Abstract: 150 words
Main: 6888
Figures: 7
Tables: 1
References: 181
Ed Summary: Blood-derived cell-free DNA are a promising source of noninvasive biomarkers for cancer, but challenges remain in pre-analytic processing, library preparation, and bioinformatic analysis that limit the accuracy of cell-free DNA diagnostics.

Limitations and opportunities of technologies for the analysis of cell-free DNA in cancer diagnostics

Ping Song*,[1] Lucia R. Wu*,[1] Yan Helen Yan,[1] Jinny X. Zhang,[1] Tianqing Chu,[2] Lawrence N. Kwong,[3] Abhijit A. Patel,[4] and David Yu Zhang#[1]

1Department of Bioengineering, Rice University, Houston, TX
2Department of Respiratory Medicine, Shanghai Chest Hospital, Shanghai Jiao Tong University, Shanghai, China
3Department of Translational Molecular Pathology, MD Anderson Cancer Center, Houston, TX
4Department of Therapeutic Radiology, Yale School of Medicine, New Haven, CT
* Equal contribution author

corresponding author: dyz1@rice.edu

Cell-free DNA (cfDNA) in the circulating blood plasma of patients with cancer contains tumour-derived DNA sequences that can serve as biomarkers for guiding therapy, for the monitoring of drug resistance, and for the early detection of cancers. However, the analysis of cfDNA for clinical diagnostic applications remains challenging because of the low concentrations of cfDNA, and because cfDNA is fragmented into short lengths and is susceptible to chemical damage. Barcodes of unique molecular identifiers have been implemented to overcome the intrinsic errors of next-generation sequencing, which is the prevailing method for highly multiplexed cfDNA analysis. However, a number of methodological and pre-analytical factors limit the clinical sensitivity of the cfDNA-based detection of cancers from liquid biopsies. In this Review, we describe the state-of-the-art technologies for cfDNA analysis, with emphasis on multiplexing strategies, and discuss outstanding biological and technical challenges that, if addressed, would substantially improve cancer diagnostics and patient care.

Dying cells release their DNA into blood plasma, where it is fragmented by nucleases to cell-free DNA (cfDNA) molecules — short (≈160 nt) double-stranded DNA fragments that are cleared from the bloodstream with a half-life between 5 and 150 min [1–3]. Because cfDNA captures a "snapshot" of dying cells throughput the whole body, it can be used to detect a broad and diverse set of diagnostic biomarkers for a variety of diseases. In particular, cfDNA has gained traction for cancer diagnostics over the past 5 years [4–13], due to the high cost and complexity often associated with the procedures based on radiology and tissue biopsies. The subset of cfDNA molecules that are tumor-derived is known as circulating tumor DNA (ctDNA), and in cancer literature cfDNA and ctDNA diagnostics are often used interchangeably because ctDNA is analyzed from a sample of cfDNA. Tumor-specific mutations can be used to distinguish ctDNA from healthy cfDNA, and the fraction of cfDNA molecules at a particular locus that bears a mutation is known as the variant allele fraction (VAF). Accurate detection and quantitation of mutation VAF is the primary technical challenge of cfDNA-based diagnostics discussed here.

Cancer diagnostics based on cfDNA face 4 unique challenges: (1) the short length of cfDNA, (2) the low concentration of cfDNA in plasma, (3) the high sequence similarity between cancer-derived and healthy human DNA, and (4) the large number of markers that must be simultaneously analyzed to achieve high clinical sensitivity. These challenges render older nucleic acid testing technologies, such as quantitative PCR methods for infectious pathogen identification, inadequate for cfDNA. Innovations in high-throughput sequencing instruments, library preparation methods, and bioinformatics pipelines been developed in the past 15 years, but cfDNA diagnostics have yet to see widespread introduction in the clinical setting in part because there is not a single dominant technology that reliably, simply, and affordably addresses all these challenges. Here, we discuss, at a detailed molecular level, the current methodological limits to cfDNA analysis, in order to stimulate further developments and scale-up in ctDNA-based diagnostics that can positively impact human health.

cfDNA diagnostics in cancer care

The United States Food and Drug Administration (FDA) defines an In Vitro Diagnostic as "tests done on samples such as blood or tissue that have been taken from the human body." Within this broader definition, there are 3 sub-definitions of diagnostics that are typically considered in the context of cancer: (1) non-FDA approved analytic tests based on patient-derived biospecimens that provide information on the presence, characteristics, or evolution of the patient's disease [14], (2) tests approved or cleared by the FDA via 510(k) [15], de novo [16], or pre-market approval pathways [17], and (3) pathology tests to definitively identify and classify a malignancy. Here, we use the first (broadest) definition of diagnostics, which includes screening, prognosis, therapy selection, and post-treatment monitoring tests, many of which have not received FDA approval or clearance, but are nonetheless informative for cancer care and are recommended by guidelines (e.g. NCCN). Typical nucleic acid tests can be separated into three main stages: preanalytical steps that result in a purified DNA sample, the analysis of the DNA sample, and the clinical interpretation of the results. Fig. 1 describes some cfDNA tests and how they would fit into the clinical diagnostic workup and treatment workflow, using non-small cell lung cancer (NSCLC) as a model disease.

Today, approximately 70% of NSCLC patients are diagnosed at late stages (III and IV) following overt clinical symptoms; only 30% of patients are diagnosed at Stage I or II, often following incidental findings or screening exams based on chest CT scans [18]. Because patient outcomes are significantly worse when NSCLC is detected at later stages, despite the advances in targeted therapies and immunotherapies [19-21], there are major initiatives in the US [22,23] and across the world [28,29] to improve and expand early screening efforts.

Currently, the most common use of cfDNA analysis is in therapy selection for Stage IIIb and IV patients. For example, EGFR mutation tests stratify patients based on likelihood of response to targeted therapies such as erlotinib [26-28] or osimertinib [29, 30]. There are over 100 cfDNA diagnostic tests in clinical trials in the US [31], and several commercial cfDNA-based laboratory-developed tests (LDTs) are being routinely ordered by oncologists [32, 33]. In addition to

specific mutation markers for resistance or sensitivity to targeted therapeutics such as tyrosine kinase inhibitors (TKI), tumor-specific DNA also can be analyzed more broadly for overall tumor mutational burden (TMB) that positively correlates with the efficacy of immunotherapies such as PD-1 [34], PD-L1 [35], and CTLA4 [36] inhibitors. The approximation of TMB from cfDNA analysis, known as blood tumor mutation burden (bTMB) [37], is an important recent use case of cfDNA-based immunotherapy guidance. Similarly, genome-wide microsatellite instability (MSI) has also been reported to be correlated with immunotherapy effectiveness [38, 39], and represents another set of promising markers for cfDNA-based therapy guidance.

As cfDNA analysis technologies advance in academic and clinical research [40-45], they are also being considered for post-treatment monitoring, including detection of recurrence and de novo resistance mutations, which may inform modification of patient therapy regimens including to combination therapies [45-48]. Recurrence monitoring has been applied in a research setting for a number of different cancer types, including breast [49, 50], colorectal [51, 52], and lung [53]. Given the active research and commercialization efforts, we expect that cfDNA-based cancer monitoring tests may soon become commercially available.

Early cancer screening via cfDNA diagnostics is a widely discussed possibility in both academia and industry [54- 56]. In a limited number of cancer types in which high-risk individuals can be identified by age, lifestyle habits, or geographic locations, cancer screening via cfDNA is becoming a reality (e.g. colorectal cancer [57, 60] and nasopharyngeal cancer [61]). However, the recent discovery of significant presence of cancer-associated mutations in healthy individuals [62, 63] suggests that it will be challenging to develop diagnostic tests with high sensitivity and specificity for pan-cancer early detection in asymptomatic populations. These challenges are exacerbated by the cost and dangers associated with applying diagnostic workups to healthy individuals (see Fig.7).

Cytosine methylation in cfDNA is being explored as an early cancer detection marker in multiple different approaches. Global methylome profiling, like with TMB for mutations, does not consider individual methylation markers at specific genomic loci; genome-wide hypomethylation has been suggested as a universal marker for cancer [64]. However, because genome-wide methylation does not inform on the location of the tumour and because epigenetic features are less conserved across cell divisions, clinical actionability is limited to the early detection of cancer. Targeted bisulfite sequencing approaches [65-67], on the other hand, are more suitable for the detection of specific cancer types. Combined with bioinformatics including haplotype phasing [68], methylation markers can be used to identify the tissue of origin based on cfDNA [69], representing a significant advantage for early detection applications.

There are numerous clinical trials that employ mutational analysis of cfDNA. Due to the long timeframes required for prospective clinical trials, many ongoing trials using cfDNA as a stratification marker remain based on low-plex PCR methods, such as NCT02418234 (non-small cell lung cancer, China), NCT00730158 (colorectal cancer, Yale), and NCT01349959 (breast cancer, NCI) [70]. With the maturation of NGS technology, multiple clinical trials based on NGS of specific gene panels using cfDNA samples have been initiated by companies like GRAIL (NCT02889978), Guardant Health (NCT03477474), and Foundation Medicine (NCT02620527).

Preanalytical factors and limitations

Preanalytical factors describe the biological variables and handling protocols of the sample and, because different protocols impact the quality, quantity, or characteristics of the DNA sample to be analyzed (Fig. 2a), they can have an especially outsized impact on the accuracy of the overall tests.

Many biological variables that impact cfDNA quantity and characteristics are difficult to fully control, for example: cfDNA is partially cleared through urine, so an individual who has recently imbibed a large amount of fluids may have lower concentrations of cfDNA; cfDNA is derived from all dying cells in the body, so an individual's physical health state, including exercise[71] and bacterial/viral infection [72-74] will affect concentrations of cfDNA; and the quantity of tumor-derived ctDNA molecules in cfDNA depend not only on the tumor mass, but also on its proximity and accessibility to the circulatory system.

In contrast to biological variables, the sample collection and handling protocols could, in principle, be fully controlled to maximize the reproducibility of results from two aliquots of the same blood sample. Ideally, fresh venous blood samples should be immediately centrifuged to separate plasma from red blood cells and buffy coat (this step is typically done twice to minimize contamination from buffy coat). Subsequently, the cfDNA should be immediately extracted from plasma, followed by downstream analysis by digital polymerase chain reaction (dPCR) or next-generation sequencing (NGS). In practice however, there are often unavoidable delays associated with transport, aliquoting and storage of the blood sample that run two main risks: chemical damage of the cfDNA (Fig. 2b), and contamination of cfDNA by leukoctye genomic DNA (Fig. 2c). DNA can in fact undergo hydrolysis, deamination, and oxidative damage both in vivo in the body and in vitro in the blood collection tube [75, 76], with more than 20 types of damage identified [77]. The products of these undesired chemical reactions are non-canonical nucleosides that can be spuriously recognized as mutations, and these processes are suspected to impact cfDNA analysis accuracy. Although there are some commercial kits that claim to repair such damage (e.g. New England Biolab's preCR), the general consensus of the field is that damage repair is imperfect and results in significant false positive variant calls. A high specificity and high yield method for reversing DNA damage could significantly improve the ultimate limits of cfDNA-based diagnostics.

In collection tubes, leukocytes in the blood slowly die and release their genomic DNA into the plasma layer; this genomic DNA contributes increased background of wild-type DNA and reduces the effective mutation (VAF), rendering detection of rare cancer mutations more difficult. Literature suggests that whole blood can be stored between 1-7 days at room temperature before significant increase in plasma DNA quantity due to leukocyte genomic DNA [78]; Streck brand blood collection tubes have been observed to increase the stability of leukocytes in collected blood samples [74, 78]. Methods

for improving the stability of leukocytes in blood, or differentiating cfDNA from leukocyte genomic DNA would improve the accuracy and reproducibility of cfDNA diagnostics.

The low quantity of cfDNA in blood and the low VAF of cancer mutations mean that Poisson sampling statistics can reduce the reproducibility of mutation VAF profiling (Fig. 2d,e). This is an ultimate limitation of cfDNA and implies that detection and quantitation of low VAF mutations can suffer from irreproducibility and lowered clinical sensitivity, regardless of the downstream analysis technology. The only way to overcome this limitation is through the use of larger quantities of cfDNA; for this reason, many commercial LDTs for cfDNA require 2 tubes of 10 mL blood as input, allowing near 100% clinical sensitivity at 1% VAF, and over 90% for 0.1% VAF. An adult human can lose up to 14% of their total 5 L blood supply without experiencing significant adverse effects, corresponding to roughly 700 mL [79]. However, it is likely to be practically difficult to collect more than about 50 mL blood from cancer patients for cfDNA analysis, given their compromised health.

Urine is another potential source of cfDNA molecules, because their small size results in a significant fraction being filtered by the kidneys into urine [80]. Adults usually pass between 800 mL and 2 L of urine per day, all of which could potentially be collected and used for extracting cfDNA without any adverse effects. However, high-yield purification and concentration of cfDNA from large volumes of urine, while limiting contamination from genomic DNA from cellular debris, is technically challenging. Additionally, cfDNA present in urine has been reported to be significantly shorter than cfDNA derived from blood [81], presenting unique challenges to cfDNA extraction yields [82, 83] as well as PCR amplification. Biochemical and/or physical methods to reliably, and affordably purify and analyse at scale cfDNA from urine could revolutionize cancer cfDNA diagnostics.

