

Challenges in Using ctDNA to Achieve Early Detection of Cancer

Imran S Haque^{1*} and Olivier Elemento²

¹ Freenome, ² Weill Cornell Medicine; *Correspondence to research@freenome.com

Abstract

Early detection of cancer is a significant unmet clinical need. Improved technical ability to detect circulating tumor-derived DNA (ctDNA) in the cell-free DNA (cfDNA) component of blood plasma via next-generation sequencing and established correlations between ctDNA load and tumor burden in cancer patients have spurred excitement about the possibilities of detecting cancer early by performing ctDNA mutation detection.

We reanalyze published data on the expected ctDNA allele fraction in early-stage cancer and the population statistics of cfDNA concentration to show that under conservative technical assumptions, high-sensitivity cancer detection by ctDNA mutation detection will require either more blood volume (150-300mL) than practical for a routine screen or variant filtering that may be impossible given our knowledge of cancer evolution, and will likely remain out of economic reach for routine population screening without multiple-order-of-magnitude decreases in sequencing cost. Instead, new approaches that integrate ctDNA mutations with multiple other blood-based analytes (such as exosomes, circulating tumor cells, ctDNA epigenetics, metabolites) as well as integration of these signals over time for each individual may be needed.

Introduction

In the last decade, rapid technological development has dramatically improved our understanding of the underlying molecular basis of cancer [Vogelstein2013]. Such insights have contributed to the development of numerous molecularly targeted therapeutic agents and improved patient outcomes. Genomic tests are now used routinely to interrogate the genomes of patients with different diseases and identify targetable alterations. A frequent application of these tests is in cancer patients. Current estimates suggest that in the late-stage setting up to 50% of patients may receive genomic test results that alter their treatment [Blumenthal2016, LimaPereira2017]. However, the clinical utility of genomic tests remains controversial and is likely to evolve as new targets and their associated therapies emerge.

The finding that tumor-derived genomic alterations are detectable in “cell-free DNA” circulating in the plasma of patients with malignancy [Sorenson1994, Nawroz1996] has inspired the development of blood-based assays for tumor genomic profiling. In particular, sequencing of

circulating cell-free tumor DNA (ctDNA) has been employed as an adjunct to DNA derived from tissue biopsies to inform treatment decision-making with the first such test (Guardant360, Guardant Health) launched in 2014 [Lanman2015]. ctDNA can be used to monitor minimal residual disease following treatment [GarciaMurillas2015, Chaudhuri2017], as well as the emergence of resistant clones prior to clinical resistance detectable via imaging [Misale2012, Diaz2012].

Not surprisingly, there is significant interest in the potential utility of ctDNA for the early non-invasive detection of cancer [Table 1], with over 1 billion USD invested in companies developing such technologies in 2017 alone [CNBC2017, GlobeNewsWire2017]. Indeed, ctDNA is detectable in some patients with early-stage cancers [Bettegowda2014, Phallen2017], but assay sensitivity has so far precluded their use for screening. The motivation for early detection of tumors is clear. It has long been appreciated that earlier detection of malignancy results in significant reduction in cancer-specific mortality [Etzioni2003, Cho2014]. This has led to screening guidelines for breast cancer via mammography and colorectal cancer via colonoscopy or stool-based assays, but such approaches are cancer-type specific. Serum protein biomarkers such as carcinoembryonic antigen (CEA) and cancer antigen 125 (CA-125) are often used to monitor disease progression; however, generalizable biomarkers that detect cancer early have remained elusive.

Aravanis and colleagues [Aravanis2017] recently proposed that the fundamental limitations for such a ctDNA-based early detection test, beyond the current state-of-the-art, are a requirement for around 100x more sequencing bandwidth and improved variant interpretation. Subsequently, in the most comprehensive study of early stage cancers to date, Phallen et al. [Phallen2017] reported the detection of somatic alterations in 50-75% of patients depending on histology. While this represents a significant advance over prior reports, achieved through both deeper panel sequencing and improved variant calling via error correction, the sensitivity remains well below that required for broad clinical implementation. Despite the obvious shortfall of this and prior approaches, there has been limited examination of the feasibility of improving sensitivity to the levels required to demonstrate clinical utility. Here, we reanalyze published ctDNA sequencing data from early-stage cancer patients and assert that statistical and physiological limitations suggest that a ctDNA-based mutational assay for early detection would be neither commercially nor biologically viable.

Targeted ctDNA mutation-detection panels require infeasibly large input volumes for early detection

Detection of tumor-derived alleles in the blood can be modeled as a binomial process: we sequence a number of independent fragments of DNA, of which only a fraction (given by the

variant allele fraction, or proportion of cfDNA at a given locus carrying a tumor-derived mutation) will be derived from the tumor. In order to detect cancer, the most sensitive test that could be built would be one that reported “positive” upon the detection of a single fragment carrying a cancer-derived allele. However, tumors exhibit a remarkable degree of mutational heterogeneity and the initiating lesions are largely unknown. This is further complicated by the presence of somatic alterations in normal tissue [Martincorena2015] and blood cells (so called clonal hematopoiesis of indeterminate potential, or CHIP) [Jaiswal2014, Genovese2014, Xie2014]. Although in practice such factors will contribute to an unacceptably high false-positive rate, this simple model sets a *lower* bound on the amount of sequencing required. If such a test were required to be 95% sensitive, this would be equivalent to requiring that in 95% of samples that carry a tumor, at least one fragment would be detected by the test. **Figure 1a** illustrates the sensitivity of a test modeled by this process, as a function of variant allele frequency (VAF) and (unique) sequencing depth.

