Monitoring Circulating-Tumor DNA

By

Timothy Butler

A DISSERTATION

Presented to the Cancer Biology Program
Oregon Health & Science University
School of Medicine
In partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

November 2016

Table of Contents

Acknowledgements

I would like to thank the members of the Spellman lab, in particular the clinical coordinators (Tara Macey, Katie Johnson-Camacho, Shaadi Tabatabaei, and Daira Melendez) who jumped through all of the IRB hoops and collected the clinical specimens that made all this research possible. I would also like to thank Dr. Myron Peto who helped get me up and running on the computational side of things. Kami Chiotti thanklessly tracked and archived all of my sequencing data. Dr. Nick Wang helped show me the ropes of making sequencing libraries. Chris Boniface assisted in a myriad of ways from extracting cfDNA and DNA from plasma and tumor tissue, creating whole-exome sequencing libraries, serving as a sounding board to bounce ideas off of, and was frequency willing to allow me to spike in my samples into one of his sequencing runs. Dr. Brett Johnson was also very generous with his time, reagents when we had a shortage, and spare sequencing capacity. I would like to thank Dr. Brian O'Roak and Sara Evans for their assistance in using their NextSeq 500 sequencer, and Dr. Bob Searles and Amy Carlos for their time and assistance at OHSU's sequencing core.

I would like to thanks the member of my dissertation committee – Dr. Matt Thayer, Dr. Chris Corless, Dr. Lisa Coussens, and Dr. Joe Gray – for their advice and guidance during the course of my graduate career. I would also like to thank Dr. Amanda McCullough for agreeing to serve as the outside member of my committee.

Special thanks go to my friends and classmates who helped me through the ups and downs of grad school: Kevin Wong and Tim and Gaby Reilly for being

excellent trivia and gaming partners, Dr. Amanda Esch for always being eager to treat this starving grad student to a happy hour, and Markus Schaaf for being the best friend I could ask for. I would like to especially thank Dr. Chris Schafer, Ashleigh Murphy, and Lauren Uhde who have edited countless drafts of various papers, qualifying exams and grants. They helped me become a better scientist and writer through their efforts. I would like to thank my parents for the support they've given me, particularly letting me live under their roof these past couple months as I finished my thesis.

Finally, I would like to thank my mentor Dr. Paul Spellman for taking a chance on me as his first graduate student. His guidance over the past five years has been instrumental in my development as a scientist. He allowed me to pursue my interests and was supportive when I inevitably hit road blocks.

List of Abbreviations

cfDNA	cell-free DNA
ctDNA	circulating-tumor DNA
AML	Acute Myeloid Leukemia
bp	base pairs
PCR	Polymerase Chain Reaction
ddPCR	droplet-digital PCR
SafeSeqS	Safe Sequencing System
iDES	integrated Digital Error Suppression
FFPE	Formalin-Fixed Paraffin Embedded
WT	Wild-Type
COSMIC	Catalogue of Somatic Mutations in Cancer
dbSNP	database of Single Nucleotide Polymorphisms
VAP	Variant Allele Percentage
CNV	Copy Number Variant
DIDA	Dual-Indexed Degenerate Adapters
SSCS	Single-Stranded Consensus Sequence
ER	Estrogen Receptor
pCR	Pathological Complete Response
AC	doxorubicin and cyclophosphamide chemotherapy
IRB	Institutional Review Board
MRI	Magnetic Resonance Imaging
Mut	Mutant Base
Ref	Reference Base
Chr	Chromosome Number

List of Figures

Chapter 3

List of Tables

Abstract

The circulating-tumor DNA (ctDNA) field has rapidly advanced over the past six years; transitioning from biological curiosity to clinical reality. The field has predominantly focused on the clinical and pharmaceutical implications of identifying targetable oncogenic mutations in the plasma. In this work, we sought to focus more broadly on determining how much information about a patient's disease could be gained from analyzing ctDNA. This has taken the form of two major studies looking at different stages of disease: metastatic and localized.

In our metastatic study we demonstrated that whole-exome sequencing of ctDNA was possible due to the high ctDNA abundance. This could provide a liquid biopsy of a patient's disease; identifying mutations of clinical and biological interest, and identifying changes in the tumor genome from the primary disease. We also identified mutations not present in any matched tumor sample, potentially identifying mutations in metastatic sites we did not have tissue from. This indicates that it is possible that ctDNA can provide more information in metastatic disease than a biopsy could. In addition to mutational information, copy number variation could be identified in a patient with sufficiently high ctDNA level, providing a more complete picture of the tumor genome. This study was among the first to demonstrate the feasibility of this approach.

To study the lower ctDNA levels present in localized diseased first required the development of a new high accuracy, high sensitivity sequencing technique. We came up with dual-indexed degenerate adaptors (DIDA). This technique used a hybrid capture approach to simultaneously assay dozens of patient-specific

mutations, while utilizing adaptors with degenerate sequences allowing for error correction to a frequency below 1 in 10k reads. In developing this technique we discovered a systemic error which occurred with this, and other related error correction techniques, we dubbed “tag swaps.” This error was responsible for systematic overestimation of sequencing depth and sensitivity. Several filtering methods were attempted, and the most aggressive one was chosen. The final version of the technique and analysis proved accurate and robust, allowing us to analyze localized disease.

We chose to track ctDNA in neoadjuvantly treated breast cancer, measuring ctDNA before, during, and after treatment. We generally saw a dramatic reduction in ctDNA during the course of treatment, however it appeared more pronounced in patients who responded to treatment than those who did not. In fact, in both patients who had tumor growth during treatment, we saw a corresponding increase in ctDNA, demonstrating a potential for early detection of progression. We were also able to detect recurrence 7 months before clinical presentation in one patient, and potentially identify the effectiveness of post-surgery radiation in another. These findings indicate that there is the potential for considerable clinical upside from pursuing ctDNA tracking in neoadjuvant treatment.

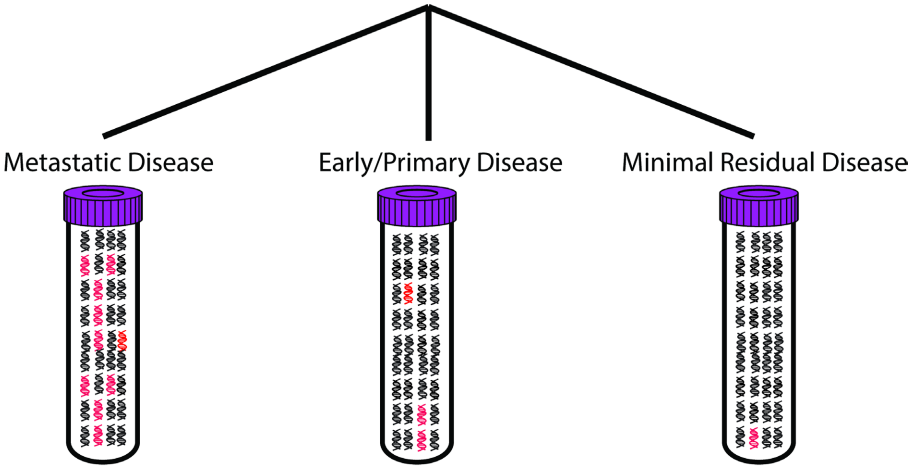
Taken together these two studies have advanced our understanding of ctDNA and identified several potential clinical and scientific avenues for further research.

Introduction

Precision cancer treatment requires knowledge of the specific molecular drivers in a tumor to select interventions. The traditional method, a small biopsy of the primary tumor, only provides a snapshot; it has limited power to identify genomic heterogeneity in the tumor or capture alterations that can occur as selective pressures on the tumor change over time. One approach to identifying the spatial and temporal complexity of tumors is the analysis of tumor DNA present in the plasma, or circulating-tumor DNA (ctDNA). ctDNA is released primarily via the apoptosis of tumor cells, which may occur throughout a tumor, potentially giving a more representative picture of the tumor genome than a single biopsy. Additionally, due to the minimally invasive nature of a blood draw, ctDNA can be serially collected to measure changes in quantity and composition over time. Currently, the ctDNA field is rapidly expanding as measured both by an increase in publications and an increase in private sector activity, over $300 million was raised in January 2016 alone.[1] This flurry of activity can be broadly divided into several somewhat overlapping areas of focus, requiring a variety of analysis methods (Fig 1):

1) ctDNA as a “liquid biopsy” to identify mutations of interest (including resistance mutations) and/or characterize tumor heterogeneity
2) serial ctDNA quantification to assess tumor burden and response to treatment
3) early detection of disease recurrence following curative treatment
4) early detection of primary disease

To understand the utility and challenges of these applications, we must first cover the basic biology of cell-free DNA (cfDNA) and ctDNA.



	Metastatic Disease	Early/Primary Disease	Minimal Residual Disease
Required Sensitivity	>5%	1 in 10^2 - 1 in 10^4	1 in 10^3 - 1 in 10^5
Analysis Methods	Whole-Genome Whole-Exome	ddPCR BEAMing SafeSeqs iDES	ddPCR BEAMing SafeSeqs iDES
Applications	Liquid Biopsy of Mets Tumor Burden Monitoring	Response to Treatment Resistance Mutation Detection	Early Recurrence Detection

Figure 1. Overview of sensitivity, methods, and applications for ctDNA analysis.

Biology of cfDNA and ctDNA

The presence of fragmented DNA in the circulating blood, dubbed cell-free DNA (cfDNA), was first described in 1948,[2] however nearly three decades passed before this DNA was analyzed in patients with cancer.[3,4] These initial observations led to the discovery that circulating cfDNA was elevated in patients with cancer. The first direct confirmation of the existence of ctDNA occurred in 1994 when mutant *NRAS* was detected in the plasma of individuals with AML.[5]

This was shortly followed by observations of *KRAS* and *TP53* mutations in individuals with solid tumors.

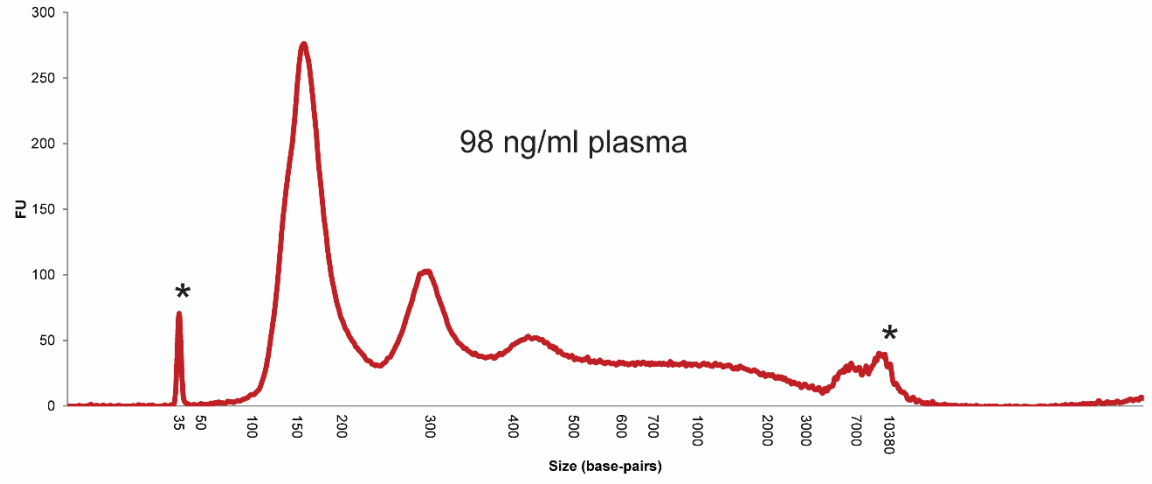


Figure 2. Stereotypical cfDNA size distribution. Agilent 2100 Bioanalyzer high-sensitivity trace of cfDNA from patient with metastatic breast cancer (patient #2 in chapter 1). Asterisks mark lower and upper size markers.

The vast majority cfDNA is highly fragmented in nature, with an average fragment length of approximately 170 bp, roughly corresponding to the length of DNA wrapped around a single nucleosome, with progressively lower quantities of cfDNA at lengths representing two and three nucleosome sizes (Fig 2).[6,7] This nucleosome association in the blood likely protects the cfDNA from activity of blood-borne nucleases.[7-9] In fact, we contributed to the first study which demonstrated that the pattern of read depth can be used to infer nucleosome occupation.[10] Genes with low expression corresponded to tightly packed nucleosomes and high read depth, while regions with highly expressed, housekeeping genes showed a ~170 bp periodicity of high read depth centered on the transcriptional start site. Determining this pattern of nucleosome

occupation can be used to estimate tissue of origin,[7] and, to a limited extent, gene expression.[9,10] There are several potential sources of cfDNA; the regular size indicates that the majority of cfDNA is the result of cell death, primarily apoptosis, though there is evidence for cfDNA release through necrosis,[11] extracellular vesicles,[12] and even active DNA release as well.[13] The tissue of origin study identified that immune cells are a major contributor to cfDNA.[7] Intriguingly, the mechanism of release may be dependent on tissue type or biological process as there is evidence that ctDNA (derived from tumor cells) is about 7 bp shorter than cfDNA (derived from normal cells).[14,15]

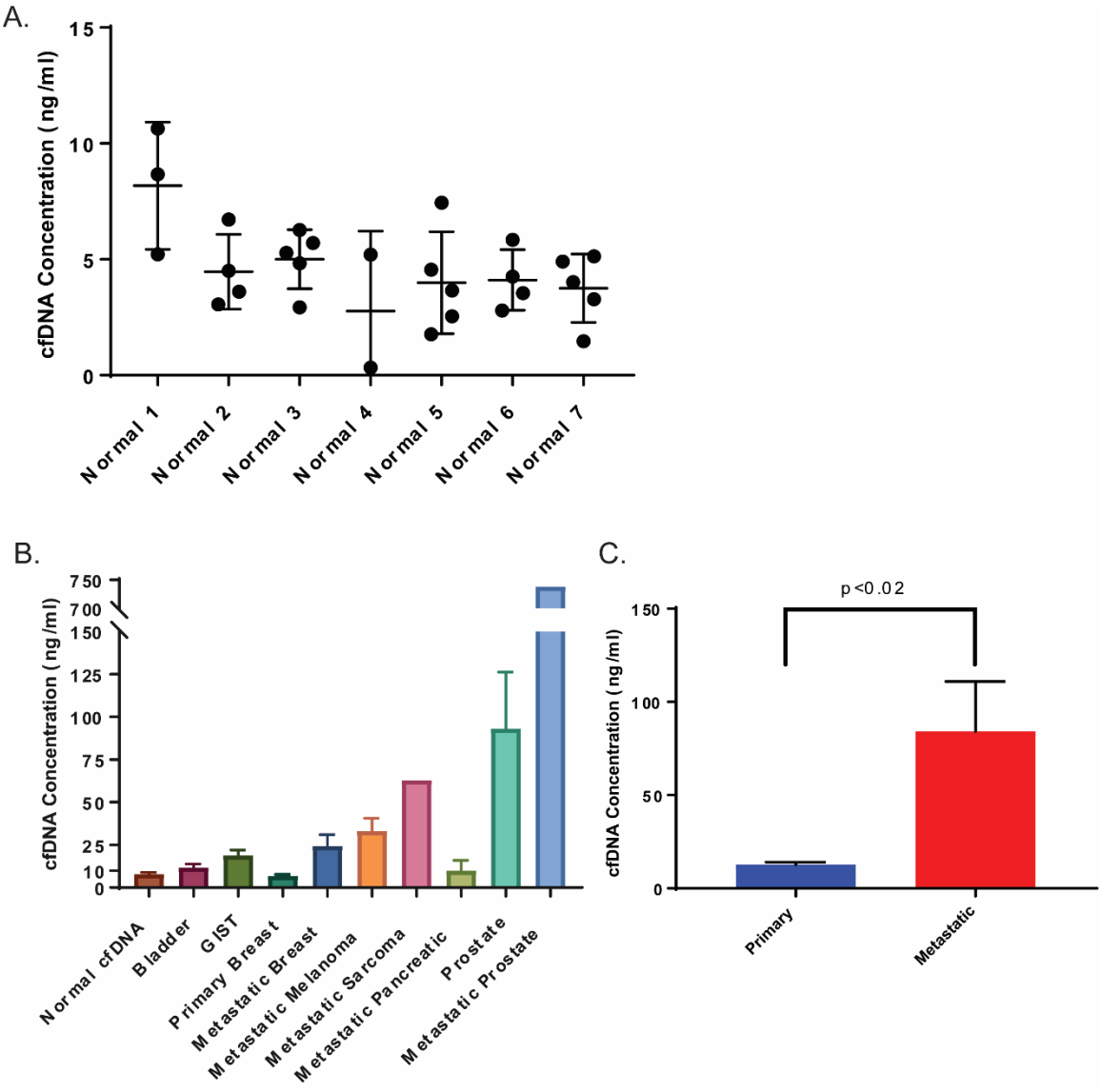


Figure 3. cfDNA concentration overview. A) Variation in cfDNA concentration of 7 normal, healthy plasma donors. Each data point represents a separate plasma donation. Mean +/- Standard deviation. B) Average cfDNA concentration for 61 samples across 10 tumor types, error bars +/- SEM. C) cfDNA is significantly elevated in metastatic disease (N=36) compared to localized primary (N=40) disease, error bars +/- SEM.

In healthy adults the level of cfDNA is typically 5-10 ng/ml of plasma, while in individuals with cancer the cfDNA concentration can be anywhere from the

normal range to over 50 times the normal levels (Fig 3).[16,17] Variation over time in the same individual can be several-fold, with many potential factors influencing the level including: exercise, time of day, and a variety of non-cancer disease states (Fig 3A).[11,18-21] Whether this variation is driven primarily by changes in cfDNA release or cfDNA clearance is poorly understood, and is deserving of further study. Differences in the cfDNA yield can dramatically alter the potential sensitivity of a given assay, and must be taken into consideration during study design. As a general rule, cfDNA from patients with metastatic disease is dramatically higher than in primary disease (Figs 3B and 3C). Tracking the loss of fetal Y-chromosome DNA following birth demonstrated that the half-life of cfDNA is around an hour,[22] with cfDNA being processed in the spleen, liver and a smaller portion excreted in the urine.[11,23,24]

The most challenging aspect of cfDNA analysis is often how little of the cfDNA is of tumor origin. Levels of ctDNA as measured by mutant allele percentage of mutations known to be present in a patient's tumor vary widely, anywhere from above 25% to well under 1 part in 10k.[16,17] The most comprehensive study to date looking at multiple tumor types and stages found 6-logs of variation in the level of ctDNA, with wide variation between and within tumor types.[17] They were able to detect ctDNA in 100 percent of some primary cancers (bladder, colorectal, and ovarian), and as little as 10 percent of glioma cases, with most cancer types detected in over half of cases. Across all tumor types both the level of ctDNA and likelihood of detection were increased in higher stage cancer, with the most dramatic improvement occurring when comparing

metastatic to localized disease. This trend of increasing ctDNA level with increasing disease stage has been widely shown in the literature. The broad variation in ctDNA level requires methods more sophisticated than conventional sequencing approaches which have error rates of around 1 in 1,000.[25] This has led to the development of a variety of high-accuracy, high-sensitivity techniques (Table 1). It is hypothesized that the fragmented nature of cfDNA, combined with the large average distance between mutations, means that detection of each mutation can be considered an independent test. So the sensitivity of an assay will be driven by the number of input cfDNA molecules, the efficiency of sequencing those input molecules, the number of mutations assayed, and the accuracy of detection. Development of new types of assays has significantly broadened the potential clinical utility of ctDNA analysis.

In addition to cfDNA, extracellular vesicles, such as exosomes, are another source of circulating nucleic acids. Exosomes are small (~100nm) vesicles derived from endosomes with potential roles in cell-cell signaling. While not a focus of this thesis, exosomes have shown promise as a source of circulating nucleic acid and protein biomarkers in cancer.[26,27]

Technique	Description	Self-Reported Accuracy	Mutations Per Assay	Refs
Digital PCR (dPCR)	Input DNA split between >1000 of wells/droplets ensuring 0 or 1 copies of template molecule present, PCR conducted in each well/droplet, presence of amplified template measured by flourescence (possibly allele-specific hybirdization), positive/negative wells/droplets counted, mutant allele frequency determined	~1 in 10^4 varies depending on platform and number of wells/droplets	Allele-specific , ~5 alleles in single assay	28
Beads, emulsion, amplification, and magnetics (BEAMing)	Emulsion PCR resulting in amplicons coating magnetic beads, magnetic separation of beads, allele-specific fluorescent hybridization to beads, read out via flow cytometry	1.6 in 10^4 to 4.3 in 10^5	Allele-specific, ~5 alleles in single assay	29,30
Safe Sequencing System (SafeSeqs)	2-Step PCR, initial multiplexed PCR using template-specific primers containing degenerate barcode and universal adapter, second PCR using universal adapter adding sample barcodes and Illumina Sequencing Adapters, error correction by creating consensus sequence from degenerate barcodes	9 in 10^6	Up to 12	31
Targeted plasma re-sequencing (Tam-Seq)	2-Step PCR, initial multiplexed PCR using temple-specific primers containing universal adapters, second PCR using universal adapter adding sample barcodes and Illumina Sequencing Adapters	1 in 1,000	~48	32
Integrated Digital Error Suppression (iDES)	Degenerate barcode ligated to template DNA, hybrid capture, error correction using degenerate barcode to create consensus sequence, subsequent computational error correction removing error-prone sequences	2 in 10^5	Limited by hybrid capture panel, 10's-1000's	33
Duplex Sequencing	Double-Stranded degenerate barcode sequence ligated to template DNA, hybrid capture, error correction to first create single-stranded consensus using degenerate barcode, then double-stranded consensus	1 in 10^8	Limited by hybrid capture panel, 10's-1000's	34

Table 1. Summary of high-accuracy ctDNA detection techniques

Liquid Biopsy

Much of the development in the ctDNA field has been geared towards the ‘”liquid biopsy;” that is, the ability to use ctDNA as an alternative to a tissue biopsy. While biopsies are routine clinical practice, they are not without their limitations, and in some situations tissue biopsies are impossible. There are clear risks associated with many tissue biopsies, as an example, a study from MD

Anderson reported adverse events in 1.6% of abdominal biopsies.[35] In addition, the genetic heterogeneity in cancers, especially in metastatic disease, means a single biopsy may not accurately reflect the genomic diversity of a tumor, potentially missing important mutational drivers.[36,37] These limitations create two overlapping clinical applications for liquid biopsies: identification of specific mutations that would inform treatment, and characterizing genetic intra-tumoral heterogeneity, which could be especially useful in metastatic disease.

Numerous studies have shown varying success in identifying mutations present in a patient's tumor with their ctDNA. In breast cancer, activating *PIK3CA* mutations were identified in ctDNA with 95% accuracy,[38] studies looking at patients with metastatic disease reported 100% accuracy.[39,40] Similar levels of success were seen in detecting *KRAS* and *EGFR* mutations in lung cancer and a variety of mutations in ovarian cancer (97% accuracy).[41] A study screening ctDNA in 157 patients with advanced cancers for mutations in *BRAF, EGFR, KRAS,* and *PIK3CA* found 83%-99% concordance with archived tumor tissue. In addition they found an association between outcome ctDNA level, patients with ctDNA frequencies above 1% had shorter median survival than those below 1%.[42] Curiously, localized pancreatic cancer does not appear to readily shed ctDNA. Despite the fact that approximately 90% of pancreatic ductal adenocarcinomas (PDAC) carry activating *KRAS* muations,[43] several studies attempting to identify *KRAS* mutations in ctDNA were unable to reliably do so in over half of patients with primary disease,[44,45] suggesting PDAC may be particularly ill suited for ctDNA analysis. This is especially unfortunate because

the extremely high prevalence of *KRAS* mutations make ctDNA detection a potential early detection tool.

ctDNA in melanoma is particularly well studied. A meta-analysis assessed levels of mutant *BRAF* in the plasma of melanoma patients enrolled in clinical trials for *BRAF*-targeted therapies in four separate studies.[46-50] Mutant *BRAF* ctDNA was detected in 76% of patients shown to have mutant *BRAF* in their tumor tissue. It was further demonstrated that there is an increase in *BRAF* ctDNA allele percentage in patients with 3 or more metastatic sites, agreeing with other studies demonstrating a positive relationship between tumor burden and ctDNA percentage. Finally, patients without detectable mutant *BRAF* ctDNA were shown to have better progression-free and overall survival compared to those patients with detectable mutant *BRAF* ctDNA. This suggests the possibility of developing a two-step screening/treatment process in which all melanoma patients undergo ctDNA *BRAF* testing. Those with detectable mutant *BRAF* would be treated with a *BRAF*-targeted therapy, and those without detectable mutant *BRAF* would receive a subsequent biopsy to confirm the *BRAF*-negative result.

Liquid biopsies also can provide sensitive detection of mutations that confer resistance to targeted therapies such as enzalutamide (targeting the androgen receptor in castration-resistance prostate cancer) and aromatase inhibitors (depriving the estrogen receptor of its ligand in estrogen-receptor positive breast cancer).[51-55] Azad and colleagues analyzed plasma from 62

metastatic castration-resistant prostate cancer patients, looking specifically for alterations to the androgen receptor (*AR*) which could explain resistance to *AR*-targeted therapies.[54] They identified *AR* point mutations and/or *AR* amplification in 65% of their cohort, demonstrating plasma-based resistance detection. Using a cohort of 171 advanced breast cancer patients, Schiavon and colleagues analyzed plasma samples taken at time of progression looking specifically at resistance mutations in the estrogen receptor (*ESR1*).[51] Patients with detectable *ESR1* mutations in plasma had a significantly reduced progression-free survival compared to those without *ESR1* mutations. They also showed *ESR1* mutations appear more commonly during treatment of metastatic disease with androgen inhibitors rather than in the adjuvant setting, possibly due to increased genetic diversity in metastatic disease. This study highlights that detecting these mutations in the ctDNA can both inform patient treatment and provide us with a better understanding of how these mutations are selected for in the first place.

The higher ctDNA levels present in metastatic disease makes analyzing a larger portion of the genome feasible. Whole-genome sequencing was successfully used to identify tumor-associated copy number changes and structural rearrangements.[56,57] The same group designed an assay to sensitively detect structural rearrangements in ctDNA as an alternative to detecting substitutions.[58] Several groups also explored using whole-exome sequencing to completely characterize the mutations present in advanced metastatic disease. In a proof of principle study, Murtaza and colleagues sequenced cfDNA from six patients with advanced cancers (breast, ovarian, and NSCLC) finding ctDNA

made up between 33-58% of the total cfDNA.[59] This high ctDNA burden allowed them to identify 93 of 119 mutations (78%) present in a matched metastatic biopsy, importantly this included an activating *PIK3CA* mutation. Interestingly, 26 additional mutations were identified uniquely in the ctDNA; they hypothesized that these were mutations present in unsampled metastatic lesions. Their final observation was that there was a positive correlation between the allele frequency in the metastatic biopsy and ctDNA, especially for those mutations also identified from a previous biopsy of the primary tumor which were also likely to be in unsampled metastatic sites. They further investigated this in a follow-up study by sampling multiple metastatic lesions from a patient with breast cancer, including five samples collected at autopsy.[60] They found four groups of mutations: 23 stem mutations (present in all metastases and the primary), 26 metastatic mutations (present in the metastases and not the primary), 108 private mutations (present in at least one, but not all metastases), and 11 plasma mutations (found in ctDNA but no tumor sample). Their ability to detect mutations in the plasma was better in the stem mutations (91%) than the private mutations (30%). In addition, they found higher ctDNA allele frequencies in the stem mutations than the private ones (20% vs 5%). These two studies demonstrated that in advanced disease whole-exome sequencing of ctDNA is a viable alternative to biopsy, particularly in cases where sampling multiple metastatic lesions is not feasible.