Low-plex approaches to cfDNA analysis

Traditional nucleic acid tests, such as those used for the detection of viral infections such as HIV and of pathogens such as MRSA, utilize low-plex instruments and assays that detect a small number of target DNA sequences. Typically, these assays are run using quantitative PCR (qPCR) [84], but other FDA-approved assays include chemiluminescence detection [85, 86], isothermal DNA amplification [87], and transcription-mediated amplification [88]. In these tests, there is typically a single binary decision that the nucleic acid test is meant to inform, such as the use of antiretroviral therapy. Cancer, in contrast, is a complex disease with many different pathways and many different treatment options. For example, there are dozens of FDA approved targeted therapies and immunotherapies for NSCLC [89]; furthermore, many drugs can be used in combination to maximise a therapeutic effect [31,32]. Consequently, more information is required than simply the presence or absence of a tumor. The presence of specific DNA mutations can not only inform therapy regimens most likely to be effective for a patient, but also provide snapshots of tumor response to the treatment including the emergence of drug-resistant tumors. The MyCancerGenome database [90] lists hundreds of mutations with known effects on cancer treatment, and databases such as the Catalog Of Somatic Mutations In Cancer database [91] and cBioPortal [92] lists over 100,000 mutations observed in cancer patients.

Because cfDNA exists in plasma in very low quantities (about 2.5 ng/mL plasma in healthy individuals, and about 10 ng/mL plasma in cancer patients), repeated low-plex testing on different sample aliquots is not practical. Consequently, low-plex tests such as those based on qPCR typically target one or a few specific mutations to guide the use of a single drug. For example, presence of the EGFR-T790M mutation in a lung cancer patient confers resistance to erlotinib [93], and informs the use of osimertinib [94], and can alternatively be used for detection of recurrence in erlotinib-resistant tumors.

Currently, digital PCR is the most used method for low-plex analysis of cfDNA. By performing end-point PCR on 20,000 individual reaction droplets, Bio-Rad's digital droplet PCR (ddPCR) allows accurate detection and quantitation of known DNA mutations without separate calibration reactions [95]. Compared to commercial qPCR amplification refractory mutation system (ARMS) assays [96], ddPCR assays achieve a significantly better VAF limit of detection by roughly 20-fold (0.05% vs. 1%), as well as more accurate VAF quantitation. However, the high cost of ddPCR instruments (list price of roughly $100,000) and the low number of installed ddPCR instruments comprise a significant challenge for the widespread adoption of ddPCR-based cfDNA assays, compared to NGS instruments with similar cost and much higher multiplexing (Fig. 3). Because ddPCR currently is only capable of analyzing one potential mutation per reaction, an unreasonably high quantity of cfDNA samples would be needed to profile many different mutations.

Other technologies for low-plex detection of mutations with low VAF in cfDNA have been described in the academic literature. These approaches include electrochemistry [97, 98], isothermal amplification with CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) [99], nanoparticles [100], and single-molecule fluorescence [101]. These emerging technologies have primarily focused on improving analytic sensitivity, by either requiring lower input cfDNA quantity or enabling lower mutation VAF. Nevertheless, more effort should be devoted to massively scaling up the multiplexing capabilities of these technologies to render them suitable for broader analysis of cfDNA for cancer diagnostics.

Sitting on the border between low-plex and massively multiplexed technologies, the Agena MassARRAY system [102] allows for many different primers to be added to the same reaction, and a potential single-nucleotide extension is performed; the potential extension products are then simultaneously analyzed via mass spectrometry. The most recent assays from Agena claim a limit of detection of 0.1% mutation VAF and detection of up to 40 different known mutations from a single sample [103]. This performance puts the MassARRAY platform at roughly the needed multiplexing capacity for actionable cancer mutation detection for diagnosis of individual cancer types; simultaneously, the low marginal cost of sample testing and the rapid workflow make the system an attractive alternative to next-generation sequencing (NGS) methods described in the rest of this review. As with digital PCR, the major adoption barrier for MassARRAY is the high up-front instrument cost ($250,000).

NGS methods for cfDNA analysis

In NGS, DNA molecules in a solution that bear pre-defined adapter sequences are randomly sampled, and the NGS instrument provides the sequences of the sampled molecules from the 5′ end, up to a defined length limit defined (the read length). For the popular Illumina NGS instruments and sequencing kits, the read length varies between 75 and 300 nt, and the number of reads (sampled molecules) varies between 4 million and 10 billion. Other major NGS platforms include Ion Torrent, Oxford Nanopore, and Pacific Biosciences (see Fig. 3a).

NGS offers orders of magnitude higher multiplexing than other nucleic acid analysis technologies and is thus the dominant approach to cfDNA analysis. Because cancer mutations follow a long-tailed distribution, a very large number of genetic loci should be simultaneously observed to ensure high clinical sensitivity [104]. Furthermore, the cost of performing NGS has been exponentially dropping (halving roughly every 18 months [105]), increasing its appeal.

Illumina NGS is commonly used for cfDNA analysis, due to the high accuracy and the low marginal cost of NGS reads (Fig. 3a). In some countries (e.g. India), Ion Torrent NGS instruments have gained higher market penetration than Illumina due to the lower upfront instrument cost. The third generation NGS platforms (Oxford Nanopore and Pacific Biosciences) [106, 107] are not currently competitive in the space of cfDNA analysis, because the primary advantage of these platforms are long read lengths of over 10,000 nt, making them a poor fit for short cfDNA with lengths of typically 160 nt.

If all the NGS reads are deployed to randomly sample all cfDNA molecules in a plasma sample, then the fraction of all reads that correspond to useful information about cancer-related genes will be very small. The human genome is over 3 x 10^9 nt long, and current cancer biology knowledge is limited to roughly 1000 possibly cancer-related genes [108], each with about 4,000 nt of protein-coding sequences, corresponding to a total of 4 x 10^6 nt. Thus, simplistically, roughly 99.9% of the NGS reads would be wasted on portions of the human genome with little cancer-relevant information, resulting in grossly increased NGS costs. Target enrichment is the process by which the composition of the cfDNA library is adjusted to increase the relative concentrations of DNA sequences corresponding to the genomic loci of interest. The two most popular methods for target enrichment today are ligation/hybrid-capture, and multiplex PCR.

In a typical ligation/hybrid-capture workflow (Fig. 3b), cfDNA first undergoes end-repair to produce flush ends with a single 3′ A tail, and then adapters are ligated to both ends of the duplex. Index primers are further appended via PCR using primers against the universal adapter sequences; this step also serves to pre-amplify the cfDNA. Subsequently, the amplicons are denatured and then hybridized to biotinylated probe sequences. Streptavidin-coated magnetic beads are then used to capture the biotinylated probes and any cfDNA amplicons bound to the probes; other cfDNA amplicons are removed via washing. The probes correspond to the genes and/or loci of interest; consequently, the capture cfDNA amplicons will be enriched in the genes/loci of interest. However, due to nonspecific binding of cfDNA amplicons to the probes and/or the magnetic beads, enrichment is imperfect. Hybrid-capture probes are available commercially via companies such as Twist Biosciences, Integrated DNA Technologies, Nimblegen, and Agilent. NGS panels for cfDNA that rely on hybrid-capture include FoundationACT [36], Guardant360 [37], and Roche Avenio [109].

Target enrichment via multiplexed PCR allows for different genes/loci to be simultaneously amplified using different PCR primer sequences [110,111], and adapters and indexes are appended afterwards using PCR or ligation (Fig. 3c). With the large number of PCR primers present, some amount of primer dimers and nonspecific amplicons from other regions of the genome are likely to form. A majority of these undesired molecules can be removed by size selection steps in the NGS library preparation process (e.g. using Agencourt AMPureXP[112]) to remove primer dimers and nonspecific amplicons with grossly different lengths than the expected amplicons. Because the amplicon concentration of the loci of interest doubles with every PCR cycle, the fold-enrichment can be significantly higher than with hybrid-capture. For example, with 20 PCR cycles, the loci of interest are enriched up to 10^6-fold, whereas it is difficult even with optimized hybrid-capture protocols to ensure that nonspecific binding is less than 1 part in 10^4. Thus, multiplexed PCR is generally able to achieve higher on-target rates than hybrid-capture, especially for smaller NGS panels. Additionally, performing multiplexed PCR is generally less complicated than hybrid-capture, with shorter total turnaround and hands-on time. The commercial Thermo Fisher Oncomine and Paragon CleanPlex NGS panels use multiplexed PCR for target enrichment [113, 114].

On-target rates of hybrid-capture methods are generally high for panels larger than 100 kilobases (kb), but low for smaller panels (the on-target rate of an NGS library is the fraction of all reads that correspond to the genes/loci of interest). In contrast, on-target rates of multiplex PCR methods are generally high for panels smaller than 10 kb, but low for larger panels. In general, ligation/hybrid-capture is preferred for large NGS panels covering over 100 kb, and multiplex PCR is preferred for small panels covering less than 10 kb (Fig. 3d).

Both methods for NGS library preparation face the same three main limitations: (1) PCR amplification and NGS read errors that result in false positive variant calls, (2) imperfect representation of the original cfDNA molecules in the NGS library, resulting in false negatives, and (3) sequencing non-uniformity that either reduces mutation sensitivity or significantly increases NGS cost (Fig. 4).

Errors in PCR amplification and in NGS can result in NGS reads containing variant sequences even when the sample is purely wild type (0% VAF). Theoretically, if the error rate was absolutely reproducible from procedure to procedure, then an unknown sample's actual mutation VAF can be mathematically computed by subtracting the expected false variant NGS reads due to errors. For example, if 2% of the reads from an NGS library contain a particular mutation, and the aggregated errors result in a reproducible 1% of NGS reads being that variant, then we can infer the true sample VAF as 2% - 1% = 1%. In practice, however, there are run-to-run variations due to different enzyme lots, slight differences in experimental temperatures, times, and concentrations, so that the aggregated error rate will not be perfectly reproducible. Furthermore, the error rate is not identical for all sequences, and can vary based both on the exact nucleotide being sequenced as well as the neighboring sequences. For this reason, it is difficult to standardize NGS panels to confidently claim mutation detection below about 1% VAF, even when using high-fidelity polymerases to

reduce PCR error and Phred quality score Q≥30 filtering to reduce NGS intrinsic error (Fig. 4a,b). To further clarify regarding Phred quality quality score, Q≥30 indicates a mean error rate of 0.1% per nucleotide, based on manufacturer specifications. However, given the large number of nucleotides sequenced and sequence biases in error rate, variant calls at ≤1% cannot be reliably made even with Q≥30 filters.

The conversion yield — the fraction of original cfDNA molecules that are represented in the final NGS library— depends on whether the library preparation method used ligation/hybrid-capture or multiplex PCR. For example, a 10 mL blood sample could contain a single-digit number of copies of DNA molecules with tumour specific mutations, so protocols with low conversion yields could end up losing all the mutant DNA molecules and report a false negative. Thus, conversion yield is an important determinant of the clinical sensitivity of cfDNA assays. Panel developers are incentivized to report optimistic numbers, based on lowest reasonable estimate of the denominator — the number of original cfDNA molecules present at a particular genomic locus. Standard DNA quantitation methods based on fluorescence of an intercalating dye (Qubit), absorbance of DNA at 260 nm (Nanodrop), and digital droplet PCR can differ by more than 2-fold depending on DNA size distribution, presence and concentration of single-stranded nucleic acids, DNA sequence, fragmentation pattern, solution buffer, and chemical impurities. Thus, conversion yield metrics today are estimates based on imperfect underlying measurement of cfDNA quantity.

For ligation/hybrid-capture, the conversion yield is primarily limited by imperfect end-repair and ligation efficiency (Fig. 4c). Because both ends of a cfDNA fragment must be ligated to adapters for the cfDNA to be amplified in subsequent steps, imperfect ligation yields have a quadratic effect on conversion yield. In some specific library preparation protocols such as DuplexSeq [121], that rely on ligation to both strands of the same cfDNA fragment for further error correction, all 4 ligation reactions must be complete for the molecule to be represented. Reported conversion yields for ligation/hybrid-capture vary between 10% and 60% [121-123], with the upper range being possibly optimistic.

For multiplex PCR, the conversion yield is primarily limited by the fraction of cfDNA molecules that cannot be amplified because the molecules do not span the length of the amplicon (Fig. 4d). Assuming the typical length of cfDNA is 160 nt [124, 125], the theoretical conversion yield of multiplex PCR can be calculated based on the length of the amplicon: A 100 nt amplicon would exhibit a conversion yield of approximately $\frac{160-100}{160} = 37.5\%$, and conversion yield would drop precipitously for longer amplicons. Furthermore, recent studies suggest the existence of a population of very short ctDNA molecules in blood [126, 127], which have previously not been systematically characterized due to limitations in DNA extraction and NGS library preparation. Multiplex PCR analysis of this short ctDNA population would likely have very low yield. Exosomal DNA also exists in blood plasma and has been reported to contain cancer-specific DNA mutations [128, 129]; exosomal DNA's longer lengths of over 2,500 nt render these fragments relatively easy to amplify by PCR.