Aravanis and colleagues have suggested a VAF target for early detection of 0.01% [Aravanis2017] based on clinical data indicating that in early-stage cancer patients the fraction of cfDNA attributable to ctDNA (“tumor fraction” or “ctDNA fraction”) is on or below this order of magnitude. Bettegowda and colleagues were able to detect ctDNA in 47-55% of 182 stage I and II cancer cases [Bettegowda2014]. Newman et al. sequenced plasma from 13 patients with non-small-cell lung cancer (NSCLC) and found that 63% of 11 patients exhibited ctDNA fraction $\leq 0.5\%$, including all four patients with stage I NSCLC; more generally, the ctDNA fraction rose with tumor volume [Newman2014]. More recently, Phallen and colleagues sequenced an 81-kb region in cfDNA to an average unique coverage of 6,182X in 138 stage I/II solid tumor patients [Phallen2017]. **Figure 2a** illustrates the distribution of tumor VAF observed in their stage I/II patients: 50% of stage I and nearly 30% of stage II patients had no mutated ctDNA observed. In the remainder, most patients had maximal ctDNA VAFs between 0.1% and 1%; however, a steep cutoff is visible in the plot below 0.1%.

The binomial model for ctDNA detection suggests that the cutoff visible in **Figure 2a** may arise from the underlying assay having insufficient depth to recover rarer alleles. **Figure 2b** shows patients sequenced by Phallen et al. by their *lowest* VAF allele, as a function of the unique coverage achieved on each patient. Also illustrated is the 95% sensitivity bounds derived from the binomial model in Figure 1a; if the models held, we would expect that these lines would represent the lower bounds on the data. Indeed, this is observed: no patients appear below the 95% sensitivity limit, and the handful of patients with alleles detected at VAF below 0.05% also have higher sequencing depth, tracking with the model predictions. Thus, published data suggests that a substantial fraction of patients (perhaps 50% of stage I cancer patients) have tumor-derived alleles present below 0.1% that have not been sequenced deeply enough to detect. This model thus suggests a unique coverage of 30,000x as a target for an early detection test (yielding 95% sensitivity for one read of a single allele at 0.01%).

We have so far assumed that it is possible to arbitrarily increase sequencing depth as required to detect rare alleles. Also illustrated in **Figure 1a** is the genomic input required for the test to

achieve this level of unique coverage (assuming 100% process efficiency from DNA extraction through sequencing): it is not possible to have 30,000 unique reads without 30,000 unique copies of the locus being sequenced. **Figure 2c** illustrates the amount of cfDNA that can be expected from healthy individuals and those with early stage cancer, drawing from data in [Phallen2017]. To interpret these data, it is critical to consider the context in which an early detection test would be run: in a mostly healthy population, as a test with an expectation of having a result delivered. Consequently, it is appropriate to consider summary statistics of the healthy distribution; since cancers are rare, most patients will be healthy. Furthermore, the mean and median are not the most useful measurements: if a test failure rate less than 5% (i.e., that 95% of prescribed tests successfully return a result) is required, this implies that the minimum amount of DNA acceptable for the test can be no higher than the 5th percentile of the population distribution. This is illustrated in **Figure 2c**: although the mean healthy individual has a plasma cfDNA concentration of 6.6 ng/mL, an individual at the 5th percentile has 2.3 ng/mL, and it is this concentration that determines the practical limitation for assay sensitivity.

Based on the minimum input established in **Figure 1a** by sampling considerations, and the population distribution from **Figure 2c**, we can now compute the total amount of plasma or blood input required for an early detection test using this approach: assuming 100% process efficiency (every molecule of cfDNA in the collection tube makes it to the sequencer), 90 ng of cfDNA are required for 95% sensitivity to detect 0.01% VAF alleles. We can expect at least 2.3 ng cfDNA/mL of plasma and therefore must collect at least 39.1 mL of plasma from each patient; since plasma constitutes ~55% of total blood volume, we must collect at least 71.1 mL of blood per patient. However, 100% process efficiency is unrealistic: sequencing involves a number of lossy steps (extraction, library preparation, target capture, sequencer loading, etc.). **Figure 1b** illustrates the constraint imposed by cfDNA concentration and process efficiency, modeling sensitivity at various VAF levels as a function of blood input volume and assuming 50% process efficiency. Even high total process efficiencies of 25-50% imply a total blood collection volume of 150-300 mL -- a substantial fraction of the approximately 5000 mL of total blood in an adult!

Thus, analysis of a binomial model that assumes constant VAF suggests that detecting alleles at 0.01% VAF with 95% confidence would require 150-300mL blood collection with 30,000x *unique* depth of sequencing coverage. Even if such amounts can in theory be collected over repeat blood draws, the logistics and limits of human physiology make the approach impractical for a population screening test. Alternative approaches that reduce the input requirements are therefore required.

Broad mutation-detection panels require infeasibly high sequencing bandwidth and are limited by somatic heterogeneity

Data from The Cancer Genome Atlas (TCGA) and other large-scale cancer genomics projects have revealed that most tumors harbor hundreds to thousands of somatic variants [Kandoth2013], a subset of which are highly recurrent across patients and tumor types [Ciriello2013]. This affords an opportunity to improve on the binomial sampling bounds derived in the previous section: If it is acceptable to detect *any* of a large number of tumor-derived mutations, instead of one or a few mutations, then for the purposes of the binomial model the effective VAF becomes less than or equal to the sum of the individual VAFs (with equality achieved if the presence of each variant is independent of the others). That is, to detect any of ten independent VAF=0.01% mutations has the same sampling difficulty as detecting a single 0.1% variant. Because this strategy is able to extract information from multiple loci, it gets more out of each genome, reducing input requirements as well as depth requirements.