Tumor Burden and Response to Treatment

Determining response to treatment traditionally relies on imaging to see if the size of the tumor changes. Quantifying ctDNA in serially collected plasma samples has several potential advantages over this. First, ctDNA assays can be cheaper than imaging and therefore can be done more frequently, providing deeper insight into the treatment response. Second, multiple mutations can be tracked simultaneously, allowing for insights into how different clones are responding to the treatment and how tumors are adapting to treatment in real-time. This could inform future treatments by identifying *de novo* resistance mechanisms. For example, sequencing ctDNA of patients who developed resistance to aromatase inhibition may have allowed the link between resistance and *ESR1* mutation to be seen earlier.

The correlation between ctDNA level and tumor burden was demonstrated by Dawson and colleagues who quantified ctDNA in 30 patients with metastatic breast cancer.[40] Similar correlations have been seen in other cancer types including: metastatic melanoma undergoing immune checkpoint blockade,[61] metastatic colorectal cancer,[32] and primary gynecological malignancies.[62] A studying looking at 39 patients representing 10 tumor types undergoing targeted therapy found a highly significant relationship between a decrease in ctDNA and an increase in time to progression.[63] The general consensus of these studies is that ctDNA analysis is insufficiently robust to completely replace imaging, but can be used in combination to gain additional insight into tumor response.

Several studies have tracked multiple mutations to assess whether specific mutations, likely derived from distinct tumor clones, are responding differently to a given therapy. Through serial whole-exome sequencing of ctDNA taken before and after various treatments, Murtaza and colleagues identified increases in several mutations including: *PIK3CA* in Paclitaxel treated breast cancer, and *EGFR* in Gefitinib treated lung cancer.[59] This increase potentially represents expansion of resistant clones and could be used as a method to identify new mutational resistance mechanisms. They took this a step further in an additional study by sequencing tissue taken from primary, metastatic, and autopsy tissue from a single breast cancer patient. They identified shared and lesion-specific variants and compared those variants to serially collected plasma samples.[60] They identified an *ERBB4* mutation as the most prevalent lesion-specific mutation in the plasma when the patient developed resistance to Lapatinib. The lesion carrying this mutation was the main site of disease progression. These studies measured the tumor evolving in response to new selective pressures, opening the door for numerous follow-on studies determining whether these responses are generalizable.

Early Detection of Recurrence

There is usually no measurable tumor burden in a patient receiving potentially curative treatment. However, many of these individuals will recur, often at distant metastatic sites seeded from the primary tumor. These metastases will carry many of the mutations that were present in the primary

tumor, although they are usually not genetically identical to the primary tumor. This presents the possibility of detecting these mutations in ctDNA as an early indicator of recurrence. This was first investigated in 2008 by using a BEAMing assay to detect individual mutations following curative surgery in colorectal cancer.[30] All 16 patients with detectable post-surgery ctDNA eventually recurred, while the 4 patients with undetectable ctDNA remained disease-free. Two more recent studies demonstrated that detectable post-surgery ctDNA serves as an early predictor of recurrence in breast cancer.[64,65] Garcia-Murillas and colleagues identified ctDNA in 12 of 12 patients who eventually recurred, with ctDNA detection preceding clinical recurrence by an average of 7.9 months. One ctDNA positive patient did not clinically recur in the 12 months of follow-up, but is likely at risk to do so. Olsson and colleagues, using a much longer monitoring period, identified ctDNA in 13 of 14 patients who recurred, with an average lead time of 11 months.

One open question in this area is whether ctDNA detection can do more than serve as a prognostic marker, but actually inform treatment. Our own work has seen the ctDNA of one breast cancer patient with detectable post-surgery ctDNA become undetectable following radiation treatment, potentially demonstrating a curative effect of that treatment. However, no study to date has analyzed whether giving patients with detectable post-curative ctDNA additional therapies reduces or delays recurrence. Patients with undetectable ctDNA may be at extremely low recurrence risk and could potentially be spared additional treatments in the adjuvant setting. Existing methods for recurrence monitoring

are limited by the abundance of ctDNA available for study, arguably, deeply multiplexed and patient-specific-assays could overcome this limitation. Post-surgery ctDNA status could also be used as an enrollment criterion for clinical trials as a way to focus on a group of patients likely to recur, thereby reducing the number of patients needed to conduct the study saving both time and money.

Early Detection of Primary Disease

The most challenging use case for ctDNA analysis is in the detection of primary disease; there is very little ctDNA present, and, unlike in recurrence detection, the mutations present in the tumor are unknown. To detect ctDNA in this early state an assay is needed that is extremely sensitive, surveys scores of potential mutations, and remains cost-effective. Newman and colleagues have attempted to address this problem by using hybrid-capture panels combined with several error-correction techniques (iDES) to accurately sequence recurrently mutated sites in various cancers, developing both cancer-specific and pan-caner panels.[33,66] They demonstrated the ability to design a Non-small-cell lung carcinoma panel on the order of 100-300 kb which can detect ctDNA in over 90% of patients, including stage 1 disease, detecting ctDNA levels as low as 1 part in 10k. While they demonstrate an effective use case for tumor genotyping and monitoring, this likely is still too high a threshold to serve as a reliable early detection method, requiring either additional depth (through more input genomes) or breadth (by sequencing a larger portion of the genome), both of which significantly increase cost. However, a methodology similar to iDES could be

utilized by a company such as the recently spawned Grail, which has set its sights on the early detection market, seeking to “Enable the early detection of cancer in asymptomatic individuals through a blood screen.”[67] Launched by Illumina, Grail is uniquely well-positioned to dramatically increase sensitivity in a cost-effective manner. While ctDNA-based early cancer detection may not yet be ready to enter the clinic, the private sector investment, and potential financial upside, will ensure continued pursuit of this goal.

Conclusions

ctDNA analysis has demonstrated varying degrees of clinical utility in liquid biopsy, resistance mutation detection, tumor burden monitoring, response to therapy, and early detection of recurrence. Additionally, the ability to serially monitor changes in the tumor genome over time and in response to specific perturbations and selective pressures, allows new scientific insights into the evolving tumor genome. It is also worth a more thorough investigation into the mechanisms and timing of ctDNA release and clearance. For example, if administration of a chemotherapy increases ctDNA release over a defined period, could that be used to improve the likelihood of detection in a patient undergoing treatment? Would a nearly homeopathic dose administered to a patient in remission increase ctDNA release thereby allowing a recurrence to be detected sooner? Results from studies answering these types of questions, combined with continued improvement and standardization of methods, have the potential to decrease the degree of difficulty for many ctDNA assays. The next step for the ctDNA field is to demonstrate that patient outcomes improve by acting on the information gained from ctDNA identification and quantification; be it administration of a new therapy upon seeing a resistance mutation, or giving an additional treatment in the adjuvant setting following detection of ctDNA. New clinical trials testing ideas such as these will be much anticipated in the ctDNA community.

Chapter 1:

Exome Sequencing of cell-free DNA from Metastatic Cancer Patients Identifies Clinically Actionable Mutations Distinct from Primary Disease

Timothy M. Butler[1], Katherine Johnson-Camacho[1], Myron Peto[1], Nicholas J. Wang[1], Tara A. Macey[1], James E. Korkola[1], Theresa M. Koppie[1], Christopher L. Corless[1], Joe W. Gray[1], Paul T. Spellman[1]*

[1] Knight Cancer Institute, Oregon Health and Sciences University, Portland, Oregon, USA

* Corresponding Author

E-mail: spellmap@ohsu.edu

Adapted from Butler et. al.[53]

Abstract

The identification of the molecular drivers of cancer by sequencing is the backbone of precision medicine and the basis of personalized therapy; however, biopsies of primary tumors provide only a snapshot of the disease's evolution, and may miss potential therapeutic targets, especially in the metastatic setting. A liquid biopsy, in the form of cfDNA sequencing, has the potential to capture the inter- and intra-tumoral heterogeneity present in metastatic disease, and, through serial blood draws, track the evolution of the tumor genome.

In order to determine the clinical utility of cfDNA sequencing we performed whole-exome sequencing on cfDNA and tumor DNA from two patients with metastatic disease; only minor modifications to our sequencing and analysis pipelines were required for sequencing and mutation calling of cfDNA. The first patient had metastatic sarcoma and 45 of 46 mutations present in the primary tumor were also found in the cfDNA. The second patient had metastatic breast cancer and sequencing identified an *ESR1* mutation in the cfDNA and metastatic site, but not in the primary tumor. This likely explains tumor progression on Anastrozole. Significant heterogeneity between the primary and metastatic tumors, with cfDNA reflecting the metastases, suggested separation from the primary lesion early in tumor evolution. This is best illustrated by an activating *PIK3CA* mutation (H1047R) which was clonal in the primary tumor, but completely absent from either the metastasis or cfDNA. Here we show that cfDNA sequencing supplies clinically actionable information with minimal risks compared to metastatic biopsies. This study demonstrates the utility of whole-

exome sequencing of cell-free DNA from patients with metastatic disease. cfDNA sequencing identified an *ESR1* mutation, potentially explaining a patient's resistance to aromatase inhibition, and gave insight into how metastatic lesions differ from the primary tumor.

Introduction

In 2014 there were over 500,000 cancer related deaths in the United States; 90% of these deaths from metastatic disease.[68,69] While cancer is characterized by clonal progression, metastatic lesions and recurrent disease can differ substantially from the primary tumor, harboring unique mutations of clinical significance.[36] Identifying these differences as they emerge requires serial sampling of the tumor genome,[70] often from multiple metastatic sites, which may have limited feasibility due to technical challenges or financial burden. Sequencing from blood plasma, however, has the potential to identify these changes without the invasiveness associated with solid tumor biopsies.[17,71,72]

Following the detection of mutant forms of *KRAS* and *NRAS* in the plasma of cancer patients, researchers have pursued cfDNA as a form of "liquid biopsy" of an individual's cancer, using it to identify oncogenic alterations in a variety of malignancies.[5,12,16,38,41,73,74] Changes in ctDNA over the course of treatment can be measured easily through serial sampling due to the minimally invasive nature of blood draws.[30,40,57,59,75] Previous studies have focused on quantifying ctDNA levels to measure disease burden,[40,58,75] searched for the emergence of resistance mutations to specific therapies,[59,76-78] tracked tumor evolution,[59] and assessed prognosis[16,79,80] and recurrence risk.[30] The detection of ctDNA requires especially sensitive methods due to its dilution by the DNA from non-cancerous cells, with variant allele percentages as low as 0.01% in early disease.[16,81,82] The study of tumors of varying types and stages has found that while ctDNA levels

vary significantly between samples, metastatic disease correlates with higher levels of cfDNA in the plasma and a higher fraction of ctDNA.[17,83] The relative abundance of cfDNA and ctDNA makes it well-suited for whole-exome sequencing[59] which, unlike panels focusing on hotspot or patient-specific mutations, has the potential to identify novel mutations, giving it unique value in the study of therapeutic resistance and tumor evolution. Whole-exome sequencing from plasma has demonstrated high levels of concordance between mutations in the tumor tissue and cfDNA in metastatic disease; however, previously this has only been shown in samples with exceptionally high ctDNA levels (33-65% of cfDNA from tumor origin), greatly limiting its clinical utility.[59]

In this study, we investigated the feasibility of whole-exome sequencing from the plasma of two patients with metastatic disease. We found that with only minor alterations to our experimental and analytical methods we could accurately recapitulate the tumor genome from plasma, identify the same clinically relevant mutations identified by sequencing tumor biopsies, and gain novel information about the evolution of the disease. These methods were sensitive in a sample with an average ctDNA variant percentage of 3.5%, indicating approximately 7.0% of cfDNA was of tumor origin (ctDNA), sufficiently low to identify ctDNA for a substantial portion of metastatic patients.[16,17,30] We conclude that cfDNA sequencing of patients with metastatic cancer lends valuable insight to the study and treatment of the disease.

Results

Patient #1. A 52-year-old female was diagnosed with primary intimal sarcoma of the pulmonary artery that was unresectable at presentation. The patient was initially treated with radiation followed by chemotherapy (Fig 4) and at this time her tumor was screened for oncogenic mutations using a multiplexed mass spectroscopy-based assay that revealed the presence of *PIK3CA* R88Q and Q546R in the primary tumor.[84] As a result, she entered a phase I clinical trial of a PI3 kinase inhibitor and had a partial response that lasted 12 months. Twenty months after diagnosis the primary tumor DNA was screened again using a targeted panel of an Ion Torrent PGM. This confirmed the *PIK3CA* mutations but also revealed *KRAS* G12R. A blood draw was taken at this time, isolating 1 ml of buffy coat and 25 mls of plasma (Table 2). At the time of blood collection the patient had numerous lesions in the lungs, pulmonary artery, and liver (Table 2). Due to the high concentration of cfDNA in the plasma (63 ng/ml), whole-exome sequencing was conducted. Based on the *KRAS* mutation, the patient was then enrolled in a phase Ib clinical trial combining MEK and PI3 kinase inhibitors. The treatment was stopped after eight months due to complications resulting from treatment, and the patient died 30 months after the initial diagnosis.

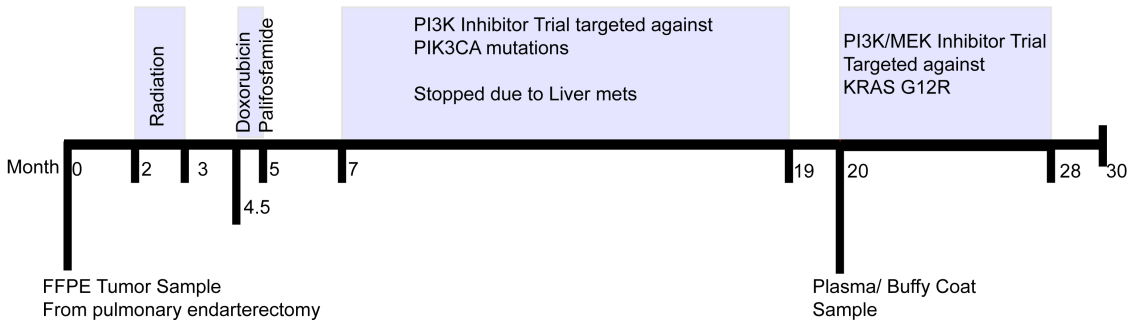


Figure 4. Overview of metastatic sarcoma patient treatment history. Patient diagnosed with intimal spindle cell sarcoma of the pulmonary artery. Treatments and sample collection indicated in months.

Sample	Primary Cancer Type	Volume Plasma Collected (ml)	cfDNA Concentration (ng/ml plasma)	Total cfDNA Extracted (ng)	Tumor Burden
Patient #1	Sarcoma	25	63	1,575	>6 chest lesions 0.5-2.6cm, 2 liver lesions 1.3 cm
Patient #2	Breast Cancer	15	98	980	>5 liver lesions 0.6-4cm, thoracic lesion in T11

Table 2. Plasma collection summary. Volume of plasma collected from single blood draw. cfDNA quantified using Quan-iT HS pico green kit. Tumor burden is at the time of the plasma collection.

Whole-exome sequencing of the primary formalin-fixed paraffin-embedded (FFPE) tumor revealed 46 somatic, exonic mutations (Fig 5A, Tables 3 and 4). We conducted whole-exome sequencing of the cfDNA (524X average depth) and, with a threshold of 1.5% variant allele percentage, identified 45 of the 46 somatic mutations present in the primary. At those 46 sites the mean sequencing depth in the cfDNA was 565X (181-1,197X). The average variant allele

percentage across these 45 mutations was 3.5%, indicating that approximately 7.40% of the plasma DNA was of tumor origin. Importantly, we identified from plasma the activating *KRAS* G12R mutation and both activating mutations in *PIK3CA* (R88Q and Q546R). Controlling for sequencing depth, number of cfDNA mutant reads, or variant allele percentage in the primary tissue did not significantly improve the correlation.

Chr	Base	Ref Base	Mut Base	Gene	AA Change	COSMIC	Tumor Depth	Tumor Variant Allele Percentage	cfDNA Read Depth	cfDNA Variant Allele Percentage
1	18023841	C	T	ARHGEF10L	T1269I	None	166	31.3	593	2.5
1	158046011	C	A	KIRREL	T54N	None	55	34.5	662	3.2
2	61436077	C	G	USP34	R2959T	None	34	17.6	622	3.7
2	101869627	G	C	C2orf29	E67D	None	70	28.6	181	2.8
2	149447882	T	G	EPC2	Y85D	None	136	19.1	1081	3.3
2	175432673	G	T	WIPF1	P420T	None	32	21.9	251	3.6
2	179599471	G	A	TTN	V4743V	None	98	16.3	887	4.3
2	179599653	G	A	TTN	R4683C	None	40	32.5	245	3.7
2	220344840	G	A	SPEG	V1774M	None	54	31.5	398	3.3
2	220424127	C	A	OBSL1	E1016*	None	147	33.3	802	3.5
3	19389236	C	T	KCNH8	P197L	None	54	33.3	386	2.1
3	62189116	T	C	PTPRG	A549A	None	68	39.7	344	3.8
3	172835203	A	C	SPATA16	L107V	None	137	38.7	943	6.9
3	178916876	G	A	PIK3CA	R88Q	COSM746	143	40.6	1113	4.6
3	178936095	A	G	PIK3CA	Q546R	COSM12459	42	42.9	391	5.1
5	191699	G	C	LRRC14B	V16L	None	155	17.4	497	3.8
5	1244353	T	A	SLC6A18	L454H	None	139	20.9	344	3.2
5	1294421	G	A	TERT	R194*	None	192	17.7	368	2.4
5	14601222	C	G	FAM105A	H71Q	None	115	19.1	703	3.6
6	27420983	C	T	ZNF184	E119K	None	64	37.5	368	3.0
6	51917924	G	A	PKHD1	T697M	None	50	32.0	480	3.3
6	87971162	G	A	ZNF292	K2605K	None	65	29.2	503	3.0
6	123714772	G	T	TRDN	Q368K	None	44	11.4	427	6.3
6	168709623	C	A	DACT2	V272L	None	58	29.3	235	5.1
8	37698931	G	A	GPR124	T1025T	None	76	28.9	279	1.8
8	75227367	G	A	JPH1	R290C	None	398	26.6	1197	4.0
9	32544159	T	C	TOPORS	K122E	None	83	44.6	777	4.5
9	130550557	C	T	CDK9	A166V	None	227	35.7	583	2.7
10	8006881	A	T	TAF3	I470F	None	133	33.8	780	3.7
10	50083161	T	C	WDFY4	S2326P	None	62	25.8	316	3.2
11	30915897	C	G	DCDC5	M288I	None	91	22.0	485	4.1
11	65412473	G	A	SIPA1	Q344Q	None	46	39.1	294	4.1
12	25398285	C	G	KRAS	G12R	COSM518	162	18.5	355	2.3
12	56845179	C	T	MIP	R226Q	None	83	15.7	765	5.2
13	39262777	G	A	FREM2	K432K	None	104	38.5	620	3.5
14	23791401	T	A	PABPN1	A121A	None	31	25.8	750	0.0
14	45716371	G	T	MIS18BP1	T40N	None	82	17.1	501	1.8
14	102792762	G	A	ZNF839	S127S	None	47	27.7	271	3.3
15	85327565	G	C	ZNF592	E553D	None	173	41.0	795	2.6
17	38955860	G	A	KRT28	R96C	None	283	24.7	1085	2.8
17	66042028	C	T	KPNA2	F496F	None	45	33.3	811	3.7
17	67129877	G	A	ABCA6	S232S	None	64	25.0	421	3.3
18	64172273	T	A	CDH19	I699F	None	136	28.7	748	3.1
19	960142	G	C	ARID3A	L248F	None	116	10.3	482	1.5
20	20243713	C	T	C20orf26	N814N	None	82	26.8	374	2.9
22	32352755	T	C	YWHAH	D239D	None	205	32.2	464	4.7
						Average	**106**	**28.2**	**565**	**3.5**

Table 3. Summary of Patient #1 somatic mutations. List of 46 somatic mutations present in primary tumor along with depth and variant allele percentage in tumor and cfDNA samples. COSMIC accession numbers listed for three mutations present in database.

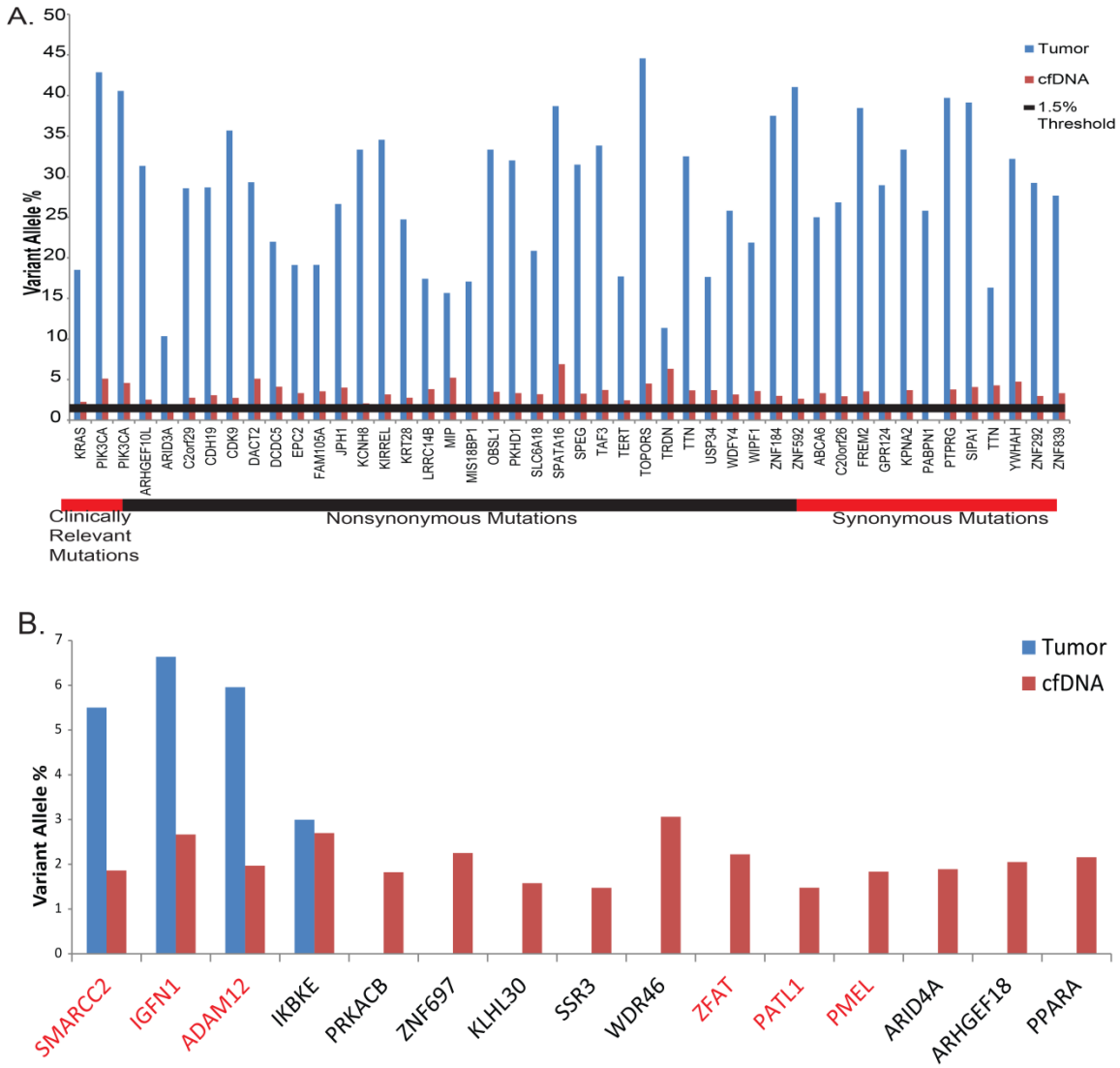


Figure 5. Patient #1 mutation calls and validation. A) Using a cutoff of 1.5% variant allele percentage, 46 of the 47 mutations present in the tumor were identified in the cfDNA. Estimating from the average variant allele percentage of 3.8%, 7.5% of the cfDNA was derived from the tumor. B) Fifteen additional mutations were called in the cfDNA which were not called in the tumor sample. Four of these are present in the tumor, but below our calling cutoff of 10% for the tumor. Genes highlighted in red text were successfully validated via sequencing on the Ion Torrent PGM. Approximately 4,000 genomes of cfDNA were used as input to the validations, giving us a lower sensitivity bound of 0.025-0.5% depending on the site-specific background error rate.

Fifteen additional mutations were identified in the cfDNA. Among these, 11 were not present in the primary number and four were present in the primary tumor (Fig 5B), but at allele frequencies below our 10% threshold for calling them in the primary tumor. These mutations were chosen for validation by sequencing on the Ion Torrent PGM where six of them were confirmed, eight failed to validate, and one did not sequence (Fig 5B). The validation rate of 43% highlights the necessity of using orthologous sequencing methods in confirming the presence of low frequency mutations in cfDNA. cfDNA variant allele percentage correlated poorly with the primary tumor (Fig 6).

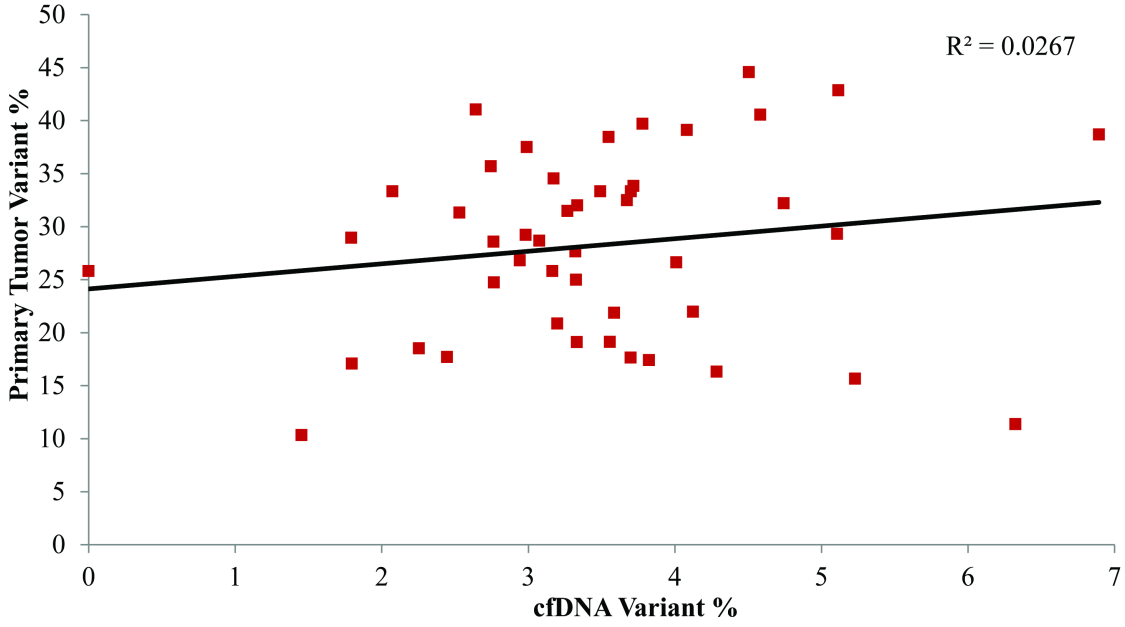


Figure 6. Variant allele percentage correlation cfDNA variant allele percentage is poorly correlated with Primary tumor variant allele percentage.

