Scientific background

Definition of clinical need: Breast cancer treatment is effective in most but not all women. Current methods for monitoring treatment (physical examination and imaging) fail in some women leading to unnecessary exposure to toxic treatment and delaying effective treatment. Breast cancer (BC) is a leading cause of death in women living in the western world. Identification of BC as a localized disease allows treatment options that often lead to a cure, making early detection crucially important. In BC, the current standard of care, neoadjuvant systemic therapy (NAST), achieves a pathological complete response (pCR) of 40-70% [1,2]. Patients achieving a pCR have an increased 5-year overall survival (OS) of 10-12% compared to those with residual disease [3,4]. However, our ability to predict who will achieve pCR is limited, and biomarkers are needed to identify which cases will achieve pCR and can be spared additional treatment, and which cases fail on upfront NAST and will require a different treatment option. Our own work and that of others has shown that detection of breast cancer cell-free DNA (cfDNA) directly during NAST can predict response, and thus merits additional study [5,6].

cfDNA released to the blood from cancer cells, holds important information on the nature of the tumor. Informative features include cancer specific mutations and copy number alterations, as well as information on the cancer cell-of-origin, which include cell-type specific DNA modifications such as 5mC. In addition, the patterns of cfDNA fragmentation, including length and fragment ends were shown to reflect cell-of-origin [7]. Thus, cfDNA emerges as a promising resource for detection and classification of cancer, for monitoring disease progression/regression and providing predictive/prognostic insights [8,9]. One of the most promising cfDNA biomarker studies for multi cancer early detection was based on targeted methylation assay of over 100,000 genomic regions in 11,000 samples, to develop a classifier which has been validated in a large clinical study and is now in widespread use for cancer detection [10]. Unlike other cfDNA biomarkers which rely on genetic alterations, 5mC can also detect collateral damage to non-cancer cells surrounding the tumor.[11]. Most of the methylation-based approaches have used bisulfite-based approaches, but immunoprecipitation-based [12] and enzymatic [13] techniques have also shown promising results. For early detection, there has been some success using low-coverage whole genome sequencing to detect copy number alterations and cancer-specific fragmentation features [14,15]. After identification and biopsy and sequencing of a tumor, personal mutations can be used bioinformatically (i.e. “tumor-informed” approaches) in combination with either deep capture sequencing [16] or whole-genome sequencing [17,18] for ultra-sensitive detection of residual disease.

We recently applied Nanopore sequencing to cfDNA (cfNano). Nanopore sequencing is unique since it can distinguish 5mC from C in native DNA sequencing, making the methylation pattern an integral part of the sequence with no need for special sample processing or PCR amplification [19]. Compared to bisulfite-sequencing which tends to fragment DNA and introduce PCR bias [13,20], our implementation of cfNano was able to preserve cancer-specific CNA, fragmentomic, and DNA methylation information to classify cancer vs. non-cancer samples [19] We propose to integrate all three layers of information into one tool to maximize the accuracy of our Nanopore-based classification, and apply it to predict and to monitor response to NAST in BC.

Research objectives and expected significance

The long-term goal is to develop Nanopore-based multimodal analysis of cfDNA as an accessible tool used in clinical oncology, which can be applied to tumor detection, classification, and response prediction. We have chosen to focus our initial study on a clinical oncology problem that is feasible and where our technology will allow us to study several resistance mechanisms which are thought to involve changes in DNA methylation and CNA. The future potential of the Nanopore WGS approach is that it provides simple sample processing and sequencing which could allow rapid point-of-care analysis. However, we will also include best-in-class capture methylation sequencing, both to validate the relatively new cfNano approach, and to determine which of the two approaches is more informative for future development. **Major goals:** (i)Develop advanced within-read methylation and copy number-aware computational approaches for identification and deconvolution of circulating tumor DNA; (ii)Develop a targeted DNA methylation hybrid capture panel for detection and profiling of breast cancer cell-free DNA in plasma; (iii)Compare Nanopore-based whole-genome analysis to the targeted capture panel. Integrate CNA, fragmentomics, and methylation using an AI system for BC subtype classification, NAST response, and toxicity.

Significance:

1. Development of a new multi-modal technology for liquid biopsy. Adoption of Nanopore sequencing is increasing rapidly for clinical applications due to its advantages, which include the low buy-in cost and portable nature of the device. Nanopore sequencing is also rapid, with recent clinical demonstrations of end to end turnaround time from sample collection to tumor classification in as little as 1-3 hours [21,22]. DNA methylation can differentiate any cell types without relying on genetic changes, this technology development will likely have an impact outside of oncology, especially in emergency medicine where rapid diagnosis can be critical.
2. Development of innovative computational methods that can be applied not only to cfNano but to cfDNA bisulfite or enzymatic methylation sequencing (Aims 1,2). We incorporate several novel and important aspects. First, we will “purify” site-specific methylation levels using cell type composition inferred from global deconvolution, along with cancer cell copy number inferred from copy number segment analysis. This will provide us not only more accurate estimates of methylation at cell-of-origin marker sites, but will allow us to monitor other biologically relevant changes during cancer progression.
3. Unmet need: When treating local BC treatment plan is set based on tumour profiling and imaging. Absence of effective tools for monitoring response leads to overtreatment and undertreatment. Both platforms we explore can provide rich data for AI but at the same time will provide a dynamic biological picture of the tumor, allowing the physician to adjust the treatment earlier and more effectively. Two common scenarios illustrate this point: i) HER2+ patients treated with anti-HER2: Loss of *ERBB2* amplification can reflect tumor eradication in response to drug or may be a resistance mechanism allowing cells to thrive. The cfDNA picture should clarify this and guide clinical action. For instance, if cfDNA shows complete clearance, the patient would require no additional therapy; if cfDNA shows some residual ctDNA with *ERBB2* amplification present (incomplete response), escalate to TDX [23]; if cfDNA shows residual

ctDNA without *ERBB2* amplification, switch treatment and/or operate; if cfDNA indicates stress on other organs (toxicity), hold treatment; ii) The second common scenario is resistance to taxene-based NAST in TNBC, was recently found to involve an unusual pattern of global DNA hypomethylation, which created a vulnerability to epigenetic (EZH2i) therapy [24]. This change in methylation should be detectable early by cfDNA profiling and prompt a consideration of EZH2i. These are two specific scenarios we will pursue, but the approach is general and can be adapted to other tumours and different clinical scenarios.