In the best case, if all mutations had the same frequency and all were independent, a panel capturing N potential variant loci would have N times the power (and require $1/N$ the genomic input) of a single-site assay. However, real-world data from tumor sequencing suggests that this best-case scenario is not likely to be achieved in practice, because mutation frequency distributions are not uniform. This then requires additional sequencing bandwidth. Newman and colleagues developed a protocol (CAPP-Seq) to sequence recurrently mutated regions in NSCLC and modeled its analytical performance versus whole-exome and whole-genome sequencing. **Table 1** shows the results of their model: more extensive panels (e.g., whole-genome vs CAPP-seq) were able to capture more mutations, but this actually reduced sensitivity (even when given additional sequencing bandwidth). Thus, although panel expansion may offer a way to reduce input requirements the non-uniform distribution of mutation frequencies implies that dramatically more sequencing depth is required to recover comparable sensitivity to assays focused on high-frequency sites.

Furthermore, a well understood problem in next-generation sequencing (NGS) is that per-base error rates in NGS reads are substantially higher than 0.01%, often in the 0.1-0.5% range [Minoche2011]. A variety of techniques utilizing molecular barcoding and computational postprocessing have been developed to reduce this error rate [Schmitt2012, Newman2016, Phallen2017] but with the tradeoff of requiring higher read depth: individual molecules are sequenced more than once, with unique molecular identifiers used to group multiple reads of a single read and correct errors by consensus. This oversampling therefore inflates the total sequencing depth (as opposed to unique sequencing) required to perform a ctDNA assay: in

TEC-Seq [Phallen2017], the reported interquartile range of oversampling was 4.77- to 9.38-fold, with a mean of 9-fold. Therefore, error correction may conservatively increase sequencing requirements fivefold.

An assay targeting an ideal (independent, uniform frequency) panel of mutations with cumulative VAF of 0.01% would require 30,000x unique coverage or, optimistically 150,000x raw depth with error correction. However, as established in the previous section, such a panel would require too much blood to be feasible. An ideal panel with 0.1% VAF might only require 3,000x unique coverage, and therefore 15-30mL blood collection; however, the CAPP-Seq experience suggests that in fact the total amount of sequencing required may be tenfold higher than that of the small panel -- equivalent to 1,500,000x coverage of the smaller panel! Thus, while panel expansion may constrain input requirements, it can only do so at an exorbitant sequencing cost.

Beyond mere technical difficulty, expanding the region of interest poses a fundamental biological challenge. Recent data exploring the mutational landscape in healthy individuals has revealed mutations in cancer driver genes at frequencies comparable to the VAF ranges being explored in early detection tests. For example, Martincorena and colleagues assayed healthy sun-exposed eyelid skin and found a dense landscape of low-level somatic variants in cancer genes such as *NOTCH1/2/3*, *TP53*, *FGFR3*, *FAT1*, and *RBM10* [Martincorena2015]. Clonal hematopoiesis associated with aging [Jaiswal2014, Genovese2014, Xie2014] also represents a major confounder of mutational heterogeneity in ctDNA. Results summarized in **Table 2** show that 10-40% of individuals carry low-level somatic mosaicism in cancer-associated genes. The presence of such variants in healthy individuals is thus a significant challenge for early cancer detection by mutation analysis because many such variants are expected to be present in the blood of older individuals (the intended population for cancer screening) and since hematopoietic stem cells dominate cfDNA signals [Snyder2016]. Moreover, the binomial limits derived above are highly optimistic - many of the detected alleles may need to be filtered out and more than one mutation will likely be required to mitigate false positives.

It is increasingly appreciated that many somatic variants are present prior to transformation and malignant outgrowth, reflective of the relatively long time period leading up to cancer formation [Tomasetti2013, Sottoriva2015]. The presence of such variants is likely to complicate the task of identifying the presence of a tumor. As the spectrum of somatic variation in a given healthy organ tends to resemble the spectrum found in tumors originating in that tissue [Hoang2016], it may be difficult to distinguish which mutations are truly indicative of a tumor's presence, and a difference in the rate at which tumor cells and the surrounding normal tissue shed cfDNA would be needed to extract signal [Hori2011].

Tumor initiating mutations will be present in all cells of the tumor (clonal), and therefore should be more readily detectable in ctDNA than subclonal mutations would be [Izumchenko2015], and may be more specific than passenger mutations that may occur in the peripheral normal. Indeed, established tumor initiating mutations and presumed cancer driver genes that are

recurrently mutated across patient tumors have formed the basis for mutational-based ctDNA assays. However, even when such canonical driver mutations (*KRAS*, *PIK3CA*, *BRAF*) are present in the tumor tissue, they may not be detectable in corresponding ctDNA using the most sensitive techniques (ddPCR or TEC-seq) [Phallen2017], thus highlighting the challenge of mutation-only based strategies. While the landscape of somatic alterations in solid tumors have been catalogued, defining the *functional* drivers of individual cancers remains a challenge due to the high degree of mutational heterogeneity and the fact that drivers are both cell-of-origin and context dependent. Further, it is well established that multiple alterations are necessary for tumor development and that they need not be exclusively genomic [Vogelstein2013]. As such, biomarkers of malignant transformation in diverse histologies remain elusive and a better understanding of the earliest events of tumorigenesis could inform early detection efforts.

Health Economics of Mutation Calling for Early Detection

Early screening tests for cancer are not novel to the medical system; existing tests provide a benchmark for the costs that payers are willing to tolerate for screening in the general population. As an example, in the USA, Medicare agreed to reimburse the Cologuard fecal screening test for colorectal cancer at \$502 every three years [Pickhardt2016]. Assuming the physical limitation of sample volume and interpretive limitations of somatic heterogeneity could be successfully overcome, reimbursement may still pose a fundamental economic threat to the viability of mutation-based ctDNA assays for early detection. This is likely to be true even for a broad test intended to screen for most common solid cancers, for which the appropriate benchmark would be the *combined* costs from each individual traditional screen.