In both ligation/hybrid-capture and multiplex PCR protocols, some loci or amplicons are sequenced to much higher depth than others (Fig. 4e), due to sequence-based hybridization kinetics that impact both PCR amplification yield and hybrid-capture efficiency. The rate constants of DNA hybridization kinetics can vary more than 3 orders of magnitude for primers/probes of the same length, at the same temperature and buffer conditions [130]. The concentrations of different PCR primers or hybridization probes can be adjusted to counteract the differences in kinetics between targets (i.e. increasing the concentrations of primers/probes that have slow kinetics) in order to make NGS read depth more uniform, but adjustment is imperfect and there is typically still a 5- to 50-fold gap between the mean and minimum sequencing depth for commercial NGS panels.

Sequencing non-uniformity, unlike sequencing error and conversion yield, increases the cost of achieving a desired sensitivity, rather than setting hard limits on sensitivity. For example, sequencing to 200x depth is typically sufficient to make confident mutations calls at 5% VAF; thus sequencing to 1000x mean depth is sufficient to ensure a 5% VAF limit of detection for all loci in an NGS panel with a 5-fold gap between mean and minimum depth. Nonetheless, for cfDNA analysis where typical commercial panels, such as the Guardant 360, consume over $1000 of NGS reads per sample (against a list price of $6000), reducing cost through improved depth uniformity remains a priority.

Unique molecular identifier technologies

In the cfDNA of cancer patients, the VAF of cancer-specific mutations can vary between 0.01% and 10%, depending on both disease stage and individual-specific disease characteristics [4]. Mutations in cfDNA may have low VAFs not only because the disease is at an early stage and the tumor mass is small, but also because subclonal mutations are present in only a subset of tumor cells. Subclonal mutations are especially important for therapy selection, because rare subclones with resistance mutations can lead to rapid treatment failure due to subclone expansion under therapy [46]. Achieving mutation VAF limit of detections of 0.1% or lower is thus critical for high clinical sensitivity for cfDNA analysis.

Unique molecular identifiers (UMIs) are currently the most popular method for overcoming the PCR and NGS errors described previously, in order to reliably detect and quantitate mutations at ≤0.1% VAF [131-134]. Recently published NanoSeq method can achieve error rate of less than 10^{-9} errors per bp [135]. SaferSeqS can detect VAF as low as 1 in 100,000 DNA termplate molecules with a background mutation rate of < $5x10^{-7}$ mutants per bp [136]. The key idea of UMIs is to attach a unique DNA sequence to each original molecule of DNA in the cfDNA sample (Fig. 5a) When the UMI is subsequently PCR amplified and sequenced, all NGS reads with the same UMI sequence can be interpreted as being derived from the same original DNA sequence. Spurious mutation reads generated from PCR or NGS errors are likely to be a minority of reads within the family of reads with the same UMI sequence (Fig. 5b), but true mutations will generate families of reads in which all or a majority of the reads have the mutation (Fig. 5c). The bioinformatic interpretation of NGS data with UMIs starts by grouping different NGS reads into "UMI families," reads on a DNA locus with identical UMI sequence. Subsequently, a "vote" is taken for each UMI family, with the identified dominant/majority sequence being accepted as the true sequence of the original DNA molecule. Using UMIs, many PCR and NGS errors can be corrected, and the mutation VAF limit of detection can be brought below the PCR and NGS error rates. UMIs can be applied in both ligation/hybrid-capture and multiplex PCR protocols, though it is more difficult for the latter when the plex number is high.

Because UMIs can only function effectively to correct PCR and NGS errors when the UMI family size is large enough to allow a majority vote, UMIs increase the required sequencing depth and cost by at least a factor of 5. Furthermore, unlike standard NGS, the amount of input cfDNA needs to be carefully controlled when UMIs are used. For example, when a DNA sample is sequenced to a mean 30,000x depth for a particular gene panel, the average UMI family size will be 10 if the input amount is 10 ng cfDNA, but will only be 2 if the input amount is 50 ng cfDNA.

The bioinformatic interpretation of UMIs is also somewhat challenging because the UMI sequences themselves could have PCR or NGS errors (Fig. 5d), which are difficult to distinguish from sequences with small UMI family sizes due to poor PCR amplification efficiency (Fig. 5e). The typical bioinformatic workflow ignores all sequence information from UMI families with fewer than either 5 or 3 reads, which effectively mitigates detection and quantitation inaccuracies due to UMI errors, but also discards information from a significant number of original cfDNA molecules, effectively reducing the conversion yield. An average UMI family size of 12 will thus result in a roughly 30% drop in effective conversion yield (Fig. 5f); using smaller UMI family sizes would significantly increase the number of original molecules whose information is discarded.

Allele enrichment technologies

Allele enrichment strategies refer to library preparation methods that seek to detect low VAF mutations by increasing the VAFs upstream of sequencing (Fig. 5g). For example, a cfDNA sample that contains 0.1% VAF of a particular mutation may generate a library that is 10% VAF in the same mutation; the latter is simple to detect and quantitate even with low-depth sequencing. Thus, in contrast to the UMI strategy that increases sequencing cost by more than 10-fold, allele enrichment methods would decrease the sequencing cost while achieving better limits of detection (Fig. 5h).

Allele enrichment can be achieved through either the removal of wildtype alleles or selective amplification of variant alleles. For example, oscillatory electrophoresis can amplify the mobility differences of DNA molecules differing by even a single nucleotide, and allows effective removal of wildtype sequences [137, 138]. Recently, wildtype-specific probes and double-strand specific nucleases have been used to selectively degrade wildtype DNA molecules [139], likewise improving VAF by depleting wildtype alleles. Selective enrichment of variant alleles is typically achieved through PCR methods in which the wildtype DNA sequences are prevented from being PCR amplified; examples of this approach include blocker PCR [140], LNA (locked nucleic acid) and PNA (peptide nucleic acid) clamp PCR [141, 142], ICE-COLD PCR [143 - 145], and blocker displacement amplification [146].

Allele enrichment technologies are not broadly applied currently despite their potential for three reasons. First, allele enrichment methods have generally not been demonstrated to perform robustly in high multiplex. Multiple allele enrichment methods have been demonstrated to work for fewer than 20-plex primers/probes [140-143, 146], but even the smallest cfDNA NGS panels today are at least 50-plex and many panels are over 1000-plex. Second, the VAF fold-enrichment needs to be stable and reproducible for accurate sample VAF quantitation. If a particular mutation's VAF is always increased 100-fold through allele enrichment3 then one can infer a 0.1% mutation VAF based on an observed 10% VAF in the NGS library; however, if the VAF fold-enrichment varies between 50 and 200 across different runs, then VAF quantitation becomes significantly less accurate. Finally, allele enrichment technologies typically struggle significantly with on-target rates, the fraction of NGS reads that map to the gene loci of interest (regardless of whether it is a wildtype or mutant. This is because while wildtype DNA sequences are removed, off-target reads such as from primer dimers and non-specific amplification of other portions of the genome are not (Fig. 5i), so high-performance allele enrichment technologies are primarily bottlenecked by off-target reads, with respect to potential savings in NGS cost.

Inaccessible cfDNA Markers

Cancer markers in DNA can be grouped by type into (1) mutations, (2) gene fusions, (3) copy number variations (including loss of heterozygosity), and (4) aneuploidy [147]. Thus far, we have primarily discussed methods for detection of mutations, including point substitutions and small insertions/deletions (≤50 nt). This is because mutations are, in some sense, the easiest of the 4 marker classes to detect in cfDNA because they exhibit qualitatively different sequences at defined coding positions within genes.

Fusions are like mutations in that they result in qualitatively different sequences, but differ in that the unique cancer-specific sequence can reside at many different DNA loci [150, 151] (Fig. 6c). For example, the breakpoint for a ROS1 gene fusion in NSCLC can occur at any of the roughly 140,000 nt in the ROS1 gene's 44 introns. Considering that a typical mutation NGS panel for cfDNA is only about 100 kb, detection of fusions at the cfDNA level is technically possible, but the full coverage of the intron regions for high clinical sensitivity is economically non-viable when applied to many potential fusions. Some commercial panels detect fusions in cfDNA by focusing on intron loci that have been documented to show higher chance of being a fusion breakpoint, but these sacrifice clinical sensitivity. For these reasons, fusions are typically detected from mature mRNA, in which the introns are spliced out and the number of possible fusion sequences is limited [152, 153].

Copy number variations (CNVs) do not typically contain any unique sequences, because the ends of duplicated regions often reside in repetitive DNA sequences [154-156]. Thus, rather than searching for the presence of a unique DNA sequence, CNV profiling requires accurate quantitation of the potentially duplicated gene, and relative to other genes. However, the stoichiometric excess of DNA corresponding to the CNV gene is very small (Fig. 6a,b), because the fraction of cfDNA that is tumor-derived can be 1% or lower. The small stoichiometric excess is often obfuscated by the Poisson distribution nature of sampling cfDNA: a typical 10 ng cfDNA sample corresponds to ≈3000 haploid genome equivalents, so the number of DNA molecules at each locus will follow a distribution with standard deviation of $\sqrt{300} \approx 55$, corresponding to almost 2%. This challenge is partially mitigated by the fact that genes are long, so multiple distinct nonoverlapping cfDNA species are available for each gene. However, technical difficulties in appending UMIs compound the challenge of statistical distribution, and current commercial cfDNA assays exhibit a CNV limit of detection of roughly 20% VAF [36, 37], resulting in very low clinical sensitivity.

Aneuploidy is similar to CNV in that typically there are no unique sequences to serve as distinctive cancer markers, but is easier to detect than CNVs due to the vastly greater number of loci for statistical comparison. For example, a gene including introns may be up to 50 kb long, but even the shortest chromosome 22 is about 50 Mb long; this 1000-fold difference can result in a 30-fold lower coefficient of variation due to Poisson sampling. For this reason, aneuploidy is routinely detected at 4% VAF from cfDNA for non-invasive prenatal diagnostics for Down's syndrome [157]. Although aneuploidy has been observed in cancer [158], aneuploidy is not currently considered clinically actionable; consequently, most commercial cancer cfDNA panels do not include assays for aneuploidy.

Accuracy requirements for cancer screening via cfDNA.
There is strong enthusiasm regarding the possible use of cfDNA markers for the early detection and screening of cancers in asymptomatic individuals [159, 160]. For example, GRAIL Inc. has raised more than $1.4 billion in funding over the past 3 years to develop cfDNA technologies and run clinical trials for early cancer detection [161]. Here, we discuss the biological, statistical, and social challenges associated with screening and early detection.

The key biological challenges of cfDNA-based cancer screening are (1) a significant fraction of healthy individuals will have low-levels of cancer-associated DNA sequences in cfDNA, and (2) a significant fraction of individuals with early-stage cancer will have undetectable cancer-specific mutations in cfDNA. The first challenge may be due to clonal hematopoiesis [162-164], somatic mutations, or somatic mosaicism, and results in false positive screening results. The second challenge may be due to tumors with poor access to the circulatory system or pathogenic mutations from cancer pathways not currently understood, resulting in false negatives. Because of these biological challenges, it is not possible for any screening test to achieve 100% specificity and 100% sensitivity (Fig. 7a).

As shown in Fig. 7b, a hypothetical a test with 80% sensitivity and 90% specificity is used to test 10,000 samples in which 0.5% of the population have early-stage cancer. Because the prior probability of early-stage cancer is low at 0.5%, the posterior probability of a test-positive individual having early-stage cancer is still at a modest 3.9% (also known as the test positive predictive value, PPV). For an early cancer screening test, the typical next step for test-positive individuals would be a diagnostic workup that includes endoscopy, X-rays, CT scans, and/or biopsies (see Fig. 1). The 3.9% PPV means that 25 unnecessary diagnostic workups will be performed for each early-stage cancer patient; depending on the medical harm and economic costs of the diagnostic workup, this may not be an acceptable tradeoff.

Thus, realization of early cancer screening via cfDNA analysis will require (1) the specificity and sensitivity of the test be very close to 100%, (2) significant improvements to the safety and cost of diagnostic workups to confirm cfDNA findings, or (3) methods for enriching the tested population, e.g. via age or family cancer history. As a point of comparison, the FDA-approved Cologuard stool assay for early detection of colorectal cancer exhibits 94% clinical sensitivity and 87% clinical specificity [61], and is recommended for individuals aged 50 or older. Furthermore, the follow-up colonscopy exam for Cologuard test-positive individuals is considered relatively safe and inexpensive. Other recent advances in early cancer screening include detection of Epstein-Barr viral DNA for nasopharyngeal cancer [62] and combinations of protein and cfDNA markers for resectable tumors [63]. Global hypomethylation [68] and specific promoter hypermethylation [73] are both promising markers for early cancer screening for multiple cancer types. However, depending on the specific cancer type, the lower disease incidence and higher medical risk of diagnostic workup renders early cancer detection a far more difficult problem from a societal level (e.g. for gliomas/brain cancers, pancreatic cancer).