Investigational team: We are an interdisciplinary team, with joint publications and required expertise for the different aspects of this project. Dr. Eden made fundamental discoveries regarding the role of global hypomethylation and genomic instability in oncogenesis, and is experienced withNanopore sequencing of genomic and cDNA, and co-developed our Nanopore cfDNA sequencing approach with Prof. Berman and Dr. Zick [19]. Prof. Berman is a pioneer in epigenetic whole-genome sequencing, producing the first deep human cancer methylome. He led The Cancer Genome Atlas (TCGA) whole-genome bisulfite sequencing project and was an analyst in the TCGA BC group. Dr. Zick is an M.D./Ph.D. Medical Oncologist with extensive applied research in BC, genome instability, cfDNA, and methylation. Working with several cfDNA groups at HUJI, Dr Zick initiated a large-scale operation for banking of plasma from cancer patients under Helsinki approval, which have been used in a number of high-profile cfDNA analysis papers, including a recent one on BC before and during NAST [5]. Prof. Kaplan is an expert in machine learning, with a strong background in DNA methylation deconvolution and cfDNA methylation analysis. His group produced the largest high quality WGBS atlas of diverse human cell types, allowing accurate fragment-level plasma deconvolution [25]. **Data management and availability.** Berman and Kaplan are experienced with sequencing consortia that involve extensive data management and availability. Prof. Berman was a PI in the NCI Genomic Data Analysis Network, a large network of groups sharing human cancer data and making it available for other researchers. A detailed explanation of our procedures is provided in the “Data management declaration”

Preliminary results.

We recently published our feasibility study of cfNano [19]. We showed that with 0.2x Nanopore sequencing coverage, we could detect copy number alterations and estimate tumor fraction using ichorCNA, and estimate tumor fraction even more sensitively using DNA methylation-based deconvolution. In the DNA methylation profiles, we detected both normal cell of origin marks such as lineage-specific transcription factor binding sites, as well as cancer-specific alterations such as global loss of methylation at “Partially Methylated Domains” (PMDs). We also showed that we could detect cancer-specific fragmentomic signals such as short mononucleosome fragments (100-150bp), and fragment end motifs such as CCCA [26,27].

The samples analyzed in [19] already include several plasma samples from the Hadassah Medical Organization (HMO) sequenced by Dr. Eden’s group. By deconvolution, using available methylation data, we detect changes in cell-type composition in patient plasma compared to healthy plasma. In patient plasma we also detect global methylation loss in published PMDs from WGBS in the TCGA project [28] (Fig. 1A-B). Using ichorCNA, we estimate tumor fraction and generate CNA which reproduces characteristic CRC CNA anueploidy profiles described in TCGA

[29] as well as focal ERBB2 amplification (Fig.1C). Whereas methylation-based deconvolution identified a small component of epithelial DNA in case HU004.01, ichorCNA could not detect any cancer DNA.

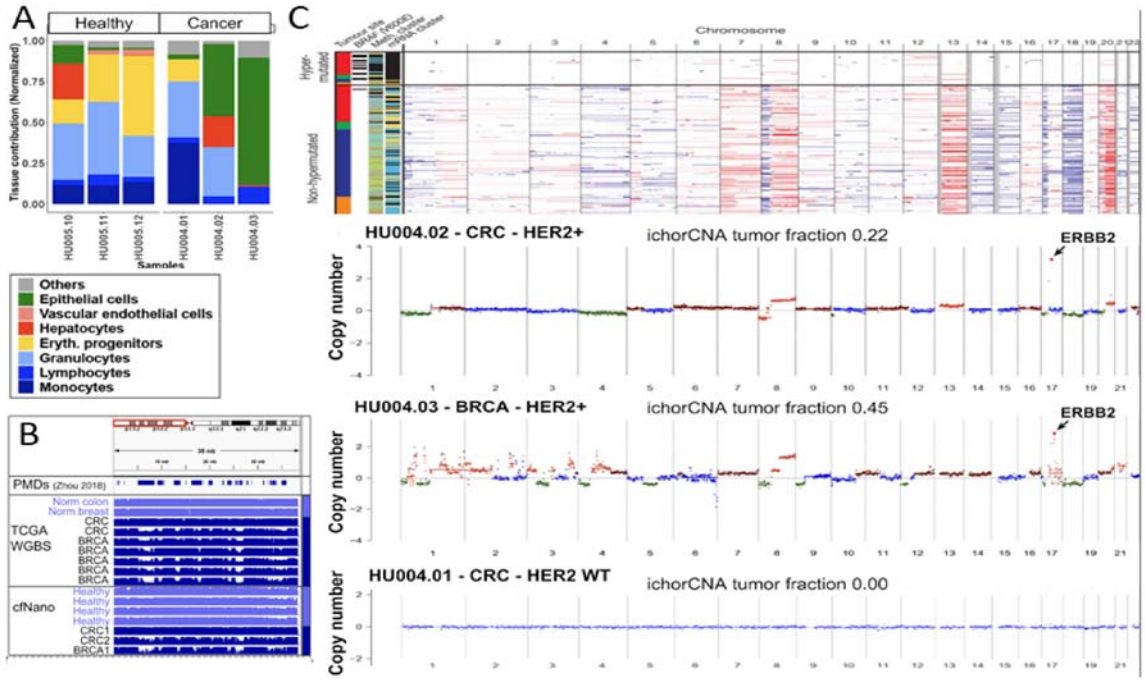

Figure 1: Estimating cell type fractions, hypomethylation and CNA a from cfNano:(**A**) Non-Negative Least Squares regression based on [19] was used to deconvolute cell types in healthy plasma cfDNA samples. (**B**) Average DNA methylation across chr16p, comparing cancer with Normal.Top: TCGA WGBS; Bottom:HUJI cfNano. PMDs from [28]. (**C**) ichorCNA, estimates tumor fraction and generates CNA profile that reproduces characteristic CRC CNA anueploidy profiles described in TCGA (Top) [29], as well as focal ERBB2 amplification.

Nanopore sequencing is a young and rapidly developing technology, with constant improvement in library preparation, sequencing chemistry, instrumentation, sequencing speed and analysis tools. Nanopore was originally optimized for longer reads. We have worked with Oxford Nanopore (ONT) experts to optimise protocols and software to better capture short reads and improve capture of methylation in short reads (see Letter of Collaboration from Spike Willcocks). We tested on cfDNA instrument configuration files from Oxford Nanopore Technology (ONT) which allow capture of short fragments from 20bp and up. ONT announced a new generation of flowcells (R10.4.1) and chemistry “Kit 14” (expected to be released in 22 Q3), with improved 99%+ basecalling accuracy and methylation data comparable to Bisulfite-seq. We are approved for early access to this new line and will begin testing for cfDNA. Our current protocol is relatively simple, and we do not expect problems sequencing new samples with this protocol. We are sequencing samples of healthy plasma mixed with serially diluted patient plasma (40% tumor fraction) to reach (0.5,1,2,4,8,16%). This dataset will serve to set the baseline sensitivity for tumor DNA detection (Nano or panel), and to monitor sensitivity during development of analysis pipeline. We expect that as more features are integrated into the analysis, sensitivity will improve. It will help us decide what coverage to aim for.

Work Plan

Aim 1: Improve cfNano computational tools. There are several major areas where we will improve the computational methods described above. The first subaim is the deconvolution of cancer DNA from mixed plasma DNA, in order to determine both the relative fraction of cancer to non-cancer cell-free DNA, as well as to purify the cancer DNA methylation levels across the genome. The second major subaim is to integrate the copy number, fragmentomic, and methylation signals into a single cancer detector.

