present a proof of concept study using targeted panel sequencing to identify mutations, TFBS accessibility, and TSS activity and use these to identify subtypes from cfDNA derived from patient xenografts and patient samples.

3.2 Results

3.2.1 *Targeted panel design and application for mutation calling*

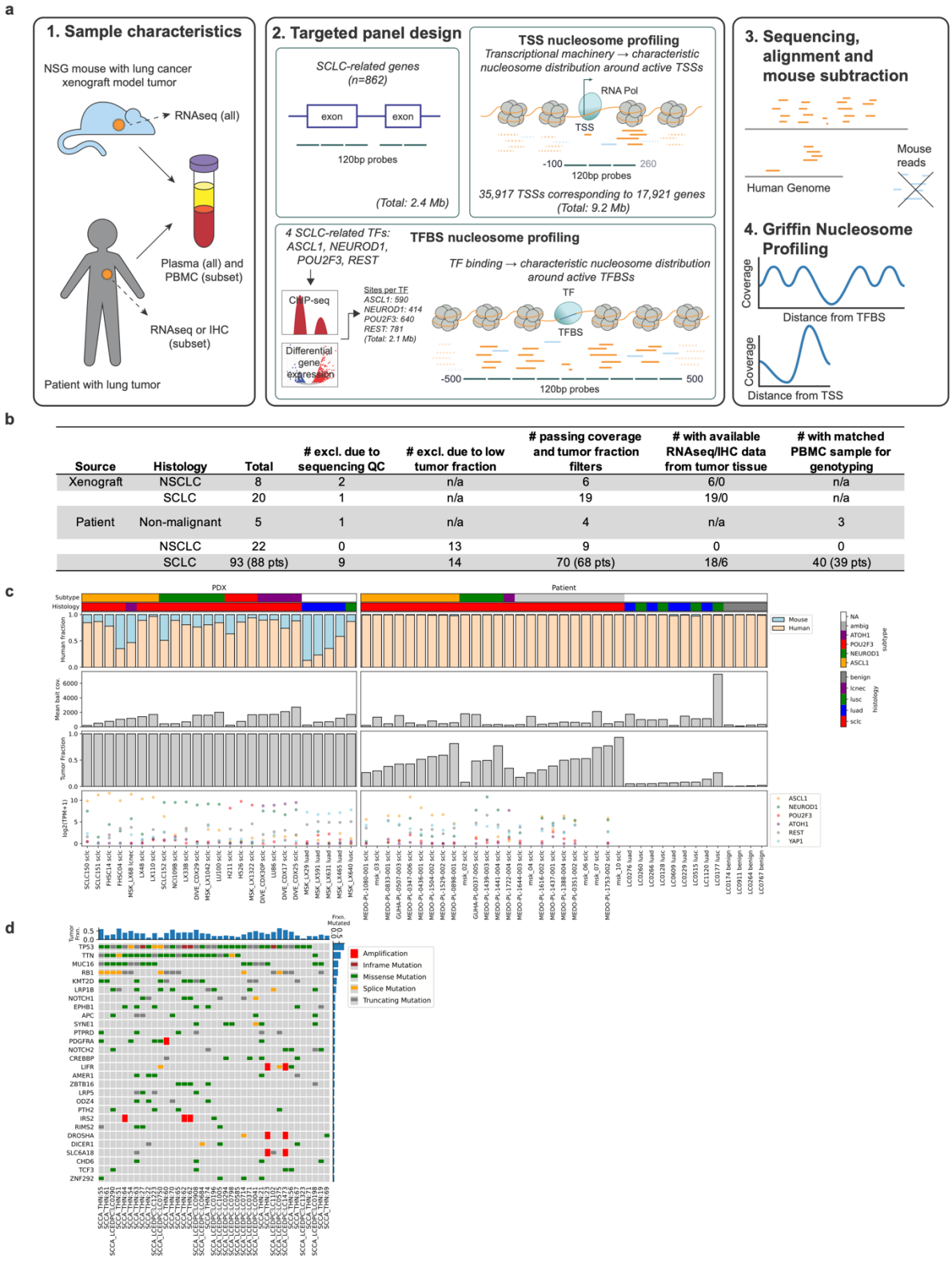
Targeted panel sequencing is a cost effective alternative to deep whole genome sequencing that has been used for cfDNA sequencing for both mutation calling[24] and for nucleosome profiling at TSS.[42] We designed a custom panel intended specifically for SCLC targeting a variety of regions including exons, TSS, and TFBSs. The exon part of the panel was designed for variant detection and covered exons from 862 cancer-mutated genes for a total of 2.4 Mb. Mutation calling from targeted panel sequencing in SCLC has previously been demonstrated using the MSK-IMPACT panel which covers 505 cancer related genes[146] and a version of this panel called MSK-ACCESS has been used to detect mutations in cfDNA[147]. However, our new panel includes a wider spectrum of cancer and SCLC-mutated genes and is integrated with other target types designed to detect both genotypic and phenotypic tumor characteristics. The TSS part of the panel was designed for nucleosome profiling to predict gene expression and captured a 360bp window spanning from 100bp upstream of the TSS to 260bp downstream of each of 35,917 TSS corresponding to 17,921 genes (9.2 Mb). Lastly, the TFBS part of the panel captured a 1000bp window centered around 2,425 TFBSs for four transcription factors. These included three TFs that define SCLC subtypes (ASCL1, NEUROD1, and POU2F3) as well as REST which is a repressor of neuroendocrine

gene expression and is upregulated in SCLC tumors with high immune related gene activity and downregulated in neuroendocrine tumors. TFBSs were selected from published ChIP-seq experiments in SCLC then further refined by selecting sites where the closest TSS was for a differentially upregulated gene in the subtype of interest (590 ASCL1, 414 NEUROD1, and 640 POU2F3 sites)[148,149]. REST sites were selected from non-tissue specific ChIP-seq experiments in the gene and transcriptional regulation database (GTRD)[150] and then further refined by requiring that the nearest TSS was differentially expressed between REST-high and REST-low lines in DepMap (781 REST sites). Overall, the targeted TFBSs spanned a total of 2.1Mb. Targeted regions were tiled with 120bp probes and a total of 13.7Mb of area was covered (Fig. 3.1a).

To test this panel, we grew 28 patient derived xenograft (PDX) lines in NGS mice and collected plasma. In addition, we obtained 120 plasma samples from 115 patients, primarily with SCLC but 22 with NSCLC and 5 with lung nodules that were determined to be benign. We isolated cfDNA from the plasma and used our custom panel to perform targeted panel sequencing (Fig. 3.1a). We aligned the reads to the human genome and removed mouse reads using a bioinformatic approach (methods)[151]. This was performed on both mouse and human samples for processing consistency. We also performed ultra-low pass (0.1x) WGS on the samples and used ichorCNA to predict tumor fraction in the human patients (Fig. 3.1c)[28]. In xenografts, nearly all cfDNA that remains after mouse subtraction is derived from the human tumor leading to nearly 100% tumor fraction while patients have some DNA that is derived from nucleated blood cells. We excluded patient samples that did not have at least 5% tumor fraction, other than patients with benign nodules which we kept as non-cancer controls. We also excluded any sample without at least 50% of targets covered at 50x depth. This left us with a total of 25 xenograft samples and 86 patient samples from 84 patients that passed quality control filters (Fig. 3.1b). These samples had a median on target coverage of 542x (Fig. 3.1c). To determine tumor phenotypes, we also isolated RNA from tumors in all xenograft models (except three POU2F3 lines with published RNA seq available) and from 18 patient tumors (Fig. 3.1c). An additional 6 patient tumors had IHC for key transcription factors which enabled us to determine the phenotype. We then performed mutation calling on the targeted exons

and performed nucleosome profiling using the Griffin pipeline around each individual targeted TSS or TFBS (Fig. 3.1a).

In the 39 SCLC patients with matched PBMC data, mutations were observed in driver genes including characteristic TP53 mutations in 36 of 39 patients (92%) and RB1 mutations in 16 of 39 (41%, Fig. 3.1d). The observation of these and other expected driver genes confirms that the targeted panel approach for mutation calling is feasible in SCLC.



Source	Histology	Total	# excl. due to sequencing QC	# excl. due to low tumor fraction	# passing coverage and tumor fraction filters	# with available RNAseq/IHC data from tumor tissue	# with matched PBMC sample for genotyping
Xenograft	NSCLC	8	2	n/a	6	6/0	n/a
	SCLC	20	1	n/a	19	19/0	n/a
Patient	Non-malignant	5	1	n/a	4	n/a	3
	NSCLC	22	0	13	9	0	0
	SCLC	93 (88 pts)	9	14	70 (68 pts)	18/6	40 (39 pts)

Figure 3.1 Study overview

Fig. 3.1 Study overview. (a) 1. Blood and tumor samples were obtained from xenograft models of SCLC or NSCLC with known phenotypes. In addition, blood and tumor samples were also obtained from patients with SCLC or NSCLC. 2. cfDNA was extracted from plasma samples and hybridized to a custom capture probe set that targeted 1000bp windows around key TFBSs, 360bp windows around TSS, and exons for key SCLC genes. 3. Captured fragments were sequenced and reads were aligned to a concatenated human and mouse reference genome. Reads which aligned to the mouse genome were removed leaving behind only human reads. Nucleosome profiling was performed using the griffin pipeline to perform GC correction and generate coverage profiles around targeted sites. **(b)** Overview of samples used in the study. Samples included PDX models of both SCLC and NSCLC tumors and patients with non-malignant nodules, NSCLC, or SCLC tumors. Samples were excluded from further analysis if they did not pass coverage filters (>50% of targets covered with at least 50x depth) and tumor fraction filters (>0.05 tumor fraction). All PDX models (SCLC and NSCLC) and 18 patient samples (SCLC only) had tumor RNA sequencing available for phenotypic characterization. An additional 6 SCLC patient samples had tumor IHC for key transcription factors allowing a phenotypic classification. In addition, 39 SCLC patients and 3 patients with non-malignant nodules had matched PBMCs for genotyping. **(c)** Sequencing metrics and metadata for samples that pass the QC filters. PDX models are shown on the left and patients on the right. SCLC samples with unknown subtype are not shown. Top row, molecular tumor subtype for SCLC. Second row, tumor histology. Third row, the fraction of cfDNA reads that aligned to the human genome. In PDX models this corresponds to the fraction of cfDNA that is derived from the human tumor. In patients, a small number of reads were removed during the mouse subtraction step either due to non-specific sequence or adapter liftover. Fourth row, mean on-target coverage. Fifth row, tumor fraction, the fraction of cfDNA reads that were derived from the tumor as calculated by ichorCNA. This is always 100% for PDX models because the non-tumor mouse reads can be removed. Bottom row, expression of key genes from tumor RNA sequencing. **(d)** Mutation calls in SCLC patients with matched PBMC.

3.2.2 *SCLC transcriptional subtypes can be distinguished from targeted panel nucleosome profiling around TFBSs*

The xenograft models allowed us to examine nucleosome profiles in pure tumor cfDNA samples without the variable mixture of blood cell derived and tumor derived cfDNA observed in patients. We began by examining the mean coverage profiles for each of the 4 targeted TFs to see if there was differential accessibility at TFBSs corresponding to the differential activity of the TFs in xenograft models. If a site is active, we expect to see a loss of coverage at the TFBSs due to increased accessibility and peaks flanking the TFBSs due to organized nucleosome protection. In inactive sites, we expect a relatively flat coverage profile due to random nucleosome protection across the entire region. In general, we observed the expected pattern with more accessible TFBSs in xenograft models where the corresponding TF is active (Fig. 3.2a). ASCL1 had the highest accessibility in models of the ASCL1 subtype, including a large cell neuroendocrine carcinoma, although it was also accessible in some other SCLC models. NEUROD1 was most accessible in NEUROD1 and ATOH1 models which both express high levels of NEUROD1. POU2F3 was specifically accessible in POU2F3 models. And REST had the highest accessibility in three SCLC models with high REST expression. In addition, non-small cell lung cancer (NSCLC) xenografts including four lung adenocarcinoma (LUAD) models and one lung squamous cell (LUSC) model showed low accessibility at the three SCLC specific TFBSs (ASCL1, NEUROD1, and POU2F3) and high accessibility at REST sites, as expected (Fig. 3.2a). Next, we quantified the accessibility at each mean TFBS profile by measuring the difference between the coverage at the TFBSs and the coverage at the flanking peaks ('central amplitude'). We examined the relationship between the expression of the key transcription factors and the central amplitude for the corresponding TFBSs in SCLC xenograft models and found a significant correlation (Pearson correlation, 2-sided p-value) for all transcription factors except ASCL1 which appeared to be also accessible in ATOH1 and POU2F3 models despite lower ASCL1 expression in these models (Fig. 3.2b). Lastly, we performed a PCA on the mean coverage profiles for the TFBSs. We found that the top 2 PCs were able to separate the four subtypes although ATOH1 samples were less distinct, possibly because our targeted sites lacked any ATOH1

specific sites (Fig. 3.2c) or because the ATOH1 subtype may have significant biological overlap with the NEUROD1 subtype because NEUROD1 is often co-expressed in ATOH1 tumors. Overall, these findings suggest that targeted panel sequencing of a relatively small number of sites (roughly 500 per TF) can capture differential accessibility around key TFBSs in SCLC and distinguish subtypes in xenograft models of SCLC.

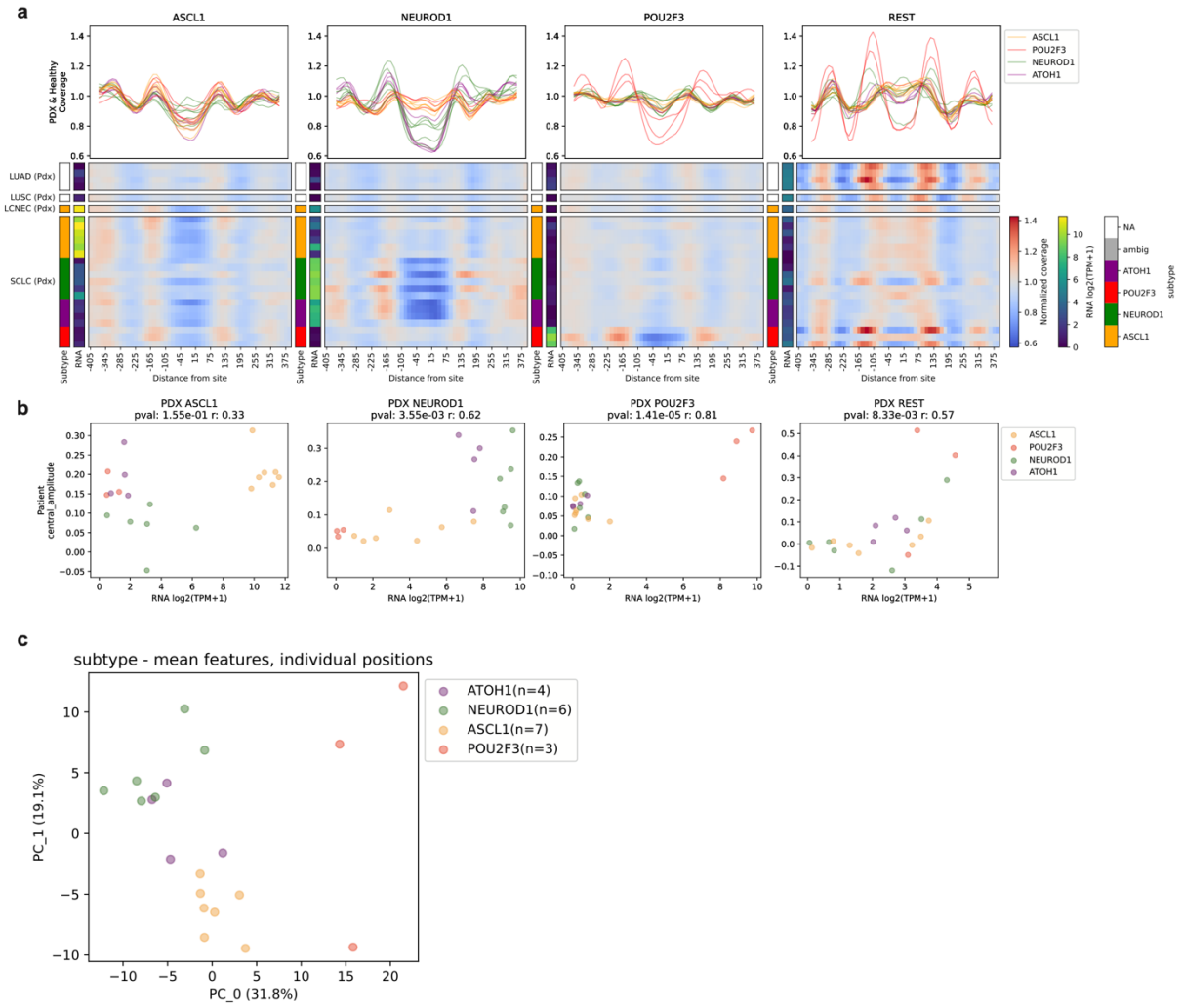


Figure 3.2 SCLC subtyping from TFBS targeted panel sites in PDX models

Fig. 3.2 SCLC subtyping from TFBS targeted panel sites in PDX models. (a) Coverage profiles around targeted TFBSs for each of 4 key transcription factors: ASCL1 (n=448 sites), NEUROD1 (n=330 sites), POU2F3 (n=471 sites) and REST (n=286 sites). Mean profiles for all targeted sites that passed

filters are shown. Plots in the top row shows the coverage in each individual PDX model colored by subtype. NSCLC models are shown in grey. Patients with non-malignant lung nodules are included in blue. Peaks in coverage correspond to the locations of nucleosomes while a loss of coverage at the center indicates a loss of nucleosome protection due to TFBS accessibility. Coverage profiles are also represented as a heatmap to better show the individual sample values. Blue indicates low coverage (due to TFBS accessibility) while red indicates high coverage (due to organized nucleosome protection). All factors are most accessible in the subtype where they are expressed. **(b)** Correlation between RNA (x axis) and amplitude of coverage dip at TFBSs (flanking nucleosome height minus center). The expression and amplitude are significantly correlated (Pearson correlation, two sided) for all NEUROD1, POU2F3, and REST. **(c)** Principal component analysis (PCA) of coverage values for the composite profiles from of all 4 targeted TFs. For each of the 4 coverage profiles, 54 features were extracted corresponding to the mean coverage value in 15 bp bins ranging from 405bp before the TF to 405bp after the TF. A PCA was then performed on all 216 features from the 4 profiles. The top two components, explaining 31.8% and 19.1% of variance respectively, are shown. These components can separate the four subtypes in PDX models with some overlap between ATOH1 and NEUROD1 models.

3.2.3 *Differential coverage profiles around TSS distinguish transcriptional subtypes*

In addition to targeting TFBSs for key TFs, we also targeted 35,917 TSSs corresponding to 17,921 genes. We then stratified all TSS coverage profiles from all xenograft models into ten groups based on the expression of each TSS in each xenograft model. We plotted the mean coverage profile in each of these groups and observed the expected pattern seen in previous studies with a loss of coverage near the TSS and a peak at the +1 nucleosome in highly expressed genes and a relatively flat profile in lowly expressed genes and the other deciles falling between these two extremes (Fig. 3.3a). CpG islands near TSS are known to play a key role in gene regulation and chromatin structure[152] so next we examined whether nucleosome profiles are influenced by the presence of a nearby CpG island. We separated TSS with adjacent CpGs (within 1,000 bp) from TSS without adjacent CpGs and plotted the mean coverage

profiles of the genes in each decile group, using the same expression cutoffs so that the profiles would be comparable. The CpG adjacent TSS were enriched for more active genes while the non-CpG adjacent TSS were enriched for less active genes, as expected. However, we noted that even among the highly expressed non-CpG adjacent genes, the coverage profiles appeared less 'active' than the coverage profiles of CpG adjacent genes with comparable expression (Fig. 3.3a). This suggests that coverage profiles are influenced by more than just gene activity and that CpG status may need to be considered when trying to predict gene expression from coverage profiles. To get a better idea of the variation between different TSS, we measured the amplitude of the signal (difference between the coverage at -45 and 120) at each TSS and generated a 2D histogram of the amplitude values for different expression levels. We observed that there was a wide distribution of amplitude values across all expression values (Fig. 3.3b). We also examined the fragment size entropy at each TSS, which has recently been shown to be related to gene expression[42]. TSS for expressed genes tend to have more variations in the fragment size distribution (higher entropy) while TSS for unexpressed genes tend to have fragments that are primarily around 167bp (lower entropy). We observed the expected trend of increased entropy with increasing expression but a wide range of values (Fig. 3.3b). Next, we decided to look at the TSS profiles for key lineage defining TFs in SCLC. In xenograft samples, we observed significant differences in coverage profiles for models that expressed a given gene compared to other models. However, strikingly, the coverage profiles at the different TSS were quite distinct from one another and generally lacked the characteristic loss in coverage at the TSS coupled with increased coverage at the +1 nucleosome (Fig. 3.3 c,d). Because of these differences in coverage profile between different TSS we developed a Euclidean distance feature to quantify how similar a coverage profile was to active models, relative to inactive models for a given TSS (methods). This bypassed the need to assume a specific profile shape in active TSS. However, it required us to exclude any TSS where there wasn't a significant difference in coverage profile between active and inactive models. We then selected genes that were putatively regulated by ASCL1, NEUROD1, ATOH1, or POU2F3 in SCLC using RNA sequencing from the CCLE database and identified a subset of 73 genes with significantly differential TSS profiles between high and low expression samples in our data. We

calculated the Euclidean distance feature for these models and used these features to perform a PCA. The first two dimensions were able to separate the four subtypes in xenograft samples (Fig. 3.3e). We also performed hierarchical clustering and were able to mostly cluster the subtypes (Fig. 3.3f). Overall, this demonstrates that there are expression specific changes in nucleosome profiles at TSS but they do not always follow the canonical shape when active. However, these shapes can still be used to identify gene expression changes that define subtypes.