Outlook

Over 70 years after its discovery [159], cfDNA is starting to impact cancer care; numerous clinical trials are in progress in North America, Europe, and Asia to assess its diagnostic utility. However, many technical challenges and opportunities remain for cfDNA diagnostics to have sufficiently high clinical sensitivity for its widespread use for cancer monitoring. Technologies for analyzing cfDNA can broadly be split into rapid low-plex methods and expensive and slow high-plex NGS-based methods. Most cfDNA diagnostic applications require information on multiple markers in order to achieve high clinical sensitivity and inform treatment strategy, making NGS-based methods the preferred choice. Simultaneously, biological, statistical, clinical, physical, chemical, and economical constraints mean that only a small portion of all potentially available information is currently accessible on commercial cfDNA panels, and opportunities are rife for improving the state-of-the-art through scientific innovation (Table 1). Minimally invasive cancer diagnostic methods hold potential because common tissue sampling techniques, such as tumor biopsies, and medical imaging techniques that require the exposure to ionizing radiation, are limited to high-risk individuals and individuals with already identified lesions. In contrast, liquid biopsy-based diagnostics are suitable for repeat sampling and can potentially be used for early cancer detection and screening. Earlier detection and continuous monitoring of patients could help stratify individuals (Fig. 7). Identifying a set of biomarkers in cfDNA with sufficient specificity and sensitivity for the early detection of cancer may be challenging if the analysis is limited to DNA mutations. Other sources of biomarkers, such as cell-free RNA [166], exosomes[167]), methylation patterns[69-71,126], protein levels [169], and circulating tumor cells (CTCs) [170,171] may likely be included in early cancer detection panels, especially if applied to asymptomatic individuals, in addition to gene fusions and CNVs that are however currently challenging to detect in cfDNA. The multiplexed detection of analytes holds great potential for improving clinical sensitivity and demonstrating clinical utility, yet will require technical advances in sample preparation and analysis methods.

Acknowledgements. The authors acknowledge Steve Skates for useful discussions regarding diagnostic economics and outcomes. LNK is supported by NIH grant P01CA163222. AAP is supported by NIH grants R01CA197486 and R01CA233364. DYZ is supported by NIH grants R01CA203964 and R01CA233364, and by CPRIT grant RP180147.

Author Contributions. PS, LRW, and DYZ wrote the paper based on in-depth discussions with all authors.

Competing interests. PS declares a competing interest in the form of consulting for NuProbe USA. LRW declares a competing interest in the form of consulting for NuProbe USA. AAP declares a competing interest in the form of consulting for and significant equity ownership in Binary Genomics, and consulting for NuProbe USA. D.Y.Z. declares a competing interest in the form of consulting for and equity ownership in NuProbe USA, Torus Biosystems, and Pana Bio.

References

[1] Lo, Y. D. et al. Rapid clearance of fetal DNA from maternal plasma. The American Journal of Human Genetics **64**, 218-224 (1999).
[2] Diehl, F. et al. Circulating mutant DNA to assess tumor dynamics. Nature medicine **14**, 985-990 (2008).
[3] Jahr, S. et al. DNA fragments in the blood plasma of cancer patients: quantitations and evidence for their origin from apoptotic and necrotic cells. Cancer Research **61**, 1659-1665 (2001).
[4] Wan, J. C. et al. Liquid biopsies come of age: towards implementation of circulating tumour DNA. Nature Reviews Cancer **17**, 223-238 (2017).
[5] Diaz Jr, L. A., & Bardelli, A. Liquid biopsies: genotyping circulating tumor DNA. Journal of Clinical Oncology **32**, 579-586 (2014).
[6] Thierry, A. R.et al. Clinical validation of the detection of KRAS and BRAF mutations from circulating tumor DNA. Nature medicine **20**, 430-435 (2014).
[7] Alix-Panabieres, C., & Pantel, K. Clinical applications of circulating tumor cells and circulating tumor DNA as liquid biopsy. Cancer discovery **6**, 479-491 (2016).
[8] Mok, T.et al. Detection and dynamic changes of EGFR mutations from circulating tumor DNA as a predictor of survival outcomes in NSCLC patients treated with first-line intercalated erlotinib and chemotherapy. Clinical Cancer Research **21**, 3196-3203 (2015).
[9] Hao, T. B. et al. Circulating cell-free DNA in serum as a biomarker for diagnosis and prognostic prediction of colorectal cancer. British journal of cancer **111**, 1482-1489 (2014).
[10] Azad, A. A., et al. Androgen receptor gene aberrations in circulating cell-free DNA: biomarkers of therapeutic resistance in castration-resistant prostate cancer. Clinical cancer research **21**, 2315-2324 (2015).
[11] Lebofsky, R., et al. Circulating tumor DNA as a non-invasive substitute to metastasis biopsy for tumor genotyping and personalized medicine in a prospective trial across all tumor types. Molecular oncology **9**, 783-790 (2015).
[12] Schwarzenbach, H., Hoon, D. S. B. & Pantel, K. Cell-free nucleic acids as biomarkers in cancer patients. Nature Reviews Cancer **11**, 426-437 (2011).
[13] Goyal, L. et al. Polyclonal secondary FGFR2 mutations drive acquired resistance to FGFR inhibition in patients with FGFR2 fusion positive cholangiocarcinoma. Cancer discovery **7**, 1-12 (2016).
[14] Distribution of In Vitro Diagnostic Products Labeled for Research Use Only or Investigational Use Only: Guidance for Industry and Food and Drug Administration Staff. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/distribution-vitro-diagnostic-productslabeled-research-use-only-or-investigational-use-only
[15] https://www.fda.gov/medical-devices/device-approvals-denials-and-clearances/510k-clearances
[16] https://www.fda.gov/medical-devices/premarket-submissions/de-novo-classification-request
[17] https://www.fda.gov/medical-devices/premarket-submissions/premarket-approval-pma
[18] Dietel, M. et al. Diagnostic procedures for non-small-cell lung cancer (NSCLC): recommendations of the European Expert Group. Thorax **71**, 177-184 (2015).
[19] Herbst, R. S. et al. Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised controlled trial. The Lancet **387**, 1540-1550 (2016).
[20] Vansteenkiste, J. et al. Final results of a multi-center, double-blind, randomized, placebo-controlled phase II study to assess the efficacy of MAGE-A3 immunotherapeutic as adjuvant therapy in stage IB/II non-small cell lung cancer (NSCLC). Journal of clinical oncology **25**, 7554-7554 (2007).
[21] Gettinger, S. N. et al. Overall survival and long-term safety of nivolumab (anti-programmed death 1 antibody, BMS-936558, ONO-4538) in patients with previously treated advanced non-small-cell lung cancer. Journal of clinical oncology, **33**, 2004-2012 (2015).
[22] Burstein, H. J. et al. Clinical cancer advances 2017: annual report on progress against cancer from the American Society of Clinical Oncology. Journal of Clinical Oncology **35**, 1341-1367 (2017).
[23] Wender, R.et al. American Cancer Society lung cancer screening guidelines. CA: a cancer journal for clinicians, **63**, 106-117 (2013).
[24] Baldwin, D. R.et al. American Cancer Society lung cancer screening guidelines.UK Lung Screen (UKLS) nodule management protocol: modelling of a single screen randomised controlled trial of low-dose CT screening for lung cancer. Thorax **66**, 308-313 (2011).
[25] Field, J. K.et al. International association for the study of lung cancer computed tomography screening workshop 2011 report. Journal of Thoracic Oncology **7**, 10-19 (2012).
[26] Zhou, C.et al. Erlotinib versus chemotherapy as first-line treatment for patients with advanced EGFR mutation-positive non-small-cell lung cancer. The lancet oncology **12**, 735-742 (2011).
[27] Gatzemeier, U.et al. Results of a phase III trial of erlotinib (OSI-774) combined with cisplatin and gemcitabine (GC) chemotherapy in advanced non-small cell lung cancer (NSCLC). Journal of Clinical Oncology **22**, 7010-7010 (2004).
[28] Spigel, D. R.et al. Final efficacy results from OAM4558g, a randomized phase II study evaluating MetMAb or placebo in combination with erlotinib in advanced NSCLC. Journal of Clinical Oncology **29**, 7505-7505 (2011).
[29] Mok, T. S.et al. Osimertinib or platinum-pemetrexed in EGFR T790M positive lung cancer. New England Journal of Medicine **376**, 629-640 (2017).
[30] Oxnard, G. R.et al. Association between plasma genotyping and outcomes of treatment with osimertinib (AZD9291) in advanced non-small-cell lung cancer. Journal of clinical oncology **34**, 3375-3382 (2016).
[31] https://clinicaltrials.gov/ Accessed August 21, 2018.
[32] Clark, T. A. et al. Analytical validation of a hybrid capture-based next-generation sequencing clinical assay for genomic profiling of cell-free circulating tumor DNA. The Journal of Molecular Diagnostics **20**, 686-702 (2018).
[33] Lanman, R. B. et al. Analytical and clinical validation of a digital sequencing panel for quantitative, highly accurate evaluation of cell-free circulating tumor DNA. PloS One **10**, e0140712 (2015).
[34] Paz-Ares, L. et al. plus chemotherapy for squamous non–small-cell lung cancer. New England Journal of Medicine **379**, 2040-2051 (2018).

[35] Fabrizio, D. et al. A blood-based next-generation sequencing assay to determine tumor mutational burden (bTMB) is associated with benefit to an anti-PD-L1 inhibitor, atezolizumab. Cancer Research **78**, 5706-5706 (2018).
[36] Hellmann, M. D., Ciuleanu, T. E., Pluzanski, A., Lee, J. S., Otterson, G. A., Audigier-Valette, C., & Paz-Ares, L. (2018). Nivolumab plus ipilimumab in lung cancer with a high tumor mutational burden. New England Journal of Medicine, **378**, 2093-2104 (2018).
[37] Gandara, D. R., Paul, S. M., Kowanetz, M., Schleifman, E., Zou, W., Li, Y., & Shames, D. S. Blood-based tumor mutational burden as a predictor of clinical benefit in non-small-cell lung cancer patients treated with atezolizumab. Nature medicine **24**, 1441-1448 (2018).
[38] Le, D. T. et al PD-1 blockade in tumors with mismatch-repair deficiency. New England Journal of Medicine **372**, 2509-2520 (2015).
[39] Le, D. T. et al. Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. Science **357**, 409-413 (2017).
[40] Rolfo, C.et al. Liquid biopsies in lung cancer: the new ambrosia of researchers. Biochimica et Biophysica Acta (BBA)-Reviews on Cancer **1846**, 539-546 (2014).
[41] Xiong, L. et al. Dynamics of EGFR mutations in plasma recapitulates the clinical response to EGFR-TKIs in NSCLC patients. Oncotarget, **8**, 63846-63856 (2017).
[42] Romano, G.et al. A preexisting rare PIK3CAE545K subpopulation confers clinical resistance to MEK plus CDK4/6 inhibition in NRAS melanoma and is dependent on S6K1 signaling. Cancer discovery **8**, 556-567 (2018).
[43] Goldberg, S. B. et al. Early assessment of lung cancer immunotherapy response via circulating tumor DNA. Clinical Cancer Research **24**, 1872-1880 (2018).
[44] Narayan, A. et al. Ultrasensitive measurement of hotspot mutations in tumor DNA in blood using error-suppressed multiplexed deep sequencing. Cancer research **72**, 3492-3498 (2012).
[45] Thress, K. S.et al. Acquired EGFR C797S mutation mediates resistance to AZD9291 in non-small cell lung cancer harboring EGFR T790M. Nature medicine, **21**, 560-562 (2015).
[46] Kwong, L. N. et al. Oncogenic NRAS signaling differentially regulates survival and proliferation in melanoma. Nature medicine **18**, 1503-1510 (2012).
[47] Wang, Z. et al. Lung adenocarcinoma harboring EGFR T790M and in trans C797S responds to combination therapy of first-and third-generation EGFR TKIs and shifts allelic configuration at resistance. Journal of Thoracic Oncology **12**, 1723-1727 (2017).
[48] Corcoran, R. B., & Chabner, B. A. Application of cell-free DNA analysis to cancer treatment. New England Journal of Medicine **379**, 1754-1765 (2018).
[49] Garcia-Murillas, I. et al. Mutation tracking in circulating tumor DNA predicts relapse in early breast cancer. Science translational medicine, **7**, 302ra133-302ra133 (2015).
[50] Dawson, S. J. et al. Analysis of circulating tumor DNA to monitor metastatic breast cancer. New England Journal of Medicine **368**, 1199-1209 (2013).
[51] Reinert, T. et al. Analysis of plasma cell-free DNA by ultradeep sequencing in patients with stages I to III colorectal cancer. JAMA oncology **5**, 1124-1131 (2019).
[52] Tie, J.et al. Circulating tumor DNA analysis detects minimal residual disease and predicts recurrence in patients with stage II colon cancer. Science translational medicine **8**, 346ra92-346ra92 (2016).
[53] Abbosh, C., Birkbak, N. J., & Swanton, C. Early stage NSCLC—challenges to implementing ctDNA-based screening and MRD detection. Nature reviews Clinical oncology **15**, 577-586 (2018).
[54] Oxnard, G. R.,et al. Noninvasive detection of response and resistance in EGFR-mutant lung cancer using quantitative next-generation genotyping of cell-free plasma DNA. Clinical cancer research **20**, 1698-1705 (2014).
[55] Phallen, J.,et al. Direct detection of early-stage cancers using circulating tumor DNA. Science translational medicine **9**, eaan2415 (2017).
[56] Lebofsky, R.,et al. Circulating tumor DNA as a non?invasive substitute to metastasis biopsy for tumor genotyping and personalized medicine in a prospective trial across all tumor types. Molecular oncology **9**, 783-790. (2015)
[57] Imperiale, T. F. et al. Multitarget stool DNA testing for colorectal-cancer screening. New England Journal of Medicine **370**, 1287-1297 (2014).
[58] Chan, K. A., Woo, J. K., King, A., Zee, B. C., Lam, W. J., Chan, S. L., & Lo, Y. M. D. Analysis of plasma Epstein-Barr virus DNA to screen for nasopharyngeal cancer. New England Journal of Medicine **377**, 513-522 (2017).
[59] Cohen, J. D. et al Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science **359**, 926-930 (2018).
[60] Lecomte, T. et al. Detection of free circulating tumor associated DNA in plasma of colorectal cancer patients and its association with prognosis. New England Journal of Medicine **370**, 1287-1297 (2014).
[61] Chan, K. A. et al. Analysis of plasma Epstein-Barr virus DNA to screen for nasopharyngeal cancer. New England Journal of Medicine **377**, 513-522 (2017).
[62] Hu, Y. *et al.* False-positive plasma genotyping due to clonal hematopoiesis. Clinical Cancer Research, **24**, 4437-4443 (2018).
[63] Heitzer, E., Haque, I. S., Roberts, C. E., & Speicher, M. R. Current and future perspectives of liquid biopsies in genomics-driven oncology. Nature Reviews Genetics, **20**, 71-88 (2019).
[64] Sina, A. A. I. et al. Epigenetically reprogrammed methylation landscape drives the DNA self-assembly and serves as a universal cancer biomarker. Nature communications **9**, 4915 (2018).
[65] Lehmann-Werman, R. et al. Identification of tissue-specific cell death using methylation patterns of circulating DNA. Proceedings of the National Academy of Sciences **113**, E1826-E1834 (2016).
[66] Deng, J. et al. Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. Nature Biotechnology **27**, 353-360 (2009).
[67] Xu, R. H. et al. Circulating tumour DNA methylation markers for diagnosis and prognosis of hepatocellular carcinoma. Nature Materials **16**, 1155-1161 (2017).