Aim1a. Reference samples for deconvolution We and others have shown that cancer DNA in plasma can be detected based on methylation patterns from the healthy cell of origin, using reference-based deconvolution [13,30,31]. For our feasibility study [19], we used a reference atlas based on methylation array data, which covers less than 3% of all CpGs in the genome and only 13% of highly cell type-specific methylation markers[25]. For breast detection, we have access to a much richer dataset of deep whole-genome (WGBS) data for FACS-purified breast epithelial cell types, blood, and other relevant cell types (Table 1). The new HUJI Atlas recently published by Prof. Kaplan contains 207 purified methylomes of purified WGBS healthy samples, sequenced at 30x [25]. In addition, we have obtained 10 FACS-purified breast cancer luminal epithelial WGBS datasets for an additional 10 donors from [32]. We also have plasma WGBS from 32 healthy donors from [33] and 30 healthy donors from [13], and will collect additional datasets from new studies. Additionally, we have downloaded 20x WGBS data of 30 primary breast tumors from BASIS [34] and 6 primary breast tumors from TCGA [28], as well as Reduced Representation Bisulfite Sequencing (RRBS) data for 1527 tumors and 244 healthy breast biopsies from METABRIC [35]. Importantly, we have already obtained access to read-level data for all datasets, which is critical for our methylation-based deconvolution method.

We will begin development of new computational methods below before our Nanopore sequencing is completed. In order to facilitate this, we will use a deep (10x) methylation sequencing dataset of cell-free DNA from 30 healthy controls, 32 pancreatic cancer cases, and 21 liver cancer cases [13]. We have obtained the raw read-level data, which allows us to analyze fragmentomic and copy number features as well as methylation. In order to approximate our lower coverage Nanopore sequencing, we will randomly downsample reads to create hundreds of virtual samples of ~0.2-1x coverage, allowing us to gauge sensitivity of our methods to low coverage. Importantly, the HUJI WGBS Atlas contains the corresponding cells-of-origin for this analysis.

Aim 1b. Within-read and copy number-aware approaches to deconvolution: The methylation ratio observed in plasma is a mixture of cancer cell DNA and DNA derived from other cells (mostly white blood cells). Deconvolution of mixtures of cell types from methylation data has been studied extensively in solid tissues [36] and to a degree in cfDNA [13,30,31,37]. Since most large methylation studies have used a methylation array platform (Infinium) which assays individual CpGs, the local coordination of multiple CpGs within individual reads has been largely neglected. For sequencing data, the power of these “within-read” approaches for deconvolution has been demonstrated with methods such as amrfinder [38], MethylPurify [39], and Methyl haplotypes [40], yet Prof. Kaplan’s group was recently among the first to apply fragment-level analysis to cfDNA deconvolution [25].

The second important factor is that the ratio of cancer to non-cancer DNA varies across the genome as a function of cancer copy number, yet this has not been taken into account in any ctDNA detection deconvolution method. This will be especially important for BC, which is largely a copy number (CN) driven cancer type [41]. Recently, the CAMDAC method was developed to explicitly model CN to deconvolute cancer from non-cancer methylation levels within tumor tissues, to provide a "purified" cancer cell methylation profile [42]. It was shown that these purified profiles had more bimodal methylation states and performed significantly better at unsupervised clustering of tumor samples from the same individual. Unfortunately, the CAMDAC method was not designed to take into account within-read methylation information, and was not applied to cell-free DNA.

The CelFIE method [37] used an Expectation-Maximization (E-M) based probabilistic model to deconvolute two or more cell types within mixed methylation sequencing data, using cell-type specific regions identified from a WGBS reference, and performed significantly better than previous approaches such as non-negative least squares [30]. However, the CelFIE model did not take within-read information into account. Prof. Berman's group has been working on an E-M model based on MethylPurify [39], which does take within-read information into account but can only model two cell types rather than multiple cell types as in CelFIE. They have also adapted the original Celfie model to incorporate within-fragment information. Both models perform significantly better than Celfie in simulation and in silico mixture experiments (data not shown). We are currently working to compare these models to the UMX within-fragment approach recently published by Prof. Kaplan [25], as well as another unpublished approach developed by his lab. We will choose the most accurate of these deconvolution models, or employ a consensus approach if they give complementary results.

Almost all current cancer cfDNA approaches assume copy number to be uniform across the genome. As we showed in our preliminary study [19] there is a strong influence of copy number on observed DNA methylation levels in the mixed plasma. We will address this by using only cell type markers located within diploid regions of the genome for deconvolution (Figure 2A). We are currently evaluating the added benefit of this copy-number aware model using simulation and in silico mixture experiments, but we expect a significant benefit. For samples where ichorCNA is able to call copy number, we will use sample-specific copy number calls. However, we expect ichorCNA to fail on many samples with lower tumor fraction. In these samples, we will rely on the large number of breast tumors analyzed for copy number status in the METABRIC consortium [25,43,44], picking regions from chromosome arms with very low rates of copy number alteration in breast cancer.

Aim 1c. Purification of cancer-specific methylation changes: We have introduced a number of improvements above that will greatly improve the accuracy and sensitivity of cancer fraction estimation. Once we have these estimates, we can use a relatively straightforward approach introduced as "Copy number-aware Methylation Deconvolution Analysis of Cancers" or CAMDAC [42], to "purify" the methylation patterns of the cancer cells. This will be important for capturing methylation changes that occur after transformation, such as global hypomethylation (Figure 2B). The CAMDAC method relies on both the global cancer fraction (tumor "purity") and copy number state. In the case of tumor tissue, these can be estimated from SNP information within CN

alterations. For cfNano, we will use the highly accurate cancer fraction derived from methylation-based deconvolution, and copy number state derived from ichorCNA. A representative methylation profile is required for the non-cancer component of the mixture, which is bioinformatically "subtracted" from the mixture. In our case, we will use methylation levels from deep (x85) WGBS sequencing of healthy plasma [45].

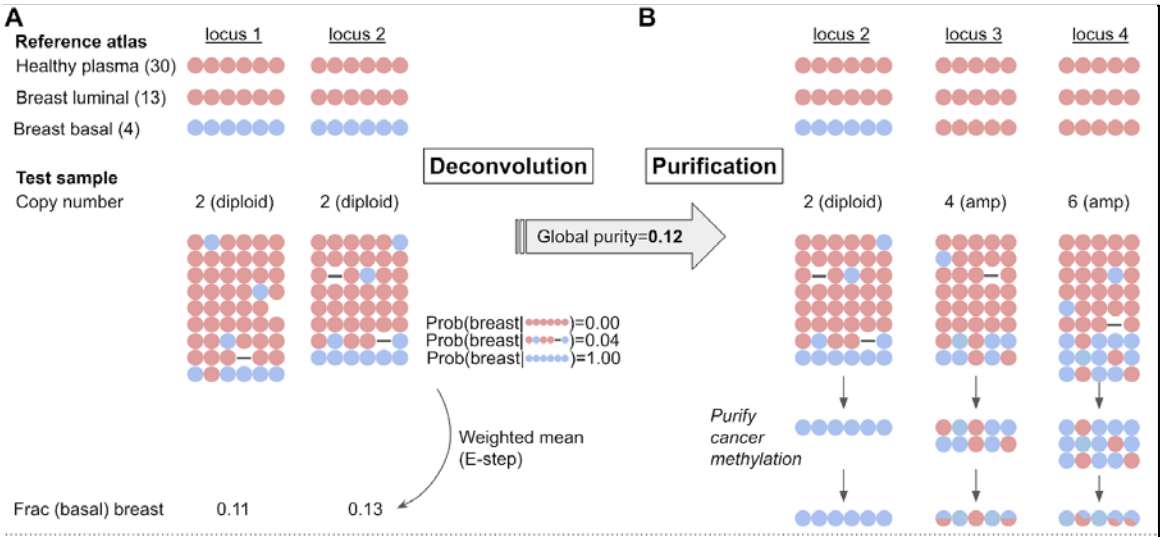

Figure 2: Copy number aware deconvolution and purification of circulating tumor DNA. (A) deconvolution using an Expectation-Maximization method based on MethylPurify and CelFIE, with modifications. We use only diploid regions based on copy number analysis to improve estimates. This results in a cell fraction (global purity). (B) Purity is used for purification in a copy-number aware subtraction of non-cancer DNA based on the CAMDAC approach.