As a final analysis we determined that ctDNA fragments (identified by the presence of a tumor-specific mutation) were on average 8 bp shorter than cfDNA fragments which mapped to the same region but did not carry the mutation (Fig 7A). Due to the fact that these mutations are likely heterozygous, roughly half of the ctDNA would not carry the mutation and therefore be misidentified as wild-type (WT). Despite this confounding factor, the difference was highly significant. To ensure this effect was not somehow caused by the presence of the mutant base, we compared the fragment lengths of WT and alternate containing reads at 2,100 dbSNP sites and did not find a significant difference (Fig 7B).

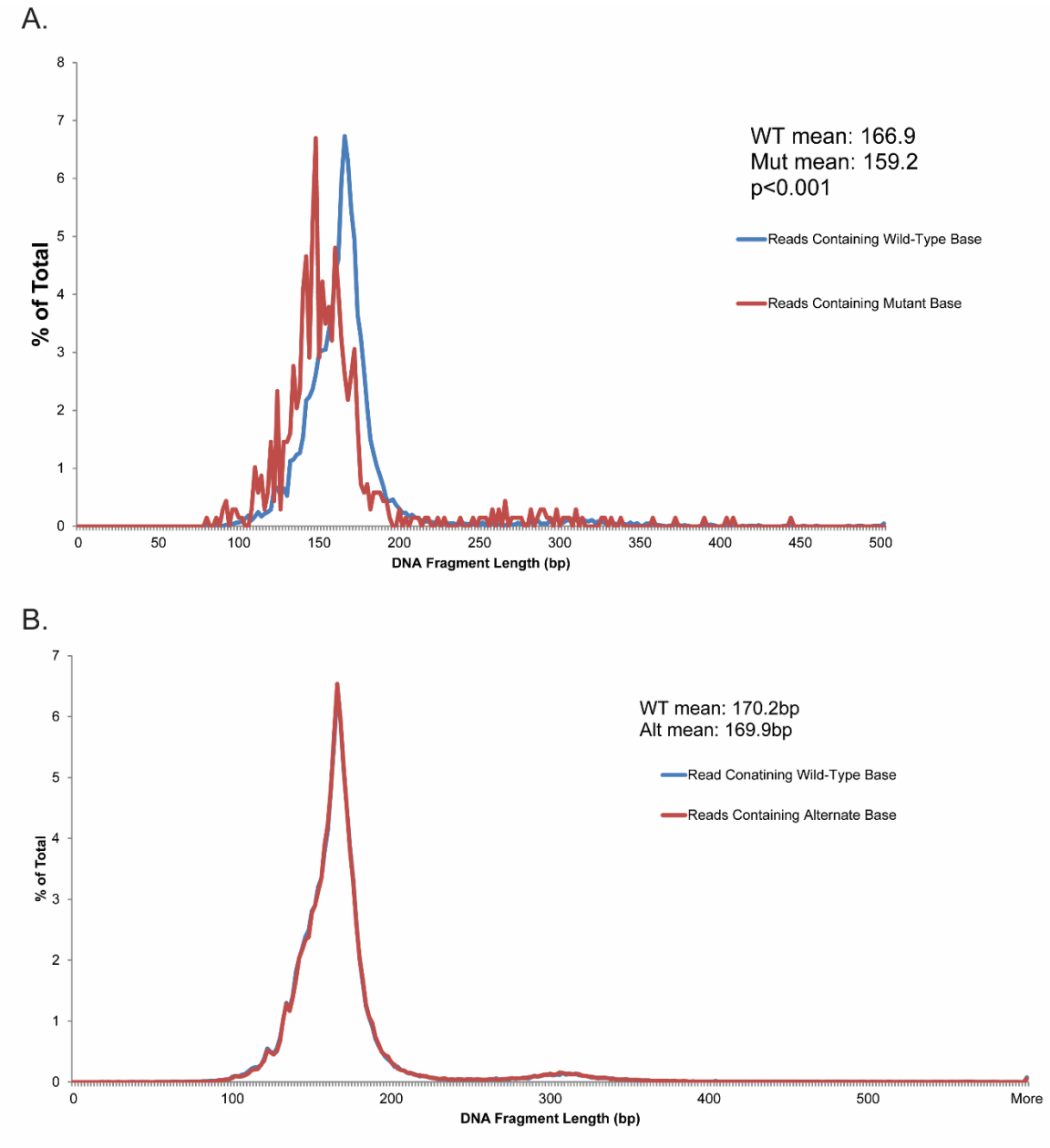


Figure 4. Comparison of DNA fragment length A) Fragment length of reads containing the WT (blue) or mutant (red) base at the 47 sites with known mutations in the primary tumor. B) Fragment length of reads containing the WT (blue) or alternate (red) base at 2,100 heterozygous dbSNP sites.

Cancer Type	Tissue	ENA Accession Number	Input DNA (ng)	Reads (Millions)	Mapped Reads (%)	Paired Reads (%)	On Target Reads (%)	PCR Duplicates (%)	Mean Sequencing Depth
Sarcoma	Buffy Coat	ERS700862	2000	392	361 (92.3)	342 (87.4)	210 (58)	29	226
Sarcoma	Primary Tumor	ERS700863	500	116	114 (98.3)	113 (96.9)	93 (81.6)	31	118
Sarcoma	cfDNA 1	ERS700864	750	250	243 (97.1)	204 (81.6)	143 (58.9)	34	160
Sarcoma	cfDNA 2	ERS700864	110	155	154.8 (99.7)	147 (94.4)	124 (80.4)	12	162
Sarcoma	cfDNA 3	ERS700864	110	186	185 (99.7)	175 (94.5)	155 (84.0)	16	203
Sarcoma	Pooled cfDNA 1-3	ERS700864	970	591	583 (98.62)	526 (89.0)	423 (72.6)	n/a	524
Breast Cancer	Buffy Coat	ERS700858	412	182	181 (99.6)	180 (98.8)	154 (84.8)	20	201
Breast Cancer	Primary Tumor	ERS700859	301	112	110.8 (99.2)	99.1 (88.8)	92 (82.8)	52	118
Breast Cancer	Metastasis	ERS700860	341	173	171.8 (99.5)	170 (98.6)	140 (81.6)	22	183
Breast Cancer	cfDNA	ERS700861	155	286	284.8 (99.5)	253 (88.5)	239 (83.8)	37	309
Sarcoma	Buffy Coat	ERS700862	2000	392	361 (92.3)	342 (87.4)	210 (58)	29	226
Sarcoma	Primary Tumor	ERS700863	500	116	114 (98.3)	113 (96.9)	93 (81.6)	31	118

Table 4. Sequencing statistics. Summary of sequencing information for all ten sequencing runs. All reads are listed in millions. Accession numbers for .bam files uploaded to European Nucleotide Archive provided, for sarcoma patient all 3 cfDNA runs were combined in a single .bam file separated by read group.

Patient #2. A 41-year-old female was diagnosed with ER+ HER2+ breast cancer, which had spread to the lymph nodes. The patient underwent neoadjuvant chemotherapy (TAC) followed by a bilateral mastectomy and oophorectomy (Fig 8A). Following surgery, the patient underwent radiation therapy and was treated with Trastuzumab for one year and Anastrozole for 33 months, until the discovery of a 4cm liver lesion and bone metastases at the 11th thoracic vertebra (T11). Additional chemotherapy and Herceptin were administered but the treatment was stopped following identification of liver metastases. At this time we collected a blood draw approximately 30 minutes

before a liver biopsy was taken and obtained an archived FFPE sample of the primary tumor. The blood draw yielded 15 mls of plasma at an average cfDNA concentration of 98 ng/ml (Table 2). Following the first plasma sample the patient underwent treatment with the anti-Her2 drug TDM1 but following an initial partial response died 62 months after initial diagnosis.

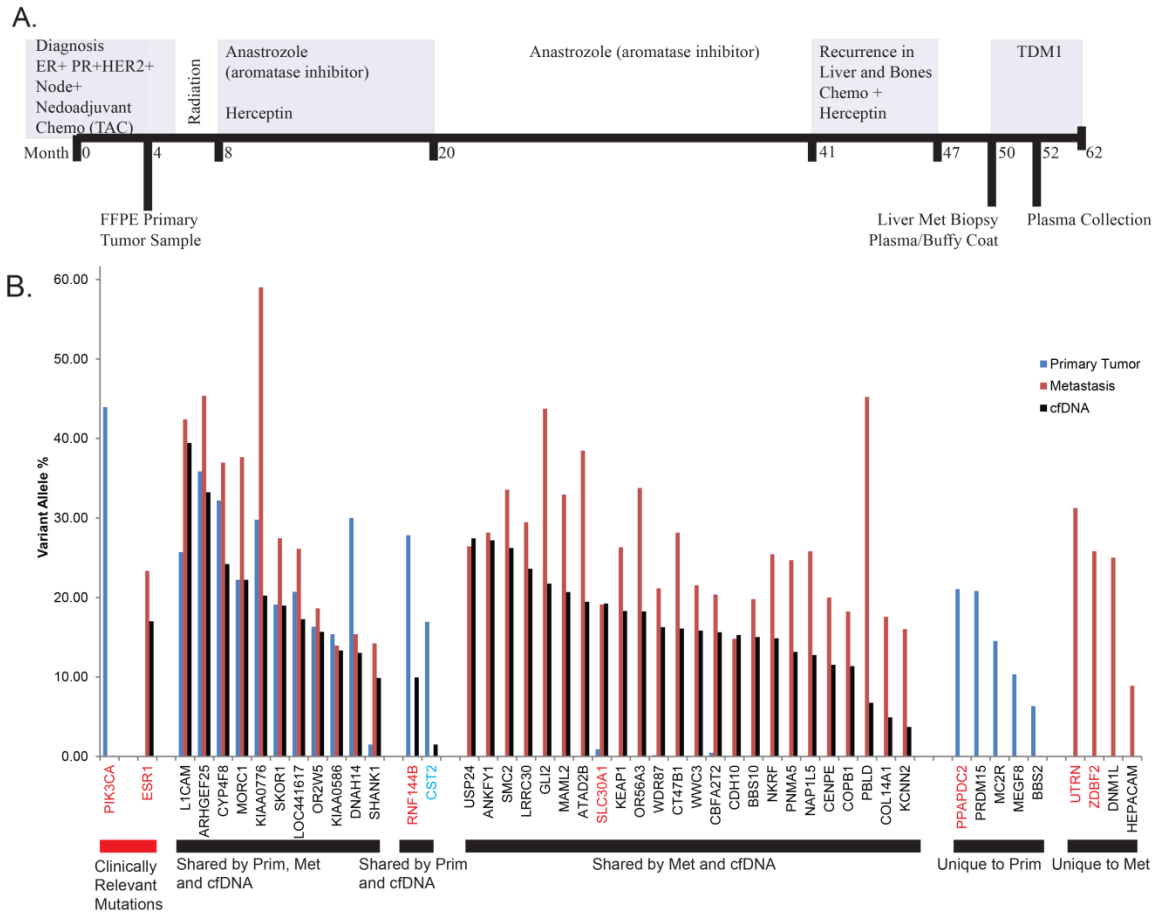


Figure 8. Patient #2 diagnosed with ER+/PR+/HER2+/Node+ breast carcinoma. A) Treatments and sample collection indicated in months. B) 48 total somatic mutations were called in the primary breast tumor and/or liver metastasis. 38 mutations were called in the cfDNA using a variant allele percentage cutoff of 1.5%. Genes in red text were successfully validated on the Ion Torrent PGM, genes in blue text failed to validate, genes with black text were not validated.

Whole-exome sequencing of the primary tumor and liver metastasis revealed a total of 48 nonsynonymous somatic mutations (Fig 8B, Table 5). Sequencing of cfDNA to an average depth of 309X identified 38 of these mutations with an average variant allele percentage of 14%, indicating approximately 28% of cfDNA was of tumor origin. cfDNA VAP correlated well with the VAP in the liver metastasis (Figs 9A and 9B), but correlated poorly with the primary tumor (data not shown). Additional deep sequencing confirmed that an activating *PIK3CA* (H1047R) mutation was present only in the primary tumor, not in the liver metastasis or cfDNA, indicating that either the mutation emerged after metastasis or was not present in the subpopulation that seeded the metastasis. Seventeen additional somatic nonsynonymous mutations were called from the plasma sample. Closer examination revealed that eight of these (47%) were unique to the plasma, potentially originating from metastatic sites not sampled (Fig 9C). Two of those mutations were selected for validation via Ion Torrent PGM, both of them successfully validated (Fig 9C).

Chr	Base	Ref Base	Mut Base	Gene	AA Change	COSMIC	Primary Depth	Primary VAP	Met Depth	Met VAP	cfDNA Depth	cfDNA VAP
1	55603253	C	T	USP24	D1046N	None	63	0.0	87	26.4	164	27.4
1	158259864	C	A	CD1C	L4M	None	118	0.0	250	0.0	690	3.0
1	211749173	G	A	SLC30A1	R361*	None	109	0.9	136	19.1	234	19.2
1	225339739	A	C	DNAH14	N1231T	None	30	30.0	52	15.4	92	13.0
1	247655216	G	A	OR2W5	G263S	None	190	16.3	381	18.6	676	15.7
2	24086321	C	T	ATAD2B	R470Q	None	79	0.0	78	38.5	185	19.5
2	54119965	T	C	PSME4	K1391E	None	22	0.0	20	35.0	49	22.4
2	80529504	C	A	LRRTM1	A481S	None	52	0.0	154	0.0	303	2.6
2	103300761	C	A	SLC9A2	A464D	None	35	0.0	36	13.9	112	4.5
2	121736125	G	A	GLI2	R495Q	None	20	0.0	48	43.8	69	21.7
2	207173222	C	T	ZDBF2	P1324S	None	81	0.0	93	25.8	220	0.0
3	108836849	C	T	MORC1	A20T	None	18	22.2	77	37.7	126	22.2
3	142681704	C	T	PAQR9	A159T	None	77	0.0	199	0.0	333	11.4
3	178952085	A	G	PIK3CA	H1047R	COSM775	132	43.9	94	0.0	122	0.0
4	89618811	G	A	NAP1L5	A32V	None	79	0.0	124	25.8	141	12.8
4	104074295	A	G	CENPE	I1049T	None	35	0.0	50	20.0	78	11.5
4	104640577	C	A	TACR3	A86S	None	106	0.0	278	0.0	441	1.8
5	24498509	G	T	CDH10	Q505K	None	49	0.0	135	14.8	242	15.3
5	79025196	C	T	CMYA5	P203L	None	113	0.0	193	0.0	257	1.6
5	113698906	G	T	KCNN2	G145V	None	278	0.0	618	16.0	725	3.7
6	18457568	A	G	RNF144B	I172V	None	194	27.8	207	0.0	422	10.0
6	96984253	G	C	KIAA0776	Q263H	None	47	29.8	61	59.0	84	20.2
6	144999662	G	A	UTRN	D2534N	None	96	0.0	96	31.3	180	0.0
6	152419926	A	G	ESR1	D538G	COSM94250	183	0.0	257	23.3	482	17.0
7	36462337	C	T	ANLN	P799S	None	60	3.3	98	1.0	254	3.1
7	98460839	G	T	TMEM130	S90R	None	188	0.5	451	0.0	551	1.6
7	142625227	C	T	TRPV5	E289K	None	163	0.0	312	0.0	297	2.7
8	121238915	C	A	COL14A1	D638E	None	108	0.0	182	17.6	427	4.9
9	4662532	C	T	PPAPDC2	P53S	None	133	21.1	180	0.0	171	0.0
9	106900435	A	C	SMC2	T1136P	None	81	0.0	146	33.6	202	26.2
9	113341504	C	T	SVEP1	R107H	None	101	1.0	204	0.0	305	1.6
10	70056047	C	A	PBLD	A87S	None	12	0.0	42	45.2	74	6.8
11	5969386	G	T	OR56A3	K270N	None	195	0.0	376	33.8	630	18.3
11	14515193	G	C	COPB1	I162M	None	99	0.0	181	18.2	264	11.4
11	35496185	G	A	PAMR1	P145S	None	24	0.0	53	3.8	94	8.5
11	35496239	G	A	PAMR1	Q163*	None	8	0.0	24	12.5	41	7.3
11	74953028	G	T	LOC441617	G213T	None	111	20.7	268	26.1	307	17.3
11	95724773	G	T	MAML2	Q752K	None	82	0.0	85	32.9	319	20.7
11	124791236	G	A	HEPACAM	S350L	None	11	0.0	45	8.9	37	0.0
12	32860333	C	T	DNM1L	H95Y	COSM938812	39	0.0	76	25.0	81	0.0
12	58009707	C	T	ARHGEF25	R443C	None	92	35.9	205	45.4	328	33.2
12	76740663	C	T	BBS10	V368M	None	142	0.0	192	19.8	293	15.0
13	38229329	C	T	TRPC4	E594K	None	11	27.3	17	47.1	36	25.0
14	58943845	G	C	KIAA0586	E804Q	None	39	15.4	43	14.0	135	13.3
15	26026228	C	T	ATP10A	A198T	None	38	0.0	91	0.0	156	3.8
15	68118582	G	A	SKOR1	R139H	None	136	19.1	357	27.5	490	19.0
16	56533701	G	A	BBS2	R506W	None	63	6.3	94	0.0	180	0.0
17	4086831	G	A	ANKFY1	A605V	None	39	0.0	103	28.2	125	27.2
17	8131877	C	T	CTC1	R1153H	None	311	0.6	304	0.0	504	2.6
17	39183231	A	G	KRTAP1-5	S59S	None	11	0.0	34	0.0	27	0.0
18	7231403	G	T	LRRC30	E89D	None	88	0.0	207	29.5	491	23.6
18	13884823	C	A	MC2R	W232L	None	55	14.5	113	0.0	243	0.0
18	74091429	C	T	ZNF516	G881S	None	257	0.0	393	0.0	566	3.2
19	10600363	G	A	KEAP1	R498*	None	22	0.0	38	26.3	71	18.3
19	15739196	C	T	CYP4F8	Synonymous	None	59	32.2	165	37.0	223	24.2
19	38377433	T	C	WDR87	E2254G	None	550	0.2	506	21.1	830	16.3
19	42863259	C	T	MEGF8	R1785W	None	58	10.3	130	0.0	222	0.0
19	46351042	C	T	SYMPK	R215H	None	38	0.0	54	0.0	69	2.9
19	51206806	G	A	SHANK1	R502*	None	200	1.5	541	14.2	881	9.9
19	55995346	C	T	ZNF628	T925M	None	460	1.1	828	0.0	811	1.4
20	23805933	C	T	CST2	D86N	None	59	16.9	132	0.0	199	1.5
20	32199056	G	A	CBFA2T2	R121H	None	426	0.5	638	20.4	819	15.6
21	32526726	C	T	TIAM1	A1004T	None	52	0.0	103	0.0	195	2.1
21	43221400	T	A	PRDM15	Stop Loss	None	96	20.8	227	0.0	302	0.0
X	10085235	T	C	WWC3	L379P	None	154	0.0	339	21.5	404	15.8
X	73961595	A	C	KIAA2022	Y933D	None	250	0.0	446	0.0	727	1.9
X	118724673	C	T	NKRF	G239S	None	178	0.0	350	25.4	471	14.9
X	120009175	G	A	CT47B1	A117V	None	38	0.0	71	28.2	87	16.1
X	152158802	T	G	PNMA5	E447D	None	165	0.0	332	24.7	608	13.2
X	153132283	C	T	L1CAM	R751H	None	70	25.7	250	42.4	644	39.4
						Average	**111**	**6.4**	**196**	**17.1**	**309**	**11.0**

Table 5. Summary of Patient #2 somatic mutations. List of 70 somatic mutations from Figure 5 detailing depth and variant allele percentage in primary, metastasis, and cfDNA samples. COSMIC accession numbers listed for three mutations present in database.

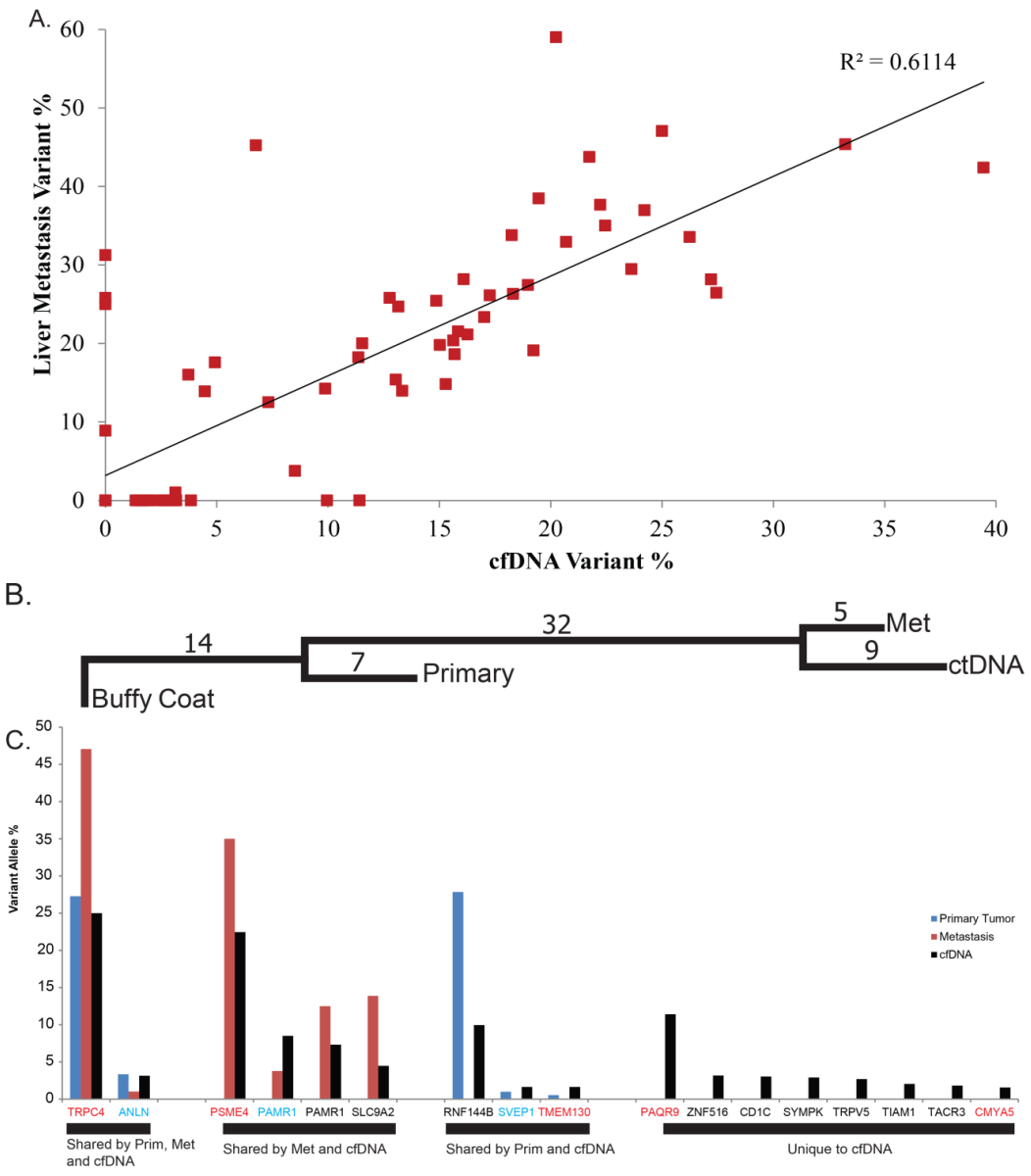


Figure 9. cfDNA and liver metastasis DNA are well correlated A) cfDNA variant allele percentage is correlated with liver metastasis variant allele percentage B) Maximum parsimony tree showing relatedness of samples, branch length are number of somatic, nonsynonymous mutations C) Seventeen additional mutations were identified uniquely in cfDNA, 9 of which have reads supporting them in the primary and/or met, but where not called due to insufficient sequencing depth or variant allele percentage. Genes in red text were successfully validated on the Ion Torrent PGM, genes in blue text failed to validate, genes with black text were not validated.

By sequencing cfDNA from plasma we are able to get a snapshot of the tumor, likely from multiple metastatic sites. The high correlation between the liver metastasis and cfDNA indicates that considerable information about the current tumor genome could be gained without the need for a biopsy. A mutation in *ESR1* (D538G), which has been shown to impart resistance to estrogen deprivation therapy, was found in both biopsies of the metastases and the cfDNA.[85,86] This mutation was not present in the initial exome sequence of the primary tumor and its absence was confirmed by subsequent validation sequencing of *ESR1* to a depth of 4,272X (Fig 10A). It is likely that the resistance of the tumor to the aromatase inhibitor Anastrozole can be explained by the mutant *ESR1*. This mutation was confirmed in a CLIA laboratory and anti-Estrogen Receptor treatments were considered between cfDNA sequencing and patient death. A total of 15 mutations were selected for validation on the Ion Torrent PGM, 13 of which were validated (Figs 8B and 9C). A second plasma sample was taken during response to TDM1 treatment (as determined by CT scan) and eight mutations present in the pre-treatment cfDNA sample were quantified in the during-treatment sample (Fig 10B). The pre-treatment cfDNA sample had a mean variant allele percentage of 13% across these eight sites while the during-treatment sample had a mean variant allele percentage of only 0.04% in the four sites containing mutant reads and no detectable mutant reads in four of the mutations tested.