It is possible to estimate the costs of running a mutation detection assay using simple assumptions about test parameters. **Table 3** estimates the input volume required and sequencing cost of a mutation-calling ctDNA-based early detection assay under highly conservative assumptions. Under these assumptions, it is evident that reasonably sized panels for tumor liquid biopsy are commercially possible with present-day technology: 15 mL sample draw (compared to 20 mL reported in a real-world validation [Lanman2015]), with sequencing cost in the tens to hundreds of USD. However, early detection appears infeasible: while small panels (e.g., the 81-kb TEC-Seq panel) have achievable sequencing costs, their input volumes are likely prohibitive. In contrast, larger panels (e.g., the 2-Mb panel reported by Razavi and colleagues [Razavi2017b]) have sequencing costs alone that are nearly tenfold the total reimbursement for existing screening tests -- even without considering the likely increase in sequencing required to exploit lower-*VAF* mutations in the expanded panel or the need to perform repeat assays to improve assay sensitivity (or evaluate a change in ctDNA abundance). While sequencing costs fell dramatically in the early days, in recent years such cost reductions have slowed dramatically, and technological advances have focused on physical or synthetic long-read technologies that do not aid sequencing of short (~170 bp) cfDNA fragments. Consequently, it is not clear that the large cost gap between value-based reimbursement and sequencing cost for ultra-deep ctDNA mutation calling could be bridged in the near term.

Summary: Clinical Performance of Early Detection by Mutation Calling

Although the challenges outlined in the previous sections are formidable, if it were possible to overcome them to significantly improve on current screening efforts, there may yet be value in pursuing this approach. However, clinical data (summarized in **Table 4**) casts doubt on this. Phallen and colleagues analyzed 138 individuals with stage I/II breast, colorectal, lung, or ovarian cancer using the TEC-Seq ctDNA 81-kb panel, at an average effective (unique, error-corrected) coverage of 6,182x [Phallen2017]. This method achieved a clinical sensitivity (fraction of clinically affected patients called positive by the assay) of 59-71% depending on disease stage, with tumor fraction and sequencing coverage the probable limiting factors.

Analysis of late-stage cancer patients, in whom ctDNA burden is higher [Bettegowda2014], offers a means to probe the limits of mutation detection. Razavi and colleagues presented data from 124 metastatic breast, lung, and prostate cancer patients using a sequencing panel covering 2.1 Mb at an average effective depth of 3,000-4,000x, achieving a clinical sensitivity of 89% [Razavi2017b]. While the 89% sensitivity achieved in metastatic patients is much higher than that achieved by similar sequencing depth (but smaller ROI) in early stage patients, it compares poorly to the performance of previously described early detection assays. Menon and colleagues described the prospective clinical validation of multimodal screening for ovarian cancer, combining measurement of serum CA-125 with ultrasound. The trial evaluated 50,078 individuals, of whom 47 had cancer within one year of trial completion (48% of cancers in stage I/II) [Menon2009]. The screening approach achieved 89.4% clinical sensitivity (42 TP detected with 5 FN discovered in 1-year followup). Imperiale and colleagues evaluated the performance of a fecal test for colorectal cancer, combining detection of occult blood with measurement of DNA markers in 9,989 individuals, achieving a clinical sensitivity of 92.3%, benchmarked relative to colonoscopy (which remains the gold standard for screening) [Imperiale2014]. Thus even given an unrealistic advantage (assuming ctDNA burdens equivalent to those present in metastatic cancer patients), the ctDNA-based mutation detection approach is unable to reach the sensitivity of existing assays for early cancer detection, making the exorbitant cost of sequencing difficult to reconcile.

Conclusions

The advent of high-throughput sequencing, coupled with the demonstration that ctDNA can be detected non-invasively in plasma at various stages of malignancy, has led to significant investment in mutation-based ctDNA assays for early detection. However, there has yet to be a systematic exploration of the statistical and biological limits of such an approach. Here we have demonstrated that intrinsic biological characteristics likely set a performance bound for mutation detection that is insufficient to achieve high sensitivity and specificity in early-stage cancer detection using ctDNA detection alone.

The fundamental limitations described by the binomial model of tumor fraction pertain to *quantity* (because tumor-derived alleles are present at very low concentrations in bulk plasma, large volumes of plasma and huge sequencing bandwidth are required to detect even *one* tumor-derived molecule) and the specificity challenges posed by somatic heterogeneity arise from the requirement to interpreting a sample from a single patient at a single time, with no other context. Relaxation of either constraint may admit viable strategies for early detection.

Box 1 summarizes a number of blood-based analytes with potential applicability for early cancer screening. While diverse, these analytes all have in common that they occupy more accessible points in the concentration-quantity space than ctDNA: options exist to find concentrated compartments of tumor-derived material, high-copy-number analytes (e.g., protein and RNA) that are present in measurable quantity even at low concentration. Particularly appealing opportunities arise in the search for signals that are not tumor-derived: if the tumor comprises 0.01% or less of circulating material, perhaps there is relevant signal in the other 99.99%, such as measurement of the body's own immunosurveillance response to the presence of 'foreign' tumor cells. Integration of germline risk information may also boost the predictive power of a phenotype classifier [Shieh2016]. Computational integration of these multi-analyte signals may further provide improved power for phenotype classification [Argelaguet2017].

Furthermore, the screening context offers a unique opportunity to use longitudinal data on single individuals to improve accuracy. As cancer screening tests are typically repeated on a 1-3 year cycle, it may become possible to use not only an individual's most recent test result but also his or her previous test results to precisely estimate changes in health status over time. Such a longitudinal testing approach may be able to use individualized data to reduce error by calibrating a predictor to individual baselines or known individual patterns of somatic variation and could also incorporate phenomena in the larger medical record. While such systems are current open areas of research, recent results on modeling of electronic medical record data suggest that the integration of structured data (such as diagnostic results) with temporal variables can improve accuracy in practice [Fiterau2017].