Importantly, CNA states from ichorCNA will be unknown for some samples, which will reduce the accuracy of non-cancer purification. For this problem, we will implement a novel method to call CNA domains based on E-M based deconvolution levels. Since each breast cell specific region yields a cancer fraction in the deconvolution step, we should be able to call domains of increased or decreased copy number using a standard CNA segmentation approach used for SNP based CNA calling such as sliding window, hidden markov model, or changepoint analysis [46]. The input to this segmentation will be both the methylation-based cancer fractions and the read coverage, which we expect to increase sensitivity over read coverage alone (which is the only input to ichorCNA). If not, we will revert to assuming diploid status for these samples.

Aim 1d. Integrate CNA, fragmentomic, and methylation features for cancer detection: In our feasibility project, we showed that we could detect cancer-specific features of DNA methylation, copy number alterations, and fragmentation from Nanopore data [19]. We have powered our study similar to other circulating tumor DNA studies that use machine learning classifiers to detect cancer vs. healthy control samples. [14,47] used a gradient tree boosting classifier to detect multiple cancer types with 1-2x whole-genome sequencing (WGS), with only the fraction of short mononucleosomes (<150bp) in non-overlapping 5Mbp genomic bins as input. Similarly, [15] used stacked/ensemble machine learning (integrating five different classes of machine learning algorithm) with inputs of the fraction of both short (<150bp) and long (300-500bp) reads, along with 4-mer fragment end motifs, to detect early stages colorectal adenocarcinomas and advanced adenomas using 4x WGS. We showed that in our feasibility study that cancer-specific end motifs (thought to derive from DNASE1L3 activity in tumors) were detectable in Nanopore cancer

samples [19]. Siejka [13] used a Support Vector Machine (SVM) classifier with inputs of short and long fragments, and cancer-specific hypomethylation based on 10x WGBS, to detect liver and pancreatic cancers.

Guided by the sample size of these earlier studies, we will use the first 3/4 of our dataset (60 breast cancer cases and 25 healthy controls) to train models, holding out the second half as an independent validation dataset. Within the training dataset, we use leave one out cross-validation to evaluate performance and tune parameters. In addition to the feature inputs suggested by [15], we will also include methylation features including global hypomethylation as in the Siejka-Zielinska study above [13]. Importantly, we believe that methylation levels that are purified based on tumor fraction and copy number status will increase the accuracy of our prediction. We will begin using gradient boosting method, but will also test random forest, Support Vector Machine, and deep learning classifiers.

Aim 1 alternative methods: Copy number analysis by ichorCNA is based on read counts alone and has limited sensitivity. More accurate methods, such as CAMDAC/ASCOT and ABSOLUTE [42], typically use heterozygous SNPs (B allele fraction). While we do not have power to call individual SNPs with our 1x Nanopore WGS, we will include a backbone of common SNPs on our capture panel (Aim 2). Since the samples will be matched, this will help us to identify complex copy number issues including whole-genome doubling events [48], and monitor their effects on deconvolution in our Nanopore data. As an example, we have one sample in our lung adenocarcinoma feasibility Nanopore study [19] where an apparent mis-estimate of copy number may be due to WGD, which occurs in approximately 70% of lung adenocarcinomas [42].

Aim 2: Development of Methylation hybrid capture panel

Nanopore-based plasma analysis will be compared to Illumina sequencing following bisulfite/enzymatic conversion and enrichment using a specialized methylation sequencing panel from TWIST Bioscience. These custom-made capture panels by TWIST are widely used (e.g. by GRAIL's GALLERI platform [10]), and allow deep sequencing (500-1000x) of 1,000-10,000 target regions across the genome. We have already designed and analyzed similar TWIST panels.

Aim 2a. Selection of target regions: In Aim 2a we will design a targeted DNA methylation panel for plasma deconvolution of breast cancer plasma before and following treatment. The panel will include ~1000 differentially methylation regions that are uniquely unmethylated regions in specific cell types and methylated elsewhere in the human body. For this, we will harness our human atlas [25], which includes dozens of DNA methylation marker regions identified for each cell type present in the plasma.We previously showed that cell-free DNA is mostly derived from blood and immune cells, with a small fraction from hepatocytes and endothelial cells [25,30]. We will include 25-50 uniquely unmethylated genomic regions for each of these cell types (monocytes, granulocytes, B, T, NL cells, macrophages, erythroblasts, megakaryocytes, hepatocytes, and endothelial cells from various tissues). For monitoring collateral damage [11], we will also include ~100 markers for normal breast epithelium (basal and luminal epithelial cells), as well as brain, lung and bone. Finally, we will use the reference BC methylome samples above (Aim 1a) to identify regions differentially methylated in breast cancer in general, or in specific breast cancer subtypes (triple negative, ER+, and HER2+). Because tumor reference data comes from bulk tumor data, we

will be careful to exclude DMRs that result from the contamination of non-cancer cells. We mainly target CpG-rich regions, where multiple CpGs are covered by each sequenced read, thus further contributing to the specificity and sensitivity of fragment-level analysis [25].