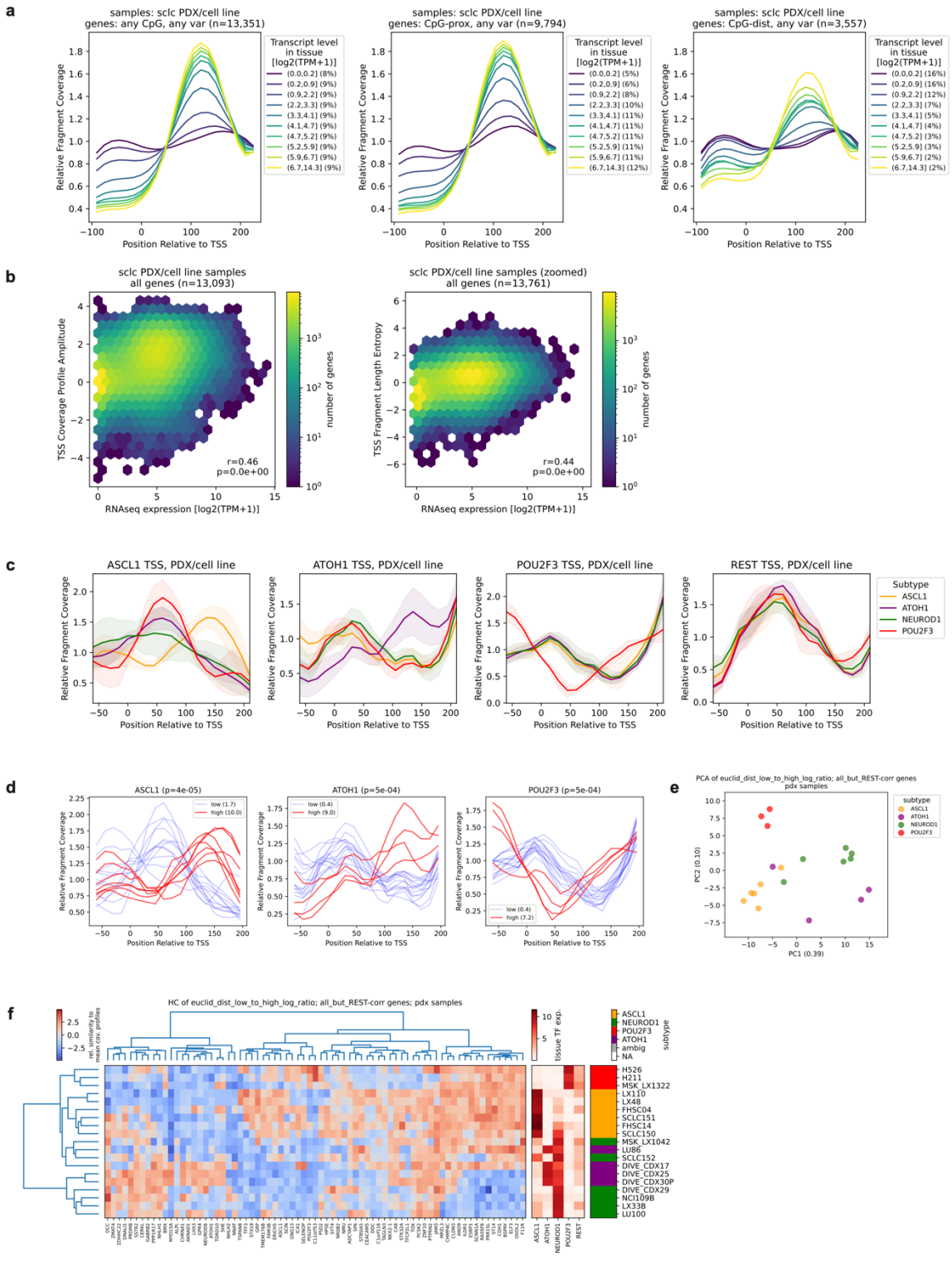


Figure 3.3 SCLC subtyping from TSS targeted panel sites in PDX models

Fig. 3.3 SCLC subtyping from TSS targeted panel sites in PDX models. (a) Mean coverage profiles for targeted TSS grouped by RNA expression decile in tumor. Left panel shows all TSS from all PDX models grouped by expression decile. Highly expressed genes have low coverage at the TSS and high coverage at the +1 nucleosome (+120bp) while unexpressed genes have a relatively flat coverage profile. Middle panel shows the subset of TSS that are adjacent to CpG islands, defined as TSS within 1,000bp of a CpG island from all PDX models. Expression cutoffs for each decile are the same as in the left panel but more genes fall into the higher categories because CpG adjacent genes tend to be expressed. Right panel shows the subset of TSS that are not adjacent to a CpG island. Cutoffs are the same, but more genes fall into the lower categories because non-CpG adjacent genes tend to have low expression. These TSS tend to have lower amplitude even when highly expressed. **(b)** 2D histogram showing the distribution of coverage profile amplitudes (+1 nucleosome coverage minus TSS coverage) across gene expression values. Left panel, all genes in all PDX models. Expressed genes tend to have higher amplitudes. Right panel, 2d histogram showing the distribution of fragment length entropy values across different gene expression values. Expressed genes tend to have somewhat higher entropy values compared to unexpressed genes. **(c)** Coverage profiles in PDX models at individual TSS for key transcription factors: ASCL1, ATOH1, POU2F3, and REST. Coverage profiles are grouped by subtype, mean ± standard deviation is shown as solid line and shading. Each profile has a unique shape which, except for REST, changes when the TSS is active. **(d)** Coverage profiles at key TSS shown for individual PDX models. Samples with high expression of the given gene have distinct profiles from those with low expression. **(e)** PCA of the Euclidean distance TSS features for the PDX models. The Euclidean distance feature for 73 TSS corresponding to subtype specific genes with differential coverage profiles were included in the PCA. The top two principal components are shown and can distinguish the four subtypes in PDX models. See methods for details on the Euclidean distance feature. **(f)** Heatmap of the Euclidean distance feature for 73 genes that are correlated to the expression of key transcription factors and have differential TSS profiles between models with high and low expression. These features cluster the PDX models into four

groups. Two groups correspond to two distinct subtypes (ASCL1 and POU2F3) while the remaining two groups contain a mixture of NEUROD1 and ATOH1 samples.

3.2.4 *Targeted panel sequencing captures patterns seen in deep whole genome sequencing*

To validate that the patterns captured by our targeted panel were a faithful representation of the patterns seen in deep whole genome sequencing (WGS) we performed deep WGS on 11 samples including 6 xenograft models (two each of ASCL1, NEUROD1, and POU2F3) and 5 patients (one each of NEUROD1 and ASCL1 and three with benign lung nodules as controls). We examined the coverage profiles around the same TFBSs as in the targeted panel and observed deep WGS coverage profiles that appeared visually similar to the targeted panel coverage profiles (Fig. 3.4a). We then quantified the central amplitude feature and found a strong correlation (Pearson r $\geq$0.98) between the targeted and deep WGS central amplitude features (Fig. 3.4b). We also examined the correlation between the deep WGS and targeted sequencing coverage values at individual sites and found that most site were well correlated with a typical Pearson r for an individual site of approximately 0.75 (Fig. 3.4c). This confirmed that our targeted panel was doing a reasonable job of capturing the coverage profiles at TFBSs. Next, we examined TSS coverage profiles in deep WGS. We looked at the coverage profiles around the TSS for the key TFs and observed visually similar profiles to the targeted panel and found that these were significantly correlated to the targeted data (for the median sample, Pearson r > 0.62, two sided p-value $\leq 2.24 \times 10^{-03}$, Fig. 3.4d). We also correlated the coverage profiles from targeted and deep for all individual TSS in all samples and found the typical correlation was very high (median Pearson r value > 0.76) (Fig. 3.4e). This confirms that the targeted TSS profiles are also a faithful representation of the deep WGS coverage profiles.

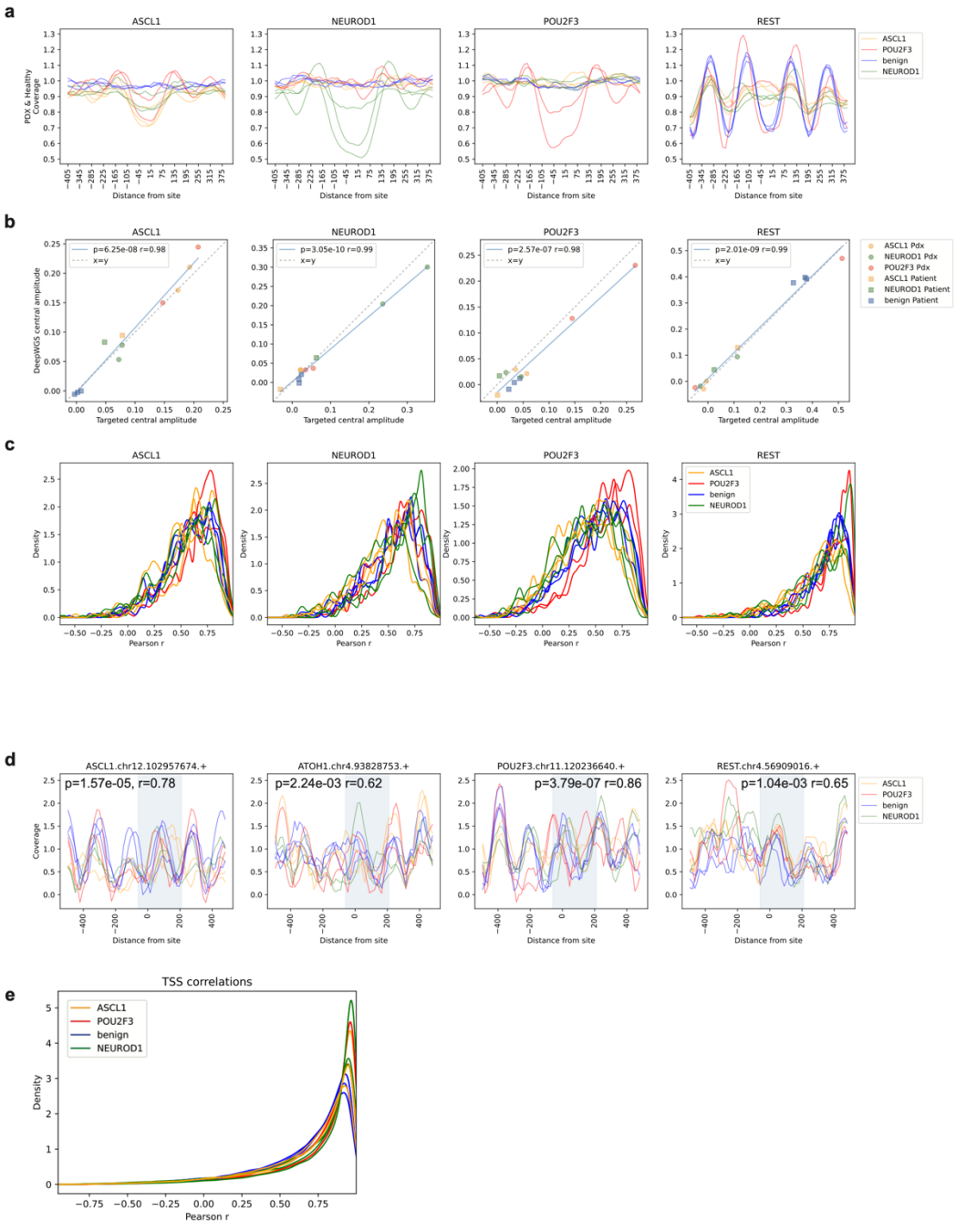


Figure 3.4 Comparison between targeted sequencing and deep WGS coverage profiles

Fig. 3.4 Comparison between targeted sequencing and deep WGS coverage profiles. (a) Coverage profiles from deep WGS around the same TFBSs as in the targeted panel. 11 samples with deep WGS are shown including 6 PDX models (2 ASCL1, 2 NEUROD1, 2 POU2F3) and 5 patient samples (3 patients with benign nodules, 1 with ASCL1 SCLC and 1 with NEUROD1 SCLC). Deep WGS TFBSs have

similar coverage profiles to targeted sequencing coverage profiles showing in Fig 2. **(b)** Pearson correlations (with two-sided p-value) between the TFBSs central amplitude feature in the deep whole genomes and the targeted panel data. **(c)** Distributions of the Pearson correlation coefficients between the targeted sequencing and deep WGS coverage profiles at individual TFBS in individual samples. Most individual sites have strong correlations between the targeted and deep WGS coverage profiles. **(d)** Deep WGS coverage profiles at TSS for key transcription factors. The area shaded in blue indicates the center of the region targeted by the panel (shown in figure 3). The coverage profiles in this region look like the targeted panel profiles. Median Pearson correlations between the targeted and deep WGS profiles for all 11 samples are listed on the plots. **(e)** Histogram of Pearson correlation coefficients between targeted and deep WGS coverage profiles at TSS shown for each of the 11 samples with deep WGS. Most TSS profiles are highly correlated.

3.2.5 *Nucleosome profiling of TSS and TFBSs can identify SCLC subtypes in patients*

After determining that both targeted TFBSs and TSS have differential coverage profiles in xenograft cfDNA, we set out to determine whether this was also true in patient samples. Unlike in xenografts, a significant portion of cfDNA in patients in derived from nucleated blood cells and this blood cell derived cfDNA has its own nucleosome profiles which can obscure subtype specific signals. Our patient samples included 24 SCLC patients with known subtypes from RNA or IHC. We also sequenced cfDNA from 5 LUAD patients, 4 LUSC patients, and 4 patients with benign nodules for comparison. Based on the tumor RNA expression, 9 of the SCLC patients had a clear ASCL1 subtype, 4 had a NEUROD1 subtype, and 1 had an ATOH1 subtype. However, the remaining 10 samples did not have a single dominant subtype and were considered ambiguous, mostly due to co-expression of ASCL1 and NEUROD1. We started by performing nucleosome profiling around the targeted TFBSs and observed increased accessibility around the active TFBSs although the signals around SCLC subtype specific TFs (ASCL1, NEUROD1, and REST) were much weaker due to the healthy blood cell contribution and stronger around REST sites which are accessible in blood cells. ASCL1, NEUROD1, and REST also had

significant correlation between the central amplitude feature and the RNA expression of a given TF, there were no POU2F3 subtype patients (Fig 5b). We next performed PCA on the TFBS coverage profiles from the xenograft samples and patients with benign nodules and found that the first two components separated the subtypes and the benign samples (Fig. 3.5c, left). We then projected the TFBS coverage features from the SCLC patients onto the same dimensions. The NEUROD1 patients fell roughly along the line between the NEUROD1 xenografts and the healthy patients while the ASCL1 patients fell roughly along the line between the ASCL1 xenografts and benign patients (Fig. 3.5c, right). This reflects the fact that the SCLC patient samples contain a mixture of cfDNA from healthy blood cells (like the patients with benign nodules) and tumor cells (like the xenografts). Next, we examined the Euclidean distance features around 73 targeted TSS with differential profiles as previously described in Fig. 3.3. Using hierarchical clustering, we were able to separate the NEUROD1 and ATOH1 samples from the ASCL1 samples (Fig. 3.5d). When we performed PCA on these same features, we also observed that the NEUROD1 samples clustered separately from the ASCL1 samples (Fig. 3.5e). This demonstrates that targeted panel sequencing of TFBSs and TSS is also able to distinguish the transcriptional subtypes in patients.

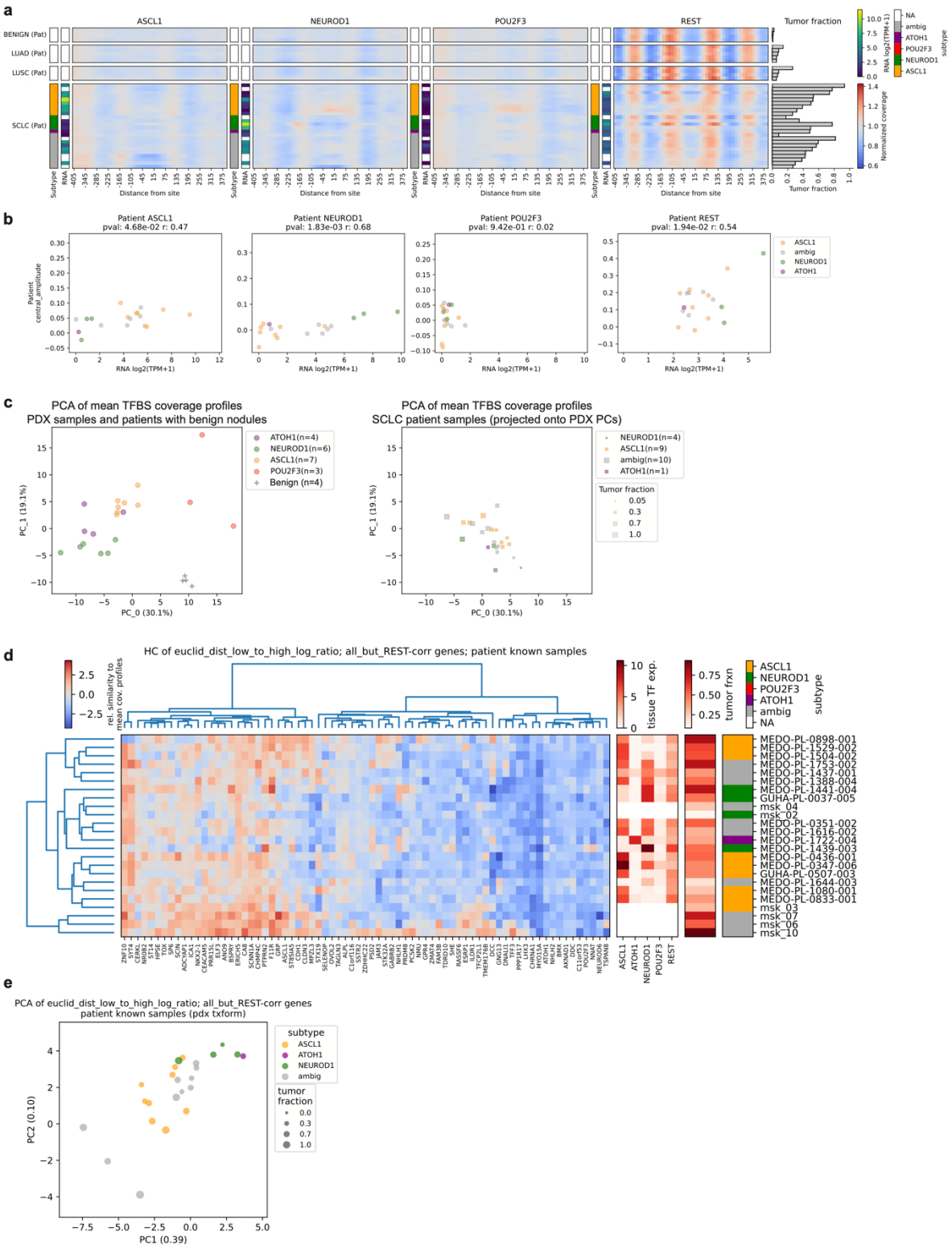


Figure 3.5 SCLC subtyping from TFBS targeted panel sites in patients

Fig. 3.5 SCLC subtyping from TFBS targeted panel sites in patients. (a) Heatmap of coverage profiles around targeted TFBS for each of 4 key transcription factors: ASCL1, NEUROD1, POU2F3 and REST. Blue indicates low coverage (due to TFBS accessibility) while red indicates high coverage (due to organized nucleosome protection). The fraction of cfDNA coming from the tumor (tumor fraction) is shown on the right side. Samples are grouped by histology. Top row shows 4 patients with benign lung nodules, second row shows patients with lung adenocarcinoma, and third row shows patients with lung squamous cell carcinoma. Bottom row shows SCLC patients with known phenotype (RNA seq or IHC available). For all factors, the TFBSs are more accessible when active. REST remains somewhat accessible in most SCLC patient samples due to contribution from healthy blood cells. **(b)** Correlation between central amplitude feature and TF expression for 18 patient samples with matched tumor RNA. Expression has been adjusted for the tumor fraction using GTEx whole blood values. **(c)** Left side, principal component analysis of the coverage profiles in SCLC PDX model cfDNA samples and patients with benign lung nodules. For each of the 4 coverage profiles, 54 features were extracted corresponding to the mean coverage value in 15 bp bins ranging from 405bp before the TF to 405bp after the TF. A PCA was then performed on all 216 features from the 4 profiles. The top two principal components separate the SCLC subtype groups and benign patient samples. Right side, projection of the SCLC patient samples onto the same PCA space as in (b), marker sizes are proportional to tumor fraction. ASCL1 samples fall roughly in the space between the benign samples and the NEUROD1 PDX models with lower tumor fraction samples being closer to the benign samples and higher tumor fraction samples being closer to the PDX models. The same trend is seen with NEUROD1 patient samples. Ambiguous subtype samples (which express both ASCL1 and NEUROD1) fall mostly in between. **(d)** Hierarchical clustering of TSS Euclidean distance features for 73 key genes with differential coverage profiles between TSS with high expression and low expression and expression correlation to one of the key transcription factors. The NEUROD1 and ATOH1 samples cluster together. **(e)** PCA of the Euclidean distance features for 73 genes. The top two principal components roughly separate the ASCL1 from the NEUROD1 samples. Marker sizes are proportional to tumor fraction.