It has long been appreciated that early detection and intervention are perhaps the most effective means for reducing cancer related mortality. This significant unmet need has long motivated the search for accurate detectors of early-stage cancer, including potential signals in tumor-derived cell-free DNA. While statistical analysis suggests that sequencing of this ctDNA alone may be insufficient to achieve clinical sufficiency in this task, the presence of additional biological signals in the blood combined with multi-analyte or longitudinal analysis methods motivates further research towards the development of clinically usable early detection protocols.

Acknowledgments

We thank Christina Curtis, Adam Drake, Kate Niehaus, Gabriel Otte, Girish Putcha, and David Weinberg for their suggestions and feedback on the manuscript.

Supplemental Information

Attached is `make_figures.py`, a Python 2.7 script. It contains an implementation of the analytical binomial model as well as the reanalysis performed on the data from [Phallen2017], and is a self-contained package to fully regenerate figures 1 and 2.

References

1. [Aravanis2017] Aravanis AM, Lee M, Klausner RD. Next-generation sequencing of circulating tumor DNA for early cancer detection. *Cell*. 168. 571-574. 2017.
2. [Argelaguet2017] Argelaguet R, Velten B, Arnol D, et al. Multi-Omics factor analysis disentangles heterogeneity in blood cancer. *bioRxiv* 217554 (2017).
3. [Best2015] Best MG et al. RNA-Seq of tumor-educated platelets enables blood-based pan-cancer, multiclass, and molecular pathway cancer diagnostics. *Cancer Cell* 28(5):666-676 (2015).
4. [Best2017] Best MG et al. Swarm intelligence-enhanced detection of non-small-cell lung cancer using tumor-educated platelets. *Cancer Cell* 32(2):238-252. (2017)
5. [Bettegowda2014] Bettegowda C et al. Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci Transl Med* 6(224). 19 Feb 2014.
6. [Blumenthal2016] Blumenthal DT et al. Clinical utility and treatment outcome of comprehensive genomic profiling in high grade glioma patients. *J Neurooncol*. 2016.
7. [Chen2016] Chen H et al. Prospective evaluation of 64 serum autoantibodies as biomarkers for early detection of colorectal cancer in a true screening setting. *Oncotarget* 2016 Mar 29; 7(13):16420-16432.
8. [Cho2014] Cho H, Mariotto AB, Schwartz LM, Luo J, Woloshin S. When do changes in cancer survival mean progress? The insight from population incidence and mortality. *J Natl Cancer Inst Monographs*, 49: 187-198. 2014.

9. [Coffelt2015] Coffelt SB et al. IL-17-producing $\gamma\delta$ T cells and neutrophils conspire to promote breast cancer metastasis. *Nature* 522:345-348 (2015).
10. [Cohen2017] Cohen JD et al. Combined circulating tumor DNA and protein biomarker-based liquid biopsy for the earlier detection of pancreatic cancers. *Proc Natl Acad Sci*. Published online AOP 2017.
11. [Ciriello2013] Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C. Emerging landscape of oncogenic signatures across human cancers. *Nat Genet* 45(10): 1127-1133. (2013)
12. [CNBC2017]
<https://www.cnbc.com/2017/05/11/guardant-raised-360-million-led-by-softbank-for-cancer-detection.html>
13. [Diaz2012] Diaz et al. The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature* 486(7404): 537-40 (2012).
14. [Fiterau2017] Fiterau M, Bhooshan S, Fries J, et al. ShortFuse: Biomedical Time Series Representation in the Presence of Structured Information. In proceedings of Machine Learning for Healthcare 2017.
[http://mucmd.org/CameraReadySubmissions/16_short_fuse_paper\(1\).pdf](http://mucmd.org/CameraReadySubmissions/16_short_fuse_paper(1).pdf)
15. [GarciaMurillas2015] Garcia-Murillas I et al. Mutation tracking in circulating tumor DNA predicts relapse in early breast cancer. *Sci Transl Med*. 7(302). 26 Aug 2015.
16. [Gentles2015] Gentles AJ et al. The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat Medicine*. 2015.
17. [Genovese2014] Genovese G et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *New Engl J Med*. 371:2477-2487 (2014).
18. [GlobeNewsWire2017]
<https://globenewswire.com/news-release/2017/03/01/929515/0/en/GRAIL-Closes-Over-900-Million-Initial-Investment-in-Series-B-Financing-to-Develop-Blood-Tests-to-Detect-Cancer-Early.html>
19. [Hanash2008] Hanash SM, Pitteri SJ, Faca VM. Mining the plasma proteome for cancer biomarkers. *Nature* 452:571-579. (2008).
20. [Hannafon2016] Hannafon BN, Trigo YD, Calloway CL et al. Plasma exosome microRNAs are indicative of breast cancer. *Breast Cancer Res* 18(1): 90. (2016).