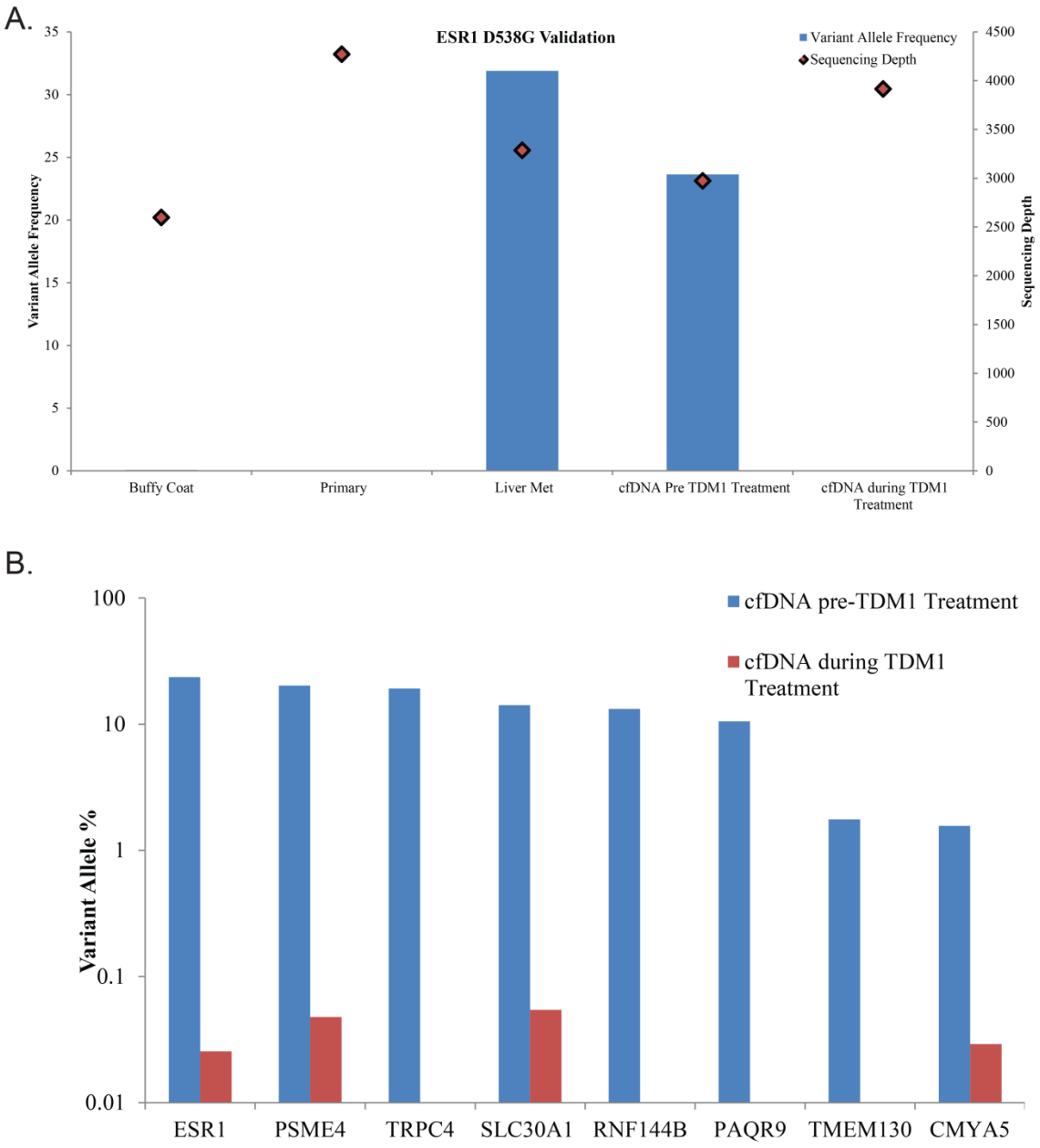


Figure 10. Patient #2 targeted resequencing. A) The *ESR1* mutation was sequenced to greater depth on the Ion Torrent PGM. B) Comparison of allele frequencies between pre- and during-TDM1 treatment cfDNA samples for eight mutations present in the pre-treatment sample.

Due to the relative high ctDNA percentage of 14%, we sought to determine if copy number variants (CNVs) could also be identified from the

cfDNA (Fig 11). Comparing the metastatic and cfDNA copy number data shows they closely correlate with each other. The CN ratios are lower in the cfDNA than the metastasis, but largely move in the same direction as would be expected from the lower tumor content between ctDNA and the metastasis.

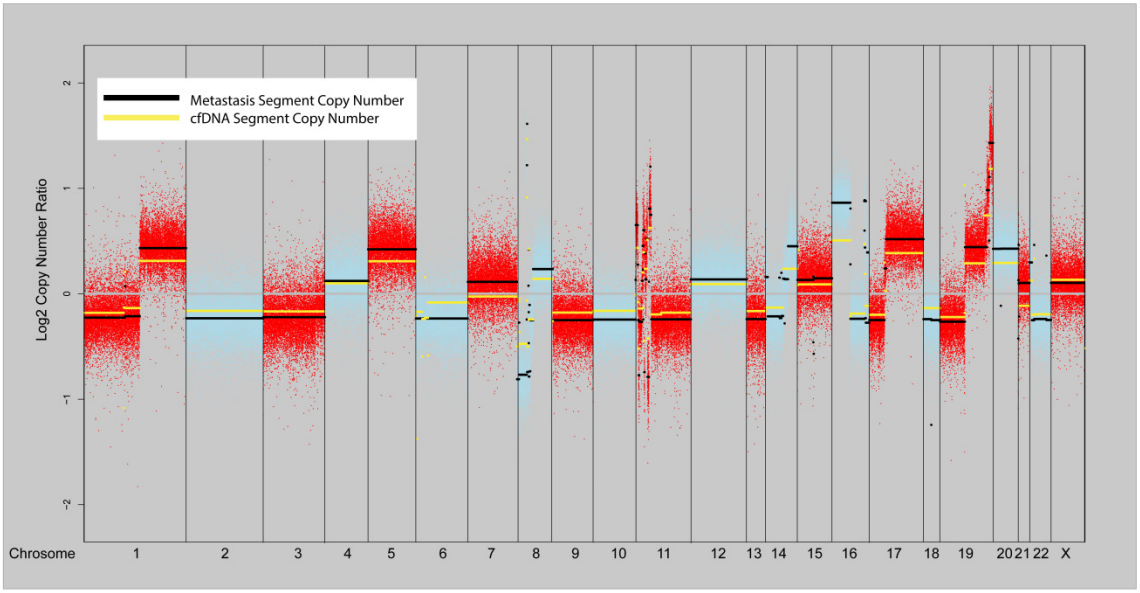


Figure 11. Copy number analysis of metastatic tumor and cfDNA. Log2 Copy number ratios of metastatic tumor (black) and cfDNA (yellow). Red/blue colors denote alternating chromosomes.

Discussion

In this study, we have demonstrated that whole-exome sequencing of cfDNA from patients with metastatic cancer can accurately identify clinically actionable mutations, and requires only minimal alterations to well-established sequencing protocols. We were able to sequence and gain valuable data from a plasma sample with a mean variant allele percentage of 3.7%, much lower than values demonstrated in previous studies and well below the frequencies of a substantial portion of metastatic cancer patients.[16,30,40,59,75] Adoption of this approach has the potential to greatly expand the utility of sequencing versus the biopsy-dependent approaches which are currently the standard of care. Mutations present in the cfDNA tightly correlated with mutations present in a synchronous metastasis sample, indicating that sequencing cfDNA can generate a more accurate picture of a patient's metastatic tumor genome than relying on a biopsy of the primary tumor. The cfDNA tightly correlates with tumor tissue taken at the time of plasma acquisition and can therefore be used to take "snapshots" of the cancer genome. Additionally, mutations unique to cfDNA were found in both patients, potentially representing lesions not sampled by biopsy. Validation via Ion Torrent sequencing confirmed that these mutations were not from normal tissue or the result of sequencing errors and were likely from sites not present in the biopsy. The inability to sample all metastatic sites within a cancer patient is a severe limitation of current sequencing techniques, and may be resolved with minimal modifications to standard sequencing procedures using cfDNA.

The finding that mutation-containing ctDNA fragments were significantly shorter than those carrying the WT sequence potentially indicates a different mechanism in the release of ctDNA compared to cfDNA release from normal cells. This finding is worthy of subsequent follow-up as it may provide insights into unique mechanisms of ctDNA release and provide potential enrichment strategies to preferentially isolate ctDNA from the plasma.

In addition to somatic mutations, CNVs were also identified from the cfDNA of patient #2. While requiring a higher ctDNA percentage than mutation identification, CNVs can provide important clinical information identifying potential therapeutic targets and representing potential resistance mechanisms. Characterizing copy number and somatic mutations can paint an even more complete picture of the tumor genome, and should be further explored to determine limits of sensitivity and reliability. Several recurrent, focal copy number aberrations, such as *HER2* amplification in breast cancer or *AR* amplification in prostate cancer can be highly amplified and therefore may be detectable in samples with ctDNA allele frequencies too low to completely characterize CNVs.

The two patients in this study had high levels of cfDNA in their plasma (Table 1), which allowed us to use over 100 ng of cfDNA to construct our sequencing libraries. However, for many patients a concentration of 10 ng of cfDNA per ml of plasma is more typical, indicating that multiple blood draws are required to get sufficient material for sequencing. Realizing this, we adopted the methods outlined in the Capp-Seq paper from the Diehn lab [75] that allows

libraries to be made more efficiently, requiring less initial input DNA. Using these methods we successfully produced complex libraries from less than 40 ng of cfDNA and successfully sequenced ~25% of the input DNA molecules (opposed to the ~1% efficiency achieved in our study). This improvement has allowed us to sequence sufficient cfDNA for nearly all our subjects.

Another advantage of sequencing cfDNA is the ability to sequence serially-collected and minimally-invasive plasma samples, allowing for near real-time monitoring of the tumor genome during treatment. The identification of emerging mutations may allow therapies to be started or stopped as soon as the tumor environment renders this advantageous. In the case of patient #2, it is possible that serial cfDNA sequencing would have identified the emergence of the *ESR1* mutation and treatment may have been adjusted from estrogen deprivation therapy (Anastrozole) to one targeting the estrogen receptor itself (*e.g.* Fulvestrant): this shift, and potentially others, may have delayed the progression of disease. In addition to looking for known resistance mechanisms, the nature of whole-exome sequencing allows for the identification of novel recurrent resistance mechanisms in a cohort of patients undergoing the same treatment, which may not be included in a targeted panel. Notably, during the response of patient #2 to TDM1 there was a dramatic reduction in the level of ctDNA, rendering it nearly undetectable by our sequencing approach. Monitoring via exome sequence during such periods would require extremely high sequencing depth, which would be prohibitively expensive with current sequencing costs.

A substantial focus has been placed on the sequencing of primary tumors and massive sequencing projects (TCGA *et al.)* have revealed a considerable amount of information about driver mutations in a variety of cancers. However, metastatic tumors, which are responsible for most patient deaths, are comparatively understudied. By sequencing primary tumors along with serially collected plasma samples it is possible to monitor metastatic progression at a genomic level. In patient #2 we observed an activating *PIK3CA* mutation in the primary tumor that was not seen in either the liver metastasis or cfDNA. It is likely that either the *PIK3CA* mutation became clonal after the metastatic process or that the mutation was not present in the metastatic clone; regardless, treatment with a PI3K inhibitor may have been effective in shrinking the primary lesion, but would have been ineffective against any of the distant metastasis. In contrast, sequencing of patient #1 showed that the cfDNA contained nearly all of the mutations identified in the primary tumor. While we were unable to get a sample of the metastasis, the low number of mutations unique to the cfDNA means it is not unreasonable to infer that there were relatively few differences between the metastasis and primary tumor. Sequencing cfDNA from larger cohort of patients may help us understand how metastatic progression varies in different tumor types and may identify therapeutically relevant patterns. The clinical utility of this method will depend largely on the systematic assignment of targeted therapies to identified cfDNA mutations.

Notably, services for cfDNA sequencing are becoming commercially available, but are based on panels and therefore have limited utility in a research

setting. We demonstrate here that there is significant value of whole-exome sequencing from cfDNA.

Subsequent Metastatic cfDNA Exome Sequencing

Following the success of our initial attempts at whole-exome sequencing of metastatic cfDNA, we received and sequenced an additional metastatic sample to screen for any variants of potential clinical interest. For this sample we only had the blood draw to work from (no matching tumor). The patient had metastatic pancreatic cancer which had an elevated cfDNA level of 22 ng/ml plasma. Using an improved set of whole-exome library creation methods, we were able to sequence the library at an improved efficiency of 25% (one-quarter of input cfDNA molecules were converted to a sequenceable library) and sequenced to an average depth of 130X. 80 Mutations were identified which passed our filtering, providing an average allele frequency of 16.2%, indicating that roughly a third of cfDNA was of tumor origin (Table 6). Of particular interest were the *TP53* and *KRAS* mutations, unfortunately this sequencing did not reveal a potential therapeutic target.

Chr	Base	Ref Base	Mut Base	Gene	AA Change	COSMIC	cfDNA Depth	cfDNA VAP
1	1284358	G	C	DVL1	L30V	None	45	11.1
1	53535777	G	A	PODN	G132S	None	75	13.3
1	163044346	C	T	RGS4	A205V	None	146	4.8
1	181727095	G	A	CACNA1E	D1448N	None	98	18.4
1	185902934	T	G	HMCN1	A602A	None	233	3.4
1	216144027	C	T	USH2A	W2299*	None	130	6.2
1	236332055	G	A	GPR137B	R155Q	None	141	19.1
2	1133348	C	G	SNTG2	L122L	None	71	7.0
2	31147091	C	T	GALNT14	R425Q	None	118	16.1
2	97039075	A	C	NCAPH	R738R	None	155	6.5
2	103068284	G	C	IL18RAP	L481F	None	227	6.2
2	169791907	C	T	ABCB11	R948H	None	170	4.7
2	222321345	T	C	EPHA4	T531A	None	61	13.1
2	230655915	T	A	TRIP12	K1415*	None	166	6.0
3	51690031	G	A	RAD54L2	R1024H	None	148	8.8
3	73440203	T	C	PDZRN3	D440G	None	111	37.8
3	151535228	C	G	AADAC	V71V	None	78	24.4
4	126389962	G	A	FAT4	R4065R	None	48	45.8
5	66462218	A	G	MAST4	E2404G	None	70	15.7
5	78573823	G	A	JMY	A375T	None	170	10.6
5	100147625	C	A	ST8SIA4	E336*	None	58	20.7
5	128844840	G	A	ADAMTS19	G267D	None	136	17.6
5	132545968	G	A	FSTL4	P544L	None	53	15.1
5	140188796	C	T	PCDHA4	A675V	None	103	9.7
5	140735432	G	A	PCDHGA4	R222H	None	114	9.6
6	37439654	C	T	CMTR1	R532W	None	43	14.0
6	100841692	G	A	SIM1	T414M	None	98	8.2
6	116429541	C	T	NT5DC1	A67V	None	67	14.9
7	98257797	C	T	NPTX2	R384R	None	121	10.7
7	105662776	C	T	CDHR3	T653I	None	211	10.0
7	127254961	C	T	PAX4	G103G	None	63	12.7
8	40011208	A	G	C8orf4	R53G	None	196	11.7
8	72951191	A	T	TRPA1	I735N	None	109	27.5
9	86570334	C	G	C9orf64	E187Q	None	69	20.3
9	116132239	C	T	BSPRY	H342H	None	100	13.0
9	120475528	C	A	TLR4	S374R	None	96	10.4
10	18266911	C	A	SLC39A12	Q278K	None	130	21.5
10	29169161	C	T	C10orf126	A102V	None	103	6.8
10	70728776	C	T	DDX21	P379S	None	171	24.6
10	124339154	T	G	DMBT1	V247G	None	40	12.5
10	129906258	G	A	MKI67	L1282L	None	148	6.8
11	36596443	C	T	RAG1	S530F	None	91	8.8
11	116719844	C	T	SIK3	D1165N	None	39	33.3
11	118869790	G	A	CCDC84	K118K	None	76	10.5
11	124794930	G	A	HEPACAM	R41C	COSM84121	52	15.4
12	25398285	C	G	KRAS	G12R	COSM517	56	33.9
12	39726830	T	G	KIF21A	D856A	None	109	24.8
12	56647525	G	A	ANKRD52	S322S	None	56	8.9
12	56722027	G	T	PAN2	Q190K	None	58	29.3
12	56722028	A	T	PAN2	T189T	None	56	28.6
12	109042552	G	A	CORO1C	F378F	None	100	8.0
13	25367267	A	C	RNF17	P341P	None	307	12.7
13	107823087	C	T	FAM155A	V379I	None	168	26.2
13	111268024	G	A	CARKD	M1I	None	82	20.7
14	21559205	C	A	ZNF219	R553S	None	32	15.6
14	74531951	C	T	ALDH6A1	G446E	None	98	48.0
15	48829969	T	C	FBN1	N192S	None	136	25.7
16	3707094	A	G	DNASE1	Q177Q	None	90	41.1
17	1581900	G	A	PRPF8	T589M	None	136	8.8
17	7578212	G	A	TP53	R213*	COSM10654	76	22.4
17	42475941	C	T	GPATCH8	R1168R	None	127	21.3
17	47044532	T	C	GIP	G21G	None	120	5.8
17	53392602	G	A	HLF	A156T	None	46	8.7
17	74869015	G	A	MGAT5B	G51R	None	48	25.0
18	5892009	G	C	TMEM200C	L18L	None	98	10.2
18	61305162	C	T	SERPINB4	G322S	None	119	19.3
19	11942503	C	T	ZNF440	P171L	None	212	7.5
19	46443413	G	A	NOVA2	T396M	None	50	10.0
19	46878881	T	C	PPP5C	H128H	None	117	5.1
19	58867670	G	A	ZNF497	C444C	None	44	36.4
20	31395614	C	T	DNMT3B	R823C	None	116	6.0
20	55209236	G	C	TFAP2C	L278F	None	102	6.9
20	62842623	C	T	MYT1	H452H	None	78	35.9
21	31744242	G	C	KRTAP13-2	S97C	None	72	11.1
22	37263465	C	T	NCF4	I101I	None	123	13.0
22	39134215	T	A	SUN2	D668V	None	92	16.3
22	45794997	T	G	SMC1B	D364A	None	146	18.5
22	46932046	A	C	CELSR1	V341G	None	109	14.7
X	100079201	G	T	CSTF2	V219V	None	83	8.4
X	144905527	G	A	SLITRK2	E528E	None	86	14.0
						Average	**106.2**	**16.2**

Table 6. Summary of somatic mutations in metastatic pancreatic cancer patient. List of 80 somatic mutations detailing depth and variant allele percentage in a cfDNA sample. COSMIC accession numbers listed for three mutations present in database.

Methods

Patient Enrollment

Written consent was obtained from two patients with metastatic cancer for enrollment in this study. The study and consent procedures were approved by the Oregon Health & Science University Institutional Review Board and in accordance with federal and institutional guidelines. Up to 40 mls of blood was collected in EDTA tubes. Plasma was isolated as described previously [30] and stored at -80°C until cfDNA was extracted using the QIAamp Circulating Nucleic Acid kit (Qiagen). Buffy coat was isolated from the same blood sample and DNA was extracted using the DNA Blood Mini kit (Qiagen). As part of the aforementioned study and consent procedure, FFPE tissue from the patient's primary tumors was acquired from archived pathology samples. Patient #1's sample was acquired from the University of Washington Pathology Department in Seattle, WA (http://www.pathology.washington.edu/clinical/dermpath/contactinfo). Patient #2's sample was acquired from Compass Oncology in Vancouver, Washington (http://compassoncology.com). FFPE tissue was extracted using the DNA FFPE Tissue kit (Qiagen). The same patient's liver metastasis was taken from a frozen core biopsy and extracted with the DNeasy Blood & Tissue kit (Qiagen).

Whole-Exome Sequencing

A minimum of 100 ng of cfDNA and 0.3-2µg of DNA from buffy coat and tumor tissue were used to create sequencing libraries. Agilent SureSelect XT reagents and protocol were used to prepare sequencing libraries. DNA from buffy coat and tumor tissue was sonicated to an average size of 150 bp using a Covaris E220. Plasma DNA samples were not sonicated, as plasma DNA is already highly fragmented. Hybrid capture was conducted using Agilent SureSelectXT Human All Exon V4+UTRs. 100 bp paired-end sequencing was conducted on an Illumina HiSeq 2000. An entire lane was dedicated to sequencing plasma DNA samples and all other libraries were sequenced two-to-a-lane. To maximize sequencing depth and avoid PCR duplicates, the plasma sample from the patient with metastatic sarcoma was made into three separate libraries, each sequenced on one full lane each, giving an average sequencing depth of 1,034X. Only a single library was needed to achieve sufficient coverage of cfDNA for patient #2.

Improved Whole-Exome Methods

Following publication of the CAPP-seq paper from the Diehn lab we adopted their library preparation method.[66] In brief, we ordered HPLC-purified, indexed sequencing adapters and blocking oligos from IDT (idtdna.com). These were used in combination with the Hyper Prep DNA Library Preparation Kit (Kapa Biosystems), at a 100:1 adapter:template ratio for cfDNA samples and 20:1 for buffy coat DNA. This library was then used as input for the Agilent SureSelect XT

hybrid-capture protocol and reagents using the all-human exon v5 set of capture baits. These libraries were paired-end 100 bp sequenced using the Illumina HiSeq 2000 platform.

Bioinformatic Analysis

In order to detect mutations we aligned HiSeq paired-end reads with hg19 human reference genome using bwa.[87] We used bwa aln to find the coordinates of input reads and then used bwa mem in order to generate alignments in a sam format.[87] We converted the sam format to bam (binary) format using Samtools import (v 0.1.19).[88] After sorting and indexing the reads in the bam formatted file, we use Picard Tools[89] MarkDuplicates to remove duplicate reads generated during the PCR amplification stage: removal is done by finding all reads that have identical 5' coordinates and keeping only the read pair with the highest base quality sums. After duplicate removal we realigned reads around SNVs and indels using the GATK Software Library.[90,91] The three libraries of the sarcoma patient were combined after PCR duplicate removal: local positions to target for realignment were called using RealignerTargetCreator and the reads were realigned using IndelRealigner. Finally, quality scores were recalibrated. This was done using GATK BaseRecalibrator and PrintReads, which binned reads based on the original quality score, the dinucleotide, and the position within the read. Sequencing statistics are summarized in Table 3 and were generated using Samtools flagstat, GAKT DepthOfCoverage, and Bedtools pairToBed.[92]

To call mutations we compared the tumor samples with the normal samples using muTect v1.1.4 using the buffy coat as a matched normal.[93] Variants were considered somatic mutations if: (a) they were not present in the dbSNP database[94] (except if the variant was also in the COSMIC database[95] eg *KRAS* and *PIK3CA* mutations), (b) there was ≥30x sequencing depth at that site in the tumor/plasma sample and ≥10x sequencing depth in the matched normal sample, (c) it had a variant allele percentage of ≥10% for the tumor samples and ≥1.5% for plasma samples, and (d) there were at least two reads containing the variant allele. Mutations in cfDNA were then further filtered out if the matched normal had >1 read supporting the mutation or the mutation was only present in one strand of the cfDNA. Impact of variants was checked using Mutation Assessor v2 (www.mutationassessor.org).

Fragment Size Analysis

Fragment size analysis was conducted using only the unsonicated cfDNA libraries. Samtools view was run at the mutation or SNP sites of interest, this printed out all the reads that mapped to that position, each read was then checked for the presence of the WT or mutant base. Average read lengths for each set of reads was measured and statistical significance determined using a student's t-test. As a comparison this analysis was also conducted on the sonicated cfDNA libraries, which did not show a significant difference between WT and mutation containing reads.

Copy-Number Analysis

Copy number analysis was conducted using a previously published method.[96] Briefly, read depth across the exome was determined and combined into adjacent segments. Average read depth for each segment was then compared to the same segment in the matched normal and converted into a log2 copy number ratio and plotted using the DNAcopy R package. Due to the lower quality of the FFPE DNA from the primary tumor, the copy number analysis was unsuccessful. For patient #1 ctDNA percentage was too low to be successfully analyzed.

Mutation Validation

Primers were designed to cover a selection of mutations identified in each patient and then used to PCR amplify buffy coat, plasma, and tumor DNA samples from both patients. For each sample, amplicons were pooled in equimolar amounts and 10-100 ng were used for library creation using the Ion Xpress Plus Fragment Library Kit. Sequencing templates were generated using emulsion PCR on the Ion OneTouch 2 using the Ion PGM Template OT2 200 kit. Up to six barcoded samples were multiplexed on Ion 316 v2 chips. Sequencing was performed on a Personal Genome Machine (PGM) sequencer (Ion Torrent) using the Ion PGM 200 v2 sequencing kit. Torrent Suite software version 4.0.2 was employed to align reads to hg19. Reads were visualized using IGV v 2.2.32 (Broad Institute) and variant allele frequencies were determined for sites previously identified via Illumina sequencing.

Chapter 2:

Development of a High-Accuracy Hybrid-Capture Based ctDNA Detection Method

Timothy M. Butler, Paul T. Spellman

Introduction

ctDNA is often present in primary and residual disease at allele frequencies well below one percent. This makes using standard next-gen sequencing methods problematic, as their per-base substitution error rates are typically around 0.1%, causing the signal from rare ctDNA mutations to become overwhelmed by the noise of sequencing error. A variety of techniques have been developed which reduced that error rate, allowing for the detection of rare ctDNA molecules. The two most popular methods are droplet-digital PCR (ddPCR) and the Safe Sequencing System (SafeSeqS). In ddPCR, DNA is separated into 10's to 100's of thousands of separate PCR reactions contained inside oil droplets each with either 0 or 1 of the DNA molecules of interest. Fluorescent, allele-specific probes indicate whether a given droplet contains the wild-type or mutant sequence. Then the fluorescent droplets are counted and the mutant allele frequency determined. This technique has a reported accuracy of 1 part in 10k, but suffers from the allele-specific nature of the assay, limiting it to analyzing at most 5 alleles simultaneously.[28] The lack of multiplexing ability severely limits the sensitivity of the assay and the reliance on a limited number of mutations is problematic when dealing with heterogeneous tumors as tracking only one or two mutations in ctDNA may not be representative of the ctDNA as a whole.

SafeSeqs is a PCR based approach which relies on adding random, degenerate sequences to the initial DNA molecules (via an initial 2-5 cycle PCR

using template-specific primers containing degenerate barcodes), then copying and sequencing those molecules multiple times to create a consensus sequence of all reads containing the same degenerate barcode.[31] This allows for errors introduced either in the PCR amplification of the library or errors made by the sequencer to be corrected, lowering the sequencing error rate from 1 in 1k to around 1 in 50k. The complex primers required for this assay are able to span up to 150 bp allowing for a single assay design to be used to detect multiple types of mutations in a given gene. However, there is only a limited ability to combine multiple sets of primers into a single assay, creating many of the same limitations to the technique as ddPCR.

In order to overcome the limitations of these techniques, we sought to develop a hybrid-capture based approach which would (like SafeSeqS) introduce degenerate sequences to our input cfDNA molecules, but then allow for hybrid-capture of dozens of mutations identified from a patient's primary tumor. Using a larger panel of mutations to identify ctDNA would improve sensitivity, and potentially allows us to detect differential changes in allele frequency between multiple mutations. The technique we developed is Dual-Indexed Degenerate Adapters (DIDA).

Results

Dual-Index Degenerate Adapters

There are two key aspects of the adapter design: the degenerate barcode and the dual-index (Fig 12A). The adapter uses the standard design for Illumina sequencing adapters, with the key difference that a portion of the indexing region uses a string of 6 degenerate bases, which following ligation to cfDNA (Fig 12B) and library amplification (Fig 12C), will create multiple copies of the same template molecule with the same degenerate sequence. This degenerate barcode is necessary because the standard method of identifying independent molecules is to rely on the position of where the reads map to. However, in high depth sequencing, multiple independent molecules can map to the exact same position, and be incorrectly identified as duplicates. Following sequencing, these copies will be grouped by degenerate barcode and mapping position (Fig 12D), then collapsed into a Single-Stranded Consensus Sequence (SSCS) (Fig 12E). We required at least three of the copies (family members) to create the SSCS. In creating the SSCS, only bases which agree in over 90% of the family members are called, positions which fail this filter are instead called an “N” for unknown base. This approach allows stochastic errors introduced during library amplification and sequencing to be filtered out, lowering the substitution error rate to approximately 1 in 10-50k.