[68] Teschendorff, A. E., & Relton, C. L. Statistical and integrative system-level analysis of DNA methylation data. Nature Reviews Genetics, **19**, 129-147 (2018).
[69] Guo, S., Diep, D., Plongthongkum, N., Fung, H. L., Zhang, K., & Zhang, K. Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. Nature genetics **49**, 635-642 (2017).
[70] https://clinicaltrials.gov/ct2/show/NCT02418234, https://clinicaltrials.gov/ct2/show/NCT00730158, https://clinicaltrials.gov/ct2/show/NCT01349959
[71] Tug, S. et al. Exercise-induced increases in cell free DNA in human plasma originate predominantly from cells of the haematopoietic lineage. Exercise immunology review **21**, 164-173 (2015).
[72] Moreira, V. G., Prieto, B., Rodr.ıguez, J. S. M., & Alvarez, F. V. Usefulness of cell-free plasma DNA, procalcitonin and C-reactive protein as markers of infection in febrile patients. Annals of clinical biochemistry **47**, 253-258 (2010).
[73] Burnham, P.et al. Urinary cell-free DNA is a versatile analyte for monitoring infections of the urinary tract. Nature communications **9**, 2412 (2018).
[74] Siljan, W. W. et al. Circulating cell-free DNA is elevated in community acquired bacterial pneumonia and predicts short-term outcome. Journal of Infection **73**, 383-386 (2016).
[75] Lindahl, T. Instability and decay of the primary structure of DNA. Nature **362**, 709-715 (1993).
[76] Richter, C., Park, J. W., & Ames, B. N. Normal oxidative damage to mitochondrial and nuclear DNA is extensive. Proceedings of the National Academy of Sciences **85**, 6465-6467 (1988).
[77] Cooke, M. S., Evans, M. D., Dizdaroglu, M., & Lunec, J. Oxidative DNA damage: mechanisms, mutation, and disease. The FASEB Journal **17**, 1195-1214 (2003).
[78] Norton, S. E., Lechner, J. M., Williams, T., & Fernando, M. R. A stabilizing reagent prevents cell-free DNA contamination by cellular DNA in plasma during blood sample storage and shipping as determined by digital PCR. Clinical biochemistry **46**, 1561-1565 (2013).
[79] Manning, J. E. in Emergency Medicine: A Comprehensive Study Guide. (ed. Tintinalli J. E.) 227 (McGraw-Hill: New York, 2004).
[80] 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. Clinical Chemistry **46**, 1078-1084 (2000).
[81] Koide, K. et al. Fragmentation of cell-free fetal DNA in plasma and urine of pregnant women. Prenatal Diagnosis **25**, 604-607 (2005).
[82] Tani, M., & Beck, S. Epigenome-wide association studies for cancer biomarker discovery in circulating cell-free DNA: technical advances and challenges. Current opinion in genetics development **42**, 48-55 (2017).
[83] Reckamp, K. L. et al. A highly sensitive and quantitative test platform for detection of NSCLC EGFR mutations in urine and plasma. Journal of Thoracic Oncology **11**, 1690-1700 (2016).
[84] Swanson, P.et al. Performance of the automated Abbott RealTime HIV-1 assay on a genetically diverse panel of specimens from London: comparison to VERSANT HIV-1 RNA 3.0, AMPLICOR HIV-1 MONITOR v1. 5, and LCx^A R○ HIV RNA Quantitative assays. Journal of virological methods **137**, 184-192 (2006).
[85] Castle, P. E. Performance of carcinogenic human papillomavirus (HPV) testing and HPV16 or HPV18 genotyping for cervical cancer screening of women aged 25 years and older: a subanalysis of the ATHENA study. The lancet oncology **12**, 880-890 (2011).
[86] Sandri, M. T. et al. Comparison of the Digene HC2 assay and the Roche AMPLICOR human papillomavirus (HPV) test for detection of high-risk HPV genotypes in cervical samples. Journal of clinical microbiology **44**, 2141-2146 (2006).
[87] Misawa, Y. et al. Application of loop-mediated isothermal amplification technique to rapid and direct detection of methicillin-resistant Staphylococcus aureus (MRSA) in blood cultures. Journal of Infection and Chemotherapy **13**, 134-140 (2007).
[88] Ethridge, S. F. et al. Performance of the Aptima HIV-1 RNA qualitative assay with 16-and 32-member specimen pools. Journal of clinical microbiology **48**, 3343-3345 (2010).
[89] https://www.fda.gov/drugs/informationondrugs/approveddrugs/ucm279174.htm
[90] https://www.mycancergenome.org/
[91] https://cancer.sanger.ac.uk/cosmic
[92] http://www.cbioportal.org/
[93] Pao, W., *et al*. Acquired resistance of lung adenocarcinomas to gefitinib or erlotinib is associated with a second mutation in the EGFR kinase domain. PLoS med, **2**, e73 (2005).
[94] Mok, T. S., *et al.* Osimertinib or platinum–pemetrexed in EGFR T790M–positive lung cancer. New England Journal of Medicine **376**, 629-640 (2017).
[95] Hindson, B. J.et al. High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. Analytical chemistry **83**, 8604-8610 (2011).
[96] Newton, C. R., Graham, A., Heptinstall, L. E., Powell, S. J., Summers, C., Kalsheker, N., & Markham, A. F. Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS). Nucleic acids research **17**, 2503-2516 (1989).
[97] Das, J. et al. An electrochemical clamp assay for direct, rapid analysis of circulating nucleic acids in serum. Nature chemistry **7**, 569-575 (2015).
[98] Lin, M., et al. Electrochemical detection of nucleic acids, proteins, small molecules and cells using a DNA-nanostructure-based universal biosensing platform. Nature Protocols **11**, 1244-1263 (2016).
[99] Gootenberg, J. S.et al. Nucleic acid detection with CRISPR-Cas13a/C2c2. Science, eaam9321 (2017).
[100] Lin, M. et al. Electrochemical detection of nucleic acids, proteins, small molecules and cells using a DNA-nanostructure-based universal biosensing platform. Nature Protocols, **11**, 1244-1263 (2016).
[101] 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. Proceedings of the National Academy of Sciences **100**, 8817-8822 (2003).

[102] Khoo, C.et al. Molecular methods for somatic mutation testing in lung adenocarcinoma: EGFR and beyond. Translational lung cancer research **4**, 126-141 (2015).
[103] Ragoussis, J. Genotyping technologies for genetic research. Annual review of genomics and human genetics **10**, 117-133 (2009).
[104] Murtaza, M. et al. Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. Nature **497**, 108-112 (2013).
[105] Wetterstrand, K. A. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcostsdata. Accessed July 19, 2018.
[106] Laver, T. et al. Assessing the performance of the oxford nanopore technologies minion. Biomolecular Detection and Quantification **3**, 1-8 (2015).
[107] Carneiro, M. O. et al. Pacific biosciences sequencing technology for genotyping and variation discovery in human data. BMC genomics **13**, 375 (2012).
[108] Taylor, A. D., Micheel, C. M., Anderson, I. A., Levy, M. A., & Lovly, C. M. The path (way) less traveled: a pathway-oriented approach to providing information about precision cancer medicine on My Cancer Genome. Translational oncology, **9**, 163-165 (2016).
[109] Diehn, M. et al. Early prediction of clinical outcomes in resected stage II and III colorectal cancer (CRC) through deep sequencing of circulating tumor DNA (ctDNA). J. Clinical Oncology **35**,3591-3591 (2017).
[110] Couraud, S. et al. Non-invasive diagnosis of actionable mutations by deep sequencing of circulating-free DNA in non-small cell lung cancer: Findings from BioCAST/IFCT-1002. Clinical Cancer Research **20**, 4613-24 (2014).
[111] Song, P. et al. Selective multiplexed enrichment for the detection and quantitation of low-fraction DNA variants via low-depth sequencing. *Nature Biomedical Engineering*, 1-12 (2021).
[112] Meyer, M., & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harbor Protocols, **2010**, pdb-prot5448, (2010).
[113] Hovelson, D. H.et al. Development and validation of a scalable next-generation sequencing system for assessing relevant somatic variants in solid tumors. Neoplasia **17**, 385-399, (2015).
[114] Dupuis, J. R.et al. HiMAP: robust phylogenomics from highly multiplexed amplicon sequencing. Molecular ecology resources **18**, 1000-1019 (2018).
[115] Potapov, V., & Ong, J. L. Examining sources of error in PCR by single-molecule sequencing. PloS One **12**, e0181128 (2017).
[116] Schirmer, M. et al. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Research **43**, e37-e37 (2015).
[117] Schirmer, M., Damore, R., Ijaz, U. Z., Hall, N., & Quince, C. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. BMC bioinformatics **17**, 125 (2016).
[118] Minoche, A. E., Dohm, J. C., & Himmelbauer, H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. Genome Biology **12**, R112 (2011).
[119] Jain, M. et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nature Biotechnology **36**, 338-345, (2018).
[120] Quail, M. A. et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC genomics **13**, 1-13 (2012).
[121] Schmitt, M. W. et al. Detection of ultra-rare mutations by next-generation sequencing. Proceedings of the National Academy of Sciences **109**, 14508-14513 (2012).
[122] Blakely, C. M. et al. Evolution and clinical impact of co-occurring genetic alterations in advanced-stage EGFR-mutant lung cancers. Nature Genetics **49**, 1693-1704 (2017).
[123] Kinde, I. et al. Detection and quantification of rare mutations with massively parallel sequencing. Proceedings of the National Academy of Sciences **108**, 9530-9535 (2011).
[124] Lo, Y. D. et al. Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. Science Translational Medicine **2**, 61ra91-61ra91 (2010).
[125] Thierry, A. R. et al. Origin and quantification of circulating DNA in mice with human colorectal cancer xenografts. Nucleic Acids Research **38**, 6159-6175 (2010).
[126] Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M., & Shendure, J. Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. Cell **164**, 57-68 (2016).
[127] Mouliere, F. et al Enhanced detection of circulating tumor DNA by fragment size analysis. Science translational medicine **10**, 466 (2018).
[128] Thakur, B. K., Zhang, H., Becker, A., Matei, I., Huang, Y., Costa-Silva, B., & Williams, C. Double-stranded DNA in exosomes: a novel biomarker in cancer detection. Cell research, **24**, 766-769 (2014).
[129] Balaj, L. et al Tumour microvesicles contain retrotransposon elements and amplified oncogene sequences. Nature communications, **2**, 180 (2011).
[130] Zhang, J. X. et al. Predicting DNA hybridization kinetics from sequence. Nature Chemistry **10**, 91-98 (2018).
[131] Newman, A. M. et al. Integrated digital error suppression for improved detection of circulating tumor DNA. Nature Biotechnology **34**, 547-555 (2016).
[132] Narayan, A. et al. Ultrasensitive measurement of hotspot mutations in tumor DNA in blood using error-suppressed multiplexed deep sequencing. Cancer Research **72**, 3492-3498 (2012).
[133] Kou, R. et al. Benefits and challenges with applying unique molecular identifiers in next generation sequencing to detect low frequency mutations. PLoS One **11**, e0146638 (2016).
[134] Jabara, C. B., Jones, C. D., Roach, J., Anderson, J. A., & Swanstrom, R. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. Proceedings of the National Academy of Sciences **108**, 20166-20171 (2011).
[135] Abascal, F.et al. Somatic mutation landscapes at single-molecule resolution. Nature, 1-6 (2021).