21. [Hoang2016] Hoang ML, Kinde I, Tomasetti C, et al. Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing. *Proc Natl Acad Sci* 113(35): 9846-9851 (2016).
22. [Hori2011] Hori SS and Gambhir SS. Mathematical model identifies blood biomarker-based early cancer detection strategies and limitations. *Sci Transl Med* 3(109): 109ra116 1-9 (2011).
23. [Imperiale2014] Imperiale TF et al. Multitarget stool DNA testing for colorectal-cancer screening. *New Engl J Med* 370:1287-1297 (2014).
24. [Izumchenko2015] Izumchenko et al. Targeted sequencing reveals clonal genetic changes in the progression of early lung neoplasms and paired circulating DNA. *Nat Commun* 6:8258 (2015)
25. [Jaiswal2014] Jaiswal et al. Age related clonal hematopoiesis associated with adverse outcomes. *New Engl J Med*. 371:2488-98 (2014).
26. [Jett2014] Jett JR, Peek LJ, Fredericks L, Jewell W, Pingleton WW, Robertson JFR. Audit of the autoantibody test, EarlyCDT-Lung, in 1600 patients: an evaluation of its performance in routine clinical practice. *Lung Cancer*. 2014 January. 83(1):51-55.
27. [Kandoth2013] Kandoth C et al. Mutational landscape and significance across 12 major cancer types. *Nature* 502: 333-339 (2013)
28. [Lanman2015] Lanman RB 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(10):e0140712 (2015).
29. [Li2017] Li W et al. 5-hydroxymethylcytosine signatures in circulating cell-free DNA as diagnostic biomarkers for human cancer. *Cell Research* 27:1243-1257 (2017).
30. [LimaPereira2017] Pereira AAL et al. Clinical utility of circulating cell-free DNA in advanced colorectal cancer. *PLoS One* 12(8):e0183949 (2017).
31. [Martincorena2015] Martincorena I et al. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* 348: 880-886 (2015).
32. [Melo2015] Melo SA, Luecke LB, Kahlert C et al. Glypican-1 identifies cancer exosomes and detects early pancreatic cancer. *Nature* 523(7559): 177-182. (2015)
33. [Menon2009] Menon U et al. Sensitivity and specificity of multimodal and ultrasound screening for ovarian cancer, and stage distribution of detected cancers: results of the

prevalence screen of the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS). *Lancet Oncol* 10:327-340 (2009).

34. [Minoche2011] Minoche AE, Dohm JC, Himmelbauer H. Evaluation of genomic highthroughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biology*. 2011;12:R112.
35. [Misale2012] Misale et al. Emergence of KRAS mutations and acquired resistance to anti-EGFR therapy in colorectal cancer. *Nature* 486(7404):532-6 (2012).
36. [Nawroz1996] Nawroz H, Koch W, Anker P, Stroun M, Sidransky D. Microsatellite alterations in serum DNA of head and neck cancer patients. *Nature Med* 2: 1035-1037 (1996).
37. [Newman2014] Newman AM et al. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nature Med*. 2014.
38. [Newman2016] Newman AM et al. Integrated digital error suppression for improved detection of circulating tumor DNA. *Nat Biotechnol* 2016.
39. [Phallen2017] Phallen J et al. Direct detection of early-stage cancers using circulating tumor DNA. *Sci Transl Med* 9. 16 Aug 2017.
40. [Pickhardt2016] Pickhardt PJ. Emerging stool-based and blood-based non-invasive DNA tests for colorectal cancer screening: The importance of cancer prevention in addition to cancer detection. *Abdom Radiol*. 2016 August; 41(8):1441-1444.
41. [Razavi2017a] Razavi P et al. Many cell-free DNA (cfDNA) mutations are derived from clonal hematopoiesis: implications for interpretation of liquid biopsy tests. Poster presented at 2017 annual meeting of American Society of Clinical Oncology (ASCO).
42. [Razavi2017b] Razavi P et al. Performance of a high-intensity 508-gene circulating-tumor DNA (ctDNA) assay in patients with metastatic breast, lung, and prostate cancer. Poster presented at 2017 annual meeting of American Society of Clinical Oncology (ASCO).
43. [Schmitt2012] Schmitt MW et al. Detection of ultra-rare mutations by next-generation sequencing. *Proc Natl Acad Sci*. 109(36):14508-14513. (2012)
44. [Senzel2009] Senzel L, Gnatenko DV, Bahou WF. The platelet proteome. *Curr Opin Hematol* 2009 Sep; 16(5); 329-333.

45. [Shieh2016] Shieh Y, Hu D, Ma L et al. Breast cancer risk prediction using a clinical risk model and polygenic risk score. *Breast Cancer Res Treat* 159(3): 513-525. (2016)
46. [Snyder2016] Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J. Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. *Cell* 164:57-68. (2016)
47. [Sorenson1994] Sorenson GD et al. Soluble normal and mutated DNA sequences from single-copy genes in human blood. *Cancer Epidemiol Biomarkers Prev.* 3: 67-71 (1994).
48. [Sottoriva2015] Sottoriva A et al. A Big Bang model of human colorectal tumor growth. *Nat Genet* 47(3): 209-216. (2015)
49. [Thakur2014] Thakur BK, Zhang H, Becker A, et al. Double-stranded DNA in exosomes: a novel biomarker in cancer detection. *Cell Res* 24(6): 766-769. (2014)
50. [Tomasetti2013] Tomasetti C, Vogelstein B, Parmigiani G. Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. *PNAS* 110(6): 1999-2004. (2013)
51. [Ulz2016] Ulz P et al. Inferring expressed genes by whole-genome sequencing of plasma DNA. *Nat Genet* 2016.
52. [Vogelstein2013] Vogelstein et al. Cancer Genome Landscapes. *Science* 339(6127): 1546-58
53. [Xie2014]. Xie et al. Age related clonal hematopoiesis associated with clonal hematopoietic expansino and malignancies. *NatMed* 20:1472-8 (2014)
54. [Xu2017] Xu YF, Hannafon BN, Zhao YD, Postier RG, Ding WQ. Plasma exosome miR-196a and miR-1246 are potential indicators of localized pancreatic cancer. *Oncotarget* 8(44): 77028-77040. (2017)

Capture panel	Average # NSCLC SNVs in capture region	Sequencing bandwidth used	Minimum VAF with 80% sensitivity to detect 2 mutant alleles
CAPP-Seq	4	1/12 lane	0.01%
WES	218		0.5%
WGS	15,659	1 lane	0.05%
			0.25%

Table 1: Effect of panel size and sequencing bandwidth on ability to detect somatically mutated alleles in plasma. Data from [Newman2014], figure 1d.