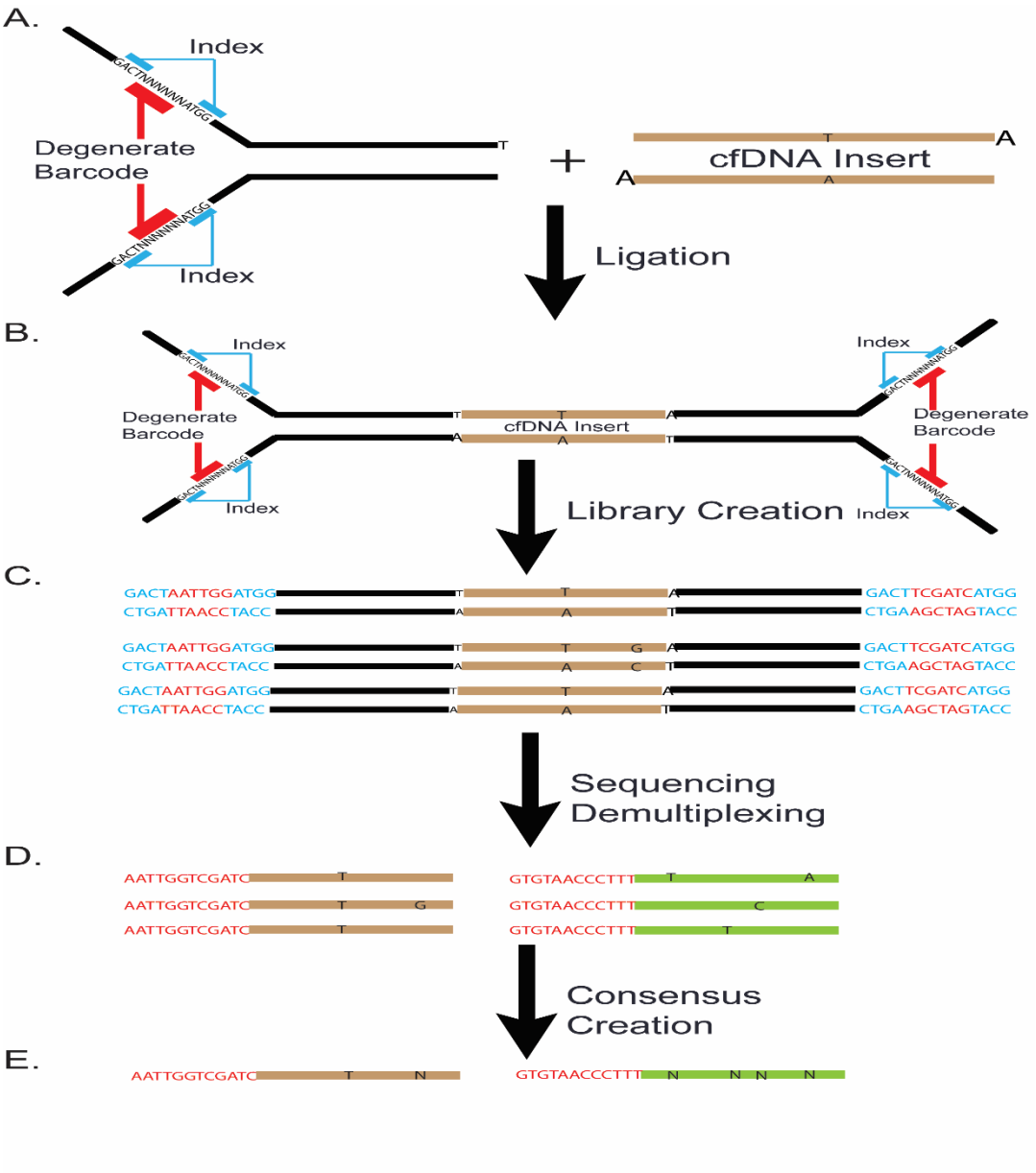


Figure 12. Overview of DIDA adapters, library creation, and consensus creation. A) Schematic of T-tailed DIDA adapter, regions in black are standard Illumina adapter sequences. B) Following ligation to A-tailed cfDNA (containing a point mutation) adapters are ligated to each side, index sequences are identical. C) Following library creation and amplification multiple copies of the same template molecule are created, a G is introduced through PCR error. Blue sequences are the indexes, red are the degenerate barcodes. D) Following sequencing and demultiplexing, reads are grouped by degenerate sequence and mapping region. Brown and green represent two different genomic regions. E) Reads are collapsed into SSCS, stochastic sequencing/PCR errors are replaced by N's.

The dual-index is an important protection against cross-contamination and misassignment. There are two major sources of this error, contamination of the barcodes (during synthesis and/or liquid handling)[97] and so-called index jumping, a process where PCR amplification creates a chimeric product through PCR-mediated recombination.[98,99] The combined effect of these factors can lead to assigning the incorrect index to a read 0.1-0.3% of the time.[97,98] This can be quite problematic when attempting to identify mutations at low allele frequencies as mutant reads from a sample with a higher ctDNA allele frequency could be misassigned to a sample with a lower ctDNA allele frequency, confusing cross-contamination for real signal. To overcome this complication, a dual-indexing strategy can be employed where each of the indexes of the Y-arm of the sequencing adapter are identical (Fig 12A). For a given read to be misassigned, both indexes would have to switch to a different pair of identical indexes. The study which proposed this method found it generated a 200-fold reduction in misassignment to 1 in 100k.[98] Utilizing this approach in our own data we found that in a run with 16 multiplexed DIDA libraries (and therefore 240 additional, incorrect index combinations), 3.4% of the total reads were assigned to an incorrect index combination (Fig 13A). Looking more closely at the incorrect indexes we found that the two most common incorrect combinations (1+10 and 10+3) accounted for roughly 6% of the total misassignments (Fig 13B). Taking the worst case scenario of a misassignment to the 10+10 index we estimate that only 1 out of 110k reads would receive this misassignment. In this hypothetical scenario, sample 1 (index 1+1) would receive a contaminated adapter containing

index 1+10, then had an “index jumping” event to create index 10+10. Assuming that this misassignment requires two independent steps, the likelihood of occurring is simply the product of the misassignment frequencies. Overall, these numbers are largely in agreement with the published literature, and gave us confidence that misassignment events would be sufficiently rare to minimize their impact on mutation detection.

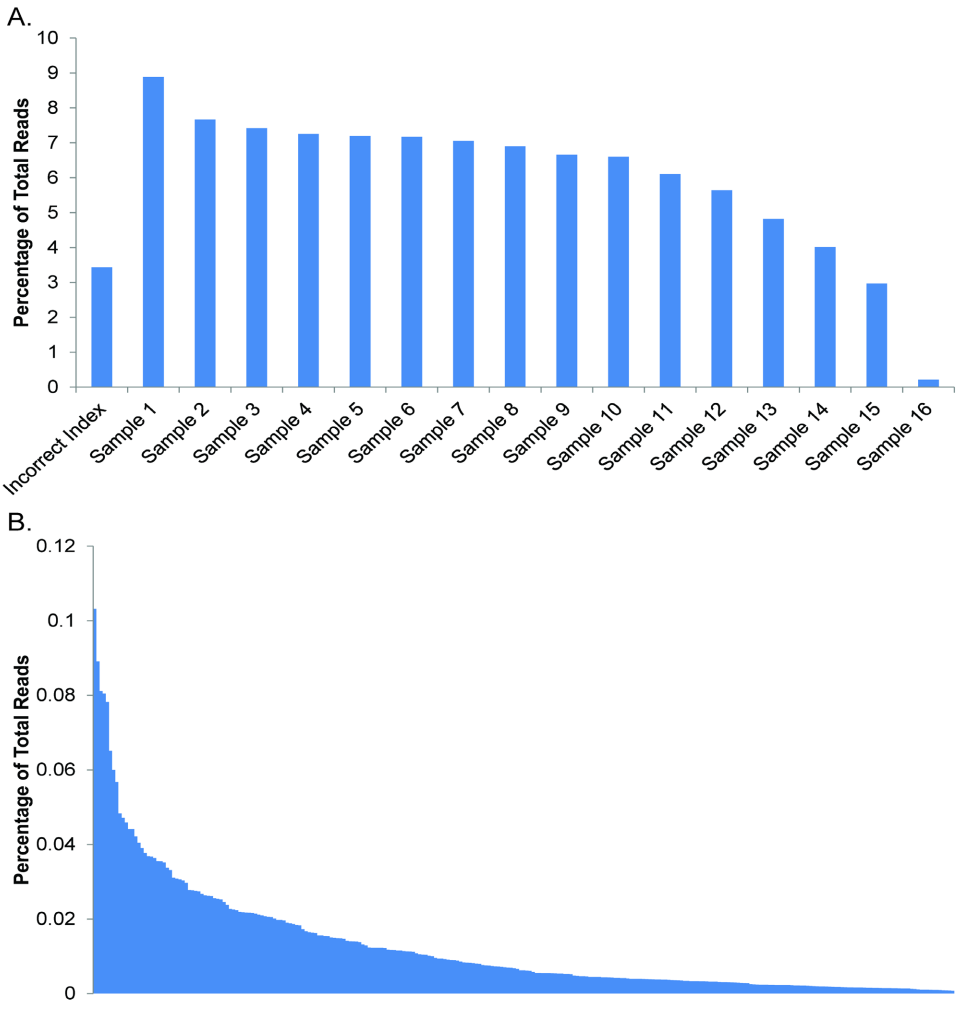


Figure 13. Distribution of dual-index assignment in DIDA run. A) Percentage of total reads assigned to each of 16 different multiplexed DIDA libraries, along with the 240 different combinations of incorrect index pairs. B) Distribution of incorrect read pairs as total reads from sequencing run in A.

Hybrid Capture

The final aspect of the assay is the hybrid capture panel. Hybrid capture allows for the enrichment of specific regions of the genome through the use of biotinylated oligos. This is necessary as it quickly becomes cost prohibitive to sequence a large genomic region to the depth necessary for rare ctDNA detection. Typical hybrid capture experiments involve a single round of hybridization, capture, and amplification before sequencing. This results in >80% of sequenced reads being “on-target” (mapping one of the regions targeted). In our assay, much smaller panels are used (~10 kb vs >1 mb), the smaller panel size requires two rounds of hybrid capture to ensure efficient on-target enrichment (Fig 14).

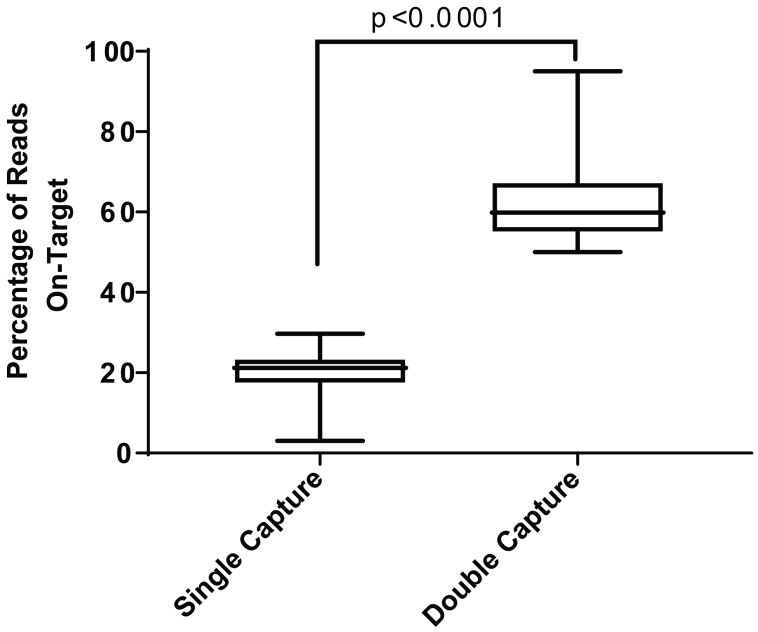


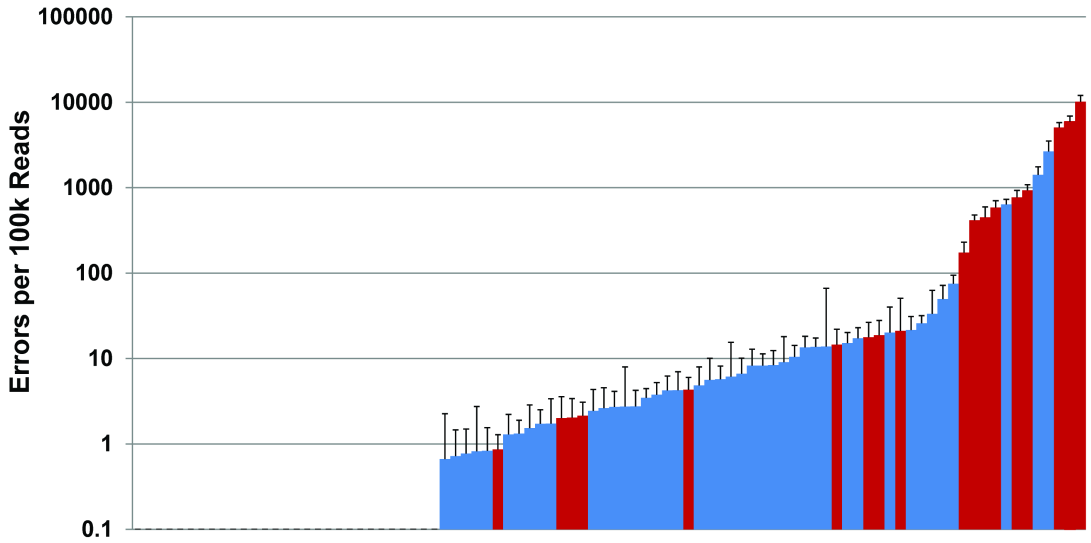
Figure 14. Benefits of two rounds of hybrid-capture. 27 DIDA libraries were sequenced following single and double hybrid-capture. The double capture protocol significantly improved on-target percentage (62% vs 20%, paired t-test).

Accuracy of DIDA Sequencing Panels

A 10.2 kb hybrid capture panel was ordered targeting 96 somatic mutations previously identified from whole-exome sequencing of 5 primary breast tumors. To test the panel's accuracy, 9 separate DIDA libraries were created from a negative control metastatic prostate cfDNA sample chosen both for its extremely high cfDNA concentration (738 ng/ml) and the absence of any mutations that the panel was designed to detect. 30-300 ng of input cfDNA was used to generate the library, allowing for extremely high depth sequencing. After pooling the 9 sets of SSCS, the average depth across the panel was 124,000X. Of the 96 sites the panel was designed to capture, only 90 were successfully captured. With 62 of those 90 sites having an error frequency of less than 1 in 10k, and 29 sites not showing a single sequencing error (Fig 15A). Looking at the entire 10.2 kb capture region, we found similar performance, with 63% of the sequenced sites having an error frequency of less than 1 in 10k (Fig 15B).

Analyzing previously sequenced whole-exome data, we found the error-prone sites were significantly ($p<.0001$) more likely to have variant reads present in those exomes (red bars Fig 15A). The most likely explanation for this is that these represent regions of the genome that are difficult to map. The red bars don't represent actual sequencing errors, but rather mapping errors. Requiring a stricter mapping quality filter in the initial exome sequencing used to design the panel would have removed 10 of the 12 most error prone sites. This filter was utilized in future panel design.

A.



B.

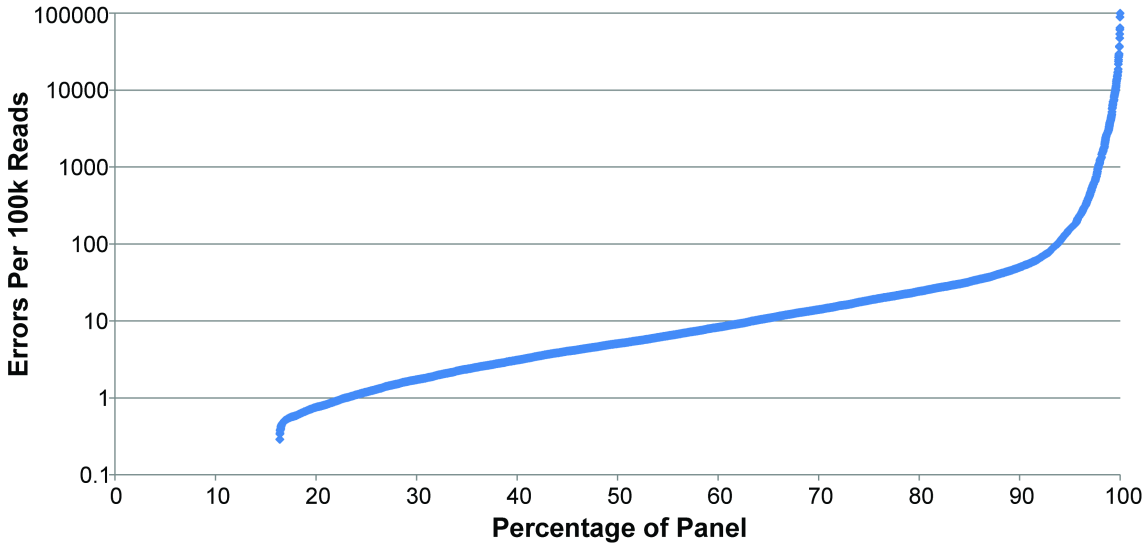


Figure 15. Accuracy of 96 site hybrid capture panel. A) Error frequency of 90 mutations across 9 separate negative control samples sequenced to an average pooled depth of 124,000X. Red bars indicate mutations for which variant reads were identified in unrelated exome-sequencing experiments. 29 sites showed no errors. Error bars +/- SEM. B) Error frequency across entire 10.2kb of panel. 16% of the panel showed no errors.

We also analyzed the distribution of errors to see whether certain bases were more or less error prone (Fig 16A). A significantly higher proportion of sites without a single detected error had A's or T's as their reference base, with fewer perfect G's, and almost no C's having perfect reads. Analyzing the frequency of specific types of errors, we again saw that A's and T's outperformed G'c and C's (Fig 16B). The prevalence of C>T errors was not surprising, as it is the result of deamination and had been previously seen in other sequencing error correction methods, whereas the C>G and G>C errors were unexpected.[100] However, despite the high error rate associated with C>G and G>C errors, these errors were relatively rare, composing only 1.3% and 3.4% of the total errors at C and G sites, respectively. These results were incorporated into subsequent panel designs, preferring WT A's and T's, and taking specific care to avoid mutations which were C>G, C>T, or G>C. This somewhat restricted the number of mutations which could be incorporated into a panel, but with 30-50 mutations present in a typical breast cancer exome we were still routinely able to identify at least 20 mutations of interest. The second 96 site capture panel we designed incorporating these rules, along with stricter mapping quality filters of the initial exome mutation calls, improved the fraction of the panel performing better than 1 error in 10k reads from 69% to 80%.

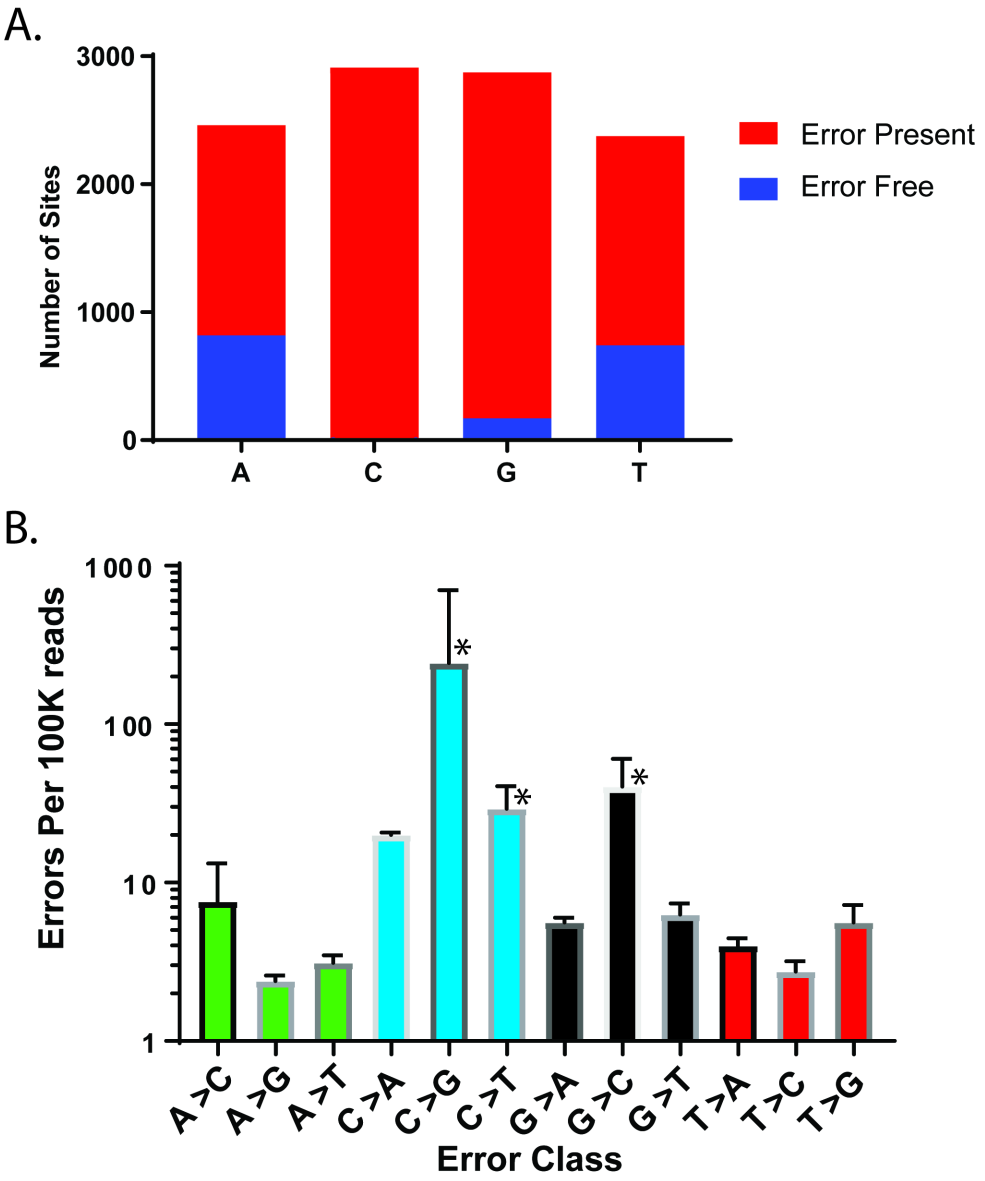


Figure 16. Distribution of sequencing errors. A) Comparison of the number of sites which are error-free based on the identity of the WT base. All pairwise comparisons except A-T are statistically significant ($p<0.0001$). B) Median error frequency for each class of sequencing error. Asterisks indicate that class is significantly different from all none asterisk classes. Error bars 95% CI.

In addition to accuracy, we were concerned about the reproducibility of the assay. As the negative control sample was specifically chosen for not sharing any mutations the panels were designed for, we relied on heterozygous SNP sites adjacent (within 200 bp) to the mutations of interest (Fig 17). The 17 SNPs

identified generally showed low variations between samples, with standard deviations ranging from 1.4%-6.8% (median of 2.6%).

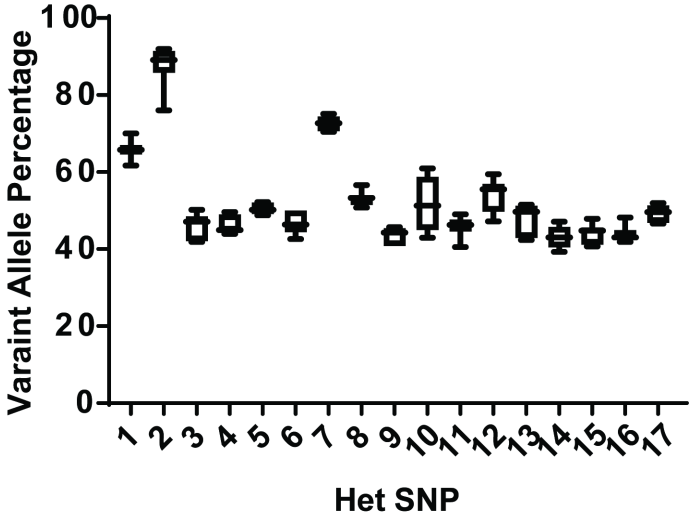


Figure 17. Reproducibility of variant allele percentages. Box and whisker plot depicting variant allele percentages for 17 heterozygous SNPs which were captured by the panel. All SNPs were within 200 bp of tumor-specific mutations the panel was designed to capture.

Identification and Filtering of Tag Swaps

We noticed that mapped SSCSs tended to cluster at identical insert positions. This is not surprising for high depth sequencing, however, at sites with mutations present, identically mapped reads tended to have the exact same sequence, which was quite surprising (Fig 18A). In this example there are 4 sets of mapping positions all showing a C>G mutation, this position in reality had 1,200 mapped SSCSs, making it extremely unlikely that the 15 mutant reads only mapped to 4 different positions, and no reads mapping at those positions had the WT base. Looking more closely at the barcodes (red sequences in Fig 18a), we noticed that SSCSs with identical mapping sites also had one half of their

barcode match exactly (shown schematically as being on either end of the read). This was likely caused by one of the barcodes being replaced during one of the amplification steps, similar to the “index jumping” phenomenon mentioned above. In the example of the blue reads, we see the right side barcodes having identical TAAG sequences, while the left differ by a single base, this is likely a sequencing or PCR error which created a new barcode. Looking at the orange reads we see the top 3 reads have identical left barcodes, and the bottom 4 reads have identical left barcodes. In this example there were likely two separate swaps which occurred, linked by the GCTG-ATTT 3rd read, allowing us to collapse these 6 reads down to 1.

The presence of these tag swaps could potentially skew the mutant allele frequencies we detect, and cause us to overestimate our sequencing depth and sensitivity. We set out to filter these reads out by collapsing SSCSs which had the entirety of either half of their barcodes match identically, and mapped to the exact same start/stop positions (Fig 18B). This filter removed on average 10% of the SSCS, with samples in the same sequencing run behaving similarly (Fig 19A). There are significantly fewer SSCSs filtered from single-captured samples vs double-captured ones (Fig 19B). A possible explanation is that the additional PCR cycles associated with the double-capture makes tag swapping more likely. Despite this additional filtering, depth remains significantly higher in the double-captured libraries (Fig 19C).