3.2.6 *Nucleosome profiling of TSS and TFBSs can distinguish SCLC from NSCLC*

Finally, although our targeted panel was designed primarily to identify transcriptional subtypes of SCLC, we decided to test whether it could also be used to distinguish SCLC from non-small cell lung cancer (NSCLC). Some NSCLC tumors transition to SCLC tumors as a mechanism of therapeutic resistance so it would be beneficial to be able to distinguish these tumor types using a blood based assay. To determine whether there were differences in the targeted nucleosome profiles between SCLC and NSCLC tumors, we performed PCA on the SCLC and NSCLC PDX models and patients with benign nodules using the features from the targeted TFBS coverage profiles. We observed that the SCLC samples clustered separately from the NSCLC samples and patients with benign nodules which clustered together likely because they both have high REST accessibility and low accessibility at SCLC specific TFs (Fig. 3.6a). We then projected the patient features into the same space and observed that the NSCLC patients clustered with the NSCLC xenografts and benign patients and the SCLC patients fell on the line between the benign patients and SCLC xenografts, recapitulating how they are a mixture of blood and tumor derived cfDNA (Fig. 3.6b). Next, we identified a list of 520 TSS with differential expression between SCLC and NSCLC and differential TSS accessibility in the xenograft models. We then calculated the Euclidean distance feature for all xenograft samples and performed PCA on these features. We observed that the SCLC xenografts clustered separately from the LUAD xenografts with the LCNEC and LUSC in between (Fig. 3.6c). When we performed hierarchical clustering, we saw similar groupings (Fig. 3.6d). Next, we performed this same Euclidean distance analysis on the SCLC and NSCLC patients. We projected these features into the same feature space as the xenograft samples and observed that the SCLC and NSCLC patients fell in the same general areas as the SCLC and NSCLC xenografts respectively (Fig. 3.6e). When we performed hierarchical clustering on these same features, we observed that many of the SCLC patients clustered together, however a subset of lower tumor fraction SCLC patients clustered with the NSCLC patients (Fig. 3.6f). Overall, these analyses suggest that it is possible to distinguish SCLC from NSCLC using this panel.

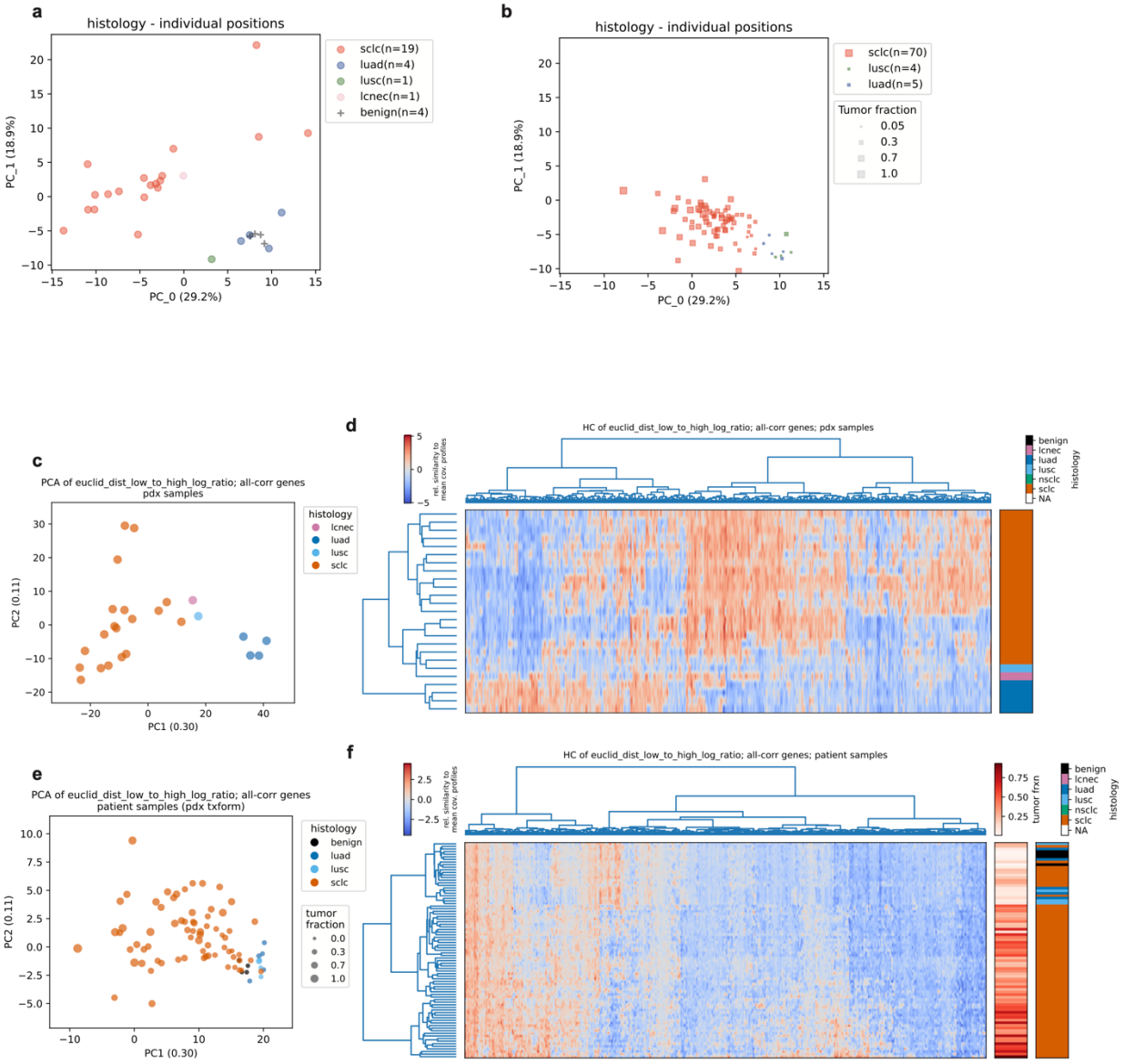


Figure 3.6 Histological subtyping in PDX models and patients

Fig. 3.6 Histological subtyping in PDX models and patients. (a) Principal component analysis of the TFBS coverage profiles in SCLC and NSCLC PDX model cfDNA samples and patients with benign lung nodules. For each of the 4 coverage profiles, 54 features were extracted corresponding to the mean coverage value in 15 bp bins ranging from 405bp before the TF to 405bp after the TF. A PCA was then performed on all 216 features from the 4 profiles. The top two principal components separate the SCLC and LCNEC PDX models from the NSCLC PDX models and benign patients. **(b)** Projection of the patient

sample TFBS coverage profiles onto the same PCA space as in (a). Marker sizes are proportional to tumor fraction. The SCLC patients fall in the space between the non-malignant patient/NSCLC PDX model cluster and the SCLC PDX model cluster with higher tumor fraction samples being generally closer to the SCLC PDX cluster. The NSCLC patient samples fall in the same space as the NSCLC PDX model and benign patients. **(c)** PCA analysis of the Euclidean distance features for 520 TSS coverage profiles with significantly different expression between SCLC and NSCLC and significantly differential coverage profiles when active and inactive. SCLC and NSCLC PDX models are shown. Lung adenocarcinoma samples (LUAD) cluster separately from the other histological subtypes. **(d)** Hierarchical clustering of the same TSS Euclidean distance features. LUAD samples cluster separately from the other samples. **(e)** Projection of the Euclidean distance TSS features from the patient samples onto the same feature space as in (c). Marker sizes are proportional to tumor fraction. SCLC samples generally cluster separately from NSCLC samples. **(f)** Hierarchical clustering of the same TSS Euclidean distance features. Tumor fraction and histology are shown on the right.

3.3 Discussion

In this study, we developed a targeted sequencing panel designed specifically for genotypic and phenotypic characterization of SCLC. The panel contained exons for identifying mutations in key SCLC-related genes, TSS for gene expression prediction from nucleosome profiling, and TFBSs for nucleosome profiling of key subtype specific TF accessibility. We applied this panel to SCLC and NSCLC cfDNA from xenograft models and patients. We demonstrated that it is possible to perform nucleosome profiling using targeted panel sequencing around TFBSs and that a relatively small number of sites (~500 per TF) is sufficient to observe differential accessibility signals in lineage defining SCLC transcription factors. We also examined signals around TSS and found that there were differential profiles around lineage defining TSS but these profiles did not have a consistent shape across TSS. We developed a Euclidean distance metric to quantify the similarity of a given TSS profile to active profiles for that TSS and showed

that this can be used to differentiate subtypes. Lastly, we showed that our panel can distinguish SCLC from NSCLC samples in both xenografts and patients.

The TFBS panel demonstrates proof of concept that TFBS accessibility signals can be captured using a targeted panel sequencing approach. Nucleosome profiling around TFBSs has shown promise for being able to predict the expression of important transcription factors in cancer but has previously required deep whole genome sequencing. A targeted panel approach could provide a cost-effective way to study cfDNA TFBS accessibility in large numbers of samples in order to better understand the variation between samples. Additionally, it provides sufficient depth to study the coverage profiles at individual sites which is not possible using other low-cost sequencing approaches like ultra-low pass whole genome sequencing. Eventually, nucleosome profiling from cfDNA could be integrated into existing clinical targeted panels for cfDNA characterization, enabling both phenotype and genotype determination from a single non-invasive panel.

The transcriptional subtypes of SCLC have only become well defined in recent years[51] and research is ongoing to better understand the distinctions between subtypes, especially in patients, and to define additional subtypes[142–145,153]. Here we show that there are observable differences in cfDNA nucleosome profiles between subtypes, a first step towards non-invasive subtype determination in the clinic. Further development of such approaches could enable subtype determination in larger cohorts of patients, potentially allowing researchers to better study subtypes and subtype specific treatment responses in-vivo.

At least two prior studies have performed nucleosome profiling using targeted panel sequencing around TSS[41,42]. Like these studies, we saw a similar global pattern around active TSS with a loss of coverage around the TSS itself and a peak of coverage at +1 nucleosomes. However, we noted that these mean TSS profiles may be heavily influenced by constitutively expressed or unexpressed genes and other genes that are not relevant to cancer. Because of this we focused more on the shapes of individual TSS and found that these varied considerably between different TSS, which is consistent with prior studies of nucleosome positioning[154]. Future studies with larger numbers of samples and more varied phenotypes

and expression patterns could help better understand these coverage profiles in cfDNA and identify better ways to predict gene expression from coverage profiles in genes with relevance in cancer. We also examined TSS fragment length entropy which had been previously shown to be correlated to gene expression in the EPIC-seq study[42]. Similar to this study, we saw a correlation with gene expression although, we didn't observe that this feature correlated better than coverage at the TSS. This may be due to the shallower sequencing depth in our study. Additionally, we demonstrated that nucleosome profiles around TSS can be observed even with a very small targeted region (360bp) which allow sequencing of TSS from most genes without needing a large, expensive panel.

Our study was limited by the relatively small number of samples we had for each subtype with only 3-7 xenograft models per subtype and even fewer patient samples with clear cut subtypes. Larger sample sets might enable us to better identify differential sites, especially those with strong signal in patients, and design better panels in the future. Additionally, our TSS method is reliant on having examples of the TSS in both active and inactive samples. A larger sample set and additional types of data such as ATAC seq and ChIP-seq for active and inactive histone marks might help us better understand the coverage profiles we are observing and build better models that don't need to be based on active and inactive samples. Additionally, although the small targeted window around each TSS allowed us to target TSS for almost all genes, a larger window could have been helpful to better observe the positions of multiple nucleosomes and compare TSS coverage against the surrounding coverage. Additionally, our analyses were confined to patient samples with at least 5% tumor fraction. Although this is common in SCLC, not all patients pass this threshold. In this study 70 of 84 (83%) of SCLC patient samples that passed sequencing QC metrics had at least 5% tumor fraction. Additionally, when comparing SCLC samples to NSCLC samples, the difference in tumor fraction between these two groups lead to large differences in signal which made clustering less informative because low tumor fraction SCLC samples tended to cluster with the NSCLC samples.

This work provides a proof of concept for using targeted panel sequencing of cfDNA to characterize small-cell lung cancer. Future studies could expand on this approach by including more

TFBSs of interest including NSCLC specific TFBSs to better distinguish these patients from healthy individuals. Additionally, panels could also be built for other cancer types such as prostate and breast cancer which also have differentially accessible sites. The TSS targeted panel could be reduced to focus on pan-cancer genes of interest rather than all TSS. This would allow targeting larger windows around each TSS without greatly increased sequencing costs. Finally, additional deep whole genome sequencing in SCLC would be beneficial in order to better identify differential TFBSs and TSS for future studies.

Overall, we have demonstrated that a targeted sequencing panel can be used to perform both genotyping and phenotyping in small-cell lung cancer PDX models and patients. We have shown that transcriptional subtypes have distinct accessibility profiles and that these can be used to distinguish subtypes. Additionally, we have demonstrated that SCLC and NSCLC can be distinguished from one another using targeted panel sites. Overall, this proof-of-concept study lays the groundwork for future development of non-invasive clinical cfDNA assays to diagnose and characterize SCLC. Non-invasive diagnostic assays could improve our understanding of tumor phenotypes in the clinic, which has previously been difficult to study due to the limited availability of tumor tissue from SCLC patients. In the longer term, this could aid in the development of targeted treatments to improve and extend patients' lives.

3.4 Methods

Targeted capture panel design

Gene selection

A panel of 862 genes was assembled manually using the following sources: (1) genes that are frequently mutated in SCLC according to publicly available databases (cBioPortal, Project GENIE), (2) cancer mutated genes from the 2019 TARGET study[155] and (3) genes that have been found in functional screens for tumor suppressor activity performed in the MacPherson lab[156].

TSS site selection

A list of human TSS from the Ensembl v97 annotation corresponding to human genome version GRCh38.p12 was downloaded from Ensembl using the BioMart tool. Gene annotations from the Gencode v31 "Basic" gene annotation were downloaded from the UCSC genome browser. These two annotations were then merged based on the Ensembl "Transcript stable ID" field. The following filters were then applied. TSSs residing on alternative contigs or chromosome Y were excluded. TSSs corresponding to transcripts not labeled "protein_coding" or "retained_intron" were removed. TSSs were then required to have at least one of the following properties: (1) transcript support level equal to 1 according to Gencode v31, (2) a single exon gene, (3) association with differential expression of the SCLC TFs ASCL1, NEUROD1, POU2F3, or REST according to lists of TF-associated genes that were published or calculated from published data, or (4) presence in the MSigDB v7.0 Hallmark pathways list. This resulted in a list of 36,379 TSSs corresponding to 18,030 unique genes. Sites were padded by 100 bp upstream and 260 bp downstream and these intervals were submitted for capture panel design. The resulting probe set adequately targeted (defined as continuous probe coverage from 90 bp upstream of the TSS to 245 bp downstream) a total of 35,917 TSSs corresponding to 17,921 genes. To facilitate gene expression prediction benchmarking and interpretation, we then used CCLE transcript-specific gene expression data (from the 2021Q1 CCLE data repository) to select a single TSS per gene associated with the most highly expressed transcript. Sites were further filtered for the ability to confidently map sequencing reads, resulting in a list of 14,062 TSSs corresponding to the same number of unique genes that were subjected to downstream analyses.

TFBS selection

Curated TFBS lists were generated individually for the SCLC TFs ASCL1, NEUROD1, POU2F3, and REST as follows. For ASCL1, NEUROD1, and POU2F3, published ChIP-seq peaks and differentially expressed gene lists (Borromeo et al, Cell Reports, 2016 for ASCL1/NEUROD1 and Huang et al, Genes Dev., 2018 for POU2F3) were intersected to generate a list of TFBSs associated with differentially expressed genes. TFBSs were defined as associated with a differentially expressed gene if they were

either within 10 kb of a differentially expressed gene TSS, or if the closest gene TSS was differentially expressed. For REST, CCLE SCLC cell line gene expression data (2019 Q2 dataset) was mined to create a coarse list of REST-associated differentially expressed genes. REST TFBSs were then extracted from the GTRD database and intersected with the top 250 most differentially expressed genes to generate a list of TFBSs for capture panel design.

Mouse subtraction and alignment

Sequenced reads were aligned using bwa-mem to a concatenated human and mouse reference genome consisting of hg38 and mm10. Read pairs where both reads are aligned to the human genomes are retained for further processing[151]. The retained reads were the realigned to the human genome using bwa-mem, duplicates were marked using picard, and base qualities were recalibrated using GATK. Picard CollectHsMetrics were used to get the on-target coverage for quality control.

ichorCNA tumor fraction

IchorCNA was used with default settings to estimate tumor fraction. The highest ranked estimate was used without manual curation.

Griffin analysis

Griffin nucleosome profiling was performed as previously described in the Griffin manuscript (Chapter 2) with a few modifications. First, when running GC correction, only positions that were on-target were considered when counting reads and GC contents. Separate GC corrections were run for the TSS targeted panel and the TFBS plus exon targeted panel. 50bp read mappability was used when identifying mappable positions (the shortest read lengths in this study were 50bp). When performing nucleosome profiling, coverage profiles were calculated in the window ±500bp from a targeted TFBS or a -100 to +240 from targeted TSS. Sites that were not fully covered by the targeted panel were excluded from further analysis. Coverage outliers were not excluded using the 'exclude_outliers' option and individual sites were saved.

REST sites alignment

REST motifs in the REST TFBSs were identified using motifmatcheR to search for instances of the HOCOMOCO REST motif from CIS-BP. 286 sites were found to have at least one REST motif. Coverage profiles were then aligned so that the motif (or the highest scoring motif if there was more than one) was in the center. Then the mean coverage profile was calculated.

Mutation calling

Plasma samples corresponding to patients for whom buffy coat (PBMC) sample was also available were subjected to variant calling. Targeted sequencing data from the plasma sample and buffy coat sample were processed using the variant callers Mutect2, Strelka, and CNVkit. Germline variation databases such as gnomAD were used to boost germline variant removal. SNVs and small indels were considered high confidence if they passed default filters for both Mutect2 and Strelka. CNV calling was limited to the autosomes. DUX4, which was recurrently called as amplified in a large fraction of samples, was excluded as a likely artifact.

TFBS central amplitude feature

The central amplitude was calculated as the mean coverage at the two peaks flanking the TFBSs (-150 to -75 bp and 75 to 150bp), minus the coverage at the TFBSs (-30 to 30bp)

CpG adjacent definitions

CpG island locations were downloaded from the UCSC table browser. The distance from each TSS to the nearest CpG island was calculated using BedTools. TSSs were considered to be CpG proximal if the nearest CpG island was <1 kb away.

Entropy measurement

At each TSS window in each sample, the distribution of the length of aligned fragments <500 bp was binned into 5 bp windows. The entropy of these binned distributions was then calculated using the scipy.stats.entropy function. Entropy scores were then Z score normalized within each individual sample.

Euclidean distance feature

For each TSS of interest, using the coverage profiles from the PDX models, we performed k-means clustering with k=2. For each sample, the Euclidean distance of the coverage profile at that TSS to the mean coverage profile of each of the two clusters was also calculated, and a single metric of relative distance was calculated as the log2 of the ratio of the distance of the coverage profile to the low expression cluster mean over the distance of the coverage profile to the high expression cluster mean. For each site, we also then tested whether the clustering yielded two distinct expression states by performing a Mann-Whitney U test of the log2(TPM+1) expression values for the gene of interest in the samples in each of the two clusters. Sites were then filtered based on the Mann-Whitney U test p-values to identify sites where the coverage profiles significantly discriminated expression in the PDX models, and which might be useful for SCLC subtype discrimination or histology prediction.

TSS amplitude measurement

TSS amplitude was defined as the difference between the relative coverage value at the +120 position and the -45 position relative to the TSS.

Identification of subtype specific gene sets (3F)

SCLC cell line gene expression data from CCLE (2021 Q2 dataset) was analyzed for the correlation of individual gene expression values with the expression of the key SCLC TFs ASCL1, ATOH1, NEUROD1, and POU2F3. TF-associated genes were defined as genes with a Pearson correlation coefficient of at least 0.6 with a given TF.

Chapter 4. Summary and future directions

4.1 Summary

In summary, this thesis describes the development of Griffin, a method for nucleosome profiling that incorporates fragment-based GC correction and is optimized for ultra-low pass whole genome sequencing. It also describes the application of Griffin for prediction of ER activity in MBC and prediction of transcriptional subtype in SCLC. This research has helped to expand the applications of nucleosome profiling from cfDNA to new cancer types and to lower coverage data, helping to pave the way for future clinical studies.

First, we developed the Griffin framework for nucleosome profiling. This work addressed the goal of studying and correcting for the impact of GC bias on cfDNA coverage profiles to better predict chromatin accessibility from low coverage WGS. By examining the GC content around TFBSs and the GC bias in hundreds of cfDNA samples, we found that correcting for GC bias reduced inter-sample variability and improved correlations to tumor fraction, especially for TFBSs with large differences in average GC content between the TFBSs and surrounding sequence. This in-depth examination of GC bias highlighted the importance of GC correction in nucleosome profiling of cfDNA. We then built the Griffin pipeline for nucleosome profiling, which incorporated a fragment-based GC correction algorithm. The Griffin pipeline was optimized for ULP-WGS, a scalable cost-effective form of sequencing. However, the Griffin pipeline is not limited to ULP-WGS sequencing, it can take bam files with any depth and can also be used on targeted panel sequencing. Additionally, the Griffin pipeline allows flexible inputs of any list of sites, allowing site selection from any type of assay. Overall, this pipeline provides a flexible tool that can be used by future researchers for nucleosome profiling studies.