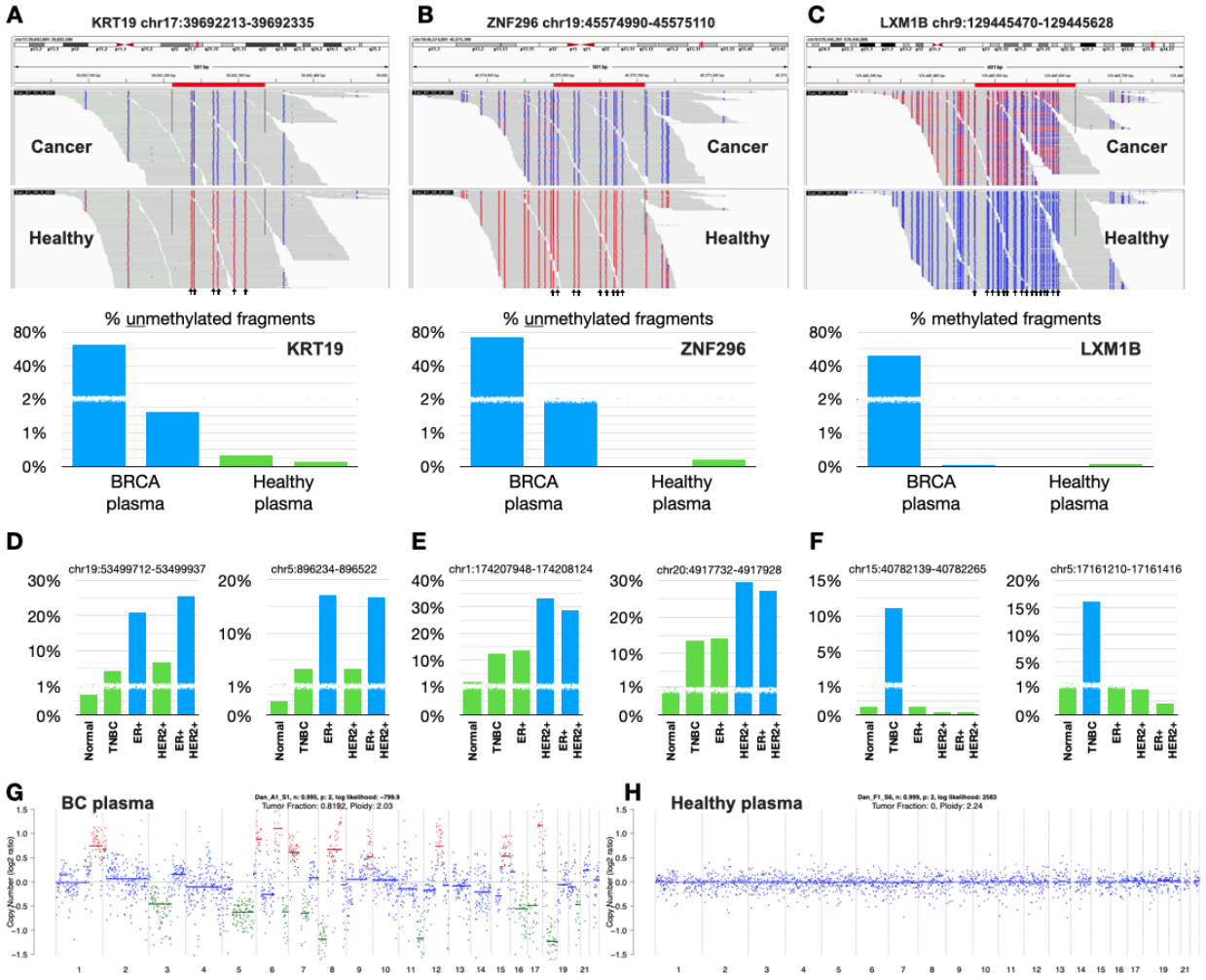

Figure 3. (A-C) Plasma cfDNA sequencing following hybrid capture at three breast cancer markers from [5]. Sequenced reads are shown in gray, with unmethylated CpGs in blue and methylated CpGs in red. Black arrows mark CpGs inside targeted regions. Fragment-level analysis, shown as barplots, highlight the percent of fragments unmethylated in all target CpGs (per molecule), across two plasma samples from breast cancer patients (one advanced and one early, shown in blue), and two healthy controls (green). **(D-F)** Fragment-level analysis from METABRIC RRBS data. Shown are (D) two genomic loci methylated in ER+ tumor samples (blue) but not in normal biopsy, TNBC or ER-/HER2+ samples (green), (E) two regions differentially methylated in HER2+ tumors (blue), compared to normal biopsy, TNBC and ER+ tumors (green), and (F) two regions differentially methylated in TNBC tumors, but not in healthy biopsy, or in ER+/HER2+ tumors. Y-axis: percent of unmethylated fragments. **(G-H)** ichorCNA analysis of hybrid capture panel off-target loci. A typical off-target rate of 20-30% provides shallow (0.2-0.4x) WGBS sequencing, which could be used for copy number analysis. Shown are the same plasma samples from (A-C), including plasma from an advanced breast cancer patient (G) and plasma from a healthy donor (H).

Our preliminary results (Figure 3A-C) show plasma DNA, captured using a similar TWIST panel designed based on the human methylation atlas. Specifically, it demonstrates the use of hybrid capture analysis at three breast cancer markers [5]. Indeed, fragment-level methylation analysis of plasma DNA identifies cancer-like fragments for two BC patients, but not in healthy individuals. As with the cfNano, initial tests with the panel will exploit plasma serial dilution set described in preliminary results, to evaluate reproducibility of panel based tumor fraction.

Aim 2b. Interpretation of panel data: tumor fraction and subtype analysis: Sequenced reads will be analyzed using wgbstools and UXM, two software packages we have recently developed for the analysis of DNA methylation atlas data [25]. Specifically, we will take a fragment-level approach by calculating the fraction of “methylated”, “unmethylated”, or “inconsistent” fragments from each genomic locus. These will be used for plasma deconvolution, highlighting the cellular composition of plasma DNA. Specifically, the entire set of cancer markers will be used for inferring the tumor load, whereas subtype-specific markers identify the tumor composition before and following treatment [5,25,30]. With an expected depth of 500-1000x across ~1000 regions, multiple tumor-derived fragments are expected to be sequenced in each locus, allowing direct unbiased comparison with the methylation signals obtained from Nanopore sequencing from the same plasma samples. Figures 3D-F shows three genomic regions that present differential methylation levels at breast cancer subtypes. These regions were selected from the METABRIC RRBS data [35] which covers ~1% of the human genome, suggesting that many more specific markers could be identified by analysis of WGBS data. Importantly, our human methylation atlas suggests their methylation pattern is unique to cancer, allowing their use as plasma cell-free DNA markers.

Aim 2c. Copy number analysis from panel off-targets (shallow WGBS): Finally, we were curious to see if hybrid capture data could also be used for copy number analysis of cell-free DNA. For this, we analyzed the read coverage of eight plasma samples analysed by TWIST panels. Targeted regions were sequenced at an average depth of 650x, while the remaining genome (off targets) was sequenced at 0.2-0.4x. As Figure 3G-H shows, this shallow WGBS of cfDNA can be used for copy number analysis, revealing multiple CNAs in the plasma of the HER2+ breast cancer patient (G) but not in healthy plasma (H). CNA analysis from Nanopore sequencing will be compared to this parallel analysis, based on this shallow genome-wide Illumina sequencing of off-targets. For a set of samples we will also perform shallow wgs on tumor DNA from FFPE to evaluate/validate the ichorCNA, although cfDNA may be considered more faithful to reality.

Aim 3: BC study for detection, subtype classification, drug response and toxicity.

Aim 3a: Collect and sequence non-cancer controls: We will collect 30 healthy controls which will serve as a baseline and allow us to quantify specificity, sensitivity, and positive predictive value (PPV) and negative predictive value (NPV) of BC at time of diagnosis. We have already more than 200 healthy women of various age and enthnic background plasma, and continue collection. It is especially important to age match controls to the extent possible, since methylation changes during aging. The number 30 was determined based on similar pre-clinical whole-genome cfDNA sequencing studies [12,13,31], where we would expect to have similar or better sensitivity.