Cohort Age	VAF threshold	Fraction of Cohort with VAF within designated range	Citation
>65yr	>10%	10%	[Genovese2014]
Avg 44yr	0.16-5.28%	16%	[Phallen2017]
<50yr	≥0.1%	10%	[Razavi2017a]
>70yr		~40%	

Table 2: Prevalence of somatic variation observed in genes associated with clonal hematopoiesis.

					Sequencing cost by panel size		
	VAF 95% sens	Corrected depth	Raw depth	Input vol. (blood)	<i>TEC-Seq</i> 58 genes 81kb	<i>Razavi et al</i> 508 genes 2,000Kb	<i>WES</i> ~20k genes 50,000Kb
Tumor liquid biopsy	0.1%	3,000x	15,000x	15mL	\$14	\$340	\$8,300
Early cancer detection	0.01%	30,000x	150,000x	150mL	\$140	\$3,400	\$83,000

Table 3: Assay requirements for tumor liquid biopsy and mutation-based early cancer detection. Assumptions as described in the main text:

- No more than 5% of samples may fail because of insufficient cfDNA quantity.
- We require 95% sensitivity to detect one read from any cancer-derived allele, assuming that one is present in the sample.
- 50% process efficiency: half of the cfDNA molecules in the input blood sample are represented in the sequencer output.
- 5x oversampling in sequencing for error correction.
- 100% on-target rate in target enrichment.
- “\$1000 genome” sequencing costs: US\$1000 / (30 x 3 Gbp) of sequencing bandwidth
- Only sequencing costs computed; all other costs (labor, equipment, facilities, depreciation, etc.) accounted at \$0.
- Panel expansion neither reduces input requirements nor increases sequencing requirements.

ctDNA tests						
Cancer type	Stage	Total Samples	Panel size	Avg Raw Depth	Avg Eff Depth	Clin. Sens.
Breast, Lung, Colorectal, Ovarian [Phallen2017]	I / II	N=138	81kb	38,589x	6,182x	59-71%
Breast, Lung, Prostate [Razavi2017b]	metastatic	N=124	2100kb	60,000x	3,000-4,000x	89%
Non-ctDNA-based tests						
Cancer Type	Stage	TP+FN / Total	Modality	FN Set	Clin. Sens.	
Ovarian [Menon2009]	48% I/II	42+5 / 50,078	CA-125 + ultrasound	1yr followup	89.4%	
Colorectal [Imperiale2014]	93% I/II/III	60+5 / 9,989	FIT + DNA	Colonoscopy	92.3%	

Table 4: Comparison of ctDNA-based cancer detection assays versus conventional screening tests.

Category	Analyte	Explanation	References
Tumor-derived material	Tumor-derived proteins and RNAs	Protein and RNA are present at higher copy number than DNA, potentially enabling detection via nonzero count even at low concentration.	[Cohen2017]
	Exosomes / microvesicles / circulating tumor cells or cell clusters	Tumor-derived bodies may contain macromolecular markers in sufficiently high concentration to detect efficiently.	[Thakur2014, Melo2015, Hannafon2016, Xu2017]
Immuno-surveillance	Platelets	Platelets contain proteins and RNAs that function in immune signaling pathways, whose composition may vary in the presence of cancer.	[Senzel2009, Best2015, Best2017]
	Cytokines, antibodies, and other immune signaling molecules	Differential cytokine and autoantibody production have been observed in cancer patients as a component of the immune response to a tumor.	[Hanash2008, Jett2014, Chen2016]
	Immune cell subpopulations	Differential composition of immune cells may indicate the presence of cancer and inform prognosis.	[Coffelt2015, Gentles2015]
Tumor microenvironment and host response	cfDNA	Patterns in cfDNA beyond sequence variation, including epigenetic modifications and fragmentation, may serve as a marker for host gene expression.	[Snyder2016, Ulz2016, Li2017]

Box 1: Biological components other than ctDNA with potential for cancer screening.

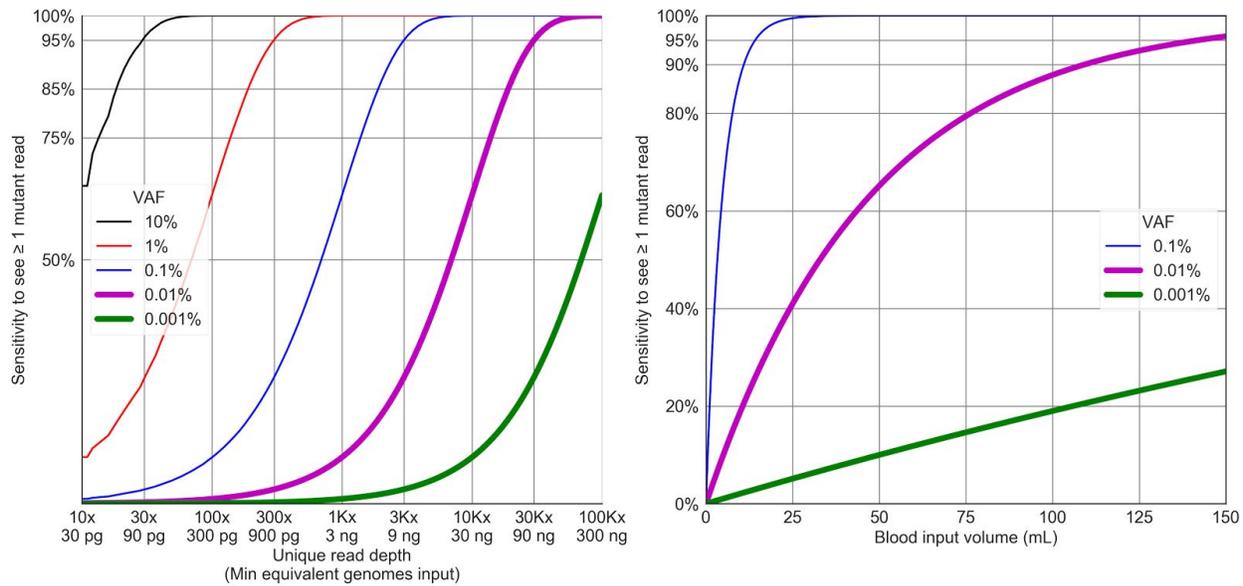


Figure 1: Binomial model for ctDNA sequencing: sensitivity achieved by a sequencing assay to detect 1 mutant allele at a given VAF.