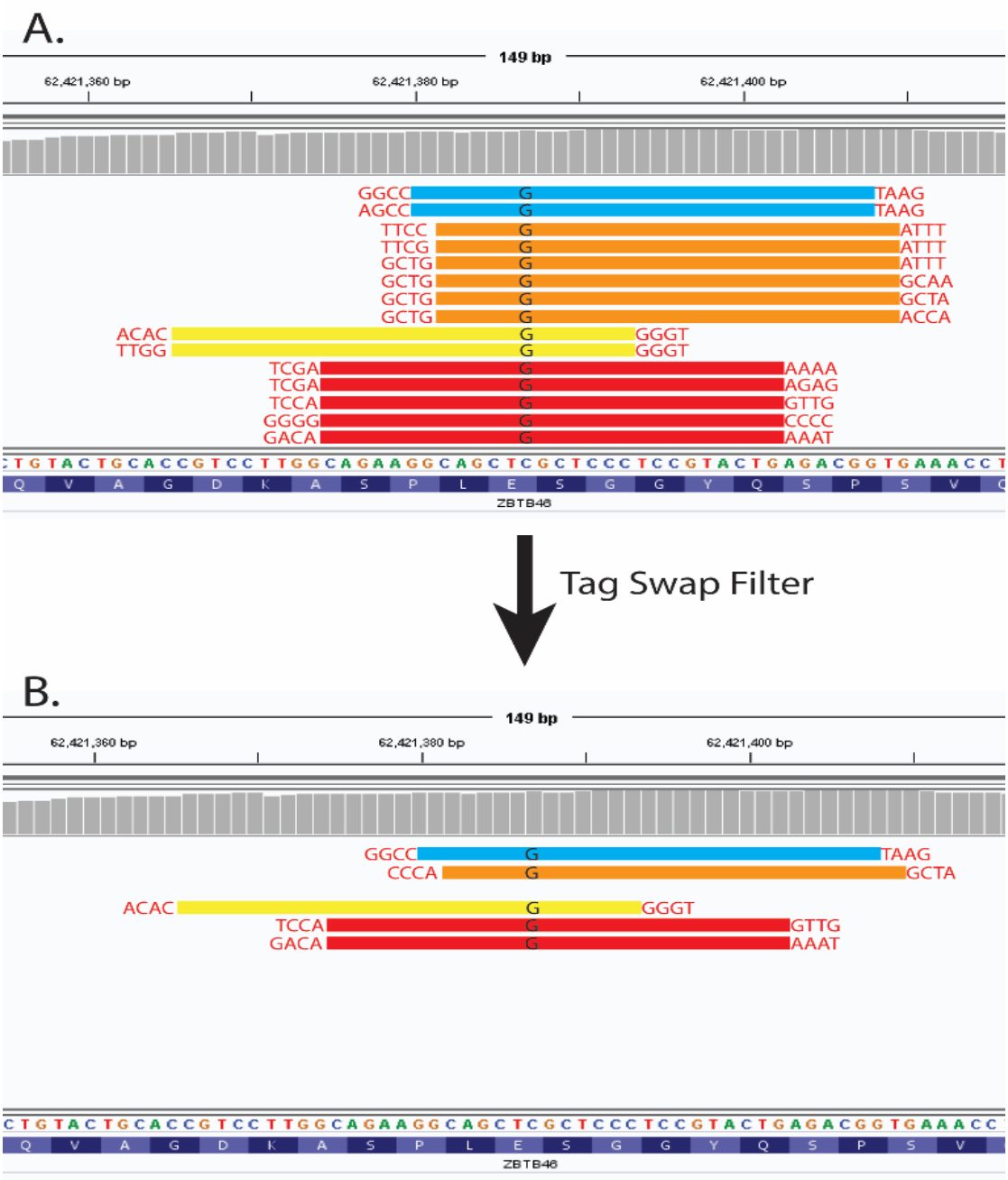


Figure 18. Identification and filtering of tag swaps. A) Representation of IGV view of SSCS reads mapping to a C>G mutation. All reads of the same color have the exact same start and stop sites. 4 red bases on either end of the read are representations of the first (left) and second (right) barcode assigned during consensus creation. Reads of the same color and with identical barcodes on either end are likely tag swaps. B) IGV view following tag swap filtering, orange and red reads may be duplicated molecules unable to be filtered out using tag swaps.

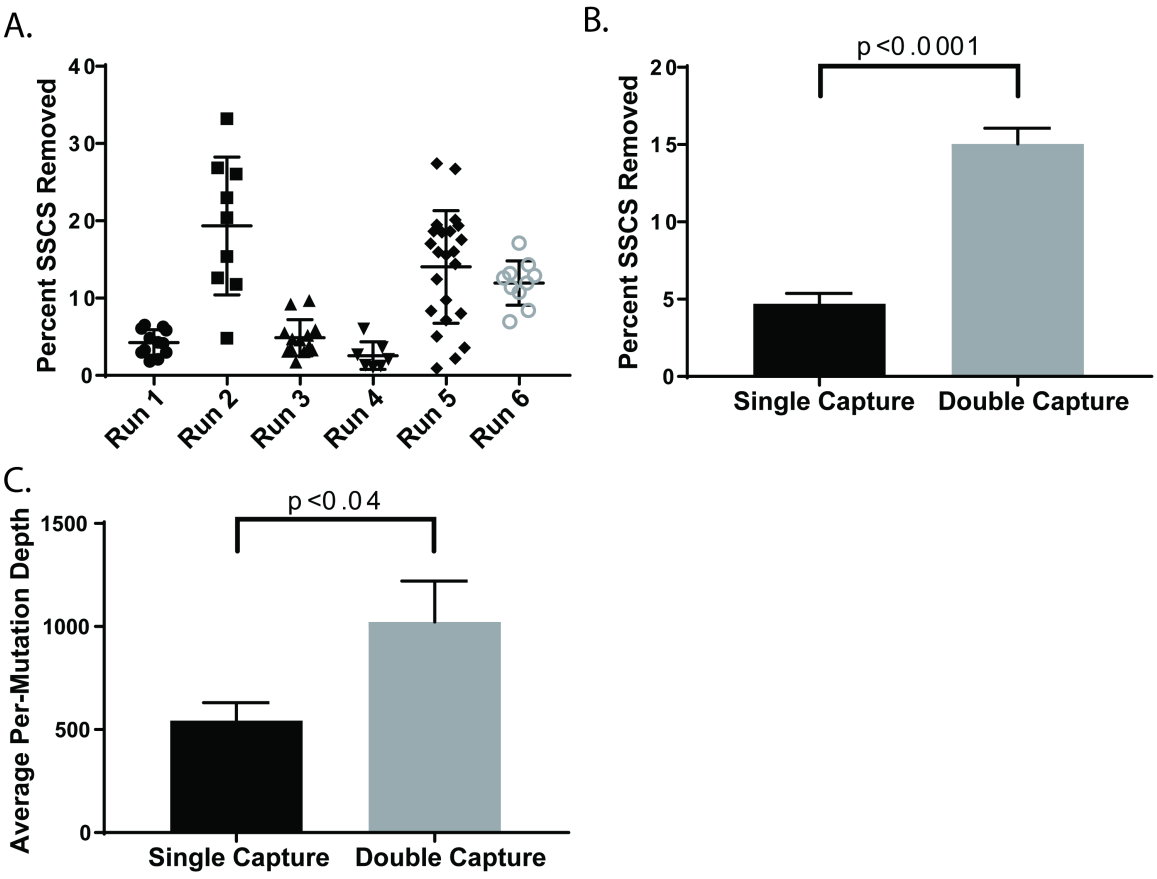


Figure 19. Tag swap filtering. A) Percent of SSCSs filtered out by the tag swap filter for each of six separate sequencing runs. Each point represents an individual DIDA sample. Error bars +/- SEM. B) Single-captured libraries have significantly fewer reads removed by the tag swap filter. Error bars +/- SEM. C) Following swap filter, double-captured libraries have significantly higher average depth.

Following tag swap filtering, there was still substantial evidence of mutations clustering in SSCSs mapping to identical positions, indicating that our tag swap filter was not identifying all potential tag swaps (red reads Fig 18b). When two or more of these SSCSs mapping to the exact same position overlapped with a heterozygous SNP site, both alleles were represented only 17% of the time (opposed to the 50% that would be expected). This indicates that

the majority of SSCSs which passed the tag swap filter, but still map to the exact same position, are likely derived from the same initial cfDNA molecule. In an effort to remove these artifacts, we made the decision to collapse SSCS which have identical cfDNA sequences and map to the exact same position regardless of their barcode sequences.

To determine whether this issue was unique to our data, or a more widespread phenomenon, we analyzed published data from the iDES paper.[33] Their barcoding strategy is similar to ours, using a 4 bp degenerate barcode in only one of the indexing regions, and two, 2 bp degenerate sequences on either side of the cfDNA insert (Fig 20). For purposes of analysis we combined the two, 2 bp barcodes into a single 4 bp barcode, allowing us to use our analysis pipeline on their data. Following SSCS creation, we determined that 6% of their consensus sequences were indeed tag swaps. From our own data we believe this is likely an underestimation of the issue, however it is possible their barcoding strategy of 2 bp barcodes on either side of the DNA insert may in fact be less prone to tag swapping.

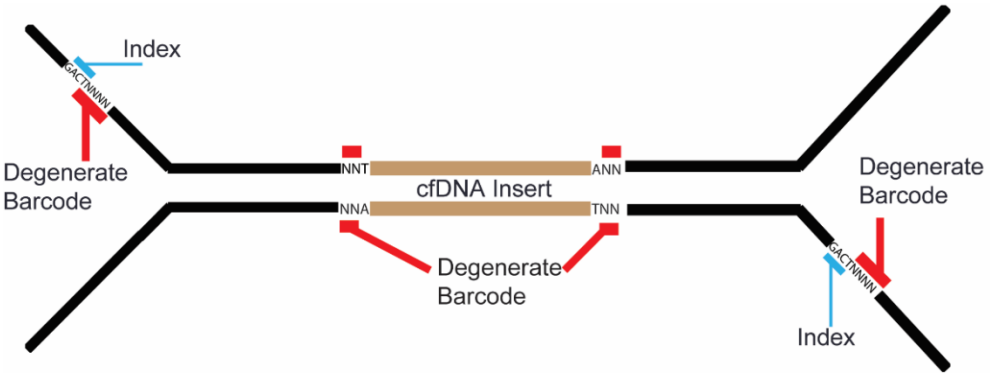


Figure 20. Overview of iDES adapter (Adapted from Newman et. al.[33])

Discussion

We were successfully able to develop an assay which reliably generates low-error SSCSs targeting up to 96 mutations of interest simultaneously. Using some relatively simple filtering metrics, we were able to design capture panels which could sequence 80% of their mutations of interest with less than 1 error in 10k reads. A subset of sites showed substantially better error frequencies, and it is possible that through more advanced filtering techniques, these extremely accurate sites could be enriched for.

We also identified a systematic source of depth overestimation using adapters containing degenerate barcodes. Experimental and bioinformatic methods need to be further optimized to minimize and efficiently filter out these tag swaps. After aggressive filtering of the tag swaps the overall efficiency of DIDA library creation dropped to an average of just under 10%. This is considerably less than the 50% efficiency reported in the iDES technique, but that number does not take into account any tag swap filtering, meaning they are likely overestimating their efficiency. Our 10% efficiency is still sufficient to reliably survey 10,000 molecules across a given patient-specific panel, a substantial improvement from our attempts using SafeSeqS.

Methods

DIDA Adapters

DIDA adapters were modified from Illumina TruSeq HT adapters, expanding the index region of the adapter to 14 bp, 6 bp of which were ordered as degenerate N's (machine mixed). To ensure minimal cross-contamination of indexes, adapters were ordered as HPLC purified, TrueGrade adapters.

Adapter sequence, X=index N=degenerate barcode:

i5 Adapter:

AATGATACGGCGACCACCGAGATCTACACXXXXNNNNNNNXXXXACACTCTTTCCCTACACGACGCTCTTCCGATC*T

i7 Adapter:

/5Phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACXXXXNNNNNNNXXXXATCTCGTATGCCGTCTTCTGCTTG

Hybrid Capture Panels

Hybrid capture panels were designed using the IDT Target Capture Probe Design tool (https://www.idtdna.com/site/order/ngs). 96 sites of interest were chosen from previously whole-exome sequenced primary breast tumors. 96, 120 bp biotinylated oligos were ordered as a 96-well plate at 8 reactions per oligo. When choosing which mutations to select, sites with A and T as the WT base were preferentially chosen over G or C sites with the exception of those mutations which were in the COSMIC database, as these mutations were of significant interest.

DIDA Library Creation

DIDA libraries were created using the Kapa Biosystems Hyper Prep kit (https://www.kapabiosystems.com/). At least 30 ng of cfDNA was used as input. Ligation occurred using a 200:1 adapter:template ratio for 16 hours at 16°C ensure a high efficiency ligation. PCR was conducted to create a 1ug library (typically 8-10 cycles). Library concentration and size was determined using the Agilent Bioanlyzer 2100 high sensitivity kit. 250 ng of the library was then combined with 250 ng from a different sample and input into the duplexed hybrid capture, allowing the remaining 750 ng to be used for subsequent hybrid captures. Hybrid capture was conducted using the IDT Hybridization and Wash kit (https://www.idtdna.com/pages/products/nextgen/target-capture/hybridization-and-wash-kit). As a cost saving measure, custom blocking oligos were ordered from IDT, inosines were placed opposite variable portions of the adapter.

i5 Blocking Oligo:

AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTiiiiiiiiiiiiiGTGTAGATCTCGGTGGTCGCCGTATCATT

i7 Blocking Oligo

CAAGCAGAAGACGGCATACGAGATiiiiiiiiiiiiiGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC

Following the first 4 hour hybridization and capture, libraries were amplified for 12 cycles and purified. The library was hybridized and captured a second time, and amplified for an additional 13 cycles. Library size was determined using the Agilent Bioanalyzer 2100 high sensitivity kit and

concentration was determined using the Kapa Biosystems Library Quantification Kit.

Sequencing

Samples were sequenced on either the Illumina HiSeq 2500, paired-end 100 bp plus 14 bp X2 indexing cycles (high capacity, rapid run mode), or Illumina NextSeq 500, paired-end 75 bp plus 14 bp X2 indexing cycles (high capacity, 150 cycle kit).

DIDA Bioinformatics Analysis Pipeline

The pipeline for analyzing DIDA data was based on the duplex sequencing pipeline developed in the Loeb lab at the University of Washington.[34] Substantial modification was required to allow it to work with our data. In brief, indexing reads (containing sample index and degenerate barcode) are prepended to each of the paired end reads. Migec checkout was used to demultiplex the samples.[101] Then a modified version of the duplex sequencing pipeline is used which:

1. Aligns reads using BWA mem
2. Groups reads by barcode, collapsing into SSCSs (requiring at least 3 reads and 90% sequence agreement)
3. SSCSs are then aligned again with bwa mem
4. SSCSs are locally realigned using GATK

5. bases from either end of the reads (containing lower quality sequences) are trimmed and replaced with N's

6. Overlapping paired end reads are clipped (to prevent double couting) using bamUtil clipOverlap (http://genome.sph.umich.edu/wiki/BamUtil:_clipOverlap)

7. Duplicates are removed using bamUtil dedup (http://genome.sph.umich.edu/wiki/BamUtil:_dedup)

8. Basecalls are made using Samtools mpileup (version 1.2), and variant allele frequencies are calculated using a custom perl script

9. On target percentages are identified using Picard Tools CalculateHsMetrics (https://broadinstitute.github.io/picard/command-line-overview.html).

Tag Swap Filter

The tag swap filter was written as a python script utilizing pysam to identify reads mapping to the exact same position, and the Distance 0.1.3 package to identify identical barcodes to be filtered.

Chapter 3:

Measuring circulating-tumor DNA Dynamics in Neoadjuvantly Treated Breast Cancer

Timothy M. Butler, Katie Johnson-Camacho, Christopher Boniface, Daira Melendez, Shaadi Tabatabaei, Paul T. Spellman

Introduction

Approximately 1 in 8 women will receive a breast cancer diagnosis in their lifetime.[102] Breast cancer can be divided into three major subtypes defined by the overexpression of estrogen receptor (ER+) and human epidermal growth factor 2-neu (HER2+) or their absence (triple-negative breast cancer, TNBC). These subtypes also correlate with gene expression signatures and prognosis: ER+ having good prognosis and HER2+ and TNBC having worse prognosis.[103,104] Breast cancer is typically treated with chemotherapy combined with surgery and, if appropriate, an agent targeting the estrogen or HER2 receptors. This regimen results in 70% of patients remaining disease-free at five years across all subtypes.[105]

Neoadjuvant (before surgery) chemotherapy has become an increasingly common treatment in breast cancer.[106] The first study testing this treatment approach showed that while patient outcomes were nearly identical, patients given neoadjuvant chemo were more likely to receive a less aggressive, breast-conservation surgery. They were also less likely to have evidence of disease in the axillary lymph nodes.[107] In addition, patients undergoing neoadjuvant treatment could be assessed for pathological complete response (pCR), the complete absence of disease following treatment. pCR is an early prognostic marker, as those who achieve pCR have significantly increased disease-free and overall survival.[108] pCR rates are not uniform across subtypes, being both more common and having more prognostic value in HER2+ and TNBC

disease.[104,108,109] Neoadjuvant chemotherapy is typically divided into two separate treatments given sequentially: the first treatment referred to as AC, which uses the DNA intercalator doxorubicin and the DNA crosslinker cyclophosphamide; the second treatment uses the microtubule inhibitor paclitaxel (Taxol). These drugs are typically administered every two weeks for 3-6 months. Taxol is typically a better tolerated therapy than AC, so in drug trials adding new agents to neoadjuvant chemotherapy, the investigative drug is typically combined with Taxol and done prior to the AC arm.[110,111]

Most patients receiving neoadjuvant chemotherapy have some response during the course of their treatment; however, a small subset shows no response (6%) or progression (3%).[112] These patients could possibly benefit from stopping treatment and moving straight to surgery. Studies have demonstrated a positive correlation between disease burden and ctDNA level. However, these studies have only analyzed this relationship during metastatic disease[32,40,59,113] or by comparing pre- and post-treatment ctDNA and associating it with treatment outcome.[58] A potential use case for ctDNA analysis is near real-time monitoring of a patient's response to treatment. This would allow for rapid feedback as to whether a given therapy is working, thereby allowing for ineffective therapies to be stopped. In addition, tracking multiple mutations of interest through ctDNA could allow for an understanding of whether certain mutations are being selected for or against during treatment.

In this study, we set out to measure ctDNA levels before, during, and after neoadjuvant treatment to determine whether early predictors of response could be seen. We find that there is a dramatic reduction in ctDNA level during the course of treatment in patients with and without pCR, possibly due to the timing of our ctDNA analysis. In the one patient who had treatment stopped due to progression, we find a consistent increase in ctDNA over the course of treatment, which was seen earlier than progression was seen clinically. This same patient showed residual ctDNA following treatment and adjuvant chemotherapy, preceding clinical detection of recurrence by over 7 months. We therefore conclude that our sampling strategy may be insufficient to reliably predict pCR, but shows promise in early prediction of progression.

Results

Study Design

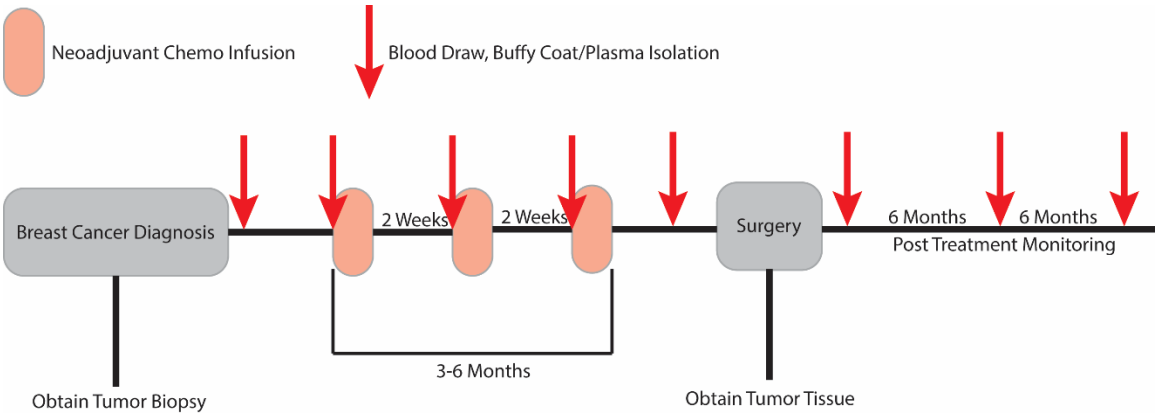


Figure 21. Plasma and tissue sampling strategy. Overview of study design collecting tumor tissue and plasma samples before, during, and after neoadjuvant chemotherapy. On-treatment plasma samples were taken in the infusion clinic before administration of drug.

This study is designed to quantify ctDNA taken before, during, and after neoadjuvantly treated breast cancer, in an effort to correlate ctDNA dynamics with treatment outcomes (Fig 21). To accomplish this, patients slated for neoadjuvant chemotherapy were identified and consented as part of an IRB approved study which allowed us to obtain samples of their tumor tissue, ability to get blood draws, and access to their medical records. A 30 ml blood draw was collected before the start of neoadjuvant chemotherapy, prior to the start of each infusion appointment, prior to surgery, following surgery, and at approximately six month intervals following surgery during routine follow up appointments. Each of these blood draws were separated into buffy coat and plasma sample and stored for further processing. In addition to blood draws, we also collected sample(s) of

the patient's tumor. Tumor samples were collected as research specific biopsies before treatment (7 patients) or sections from archived diagnostic biopsies (1 patient), for two patients we also collected surgical tissue. The tumor tissue was used to conduct whole-exome sequencing to generate a list of tumor-specific mutations, 10-20 of which were ordered as part of a hybrid capture panel for ctDNA analysis. The serially collected plasma samples was analyzed to quantify the ctDNA relying on these tumor-specific mutations using the DIDA high accuracy sequencing method. A subset of the samples were also analyzed using PCR-based high accuracy sequencing method called SafeSeqS, with which we only assayed one mutation at a time. We then compared the ctDNA frequencies with the patient's clinical data to see if there was a correlation with treatment outcome.

As part of this study we consented 18 patients. We were unable to get tumor tissue released for 5 of the patients, treatment has not yet been completed for 3 patients, and hybrid capture panels had not yet been ordered for 3 of the patients. The remaining 7 patients analyzed in detail represented all 3 major breast cancer subtypes with 2 patients having achieved pCR (Table 7).

Patient ID	Tumor Grade	Path CR	Pre-Treatment Dimensions	Post-Treatment Dimensions	Node Positive
ER-1	3	No	2.2 X 2.0	0.8 X 0.5	No
ER-2	1	No	2.7 X 2.3	4.5 X 4.0	Yes
ER+HER2-1	2	No	5.8 X 5.7	4.3 X 2.1	Yes
ER+HER2-2	3	No	2.2 X 2.0	3.4 X 2.9	Yes
Her2-1	3	Yes	2.2 X 1.8	0	No
Triple Neg-1	2	Yes	5.3 X 4.5	0	No
Triple Neg-2	3	No	2.7 X 2.3	5.0 X 2.5	Yes

Table 7. Enrolled patient characteristics. Patient ID's are described by their breast cancer subtype. ER+HER2 are patients positive for both ER and HER2. Pre- and post-treatment dimensions are the largest two dimensions of the tumor by MRI. Node positive refers to presence of lymph node metastases from the surgical specimen.

cfDNA Concentration Increases During Treatment

The enrolled patients were treated using three different regimen: standard neoadjuvant regiment of AC followed by Taxol (2), patients that were part of the ISPY2-TRIAL which starts with the Taxol treatment combined with and investigational drug followed by AC (3), and patients who underwent an anti-HER2 treatment followed by AC (2). We first sought to determine if there were any differences in the cfDNA concentrations under different treatment conditions (Fig 22). We found that the cfDNA concentration was significantly elevated in the plasma for both AC (average of 4-fold) and Taxol (average of 3-fold) treatments. It is possible that the increased concentration is the result of cell death (of both normal and tumor cell) resulting from these therapies. This additional cfDNA gave us improved yields from our plasma collection. However, if the additional cfDNA was primarily from healthy cells, it would make detection of ctDNA more difficult by diluting the signal. To overcome some of the difficulties of varying cfDNA concentration, we opted to calculated ctDNA detection as mutant

genomes per ml of plasma, rather than simply the mutant allele frequency detected. This was accomplished by simply correcting the allele frequency by the cfDNA concentration.

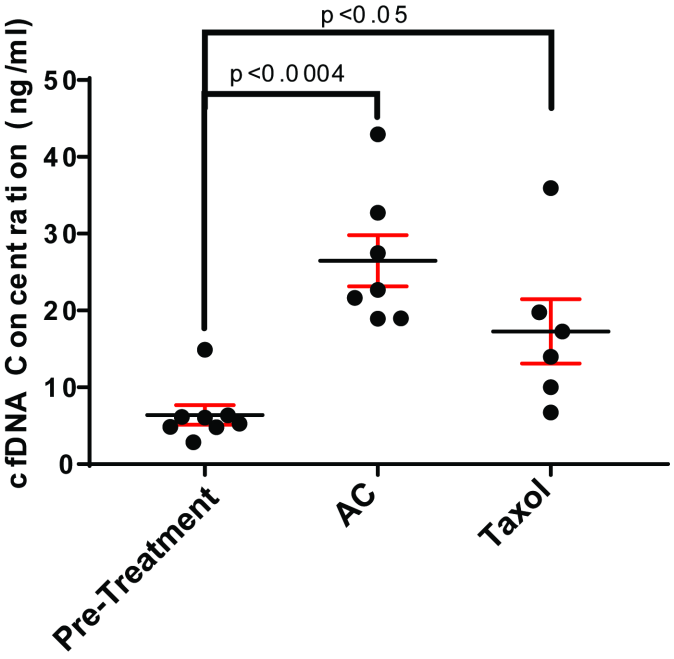


Figure 22. cfDNA Concentration increases during neoadjuvant chemotherapy. Each point represents the average cfDNA concentration for all plasma samples from a single patient under a single treatment condition. Each of the 8 patients had at least one pre-treatment and one of AC or Taxol as part of treatment, 4 had both. Error bars +/- SEM.

Pre-Treatment ctDNA concentration

Using the patient-specific DIDA panels, pre-treatment ctDNA was quantified in each of the patients (Fig 23). ctDNA was detectable in each of the samples at average allele percentages between 0.016-0.76%. Our 100% detection rate is somewhat higher than the 50% previously reported in localized breast cancer, and may highlight the advantages of using a larger panel over

tracking a single mutation.[17] The comparison between ctDNA allele percentage (Fig 23A) and mutant genome per ml (Fig 23B) shows the impact of adjusting the allele frequency by cfDNA concentration, creating less variation between the samples. Patient ER-1 had only a single mutant read detected of the >6,000 combined reads across the panel; it is possible this is a false positive. Even if this is a true mutation, the extremely low frequency makes detection difficult and seeing a dynamic range nearly impossible. In fact, all subsequent plasma samples had no detectable ctDNA (data not shown). The remaining 6 patients all had multiple mutant reads detected across multiple different tumor-specific mutations, ranging from 8-15 detected mutations. Detection of ctDNA in these pre-treatment samples was used as a positive control for the mutations in the panel, mutations not seen in the pre-treatment sample, or present in a subsequent sample, were excluded from analysis. The mutations failing the positive control filter are possibly false-positive mutations which were never present in the tumor, or belong to tumor clones not readily shedding their DNA into the plasma.