Aim 3b. Collect and sequence neoadjuvant breast cancer patients: Women treated with BC NAST in HMC will be offered to participate in the study. In women who consent we will draw blood before biopsy, three times during NAST (before treatment); and prior breast surgery (Fig. 4A). We aim to collect and sequence 40 BC patients/year for 2 years, to complete sequencing by year 3. This includes at least 15HER2+ and 15TNBC cases per year, to allow us to study particular resistance scenarios of interest (loss of ERBB2 amplification, and global hypomethylation).

Our research is conducted as part of an approved clinical trial "Characterization of genetic

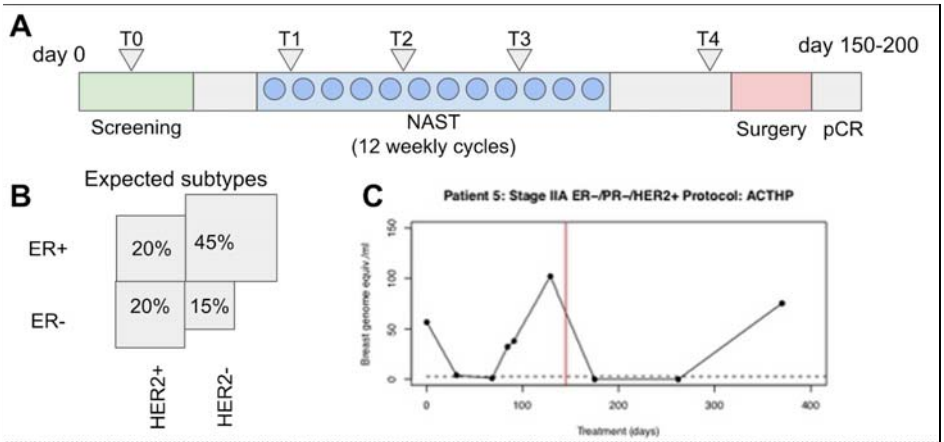

Figure 4: cfNano study of neoadjuvant breast cancer treatment. (A) Planned time course of blood collections at timepoints T0-T4. (B) Expected subtypes based on previous patient profiles. (C) Example of a similar timecourse of neoadjuvant treatment from our earlier cfDNA methylation study (reproduced from Moss, Zick et al. Annals Onc. 2020.)

material, including total genome/exome sequencing, from tumor cells as a predictive and prognostic measure, and as a tool to follow up treatment in cancer patients" study HMO-346-12. HMO is a leading center in Israel for diagnosis and treatment of BC, performing over 500 breast biopsies a year. Dr. Zick's group has collected and banked plasma samples for cfDNA analysis for the past 10 years, an effort involving a study coordinator, nurse practitioner, and lab assistant who process and link biospecimens to clinical data in a RedCap research database.

Aim 3c. Evaluate detection of known breast cancer CNA markers: Different types of cancer have different relative contributions of single-nucleotide mutations vs. copy number alterations to oncogenesis. These have been described as M-class cancers dominated by mutations and C-class cancers dominated by copy number events, and BC is among the prototypical C-class cancers [41,49]. Along with global RNA expression, global copy number profiles are highly informative for classifying breast cancer cases into clinically-relevant subtypes [43,44]. Notably, most ER-/HER2+ cases fall into a single cluster, as do most ER-/HER2- (triple-negative, TNBC). In a recent long-term BC study, these integrated copy number (IntClust) clusters could identify subgroups of both ER+ and ER- cases at high risk of relapse up to 20 years after diagnosis, which improved the prediction of late, distant relapse beyond what is possible with clinical covariates [44]. We expect a significant number of cases of each of the four major clinical subtypes of breast cancer in our cohort (Fig. 4B). We will call CN segments using ichorCNA [50] and classify cases based on IntClust clusters, and quantify our ability to make confident assignments. Dr. Zick's group routinely performs low-coverage (2x) Illumina whole genome sequencing (lcWGS) on FFPE tissues that are banked by the Pathology department to identify CNA. We will obtain and sequence tumor tissue for as many cases as possible to compare to cfDNA IntClust assignments.

Detection of clinically actionable amplifications, especially *ERBB2*, is critical for targeted treatment of BC. In our preliminary Nanopore sequencing, we tested a *ERBB2*-positive colorectal cancer (CRC) patient, and detected a highly amplified *ERBB2* by our ichorCNA pipeline (Fig. 1c). Using ichorCNA, we will analyze our sensitivity to detect *ERBB2* amplification and another potentially targetable amplification, *FGFR1,* which was found to be amplified in similar numbers of BC patients *as ERBB2* in The Cancer Genome Atlas cohort [49]. The FGFR1 inhibitor Erdafitinib has been approved for bladder cancer and is currently in a clinical trial for treatment of BRCA [51].

Importantly, we seek to detect anti-HER2 sensitivity in resistance. Therefore, we will evaluate our ability to call *ERBB2* amplification in samples that have residual ctDNA during and after treatment,

when the total fraction of ctDNA is low and more challenging to analyze. In cases where standard CNA detection by ichorCNA fails, we believe adding methylation and fragmentomic features can increase sensitivity. In our cfNano samples, we have observed that proportion of short mononucleosome fragments (100-150bp) increases within the ERBB2 amplicon, due to the increased fraction of cancer-derived DNA. The ERBB2 amplicon is generally long enough (200kb) to detect cancer-specific differentially methylated regions as well as sequence-specific signatures of global hypomethylation that we developed [28], so that we should be able to define a combined model of amplification that includes increased read depth overall (ichorCNA), as well as increased frequencies of cancer-specific fragmentation features (short mononucleosomes) and cancer-specific methylation. We will quantify our accuracy to call potentially actionable alterations (*ERBB2*, FGFR1) using the lcWGS data on matched FFPE samples described above.

Aim 3d: Machine learning classifier of treatment response: Here, we will evaluate the ability of cfDNA to predict specific endpoints of NAST response. First, we will use cfDNA features to predict pathological complete response (pCR), both across all cases, and within each of the three major NAST treatment groups. We will then evaluate the relative contribution of each cfNano feature (methylation tumor fraction, copy number alterations, and fragmentomic features including length distribution and fragment-end motifs) by building a regression model that incorporates each individual feature [13]. Additionally, we will train machine learning classifiers to integrate these features, using decision trees, random forests, gradient boosting trees, or Support Vector Machines [52]. Additionally, we will develop fragment-level multimodal classifiers, that will integrate information about the position of each plasma fragment (e.g. in CNA regions), its methylation status, its length and end motifs, and calculate the likelihood of this fragment to have been released from a cancer or a healthy cell. Each classifier will be trained independently for cfNano data and DNA capture panel data, allowing us to compare the predictive power of the two techniques, and determine whether they are redundant or complementary.