Griffin was designed to take flexible site inputs, and this allowed us to address our second goal of characterizing sites from different selection assays and determining the best sites for specific applications. In particular, we examined whether differential ATAC-seq sites were able to differentiate cfDNA from

patients with ER+ and ER- breast cancer. We found that these sites were better able to distinguish these MBC subtypes than TFBSs from ChIP-seq. The finding that differential ATAC-seq sites performed well provides a strategy for the development of new assays to distinguish tumor subtypes from cfDNA in a variety of cancer types, even when the subtypes aren't driven by known transcription factors.

Our third goal was to apply nucleosome profiling to new cancer types and subtypes in order to expand the range of cancer patients that could benefit from this non-invasive technology. To address this, we first demonstrated that nucleosome profiling can be used to predict ER status in MBC. ER status is important for both selecting targeted treatment and predicting prognosis[44], so the ability to predict it from cfDNA could provide benefit to patients. Secondly, we demonstrated that nucleosome profiling can be used to distinguish transcriptional subtypes in SCLC. These subtypes have only been described recently and there is limited clinical data about the prevalence of these subtypes in patient tumors or their clinical significance[51]. This is largely because SCLC is often detected when it has already progressed to an advanced stage and tumor biopsies are not routine standard of care[157]. Because of this, a non-invasive assay to predict subtype from cfDNA would enable better characterization of the subtypes in the clinic when tumor tissue isn't available. Our proof-of-concept study demonstrated that SCLC subtypes can be determined from cfDNA, paving the way for larger future clinical studies.

4.2 Future directions

The development of the Griffin pipeline and demonstration of tumor subtyping from cfDNA in MBC and SCLC are first steps in several exciting new areas of research on non-invasive phenotyping from cfDNA. However, much work remains to be done before these technologies will reach the clinic and impact patients' lives.

First, we found that differential ATAC sites were able to distinguish ER status in MBC, however this was only possible in patients with at least 5% tumor fraction, which excludes many patients with early stage and more treatable disease. ATAC-seq sites performed better than TFBSs but additional research is needed to further define the best sites for subtyping in this cancer type and others, and to

optimize this method to perform well on lower tumor fraction samples. One way to achieve this could be to identify sites directly from cfDNA, which would enable researchers to discover the sites that are most differential in cfDNA itself and most likely to have clear signals in patients with low tumor fraction. This will require more patient samples with well characterized tumors as well as deeper sequencing of cfDNA so that analysis can be performed directly on this data type.

Secondly, the demonstration of ER status prediction from cfDNA is an exciting proof-of-concept but additional studies are needed to determine whether this approach will work in the clinic, especially in patients with ER loss or heterogeneous tumors. Although we found evidence that some of these misclassifications may have been due to tumor heterogeneity within individual patients, it is also possible that there are differences in chromatin structure between primary ER negative tumors and those with ER loss which contributed to the misclassifications of these samples. More studies are needed to explore the cause of these misclassifications. Our study was limited by a relatively small number of patients (n=9) with known ER loss. Future studies including more patients with ER loss could better examine the nucleosome profiles in these patients and determine whether there are differences between nucleosome profiles in ER-primary and ER loss tumors. These profiles could then be used to build classifiers specifically to predict ER loss in the context of an ER+ primary. These sorts of classifiers could be useful in the clinic because patients with metastatic recurrence don't always undergo metastatic biopsy and may benefit from treatment tailored to the metastatic subtype. Additionally, more studies are needed to confirm the contribution of heterogeneous tumors to cfDNA. Although we observed fluctuations in the predicted ER status in several patients who had metastatic biopsies with both ER+ and ER- subtypes, our study was not able to confirm the source of the cfDNA or definitively determine whether these metastases were driving the changes in predicted ER status. Future studies with deeper sequencing of cfDNA and sampling of multiple metastases for both phenotypic and genotypic characterization could be used to determine whether the phenotypes predicted from cfDNA match the phenotypes of the contributing metastases. These studies could also help clarify whether cfDNA provides a more complete picture of tumor heterogeneity within a patient than individual tumor biopsies. If future studies demonstrate that

cfDNA can reliably predict tumor phenotypes and heterogeneity in MBC, prospective clinical studies could be designed to assess whether selecting treatment from cfDNA predicted subtype leads to better outcomes.

In SCLC, we demonstrated that transcriptional subtypes have different coverage profiles around key TFBSs and TSS, however more work is needed to link these subtypes to treatment responses or prognosis. Future studies with larger numbers of patients could be used to develop better subtype classifiers from cfDNA. Additionally, collecting cfDNA from patients enrolled in clinical trials and performing nucleosome profiling on these samples would enable correlation of subtypes with prognosis and treatment response, even in the absence of tumor tissue.

4.3 Concluding remarks

Ultimately, cfDNA has the potential to be used for non-invasive cancer characterization. Tumor genotyping from cfDNA is already available in the clinic but tumor phenotyping is at an earlier stage of development. Tumor tissue is still necessary for assays such as histology and immunohistochemistry to determine tumor phenotypes. However, new studies, including those described in this thesis are beginning to demonstrate the possibility of predicting tumor phenotypes directly from cfDNA. This raises the exciting possibility that one day cfDNA may be able to replace tumor biopsies for both genotyping and phenotyping. This would allow more frequent monitoring of tumors with fewer invasive procedures, allowing earlier administration of targeted treatments and leading to better outcomes for patients. Although there remains much work to be done before this potential is realized, this thesis and other studies have clearly demonstrated that nucleosome profiling of cfDNA is a promising approach for non-invasive cancer diagnostics with the potential to transform patients' lives.

References

1. Wan, J. C. M. *et al.* Liquid biopsies come of age: Towards implementation of circulating tumour DNA. *Nature Reviews Cancer* **17**, 223–238 (2017).

2. Leon, S. A., Shapiro, B., Sklaroff, D. M. & Yaros, M. J. Free DNA in the serum of cancer patients and the effect of therapy. *Cancer Res* **37**, 646–650 (1977).

3. Lui, Y. Y. N. *et al.* Predominant hematopoietic origin of cell-free dna in plasma and serum after sex-mismatched bone marrow transplantation. *Clinical Chemistry* **48**, 421–427 (2002).

4. Sun, K. *et al.* Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proceedings of the National Academy of Sciences* **112**, E5503–E5512 (2015).

5. Stroun, M. *et al.* Neoplastic characteristics of the DNA found in the plasma of cancer patients. *Oncology* **46**, 318–322 (1989).

6. Sorenson, G. D. *et al.* Soluble normal and mutated DNA sequences from single-copy genes in human blood. *Cancer Epidemiol Biomarkers Prev* **3**, 67–71 (1994).

7. Jahr, S. *et al.* DNA Fragments in the Blood Plasma of Cancer Patients: Quantitations and Evidence for Their Origin from Apoptotic and Necrotic Cells1. *Cancer Research* **61**, 1659–1665 (2001).

8. Watanabe, T., Takada, S. & Mizuta, R. Cell-free DNA in blood circulation is generated by DNase1L3 and caspase-activated DNase. *Biochemical and Biophysical Research Communications* **516**, 790–795 (2019).

9. Han, D. S. C. & Lo, Y. M. D. The Nexus of cfDNA and Nuclease Biology. *Trends in Genetics* **37**, 758–770 (2021).

10. Rostami, A. *et al.* Senescence, Necrosis, and Apoptosis Govern Circulating Cell-free DNA Release Kinetics. *Cell Reports* **31**, (2020).

11. Heitzer, E., Auinger, L. & Speicher, M. R. Cell-Free DNA and Apoptosis: How Dead Cells Inform About the Living. *Trends in Molecular Medicine* **26**, 519–528 (2020).

12. Kahlert, C. *et al.* Identification of double-stranded genomic DNA spanning all chromosomes with mutated KRAS and p53 DNA in the serum exosomes of patients with pancreatic cancer. *J Biol Chem* **289**, 3869–3875 (2014).

13. Stroun, M., Lyautey, J., Lederrey, C., Olson-Sand, A. & Anker, P. About the possible origin and mechanism of circulating DNA apoptosis and active DNA release. *Clin Chim Acta* **313**, 139–142 (2001).

14. Abolhassani, M., Tillotson, J. & Chiao, J. Characterization of the release of DNA by a human leukemia-cell line hl-60. *Int J Oncol* **4**, 417–421 (1994).

15. Cheng, T. H. T. *et al.* DNase1 Does Not Appear to Play a Major Role in the Fragmentation of Plasma DNA in a Knockout Mouse Model. *Clinical Chemistry* **64**, 406–408 (2018).

16. Serpas, L. *et al.* Dnase1l3 deletion causes aberrations in length and end-motif frequencies in plasma DNA. *Proceedings of the National Academy of Sciences of the United States of America* **116**, 641–649 (2019).

17. Lo, Y. M. D., Han, D. S. C., Jiang, P. & Chiu, R. W. K. Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies. *Science* **372**, eaaw3616 (2021).

18. Han, D. S. C. *et al.* The Biology of Cell-free DNA Fragmentation and the Roles of DNASE1, DNASE1L3, and DFFB. *The American Journal of Human Genetics* **106**, 202–214 (2020).

19. Diehl, F. *et al.* Detection and quantification of mutations in the plasma of patients with colorectal tumors. *Proceedings of the National Academy of Sciences* **102**, 16368–16373 (2005).

20. Cohen, J. D. *et al.* Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* **359**, 926–930 (2018).

21. Lennon, A. M. *et al.* Feasibility of blood testing combined with PET-CT to screen for cancer and guide intervention. *Science* **9601**, eabb9601 (2020).

22. Crowley, E., Di Nicolantonio, F., Loupakis, F. & Bardelli, A. Liquid biopsy: monitoring cancer-genetics in the blood. *Nature reviews. Clinical oncology* **10**, 472–84 (2013).

23. Kimura, H. *et al.* Detection of Epidermal Growth Factor Receptor Mutations in Serum as a Predictor of the Response to Gefitinib in Patients with Non–Small-Cell Lung Cancer. *Clinical Cancer Research* **12**, 3915–3921 (2006).

24. Ignatiadis, M., Sledge, G. W. & Jeffrey, S. S. Liquid biopsy enters the clinic - implementation issues and future challenges. *Nat Rev Clin Oncol* **18**, 297–312 (2021).

25. Garcia-Murillas, I. *et al.* Assessment of Molecular Relapse Detection in Early-Stage Breast Cancer. *JAMA Oncol* **5**, 1473–1478 (2019).

26. Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol* **30**, 413–421 (2012).

27. Ha, G. *et al.* TITAN: Inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Research* (2014) doi:10.1101/gr.180281.114.

28. Adalsteinsson, V. A. *et al.* Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nature Communications* **8**, 1324 (2017).

29. Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M. & Shendure, J. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell* **164**, 57–68 (2016).

30. Jiang, P. *et al.* Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. *Proc Natl Acad Sci U S A* **112**, E1317-1325 (2015).

31. Mouliere, F. *et al.* High Fragmentation Characterizes Tumour-Derived Circulating DNA. *PLOS ONE* **6**, e23418 (2011).

32. Underhill, H. R. *et al.* Fragment Length of Circulating Tumor DNA. *PLoS Genetics* **12**, 1–24 (2016).

33. Mouliere, F. *et al.* Enhanced detection of circulating tumor DNA by fragment size analysis. *Science Translational Medicine* **10**, eaat4921 (2018).

34. Chabon, J. J. *et al.* Integrating genomic features for non-invasive early lung cancer detection. *Nature* **580**, 245–251 (2020).

35. Cristiano, S. *et al.* Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature* **570**, 385–389 (2019).

36. Mathios, D. *et al.* Detection and characterization of lung cancer using cell-free DNA fragmentomes. *Nat Commun* **12**, 5060 (2021).

37. Sun, K. *et al.* Size-tagged preferred ends in maternal plasma DNA shed light on the production mechanism and show utility in noninvasive prenatal testing. *Proc Natl Acad Sci U S A* **115**, E5106–E5114 (2018).

38. Shi, J., Zhang, R., Li, J. & Zhang, R. Size profile of cell-free DNA: A beacon guiding the practice and innovation of clinical testing. *Theranostics* **10**, 4737–4748 (2020).

39. Jiang, P. *et al.* Plasma DNA end-motif profiling as a fragmentomic marker in cancer, pregnancy, and transplantation. *Cancer Discovery* **10**, 664–673 (2020).

40. Ulz, P. *et al.* Inferring expressed genes by whole-genome sequencing of plasma DNA. *Nature genetics* **48**, 1273–8 (2016).

41. Zhu, G. *et al.* Tissue-specific cell-free DNA degradation quantifies circulating tumor DNA burden. *Nature Communications* **12**, 2229 (2021).

42. Esfahani, M. S. *et al.* Inferring gene expression from cell-free DNA fragmentation profiles. *Nat Biotechnol* **40**, 585–597 (2022).

43. Chung, C. & Abboud, K. Targeting the androgen receptor signaling pathway in advanced prostate cancer. *American Journal of Health-System Pharmacy* **79**, 1224–1235 (2022).

44. Rivenbark, A. G., O'Connor, S. M. & Coleman, W. B. Molecular and cellular heterogeneity in breast cancer: Challenges for personalized medicine. *American Journal of Pathology* **183**, 1113–1124 (2013).

45. Ulz, P. *et al.* Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection. *Nature Communications* **10**, 4666 (2019).

46. Herberts, C. *et al.* Deep whole-genome ctDNA chronology of treatment-resistant prostate cancer. *Nature* (2022) doi:10.1038/s41586-022-04975-9.

47. Sun, K. *et al.* Orientation-aware plasma cell-free DNA fragmentation analysis in open chromatin regions informs tissue of origin. *Genome Research* **29**, 418–427 (2019).

48. Peneder, P. *et al.* Multimodal analysis of cell-free DNA whole-genome sequencing for pediatric cancers with low mutational burden. *Nature Communications* **12**, 3230 (2021).

49. Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research* **40**, e72–e72 (2012).

50. Wang, J. *et al.* Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res* **22**, 1798–1812 (2012).

51. Rudin, C. M. *et al.* Molecular subtypes of small cell lung cancer: a synthesis of human and mouse model data. *Nature Reviews Cancer* **19**, 289–297 (2019).

52. Heitzer, E., Auinger, L. & Speicher, M. R. Cell-Free DNA and Apoptosis: How Dead Cells Inform About the Living. *Trends in Molecular Medicine* **26**, 519–528 (2020).

53. Diehl, F. *et al.* Circulating mutant DNA to assess tumor dynamics. *Nature medicine* **14**, 985–90 (2008).

54. Maheswaran, S. *et al.* Detection of mutations in EGFR in circulating lung-cancer cells. *The New England journal of medicine* **359**, 366–77 (2008).

55. Wan, J. C. M. *et al.* Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nature Reviews Cancer* **17**, 223–238 (2017).

56. Cohen, J. D. *et al.* Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science (New York, N.Y.)* **359**, 926–930 (2018).

57. McDonald, B. R. *et al.* Personalized circulating tumor DNA analysis to detect residual disease after neoadjuvant therapy in breast cancer. *Science Translational Medicine* **11**, eaax7392 (2019).

58. Parsons, H. A. *et al.* Sensitive detection of minimal residual disease in patients treated for early-stage breast cancer. *Clinical Cancer Research* clincanres.3005.2019 (2020) doi:10.1158/1078-0432.ccr-19-3005.

59. Murtaza, M. *et al.* Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. *Nature* **497**, 108–112 (2014).

60. Zviran, A. *et al.* Genome-wide cell-free DNA mutational integration enables ultra-sensitive cancer monitoring. *Nat Med* **26**, 1114–1124 (2020).

61. Adalsteinsson, V. A. *et al.* Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nature Communications* **8**, (2017).

62. Stover, D. G. *et al.* Association of Cell-Free DNA Tumor Fraction and Somatic Copy Number Alterations With Survival in Metastatic Triple-Negative Breast Cancer. *Journal of Clinical Oncology* JCO.2017.76.003 (2018).

63. Choudhury, A. D. *et al.* Tumor fraction in cell-free DNA as a biomarker in prostate cancer. *JCI Insight* **3**, (2018).

64. Sumanasuriya, S. *et al.* Elucidating Prostate Cancer Behaviour During Treatment via Low-pass Whole-genome Sequencing of Circulating Tumour DNA. *European Urology* **80**, 243–253 (2021).

65. Wyatt, A. W. *et al.* Concordance of Circulating Tumor DNA and Matched Metastatic Tissue Biopsy in Prostate Cancer. *JNCI: Journal of the National Cancer Institute* **110**, 78–86 (2018).

66. Viswanathan, S. R. *et al.* Structural Alterations Driving Castration-Resistant Prostate Cancer Revealed by Linked-Read Genome Sequencing. *Cell* **174**, 433-447.e19 (2018).

67. Beltran, H. *et al.* Divergent clonal evolution of castration-resistant neuroendocrine prostate cancer. *Nature Medicine* **22**, 298–305 (2016).

68. Bluemn, E. G. *et al.* Androgen Receptor Pathway-Independent Prostate Cancer Is Sustained through FGF Signaling. *Cancer cell* **32**, 474-489.e6 (2017).

69. Aggarwal, R. *et al.* Clinical and Genomic Characterization of Treatment-Emergent Small-Cell Neuroendocrine Prostate Cancer: A Multi-institutional Prospective Study. *JCO* **36**, 2492–2503 (2018).

70. Quintanal-Villalonga, A. *et al.* Multi-omic analysis of lung tumors defines pathways activated in neuroendocrine transformation. *Cancer Discov* (2021) doi:10.1158/2159-8290.CD-20-1863.

71. Niederst, M. J. *et al.* RB loss in resistant EGFR mutant lung adenocarcinomas that transform to small-cell lung cancer. *Nat Commun* **6**, 6377 (2015).

72. Van Poznak, C. *et al.* Use of Biomarkers to Guide Decisions on Systemic Therapy for Women With Metastatic Breast Cancer: American Society of Clinical Oncology Clinical Practice Guideline. *JCO* **33**, 2695–2704 (2015).

73. Bianchini, G., Balko, J. M., Mayer, I. A., Sanders, M. E. & Gianni, L. Triple-negative breast cancer: challenges and opportunities of a heterogeneous disease. *Nat Rev Clin Oncol* **13**, 674–690 (2016).

74. McAnena, P. F. *et al.* Breast cancer subtype discordance: impact on post-recurrence survival and potential treatment options. *BMC Cancer* **18**, 203 (2018).

75. Hulsbergen, A. F. C. *et al.* Subtype switching in breast cancer brain metastases: a multicenter analysis. *Neuro-Oncology* **22**, 1173–1181 (2020).

76. Schrijver, W. A. M. E. *et al.* Receptor Conversion in Distant Breast Cancer Metastases: A Systematic Review and Meta-analysis. *JNCI: Journal of the National Cancer Institute* **110**, 568–580 (2018).

77. Lindström, L. S. *et al.* Clinically used breast cancer markers such as estrogen receptor, progesterone receptor, and human epidermal growth factor receptor 2 are unstable throughout tumor progression. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **30**, 2601–8 (2012).

78. Aurilio, G. *et al.* A meta-analysis of oestrogen receptor, progesterone receptor and human epidermal growth factor receptor 2 discordance between primary breast cancer and metastases. *European Journal of Cancer* **50**, 277–289 (2014).

79. Hoefnagel, L. D. C. *et al.* Receptor conversion in distant breast cancer metastases. *Breast cancer research : BCR* **12**, R75 (2010).

80. Shah, S. P. *et al.* The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486**, 395–9 (2012).

81. Lindström, L. S. *et al.* Intratumor Heterogeneity of the Estrogen Receptor and the Long-term Risk of Fatal Breast Cancer. *JNCI: Journal of the National Cancer Institute* **110**, 726–733 (2018).

82. Ulz, P. *et al.* Inferring expressed genes by whole-genome sequencing of plasma DNA. *Nature Genetics* **48**, 1273–1278 (2016).

83. Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M. & Shendure, J. Cell-free DNA Comprises an in Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell* **164**, 57–68 (2016).

84. Zhu, G. *et al.* Tissue-specific cell-free DNA degradation quantifies circulating tumor DNA burden. *Nature Communications* **12**, 2229 (2021).

85. Sun, K. *et al.* Orientation-aware plasma cell-free DNA fragmentation analysis in open chromatin regions informs tissue of origin. *Genome research* **29**, 418–427 (2019).

86. Jiang, P. *et al.* Plasma DNA End-Motif Profiling as a Fragmentomic Marker in Cancer, Pregnancy, and Transplantation. *Cancer Discov* **10**, 664–673 (2020).

87. Lo, Y. M. D., Han, D. S. C., Jiang, P. & Chiu, R. W. K. Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies. *Science* **372**, (2021).

88. Lai, B. *et al.* Principles of nucleosome organization revealed by single-cell micrococcal nuclease sequencing. *Nature* **562**, 281–285 (2018).

89. Cristiano, S. *et al.* Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature* **570**, 385–389 (2019).

90. Peneder, P. *et al.* Multimodal analysis of cell-free DNA whole-genome sequencing for pediatric cancers with low mutational burden. *Nat Commun* **12**, 3230 (2021).

91. Mouliere, F. *et al.* Enhanced detection of circulating tumor DNA by fragment size analysis. *Science Translational Medicine* **10**, eaat4921 (2018).