[136] Cohen, J. D. et al. Detection of low-frequency DNA variants by targeted sequencing of the Watson and Crick strands. Nature Biotechnology, 1-8 (2021).
[137] Pel, J., et al. Nonlinear electrophoretic response yields a unique parameter for separation of biomolecules. Proceedings of the National Academy of Sciences **106**, 14796-14801 (2009).
[138] Kidess, E. et al. Mutation profiling of tumor DNA from plasma and tumor tissue of colorectal cancer patients with a novel, high-sensitivity multiplexed mutation detection platform. Oncotarget **6**, 2549 (2015).
[139] Song, C., Liu. et al. Elimination of unaltered DNA in mixed clinical samples via nuclease-assisted minor-allele enrichment. Nucleic acids research **44**, e146-e146. (2016).
[140] Seyama, T. et al. A novel blocker-PCR method for detection of rare mutant alleles in the presence of an excess amount of normal DNA. Nucleic Acids Research **20**, 2493-2496 (1992).
[141] Arcila, M., Lau, C., Nafa, K., & Ladanyi, M. Detection of KRAS and BRAF mutations in colorectal carcinoma: Roles for high-sensitivity locked nucleic acid-PCR sequencing and broad-spectrum mass spectrometry genotyping. The Journal of Molecular Diagnostics **13**, 64-73 (2011).
[142] Orum, H. et al. Single base pair mutation analysis by PNA directed PCR clamping. Nucleic Acids Research **21**, 5332-5336 (1993).
[143] Milbury, C. A., Li, J., & Makrigiorgos, G. M. Ice-COLD-PCR enables rapid amplification and robust enrichment for low-abundance unknown DNA mutations. Nucleic acids research **39**, e2 (2011).
[144] Li, J. et al. Replacing PCR with COLD-PCR enriches variant DNA sequences and redefines the sensitivity of genetic testing. Nature Medicine **14**, 579-584 (2008).
[145] Zuo, Z.et al. Application of COLD-PCR for improved detection of KRAS mutations in clinical samples. Modern pathology, **22**, 1023-1031 (2009).
[146] Wu, L. R. et al. Multiplexed enrichment of rare DNA variants via sequence-selective and temperature-robust amplification. Nature Biomedical Engineering **1**, 714-723 (2017).
[147] Ciriello, G. et al. Emerging landscape of oncogenic signatures across human cancers. Nature Genetics, **45**, 1127-1133 (2013).
[148] Leary, R. J. et al. Development of personalized tumor biomarkers using massively parallel sequencing. Science translational medicine **2**, 20ra14-20ra14 (2010).
[149] Snyder, M. W.et al. Copy-number variation and false positive prenatal aneuploidy screening results. New England Journal of Medicine **372**, 1639-1645 (2015).
[150] Mertens, F., Johansson, B., Fioretos, T., & Mitelman, F. The emerging complexity of gene fusions in cancer. Nature Reviews Cancer **15**, 371-381 (2015).
[151] Wang, Q. et al. Application of next generation sequencing to human gene fusion detection: computational tools, features and perspectives. Briefings in Bioinformatics **14**, 506-519 (2012).
[152] Maher, C. A. et al. Transcriptome sequencing to detect gene fusions in cancer. Nature 458, 97 (2009).
[153] Zhao, Q. et al. Transcriptome-guided characterization of genomic rearrangements in a breast cancer cell line. Proceedings of the National Academy of Sciences **106**, 1886-1891 (2009).
[154] Zack, T. I. et al. Pan-cancer patterns of somatic copy number alteration. Nature Genetics 45, 1134 (2013).
[155] Whale, A. S. et al. Comparison of microfluidic digital PCR and conventional quantitative PCR for measuring copy number variation. Nucleic Acids Research **40**, e82-e82 (2012).
[156] Pinkel, D. et al. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nature Genetics **20**, 207-211 (1998).
[157] Chiu, R. W. et al. Non-invasive prenatal assessment of trisomy 21 by multiplexed maternal plasma DNA sequencing: large scale validity study. BMJ **342**, c7401 (2011).
[158] Kops, G. J., Weaver, B. A., & Cleveland, D. W. On the road to cancer: aneuploidy and the mitotic checkpoint. Nature Reviews Cancer **5**, 773-785 (2004).
[159] Mandel, P. & Metais, P. Les acides nucleiques du plasma sanguin chez la homme. C. R. Seances Soc. Biol. Fil. **142**, 241-243 (1948).
[160] Bettegowda, C. et al. Detection of circulating tumor DNA in early-and late-stage human malignancies. Science translational medicine, **6**, 224ra224-224ra224 (2014).
[161] Cohen, J. D. et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science **359**, 926-930 (2018).
[162] Aravanis, A. M., Lee, M., & Klausner, R. D. Next-generation sequencing of circulating tumor DNA for early cancer detection. Cell **168**, 571-574 (2017).
[163] Razavi, P. et al. Cell-free DNA (cfDNA) mutations from clonal hematopoiesis: Implications for interpretation of liquid biopsy tests. J. Clinical Oncology **35**, 11526-11526 (2017).
[164] Genovese, G. et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. New England Journal of Medicine **371**, 2477-2487 (2014).
[165] Mouliere, F., & Rosenfeld, N. Circulating tumor-derived DNA is shorter than somatic DNA in plasma. Proceedings of the National Academy of Sciences **112**, 3178-3179 (2015).
[166] Schwarzenbach, H., Nishida, N., Calin, G. A., & Pantel, K. Clinical relevance of circulating cell-free microRNAs in cancer. Nature Reviews Clinical Oncology **11**, 145-156 (2014).
[167] Azmi, A. S., Bao, B., & Sarkar, F. H. Exosomes in cancer development, metastasis, and drug resistance: a comprehensive review. Cancer and Metastasis Reviews **32**, 623-642 (2013).
[168] Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M., & Shendure, J. Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. Cell **164**, 57-68 (2016).
[169] Cohen, J. D. et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science **359**, 926-930 (2018).
[170] De Mattos-Arruda, L. et al. Circulating tumour cells and cell-free DNA as tools for managing breast cancer. Nature reviews Clinical oncology **10**, 377-389, (2013).

[171] De Bono, et al. Circulating tumor cells predict survival benefit from treatment in metastatic castration-resistant prostate cancer. Clinical cancer research **14**, 6302-6309. (2008).

TABLE I: Current methods and challenges in cfDNA analysis

Challenge	Solution	Comments
DNA damage	DNA repair	Need high-yield method to reverse DNA oxidation and deamination
cfDNA sampling stochasticity	Larger blood volumes	Concern for patient health
cfDNA sampling stochasticity	Urine cfDNA	Need for extracting cfDNA from large volumes; process short cfDNA
Detecting mutations with ≤1% VAF	Digital PCR	Single-plex, only known mutations
Detecting mutations with ≤1% VAF	Mass spectrometry	Medium-plex, only known mutations
Detecting mutations with ≤1% VAF	NGS with UMIs	Expensive and low conversion yield
Detecting mutations with ≤1% VAF	NGS with allele enrichment	Low-plex, inaccurate quantitation, and low on-target rates
High conversion yield from cfDNA	N/A	Need high-yield end-repair and ligation
NGS depth uniformity	Primer/probe conc. tuning	labor-intensive and imperfect uniformity
Detecting fusions in cfDNA	Very large NGS panel	Very expensive because introns are long
Detecting CNVs in cfDNA	NGS, ddPCR	No current solution for detection ≤5% VAF
Rapid cfDNA diagnostics	Nanopore sequencing	High error rates and expensive reads
Affordable cfDNA diagnostics	N/A	Current cfDNA NGS panels have list price over $4,000

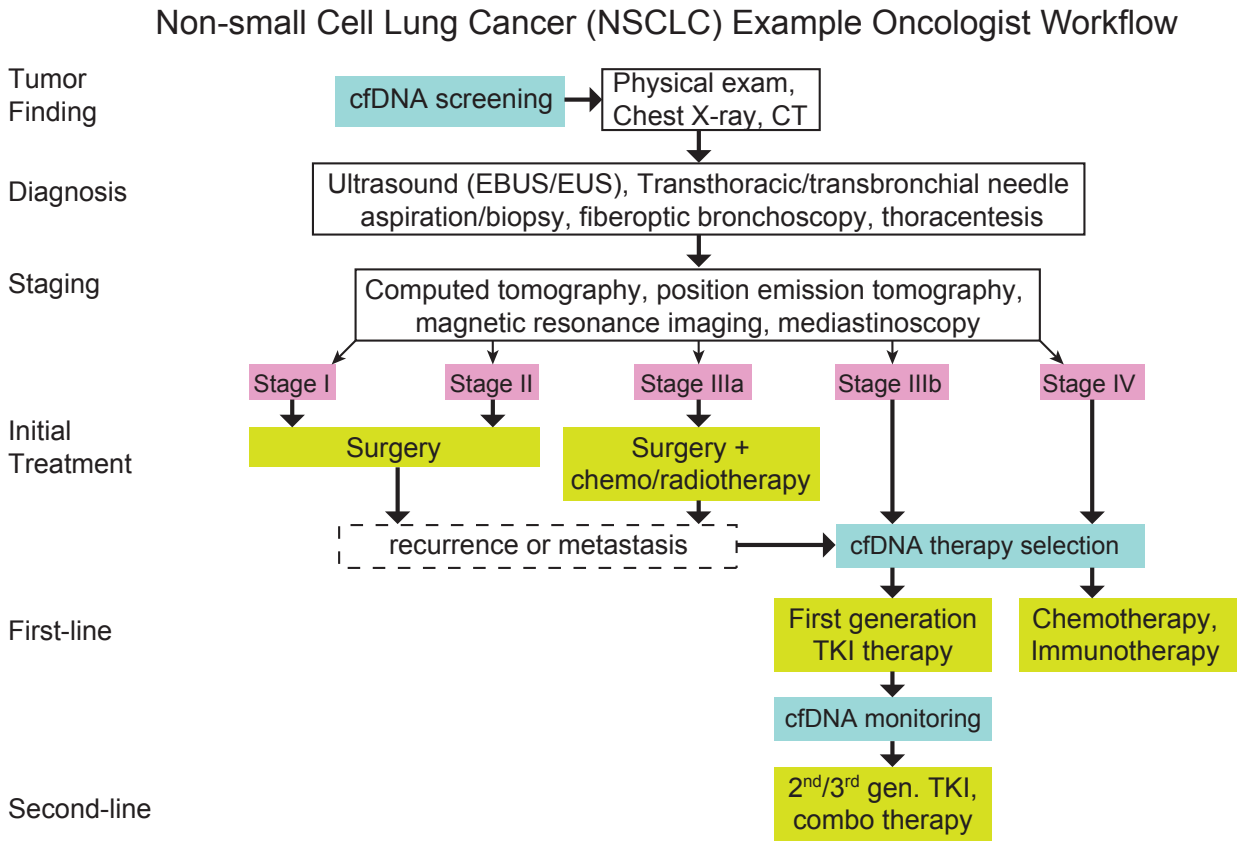

FIG. 1: Roles of cell-free DNA (cfDNA) tests in non-small cell lung cancer (NSCLC) clinical diagnostics and treatment workflow. cfDNA diagnostics can play in cancer care, we show NSCLC as an example. cfDNA screening can be used for tumor finding with combination of physical exam, Chest X-ray, and CT. 70% of NSCLC patients are diagnosed at late stage (III and IV) following overt clinical symptoms. The most common use of cfDNA analysis is in therapy selection for stage IIIb and IV patients. The technologies of cfDNA analysis can be used for post-treatment monitoring, including detection of recurrence and de novo resistance mutations.