HER2-positive cases that do not respond well present an opportunity in this study. Importantly, up to 27% of initially HER2+ cases will be HER2- in resistance [53]. We will have pathology and tumor lcWGS for most of the post-NAST tumors, and we will be able to quantify the accuracy of our cfDNA assays to detect ERBB2 loss during treatment. However, we expect to see cases of disagreement where cfDNA indicates ERBB2 remains amplified, but intra-tumor heterogeneity leads to a single point sampling of the tumor to be negative. In one major study, amplifications found by cfDNA were not found in the matched tumor biopsy in up to 78% of cases [8].

Conclusion: The cfDNA-assisted approach can help guide the treatment plan with a very short turnaround time, and we will test a well established methylation-based method with a new and promising point-of-care one. These can assist the physician and patient to choose which treatment is needed at every time point, by monitoring multiple genomic and epigenomic levels before treatment failure and adverse events become clinically evident. These methods will also shed light on the biology tumour resistance and guide new therapeutic strategies. Breast cancer is studied because it is the most common and best characterised tumour, but this methodology can be adapted to other treatment settings and tumour types. As new cancer therapeutic options continue to give hope for patients, the cfDNA guide to choose when and how to use them is at hand.

Bibliography

1. Gianni L, Pienkowski T, Im YH, Roman L, Tseng LM, Liu MC, et al. Efficacy and safety of neoadjuvant pertuzumab and trastuzumab in women with locally advanced, inflammatory, or early HER2-positive breast cancer (NeoSphere): a randomised multicentre, open-label, phase 2 trial. Lancet Oncol. 2012;13: 25–32.

2. van Ramshorst MS, van der Voort A, van Werkhoven ED, Mandjes IA, Kemper I, Dezentje VO, et al. Neoadjuvant chemotherapy with or without anthracyclines in the presence of dual HER2 blockade for HER2-positive breast cancer (TRAIN-2): a multicentre, open-label, randomised, phase 3 trial. Lancet Oncol. 2018;19: 1630–1640.

3. Jackisch C, Stroyakovskiy D, Pivot X, Ahn JS, Melichar B, Chen SC, et al. Subcutaneous vs Intravenous Trastuzumab for Patients With ERBB2-Positive Early Breast Cancer: Final Analysis of the HannaH Phase 3 Randomized Clinical Trial. JAMA Oncol. 2019;5: e190339.

4. Huober J, Holmes E, Baselga J, de Azambuja E, Untch M, Fumagalli D, et al. Survival outcomes of the NeoALTTO study (BIG 1-06): updated results of a randomised multicenter phase III neoadjuvant clinical trial in patients with HER2-positive primary breast cancer. Eur J Cancer. 2019;118: 169–177.

5. Moss J, Zick A, Grinshpun A, Carmon E, Maoz M, Ochana BL, et al. Circulating breast-derived DNA allows universal detection and monitoring of localized breast cancer. Ann Oncol. 2020;31: 395–403.

6. Magbanua MJM, Swigart LB, Wu H-T, Hirst GL, Yau C, Wolf DM, et al. Circulating tumor DNA in neoadjuvant-treated breast cancer reflects response and survival. Ann Oncol. 2021;32: 229–239.

7. Lo YMD, Han DSC, Jiang P, Chiu RWK. Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies. Science. 2021;372. doi:10.1126/science.aaw3616

8. Parikh AR, Leshchiner I, Elagina L, Goyal L, Levovitz C, Siravegna G, et al. Liquid versus tissue biopsy for detecting acquired resistance and tumor heterogeneity in gastrointestinal cancers. Nat Med. 2019;25: 1415–1421.

9. Awad MM, Liu S, Rybkin II, Arbour KC, Dilly J, Zhu VW, et al. Acquired Resistance to KRAS Inhibition in Cancer. N Engl J Med. 2021;384: 2382–2393.

10. Klein EA, Richards D, Cohn A, Tummala M, Lapham R, Cosgrove D, et al. Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set. Ann Oncol. 2021;32: 1167–1177.

11. Lubotzky A, Zemmour H, Neiman D, Gotkine M, Loyfer N, Piyanzin S, et al. Liquid biopsy reveals collateral tissue damage in cancer. JCI Insight. 2022;7. doi:10.1172/jci.insight.153559

12. Shen SY, Singhania R, Fehringer G, Chakravarthy A, Roehrl MHA, Chadwick D, et al. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. Nature. 2018;563: 579–583.

13. Siejka-Zielińska P, Cheng J, Jackson F, Liu Y, Soonawalla Z, Reddy S, et al. Cell-free DNA TAPS provides multimodal information for early cancer detection. Sci Adv. 2021;7: eabh0534.

14. Cristiano S, Leal A, Phallen J, Fiksel J, Adleff V, Bruhm DC, et al. Genome-wide cell-free DNA fragmentation in patients with cancer. Nature. 2019;570: 385–389.

15. Ma X, Chen Y, Tang W, Bao H, Mo S, Liu R, et al. Multi-dimensional fragmentomic assay for ultrasensitive early detection of colorectal advanced adenoma and adenocarcinoma. J

Hematol Oncol. 2021;14: 175.

16. Tie J, Cohen JD, Wang Y, Christie M, Simons K, Lee M, et al. Circulating Tumor DNA Analyses as Markers of Recurrence Risk and Benefit of Adjuvant Therapy for Stage III Colon Cancer. JAMA Oncol. 2019;5: 1710–1717.

17. Wan JCM, Heider K, Gale D, Murphy S, Fisher E, Mouliere F, et al. ctDNA monitoring using patient-specific sequencing and integration of variant reads. Sci Transl Med. 2020;12. doi:10.1126/scitranslmed.aaz8084

18. Widman AJ, Shah M, Øgaard N, Khamnei CC, Frydendahl A, Deshpande A, et al. Machine learning guided signal enrichment for ultrasensitive plasma tumor burden monitoring. bioRxiv. 2022. p. 2022.01.17.476508. doi:10.1101/2022.01.17.476508

19. Katsman E, Orlanski S, Martignano F, Fox-Fisher I, Shemer R, Dor Y, Zick A, Eden A, Petrini I, Conticello SG, Berman BP. Detecting cell-of-origin and cancer-specific features of cell-free DNA with Nanopore sequencing. Genome Biol.

20. Erger F, Nörling D, Borchert D, Leenen E, Habbig S, Wiesener MS, et al. cfNOMe — A single assay for comprehensive epigenetic analyses of cell-free DNA. Genome Med. 2020;12: 1–14.

21. Kuschel LP, Hench J, Frank S, Hench IB, Girard E, Blanluet M, et al. Robust methylation-based classification of brain tumors using nanopore sequencing. medRxiv. 2021; 2021.03.06.21252627.

22. Djirackor L, Halldorsson S, Niehusmann P, Leske H, Kuschel LP, Pahnke J, et al. EPCT-15. RAPID EPIGENOMIC CLASSIFICATION OF BRAIN TUMORS ENABLES INTRAOPERATIVE NEUROSURGICAL RISK MODULATION. Neuro Oncol. 2021;23: i50–i50.

23. Cortes J, Kim SB, Chung WP, Im SA, Park YH, Hegg R, et al. Trastuzumab Deruxtecan versus Trastuzumab Emtansine for Breast Cancer. N Engl J Med. 2022;386: 1143–1154.