a) Upper bound on sensitivity as a function of sequencing depth; note logarithmic x-axis. Also shown is the minimum amount of unique DNA input required for sequencing, assuming 3pg haploid genome mass and 100% process efficiency.

b) Sensitivity as a function of blood input volume, assuming 2.3ng cfDNA/mL plasma, plasma volume 55% of blood volume, and 50% process efficiency.

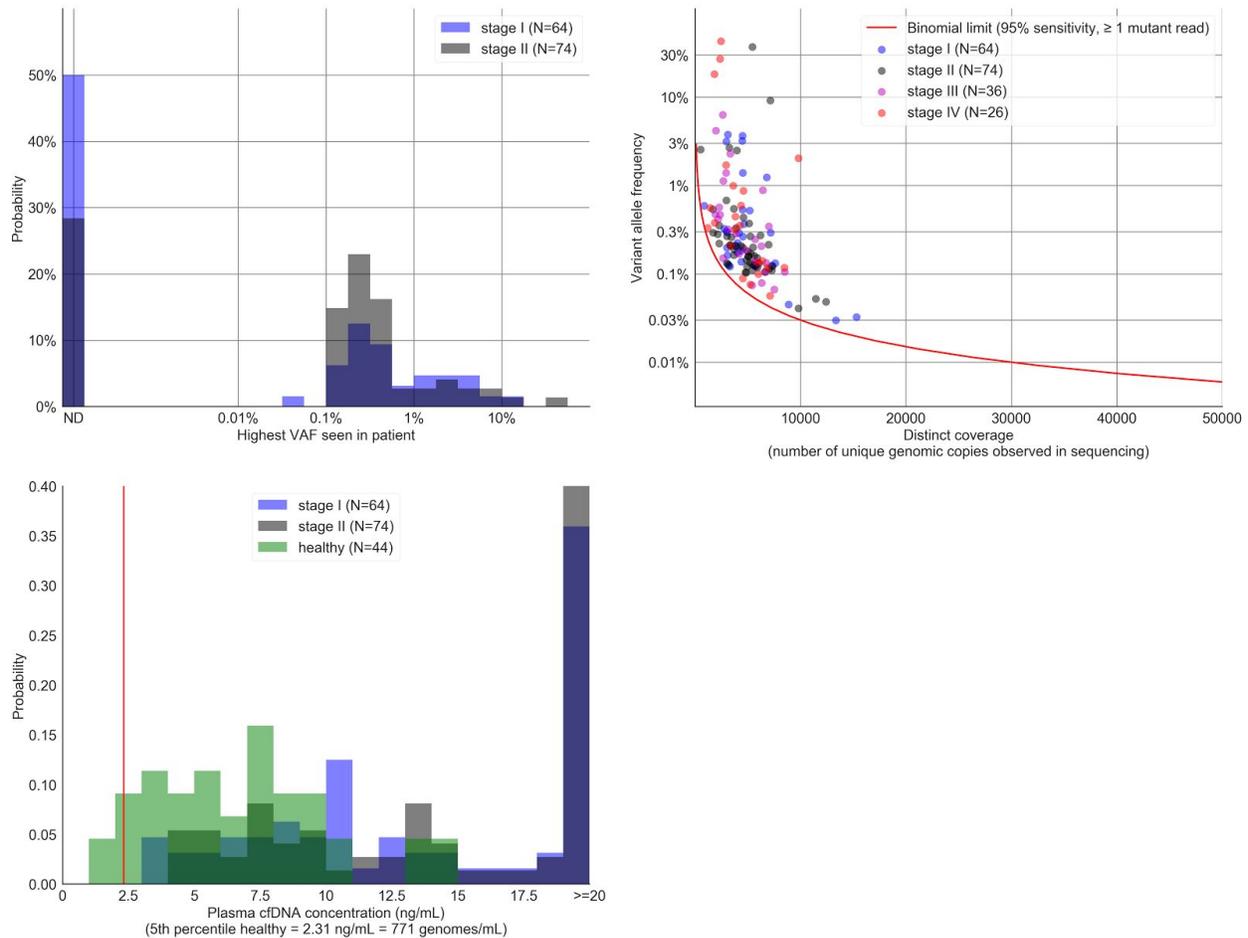


Figure 2: Analysis of ctDNA characteristics from early-stage cancer patients in Phallen et al. [Phallen2017]. Panels clockwise from top left.

a) Highest per-patient VAF observed for any cancer-related variant in stage I/II colorectal, breast, ovarian, or lung cancer patients, as measured by the TEC-Seq protocol. Samples listed as “ND” had no cancer-derived alleles observed. In patients with multiple cancer alleles detected in plasma, the highest VAF is shown.

b) Comparison of observed VAF to binomial model. Dots correspond to the VAF of the lowest-frequency cancer-derived variant detected and the unique depth of coverage for that patient. Red and blue curves respectively show the VAF expected to be detected with 95% confidence by the binomial model in Figure 1.

c) Distribution of cfDNA concentrations observed in healthy individuals and stage I/II cancer patients. Red line is drawn at the 5th percentile of the distribution of healthy individual cfDNA concentration.