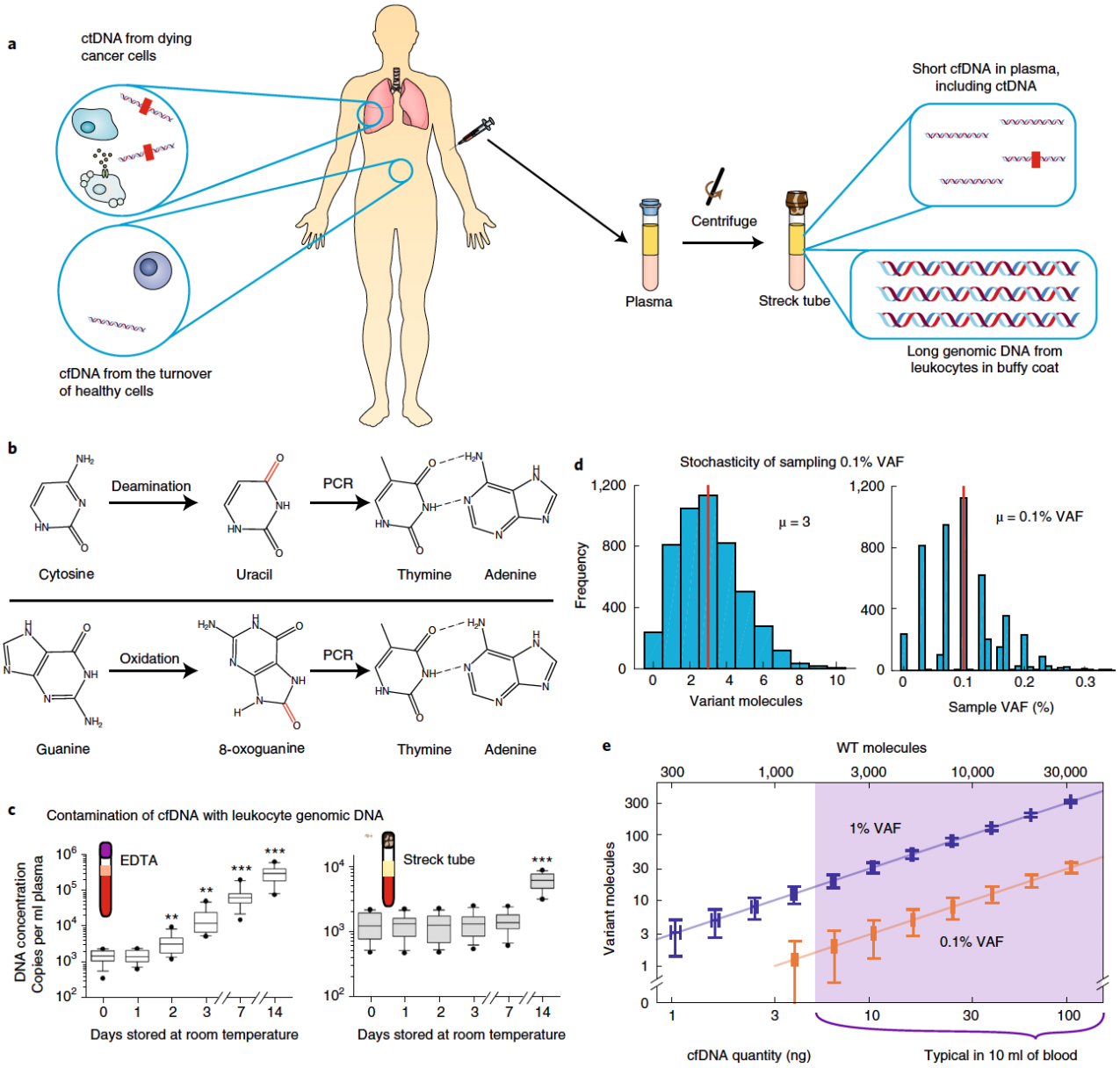

FIG. 2: Pre-analytical factors impacting the accuracy of cfDNA analysis. **(a)** Whereas the buffy coat layer of blood is rich in genomic DNA from peripheral mononuclear blood cells (PBMCs), blood plasma contains relatively low quantities of extracellular DNA, known as cfDNA. cfDNA is derived from dying cells from the entire body including both healthy cells (white) and from tumor cells (brown) dying from apoptosis, necrosis, and immune cytotoxicity. Thus, only a small fraction of cfDNA comprises tumor-derived circulating tumor DNA (ctDNA). The red box inside the dsDNA denotes tumor-specific mutations in ctDNA. **(b)** cfDNA in blood may be damaged during sample collection, transport, and storage, resulting in modified nucleosides that are incorrectly recognized by DNA polymerases during PCR amplification, resulting in amplicon DNA sequences with variants that may be interpreted as cancer-specific mutations. Illustrated here are cytosine deamination and guanine oxidation, the two most commonly observed types of DNA damage. **(c)** Contamination of cfDNA with genomic DNA from leukocytes. Except in the case of blood cancers, genomic DNA from leukocytes will not contain ancer-specific ctDNA. Thus, contamination of cfDNA with leukocyte genomic DNA will dilute the fraction of cfDNA that contain useful information, rendering downstream DNA mutation analysis more difficult and variant allele fraction (VAF) quantitation less accurate. Image adapted from [78]. **(d)** Poisson distribution of tumor mutation molecules and VAF in a blood sample. An adult human has roughly 5 L of blood in circulation, and sampling 10 mL of blood for cfDNA analysis introduces VAF significant variation due to small number statistics. Assuming a “ground truth” of 0.1% cancer mutation VAF in the entire 5 L blood supply and a 10 ng sample of cfDNA in a 4 mL plasma sample, the number of cancer mutation molecules present will range between 0 and 10, corresponding to an observed VAF range from 0% to 0.3% for any given DNA locus. No technology improvements can transcend this sampling variation; only the use of larger volume blood samples can mitigate this VAF irreproducibility challenge. Matlab code used to generate these results is available in the Supplementary Information. (e) Visualization of molecule number variations due to cfDNA sampling. The vertical and horizontal error bars show analytically calculated standard deviation values for different cfDNA input quantities and mutation VAFs.

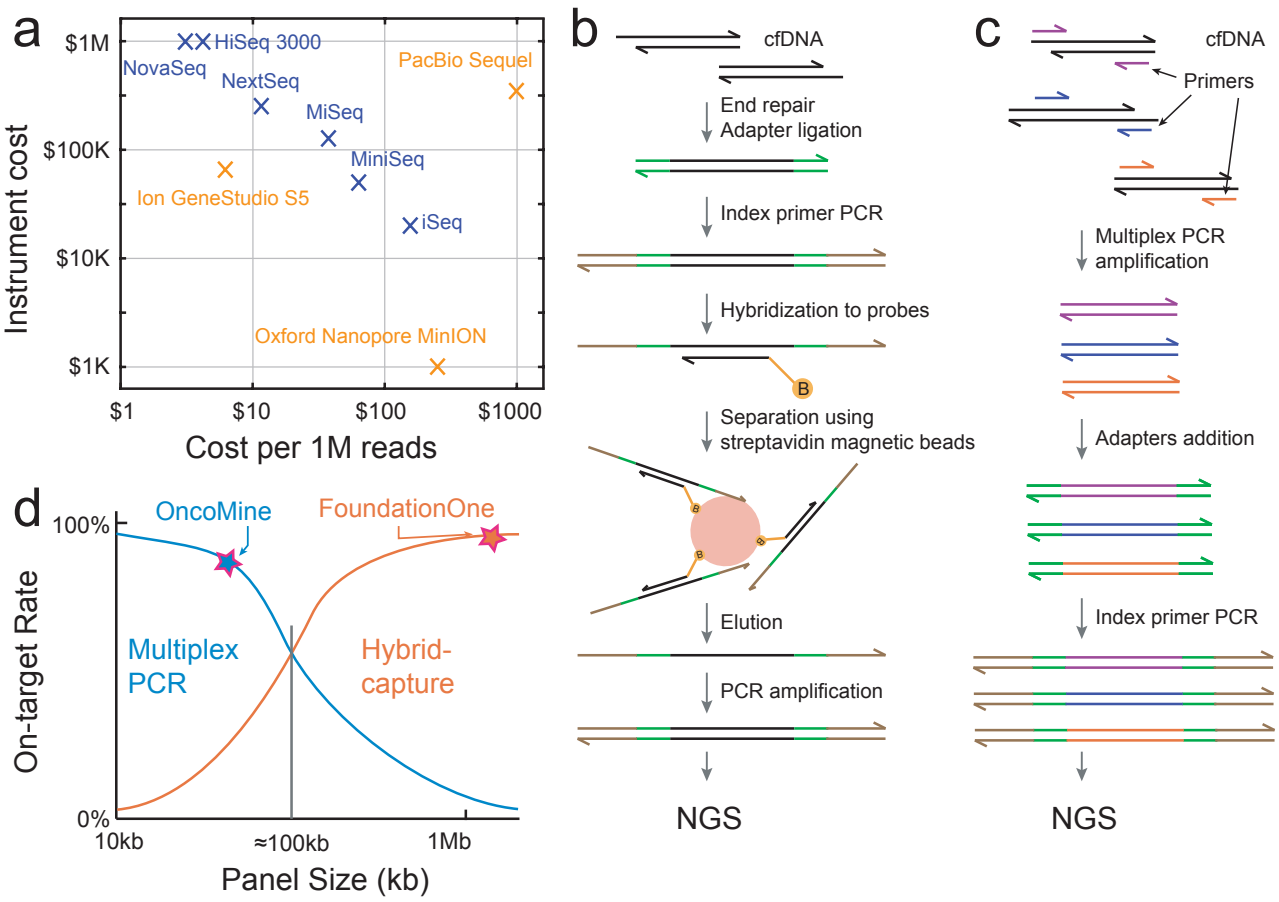

FIG. 3: NGS for cfDNA analysis. **(a)** Current sequencing platforms and their cost. **(b)** Example NGS library preparation workflow of target enrichment via ligation/hybrid-capture. **(c)** Example NGS library preparation workflow of target enrichment via multiplex PCR. **(d)** Comparison of hybrid-capture and multiplex PCR target enrichment in terms of on-target rate and panel size. The approximate panel sizes and on-target rates of 2 commercial panels, Oncomine (multiplex PCR) and Guardant360 (hybrid-capture) are displayed.

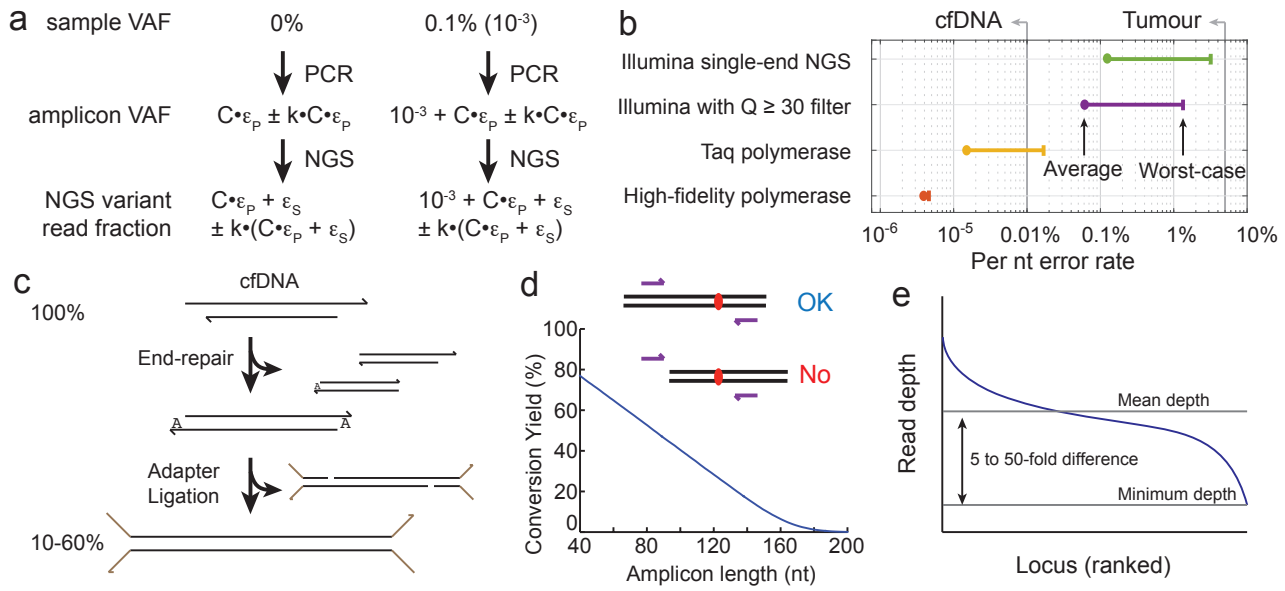

FIG. 4: Primary limitations of NGS-based cfDNA analysis. **(a)** DNA polymerase misincorporation errors and NGS intrinsic errors limit the mutation limit of detection. The limit of detection is defined as the lowest variant allele frequency (VAF) of a mutation that can be confidently distinguished from a purely wildtype sample (0% VAF). The PCR misincorporation rate (ε_P), the number of PCR cycles (C), and the NGS intrinsic error rate (ε_S) all increase the fraction of NGS reads that correspond to variant sequences for a 0% VAF sample. Due to variations in the error rates depending on experimental protocol minutiae, the actual fraction of NGS reads corresponding to variants will vary from run to run. Consequently, the combined error rate (C x ϵP + ϵS) should be significantly (e.g. 2-fold or more) lower than the limit of detection. **(b)** Typical error rates for PCR amplification and NGS [115-120]. All error rates exhibit some sequence bias; plotted here are average error rates and worst-case error rates. Single-pass NGS intrinsic errors are lowest for Illumina platforms; average sequencing error rates for single-pass sequencing by Ion Torrent, Pacific Biosciences, and Oxford Nanopore range from 1% to 20%. DNA polymerase error rates shown are per extension; through the cours e of a PCR reaction, the misincorporation rate is multiplied by the number of cycles. High-fidelity polymerases refer to enzymes with 3′ > 5′ proofreading capabilities; the three high-fidelity polymerases most frequently used for NGS are Phusion, NEB Q5, and KAPA HiFi. The vertical lines labelled cfDNA and Tumour indicate the currently achievable VAF limit of detections based on NGS. **(c)** Imperfect end-repair and ligation efficiency limits the conversion yield of cfDNA for ligation hybrid-capture protocols. The conversion yield is the fraction of the original cfDNA molecules represented as amplicons in the NGS library. For ligation hybrid-capture workflows, conversion yield is primarily bottlenecked by ligation efficiency, and secondarily by DNA extraction and DNA end-repair. We note that conversion yields listed in literature are typically high-end estimates, because there are different ways for estimating total quantity of input cfDNA that can differ by a factor of 2 or more. **(d)** cfDNA breakpoints limit the conversion yield of cfDNA for multiplex PCR protocols. Long amplicons have a high probability of not being able to amplify the original cfDNA molecule of interest, due to the original molecule not spanning the nucleotides of the desired amplicon. **(e)** Non-uniformity increases total NGS reads needed to ensure a minimum depth needed for achieving a defined limit of detection. The mean NGS read depth can be calculated as the total NGS reads multiplied by the on-target rate and divided by the amplicons/loci; however, the minimum depth can be a factor of 5 to 50 lower than the mean depth. Because sequencing depth limits analytical sensitivity to low VAF mutations, some mutations will have worse (higher) VAF limits of detection than others. Matlab code used to generate these results is available in the Supplementary Information.