24. Deblois G, Tonekaboni SAM, Grillo G, Martinez C, Kao YI, Tai F, et al. Epigenetic Switch-Induced Viral Mimicry Evasion in Chemotherapy-Resistant Breast Cancer. Cancer Discov. 2020;10: 1312–1329.

25. Loyfer N, Magenheim J, Peretz A, Cann G, Bredno J, Klochendler A, et al. A human DNA methylation atlas reveals principles of cell type-specific methylation and identifies thousands of cell type-specific regulatory elements. bioRxiv. 2022. p. 2022.01.24.477547. doi:10.1101/2022.01.24.477547

26. Jiang P, Sun K, Peng W, Cheng SH, Ni M, Yeung PC, et al. Plasma DNA End-Motif Profiling as a Fragmentomic Marker in Cancer, Pregnancy, and Transplantation. Cancer Discov. 2020;10: 664–673.

27. Chan RWY, Serpas L, Ni M, Volpi S, Hiraki LT, Tam L-S, et al. Plasma DNA Profile Associated with DNASE1L3 Gene Mutations: Clinical Observations, Relationships to Nuclease Substrate Preference, and In Vivo Correction. Am J Hum Genet. 2020;107: 882–894.

28. Zhou W, Dinh HQ, Ramjan Z, Weisenberger DJ, Nicolet CM, Shen H, et al. DNA methylation loss in late-replicating domains is linked to mitotic cell division. Nat Genet. 2018;50: 591–602.

29. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012;487: 330–337.

30. Moss J, Magenheim J, Neiman D, Zemmour H, Loyfer N, Korach A, et al. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. Nat Commun. 2018;9: 5068.

31. Cheng THT, Jiang P, Teoh JYC, Heung MMS, Tam JCW, Sun X, et al. Noninvasive Detection of Bladder Cancer by Shallow-Depth Genome-Wide Bisulfite Sequencing of Urinary Cell-Free DNA for Methylation and Copy Number Profiling. Clin Chem. 2019;65: 927–936.

32. Senapati P, Miyano M, Sayaman RW, Basam M, Trac C, Leung A, et al. Aging leads to DNA methylation alterations associated with loss of lineage fidelity and breast cancer in mammary luminal epithelial cells. bioRxiv. 2021. p. 2020.06.26.170043. doi:10.1101/2020.06.26.170043

33. Sun K, Jiang P, Chan KCA, Wong J, Cheng YKY, Liang RHS, et al. Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. Proc Natl Acad Sci U S A. 2015;112: E5503–12.

34. Brinkman AB, Nik-Zainal S, Simmer F, Rodríguez-González FG, Smid M, Alexandrov LB, et al. Partially methylated domains are hypervariable in breast cancer and fuel widespread CpG island hypermethylation. Nat Commun. 2019;10: 1–10.

35. Batra RN, Lifshitz A, Vidakovic AT, Chin S-F, Sati-Batra A, Sammut S-J, et al. DNA methylation landscapes of 1538 breast cancers reveal a replication-linked clock, epigenomic instability and cis-regulation. Nat Commun. 2021;12: 5406.

36. Chakravarthy A, De Carvalho DD. Using epigenetic data to estimate immune composition in admixed samples. Methods Enzymol. 2020;636: 77–92.

37. Caggiano C, Celona B, Garton F, Mefford J, Black B, Lomen-Hoerth C, et al. Estimating the rate of cell type degeneration from epigenetic sequencing of cell-free DNA. Cold Spring Harbor Laboratory. 2020. p. 2020.01.15.907022. doi:10.1101/2020.01.15.907022

38. Fang F, Hodges E, Molaro A, Dean M, Hannon GJ, Smith AD. Genomic landscape of human allele-specific DNA methylation. Proceedings of the National Academy of Sciences. 2012;109: 7332–7337.

39. Zheng X, Zhao Q, Wu H-J, Li W, Wang H, Meyer CA, et al. MethylPurify: tumor purity deconvolution and differential methylation detection from single tumor DNA methylomes. Genome Biology 2014 15:7. 2014;15: 1–13.

40. Guo S, Diep D, Plongthongkum N, Fung H-L, Zhang K, Zhang K. Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. Nat Genet. 2017;49: 635.

41. Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C. Emerging landscape of oncogenic signatures across human cancers. Nat Genet. 2013;45: 1127–1133.

42. Cadieux EL, Tanić M, Wilson GA, Baker T, Dietzen M, Dhami P, et al. Copy number-aware deconvolution of tumor-normal DNA methylation profiles. Cold Spring Harbor Laboratory. 2020. p. 2020.11.03.366252. doi:10.1101/2020.11.03.366252

43. Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486: 346–352.

44. Rueda OM, Sammut S-J, Seoane JA, Chin S-F, Caswell-Jin JL, Callari M, et al. Dynamics of breast-cancer relapse reveal late-recurring ER-positive genomic subgroups. Nature. 2019;567: 399–404.

45. Fox-Fisher I, Piyanzin S, Ochana BL, Klochendler A, Magenheim J, Peretz A, et al. Remote immune processes revealed by immune-derived circulating cell-free DNA. Elife. 2021;10. doi:10.7554/eLife.70520

46. Ross EM, Haase K, Van Loo P, Markowetz F. Allele-specific multi-sample copy number

segmentation in ASCAT. Bioinformatics. 2021;37: 1909–1911.

47. Mathios D, Johansen JS, Cristiano S, Medina JE, Phallen J, Larsen KR, et al. Detection and characterization of lung cancer using cell-free DNA fragmentomes. Nat Commun. 2021;12: 5060.

48. Taylor AM, Shih J, Ha G, Gao GF, Zhang X, Berger AC, et al. Genomic and Functional Approaches to Understanding Cancer Aneuploidy. Cancer Cell. 2018;33: 676–689.e3.

49. Sanchez-Vega F, Mina M, Armenia J, Chatila WK, Luna A, La KC, et al. Oncogenic Signaling Pathways in The Cancer Genome Atlas. Cell. 2018;173: 321–337.e10.

50. Adalsteinsson VA, Ha G, Freeman SS, Choudhury AD, Stover DG, Parsons HA, et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. Nat Commun. 2017;8: 1324.

51. Mayer IA, Haley BB, Abramson VG, Brufsky A, Rexer B, Stringer-Reasor E, et al. Abstract PD1-03: A phase Ib trial of fulvestrant + CDK4/6 inhibitor (CDK4/6i) palbociclib + pan-FGFR tyrosine kinase inhibitor (TKI) erdafitinib inFGFR-amplified/ ER+/ HER2-negative metastatic breast cancer (MBC). Poster Spotlight Session Abstracts. American Association for Cancer Research; 2021. doi:10.1158/1538-7445.sabcs20-pd1-03

52. Liu Y, Siejka-Zielińska P, Velikova G, Bi Y, Yuan F, Tomkova M, et al. Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. Nat Biotechnol. 2019;37: 424–429.

53. Walter V, Fischer C, Deutsch TM, Ersing C, Nees J, Schutz F, et al. Estrogen, progesterone, and human epidermal growth factor receptor 2 discordance between primary and metastatic breast cancer. Breast Cancer Res Treat. 2020;183: 137–144.