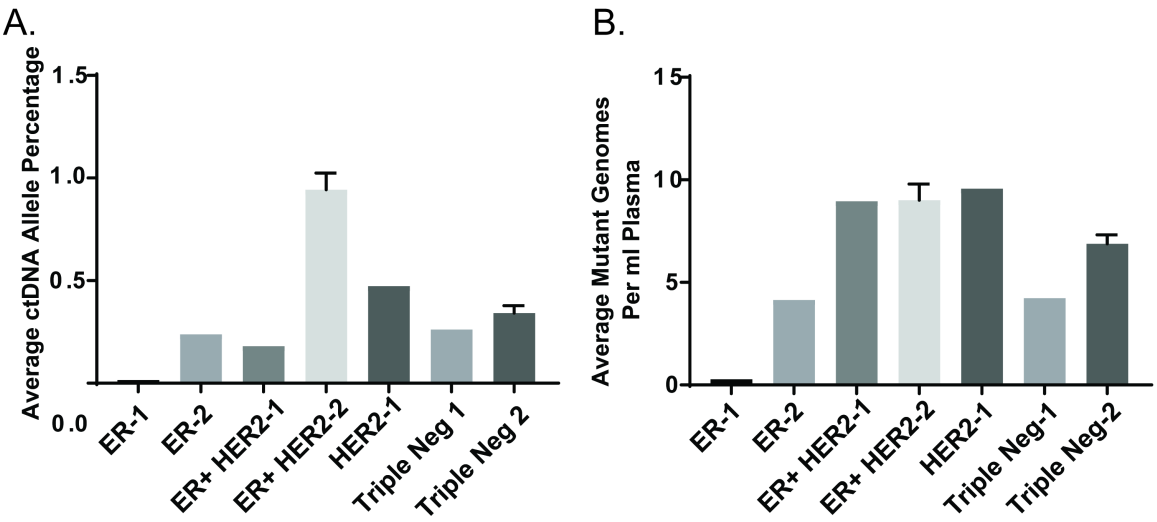


Figure 23. Average pre-treatment ctDNA levels. Average ctDNA percentage (A) or mutant genomes per ml (B) from DIDA sequencing panels. Sample Triple Neg-2 had two pre-treatment plasma samples on different days.

ctDNA Dynamics in pCR Patients

For the first pCR patient, Triple Neg-1, was part of the ISPY-2 TRIAL and received Ganiumab in addition to Taxol as the first treatment. We utilized a single gene SafeSeqS assay targeting a mutation in RTN2 (serial DIDA results are still planned). Unfortunately, this particular mutation was not successfully captured by the DIDA panel used on Triple Neg-1's pre-treatment sample, so a direct comparison is not possible. Comparing the two methods, the DIDA panel yielded a slightly higher mutant genome measurement than SafeSeqS (4.2 vs 2.3). It is possible this difference is due to sampling multiple mutations in the DIDA panel, different biases in the techniques, sampling noise, or some combination of all three. Tracking the ctDNA measurement over time shows a dramatic decrease in ctDNA level at the start of treatment which quickly becomes, and remains, undetectable (Fig 24). This is the kind of pattern we expected from a patient with

pCR, at some point during treatment the tumor completely disappears and this is reflected in the absence of ctDNA. The reduction that occurs at day 20 is just above the limit of detection of this assay, so it is not possible to know for sure whether ctDNA at subsequent time points is just below the limit of detection or completely absent.

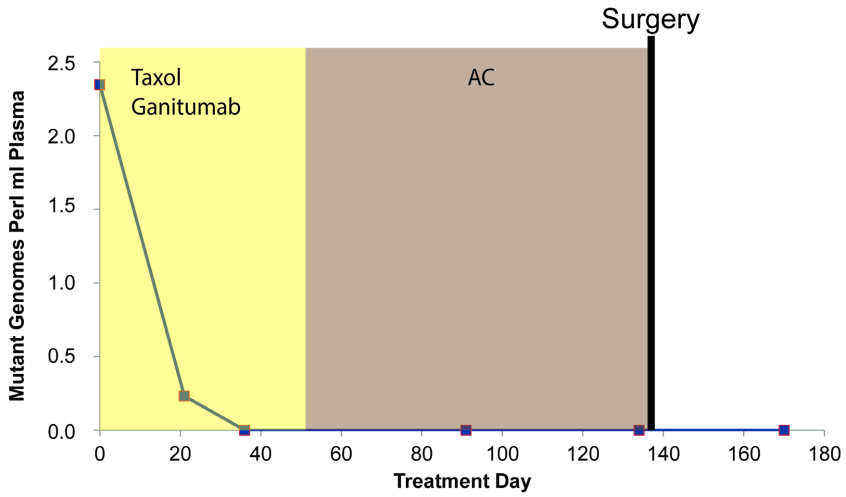


Figure 24. Triple Neg-1 ctDNA dynamics. ctDNA as measured by RTN2 SafeSeqS assay. Following day 20 time point, ctDNA remained undetectable. Ganitumab is a monoclonal antibody targeting IGF-1R.

For the second pCR patient, Her2-1, we were able to deploy a 19 gene DIDA panel, with the 15 mutations generating less than 1 error in 10k reads included in our analysis (Table 8). This patient showed a similar reduction during the first therapy arm (targeting HER2), but also had detectable ctDNA during the AC treatment (Fig 25). Again, the ctDNA levels were near the limit of detection of this assay (10k total depth across the panel). Of note, despite having similar ctDNA allele percentages, the day 23 and 109 time points yielded significantly

different mutant genome calculations (Fig 5B). This is due to the over 4-fold increase in cfDNA concentration which occurred between the first and second treatments. In contrast to Triple Neg-1, ctDNA was seen during the second treatment, potentially indicating residual tumor which was eventually eliminated by the AC treatment.

From the two pCR patients we saw a ctDNA reduction following start of treatment, leading to eventual disappearance. ctDNA remained undetectable following surgery, and neither patient has shown any clinical recurrence. The presence of ctDNA in the AC arm of Her2-1 may indicate that the AC treatment was responsible for the pCR, further exploration of identically treated patients is necessary to expand on these results.

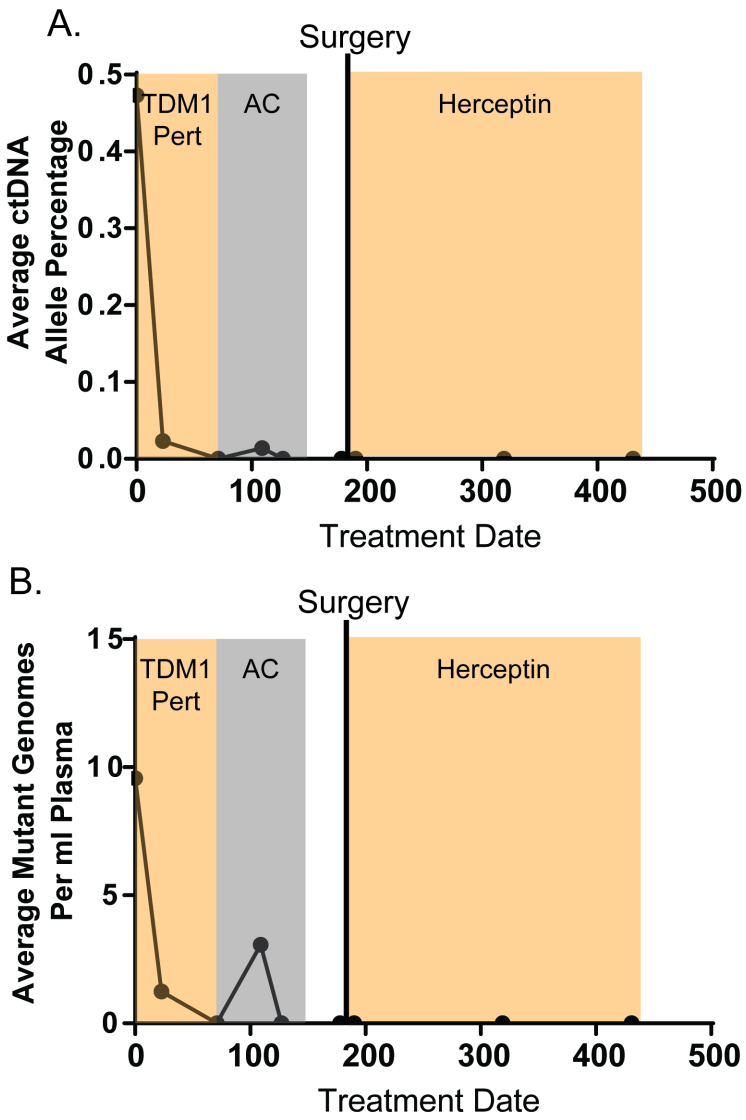


Figure 25. Her2-1 ctDNA dynamics. A) ctDNA allele percentage across serially collected plasma samples, average calculated from 15, high accuracy mutations. B) Mutant genomes per ml plasma calculation, correcting percentage calculations by the cfDNA concentration at each time point. Pert is abbreviation for the anti-Her2 monoclonal antibody Pertuzamab.

Chr	Position	Ref Base	Mut Base	Gene	Errors Per 100k Reads
4	68606283	G	C	GNRHR	0.00
4	68606341	G	C	GNRHR	0.00
4	1.64E+08	G	C	NPY5R	0.00
5	1.15E+08	G	C	CCDC112	0.00
12	1.18E+08	G	C	NOS1	0.00
18	10689765	G	T	FAM38B	0.00
20	62421387	C	G	ZBTB46	0.00
1	2.05E+08	C	T	CNTN2	2.00
6	26199793	G	A	HIST1H2BF	2.75
7	5541095	C	T	FBXL18	2.76
17	34854330	G	A	MYO19	5.76
18	58039394	G	T	MC4R	8.27
X	70389754	G	A	NLGN3	8.41
11	20177994	G	A	DBX1	9.06
6	31838426	G	A	SLC44A4	18.83
19	50365659	G	A	PNKP	20.15
X	1.06E+08	C	A	RNF128	21.10
17	7577550	C	T	TP53	23.90
19	44792318	T	A	ZNF235	416.59

Table 8. Her2-1, 19 gene DIDA panel. Only mutations with less than 10 errors per 100k reads were included as part of the analysis.

ctDNA Dynamics in non-pCR Patients

Patient ER-2 did not achieve pCR, in fact the pre-surgery MRI revealed a larger tumor than at the start of treatment, likely indicating progression during treatment (Table 7). Utilizing a 22-gene DIDA panel (Table 9) we were able to see evidence for this increase reflected in the ctDNA (Fig 26). After an initial decrease and disappearance during AC, the ctDNA increases over the course of the Taxol time points. This potentially shows that while AC was effective in reducing the ctDNA level (and possibly shrinking the tumor), Taxol was ineffective and possibly even permitted tumor growth. ctDNA remained detectable 11 days following surgery, indicating a potential increased risk of recurrence. However, following a 50 Gray chest irradiation, ctDNA was

undetectable and remained undetectable a year later. It is possible that the radiation treatment removed any residual tumor cells contributing detectable ctDNA. Surprisingly, we saw a dramatic decrease in ctDNA between the first and second pre-treatment ctDNA samples. It is unlikely this reflects any change in the tumor size, and we do not have enough examples of multiple pre-treatment ctDNA time points to know if this kind of variation is typical.

Chr	Position	Ref Base	Mut Base	Gene	Errors Per 100k Reads
1	157494285	G	T	FCRL5	0.00
2	202136322	A	G	CASP8	0.00
15	77025665	A	T	SCAPER	0.00
20	61299857	G	A	SLCO4A1	0.00
2	74425738	T	A	MTHFD2	0.00
5	95103456	A	G	RHOBTB3	0.00
8	29194273	A	C	DUSP4	0.00
14	63841230	T	C	PPP2R5E	0.00
18	29057302	T	C	DSG3	0.00
17	33768040	G	A	SLFN13	0.72
12	122702878	T	C	DIABLO	1.30
8	98943465	G	A	MATN2	1.32
1	228444416	C	T	OBSCN	1.72
19	44097516	G	A	IRGQ	1.73
11	65161973	T	G	FRMD8	2.55
17	71189512	C	G	COG1	2.64
18	19154544	C	T	ESCO1	2.73
8	52733196	C	A	PCMTD1	3.48
14	77242554	C	T	VASH1	4.25
17	7577121	G	A	TP53	5.58
18	48255582	C	T	MAPK4	10.45
1	152282617	G	A	FLG	33.53

Table 9. ER-2, 22 gene DIDA panel.

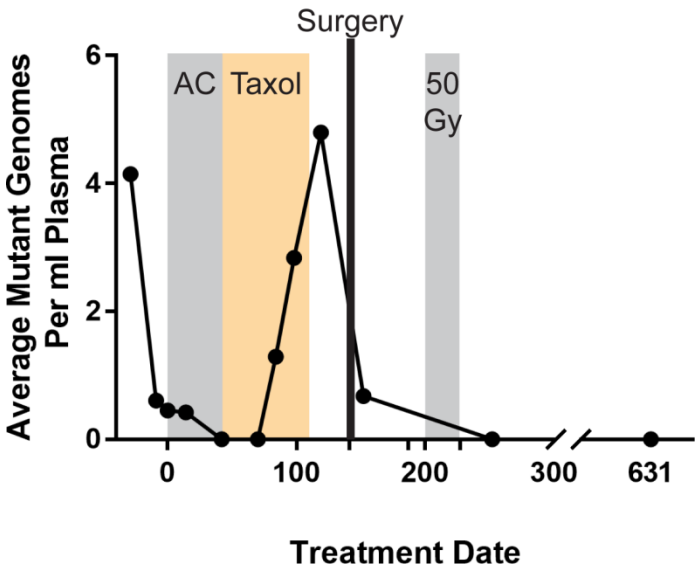


Figure 26. ER-2 ctDNA dynamics. 50 Gray of radiation were administered in the post-surgery setting.

Two additional non-pCR patients, ER+HER2-1 and ER+HER2-2 did not show a consistent increase in ctDNA during treatment, but had ctDNA stochastically detected throughout treatment (Fig 27). ER+HER2-2 had an initial increase in ctDNA level, followed by a decrease to just at the limit of detection of the assay. ctDNA was detectable pre-surgery as well as two days following surgery (Fig 27A). This post-surgery ctDNA may represent residual tumor cells, or simply be leftover ctDNA as has been seen in another study looking at ctDNA shortly after surgery.[30] The day 484 time point is based on a single mutant read and may represent a false positive, or indicate a small amount of residual disease. ER+HER2-1 had a relatively poor performing panel, with only 6 mutations passing our accuracy filter, giving us an average total panel depth of only 2,000-4,000X, a 2-5 fold reduction in sensitivity compared to the other patients. Despite this, there were three on-treatment time points with detectable

ctDNA (Fig 27B). Following surgery, ctDNA was undetectable and remained so.

Neither ER+HER-1 nor -2 have recurred.

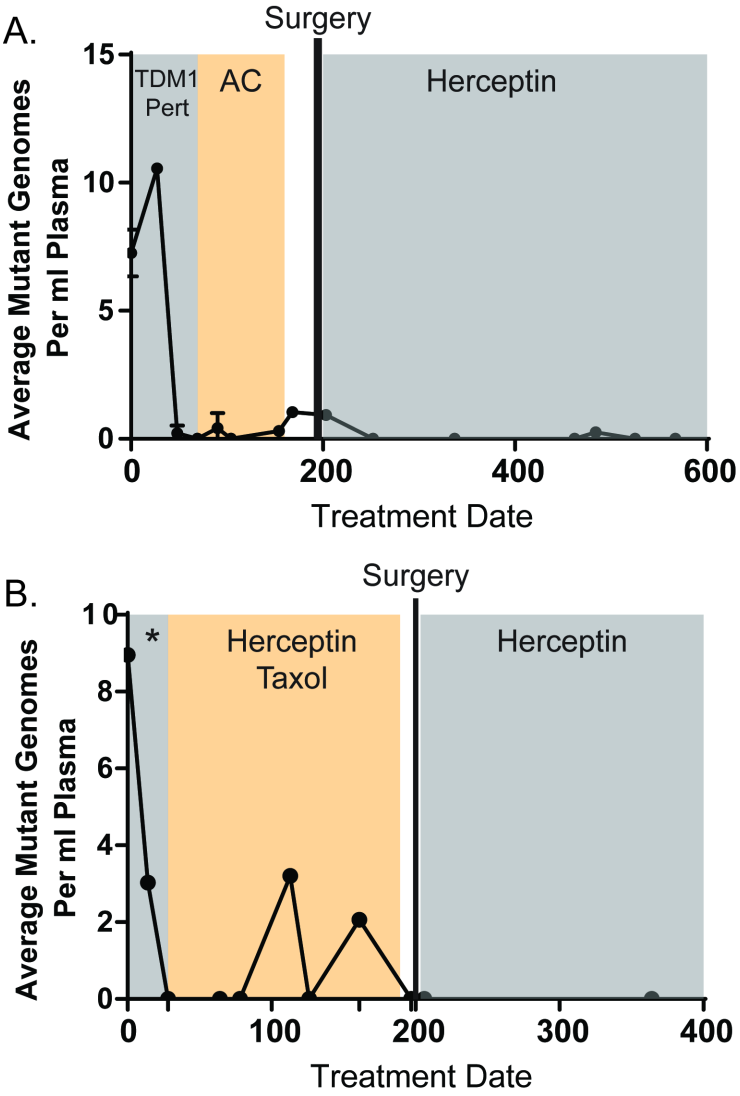


Figure 27. ER+HER2-2 and ER+HER2-1 ctDNA dynamics. A) Patient ER+HER2-2. Day 200, post-surgery time point was collected 2 days following surgery. Error bars +/- SEM B) Patient ER+HER2-1. *Patient treated with Taxol, Herceptin, and MK2206 (*AKT* inhibitor), MK2206 treatment stopped due to toxicity issues.

Chr	Position	Ref Base	Mut Base	Gene	Errors Per 100k Reads
2	85628387	G	C	CAPG	0
10	37425504	G	T	ANKRD30A	0
10	127414256	A	G	C10orf137	0
22	43924660	C	G	EFCAB6	0
22	43926736	C	G	EFCAB6	0
17	7577141	C	A	TP53	4.034698
16	23716446	G	A	ERN2	16.99428
4	72101926	G	C	SLC4A4	123.8673
10	46179929	C	G	IGR	255.102
12	87475	T	C	IGR	580.6755
13	114058885	G	T	IGR	753.1494
X	119572319	T	A	LAMP2	880.8885
1	121478645	T	C	IGR	1779.256
9	39078814	C	A	CNTNAP3	2329.322
2	31414288	C	A	CAPN14	2909.179
19	56284464	A	G	RFPL4AL	3548.6
14	20098038	A	G	RP11	4919.302
16	32487123	T	C	IGR	14611.74

Table 10. ER+HER2-1, 18 gene DIDA panel

Chr	Position	Ref Base	Mut Base	Gene	Errors Per 100k Reads
1	86171847	C	G	ZNHIT6	0.00
2	80772064	C	A	CTNNA2	0.00
5	16179191	T	G	MARCH11	0.00
8	1.01E+08	C	T	RGS22	0.00
9	1.37E+08	C	G	BRD3	0.00
11	6190651	A	G	OR52B2	0.00
11	6977691	T	G	ZNF215	0.00
12	1.21E+08	G	C	CCDC64	0.00
15	33442736	C	G	FMN1	0.00
17	7577114	C	T	TP53	0.00
17	8079184	A	C	TMEM107	0.00
20	31659933	C	T	BPIFB3	3.40
16	72831707	C	T	ZFHX3	26.48
13	1.13E+08	C	T	SOX1	37.68
16	15178612	C	T	RRN3	1566.48

Table 11. Triple Neg-2, 15 gene DIDA panel.

Detection of Minimal Residual Disease

Our final patient, Triple Neg-2, stopped neoadjuvant treatment early due to discovery of metastasis in the lymph nodes, indicating disease progression on treatment. Using our 15 gene DIDA panel (Table 11), ctDNA was seen to increase throughout the AC treatment, ending up higher than the pre-treatment samples (Fig 28). Following surgery, ctDNA was barely detectable (only 2 mutant reads seen) and increased slightly following adjuvant Taxol treatment (4 mutant reads). The detectable post-treatment ctDNA indicated the patient was at increased risk for recurrence, and in fact 7 months later was diagnosed with bone metastases. A plasma sample taken at a follow up appointment two weeks prior to diagnosis, showed a dramatic increase in ctDNA level as expected with metastatic disease. The metastatic ctDNA had 14 of the 15 panel mutations present. The high ctDNA allele percentage of 15% makes it possible to conduct whole-exome sequencing on this time point to look for any metastasis-specific mutations. In contrast to ER-2, the two pre-treatment plasma samples showed very similar amounts of ctDNA.

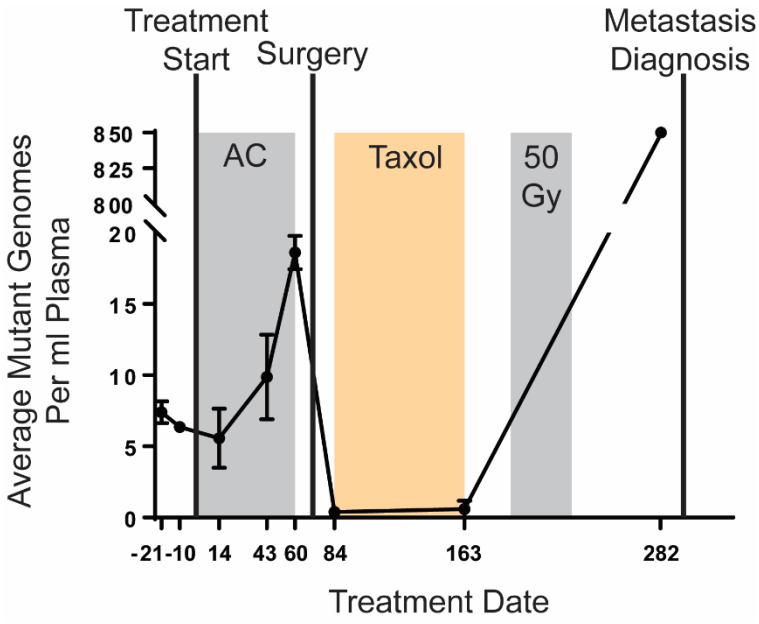


Figure 28. Triple Neg-3 ctDNA dynamics. Error bars +/- SEM, time points -10 and 84 have error bars too small to display.

Discussion

In this study we demonstrated the ability to design patient-specific hybrid capture panels to sensitively detect ctDNA before, during, and after neoadjuvant chemotherapy. We also demonstrated a significant increase in total cfDNA during either Taxol or AC treatment, an observation that may serve as a proxy for increased death of normal cells.

In every patient, except Triple Neg-2, we saw a dramatic decrease in ctDNA level at some point during neoadjuvant treatment. This decrease was near the limit of sensitivity of the assay and likely meant that several of the undetectable time points had ctDNA present below the limit of detection rather than not present at all. It also meant that ctDNA detection for many of these time points was based on detection of fewer than 5 mutant reads, making estimates of ctDNA level quite noisy. Improving library efficiency and therefore sequencing depth and/or inputting more cfDNA into the assay may help address some of these detection limit issues.

The two pCR patients appeared to have fewer detectable on-treatment ctDNA time points than the non-pCR patients, potentially indicating a relationship between pCR and ctDNA. However, the similar ctDNA reduction also seen in ER-2, and ER+Her2-1 and -2 makes differentiating the ctDNA response between pCR and non-pCR difficult, especially based only on the early time points. With only two pCR patients it is impossible to know for certain whether ctDNA can predict pCR, but these results warrant further study.

In two patients, Triple Neg-2 and ER-2, we saw increasing ctDNA levels across three consecutive treatment time points, which corresponded to increased tumor size. In Triple Neg-2 progression was identified clinically and the treatment plan was altered. Taken together these results indicate a role for ctDNA analysis in identifying tumor progression and stopping ineffective therapies. It might also be possible to track ctDNA in identically treated individuals to see if different arms of neoadjuvant therapy are more or less effective with different tumor subtypes. In both these patients, we also detected ctDNA following surgery, a result which has been shown to indicate a risk of recurrence.[65] In Triple Neg-2 this recurrence occurred 7 months following the ctDNA detection. Detection of ctDNA before and after adjuvant Taxol and radiation likely indicated those treatments were ineffective. In contrast ER-2 had their post-surgery ctDNA disappear following radiation treatment potentially demonstrating its effectiveness. This highlights a potential role for post-surgery ctDNA analysis in prediction of recurrence and assessment of adjuvant treatment effectiveness. One could imagine a scenario of using multiple adjuvant treatments until ctDNA became undetectable. Similarly, an adjuvant treatment could be avoided if ctDNA was never detected post-surgery.

The plasma sampling strategy we used collected blood prior to chemotherapy infusion. This strategy was chosen for simplicity as the patients were already in the clinic. However, it is possible the timing of collection, two weeks after the previous drug administration, hurt our ability to detect ctDNA. A sampling schedule that collected blood two or three days after infusion might

have detected ctDNA from actively dying cells. However, this signal may no longer be present after two weeks. Additional studies looking at the timing of ctDNA release following drug administration could provide valuable insight into how to best detect on-treatment ctDNA, and could measure the kinetics of tumor cell death.

These results highlight the potential of neoadjuvant ctDNA analysis to identify pCR, tumor progression, measure adjuvant treatment effectiveness, and predict recurrence. Additional work is needed to expand on and replicate these findings, and develop standardized methods of assaying ctDNA. It is possible that similarly promising results may be found in additional tumor types or treatment regimens, opening the door for ctDNA analysis to contribute to patient treatment decisions and improving clinical outcomes.

Methods

Patient Enrollment and Sample Collection

Written consent was obtained from patients as part of two studies approved by Oregon Health & Science University's Institutional Review board: Breast Cancer Registry (IRB# 8314) or Tumor in Blood (IRB# 10163). Up to 30 mls of blood were collected in 5, 6 ml purple-capped EDTA tubes. Plasma was isolated by first spinning at 1,000 g for 10 mins, separating the top plasma layer into 1 ml aliquots, then spinning those aliquots at 15,000 g for 10 mins, transferring the supernatant to cryovials and storing at -80°C. Buffy coat was isolated from the intermediate blood layer following the first spin, and also stored at -80°C. cfDNA was extracted using the QIAamp Circulating Nucleic Acid kit (Qiagen). Buffy coat DNA was extracted using the DNA Blood Mini kit (Qiagen).

Tumor tissue was obtained from a core needle biopsy of the primary tumor, which was placed in OCT and stored at -80°C. Prior to extraction the OCT block was sent out for sectioning and path review. DNA was extracted using the DNeasy blood and tissue kit (Qiagen). For 1 patient we received 10 um sliced of an archived FFPE diagnostic biopsy, DNA was extracted using the QIAamp DNA FFPE tissue kit (Qiagen).

DNA was quantified using the Kapa hgDNA Quantification and QC kit (Kapa Biosystems).

Whole-Exome Sequencing

Whole exome sequencing was conducted using HPLC purified, dual-index adapters ordered from IDT (idtdna.com). These were used in combination with the Hyper Prep DNA Library Preparation Kit (Kapa Biosystems), at a 10:1 adapter:template ratio. This library was then used as input for the Agilent SureSelect XT hybrid-capture protocol and reagents using the all-human exon v5 set of capture baits. These libraries were either paired-end 100 bp sequenced using the Illumina HiSeq 2500 platform, or paired-end 75 bp sequenced using the Illumina NextSeq 500 platform.