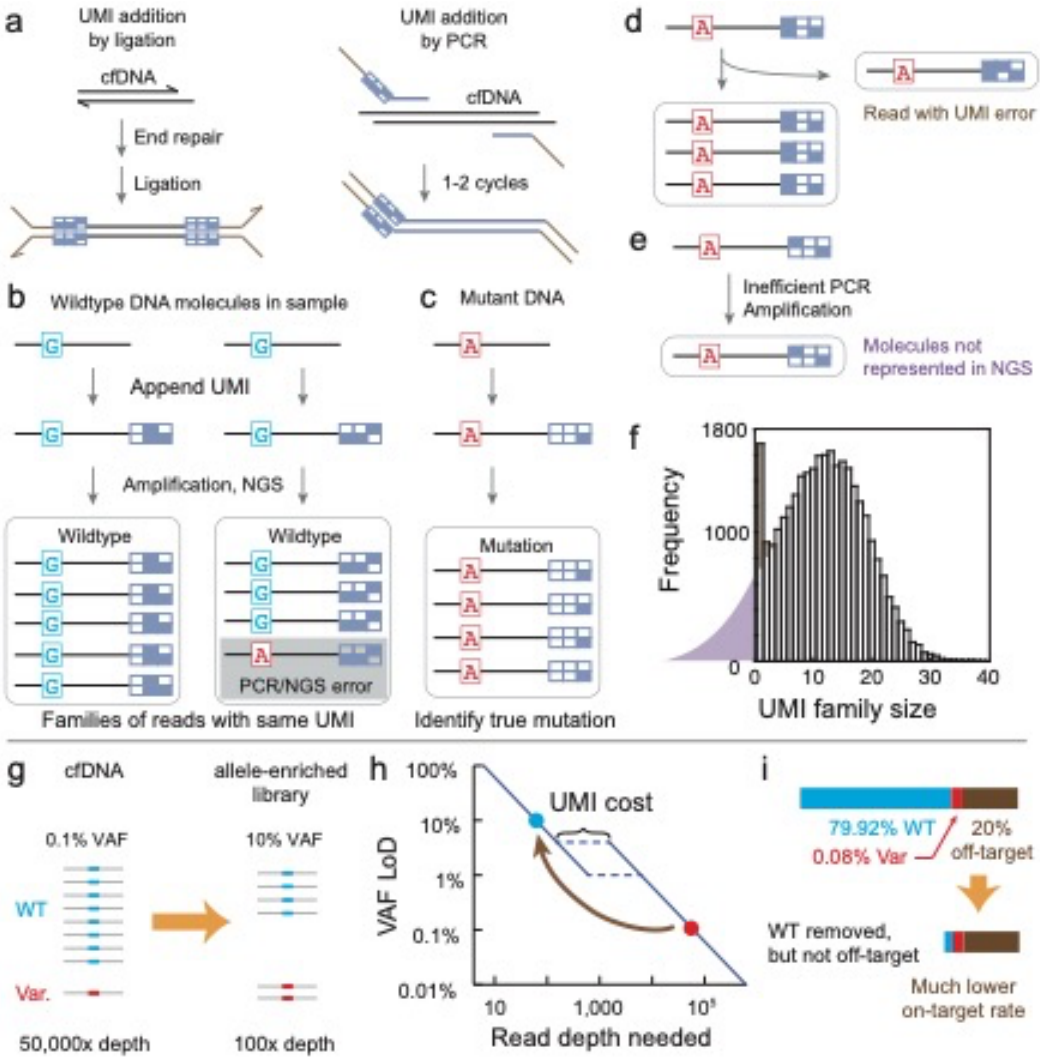

FIG. 5: Methods for the accurate detection of mutations with ≤1% in VAF. **(a)** Unique molecular identifiers (UMIs) are used to overcome the limits of detection imposed by PCR misincorporation errors and NGS intrinsic errors. UMIs are unique sequences that are attached to each original cfDNA molecule, and here displayed as a 2D barcode. UMIs are typically randomer degenerate sequences such as "NNNNNN," but can also comprise specific designed DNA sequences with error-correction or error-detection properties. As illustrated here, UMIs can be incorporated into both ligation/hybrid-capture and multiplex PCR protocols. Adapter sequences are shown in brown. **(b)** UMIs correct most PCR and NGS errors. In the absence of UMIs, the NGS results showing 1 mutant A (red) read and 8 wildtype G (cyan) reads may suggest that the A mutation has a VAF of 11%. By sorting different reads by UMIs, we can bioinformatically determine that all 9 reads were derived from 2 original DNA molecules in the sample, which are both likely to be wildtype in sequence. However, there is a small chance that the right group of reads were actually derived from an early stage PCR error. **(c)** True mutations, on the other hand, will likely be represented by a family of reads with the same UMI that predominantly correspond to the same mutation sequence. **(d)** The UMI strategy is imperfect in error correction however, because PCR or NGS errors can occur within the UMI barcode sequence. The reads bearing UMI errors cannot be easily distinguished from a true family of UMIs corresponding to reads derived from another original cfDNA molecule. Thus, both the number of mutant and wild-type molecules may be overestimated. **(e)** Another limitation of UMIs is that different UMI sequences can have significant and unpredictable impact on PCR amplification efficiency, resulting in some molecules being poorly amplified and thus not well represented in the NGS library. This results in an effectively lower conversion yield than without UMIs, and can yield false negative results. **(f)** Typical distribution of UMI family sizes for an NGS library; these results are from a 3 Mb ligation/hybrid-capture panel. The median UMI family size is roughly 13, and the UMI family size roughly follows a normal distribution. The distribution of UMI family sizes suggest that a significant fraction (≈ 20%) of UMI families are not represented in the library (shaded in purple), due to the UMI amplification bias described in panel (e). Furthermore, the number of UMI families with size 1 and 2 is unusually high; the excess families (shaded in brown) are likely UMI errors described in panel (d). Because it is not possible to distinguish which of the UMI families of size 1 and 2 are UMI errors, a typical bioinformatic workflow will ignore all UMI families with size less than 3, resulting in an even greater loss of effective conversion yield. **(g)** Allele enrichment seeks to increase the representation of the NGS library by the variant alleles (e.g. cancer mutations). This is typically accomplished either through selective removal of wild-type alleles via probe hybridization [137] or enzymatic degradation [139], or selective PCR amplification of variant alleles [143, 146]. **(h)** NGS read depth required for different VAF sensitivities. There is a non-linear increase in depth required between 1% and 5% VAF sensitivity, due to the overhead required for UMIs to suppress NGS intrinsic error. Thus, enriching variant VAF from 0.1% to 10% can, in principle, reduce the required NGS reads by more than 500-fold. LoD, limit of detection. **(i)** The reads savings provided by allele enrichment is typically bottlenecked by on-target rate. By depleting the majority of wild-type reads in an NGS library, the relative fraction of off-target reads (e.g. nonspecific amplification of other genomic loci, primer dimers) becomes significantly higher in an allele-enriched library. In the illustrated case where the original library's on-target fraction is a reasonable 80%, the NGS reads saving is limited to a factor of 5, even if all on-target wild-type molecules/reads are perfectly removed. Thus, NGS libraries need to be close to 100% on-target rates in order to fully realize the potential of allele enrichment.

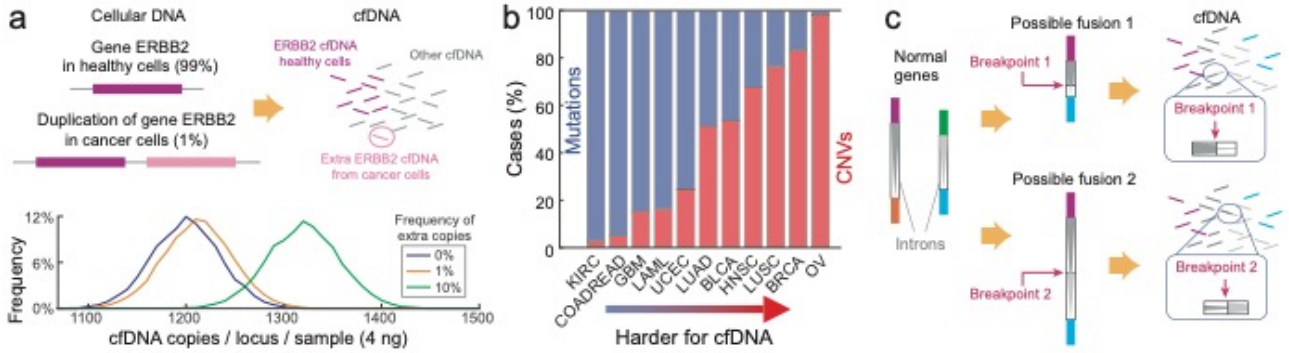

FIG. 6: CNVs and gene fusions are challenging biomarkers for cfDNA analysis. **(a)** Because the fraction of all cfDNA that is tumor-derived is frequently under 1%, the stochasticity of molecular sampling renders CNVs difficult to distinguish from regular samples. Plotted here are the expected distribution of number of molecules present per locus in a 4 ng sample of cfDNA for 0%, 1%, and 10% VAFs (based on 5000 stochastic simulations). Even for 10% VAF sample, the molecular count overlaps significantly with the 0% VAF sample, resulting in imperfect clinical sensitivity and specificity for a CNV call based on a single locus. Matlab code used to generate these results is available in the Supplementary Information. **(b)** The distribution of mutation vs. CNV markers vary drastically by cancer type, with some diseases such as ovarian cancer bearing almost exclusively CNV markers [147]. Thus, there is a pressing unmet need to reliably quantitate CNVs in cfDNA. Shown are TCGA abbreviations: KIRC = kidney renal clear cell carcinoma, COADREAD = colon adenocarcinoma and rectum adenocarcinoma, GBM = glioblastoma multiforme, LAML = acute myeloid leukemia, UCEC = uterine corpus endometrial carcinoma, LUAD = lung adenocarcinoma, BLCA = bladder urothelial carcinoma, HNSC = head and neck squamous cell carcinoma, LUSC = lung squamous cell carcinoma, BRCA = breast invasive carcinoma, OV = ovarian carcinoma. **(c)** Gene fusions [148] are difficult to detect in cfDNA because the fusion breakpoints can occur at any of a very large number of intron positions. Given the long lengths of introns and variable nature of fusion components, a very large (> 200kb) intron panel would be needed for high clinical sensitivity for a single fusion type (e.g. EML4-ALK). In contrast, fusions are more easily detected in RNA using tissue biopsies, because exon splicing results in a much smaller number of well-defined mature mRNA sequences.

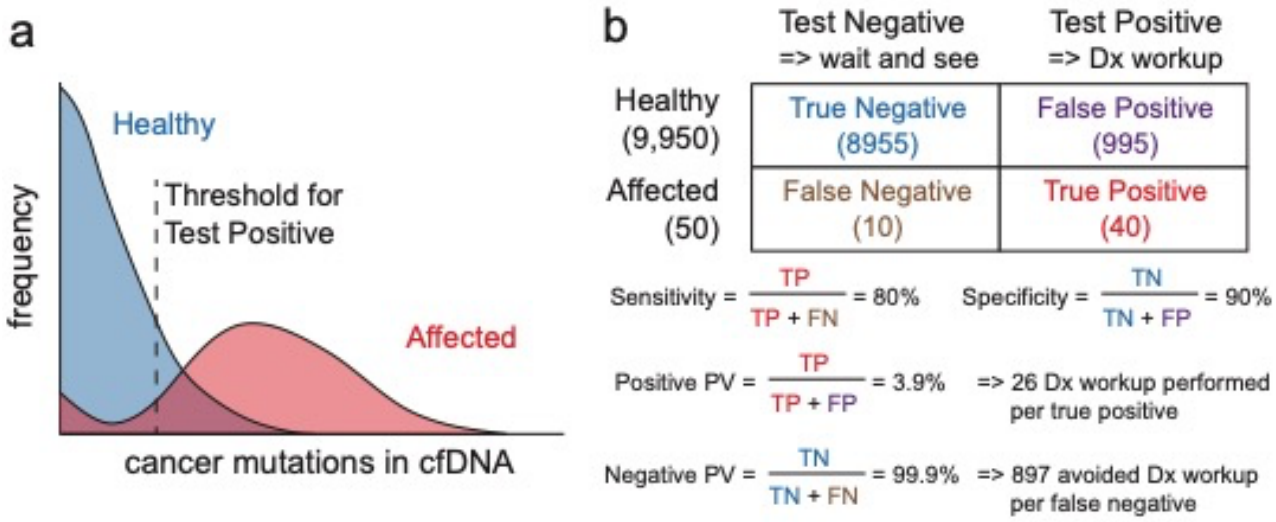

FIG. 7: Accuracy requirements for the screening of early cancers via cfDNA analysis. **(a)** Hypothetical distribution of cancer mutations in cfDNA for healthy and affected individuals. The test will report a positive when the observed mutations and VAF combinations exceed some threshold, and will have both false positives (healthy individuals above the threshold) and false negatives (affected individuals below the threshold). The threshold can be moved to change the tradeoff between clinical sensitivity and specificity. **(b)** Hypothetical test results for a test with 80% sensitivity and 90% specificity, for a tested population of 10,000 in which 0.5% of the population have early-stage cancer.