92. Underhill, H. R. *et al.* Fragment Length of Circulating Tumor DNA. *PLOS Genet* **12**, 426–37 (2016).

93. Markus, H. *et al.* Analysis of recurrently protected genomic regions in cell-free DNA found in urine. *Science Translational Medicine* **13**, (2021).

94. Budhraja, K. K. *et al.* Analysis of fragment ends in plasma DNA from patients with cancer. *medRxiv* 2021.04.23.21255935 (2021) doi:10.1101/2021.04.23.21255935.

95. Ulz, P. *et al.* Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection. *Nature Communications* **10**, 4666 (2019).

96. Mathios, D. *et al.* Detection and characterization of lung cancer using cell-free DNA fragmentomes. *Nat Commun* **12**, 5060 (2021).

97. Wang, J. *et al.* Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.* **22**, 1798–1812 (2012).

98. Yevshin, I., Sharipov, R., Kolmykov, S., Kondrakhin, Y. & Kolpakov, F. GTRD: A database on gene transcription regulation - 2019 update. *Nucleic Acids Research* **47**, D100–D105 (2019).

99. Albergaria, A. *et al.* Expression of FOXA1 and GATA-3 in breast cancer: the prognostic significance in hormone receptor-negative tumours. *Breast Cancer Research* **11**, R40 (2009).

100. Corces, M. R. *et al.* The chromatin accessibility landscape of primary human cancers. *Science* **362**, eaav1898 (2018).

101. Satpathy, A. T. *et al.* Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nature Biotechnology* **37**, 925–936 (2019).

102. Ahuno, S. T. *et al.* Circulating tumor DNA is readily detectable among Ghanaian breast cancer patients supporting non-invasive cancer genomic studies in Africa. *npj Precis. Onc.* **5**, 1–8 (2021).

103. Bujak, A. Z. *et al.* Circulating tumour DNA in metastatic breast cancer to guide clinical trial enrolment and precision oncology: A cohort study. *PLOS Medicine* **17**, e1003363 (2020).

104. Weber, Z. T. *et al.* Modeling clonal structure over narrow time frames via circulating tumor DNA in metastatic breast cancer. *Genome Medicine* **13**, 89 (2021).

105. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014).

106. Wu, S. J. *et al.* Single-cell CUT&Tag analysis of chromatin modifications in differentiation and tumor progression. *Nat Biotechnol* **39**, 819–824 (2021).

107. Pierce, S. E., Granja, J. M. & Greenleaf, W. J. High-throughput single-cell chromatin accessibility CRISPR screens enable unbiased identification of regulatory networks in cancer. *Nat Commun* **12**, 2969 (2021).

108. Beltran, H. *et al.* Circulating tumor DNA profile recognizes transformation to castration-resistant neuroendocrine prostate cancer. *J Clin Invest* **130**, 1653–1668 (2020).

109. Wu, A. *et al.* Genome-wide plasma DNA methylation features of metastatic prostate cancer. *J Clin Invest* **130**, 1991–2000 (2020).

110. Shen, S. Y. *et al.* Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature* **563**, 579–583 (2018).

111. Liu, M. C. *et al.* Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Annals of Oncology* **31**, 745–759 (2020).

112. Larson, M. H. *et al.* A comprehensive characterization of the cell-free transcriptome reveals tissue- and subtype-specific biomarkers for cancer detection. *Nature Communications* **12**, 2357 (2021).

113. Kang, S. *et al.* CancerLocator: non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA. *Genome Biology* **18**, 53 (2017).

114. Chan, K. C. A. *et al.* Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing. *Proceedings of the National Academy of Sciences* **110**, 18761–18768 (2013).

115. Stover, D. G. *et al.* Association of Cell-Free DNA Tumor Fraction and Somatic Copy Number Alterations With Survival in Metastatic Triple-Negative Breast Cancer. *JCO* **36**, 543–553 (2018).

116. Group (EBCTCG), E. B. C. T. C. Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: patient-level meta-analysis of randomised trials. *The Lancet* **378**, 771–784 (2011).

117. Hefti, M. M. *et al.* Estrogen receptor negative/progesterone receptor positive breast cancer is not a reproducible subtype. *Breast Cancer Research* **15**, R68 (2013).

118. Slamon, D. J. *et al.* Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science* **235**, 177–182 (1987).

119. Guan, X. *et al.* Longitudinal HER2 amplification tracked in circulating tumor DNA for therapeutic effect monitoring and prognostic evaluation in patients with breast cancer. *The Breast* **49**, 261–266 (2020).

120. Nielsen, T. O. *et al.* A Comparison of PAM50 Intrinsic Subtyping with Immunohistochemistry and Clinical Prognostic Factors in Tamoxifen-Treated Estrogen Receptor–Positive Breast Cancer. *Clinical Cancer Research* **16**, 5222–5232 (2010).

121. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research* **44**, W160–W165 (2016).

122. Karimzadeh, M., Ernst, C., Kundaje, A. & Hoffman, M. M. Umap and Bismap: quantifying genome and methylome mappability. *Nucleic Acids Research* **46**, e120–e120 (2018).

123. Dale, R. K., Pedersen, B. S. & Quinlan, A. R. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* **27**, 3423–3424 (2011).

124. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

125. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).

126. Seabold, S. & Perktold, J. Statsmodels: Econometric and Statistical Modeling with Python. in 92–96 (2010). doi:10.25080/Majora-92bf1922-011.

127. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43**, 491–498 (2011).

128. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. **00**, 1–3 (2013).

129. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, 1–4 (2021).

130. Weirauch, M. T. *et al.* Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. *Cell* **158**, 1431–1443 (2014).

131. Goldman, M. J. *et al.* Visualizing and interpreting cancer genomics data via the Xena platform. *Nat Biotechnol* **38**, 675–678 (2020).

132. Clarke, D. J. B. *et al.* Appyters: Turning Jupyter Notebooks into data-driven web apps. *Patterns* **2**, 100213 (2021).

133. Meuleman, W. *et al.* Index and biological spectrum of human DNase I hypersensitive sites. *Nature* **584**, 244–251 (2020).

134. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

135. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* **17**, 261–272 (2020).

136. Vallat, R. Pingouin: statistics in Python. *Journal of Open Source Software* **3**, 1026 (2018).

137. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

138. Gazdar, A. F., Bunn, P. A. & Minna, J. D. Small-cell lung cancer: What we know, what we need to know and the path forward. *Nature Reviews Cancer* **17**, 725–737 (2017).

139. Blackhall, F. *et al.* Will liquid biopsies improve outcomes for patients with small-cell lung cancer? *The Lancet Oncology* **19**, e470–e481 (2018).

140. van Meerbeeck, J. P., Fennell, D. A. & De Ruysscher, D. K. Small-cell lung cancer. *The Lancet* **378**, 1741–1755 (2011).

141. Rudin, C. M. *et al.* Molecular subtypes of small cell lung cancer: a synthesis of human and mouse model data. *Nature Reviews Cancer* **19**, 289–297 (2019).

142. Baine, M. K. *et al.* SCLC Subtypes Defined by ASCL1, NEUROD1, POU2F3, and YAP1: A Comprehensive Immunohistochemical and Histopathologic Characterization. *Journal of Thoracic Oncology* **15**, 1823–1835 (2020).

143. Lissa, D. *et al.* Heterogeneity of neuroendocrine transcriptional states in metastatic small cell lung cancers and patient-derived models. *Nat Commun* **13**, 2023 (2022).

144. Gay, C. M. *et al.* Patterns of transcription factor programs and immune pathway activation define four major subtypes of SCLC with distinct therapeutic vulnerabilities. *Cancer Cell* **39**, 346-360.e7 (2021).

145. Simpson, K. L. *et al.* A biobank of small cell lung cancer CDX models elucidates inter- and intratumoral phenotypic heterogeneity. *Nature Cancer* **1**, 437–451 (2020).

146. Febres-Aldana, C. A. *et al.* Rb Tumor Suppressor in Small Cell Lung Cancer: Combined Genomic and IHC Analysis with a Description of a Distinct Rb-Proficient Subset. *Clinical Cancer Research* OF1–OF12 (2022) doi:10.1158/1078-0432.CCR-22-1115.

147. Rose Brannon, A. *et al.* Enhanced specificity of clinical high-sensitivity tumor mutation profiling in cell-free DNA via paired normal sequencing using MSK-ACCESS. *Nat Commun* **12**, 3770 (2021).

148. Borromeo, M. D. *et al.* ASCL1 and NEUROD1 Reveal Heterogeneity in Pulmonary Neuroendocrine Tumors and Regulate Distinct Genetic Programs. *Cell Reports* **16**, 1259–1272 (2016).

149. Huang, Y.-H. *et al.* POU2F3 is a master regulator of a tuft cell-like variant of small cell lung cancer. *Genes & Development* **32**, 915–928 (2018).

150. Yevshin, I., Sharipov, R., Kolmykov, S., Kondrakhin, Y. & Kolpakov, F. GTRD: A database on gene transcription regulation - 2019 update. *Nucleic Acids Research* **47**, D100–D105 (2019).

151. Jo, S.-Y., Kim, E. & Kim, S. Impact of mouse contamination in genomic profiling of patient-derived models and best practice for robust analysis. *Genome Biology* **20**, 231 (2019).

152. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev* **25**, 1010–1022 (2011).

153. Caeser, R. *et al.* Genomic and transcriptomic analysis of a library of small cell lung cancer patient-derived xenografts. *Nat Commun* **13**, 2144 (2022).

154. Kundaje, A. *et al.* Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. *Genome Res.* **22**, 1735–1747 (2012).

155. Rothwell, D. G. *et al.* Utility of ctDNA to support patient selection for early phase clinical trials: the TARGET study. *Nature Medicine* **25**, (2019).

156. Augert, A. *et al.* MAX Functions as a Tumor Suppressor and Rewires Metabolism in Small Cell Lung Cancer. *Cancer Cell* **38**, 97-114.e7 (2020).

157. Gazdar, A. F., Bunn, P. A. & Minna, J. D. Small-cell lung cancer: What we know, what we need to know and the path forward. *Nature Reviews Cancer* **17**, 725–737 (2017).

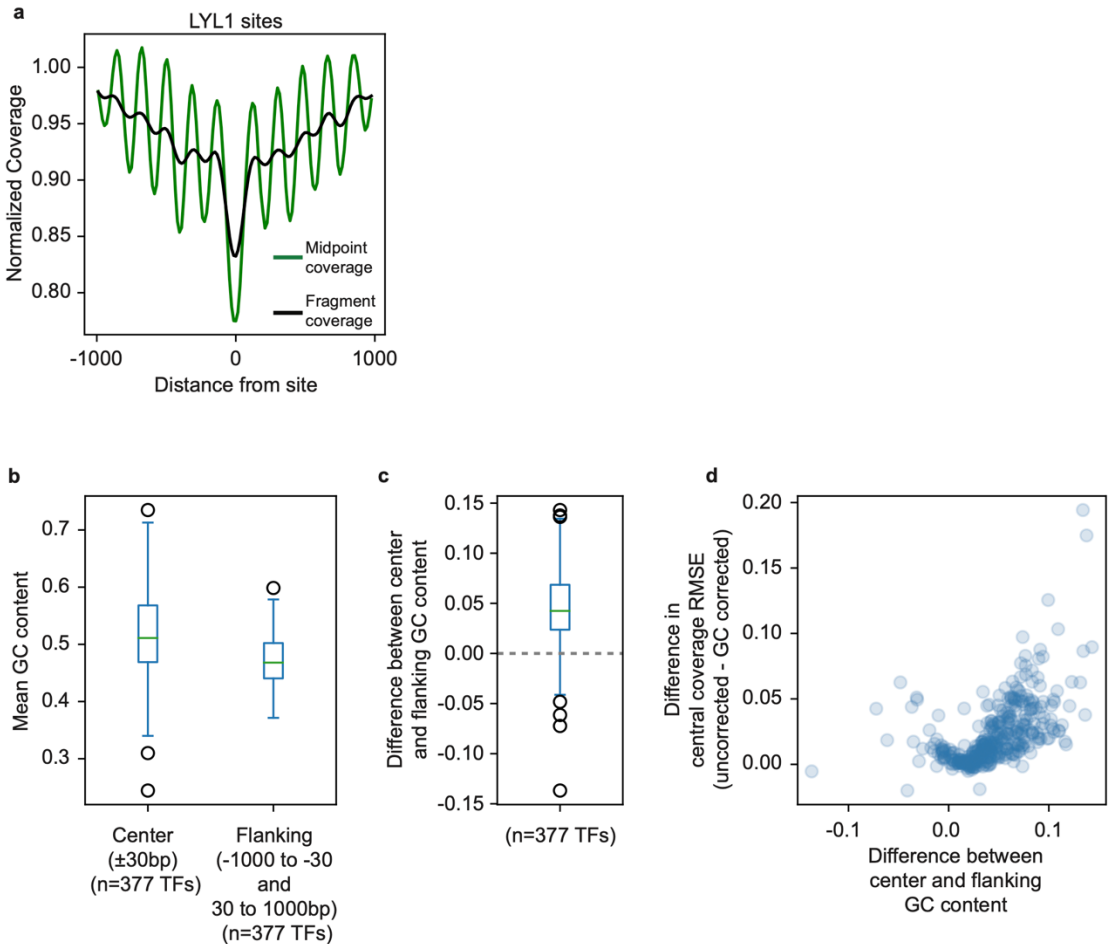
158. Dorritie, K. A., McCubrey, J. A. & Johnson, D. E. STAT transcription factors in hematopoiesis and leukemogenesis: opportunities for therapeutic intervention. *Leukemia* **28**, 248–257 (2014).

159. Liu, T. M., Lee, E. H., Lim, B. & Shyh-Chang, N. Concise Review: Balancing Stem Cell Self-Renewal and Differentiation with PLZF. *Stem Cells* **34**, 277–287 (2016).

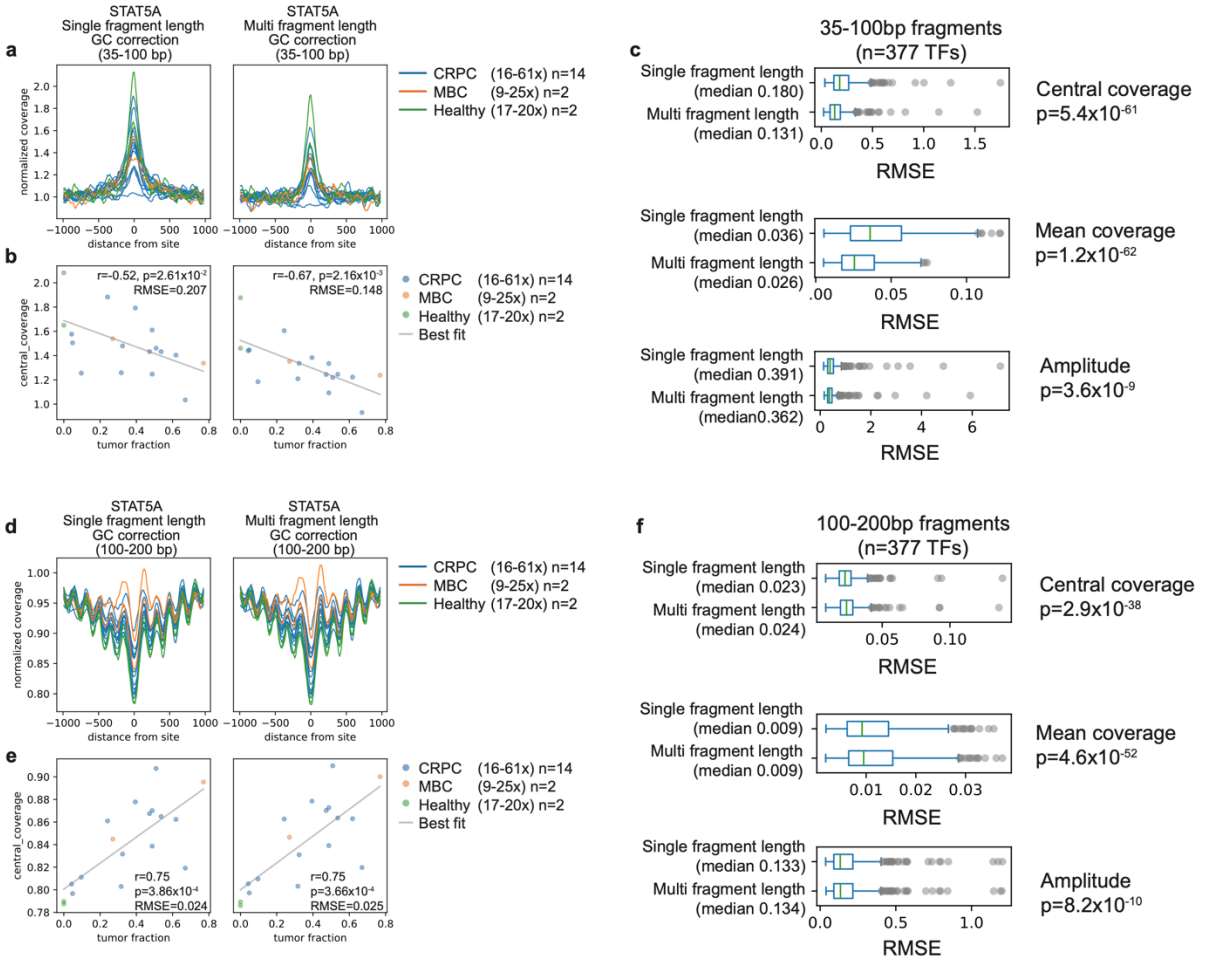
160. Ahuno, S. T. *et al.* Circulating tumor DNA is readily detectable among Ghanaian breast cancer patients supporting non-invasive cancer genomic studies in Africa. *NPJ Precis Oncol* **5**, 83 (2021).

161. Bujak, A. Z. *et al.* Circulating tumour DNA in metastatic breast cancer to guide clinical trial enrolment and precision oncology: A cohort study. *PLOS Medicine* **17**, e1003363 (2020).

Appendix

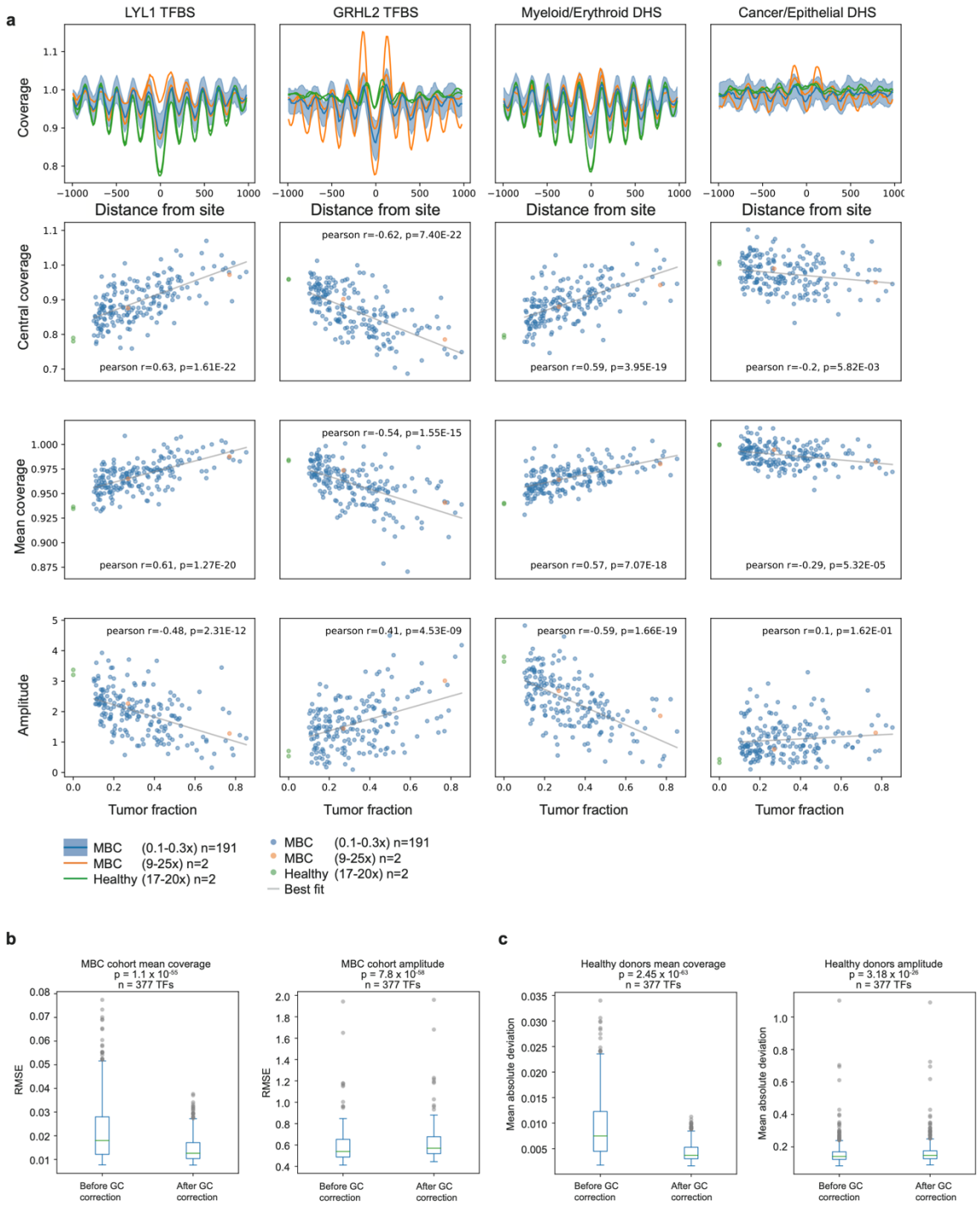


Supplementary Fig. 1: (a) Comparison of midpoint coverage profiles with fragment coverage profiles. The midpoint coverage profile was computed using Griffin for the top 10,000 LYL1 sites in a healthy donor sample (HD_45). Griffin computes coverage by counting the number of midpoints that overlap each site. The fragment coverage was computed using a modified version of Griffin which counted the number of fragments overlapping each position. (**b)** Boxplot of the mean GC content around the top 10,000 TFBSs for each of 377 TFs. For each list of TFBSs, the GC content was calculated in two windows: center (±30bp from the TFBSs) and flanking (±1000bp from the TFBS, excluding the center region ±30bp). The boxed range represents the median ± IQR of the 377 TFs, whiskers represent the range of the non-outlier data (maximum extent is 1.5x the IQR). Outliers are plotted as circles. (**c)** Boxplot of the difference between the mean center GC content and mean flanking GC content for each of the 377 TFs. Box elements are the same as in (b). (**d)** Scatter plot of the difference between center and flanking GC content and the difference in the RMSE before and after GC correction.