DIDA Library Preparation and Sequencing

30-50 ng of cfDNA were used as input for creating DIDA libraries as outlined in the “Development of a high-accuracy hybrid-capture based ctDNA detection method” section of this thesis. Double-hybrid capture was conducted on 96-oligo hybrid capture panels designed to capture the mutations identified from whole exome sequencing. These panels were combinations of 4-5 patient-specific mutation sets. These libraries were either paired-end 100 bp sequenced using the Illumina HiSeq 2500 platform, or paired-end 75 bp sequenced using the Illumina NextSeq 500 platform.

SafeSeqS Library Preparation and Sequencing

The SafeSeqS library was prepared as described previously.[31] Briefly, a pair of mutation-specific primers were ordered which in addition to 20 template-

specific bases, contained 12 degenerate N's and the first half of the Illumina Sequencing adapter. This was subjected to 4 cycles of PCR using Phusion Hotstart II polymerase. Following PCR, unused primers were removed using RecJf nuclease, and cleaned up using AMPureXP DNA binding beads. A second round of 30 PCR cycles was conducted using primers against the Illumina sequence and containing an Illumina sample index and the remainder of the adapter. This was then purified again with AMPureXP beads, and sequenced spiked-in to another library at <1% and sequence don the Illumina NextSeq 500.

Bioinformatics Analysis

Whole-exome sequencing data were analyzed as described previously.[53] Mutations were called using Mutect v 1.1.17, filtering to only include reads with a mapping quality of 20 or greater.

DIDA data was analyzed using the pipeline described in the "Development of a high-accuracy hybrid-capture based ctDNA detection method" section of this thesis. SafeSeqs libraries were analyzed using the same pipeline excepting the duplicate removal steps.

Summary and Conclusions

In this work we have demonstrated the ability to improve our understanding of a patient's tumor through the use of ctDNA sequencing. In metastatic disease, ctDNA levels above 2% allow for the use of whole-exome sequencing to identify the majority of the mutations present in matched tumor samples. This allows for the use of ctDNA sequencing as a "liquid biopsy," potentially eliminating the need for surgical biopsy tissue. This sequencing identified clinically actionable mutations in genes such as *PIK3CA*, *KRAS*, and *ESR1*. Comparison to the primary tumor provided information about the evolution of the tumor over time, and identified a potential resistance mechanism to aromatase inhibition.

To study primary disease we developed a high-accuracy, high-sensitivity DIDA assay to simultaneously identify dozens of patient-specific mutations. This assay consistently achieved accuracy of better than 1 error in 10k reads, allowing us to identify rare ctDNA mutations. The dual-indexing approach minimized the risk of cross-contamination or incorrect index assignment. In the process of analyzing the data we discovered a previously unreported source of error in the form of "tag swaps." Removing this error led to a more accurate estimate of the sequencing depth, and therefore sensitivity, of the assay.

Finally, we demonstrated that ctDNA can be detected and tracked before, during, and after neoadjuvantly treated breast cancer. The results suggested that tumor growth was reflected in a corresponding increase in the ctDNA, potentially

allowing for early detection of disease progression. Conversely, pCR patients showed a dramatic decrease in the ctDNA levels. These findings need to be expanded upon in follow up studies in order to determine whether neoadjuvant ctDNA analysis has the ability to predict treatment response, but these preliminary results are promising. In the post-surgery setting, two patients had detectable ctDNA, putting them at increased risk of recurrence. In one patient radiation treatment eliminated the ctDNA signal and the patient remains disease free over a year later. In the second patient, ctDNA remained detectable following adjuvant Taxol and radiation, leading to clinical recurrence 7 months after surgery. These results indicate a role of ctDNA in the post-surgery setting in assessing the effectiveness of adjuvant treatments.

Overall these studies advanced the ctDNA field by identifying a variety of use cases for ctDNA analysis. In the metastatic setting, whole-exome sequencing can paint a near complete picture of the tumor genome, identifying clinically relevant mutations, and allowing for inferences to be made about tumor evolution. In the neoadjuvant setting, ctDNA response may serve as an early indicator of response to treatment, allowing for ineffective therapies to be stopped. In the minimal residual disease setting, effectiveness of adjuvant therapies can be assessed through ctDNA tracking, potentially delaying or preventing recurrence by treating until ctDNA disappears. It is this last scenario which has the potential to most dramatically impact patient outcomes. However, a larger clinical trial is necessary to determine whether making treatment decisions based on ctDNA analysis can improve patient outcomes.

References

1 Sullivan, L. L. *Liquid Biopsy Market Gets Off to Big Start in 2016*, <http://blog.bccresearch.com/liquid-biopsy-market-gets-off-to-big-start-in-2016> (2016).

2 Mandel P Fau - Metais, P. & Metais, P. Les acides nucléiques du plasma sanguin chez l'homme. *C R Seances Soc Biol Fil*, doi:D - CLML: 4815:805a OTO - NLM (1948).

3 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).

4 Stroun, M., Anker, P., Lyautey, J., Lederrey, C. & Maurice, P. A. Isolation and characterization of DNA from the plasma of cancer patients. *Eur J Cancer Clin Oncol* **23**, 707-712 (1987).

5 Vasioukhin, V. *et al.* Point mutations of the N-ras gene in the blood plasma DNA of patients with myelodysplastic syndrome or acute myelogenous leukaemia. *British Journal of Haematology* **86**, 774-779, doi:10.1111/j.1365-2141.1994.tb04828.x (1994).

6 Mouliere, F. & Rosenfeld, N. Circulating tumor-derived DNA is shorter than somatic DNA in plasma. *Proc Natl Acad Sci U S A* **112**, 3178-3179, doi:10.1073/pnas.1501321112 (2015).

7 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, doi:10.1016/j.cell.2015.11.050 (2016).

8 Lo, Y. M. *et al.* Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Sci Transl Med* **2**, 61ra91, doi:10.1126/scitranslmed.3001720 (2010).

9 Ulz, P. *et al.* Inferring expressed genes by whole-genome sequencing of plasma DNA. *Nat Genet* **48**, 1273-1278, doi:10.1038/ng.3648 (2016).

10 Ivanov, M., Baranova, A., Butler, T., Spellman, P. & Mileyko, V. Non-random fragmentation patterns in circulating cell-free DNA reflect epigenetic regulation. *BMC Genomics* **16 Suppl 13**, S1, doi:10.1186/1471-2164-16-S13-S1 (2015).

11 Jung, K., Fleischhacker, M. & Rabien, A. Cell-free DNA in the blood as a solid tumor biomarker--a critical appraisal of the literature. *Clin Chim Acta* **411**, 1611-1624, doi:10.1016/j.cca.2010.07.032 (2010).

12 Schwarzenbach, H., Hoon, D. S. & Pantel, K. Cell-free nucleic acids as biomarkers in cancer patients. *Nat Rev Cancer* **11**, 426-437, doi:10.1038/nrc3066 (2011).

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 Jiang, P. *et al.* Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. *Proc Natl Acad Sci U S A* **112**, E1317-1325, doi:10.1073/pnas.1500076112 (2015).

15 Underhill, H. R. *et al.* Fragment Length of Circulating Tumor DNA. *PLoS Genet* **12**, e1006162, doi:10.1371/journal.pgen.1006162 (2016).

16 Diehl, F. *et al.* Detection and quantification of mutations in the plasma of patients with colorectal tumors. *Proc Natl Acad Sci U S A* **102**, 16368-16373, doi:10.1073/pnas.0507904102 (2005).

17 Bettegowda, C. *et al.* Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci Transl Med* **6**, 224ra224, doi:10.1126/scitranslmed.3007094 (2014).
18 Pokrywka, A. *et al.* The influence of hypoxic physical activity on cfDNA as a new marker of vascular inflammation. *Arch Med Sci* **11**, 1156-1163, doi:10.5114/aoms.2015.56341 (2015).
19 Xin, Y. *et al.* Circulating cell-free DNA indicates M1/M2 responses during septic peritonitis. *Biochem Biophys Res Commun* **477**, 589-594, doi:10.1016/j.bbrc.2016.06.092 (2016).
20 Macher, H. *et al.* Role of early cell-free DNA levels decrease as a predictive marker of fatal outcome after severe traumatic brain injury. *Clin Chim Acta* **414**, 12-17, doi:10.1016/j.cca.2012.08.001 (2012).
21 Tsai, N. W. *et al.* The value of serial plasma nuclear and mitochondrial DNA levels in patients with acute ischemic stroke. *Clin Chim Acta* **412**, 476-479, doi:10.1016/j.cca.2010.11.036 (2011).
22 Lo, Y. M. *et al.* Rapid clearance of fetal DNA from maternal plasma. *Am J Hum Genet* **64**, 218-224, doi:10.1086/302205 (1999).
23 Korabecna, M. *et al.* Cell-free plasma DNA during peritoneal dialysis and hemodialysis and in patients with chronic kidney disease. *Ann N Y Acad Sci* **1137**, 296-301, doi:10.1196/annals.1448.014 (2008).
24 Botezatu, I. *et al.* Genetic analysis of DNA excreted in urine: a new approach for detecting specific genomic DNA sequences from cells dying in an organism. *Clin Chem* **46**, 1078-1084 (2000).
25 Ross, M. G. *et al.* Characterizing and measuring bias in sequence data. *Genome Biol* **14**, R51, doi:10.1186/gb-2013-14-5-r51 (2013).
26 Yanez-Mo, M. *et al.* Biological properties of extracellular vesicles and their physiological functions. *J Extracell Vesicles* **4**, 27066, doi:10.3402/jev.v4.27066 (2015).
27 Minciacchi, V. R., Freeman, M. R. & Di Vizio, D. Extracellular vesicles in cancer: exosomes, microvesicles and the emerging role of large oncosomes. *Semin Cell Dev Biol* **40**, 41-51, doi:10.1016/j.semcdb.2015.02.010 (2015).
28 Milbury, C. A. *et al.* Determining lower limits of detection of digital PCR assays for cancer-related gene mutations. *Biomolecular Detection and Quantification* **1**, 8-22, doi:10.1016/j.bdq.2014.08.001 (2014).
29 Dressman, D., Yan, H., Traverso, G., Kinzler, K. W. & Vogelstein, B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc Natl Acad Sci U S A* **100**, 8817-8822, doi:10.1073/pnas.1133470100 (2003).
30 Diehl, F. *et al.* Circulating mutant DNA to assess tumor dynamics. *Nat Med* **14**, 985-990, doi:10.1038/nm.1789 (2008).
31 Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K. W. & Vogelstein, B. Detection and quantification of rare mutations with massively parallel sequencing. *Proc Natl Acad Sci U S A* **108**, 9530-9535, doi:10.1073/pnas.1105422108 (2011).
32 Tie, J. *et al.* Circulating tumor DNA as an early marker of therapeutic response in patients with metastatic colorectal cancer. *Ann Oncol* **26**, 1715-1722, doi:10.1093/annonc/mdv177 (2015).
33 Newman, A. M. *et al.* Integrated digital error suppression for improved detection of circulating tumor DNA. *Nat Biotechnol* **34**, 547-555, doi:10.1038/nbt.3520 (2016).

34 Kennedy, S. R. *et al.* Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat Protoc* **9**, 2586-2606, doi:10.1038/nprot.2014.170 (2014).
35 Overman, M. J. *et al.* Use of research biopsies in clinical trials: are risks and benefits adequately discussed? *J Clin Oncol* **31**, 17-22, doi:10.1200/JCO.2012.43.1718 (2013).
36 Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med* **366**, 883-892, doi:10.1056/NEJMoa1113205 (2012).
37 Xu, X. *et al.* Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell* **148**, 886-895, doi:10.1016/j.cell.2012.02.025 (2012).
38 Board, R. E. *et al.* Detection of PIK3CA mutations in circulating free DNA in patients with breast cancer. *Breast Cancer Res Treat* **120**, 461-467, doi:10.1007/s10549-010-0747-9 (2010).
39 Higgins, M. J. *et al.* Detection of tumor PIK3CA status in metastatic breast cancer using peripheral blood. *Clin Cancer Res* **18**, 3462-3469, doi:10.1158/1078-0432.CCR-11-2696 (2012).
40 Dawson, S. J. *et al.* Analysis of circulating tumor DNA to monitor metastatic breast cancer. *N Engl J Med* **368**, 1199-1209, doi:10.1056/NEJMoa1213261 (2013).
41 Forshew, T. *et al.* Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. *Sci Transl Med* **4**, 136ra168, doi:10.1126/scitranslmed.3003726 (2012).
42 Janku, F. *et al.* Actionable mutations in plasma cell-free DNA in patients with advanced cancers referred for experimental targeted therapies. *Oncotarget* **6**, 12809-12821, doi:10.18632/oncotarget.3373 (2015).
43 Waddell, N. *et al.* Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature* **518**, 495-501, doi:10.1038/nature14169 (2015).
44 Hadano, N. *et al.* Prognostic value of circulating tumour DNA in patients undergoing curative resection for pancreatic cancer. *Br J Cancer* **115**, 59-65, doi:10.1038/bjc.2016.175 (2016).
45 Sausen, M. *et al.* Clinical implications of genomic alterations in the tumour and circulation of pancreatic cancer patients. *Nat Commun* **6**, 7686, doi:10.1038/ncomms8686 (2015).
46 Santiago-Walker, A. *et al.* Correlation of BRAF Mutation Status in Circulating-Free DNA and Tumor and Association with Clinical Outcome across Four BRAFi and MEKi Clinical Trials. *Clin Cancer Res* **22**, 567-574, doi:10.1158/1078-0432.CCR-15-0321 (2016).
47 Ascierto, P. A. *et al.* Phase II trial (BREAK-2) of the BRAF inhibitor dabrafenib (GSK2118436) in patients with metastatic melanoma. *J Clin Oncol* **31**, 3205-3211, doi:10.1200/JCO.2013.49.8691 (2013).
48 Long, G. V. *et al.* Dabrafenib in patients with Val600Glu or Val600Lys BRAF-mutant melanoma metastatic to the brain (BREAK-MB): a multicentre, open-label, phase 2 trial. *Lancet Oncol* **13**, 1087-1095, doi:10.1016/S1470-2045(12)70431-X (2012).
49 Hauschild, A. *et al.* Dabrafenib in BRAF-mutated metastatic melanoma: a multicentre, open-label, phase 3 randomised controlled trial. *The Lancet* **380**, 358-365, doi:10.1016/s0140-6736(12)60868-x (2012).
50 Flaherty, K. T. *et al.* Improved survival with MEK inhibition in BRAF-mutated melanoma. *N Engl J Med* **367**, 107-114, doi:10.1056/NEJMoa1203421 (2012).

51 Schiavon, G. *et al.* Analysis of ESR1 mutation in circulating tumor DNA demonstrates evolution during therapy for metastatic breast cancer. *Sci Transl Med* **7**, 313ra182, doi:10.1126/scitranslmed.aac7551 (2015).
52 Sefrioui, D. *et al.* Short report: Monitoring ESR1 mutations by circulating tumor DNA in aromatase inhibitor resistant metastatic breast cancer. *Int J Cancer* **137**, 2513-2519, doi:10.1002/ijc.29612 (2015).
53 Butler, T. M. *et al.* Exome Sequencing of Cell-Free DNA from Metastatic Cancer Patients Identifies Clinically Actionable Mutations Distinct from Primary Disease. *PLoS One* **10**, e0136407, doi:10.1371/journal.pone.0136407 (2015).
54 Azad, A. A. *et al.* Androgen Receptor Gene Aberrations in Circulating Cell-Free DNA: Biomarkers of Therapeutic Resistance in Castration-Resistant Prostate Cancer. *Clin Cancer Res* **21**, 2315-2324, doi:10.1158/1078-0432.CCR-14-2666 (2015).
55 Lallous, N. *et al.* Functional analysis of androgen receptor mutations that confer anti-androgen resistance identified in circulating cell-free DNA from prostate cancer patients. *Genome Biol* **17**, 10, doi:10.1186/s13059-015-0864-1 (2016).
56 Leary, R. J., Sausen, M., Diaz, L. A., Jr. & Velculescu, V. E. Cancer detection using whole-genome sequencing of cell free DNA. *Oncotarget* **4**, 1119-1120, doi:10.18632/oncotarget.1183 (2013).
57 Leary, R. J. *et al.* Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing. *Sci Transl Med* **4**, 162ra154, doi:10.1126/scitranslmed.3004742 (2012).
58 Leary, R. J. *et al.* Development of personalized tumor biomarkers using massively parallel sequencing. *Sci Transl Med* **2**, 20ra14, doi:10.1126/scitranslmed.3000702 (2010).
59 Murtaza, M. *et al.* Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. *Nature* **497**, 108-112, doi:10.1038/nature12065 (2013).
60 Murtaza, M. *et al.* Multifocal clonal evolution characterized using circulating tumour DNA in a case of metastatic breast cancer. *Nat Commun* **6**, 8760, doi:10.1038/ncomms9760 (2015).
61 Lipson, E. J. *et al.* Circulating tumor DNA analysis as a real-time method for monitoring tumor burden in melanoma patients undergoing treatment with immune checkpoint blockade. *J Immunother Cancer* **2**, 42, doi:10.1186/s40425-014-0042-0 (2014).
62 Pereira, E. *et al.* Personalized Circulating Tumor DNA Biomarkers Dynamically Predict Treatment Response and Survival In Gynecologic Cancers. *PLoS One* **10**, e0145754, doi:10.1371/journal.pone.0145754 (2015).
63 Frenel, J. S. *et al.* Serial Next-Generation Sequencing of Circulating Cell-Free DNA Evaluating Tumor Clone Response To Molecularly Targeted Drug Administration. *Clin Cancer Res* **21**, 4586-4596, doi:10.1158/1078-0432.CCR-15-0584 (2015).
64 Olsson, E. *et al.* Serial monitoring of circulating tumor DNA in patients with primary breast cancer for detection of occult metastatic disease. *EMBO Mol Med* **7**, 1034-1047, doi:10.15252/emmm.201404913 (2015).
65 Garcia-Murillas, I. *et al.* Mutation tracking in circulating tumor DNA predicts relapse in early breast cancer. *Sci Transl Med* **7**, 302ra133, doi:10.1126/scitranslmed.aab0021 (2015).
66 Newman, A. M. *et al.* An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat Med* **20**, 548-554, doi:10.1038/nm.3519 (2014).

67 Illumina. *Grail A Revolution in Early Cancer Detection*, <http://www.illumina.com/content/dam/illumina-marketing/documents/company/investor-relations/investor_presentations/grail.investor.presentation.pdf> (2016).
68 Howlader N, N. A., Krapcho M, Garshell J, Miller D, Altekruse SF, Kosary CL, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis DR, Chen HS, Feuer EJ, Cronin KA (eds). (National Cancer Institue).
69 Spano, D., Heck, C., De Antonellis, P., Christofori, G. & Zollo, M. Molecular networks that regulate cancer metastasis. *Seminars in cancer biology* **22**, 234-249, doi:10.1016/j.semcancer.2012.03.006 (2012).
70 Aparicio, S. & Caldas, C. The implications of clonal genome evolution for cancer medicine. *The New England journal of medicine* **368**, 842-851, doi:10.1056/NEJMra1204892 (2013).
71 Diaz, L. A., Jr. *et al.* The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature* **486**, 537-540, doi:10.1038/nature11219 (2012).
72 Misale, S. *et al.* Emergence of KRAS mutations and acquired resistance to anti-EGFR therapy in colorectal cancer. *Nature* **486**, 532-536, doi:10.1038/nature11156 (2012).
73 Sorenson, G. D. *et al.* Soluble normal and mutated DNA sequences from single-copy genes in human blood. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* **3**, 67-71 (1994).
74 Chan, K. C. *et al.* Cancer genome scanning in plasma: detection of tumor-associated copy number aberrations, single-nucleotide variants, and tumoral heterogeneity by massively parallel sequencing. *Clinical chemistry* **59**, 211-224, doi:10.1373/clinchem.2012.196014 (2013).
75 Newman, A. M. *et al.* An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nature medicine*, doi:10.1038/nm.3519 (2014).
76 Misale, S. *et al.* Blockade of EGFR and MEK intercepts heterogeneous mechanisms of acquired resistance to anti-EGFR therapies in colorectal cancer. *Science translational medicine* **6**, 224ra226, doi:10.1126/scitranslmed.3007947 (2014).
77 Nakamura, T. *et al.* Application of a highly sensitive detection system for epidermal growth factor receptor mutations in plasma DNA. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer* **7**, 1369-1381, doi:10.1097/JTO.0b013e31825f2821 (2012).
78 Taniguchi, K. *et al.* Quantitative detection of EGFR mutations in circulating tumor DNA derived from lung adenocarcinomas. *Clinical cancer research : an official journal of the American Association for Cancer Research* **17**, 7808-7815, doi:10.1158/1078-0432.CCR-11-1712 (2011).
79 Shinozaki, M. *et al.* Utility of circulating B-RAF DNA mutation in serum for monitoring melanoma patients receiving biochemotherapy. *Clinical cancer research : an official journal of the American Association for Cancer Research* **13**, 2068-2074, doi:10.1158/1078-0432.CCR-06-2120 (2007).
80 Swisher, E. M. *et al.* Tumor-specific p53 sequences in blood and peritoneal fluid of women with epithelial ovarian cancer. *American journal of obstetrics and gynecology* **193**, 662-667, doi:10.1016/j.ajog.2005.01.054 (2005).

81 Crowley, E., Di Nicolantonio, F., Loupakis, F. & Bardelli, A. Liquid biopsy: monitoring cancer-genetics in the blood. *Nat Rev Clin Oncol* **10**, 472-484, doi:10.1038/nrclinonc.2013.110 (2013).
82 Fleischhacker, M. & Schmidt, B. Circulating nucleic acids (CNAs) and cancer--a survey. *Biochim Biophys Acta* **1775**, 181-232, doi:10.1016/j.bbcan.2006.10.001 (2007).
83 Frattini, M. *et al.* Quantitative and qualitative characterization of plasma DNA identifies primary and recurrent colorectal cancer. *Cancer letters* **263**, 170-181, doi:10.1016/j.canlet.2008.03.021 (2008).
84 Beadling, C. *et al.* Multiplex mutation screening by mass spectrometry evaluation of 820 cases from a personalized cancer medicine registry. *The Journal of molecular diagnostics : JMD* **13**, 504-513, doi:10.1016/j.jmoldx.2011.04.003 (2011).
85 Robinson, D. R. *et al.* Activating ESR1 mutations in hormone-resistant metastatic breast cancer. *Nature genetics* **45**, 1446-1451, doi:10.1038/ng.2823 (2013).
86 Toy, W. *et al.* ESR1 ligand-binding domain mutations in hormone-resistant breast cancer. *Nature genetics* **45**, 1439-1445, doi:10.1038/ng.2822 (2013).
87 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).
88 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).
89 http://picard.sourceforge.net/.
90 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**, 491-498, doi:10.1038/ng.806 (2011).
91 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).
92 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).
93 Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology* **31**, 213-219, doi:10.1038/nbt.2514 (2013).
94 Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308-311 (2001).
95 Forbes, S. A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* **43**, D805-811, doi:10.1093/nar/gku1075 (2015).
96 Lonigro, R. J. *et al.* Detection of Somatic Copy Number Alterations in Cancer Using Targeted Exome Capture Sequencing. *Neoplasia* **13**, 1019-IN1021, doi:10.1593/neo.111252 (2011).
97 Quail, M. A. *et al.* SASI-Seq: sample assurance Spike-Ins, and highly differentiating 384 barcoding for Illumina sequencing. *BMC Genomics* **15**, 110, doi:10.1186/1471-2164-15-110 (2014).
98 Kircher, M., Sawyer, S. & Meyer, M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res* **40**, e3, doi:10.1093/nar/gkr771 (2012).
99 Schnell, I. B., Bohmann, K. & Gilbert, M. T. Tag jumps illuminated--reducing sequence-to-sample misidentifications in metabarcoding studies. *Mol Ecol Resour* **15**, 1289-1303, doi:10.1111/1755-0998.12402 (2015).

100 Lou, D. I. *et al.* High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. *Proc Natl Acad Sci U S A* **110**, 19872-19877, doi:10.1073/pnas.1319590110 (2013).
101 Shugay, M. *et al.* Towards error-free profiling of immune repertoires. *Nat Methods* **11**, 653-655, doi:10.1038/nmeth.2960 (2014).
102 Howlader, N. *et al.* US incidence of breast cancer subtypes defined by joint hormone receptor and HER2 status. *J Natl Cancer Inst* **106**, doi:10.1093/jnci/dju055 (2014).
103 Dai, X. *et al.* Breast cancer intrinsic subtype classification, clinical use and future trends. *Am J Cancer Res* **5**, 2929-2943 (2015).
104 Prat, A. *et al.* Response and survival of breast cancer intrinsic subtypes following multi-agent neoadjuvant chemotherapy. *BMC Med* **13**, 303, doi:10.1186/s12916-015-0540-z (2015).
105 Comparisons between different polychemotherapy regimens for early breast cancer: meta-analyses of long-term outcome among 100 000 women in 123 randomised trials. *The Lancet* **379**, 432-444, doi:10.1016/s0140-6736(11)61625-5 (2012).
106 Haddad, T. C. & Goetz, M. P. Landscape of neoadjuvant therapy for breast cancer. *Ann Surg Oncol* **22**, 1408-1415, doi:10.1245/s10434-015-4405-7 (2015).
107 Rastogi, P. *et al.* Preoperative chemotherapy: updates of National Surgical Adjuvant Breast and Bowel Project Protocols B-18 and B-27. *J Clin Oncol* **26**, 778-785, doi:10.1200/JCO.2007.15.0235 (2008).
108 von Minckwitz, G. *et al.* Definition and impact of pathologic complete response on prognosis after neoadjuvant chemotherapy in various intrinsic breast cancer subtypes. *J Clin Oncol* **30**, 1796-1804, doi:10.1200/JCO.2011.38.8595 (2012).
109 Esserman, L. J. *et al.* Pathologic complete response predicts recurrence-free survival more effectively by cancer subset: results from the I-SPY 1 TRIAL--CALGB 150007/150012, ACRIN 6657. *J Clin Oncol* **30**, 3242-3249, doi:10.1200/JCO.2011.39.2779 (2012).
110 Park, J. W. *et al.* Adaptive Randomization of Neratinib in Early Breast Cancer. *N Engl J Med* **375**, 11-22, doi:10.1056/NEJMoa1513750 (2016).
111 Rugo, H. S. *et al.* Adaptive Randomization of Veliparib-Carboplatin Treatment in Breast Cancer. *N Engl J Med* **375**, 23-34, doi:10.1056/NEJMoa1513749 (2016).
112 Caudle, A. S. *et al.* Predictors of tumor progression during neoadjuvant chemotherapy in breast cancer. *J Clin Oncol* **28**, 1821-1828, doi:10.1200/JCO.2009.25.3286 (2010).
113 Madic, J. *et al.* Circulating tumor DNA and circulating tumor cells in metastatic triple negative breast cancer patients. *Int J Cancer* **136**, 2158-2165, doi:10.1002/ijc.29265 (2015).