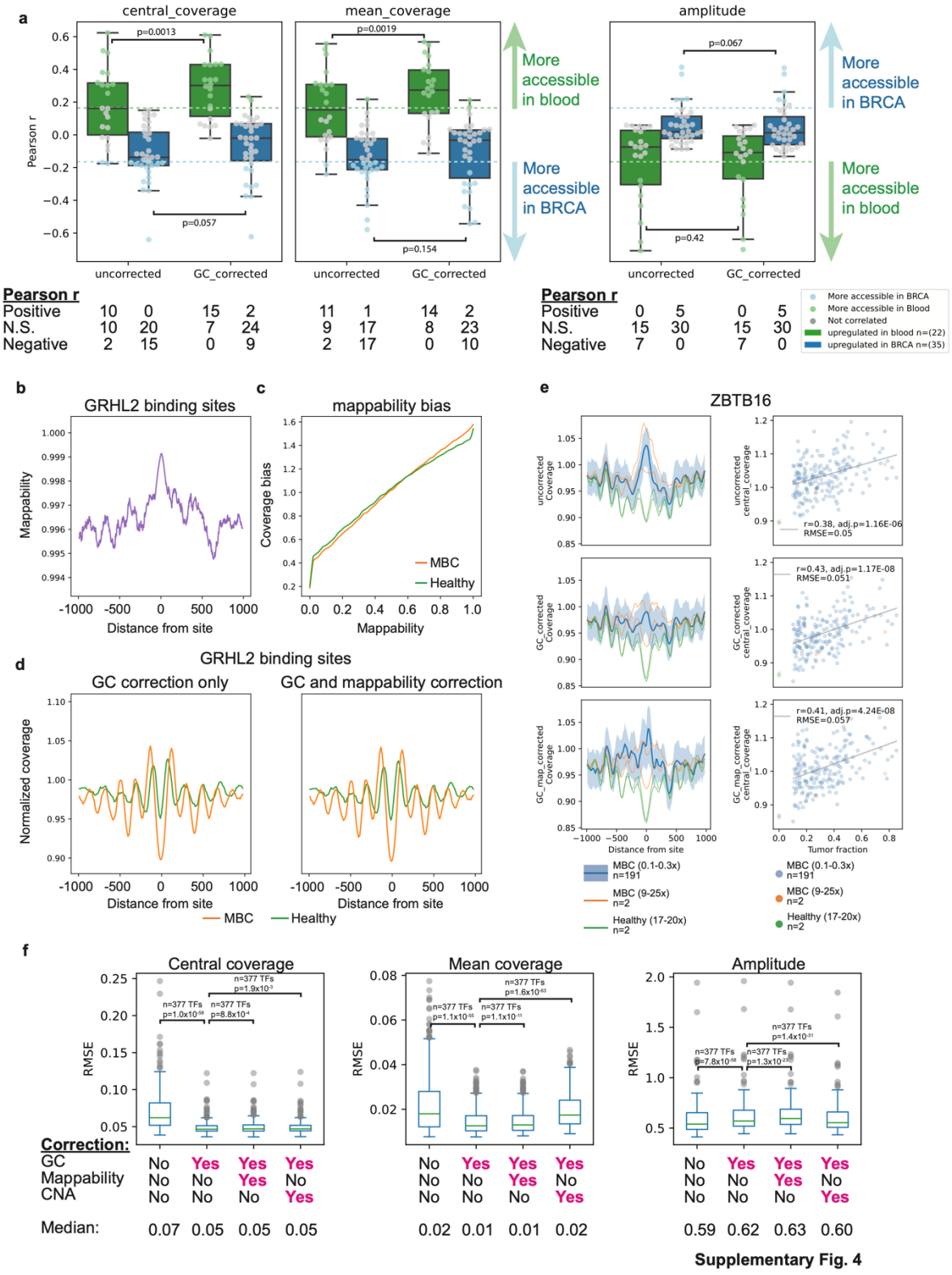
Supplementary Fig. 2: (a) Composite coverage profiles derived from short cfDNA fragments (35-100bp) of 10,000 STAT5A sites with single fragment length (left) and multi fragment length (right) GC correction approaches (see methods), shown for 14 CRPC samples with deep WGS (16-61x WGS, blue), two MBC samples (9-25x WGS, orange), and two healthy donors (17-20x WGS, green). Short cfDNA fragments are derived from TF sites that are actively bound and protected by the TF itself[29]. For STAT5A, which is associated with hematopoiesis[158], we observe the expected increase in coverage at binding sites in healthy donors and lower tumor fraction samples. **(b)** cfDNA tumor fraction and short fragment central coverage correlation for STAT5A, shown for the same deep WGS samples as in (a). cfDNA contains a mixture of tumor and blood cells; therefore, central coverage values are expected to be negatively correlated with tumor fraction for STAT5A (higher coverage represents increased TF binding). The multi fragment length approach leads to a stronger correlation based on Pearson's r correlation coefficient and p-value (2 sided). Root mean squared error (RMSE) of the linear fit is shown. **(c)** Boxplots showing the distribution of the RMSE (linear fit between each of the three features for short fragments (35-100bp) and tumor fraction across the 377 TFs, for single fragment length and multi fragment length GC correction for the same deep WGS samples as in (a). For central coverage 360 of 377 factors had lower RMSE with multi fragment length GC correction (95%), for mean coverage 366 of 377 (97%) and for amplitude 226 of 377 (60%). p-values were calculated using the Wilcoxon signed-rank test (two-sided). The boxed range represents the median ± IQR, whiskers represent the range of the non-outlier data (maximum extent is 1.5x the IQR). Outliers are plotted in grey. **(d)** Composite coverage profiles derived from nucleosome sized cfDNA

fragments (100-200bp) of 10,000 STAT5A sites with single fragment length (left) and multi fragment length (right) GC correction approaches, shown for the same samples as (a). For nucleosome sized fragments, lower 'central coverage' corresponding to greater site accessibility in the healthy donor samples is expected because STAT5A is a transcription factor associated with hematopoiesis. **(e)** cfDNA tumor fraction and short fragment central coverage correlation for STAT5A, shown for the same samples as in (a). Central coverage values for nucleosome sized fragments are expected to be positively correlated with tumor fraction for STAT5A (lower represents greater accessibility). Root mean squared error (RMSE) of the linear fit is shown. **(f)** Boxplots showing the distribution of the RMSE (linear fit between each of the three features for nucleosome sized fragments (100-200bp) and tumor fraction across the 377 TFs, for single fragment length and multi fragment length GC correction. For central coverage 86 out of 377 factors had lower RMSE after correction (95%), for mean coverage 44 of 377 (97%) and for amplitude 145 of 377 (60%). p-values were calculated using the Wilcoxon signed-rank test (two-sided). Box elements are the same as in (c).



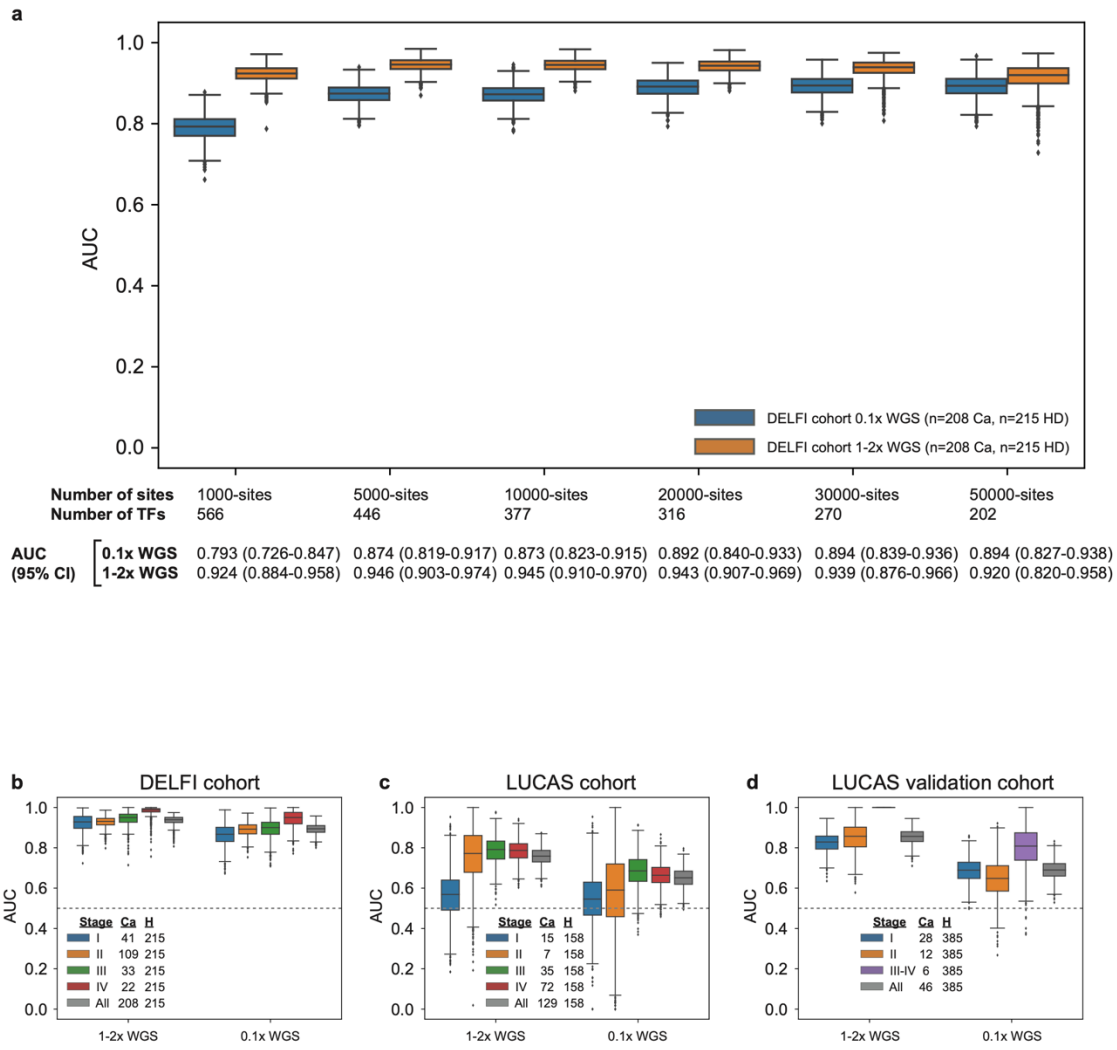
Supplementary Fig. 3: Correlations between tumor fraction and features. **(a)** Top row: Nucleosome profiles for 4 site types that are healthy blood specific (LYL1 and Myeloid/Erythroid DNase Hypersensitivity Sites (DHS)) or cancer specific (GRHL2 and Cancer/Epithelial DHS). Each profile is from the top 10,000 sites. For 191 ULP-WGS MBC samples with ≥0.1 tumor fraction,[28] the median coverage ± IQR is shown. Two healthy donor samples and two deep-WGS MBC samples are included for illustration. Second, third, and fourth rows: Correlation between tumor fraction and central coverage (second row), mean coverage (third row), and amplitude (fourth row) for the 191 MBC samples. Pearson's r and p-value (two sided) are shown for each correlation. Because central coverage and mean coverage are reduced when

a site is accessible, these features are positively correlated with tumor fraction for blood specific sites and negatively correlated with tumor fraction for cancer specific sites. Amplitude is increased when a site is accessible, so this feature is expected to be negatively correlated with tumor fraction for blood specific sites and positively correlated with tumor fraction for tumor specific sites. For all plots, healthy and deep-WGS MBC samples are included for illustration and not included in statistics. **(b)** Boxplots showing the distribution of the RMSE (linear fit between mean coverage and tumor fraction in the MBC ULP-WGS dataset [0.1-0.3x, n=191]) across the 377 TFs, before and after GC correction. The boxed range represents the median ± IQR, whiskers represent the range of the non-outlier data (maximum extent is 1.5x the IQR). Outliers are plotted in grey. p-value was calculated using the Wilcoxon signed-rank test (two-sided). 325 of 377 factors (86%) have a lower RMSE post GC correction for mean coverage and 32 of 377 (8.5%) for amplitude. **(c)** Boxplots showing the distribution of the mean absolute deviation (of the mean coverage and amplitude across 215 healthy donors [1-2x WGS]) across the 377 TFs, before and after GC correction. Box elements are the same as (b). p-value was calculated using the Wilcoxon signed-rank test (two-sided). 372 of 377 factors (99%) have a lower RMSE post GC correction for mean coverage and 89 of 377 (24%) for amplitude.



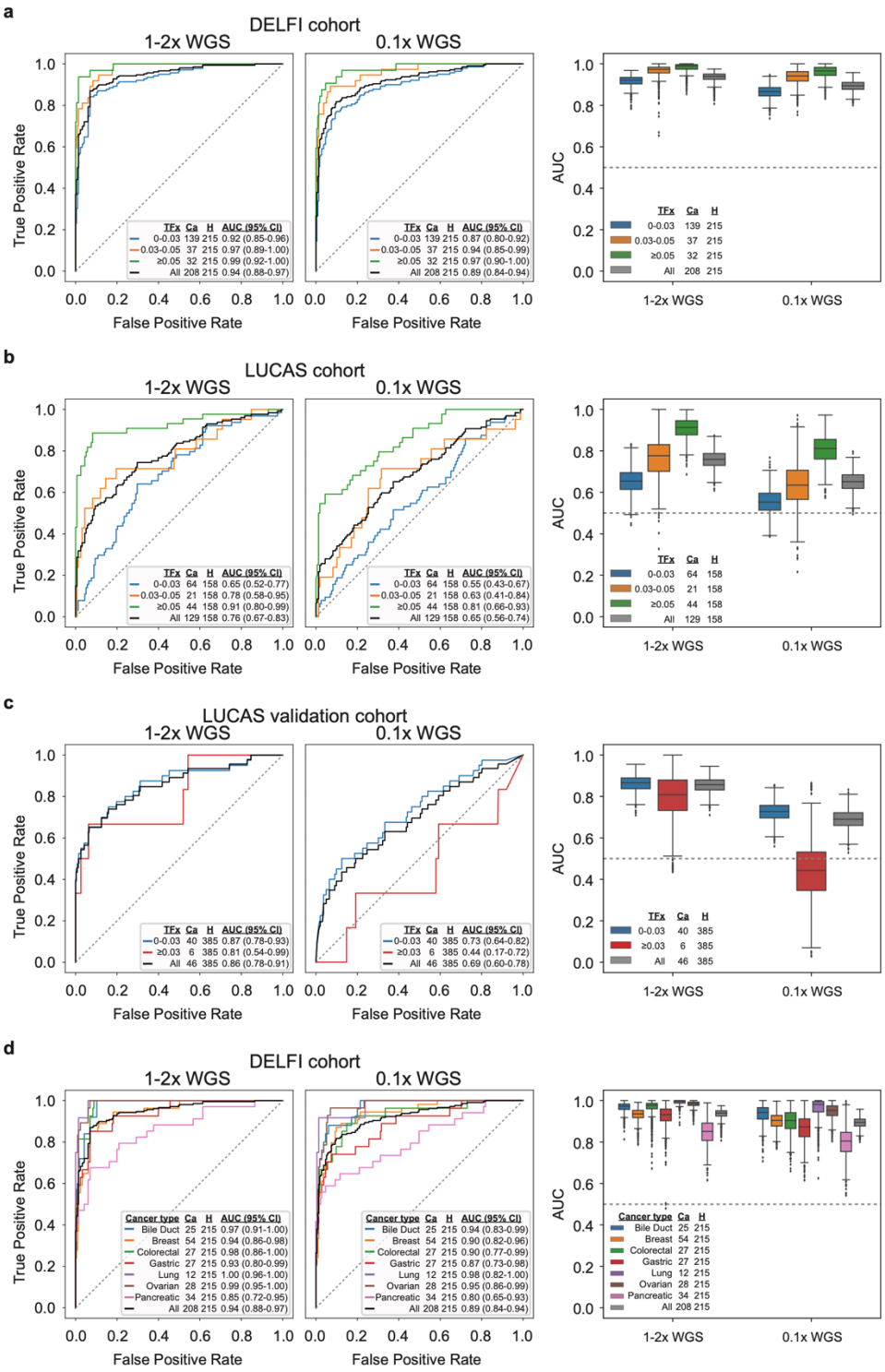
Supplementary Fig. 4: (a) Boxplots of Pearson correlation coefficients for TFs that are differentially expressed between blood cells and breast cancer (BRCA) cells. TFs were identified using differential gene

expression analysis and further filtered to remove TFs which shared many sites with differential TFs from the opposite group (methods). After these filters there were a total of n=22 blood TFs (upregulated in blood relative to cancer, green boxes) and n=35 BRCA TFs (upregulated in cancer relative to blood, blue boxes). Individual values for correlation coefficients between tumor fraction and cfDNA features (centeral coverage, mean coverage, or amplitude) for n=191 metastatic breast cancer samples were plotted. A subset of factors were significantly correlated with tumor faction and these correlations tended to be in the expected direction. Points were colored green (more accessible in blood) or blue (more accessible in cancer) if the Pearson correlation was significant after FDR correction. For blood factors, central coverage, and mean coverage (left and middle panels) correlation tended to be positively correlated to tumor fraction indicating more accessibility in blood cells and GC correction significantly increased this correlation (Wilcoxon signed rank test, two sided). Amplitude was negatively correlated, as expected, but not significantly impacted by GC correction. For BRCA factors, the opposite trend was observed with negative correlations for mean coverage and central coverage and a positive correlation for amplitude. GC correction did not significantly impact these correlations. Box elements are the same as (f). **(b)** Aggregated mean mappability at 10,000 GRHL2 binding sites and its surrounding 2kb region showing a slight increase in mappability (Umap multi-read mappability track for 100bp) at the site center. **(c)** cfDNA mappability bias is unique to each sample. Mappability bias computed for cfDNA from a healthy donor (HD_46; green) and a metastatic breast cancer (MBC_315; orange). **(d)** Composite coverage profile of 10,000 GRHL2 binding sites before and after mappability correction, shown for HD_46 (green) and MBC_315 (orange). There is minimal change in coverage profile after mappability correction. **(e)** Composite coverage profiles of 10,000 ZBTB16 sites before correction (top row), after GC correction only (middle row) and after GC and mappability correction (bottom row), shown for two MBC samples with deep WGS (9-25x, orange), two healthy donors (17-20x, green), and 191 MBC samples with ULP-WGS (0.1-0.3x, blue). Median +/- IQR of 191 ULP-WGS samples is shown with blue shading. Lower 'central coverage' corresponding to greater site accessibility in the healthy donor samples is expected because ZBTB16 is a transcription factor associated with hematopoiesis[159]. However, the addition of mappability correction leads to increased noise (unexpected spikes in coverage) relative to the GC corrected profile. After GC correction, the correlation between tumor fraction and central coverage (for the MBC ULP-WGS samples) is stronger based on Pearson's r correlation coefficient (two sided), however, the addition of mappability correction (bottom row) weakens the correlation. **(f)** Boxplots showing the distribution of the RMSE (Root mean squared error; linear fit between central coverage and tumor fraction in the MBC ULP-WGS dataset [0.1-0.3x, n=191]) across the n=377 TFs, before and after GC, mappability, and copy number alteration (CNA) correction. GC correction leads to a large improvement (decrease) in median RMSE for central coverage and mean coverage but not amplitude (Figure 2f, Supplementary Fig. 3b). However, mappability and CNA correction both have a more modest effect on all features leading to a slight but significant increase in RMSE for all features except amplitude at CNA corrected sites. The boxed range represents the median ± IQR, whiskers represent the range of the non-outlier data (maximum extent is 1.5x the IQR). Outliers are plotted in grey. p-values were calculated using the Wilcoxon signed-rank test (two-sided).



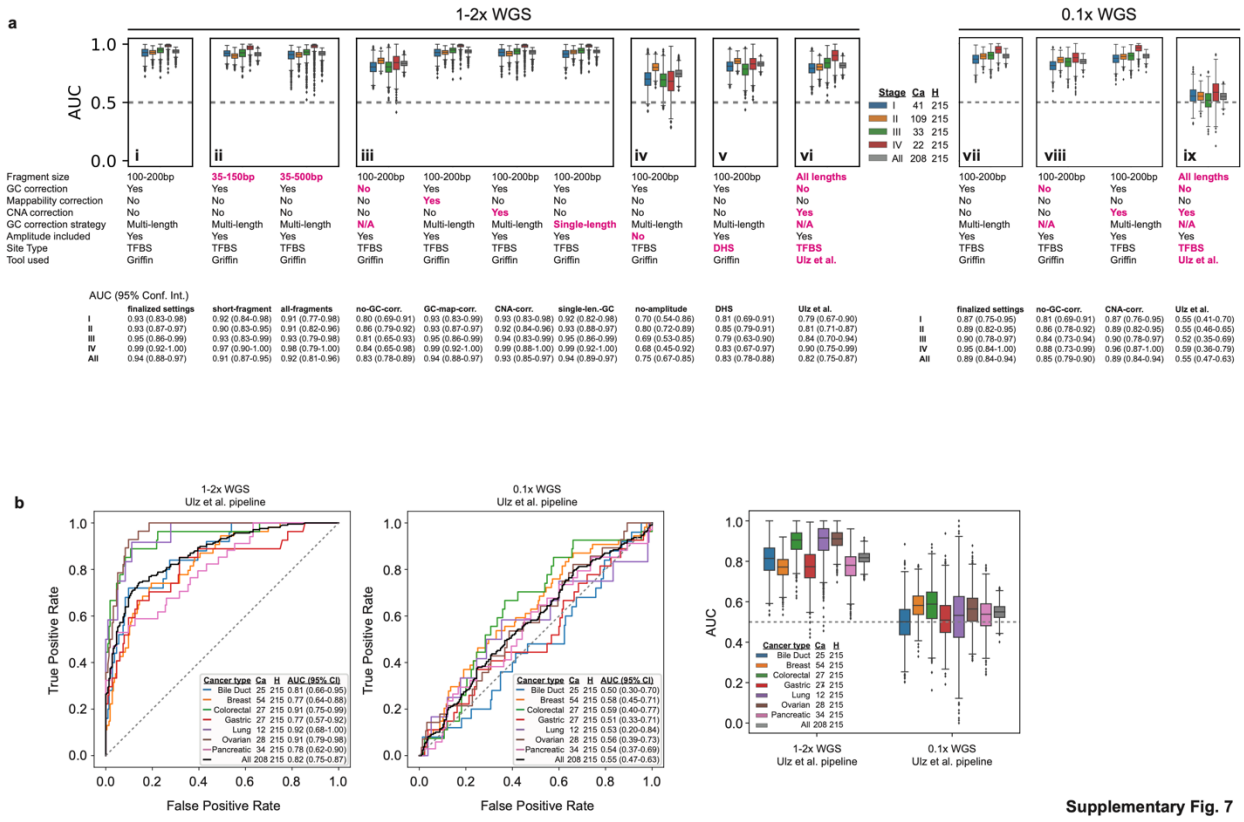
Supplementary Fig. 5: Cancer detection performance metrics. **(a)** Distribution of AUC values for 1,000 bootstrap iterations of the cancer detection logistic regression model on the DELFI cohort (n=208 cancer (Ca) cfDNA samples, n=215 healthy donor (H) cfDNA samples) using different numbers of TFBS per TF when running Griffin (methods). Performance for 1-2x WGS data (orange) and 0.1x downsampled WGS data (blue) are shown. The boxed range represents the median ± IQR, whiskers represent the range of the non-outlier data (maximum extent is 1.5x the IQR). Outliers are plotted as grey diamonds. 95% confidence intervals (CIs) are printed below the plot and were obtained from 1,000 bootstrap iterations **(b)** Distribution of AUC values for 1,000 bootstrap iterations of the cancer detection logistic regression model on the DELFI cohort [same samples as in (a)] for 1-2x WGS data and 0.1x WGS data. Values are shown for each stage and overall. Corresponds to the data and values shown in Main Figure 3a. Box elements are the same as in (a). **(c)** Distribution of AUC values for 1,000 bootstrap iterations of the cancer detection logistic regression model on the LUCAS cohort (n=129 cancer (Ca) cfDNA samples, n=158 non-cancer (H) cfDNA samples) for 1-2x WGS data and 0.1x WGS data. Values are shown for each stage and overall. Corresponds to the data and values shown in Main Figure 3c. Box elements are the same as in (a). **(d)** Distribution of AUC values for 1,000 bootstrap iterations of the cancer detection logistic regression model on the LUCAS validation cohort (n=28 cancer (Ca) cfDNA samples, n=385 healthy donor (H) cfDNA samples) for 1-2x WGS data and 0.1x WGS data. Values are shown for each stage (III and IV are combined due to small

number of samples) and overall. Corresponds to the data and values shown in Main Figure 3d. Box elements are the same as in (a).



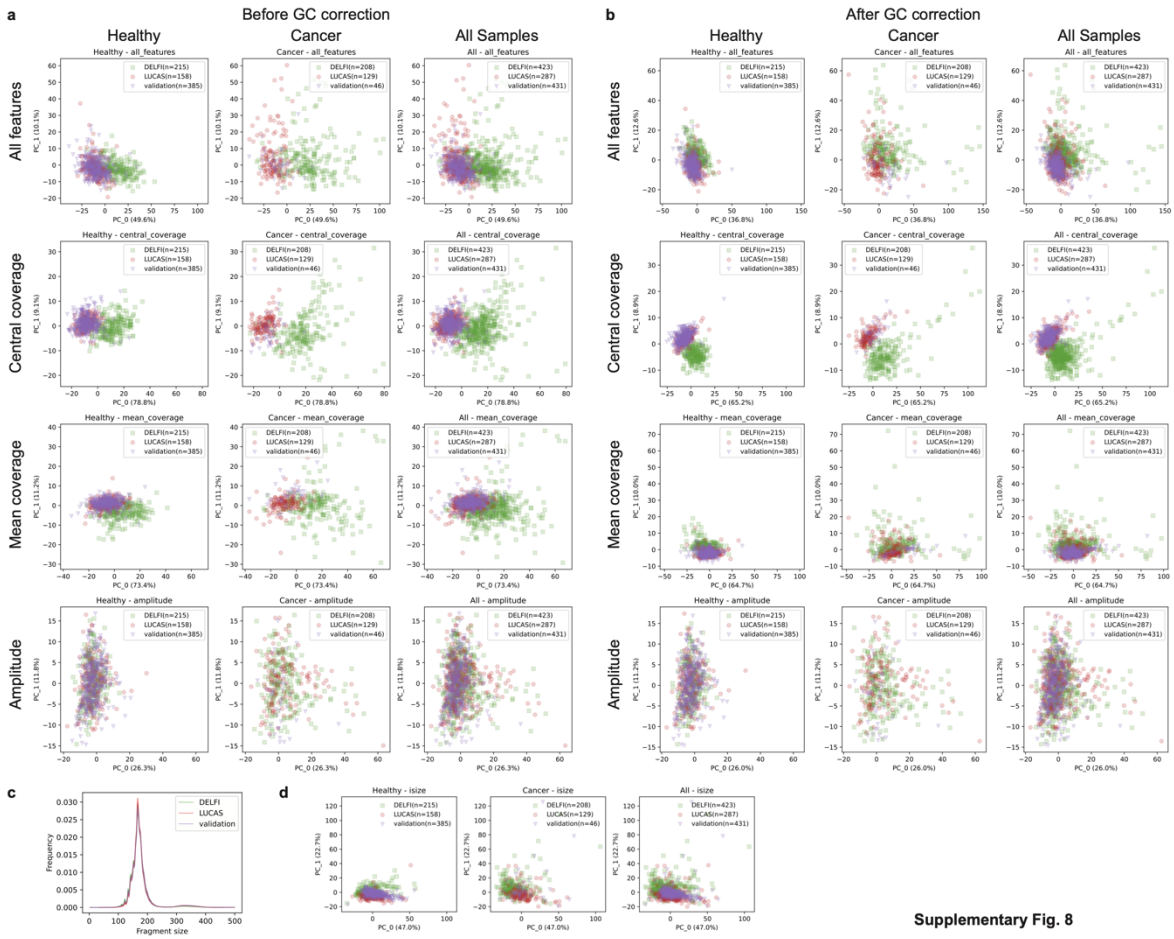
Supplementary Fig. 6: Cancer detection performance metrics by tumor fraction (TFx) and cancer type. Receiver operator characteristic (ROC) curves for logistic regression classification of cancer (Ca) vs. healthy controls (H) are shown for three cohorts, **(a,d)** the DELFI cohort[89], **(b)** the LUCAS cohort, and **(c)** the LUCAS validation cohort[36]. Logistic regression was performed on the top PCA components which explained 80% of the variance in the features (central coverage, mean coverage, and amplitude) extracted

from nucleosome profiles around 30,000 TFBSs for each of 270 TFs. ROC for cancer grouped by TFx vs. healthy **(a-c)** or cancer grouped by cancer type vs. healthy **(d)** are shown. Duodenal cancer (n=1) is not shown as a separate cancer type. For each cohort, performance is shown for both the original low pass (1-2x) WGS (left panel) and ultra-low pass (0.1x) WGS (middle panel) generated by in-silico downsampling. 95% confidence intervals (CIs) were obtained from 1,000 bootstrap iterations. The right panel for each cohort contains boxplots of the AUC values for the bootstrap iterations. The boxed range represents the median ± IQR, whiskers represent the range of the non-outlier data (maximum extent is 1.5x the IQR). Outliers are shown as grey diamonds.

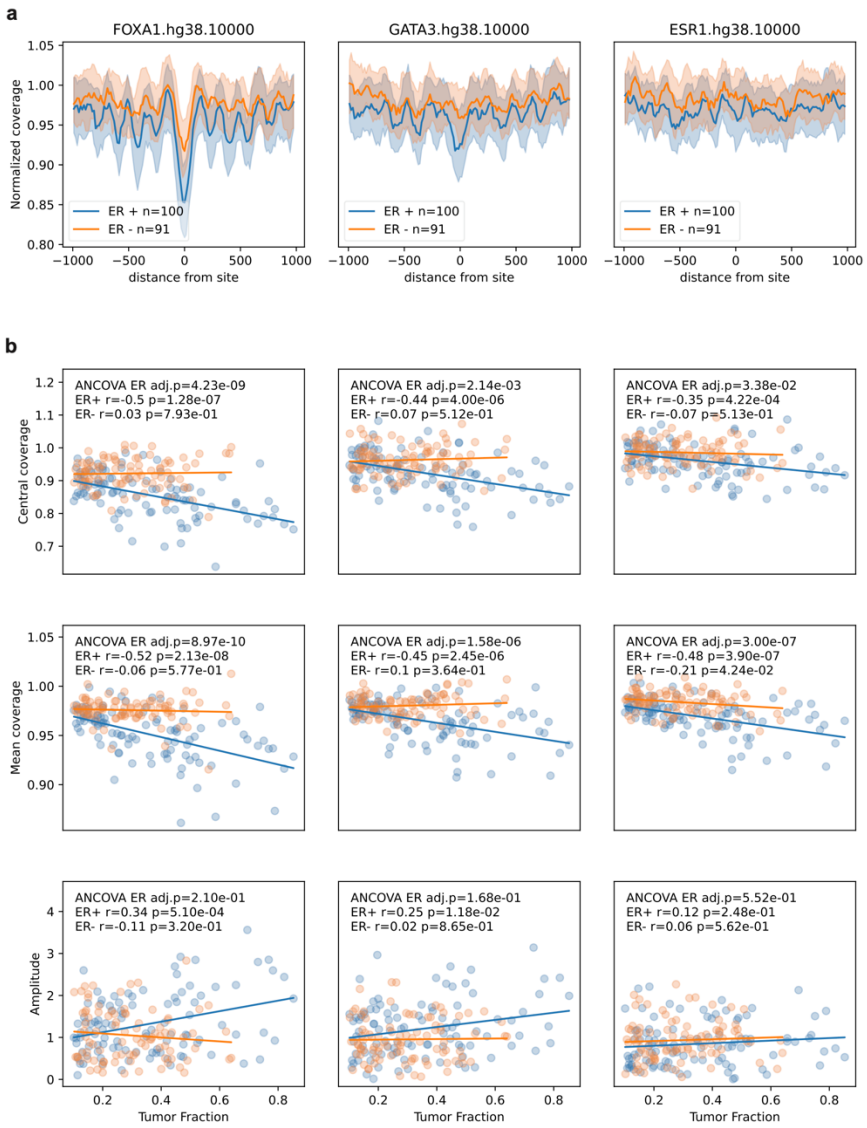


Supplementary Fig. 7: Evaluation of various configurations and comparisons of Griffin for cancer detection. **(a)** Boxplots of the AUC values for 1,000 bootstrap iterations of the logistic regression classifier using various configurations and comparisons of Griffin on the DELFI cohort (n=208 cancer (Ca) cfDNA samples, n=215 healthy donor (H) cfDNA samples) in the original 1-2x WGS data and downsampled data (0.1x WGS) grouped by stage. The boxed range represents the median ± IQR, whiskers represent the range of the non-outlier data (maximum extent is 1.5x the IQR). Outliers are shown as grey diamonds. **(i)** finalized Griffin configuration with parameters settings and median AUC for the 1,000 bootstraps (with 95% CI) listed underneath for comparison to other configurations. Logistic regression model was trained on top PCA features extracted from mean nucleosome profiles around 30,000 TFBSs for each of 270 TFs (total of 810 features prior to PCA dimensionality reduction), see methods. **(ii)** same Griffin analysis as in (i) but using two different fragment size ranges, short fragments, which are known to be enriched in cancer[33,35] (35-150bp), and a wider range of fragment sizes encompassing most cfDNA fragments (35-500bp). Both have decreased performance (0.91 and 0.92 AUC, respectively). **(iii)** Same Griffin analysis as in (i) but without GC correction (left), with an added mappability correction step (middle left), with an added CNA correction step (middle right), and with a different GC correction approach using a single fragment length (right). **(iv)** Same as (i) but with the amplitude features excluded from the model (only used central coverage and mean coverage features). **(v)** Griffin with standard parameters from (i) applied to the top 10,000 sites around 16 types of tissue specific DNAse hypersensitivity sites (see methods). **(vi)** Pipeline developed by Ulz and colleagues. This pipeline uses all fragment sizes (the vast majority of which are between 35 and 500bp) and collects coverage profiles around the top 1,000 sites for 504 transcription factors and extracts one feature (High frequency range) per feature. Dimensionality was reduced with PCA using the same approach as described in (i) and the top features were put into the logistic regression model. **(vii-xi)** Performance of selected configurations on 0.1x WGS data. **(b)** Receiver operator characteristic (ROC) curve for logistic regression classification of cancer vs. healthy controls using the pipeline from Ulz et al on the DELFI

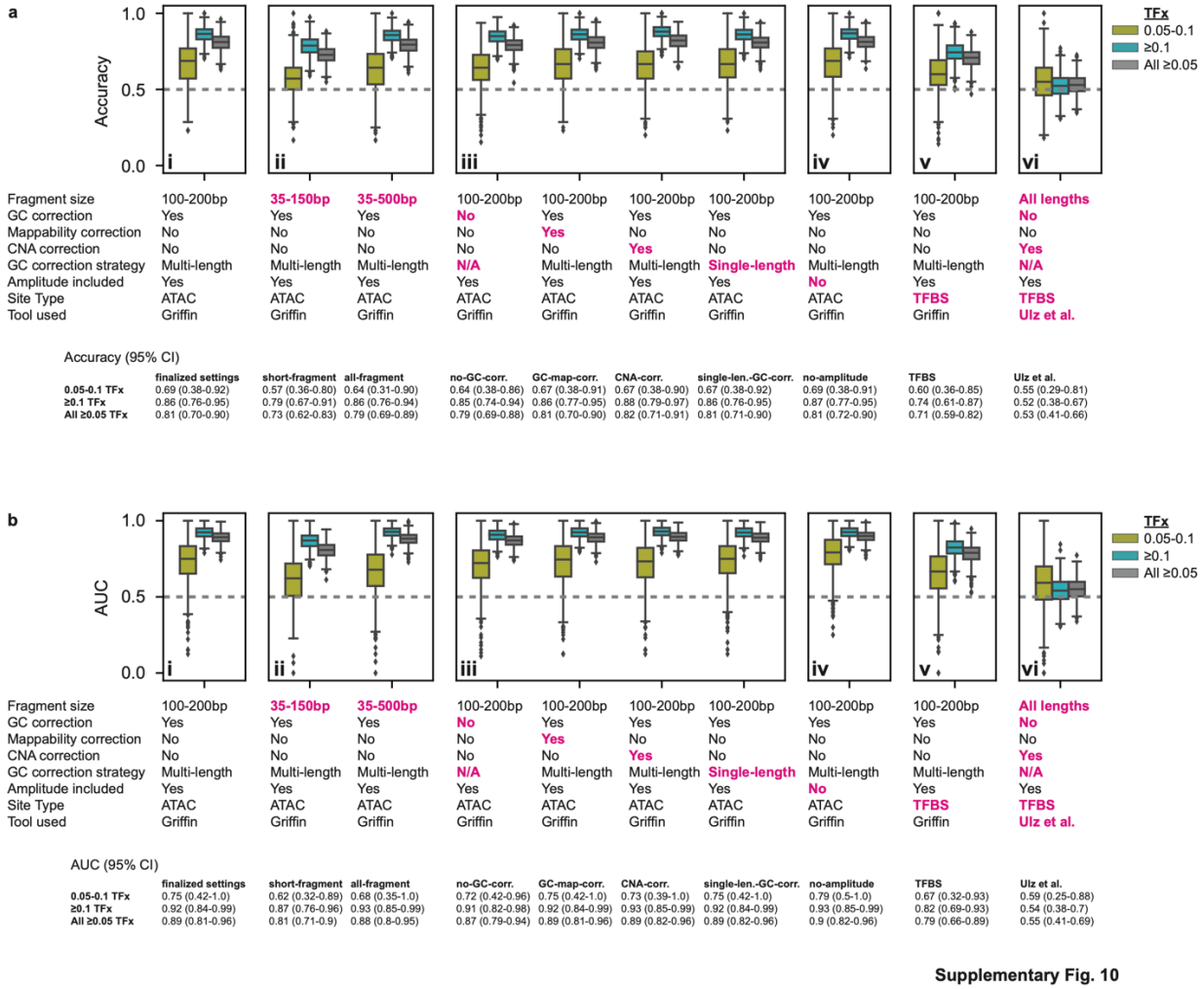
cohort[89] in 1-2x WGS data (left) and 0.1x WGS data (middle). ROC for each cancer type vs. healthy are shown. 95% confidence intervals (CIs) were obtained by bootstrapping. Duodenal cancer (n=1) is not shown as a separate cancer type. The right panel for each cohort contains boxplots of the AUC values for 1,000 bootstrap iterations. Box elements are the same as in (a).



Supplementary Fig. 8: Principal component analysis (PCA) on Griffin features for cancer detection cohorts before GC correction **(a)** and after GC correction **(b)**. For each cancer detection cohort (DELFI, LUCAS, and LUCAS validation) Griffin analysis was performed on 30,000 TFBSs each for 270 TFs and 3 features (central coverage, mean coverage, and amplitude) were extracted from each profile for a total of 810 features. Top row, a PCA was performed on all features for all samples and the top two components were plotted for healthy samples from all three cohorts (left), cancer samples (middle) and all samples (right). The DELFI cohort clustered away from the other cohorts indicating systematic difference between the DELFI cohort and other cohorts. Next, PCA was performed separately on each of the 3 feature types: central coverage (second row), mean coverage (third row), and amplitude (bottom row) which revealed that the difference between the DELFI cohort and other cohorts was primarily due to differences in the central coverage. Percentage of variance explained by each PC is labeled on the axes. **(c)** Mean normalized fragment size profiles for the three cohorts. Shading indicates IQR. **(d)** PCA of the fragment size profiles in the three cohorts. Top two PCs are shown for healthy samples (left), cancer samples (middle), and all samples (right).

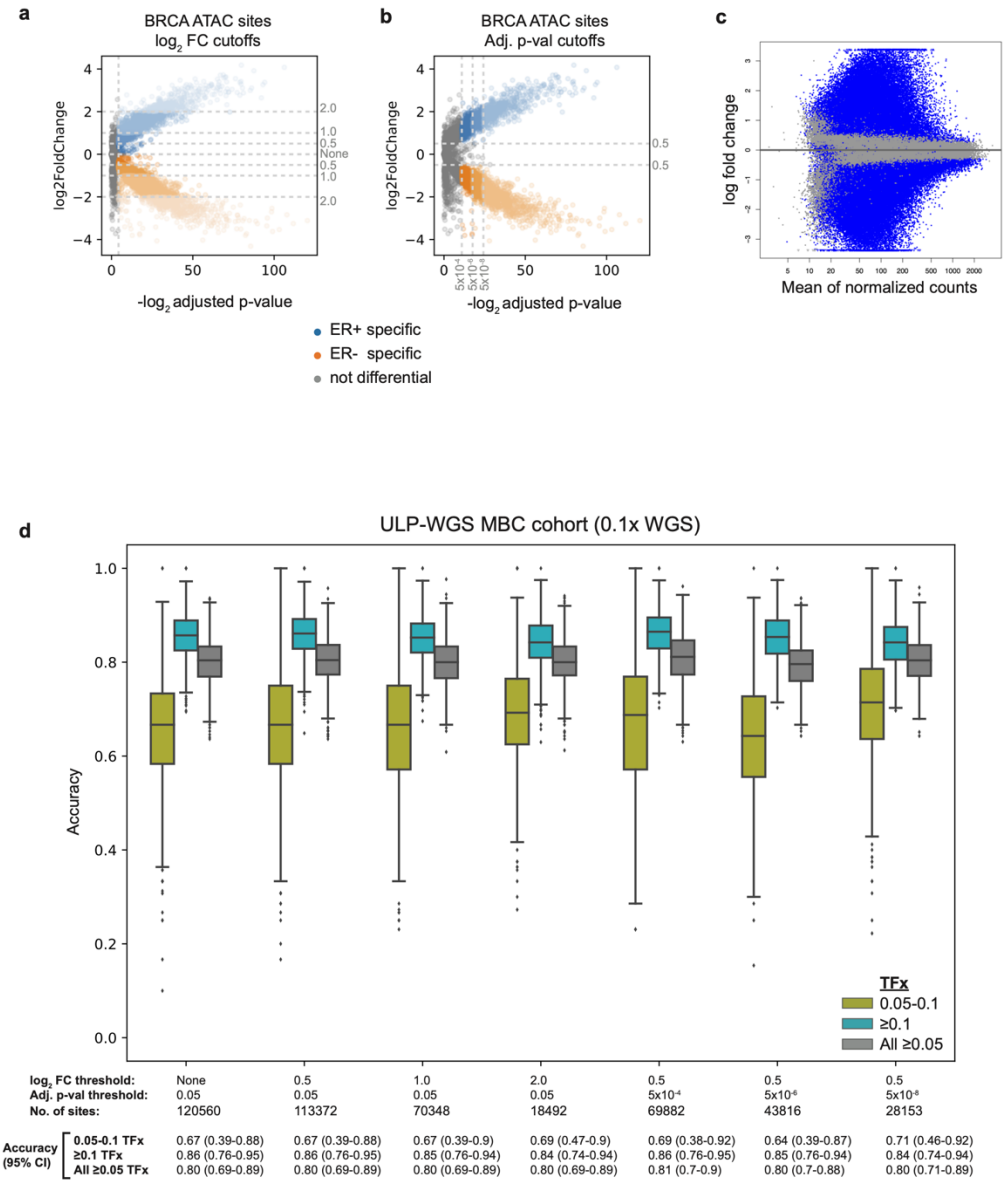


Supplementary Fig. 9: (a) Coverage profiles for the top 10,000 TFBSs for each of 3 key ER positive specific transcription factors, FOXA1, GATA3, and ESR1. Median ± IQR shown for 100 ER positive and 91 ER negative ULP-WGS MBC samples with ≥ 0.1 tumor fraction.[28] **(b)** Correlation between the three features and the tumor fractions for the coverage profiles shown in (a). Top row: central coverage, middle row: mean coverage, bottom row: amplitude. ANCOVA p-values (two sided) are shown for the ER status after accounting for the tumor fraction as a covariate. Benjamini-Hochberg FDR correction was used for multi test correction (See Methods). Pearson r and p-values (two sided) for the correlation between tumor fraction and feature are shown for ER+ and ER- separately.



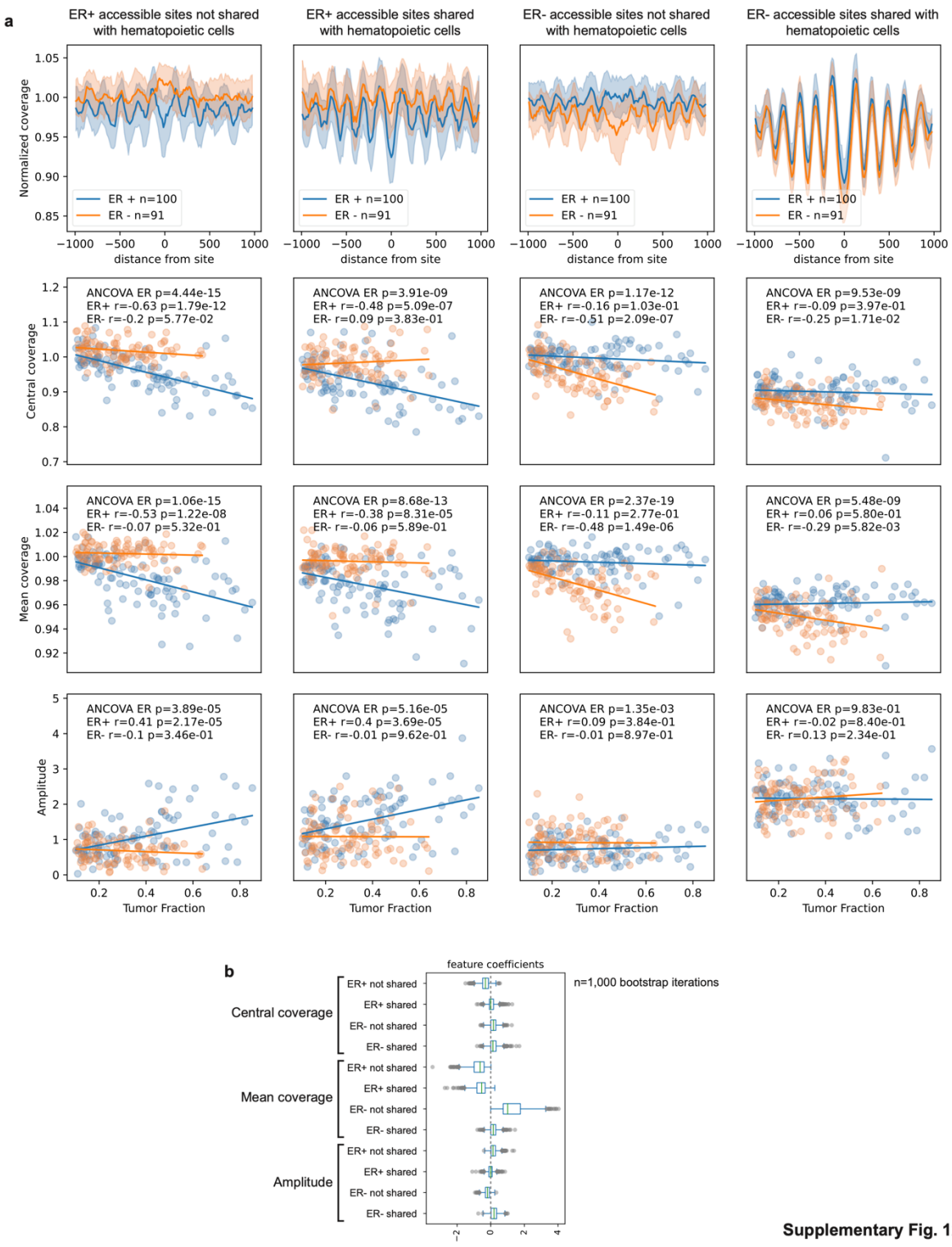
Supplementary Fig. 10: Evaluation of various configurations and comparisons of Griffin for ER status prediction. **(a)** Boxplots of the accuracy values for 1,000 bootstrap iterations of the logistic regression classifier using various configurations and comparisons of Griffin on the MBC cohort (n=139, 0.1x WGS data). Accuracy is shown for patients grouped by tumor fraction (TFx), 0.05 – 0.1 (ER+, n=24; ER-, n=14) and ≥ 0.1 (ER+, n=50; ER-, n=51), and for all patients with ≥0.05 TFx. The boxed range represents the median ± IQR, whiskers represent the range of the non-outlier data (maximum extent is 1.5x the IQR). Outliers are shown as grey diamonds. **(i)** finalized Griffin configuration with parameters settings and median accuracy for the 1,000 bootstrap iterations (with 95% CI) listed underneath for comparison to other configurations. Logistic regression model was trained on the differential ATAC features (total of 12 features), see methods. **(ii)** same Griffin analysis as in (i) but using two different fragment size ranges, short fragments which are known to be enriched in cancer[33,35] (35-150bp) and a wider range of fragment sizes encompassing most cfDNA fragments (35-500bp). **(iii)** Same Griffin analysis as in (i) but without GC correction (left), with an added mappability correction step (middle left), with an added CNA correction step (middle right), and with a different GC correction approach using a single fragment length (right). **(iv)** Same as (i) but with the amplitude features excluded from the model (only used central coverage and mean coverage). **(v)** Griffin with standard parameters from (i) applied to the top 30,000 TFBSs for 270 TFs. Features were extracted and dimensionality was reduced with PCA using the same approach as described in the methods for cancer detection (see methods). **(vi)** Regression model on the outputs of the nucleosome profiling pipeline developed by Ulz and colleagues. This pipeline uses all fragment sizes (the vast majority

of which are between 35 and 500bp) and collects coverage profiles around the top 1,000 sites for 504 transcription factors and extracts one feature (High frequency range) per feature. Dimensionality was reduced with PCA using the same approach as described in the methods for cancer detection using TFBSs and the top features were put into the logistic regression model. **(b)** Same as (a) but showing AUC rather than accuracy.



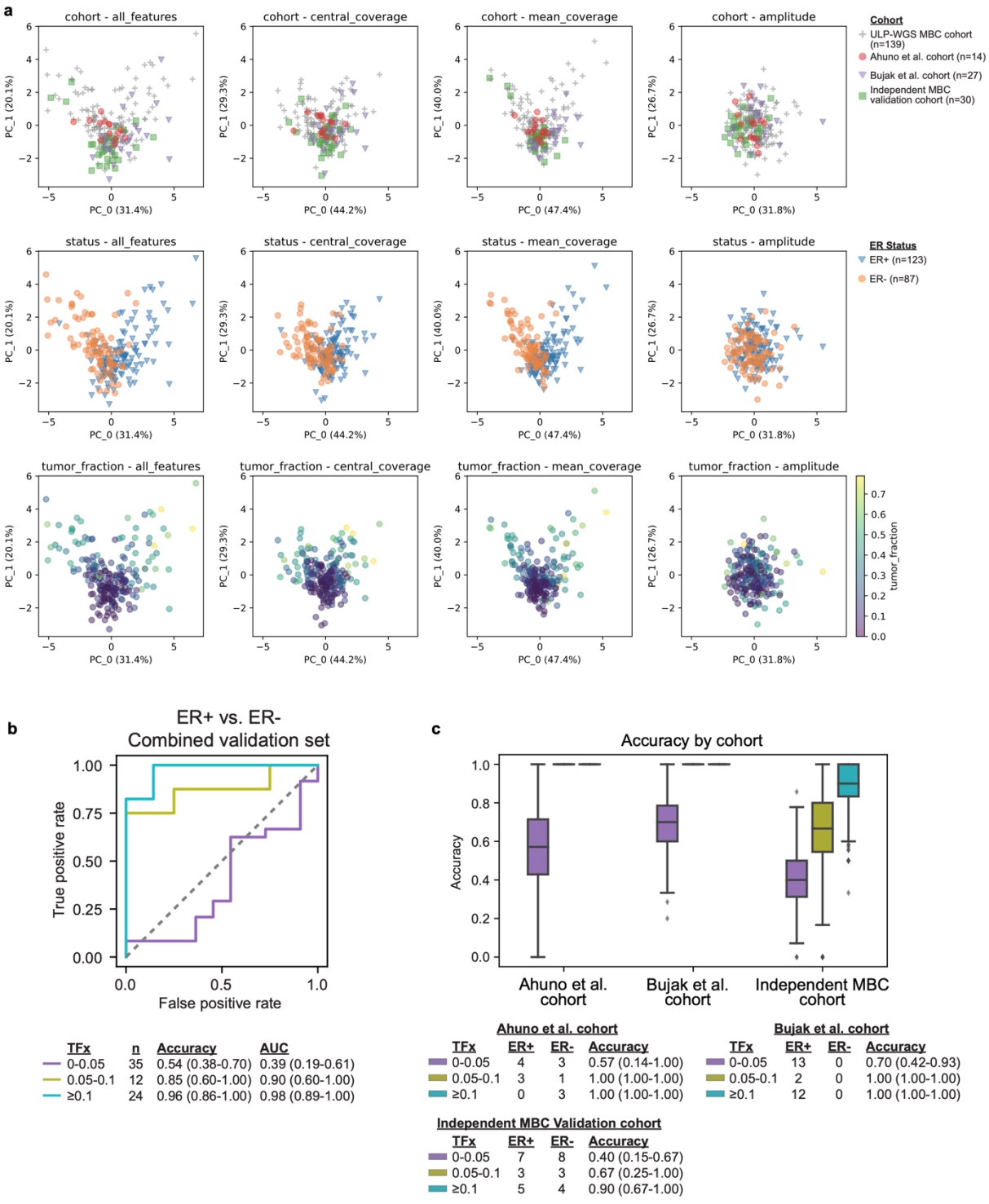
Supplementary Fig. 11: Evaluation of different cutoffs for differential ATAC site selection. **(a)** ER+ and ER- differential open chromatin sites were selected from assay for transposase-accessible chromatin using sequencing (ATAC-seq) data from ER+ (n=44) and ER- (n=15) breast cancer (BRCA) tumors in The Cancer Genome Atlas (TCGA).[100] Differential sites were identified using the DESeq2 software[105] which uses a Wald test with Benjamini-Hochberg FDR correction to calculate the adjusted p-value and $\log_2$ fold-change (FC) for each site. The butterfly plot displays the adjusted p-values and $\log_2$ fold-change values for ATAC sites. Several different cutoffs (None, 0.5, 1.0, 2.0) were considered and are shown with dashed lines. For all $\log_2$ FC cutoffs, an adjusted p-value cutoff of $5x10^{-2}$ was also used. Sites that met the criteria for being differential are shaded in blue (ER+) or orange (ER-). **(b)** Same as (a) but with various adjusted p-value cutoffs from DESeq2 ($5x10^{-4}$, $5x10^{-6}$,$5x10^{-8}$). For all cutoffs, a $\log_2$ fold-change cutoff of 0.5 was also used. **(c)** M-A plot showing the relationship between number of read counts at the sites and $\log_2$ fold change. Output from DESeq2. **(d)** Boxplot of accuracy values for 1,000 bootstrap iterations of the ER status prediction logistic regression model on the ULP-WGS MBC cohort using different DESeq2 cutoffs for selecting differential sites. Accuracy is shown for patients grouped by tumor fraction (TFx), 0.05 – 0.1

(ER+, n=24; ER-, n=14) and ≥ 0.1 (ER+, n=50; ER-, n=51), and for all patients with ≥0.05 TFx. 95% CIs were obtained by bootstrapping. The performance was generally similar for all cutoffs but there was a slightly higher performance when using a $\log_2$ FC of 0.5 and an adjusted p-value cutoff of $5x10^{-4}$ so this cutoff was used for further analysis. The boxed range represents the median ± IQR, whiskers represent the range of the non-outlier data (maximum extent is 1.5x the IQR). Outliers are plotted as grey diamonds.



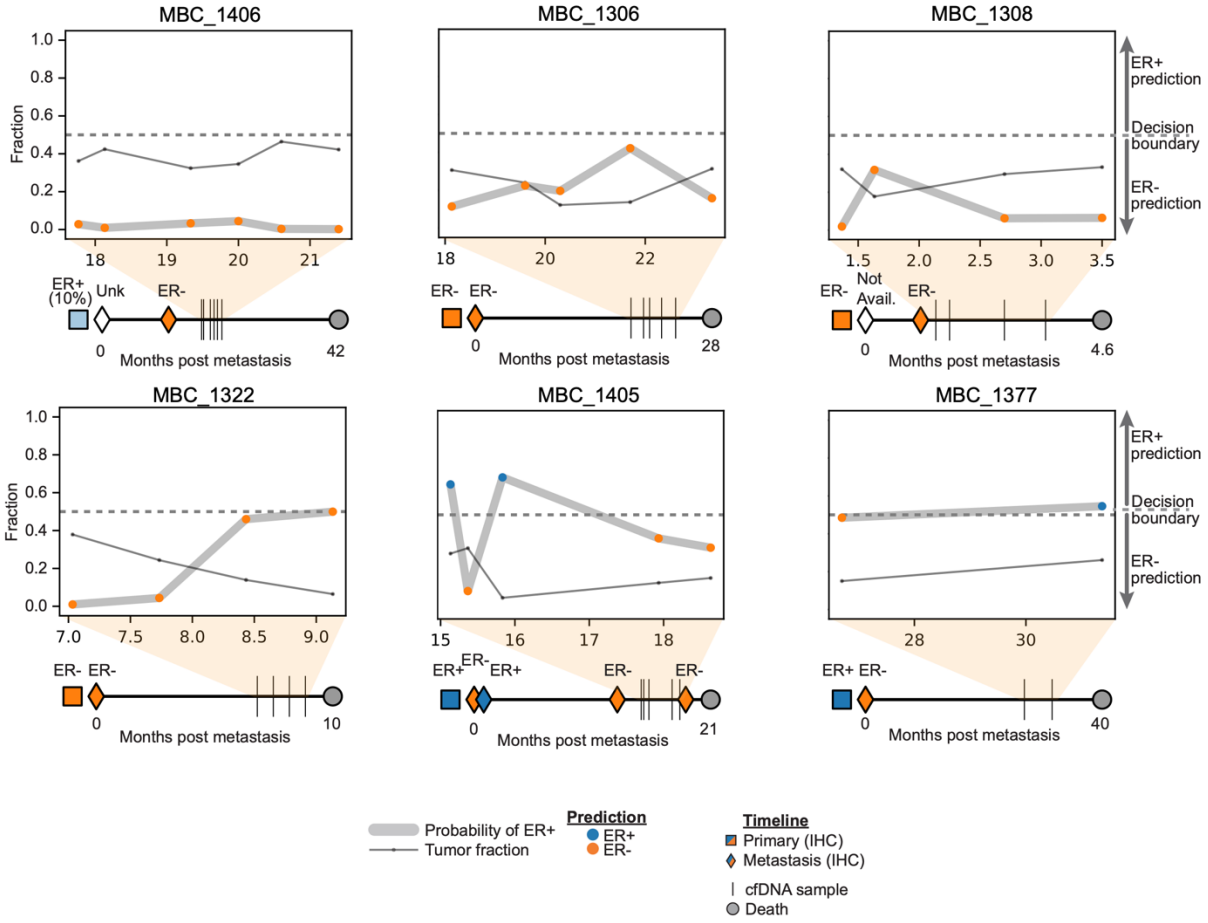
Supplementary Fig. 12: Differential ATAC seq features identified using the optimal $5x10^{-4}$ adjusted p-value cutoff from DESeq2 **(a)** First row: Griffin coverage profiles for ER subtype differential ATAC-seq sites in ER+ (n=100) and ER- (n=91) ULP-WGS MBC samples with ≥ 0.10 tumor fraction.[28] Median +/- IQR is shown. Second, third, and fourth rows: correlation between tumor fraction and central coverage

(second row), mean coverage (third row), and amplitude (fourth row), respectively, for ER+ (n=100) and ER- (n=91) samples. ANCOVA p-values (two sided) for the ER status after accounting for the tumor fraction are shown on the plots (See methods, ANCOVA). Pearson r and p-values (two sided) for the correlation between tumor fraction and feature are shown for ER+ and ER- separately. **(b)** Boxplot of logistic regression feature coefficients for each of 1,000 bootstrap iterations. 'not shared' sites are not shared with hematopoietic cells, 'shared' are shared with hematopoietic cells. The boxed range represents the median ± IQR, whiskers represent the range of the non-outlier data (maximum extent is 1.5x the IQR). Outliers are plotted in grey.



Supplementary Fig. 13: MBC validation set performance **(a)** Principal component analysis (PCA) on Griffin features for breast cancer samples from the initial ULP-WGS cohort and three validation cohorts. For each cohort Griffin analysis was performed on differential ATAC seq sites identified by DESeq2 using the $5x10^{-4}$ adjusted p-value cutoff. 3 features were extracted from each profile for a total of 12 features. A PCA was performed on all 12 features (first column), central coverage features only (second column), mean coverage features (third columns), or amplitude features (fourth column). This PCA was

then colored by cohort (top row) to look for batch effects, but batch effects were not observed in the top two principal components. The PCA was also colored by ER status (second row), demonstrating that the first PC (PC_0, x axis) appears to correspond to status. Finally, the PCA was colored by tumor fraction (third row) indicating that the second PC (PC_1, y axis) corresponds to tumor fraction. Percentage of variance explained by each PC is labeled on the axes. **(b)** Receiver operator characteristic (ROC) curve for a logistic regression model predicting ER+ and ER- subtype on 71 breast cancer samples from the three validation cohorts[160,161]. Model was trained on the initial ULP-WGS cohort and applied to the three validation cohorts. ROC curve, accuracy and AUC are shown for all patients and for patients grouped by tumor fraction (TFx) similar to Figure 4e but including performance for samples below 0.05 TFx. 95% CIs were obtained by bootstrapping. For patients with multiple samples, the first sample was used. **(c)** Boxplot of the accuracy of the model on the validation cohorts grouped by tumor faction and cohort. Confidence intervals were obtained via bootstrapping (n=1000 bootstrap iterations). The boxed range represents the median ± IQR, whiskers represent the range of the non-outlier data (maximum extent is 1.5x the IQR). Outliers are plotted in grey.



Supplementary Fig. 14: Timelines of disease progression for 6 patients with multiple metastatic biopsies and multiple cfDNA samples. ER+ prediction probability (thick grey line), and tumor fraction (thin grey line) are shown for all cfDNA samples that passed the >0.05 tumor fraction and 0.1x coverage thresholds. Decision boundary for ER+ (≥0.5) and ER- (<0.5) is indicated with dotted line. Timelines in months from metastatic diagnosis to death are shown for each patient. The square indicates primary status and timeline from primary is not to scale. Metastatic biopsies are shown with diamonds, blue indicates an ER+ biopsy, light blue an ER low biopsy (1-10% ER positive), orange an ER- biopsy, and white a biopsy where status was not available from clinical records.